OPEN ACCESS Freely available online

PLoS one

Prediction of Pharmacological and Xenobiotic Responses to Drugs Based on Time Course Gene Expression Profiles

Tao Huang[3,4☯], WeiRen Cui[1,2☯], LeLe Hu[1], KaiYan Feng[4], Yi-Xue Li[3,4*], Yu-Dong Cai[1*]

1 Institute of Systems Biology, Shanghai University, Shanghai, People's Republic of China, **2** Centre for Computational Systems Biology, Fudan University, Shanghai, People's Republic of China, **3** Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China, **4** Shanghai Center for Bioinformation Technology, Shanghai, People's Republic of China

Abstract

More and more people are concerned by the risk of unexpected side effects observed in the later steps of the development of new drugs, either in late clinical development or after marketing approval. In order to reduce the risk of the side effects, it is important to look out for the possible xenobiotic responses at an early stage. We attempt such an effort through a prediction by assuming that similarities in microarray profiles indicate shared mechanisms of action and/or toxicological responses among the chemicals being compared. A large time course microarray database derived from livers of compound-treated rats with thirty-four distinct pharmacological and toxicological responses were studied. The mRMR (Minimum-Redundancy-Maximum-Relevance) method and IFS (Incremental Feature Selection) were used to select a compact feature set (141 features) for the reduction of feature dimension and improvement of prediction performance. With these 141 features, the Leave-one-out cross-validation prediction accuracy of first order response using NNA (Nearest Neighbor Algorithm) was 63.9%. Our method can be used for pharmacological and xenobiotic responses prediction of new compounds and accelerate drug development.

Citation: Huang T, Cui W, Hu L, Feng K, Li Y-X, et al. (2009) Prediction of Pharmacological and Xenobiotic Responses to Drugs Based on Time Course Gene Expression Profiles. PLoS ONE 4(12): e8126. doi:10.1371/journal.pone.0008126

Editor: Hany A. El-Shemy, Cairo University, Egypt

Received September 7, 2009; **Accepted** November 10, 2009; **Published** December 2, 2009

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yxli@sibs.ac.cn (YL); cai_yud@yahoo.com.cn (YC)

☯ These authors contributed equally to this work.

Introduction

With drug discovery now being driven primarily by bio-chemistry and high-throughput screening, the biological effects and, in particular, the pharmacology and toxicity of new compounds are required to be studied and evaluated properly before they are released. However, it is impossible to test every detail of a new compound in vitro. It is necessary to predict the possible effects of new drugs, and then experimental examinations can be initiated and orientated, resulting in a new subject of study – toxicogenomics [1,2,3,4,5] (combining the toxicology with some high-throughput technologies) – which enables us to ask some detailed questions about the possible drug effects very early on, thereby fundamentally changing the traditional approaches for the drug discovery. Microarray profiles have been used extensively in some basic biological researches, biomarker determination, pharmacology, drug target selectivity, development of prognostic tests and determination of disease-subclass, as well as in toxicogenomics. Microarray profile will also be used as the input data of pharmacological and xenobiotic response for this study. The livers play many roles in the body functioning, such as the control and synthesis of critical blood constituents including glucose, free-fatty acids, ketone bodies, amino acids, hormones, clotting factors, and inflammatory mediators [6]. The liver is critical in defense against certain infectious organisms and toxins, entered from the gastrointestinal tract [7]. Therefore, data from the liver xenobiotic and pharmacological responses are used for analysis in the study.

Both the pharmaceutical industry and the Regulatory Authorities are, despite the increasing effort to develop safer drugs, concerned by the risk of unexpected side effects observed in the later steps of the development of new drugs, either in late clinical development or after marketing approval. In order to reduce the risk of the side effects, it is important to look out for the possible xenobiotic responses at an early stage. We attempt such an effort through a prediction by assuming that similarities in microarray profiles indicate shared mechanisms of action and/or toxicological responses among the chemicals being compared [8,9,10] since it has been demonstrated that compounds with similar pharmacological or toxicological effects produced similar gene expression profiles either in vitro [11] or in vivo [12,13] exposure conditions. Because one drug may have multiple responses during the regulatory time-course studies, the prediction should allow one data to be allocated to multiple classes, i.e. a multiple-target classification/prediction problem. 34 categories of pharmacological and toxicological effects were adopted (Refer to Table 1) to be the targets of each molecular compound. These categories are divided according to the body and organ weight (BO), histopathology (H), clinical pathology (CP) and structural activity class (SAC).

