Chapter 15

A Systems Biology Approach for Identifying Hepatotoxicant Groups Based on Similarity in Mechanisms of Action and Chemical Structure

Dennie G.A.J. Hebels, Axel Rasche, Ralf Herwig, Gerard J.P. van Westen, Danyel G.J. Jennen, and Jos C.S. Kleinjans

Abstract

When evaluating compound similarity, addressing multiple sources of information to reach conclusions about common pharmaceutical and/or toxicological mechanisms of action is a crucial strategy. In this chapter, we describe a systems biology approach that incorporates analyses of hepatotoxicant data for 33 compounds from three different sources: a chemical structure similarity analysis based on the 3D Tanimoto coefficient, a chemical structure-based protein target prediction analysis, and a cross-study/cross-platform meta-analysis of in vitro and in vivo human and rat transcriptomics data derived from public resources (i.e., the diXa data warehouse). Hierarchical clustering of the outcome scores of the separate analyses did not result in a satisfactory grouping of compounds considering their known toxic mechanism as described in literature. However, a combined analysis of multiple data types may hypothetically compensate for missing or unreliable information in any of the single data types. We therefore performed an integrated clustering analysis of all three data sets using the R-based tool iClusterPlus. This indeed improved the grouping results. The compound clusters that were formed by means of iClusterPlus represent groups that show similar gene expression while simultaneously integrating a similarity in structure and protein targets, which corresponds much better with the known mechanism of action of these toxicants. Using an integrative systems biology approach may thus overcome the limitations of the separate analyses when grouping liver toxicants sharing a similar mechanism of toxicity.

Key words Systems biology, 3D Tanimoto, Protein targets, Meta-analysis, iClusterPlus, Hepatotoxicity, Chemical structure, Mechanism of action, Similarity, diXa

1 Introduction

Systems biology is an interdisciplinary field of study that focuses on complex interactions within biological systems. It uses a holistic approach that aims at integrating data from multiple sources to study the interactions between the components of biological systems and gain a wider understanding of how these interactions give rise to the function and behavior of that system, e.g., a pathway, a cell, etc.

Emilio Benfenati (ed.), *In Silico Methods for Predicting Drug Toxicity*, Methods in Molecular Biology, vol. 1425,
DOI 10.1007/978-1-4939-3609-0_15, © Springer Science+Business Media New York 2016

In other words, instead of taking apart a system and studying each of its individual components, systems biology focuses on integrating all these parts to reach a new level of understanding under the assumption that the whole is more than the sum of its parts.

Omics technologies are particularly useful for this purpose since they cover a large part of the changes in a certain part of the system, such as the transcriptome, the proteome, or the metabolome, thereby aiding the systems biology approach. However, despite the vast amount of information obtained from omics techniques, single omics analysis still does not always provide sufficient information to understand the behaviors of, for example, a cellular system. Therefore, a combination of multiple omics analyses and/or other data sources, the multi-omics (or multi-data source) approach, is needed to acquire a more precise picture of a system [1–5]. Combining multiple data types also has the advantage of being able to compensate for missing or unreliable information in any of the single data types and decreases the likelihood of false-positive findings.

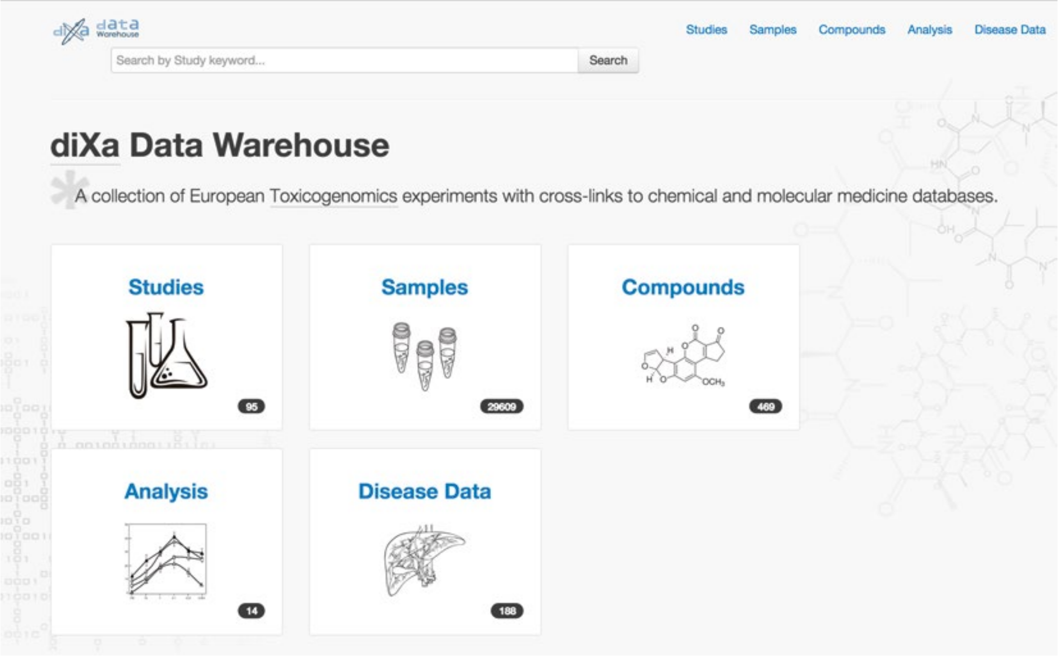
In the field of hepatotoxicity, systems biology approaches are also receiving much attention [6–11]. Given the liver's vital role as a detoxification organ, it is not surprising that hepatotoxicity is the most prominent adverse reaction against drugs. As a result many newly developed candidate drugs fail in preclinical or clinical trials which is associated with a huge financial drain considering that the costs to develop a fully approved drug are around $800 million [12]. Failure to pick up hepatotoxicity in early stages is also contributable to the idiosyncratic nature of many adverse reactions, i.e., unusual individual reactions with very low frequency likely associated with differences in genetic make-up between individuals [13]. New screening methods, able to detect (idiosyncratic) drug-induced liver injury in the early stages of the research process, represent an important step toward efficient new drug development. Despite their poor predictive accuracy, animal models are still considered the gold standard toxicological approach for evaluating chemical toxicity and contribute substantially to the high costs involved in drug development [14]. In vitro systems are therefore increasingly studied with the ultimate goal of replacing animal models. Because of the time-saving nature and practicality of such systems, they are especially well suited to study drug metabolism, measure enzyme kinetics, evaluate toxicity mechanisms, and examine dose–response relationships using systems biology approaches [15]. The systems biology "map" of a hepatotoxic compound of interest may serve as a profile of its (idiosyncratic) toxicological mechanism. Studying large compilations of such compound profiles can thus assist in finding groups of compounds with similar (toxicological) mechanisms of action by comparing profiles and thereby assist in the early identification and elimination of compounds with a potential hepatotoxic effect.

In this chapter we will demonstrate a systems biology approach focused on compiling compound profiles from multiple data sources in order to group toxic compounds based on similarity. There are many data types available which can be used to obtain such similarity measures. Here, transcriptomics and proteomics data are of particular interest. While such omics data are excellent sources to explore the biological signaling cascades involved in hepatotoxic responses, including sources that focus more on the chemical similarities of the compounds may contribute significantly to the grouping of compounds with comparable hepatotoxic mechanisms. Given the crucial role of chemical structures with respect to xenobiotic metabolism in the liver, quantifying the chemical similarity of molecules is a very active field of research. In our multi-data source systems biology approach, we will therefore focus on a combination of these two approaches. A test data set will be used to illustrate an integrative analysis approach of a transcriptomics analysis and two chemical structure-based analyses. These three analysis approaches will first be explained in more detail separately. They involve a chemical structure similarity analysis based on the 3D Tanimoto coefficient, a chemical structure-based protein target prediction analysis, and a comprehensive transcriptomics meta-analysis. A hierarchical clustering-based grouping of the analysis results will be used to discuss the limitations of the individual methods by comparing the outcome with the known mode of action as described in literature. The multi-omics tool iClusterPlus will subsequently be presented as a means of overcoming these limitations and integrating multiple sources of information to improve grouping of similarly acting hepatotoxic compounds.

2 Data Set

To demonstrate the application of multisource data analysis on hepatotoxicity data, we queried the Data Infrastructure for Chemical Safety Assessment (diXa) data warehouse [16]. diXa is a recently created robust and sustainable infrastructure designed for storing toxicogenomics data. The warehouse is designed to store any type of omics data for every disease of interest and currently mostly contains transcriptomics data on hepatotoxicants and nephrotoxicants. The warehouse is connected to a portal with links to chemical information and molecular and phenotype data. diXa is publicly available through a user-friendly web interface, and new data can be readily deposited into diXa (http://wwwdev.ebi.ac.uk/fg/dixa/index.html, Fig. 1).

A selection of studies stored within diXa was downloaded to present as a use case in this chapter. The selection was based on an initial exploration of the data sets where we set out to include data covering a wide range of experimental conditions (several doses and exposure times, in vitro and in vivo studies) and multiple species (rat and human). To improve data comparability, only studies using the same microarray platform (Affymetrix) were considered. Gene annotations were adjusted to their corresponding orthologues between species where needed. Using these criteria, nine studies were selected covering a total of 33 compounds as shown in Table 1.



diXa Data Warehouse

A collection of European Toxicogenomics experiments with cross-links to chemical and molecular medicine databases.

Fig. 1 The diXa data warehouse web portal provides immediate access to a wide range of transcriptomics studies

Table 1
Overview of studies included in analysis and the full list of hepatotoxic compounds

Project	Species	In vitro/in vivo	Cell/tissue type
carcinoGENOMICS	Homo sapiens	In vitro	HepaRG
	Homo sapiens	In vitro	HepG2
	Rattus norvegicus	In vitro	Primary rat hepatocytes
DrugMatrix	*Rattus norvegicus*	In vitro	Primary rat hepatocytes
	Rattus norvegicus	In vivo	Liver tissue

(continued)

Table 1
(continued)

Project	Species	In vitro/in vivo	Cell/tissue type
Predictomics	Homo sapiens	In vitro	HepG2
TG-GATEs	Homo sapiens	In vitro	Primary human hepatocytes
	Rattus norvegicus	In vitro	Primary rat hepatocytes
	Rattus norvegicus	In vivo	Liver tissue
Hepatotoxic compounds			
1-Naphthyl isothiocyanate	Cyclophosphamide	Gemfibrozil	Phenobarbital
Acetaminophen	Danazol	Ketoconazole	Pirinixic acid
Aflatoxin B1	Diclofenac	Lomustine	Simvastatin
Allyl alcohol	Doxorubicin	Methapyrilene	Sulindac
Amiodarone	Ethanol	Nifedipine	Tamoxifen
Azathioprine	Ethinyl estradiol	Nimesulide	Tetracycline
Carbon tetrachloride	Fenofibrate	*N*-nitrosodimethylamine	Tolbutamide
Clofibrate	Fluphenazine	Pemoline	Valproic acid
Clomipramine			

Elaborate descriptions of all studies can be found in the diXa data warehouse (http://wwwdev.ebi.ac.uk/fg/dixa/index.html)

3 Tanimoto Similarity Score

Structural similarities between compounds may reflect similar mechanisms of action. Quantifying the similarity of two molecules is therefore a key concept in cheminformatics and pharmaceutical research. Although a close similarity between compounds can never guarantee an overlap in the mechanism of action, there is a strong correlation between the presence of certain structural subunits in a molecule and the eventual biological effect, which is a relationship that is often explored during the development of new pharmaceutical compounds. The Tanimoto coefficient [17] is a frequently used measure of chemical similarity and will be applied here to focus purely on the overlap in chemical properties of the compounds in the test data set.

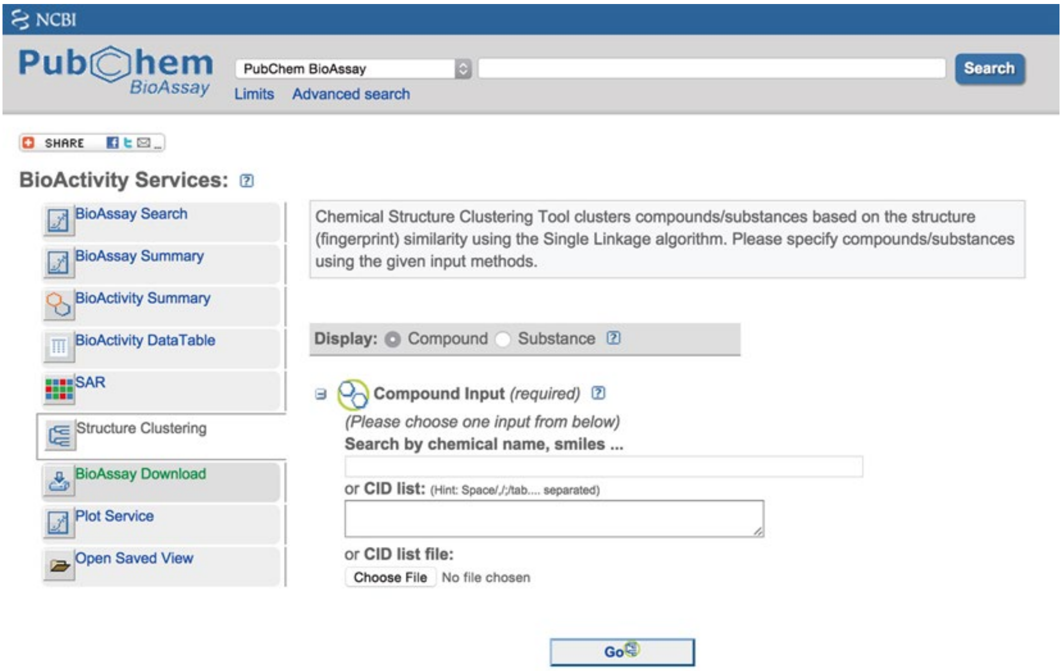
3.1 Tanimoto Coefficient Procedure

Calculation of Tanimoto coefficient similarity scores can be performed in PubChem, which is an open repository for small molecules and their experimental biological activities [17]. Generating

Tanimoto scores is a very straightforward procedure and requires a list of compounds (compound name, PubChem compound database identifier (CID)) which can be uploaded. Subsequently structural similarity data will be calculated between each pair of compounds (https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=clustering, Fig. 2). This resulting structure similarity matrix is then clustered using the single-linkage clustering algorithm.

The structural similarity in PubChem is either based on the Tanimoto score calculated from the 2D structure fingerprint or the 3D shape/feature similarity [17, 18]. The 2D structure fingerprint is based on an ordered list of binary substructures (i.e., fragments of a chemical structures) for chemical structures, in which each substructure is counted as either present or not present in the compound under investigation (e.g., an atomic element count, a type of ring system, atom pairing, atom neighbors, etc.). These fingerprints are used by PubChem for similarity neighboring and similarity searching [17].

A defining characteristic of 3D similarity methods, compared to 2D methods, is that they are applied at a conformer level instead of a compound level, thereby making it possible to consider the various distinct molecular conformations a compound can adopt in 3D space which may have biological relevance [19]. PubChem3D makes a distinction between two 3D similarity measures, i.e.,



Chemical Structure Clustering Tool clusters compounds/substances based on the structure (fingerprint) similarity using the Single Linkage algorithm. Please specify compounds/substances using the given input methods.