Machine learning and data mining methods have been widely used in the computational biology and bioinformatics area. Many researchers have made lots of efforts to develop useful algorithms and software to investigate various biology problems such as protein post-translation modification, bio-

Table 1. The Characteristics of 34 responses.

Unique ID	Type	Category	Description	Drugs
SV0567082R5RU	T	CP	Absolute monocyte increase	1-NAPHTHYL ISOTHIOCYANATE, IBUPROFEN, SULINDAC, FLUCONAZOLE, NAPROXEN, ITRACONAZOLE, 4,4′-METHYLENEDIANILINE, ERYTHROMYCIN, GERANIOL, CHOLECALCIFEROL, OXICONAZOLE, CITRIC ACID, LANSOPRAZOLE, GENTIAN VIOLET, CHLOROXYLENOL, PRAZIQUANTEL, CARBAMAZEPINE, NYSTATIN, PRAMOXINE, KETOROLAC, PRALIDOXIME CHLORIDE, BENZETHONIUM CHLORIDE, ROFLUMILAST, IBUFENAC
SV0567098R5RU	T	CP	Creatinine increase	IBUPROFEN, NIMESULIDE, CISPLATIN, CHLOROFORM, LOMEFLOXACIN, FLUOXETINE, PROPYLTHIOURACIL, TICLOPIDINE, PRIMAQUINE, ANISINDIONE, SULFADIAZINE, COLISTIN, PYROGALLOL, TACRINE, ETODOLAC, ROXITHROMYCIN, AMIODARONE, NAFENOPIN
SV0567149R5RU	T	CP	Albumin increase	KETOCONAZOLE, FENOFIBRATE, LOVASTATIN, PREDNISOLONE, PRAVASTATIN, AMOXAPINE, ISONIAZID, TOLAZAMIDE, DEFERIPRONE, PRIMIDONE, MEGESTROL ACETATE, PIRINIXIC ACID, BUPROPION, BETAMETHASONE, FLUDROCORTISONE ACETATE, HYDROCORTISONE, NAFENOPIN
SV0562011R5RU	T	CP	Mean corpuscular hemoglobin concentration decrease (diagnostic, 3–7D time points)	CORTISONE, NIMETAZEPAM, THALIDOMIDE, ETODOLAC, ROXITHROMYCIN, ETHISTERONE, OXYMETHOLONE, 2,3,7,8-TETRACHLORODIBENZO-P-DIOXIN, 3-METHYLCHOLANTHRENE, PHENOBARBITAL, BETA-NAPHTHOFLAVONE, PERHEXILINE, ETHYLESTRENOL, CELECOXIB, ROFECOXIB, BENOXAPROFEN
SV0571010R5RU	T	CP	Mean corpuscular hemoglobin concentration decrease (predictive, 0.25–1D time points)	ROFECOXIB, ETODOLAC, ROXITHROMYCIN, NIMETAZEPAM, CORTISONE, THALIDOMIDE, OXYMETHOLONE, ETHISTERONE, BETA-NAPHTHOFLAVONE, 3-METHYLCHOLANTHRENE, 2,3,7,8-TETRACHLORODIBENZO-P-DIOXIN, BENOXAPROFEN, ETHYLESTRENOL, PERHEXILINE, CELECOXIB
SV0567088R5RU	T	CP	Glucose increase	KETOCONAZOLE, DEXAMETHASONE, THIOGUANINE, METHOTREXATE, CYCLOSPORIN A, CARMUSTINE, NAPROXEN, CYPROTERONE ACETATE, PROMAZINE, ISONIAZID, PYROGALLOL, BETAMETHASONE, HYDROCORTISONE, FLUOCINOLONE ACETONIDE
SV0650093R5RU	T	H	Liver- centrilobular , inflammatory cell infiltrate, mixed cell	ASPIRIN, LEFLUNOMIDE, PENICILLAMINE, CARBOPLATIN, BIS(2-ETHYLHEXYL)PHTHALATE, CHLOROFORM, CLOFIBRIC ACID, CARBIMAZOLE, AMINOSALICYLIC ACID, ISONIAZID, PYRAZINAMIDE, ACETAMINOPHEN, 3-METHYLCHOLANTHRENE, BETA-NAPHTHOFLAVONE, ALPHA-NAPHTHOFLAVONE, 2,3,7,8-TETRACHLORODIBENZO-P-DIOXIN
SV0567153R5RU	T	CP	Total protein increase	CARMUSTINE, N,N-DIMETHYLFORMAMIDE, KETOCONAZOLE, AZATHIOPRINE, CYPROTERONE ACETATE, PREDNISOLONE, PYROGALLOL, CORTISONE, ETHISTERONE, MEGESTROL ACETATE, BETAMETHASONE, FLUDROCORTISONE ACETATE, ETHYLESTRENOL, HYDROCORTISONE
SV0635003R5RU	T	CP	Leukocyte count increase	IBUPROFEN, 1-NAPHTHYL ISOTHIOCYANATE, BROMHEXINE, GENTIAN VIOLET, CHLOROXYLENOL, NYSTATIN, PRAMOXINE, TICRYNAFEN, BENZETHONIUM CHLORIDE
SV0562020R5RU	T	CP	Hemoglobin decrease	IBUPROFEN, SULINDAC, DEXAMETHASONE, THIOGUANINE, NIMESULIDE, HYDROXYUREA, CYTARABINE, INDOMETHACIN, DICLOFENAC, MELOXICAM, SULFISOXAZOLE, LIPOPOLYSACCHARIDE E. COLI O55:B5, TRICHLOROACETIC ACID, PYROGALLOL, ETODOLAC, BROMFENAC, KETOROLAC, PIOGLITAZONE, BENOXAPROFEN
SV0643003R5RU	T	BO	Relative liver weight decrease	SIMVASTATIN, ATORVASTATIN, DICLOFENAC, TAMOXIFEN, TOSUFLOXACIN, LOMEFLOXACIN, 3-METHYLCHOLANTHRENE, BETA-NAPHTHOFLAVONE, ALPHA-NAPHTHOFLAVONE, 2,3,7,8-TETRACHLORODIBENZO-P-DIOXIN, CYCLOPROPANE CARBOXYLIC ACID
SV0562050R5RU	T	CP	Alkaline phosphatase decrease	SODIUM ARSENITE, KETOCONAZOLE, METHOTREXATE, MITOMYCIN C, ETODOLAC, KETOROLAC, CYCLOPROPANE CARBOXYLIC ACID, ROFLUMILAST
SV0643002R5RU	T	BO	Relative spleen weight decrease	CHLORAMBUCIL, DEXAMETHASONE, THIOGUANINE, ROSIGLITAZONE, DOXORUBICIN, LEFLUNOMIDE, KETOCONAZOLE, METHOTREXATE, BETAMETHASONE, HYDROCORTISONE, EPIRUBICIN, FLUOCINOLONE ACETONIDE, DAUNORUBICIN, CYCLOPROPANE CARBOXYLIC ACID
SV0562014R5RU	T	CP	Mean corpuscular hemoglobin decrease (diagnostic, 3–7D time points)	ETODOLAC, ROXITHROMYCIN, 2,3,7,8-TETRACHLORODIBENZO-P-DIOXIN, 3-METHYLCHOLANTHRENE, PHENOBARBITAL, BETA-NAPHTHOFLAVONE, CYCLOPROPANE CARBOXYLIC ACID
SV0562026R5RU	T	CP	Leukocyte count decrease	CHLORAMBUCIL, VALPROIC ACID, THIOGUANINE, CYTARABINE, DOXORUBICIN, LEFLUNOMIDE, IFOSFAMIDE, CARMUSTINE, METHOTREXATE, PROCARBAZINE, MITOMYCIN C, INDOMETHACIN, ETODOLAC, EPIRUBICIN, DAUNORUBICIN, CYCLOPROPANE CARBOXYLIC ACID
SV0650033R5RU	T	H	Liver-periportal, hypertrophy	DEXAMETHASONE, ZOMEPIRAC, DICLOFENAC, MELOXICAM, MITOMYCIN C, INDOMETHACIN, MESTRANOL, ETODOLAC, KETOROLAC, CARVEDILOL, EPIRUBICIN
SV0567174R5RU	T	CP	Absolute basophil increase	3-METHYLCHOLANTHRENE, BETA-NAPHTHOFLAVONE, ALPHA-NAPHTHOFLAVONE, 2,3,7,8-TETRACHLORODIBENZO-P-DIOXIN, PERHEXILINE, ETHYLESTRENOL, CELECOXIB, ROFECOXIB, BENOXAPROFEN
SV0642001R5RU	T	BO	Relative liver weight increase	DEXAMETHASONE, ITRACONAZOLE, KETOCONAZOLE, CYPROTERONE ACETATE, ARTEMISININ, GENTIAN VIOLET, BETAMETHASONE, HYDROCORTISONE, FLUOCINOLONE ACETONIDE
SV0651106R5RU	T	H	Liver-diffuse, cytoplasm, eosinophilia	BEZAFIBRATE, FENOFIBRATE, FLUVASTATIN, CERIVASTATIN, ERYTHROMYCIN, AMINOSALICYLIC ACID, PIRINIXIC ACID, VINBLASTINE
SV0575020R5RU	T	CP	Lipase increase	ATORVASTATIN, BISPHENOL A, KETOCONAZOLE, CLOTRIMAZOLE, BITHIONOL, FLUVASTATIN, NITRAZEPAM