Display: ● Compound ○ Substance

Compound Input *(required)*

(Please choose one input from below)

Search by chemical name, smiles ...

or CID list: (Hint: Space/,/;/tab.... separated)

or CID list file:

Choose File No file chosen

Fig. 2 The PubChem chemical structure clustering tool which generates a clustering dendrogram based on calculated Tanimoto scores (2D and 3D) for any list of compounds

shape-Tanimoto (ST) and color-Tanimoto (CT). The ST score is a measure of shape similarity, while the CT score quantifies the similarity of 3D orientation of functional groups or features by checking the overlap of fictitious "color" atoms which represent the six functional group types: hydrogen-bond donors, hydrogen-bond acceptors, anion, cation, hydrophobes, and rings. The ST and CT similarity metrics attempt to cover key aspects important for locating chemical structures that may have similar biological activity. In other words, the ST helps to identify compounds that can adopt a particular 3D shape (e.g., of a neurotransmitter bound in a particular conformational orientation in a postsynaptic membrane protein pocket), while the CT helps to identify compounds with similar 3D orientation of molecular features (e.g., necessary for making a hydrogen or ionic bond interaction of a neurotransmitter with its receptor). The assumption is then that compounds with highly similar 3D shape and feature orientations may also display similarities in their biological activity [19].

Given the importance of biological activity with respect to hepatotoxicity, in this chapter we will focus on the 3D Tanimoto scores. CID identifiers of the 33 compounds in our test data set were retrieved from PubChem, and 3D Tanimoto scores were calculated using the default options of a combined shape (ST) and feature (CT) similarity score (optimized for shape), which was followed by a clustering analysis (see paragraph 6).

4 Protein Target Analysis

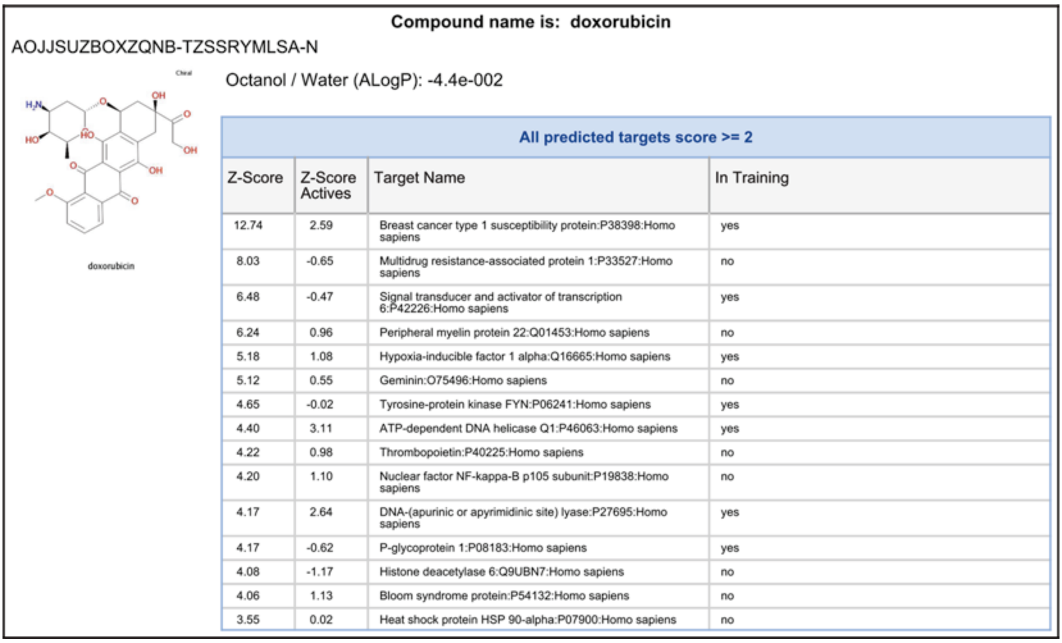
Biological relevance and investigation of mode of action require an understanding of the proteins to which the compounds bind. Based on the chemical structure of the compounds, we can predict their interaction partners (protein targets) in an organism. This is done by comparing the structure of the compound to large curated literature-based databases of known compound–protein interactions such as DrugBank, ChEMBL, the Human Metabolome Database, and the Therapeutic Target Database [20–23]. In this chapter, we use the data in the ChEMBL database release 17 containing approximately 12 million data points [24, 25].

4.1 Protein Target Procedure

A multi-category naive Bayes statistical model trained on ChEMBL database release 17 was used for target prediction [25]. The compound structural features were encoded using extended-connectivity fingerprints with a diameter of six covalent bonds (ECFP6) as implemented in Pipeline Pilot (version 8.5, Accelrys Software Inc.) [26]. Target classes were limited to single protein targets with at least 30 active compounds (to ensure a robust model). Active was defined as having an activity better than 10 μM where the activity type was restricted to Ki, Kd, IC50, AC50, or EC50. In total

690,853 data points were used to construct the model. A multicategory model was then built for each of these proteins; herein relevant molecular features correlated to activity were identified by comparing the structure of actives per protein versus all of the other compounds (assumed inactive). Subsequently, each compound was scored with all 1282 models, and a ranked list of up to the top ten predicted protein targets for each compound was generated.

However, due to large differences in available data points per target (e.g., adenosine A2A receptor versus solute carrier organic anion transporter 1B1) and differences in average compound size per target (e.g., metabotropic glutamate receptors versus thrombin), the raw Bayesian score can differ significantly per protein target (per model class). To make the scores comparable, they were standardized in the form of *z*-scores [27]. The score per compound–protein pair was obtained for predictions by subtracting the mean score for the protein considered from the raw score and dividing this over the standard deviation for that protein (e.g., [1]). To obtain these values, after model training, the model was used to score all compounds in ChEMBL release 17 (1.3 million compounds). From this, a mean score per target and standard deviation per target were derived. Similarly, the mean score and standard deviation of compounds known to be active on a protein were calculated. After model predictions, targets with a standard score ≥2 were considered as a significant protein target for the



Compound name is: doxorubicin

AOJJSUZBOXZQNB-TZSSRYMLSA-N

Octanol / Water (ALogP): -4.4e-002

All predicted targets score >= 2

Z-Score	Z-Score Actives	Target Name	In Training
12.74	2.59	Breast cancer type 1 susceptibility protein:P38398:Homo sapiens	yes
8.03	-0.65	Multidrug resistance-associated protein 1:P33527:Homo sapiens	no
6.48	-0.47	Signal transducer and activator of transcription 6:P42226:Homo sapiens	yes
6.24	0.96	Peripheral myelin protein 22:Q01453:Homo sapiens	no
5.18	1.08	Hypoxia-inducible factor 1 alpha:Q16665:Homo sapiens	yes
5.12	0.55	Geminin:O75496:Homo sapiens	no
4.65	-0.02	Tyrosine-protein kinase FYN:P06241:Homo sapiens	yes
4.40	3.11	ATP-dependent DNA helicase Q1:P46063:Homo sapiens	yes
4.22	0.98	Thrombopoietin:P40225:Homo sapiens	no
4.20	1.10	Nuclear factor NF-kappa-B p105 subunit:P19838:Homo sapiens	no
4.17	2.64	DNA-(apurinic or apyrimidinic site) lyase:P27695:Homo sapiens	yes
4.17	-0.62	P-glycoprotein 1:P08183:Homo sapiens	yes
4.08	-1.17	Histone deacetylase 6:Q9UBN7:Homo sapiens	no
4.06	1.13	Bloom syndrome protein:P54132:Homo sapiens	no
3.55	0.02	Heat shock protein HSP 90-alpha:P07900:Homo sapiens	no

Fig. 3 Example output from the protein target analysis for the compound doxorubicin, showing compound structure and InChI key and the top 15 protein target *z*-scores and *z*-score actives

compound in question and reflect the enrichment of the score over randomness (i.e., all compounds in ChEMBL release 17) for the specific target of that compound in terms of standard deviations. Likewise *z*-score actives are calculated which show the difference a compound scores on this target compared to the average score of known actives for that protein. An example output for the compound doxorubicin is shown in Fig. 3. For further analyses as presented in paragraph 6, all calculated *z*-scores (significant and nonsignificant) were taken into account.