Table 1. Cont.

Unique ID	Type	Category	Description	Drugs
SV0571053R5RU	T	CP	Absolute lymphocyte decrease	CHLORAMBUCIL, THIOGUANINE, DEXAMETHASONE, DOXORUBICIN, KETOCONAZOLE, BETAMETHASONE, FLUDROCORTISONE ACETATE, HYDROCORTISONE, EPIRUBICIN, FLUOCINOLONE ACETONIDE, DAUNORUBICIN
SV0650143R5RU	T	H	Liver-periportal, fibrosis	1-NAPHTHYL ISOTHIOCYANATE, CARMUSTINE, LOMUSTINE, 4,4′-METHYLENEDIANILINE, CROTAMITON
SV0562116R5RU	T	CP	Glucose decrease	1-NAPHTHYL ISOTHIOCYANATE, CLOTRIMAZOLE, NALOXONE, BETA-NAPHTHOFLAVONE, ALPHA-NAPHTHOFLAVONE
SV0650106R5RU	T	H	Liver- hepatocyte, peri-portal, lipid accumulation	MICONAZOLE, ECONAZOLE, MIFEPRISTONE, ALPHA-NAPHTHOFLAVONE
SV0650121R5RU	T	H	Liver- hepatocyte, centri-lobular, lipid accumulation, microvesicular	SULINDAC, MICONAZOLE, INDOMETHACIN, CHLOROFORM
SV0599196R5RU	P	SAC	GR-MR agonist	DEXAMETHASONE, PREDNISOLONE, CORTISONE, BETAMETHASONE, FLUDROCORTISONE ACETATE, HYDROCORTISONE, FLUOCINOLONE ACETONIDE
SV0614125R5RU	T	SAC	Toxicant, DNA alkylator	N-NITROSODIETHYLAMINE, HYDRAZINE, 2-ACETYLAMINOFLUORENE, 4,4′-METHYLENEDIANILINE, AFLATOXIN B1, N-NITROSODIMETHYLAMINE
SV0614137R5RU	P	SAC	Estrogen receptor agonist, steroidal	ETHINYLESTRADIOL, BETA-ESTRADIOL, BETA-ESTRADIOL 3-BENZOATE, ESTRIOL, MESTRANOL
SV0614148R5RU	P	SAC	PPAR a agonist, fibric acid	GEMFIBROZIL, BEZAFIBRATE, CLOFIBRIC ACID, PIRINIXIC ACID, NAFENOPIN
SV0599539R5RU	P	SAC	H+/K+ ATPase inhibitor	OMEPRAZOLE, PANTOPRAZOLE, LANSOPRAZOLE, RABEPRAZOLE
SV0614270R5RU	P	SAC	PDE4 inhibitor	PICLAMILAST, ROFLUMILAST, ROLIPRAM, SCH-351591
SV0599291R5RU	T	SAC	Toxicant, heavy metal (3, 5 and 7D, other non- metal toxicants in negative class)	SODIUM ARSENITE, LEAD(IV) ACETATE, LEAD (II) ACETATE
SV0614202R5RU	T	SAC	Toxicant, heavy metal (0.25–7D allowed, other toxicants not in negative class)	SODIUM ARSENITE, LEAD (II) ACETATE, LEAD(IV) ACETATE
SV0614084R5RU	P	SAC	HMG-CoA reductase inhibitors	ATORVASTATIN, FLUVASTATIN, CERIVASTATIN

The "type" column indicates the toxicity-type (T) or the pharmacology-type (P); there are four categories of responses presented, body and organ weight (BO), histopathology (H), clinical pathology (CP) and structure activity class (SAC).
doi:10.1371/journal.pone.0008126.t001

molecular function classification, protein subcellular locations and protein-DNA interaction [14,15,16,17,18,19,20,21,22, 23,24,25,26].

In this research, we present a classification of the liver toxicogenomic data [27] to support decision making of drug classification, or biomarkers when a new compound is entered for examination. The following sections will describe the microarray data obtained for the study, the analytical machine learning method which include the classification model and feature selection approach mRMR (Minimum-Redundancy-Maximum-Relevance), the results of the prediction and some discussions.

Materials and Methods

Original Data Set

The data used in this work are the time-series microarray data that are extracted from a large liver xenobiotic and pharmacological response database of Iconix Biosciences. The data are publicly available at GEO http://www.ncbi.nlm.nih.gov/geo under accession number GSE8858. The initial data set consists of 1695 individual animal studies and 5288 microarrays. GE Healthcare/Amersham Biosciences CodeLink UniSet Rat I Bioarray, layout EXP5280X2-584, layout EXP5280X2-613 and layout EXP5280X2-648 containing about 10000 probes was used to analyze the global gene expression in the livers of compound-treated rats. Only treatments with gene expression data of day 1, 3 and 5 were involved in our analysis, including 402 treatments with 306 compounds.

Data Construction

First, we get a list of 10399 common probe sets between GE Healthcare/Amersham Biosciences CodeLink UniSet Rat I Bioarray, layout EXP5280X2-584, layout EXP5280X2-613 and layout EXP5280X2-648. Secondly, the gene expression profiles of 402 treatments on day 1, 3 and 5 were obtained from corresponding 3563 microarrays by averaging the duplicated experiments. Then, the control probe sets and probe sets without GenBank Accession number were excluded. The probe sets with more than 30% missing value were also excluded. This yields a subset of 9852 probes. After probe filtering, the missing expression data were imputed using nearest neighbor averaging. Finally, we normalized the expression data of 402 treatments on day 1, 3 and 5 using quantile method.

Thus expression data of 9852 genes of each day (day 1, 3 or 5) were involved in our study, producing 9852*3 = 29556 features for each of the 402 samples. Each sample is to be allocated into the 34 categories listed in Table 1, with the allowance of multiple entries into the categories, using the 29556 features.

Minimum Redundancy Maximum Relevance Feature Selection

Minimum-Redundancy-Maximum-Relevance (mRMR) [28] is a widely used method for feature selection. The goal of mRMR is

to select a feature subset that can best characterize the statistical property of a target classification variable, subject to the constraint that these features are mutually as dissimilar to each other as possible, but marginally as similar to the classification variable as possible.

The feature which has maximum relevance with the target variable and minimum redundancy within the features is defined as a "good" feature. Mutual information (MI) is defined to describe both relevance and redundancy:

$$I(x,y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad (1)$$

Where x and y are two vectors; $p(x,y)$ is the joint probabilistic density; $p(x)$ and $p(y)$ are the marginal probabilistic densities.

The whole vector set is defined as Ω, The selected vector set with m vectors is defined as Ω_s, and the to-be-selected vector set with n vectors is defined as Ω_t. Relevance D of a feature f in Ω_t can be calculated by Eq (2):

$$D=I(f,c) \quad (2)$$

Here c is a classification variable.

Redundancy R of a feature f in Ω_t with all the features in Ω_s can be calculated by Eq (3):

$$R=\frac{1}{m}\sum_{f_i\in\Omega_s}I(f,f_i) \quad (3)$$

mRMR function maximize relevance and minimize redundancy by integrating Eq (2) and Eq (3):

$$\max_{f_j\in\Omega_t}\left[I(f_j,c)-\frac{1}{m}\sum_{f_i\in\Omega_s}I(f_j,f_i)\right](j=1,2,...,n) \quad (4)$$

After the pre-evaluation procedure, a feature set S is provided:

$$S=\left[f_0',f_1',...,f_h',...,f_{N-1}'\right] \quad (5)$$

the feature index reflects the evaluations for feature. The feature which fits the Eq(4) better will be added to the set S earlier. For example, If a<b, f_a is considered to be better than f_b.