5 Gene Expression Meta-Analysis

An inherent problem of heterogeneous data sets is the experiment-specific variation which cannot be controlled for in a post hoc analysis. These variations stem from a variety of sources such as the use of different cell culture assays, differences in compound concentration and exposure time, and the use of different species (Table 1). To compensate for such variations, cross-study/cross-platform gene expression meta-analysis is a valid strategy to extract consistent information from a set of individual studies across a wide range of experimental conditions, including in vitro and in vivo data. In fact, combining data from in vitro and in vivo studies on liver carcinogens with gene expression data from human liver cancers was shown to improve carcinogenicity prediction [28]. Meta-analysis has been frequently applied in diseases with complex phenotypes such as cancer [29], Down syndrome [30], and diabetes mellitus type 2 [31]. A meta-analysis approach on hepatotoxicity-associated transcriptomics data can therefore be very valuable given the vast amount of heterogeneous data sets available in literature.

5.1 Meta-analysis Procedure

All experiments in the data set (*see* Table 1) have a case–control design comparing two groups of replicate samples. These groups are denoted as treatment and control, respectively, and constitute a test case. For a test case, the generated chips are normalized with each other using the R/Bioconductor framework.

The normalization accounts for three major influence factors in the hybridization data: background expression, probe binding affinity, and measurement variation. GC-RMA corrects for such effects [32]. In the background correction, it takes into account the GC content of the probe sequences, i.e., the number of G or C nucleotides in the sequence. A higher GC content is associated with a higher binding affinity of the probes due to three instead of two covalent bindings for single nucleotides. GC-RMA contains a position-specific model correcting the binding affinity between probes. Between chips unwanted effects are introduced by RNA extraction, pipetting, temperature fluctuations, hybridization efficiency, and more. To reduce these effects, the quantile normalization is implemented in GC-RMA. Finally probe intensities are summarized into probe set expressions. GC-RMA uses median polish which proposes a linear model of a baseline hybridization with two factors, a probe effect and an array effect [33]. The model is fitted robustly with a median decomposition.

An advantage of the Affymetrix array design is the possibility to calculate a presence tag, i.e., the probability that the corresponding gene is effectively expressed and active in the sample under study. Non-expressed genes confuse the results with low intensities leading to high, unmotivated fold changes. The presence tag, or detection *p*-value, is based on a comparison of raw perfect-match values and corresponding mismatch values. Using a robust Wilcoxon test yields a *p*-value for each probe set which indicates whether or not the perfect-match probe signals are different from the mismatch probe signals and thus allows judging the expression of the corresponding gene.

Necessary for any meta-analysis is the consolidation of the different identifier types, different species, or different arrays [34]. The Ensembl database provides a stable reference for microarray studies (http://www.ensembl.org; version 74) and enables orthologue gene searches to allow for the combination of human and rat data. Since comparability of chip studies is hindered by the total number of probes and preprocessing issues between manufacturers, the analysis in this chapter constrains on Affymetrix arrays for case–control studies. Expression results from the arrays are mapped to Ensembl by the custom chip definition file (CDF) annotations [35].

The computation of gene expressions and presence tags is followed by a gene-wise evaluation of treatment versus control expressions. Expressions are assessed by two criteria: presence and alteration. For the approach of a meta-analysis, as presented in this chapter, the two criteria are condensed into a single score for every gene. The test case score *St* of a gene is computed as follows:

$$S_t = \begin{Bmatrix} \left|\log_2(r)\right|(1-10p) & ,p \le 0.1 \\ 0 & ,\text{else} \end{Bmatrix}$$

Here, *r* is the fold change and *p* is the average detection *p*-value. Thus, the fold change is corrected with its effective expression activity. The gene expression alteration in every study test case *t* is quantified with this score.

For every gene *g*, the scores from compound-specific studies are summarized constituting a gene–compound score *Sgc*:

$$S_{g-c} = \frac{T_g}{T_{g-c}} \sum_{g-c} S_t$$

So we sum up the gene scores over all test cases related to compound *c*. The sum is weighted by the quotient of *Tg* the number of test cases with gene scores divided by *Tgc* the number of test cases with scores for gene *g* and compound *c*. This weight compensates for genes which are not represented on every Affymetrix array, which is, for example, relevant for nonhomologous genes between human and rat. The results are discussed in the next paragraph together with the results of the other two analysis approaches.

6 Results of Individual Data Analysis Approaches

The Tanimoto 3D similarity scores are automatically processed in a clustering analysis, the results of which are shown in Fig. 4a. For comparison purposes the protein target *z*-scores and meta-analysis gene scores were also hierarchically clustered; this is shown in Fig. 4b, c (both Ward's clustering, using the "minimum increase in the sum of squares for error" method). This also allows for a more straightforward comparison of the individual analysis results with the integrative analysis covered in the next paragraph.

If we compare the two analysis methods based on chemical structure, i.e., the Tanimoto similarity scores and protein target *z*-scores, there is a number of subclusters that appear to correspond with certain protein target clusters. However, it is also apparent that the protein target scores tend to cluster into more distinct groups of compounds, whereas the Tanimoto dendrogram

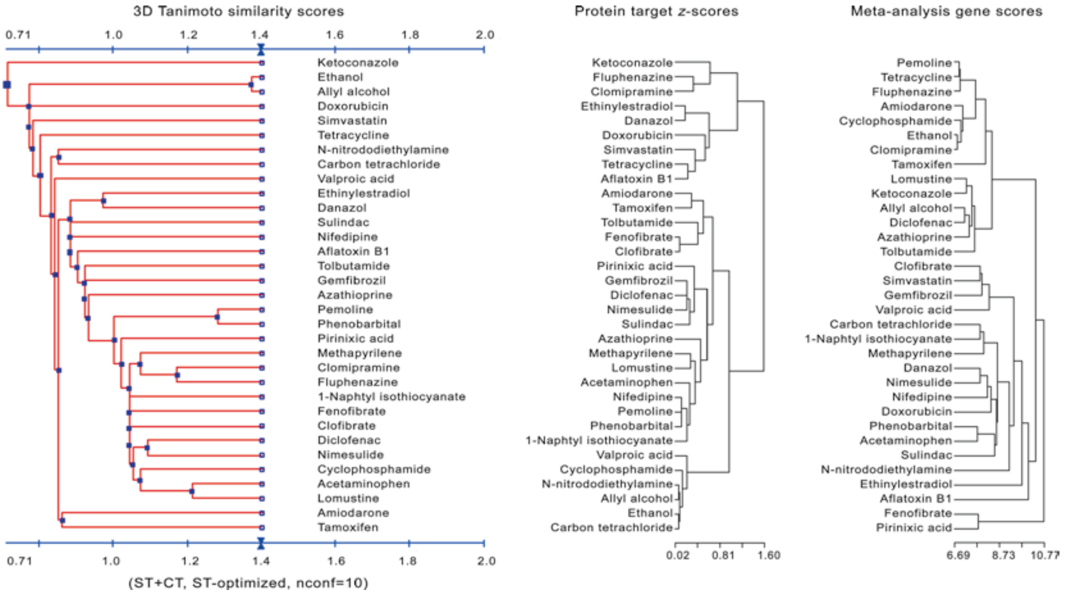


Fig. 4 Clustering dendrograms of the Tanimoto similarity scores (**a**), the protein target *z*-scores (**b**), and the meta-analysis gene scores (**c**)

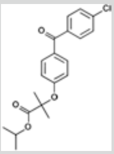
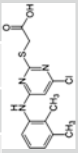
does not form any separate groups with the exception of the duo clusters ethanol/allyl alcohol and amiodarone/tamoxifen. These two duo clusters can also be readily recognized in the protein target dendrogram. Other small clusters which can also be distinguished in the Tanimoto dendrogram include ethinyl estradiol/danazol and pemoline/phenobarbital.

Strong disagreements between analyses become apparent when the meta-analysis is taken into account. Its dendrogram's clusters are quite inconsistent with the Tanimoto and protein score dendrograms, and no immediate overlap can be seen (Fig. 4). So the question arising is which one of these analyses is right? There is of course no straightforward answer to this. If we consider some of the grouped compounds in these dendrograms and compare them with what can be found in literature regarding known mechanism of action, we see that all three analyses cluster compounds as might be expected. We will use the following examples to illustrate this:

(a) Fenofibrate and pirinixic acid.

The meta-analysis suggests that fenofibrate and pirinixic acid induce a similar gene expression response, which indeed makes sense given the fact that they are both peroxisome proliferator-activated receptor alpha (PPARA) agonists [36]. The Tanimoto score analysis does not consider these compounds to be structurally related. Of course structural dissimilarity does not exclude the possibility of having a similar biological effect and vice versa. Small identical substructures in two molecules can already be enough to exert a similar effect even when the overall composition is very different. Conversely, a good example of compounds

Table 2
List of significant protein targets (*z*-score >2) for fenofibrate and pirinixic acid compounds

Fenofibrate							Pirinixic acid		
Structure	Protein (HGNC)	*z*-Score	Protein (HGNC)	*z*-Score	Protein (HGNC)	*z*-Score	Protein (HGNC)	*z*-Score	Structure
	LSS	3.23	GLP1R	2.42	TRPA1	2.28	PTGES	4.82	
	PPARA	2.94	IGFBP3	2.39	CACNA1H	2.24	ALOX5	3.00	
	FFAR2	2.93	PPARD	2.39	P2RY1	2.22	PLA2G7	2.32	
	SCN2A	2.80	AKR1C2	2.39	CTSG	2.21	AKR1C2	2.24	
	SCN10A	2.79	CYP26A1	2.36	ELOVL6	2.20	CSNK2A2	2.16	
	PPARG	2.79	VCAM1	2.34	SRD5A2	2.13	PPARG	2.16	
	GIPR	2.60	ICAM1	2.29	SELE	2.06	CXCR2	2.10	
	ELANE	2.59	UTS2	2.28	MMP14	2.04	CTSA	2.00	

with high structural similarity but entirely different effects are the enantiomers of thalidomide; S-thalidomide is a severe teratogen, while R-thalidomide is a sedative with no teratogenic action. This difference in structure between fenofibrate and pirinixic acid also partially explains why the protein target analysis does not group these compounds together since this analysis is also based on the chemical (2D) structure using the ECFP6 fingerprints. However, if we take a closer look at the calculated *z*-scores of this analysis, there are also some inconsistencies with literature. Despite the fact that both compounds are PPARA agonists, PPARA is only a significant protein target for fenofibrate, not pirinixic acid. Another interesting observation is the significance of PPARD and PPARG for fenofibrate when this compound is usually not considered an agonist for these two PPARs [37]. The two top-scoring protein targets for pirinixic acid, prostaglandin E2 synthase-1 (PGES-1) and 5-lipoxygenase (ALOX5), also show an inconsistency with literature (Table 2). PGES and ALOX5 are only protein targets for pirinixic acid after substantial modification of the structure to an aminothiazole-featured pirinixic acid [38]. It thus appears that protein targets do not always reflect literature accurately, which may be related to drawbacks of the manual curation on which the algorithm is dependent.

(b) Clofibrate, gemfibrozil, valproic acid, and simvastatin.

The compounds clofibrate, gemfibrozil, valproic acid, and simvastatin form an obvious cluster in the meta-analysis but are completely scattered across the Tanimoto and protein target dendrograms. Clofibrate and gemfibrozil are PPARA agonists, while simvastatin, a statin compound, increases expression of PPARA and as such can have a similar effect [39]. Indeed there appears to be a cross-talk of statin signaling pathways and (agonist-induced) PPARA activity, and combination therapies of fibrates and statins are being used to treat dyslipidemia [40–42]. Valproic acid has a different mechanism of action and is used as an anticonvulsant and mood-stabilizing drug which has been attributed to the blockade of voltage-dependent sodium channels and increased brain levels of gamma-aminobutyric acid (GABA) [43]. However, it has also been found to be an activator of PPARD, but not PPARA or PPARG, although it is not a direct PPARD ligand [44, 45]. Valproic acid can therefore interact in the PPAR signaling cascades, which explains its similarity in gene expression with the other three compounds. Despite this similarity in gene expression and evidence in literature for overlap in mechanism of action, the Tanimoto and protein target analyses do not consider these compounds to be similar in their effect. However, a visual inspection of the molecular structures of these compounds does reveal a struc-

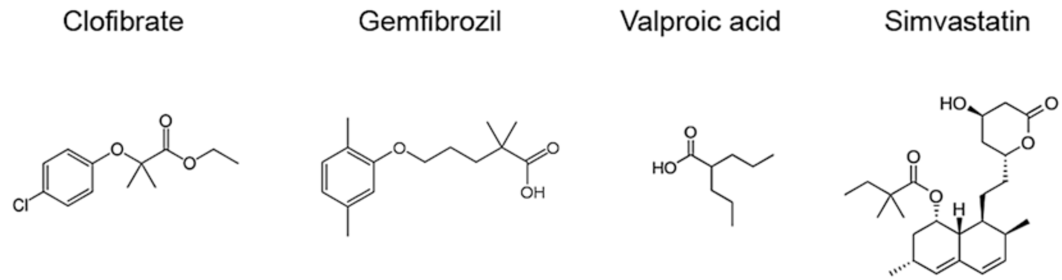


Fig. 5 Molecular structures of clofibrate, gemfibrozil, valproic acid, and simvastatin

tural similarity, especially between clofibrate and gemfibrozil (Fig. 5). Moreover, the carboxylic (pentanoic) acid moiety in these two compounds is also present in valproic acid. This moiety is essential for fibrates to function as PPAR agonists [46]. Since both the 3D Tanimoto analysis and the 2D ECFP6-based protein target analysis take the entire structure into account, essential substructures that convey the similarity in working mechanism could be masked by a dissimilarity in the remainder of the molecule. Smaller molecules with structural similarities can therefore be expected to cluster together more readily as can be seen in the next example.

(c) Allyl alcohol, ethanol, carbon tetrachloride, and *N*-nitrosodimethylamine.

The compounds allyl alcohol, ethanol, carbon tetrachloride, and *N*-nitrosodimethylamine form clusters in the Tanimoto and protein target dendrograms but are completely scattered across the meta-analysis. Indeed their structures are very similar as shown in Fig. 6 which also contributes to the overlap in protein targets. While structural similarity does not guarantee similar gene expression responses, literature review does suggest that these compounds should share some common mode of action. For example, all four compounds are metabolized by the cytochrome P450 metabolizing enzyme CYP2E1 and/or alcohol dehydrogenase (ADH) causing oxidative stress which (partially) explains their hepatotoxic effects [47–52]. An explanation for the scattered clustering in the meta-analysis could lay in the fact that some essential information in the gene expression meta-analysis may get lost since we found that some compounds did cluster similarly to the protein target *z*-scores and Tanimoto scores when a distinction was made based on, for example, dose and exposure time (results not shown). Of course this is inherent to the approach of the meta-analysis, but could lead to problems with group identification if transcriptomic responses differ greatly between experimental conditions.

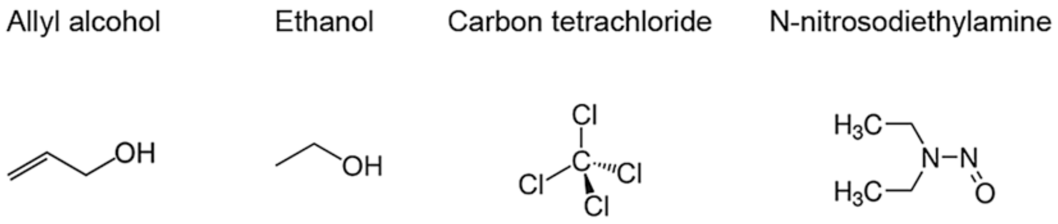


Fig. 6 Molecular structures of allyl alcohol, ethanol, carbon tetrachloride, and *N*-nitrosodimethylamine

7 Combined Analysis Using iClusterPlus

The individual analyses presented above reveal a number of shortcomings, which include (a) disagreements with the described mechanisms of action of compounds with respect to identified protein targets, (b) important similarities in compound substructures which are missed, and (c) a loss of important information when performing a cross-study/cross-platform meta-analysis. These limitations may be overcome by running an integrative clustering that takes into account all data in one single analysis and can resolve the considerable heterogeneity present in individual data sets. iClusterPlus is an R-based tool specifically designed for such a multi-data source integration using a joint latent variable model [53]. It is designed to perform pattern discovery that can integrate diverse data types such as binary values (e.g., somatic mutation data), categorical values (e.g., copy number gain, normal, loss), and continuous values (e.g., gene expression, protein levels) (Fig. 7).

Given multiple data types (e.g., gene expression, Tanimoto scores, protein target data, etc.) measured in the same set of samples and specified sparsity parameter values, iClusterPlus uses generalized linear regression to fit a regularized latent variable model-based clustering that generates an integrated cluster assignment based on joint inference across data types. The common set of latent variables represents distinct driving factors, which, geometrically speaking, form a set of principal coordinates that span a lower dimensional integrated subspace and collectively capture major biological variations, enabling rigorous analysis of the integrated genomic data [53]. The iClusterPlus package is available for download from the open-source software framework Bioconductor (http://www.bioconductor.org/).

7.1 iClusterPlus Results

Compounds with similar toxicity and/or mode of action were grouped using iClusterPlus by integrating meta-analysis gene scores, structural similarities, and protein target predictions. In order to guarantee that each data type has the same weight in the analyses, scaled Euclidian distances were used for meta-analysis gene scores and target predictions in the range of 0–1 (0 = most similar; 1 = most dissimilar), and for structural similarities the 3D Tanimoto scores were used in the range 0–2 (0 = most dissimilar; 2 = most similar).

The iClusterPlus analysis was performed using default settings except for the number of CPUs used for parallel computing (30 CPUs) and the lasso parameter λ which was rescaled to be between 0 and 0.1. These settings were used to determine the optimal number of clusters by calculating the percentage of total variation explained by the model for 2–21 clusters. The percentage explained variation typically increases as more clusters are introduced. The optimal number of clusters is where the curve of percentage explained variation levels off. Figure 8 shows the curve for the analysis with the three data types combined, where 16 clusters are indicated as the optimum number of clusters.

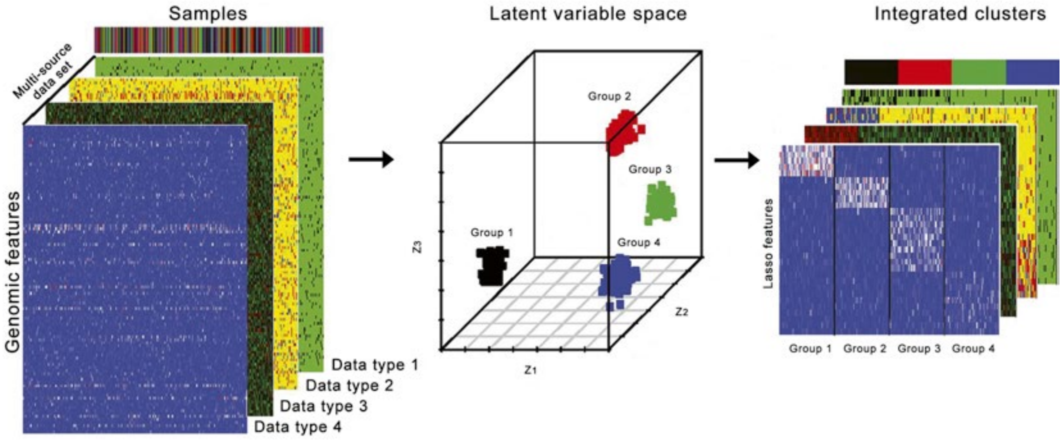


Fig. 7 The basic principle of iClusterPlus analysis. Adapted with PNAS permission from Ref. [53]

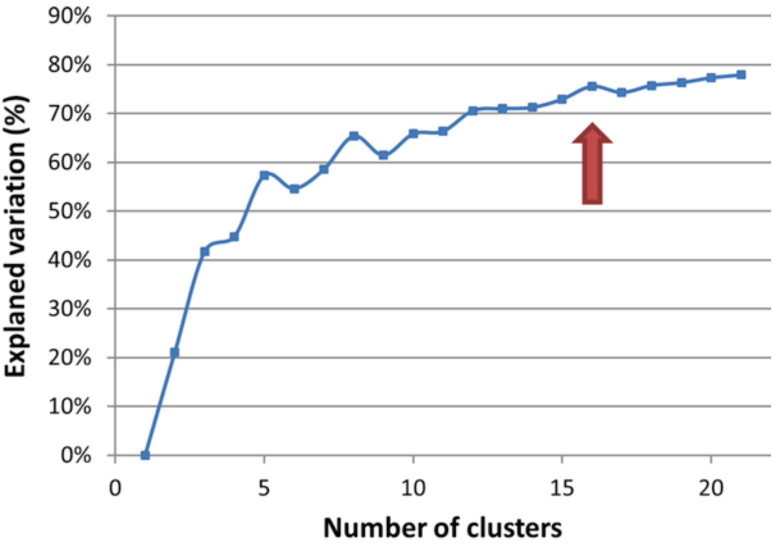
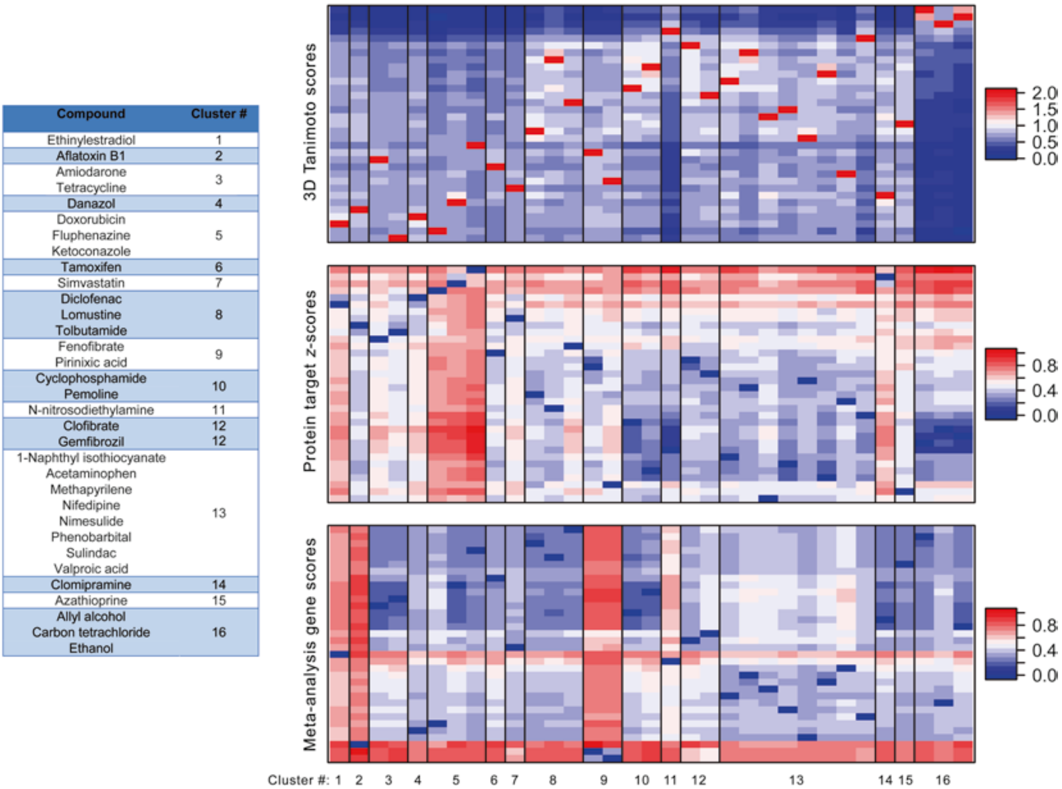


Fig. 8 Percentage explained variation curve for the analysis with the three data types combined. The arrow indicates the optimal number of clusters

There indeed appears to be a better grouping of compounds when all three approaches are combined (Fig. 9). For example, fenofibrate and pirinixic acid now cluster together (cluster #9) where they previously did only in the meta-analysis (Fig. 4). Protein targets in this case did not fully reflect the literature (which provides sufficient evidence for a similar mechanism of action), and the structures, while having some similarities, were found to be considered as different when taking into account the whole structure in the Tanimoto score analysis.

Clofibrate, gemfibrozil, simvastatin, and valproic acid previously grouped together in the meta-analysis which was supported by literature to a certain degree (all involved in peroxisome signaling), but structurally they are more dissimilar, and their protein targets are different because they work through different mechanisms (i.e., clofibrate and gemfibrozil are PPARA agonists, while simvastatin increases PPARD expression and valproic acid affects PPARD signaling). This is now much better reflected by the clustering in Fig. 9 where clofibrate and gemfibrozil cluster together



Compound	Cluster #
Ethinylestradiol	1
Aflatoxin B1	2
Amiodarone Tetracycline	3
Danazol	4
Doxorubicin Fluphenazine Ketoconazole	5
Tamoxifen	6
Simvastatin	7
Diclofenac Lomustine Tolbutamide	8
Fenofibrate Pirinixic acid	9
Cyclophosphamide Pemoline	10
N-nitrosodiethylamine	11
Clofibrate	12
Gemfibrozil	12
1-Naphthyl isothiocyanate Acetaminophen Methapyrilene Nifedipine Nimesulide Phenobarbital Sulindac Valproic acid	13
Clomipramine	14
Azathioprine	15
Allyl alcohol Carbon tetrachloride Ethanol	16

3D Tanimoto scores
2.0
1.5
1.0
0.5
0.0
Protein target z-scores
0.8
0.4
0.0
Meta-analysis gene scores
0.8
0.4
0.0
Cluster #: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Fig. 9 iClusterPlus results, showing the grouping of the 33 compounds in the data set based on an integrated multisource analysis of protein target *z*-scores (Euclidian distances), meta-analysis gene scores (Euclidian distances), and 3D Tanimoto scores. The order of the compounds in the table corresponds with the column order in the clustering heatmap (cluster #12), gemfibrozil forms a separate group (cluster #7), and valproic acid is clustered together with a set of other compounds (cluster #13). These compounds include the COX-2-selective, nonsteroidal anti-inflammatory drug nimesulide which is known to affect both GABA neurotransmission and PPARD signaling just like valproic acid [43, 54, 55] and phenobarbital, which is also an anticonvulsant that interacts with the GABAergic response [56].

According to literature, allyl alcohol, carbon tetrachloride, ethanol, and *N*-nitrosodimethylamine all have a somewhat similar metabolic mechanism and toxicity (CYP2E1/ADH metabolism, oxidative stress). Indeed these compounds had similar protein targets and a similarity in structure (Fig. 4, ethanol and allyl alcohol form a group and carbon tetrachloride and *N*-nitrosodimethylamine), but this was not reflected by the meta-analysis data. However, when separate doses and time points were investigated, this grouping was better (results not shown). The iClusterPlus analysis now also shows a much better grouping of these compounds with only NDEA forming a separate group (#11).

It thus appears that an integrated analysis of data from multiple sources potentially leads to an improved clustering of related hepatotoxic compounds.

8 Conclusion

In this chapter, we have presented an approach that focuses on integrating hepatotoxic compound-induced gene expression and (protein target-directed) chemical structural patterns in order to evaluate whether they can complement each other. The presented examples show that grouping compounds based solely on cross-study/cross-platform gene expression, 3D chemical structure, or protein targets can result in wrongly clustered compounds which have different toxicity or mode of action. To overcome these limitations, iClusterPlus is shown to be a promising tool for integrating data from several distinct sources and improving the clustering of related compounds which share a common mechanism of action. It should be pointed out though that evaluation of the identified groups is needed by (literature-based) expert judgment. Still, a systems biology approach where multiple data sources are used, especially when these data types focus on different aspects of compound (hepato)toxicity and/or chemistry, appears to be a promising way of handling big data sets and promoting the development of new pharmaceutical compounds. The flexibility of iClusterPlus with regard to data set types (e.g., binary, categorical, and continuous values) allows for many data sets to be included in the analysis if considered toxicologically relevant. Inclusion of other data sources, such as proteomics or fragment-based fingerprint methods, is likely to further improve the grouping of similar compounds.

Acknowledgments

This work was supported by the European Commission under its 7th Framework Programme with the Grant "diXa" (283775) and by the European Chemical Industry Council - Long-Range Research Initiative (CEFIC-LRI) with the Grant "DECO2" (AIMT4-UM).

References

1. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 16:85–97
2. Holzinger ER, Ritchie MD (2012) Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. Pharmacogenomics 13:213–222
3. Reif DM, White BC, Moore JH (2004) Integrated analysis of genetic, genomic and proteomic data. Expert Rev Proteomics 1:67–75
4. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, Beyene J (2009) Data integration in genetics and genomics: methods and challenges. Hum Genom Proteomics. DOI: 10.4061/2009/869093
5. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11:476–486
6. Shon J, Abernethy DR (2014) Application of systems pharmacology to explore mechanisms of hepatotoxicity. Clin Pharmacol Ther 96:536–537
7. Howell BA, Siler SQ, Watkins PB (2014) Use of a systems model of drug-induced liver injury (DILIsym((R))) to elucidate the mechanistic differences between acetaminophen and its less-toxic isomer, AMAP, in mice. Toxicol Lett 226:163–172
8. Bhattacharya S, Shoda LK, Zhang Q, Woods CG, Howell BA, Siler SQ, Woodhead JL, Yang Y, McMullen P, Watkins PB, Andersen ME (2012) Modeling drug- and chemical-induced hepatotoxicity with systems biology approaches. Front Physiol 3:462
9. Chen M, Vijay V, Shi Q, Liu Z, Fang H, Tong W (2011) FDA-approved drug labeling for the study of drug-induced liver injury. Drug Discov Today 16:697–703
10. Cui Y, Paules RS (2010) Use of transcriptomics in understanding mechanisms of drug-induced toxicity. Pharmacogenomics 11:573–585
11. Giuliano KA, Gough AH, Taylor DL, Vernetti LA, Johnston PA (2010) Early safety assessment using cellular systems biology yields insights into mechanisms of action. J Biomol Screen 15:783–797
12. DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22:151–185
13. Senior JR (2008) What is idiosyncratic hepatotoxicity? What is it not? Hepatology 47: 1813–1815
14. Holmes AM, Creton S, Chapman K (2010) Working in partnership to advance the 3Rs in toxicity testing. Toxicology 267:14–19
15. Soldatow VY, Lecluyse EL, Griffith LG, Rusyn I (2013) In vitro models for liver toxicity testing. Toxicol Res (Camb) 2:23–39
16. Hendrickx DM, Aerts HJ, Caiment F, Clark D, Ebbels TM, Evelo CT, Gmuender H, Hebels DG, Herwig R, Hescheler J, Jennen DG, Jetten MJ, Kanterakis S, Keun HC, Matser V, Overington JP, Pilicheva E, Sarkans U, Segura-Lepe MP, Sotiriadou I, Wittenberger T, Wittwehr C, Zanzi A, Kleinjans JC (2015) diXa: a data infrastructure for chemical safety assessment. Bioinformatics 31:1505–1507
17. Bolton EE, Chen J, Kim S, Han L, He S, Shi W, Simonyan V, Sun Y, Thiessen PA, Wang J, Yu B, Zhang J, Bryant SH (2011) PubChem3D: a new resource for scientists. J Cheminform 3:32
18. Kim S, Bolton EE, Bryant SH (2012) Effects of multiple conformers per compound upon 3-D similarity search and bioassay data analysis. J Cheminform 4:28
19. Kim S, Bolton EE, Bryant SH (2011) PubChem3D: biologically relevant 3-D similarity. J Cheminformatics 3:26
20. Jenkins JL, Bender A, Davies JW (2006) In silico target fishing: predicting biological tar-

gets from chemical structure. Drug Discov Today Tech 3:413–421

21. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model 46:1124–1133
22. Southan C, Sitzmann M, Muresan S (2013) Comparing the chemical structure and protein content of ChEMBL, DrugBank, human metabolome database and the therapeutic target database. Mol Informat 32:881–897
23. Mugumbate G, Abrahams KA, Cox JA, Papadatos G, van Westen G, Lelievre J, Calus ST, Loman NJ, Ballell L, Barros D, Overington JP, Besra GS (2015) Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and in vitro validation. PLoS One 10:e0121492
24. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
25. ChEMBL Team (2013) ChEMBL release 17. DOI: 10.6019/CHEMBL.database.17
26. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754
27. Kreyszig E (1979) Applied mathematics. Wiley Press, New York
28. Caiment F, Tsamou M, Jennen D, Kleinjans J (2014) Assessing compound carcinogenicity in vitro using connectivity mapping. Carcinogenesis 35:201–207
29. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 101:9309–9314
30. Vilardell M, Rasche A, Thormann A, Maschke-Dutz E, Perez-Jurado LA, Lehrach H, Herwig R (2011) Meta-analysis of heterogeneous down syndrome data reveals consistent genome-wide dosage effects related to neurological processes. BMC Genomics 12:229
31. Rasche A, Al-Hasani H, Herwig R (2008) Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. BMC Genomics 9:310
32. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193
33. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31:e15
34. Rasche A, Yildirimman R, Herwig R (2009) Integrative analysis of microarray data: a path for systems toxicology, General, applied and systems toxicology. Wiley, Hoboken, NJ
35. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 33:e175
36. Guo L, Fang H, Collins J, Fan XH, Dial S, Wong A, Mehta K, Blann E, Shi L, Tong W, Dragan YP (2006) Differential gene expression in mouse primary hepatocytes exposed to the peroxisome proliferator-activated receptor alpha agonists. BMC Bioinformatics 7(Suppl 2):S18
37. Ogata M, Tsujita M, Hossain MA, Akita N, Gonzalez FJ, Staels B, Suzuki S, Fukutomi T, Kimura G, Yokoyama S (2009) On the mechanism for PPAR agonists to enhance ABCA1 gene expression. Atherosclerosis 205:413–419
38. Hanke T, Dehm F, Liening S, Popella SD, Maczewsky J, Pillong M, Kunze J, Weinigel C, Barz D, Kaiser A, Wurglics M, Lammerhofer M, Schneider G, Sautebin L, Schubert-Zsilavecz M, Werz O (2013) Aminothiazole-featured pirinixic acid derivatives as dual 5-lipoxygenase and microsomal prostaglandin E2 synthase-1 inhibitors with improved potency and efficiency in vivo. J Med Chem 56:9031–9044
39. Seo M, Inoue I, Ikeda M, Nakano T, Takahashi S, Katayama S, Komoda T (2008) Statins activate human PPARalpha promoter and increase PPARalpha mRNA expression and activation in HepG2 cells. PPAR Res 2008:316306
40. Paumelle R, Blanquart C, Briand O, Barbier O, Duhem C, Woerly G, Percevault F, Fruchart JC, Dombrowicz D, Glineur C, Staels B (2006) Acute antiinflammatory properties of statins involve peroxisome proliferator-activated receptor-alpha via inhibition of the protein kinase C signaling pathway. Circ Res 98:361–369
41. Wierzbicki AS, Mikhailidis DP, Wray R, Schacter M, Cramb R, Simpson WG, Byrne CB (2003) Statin-fibrate combination: therapy for hyperlipidemia: a review. Curr Med Res Opin 19:155–168

42. Barnett J, Viljoen A, Wierzbicki AS (2013) The need for combination drug therapies in patients with complex dyslipidemia. Curr Cardiol Rep 15:391

43. Chateauvieux S, Morceau F, Dicato M, Diederich M (2010) Molecular and therapeutic potential and toxicity of valproic acid. J Biomed Biotechnol. http://www.ncbi.nlm.nih.gov/pubmed/20798865

44. Lampen A, Carlberg C, Nau H (2001) Peroxisome proliferator-activated receptor delta is a specific sensor for teratogenic valproic acid derivatives. Eur J Pharmacol 431:25–33

45. Ren H, Aleksunes LM, Wood C, Vallanat B, George MH, Klaassen CD, Corton JC (2010) Characterization of peroxisome proliferator-activated receptor alpha – independent effects of PPARalpha activators in the rodent liver: di-(2-ethylhexyl) phthalate also activates the constitutive-activated receptor. Toxicol Sci 113:45–59

46. Kliewer SA, Xu HE, Lambert MH, Willson TM (2001) Peroxisome proliferator-activated receptors: from genes to physiology. Recent Prog Horm Res 56:239–263

47. Jia R, Cao LP, Du JL, Wang JH, Liu YJ, Jeney G, Xu P, Yin GJ (2014) Effects of carbon tetrachloride on oxidative stress, inflammatory response and hepatocyte apoptosis in common carp (Cyprinus carpio). Aquat Toxicol 152:11–19

48. Jimenez-Lopez JM, Cederbaum AI (2005) CYP2E1-dependent oxidative stress and toxicity: role in ethanol-induced liver injury. Expert Opin Drug Metab Toxicol 1:671–685

49. Jaeschke H, Gores GJ, Cederbaum AI, Hinson JA, Pessayre D, Lemasters JJ (2002) Mechanisms of hepatotoxicity. Toxicol Sci 65:166–176

50. Yang JW, Shin JS, Lee JJ, Chang HI, Kim CW (2001) In vitro model using mouse hepatocytes for study of alcohol stress. Biosci Biotechnol Biochem 65:1528–1533

51. Kwolek-Mirek, M., R. Zadrag-Tecza, S. Bednarska and G. Bartosz (2014) Acrolein-Induced Oxidative Stress and Cell Death Exhibiting Features of Apoptosis in the Yeast Saccharomyces cerevisiae Deficient in SOD1. Cell Biochem Biophys 71:1525–1536

52. Kujawska M, Ignatowicz E, Murias M, Ewertowska M, Mikolajczyk K, Jodynis-Liebert J (2009) Protective effect of red beetroot against carbon tetrachloride- and N-nitrosodiethylamine-induced oxidative stress in rats. J Agric Food Chem 57:2570–2575

53. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A 110:4245–4250

54. Zhu B, Bai R, Kennett MJ, Kang BH, Gonzalez FJ, Peters JM (2010) Chemoprevention of chemically induced skin tumorigenesis by ligand activation of peroxisome proliferator-activated receptor-beta/delta and inhibition of cyclooxygenase 2. Mol Cancer Ther 9:3267–3277

55. Dhir A, Naidu PS, Kulkarni SK (2007) Neuroprotective effect of nimesulide, a preferential COX-2 inhibitor, against pentylenetetrazol (PTZ)-induced chemical kindling and associated biochemical parameters in mice. Seizure 16:691–697

56. Ghio L, Cervetti A, Respino M, Belvederi Murri M, Amore M (2014) Management and treatment of gamma butyrolactone withdrawal syndrome: a case report and review. J Psychiatr Pract 20:294–300