Prediction Model

With the mRMR selected features, Nearest Neighbor Algorithm (NNA) [29] is used to classify the data into the above mentioned categories. NNA allocates a new data into categories by comparing the features of the data with the features of those that have known categories. The similarity between two vectors p_x, p_y is defined as [25]:

$$D(p_x,p_y)=1-\frac{p_x\cdot p_y}{||p_x||\cdot||p_y||} \quad (6)$$

where $p_x\cdot p_y$ is the inner product of p_x and p_y, and $||p||$ is the module of vector p. p_x and p_y are considered to be more similar if $D(p_x,p_y)$ is smaller.

Traditionally, NNA chooses to classify the new pattern p_t into the class of its nearest neighbor which has the smallest $D(p_n,p_t)$. That is:

$$D(p_n,p_t)=\min\{D(p_1,p_t),D(p_2,p_t),...,D(p_z,p_t),...,D(p_N,p_t)\}\quad(z\neq t) \quad (7)$$

where N represents the number of training samples.

Because this research is about multi-target classification i.e. a data can belong to more than one category, the prediction model needs to be adjusted to cope with the multi-target problem. In the prediction of multi-targets, if $D(p_m,p_t)<D(p_n,p_t)$, it means that p_t is closer to p_m than to p_n. Thus we rank the predicted classes of each drug data as:

$$\begin{aligned}&class\ i\le class\ j \quad if\ D(p_i,p_t)\le D(p_j,p_t)\\&D(p_i,p_t)=\min\{D(p_1,p_t),D(p_2,p_t),...,D(p_Z,p_t),...D(p_N,p_t)\}\\&\qquad(z\neq t,p_z\in class\ i)\end{aligned} \quad (8)$$

From Eq. (8), we can get a list with the most likely class (defined as order-1 response) to be in the first position, and the second likely class (defined as order-2 response) to be in the second position, and so on.

Jackknife Cross-Validation Method

Jackknife Cross-Validation Method [14,18] is an effective and objective way to evaluate statistical predictions. Each sample in the data set is in turn knocked out and tested by the predictor trained by the other samples remaining in the data set. During the process, every sample is used not only for the training, but also for the testing. The prediction accuracy Q for overall samples was used to evaluate the performance of predictor:

$$Q=\frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative, respectively.

Incremental Feature Selection (IFS)

mRMR only provides a list of features by sorting the features according to their importance to the prediction without telling how many fore features in the list should be selected. The fore features are selected by testing all possible feature sets, and choosing the feature set that achieves the best prediction rate. A possible feature subset S_i can be expressed by the following equation.

$$S_i=\{f_0,f_1,...,f_i\}(0\le i\le N-1) \quad (10)$$

The initial feature subset is $S_0=\{f_0\}$, and the last feature subset is $S_{N-1}=\{f_0,f_1,...,f_{N-1}\}$ which includes all the features. Jackknife test is then used to obtain the accurate prediction rates of all the feature subsets. The one that achieves the highest prediction accuracy is considered to be optimized feature set selected by IFS. We can plot a curve, called IFS curve, with index i as its x-axis and the overall accurate rate as its y-axis.

Results

IFS Curves of the Drug Responses

Because a drug may have several pharmaceutical responses, Eq. 8 is used to rate all the available responses. We only take the first

three responses for every drug. And more will be available if they are needed in a future research. The cumulated prediction accuracies of first one, two, and three responses using different number of features, are shown in Figure 1, evaluated by jackknife cross-validation test. The highest prediction accuracy of first order response was 63.9% with 141 features. The highest cumulated prediction accuracies of first two responses and first three responses were also achieved with these 141 features. The detailed information of the IFS procedure and these 141 features can be found in Table S1 and Table S2.

IFS Feature Selection and the Prediction Accuracy

141 features are selected as the result according to the IFS curves. Using these 141 features, the highest prediction accuracy for the first order response is 63.9%, evaluated by jackknife cross-validation test. Unfortunately, the prediction accuracy is rather low, which might be due to the sparse data points in the high-dimensional feature space. More samples could be used in a future research to study how much the prediction accuracy is affected by the number of samples available for training and predicting the prediction model. And the biological relevance of these 141 features was explored by KEGG and GO category enrichment analysis.

The KEGG category enrichment analysis (see Table S3) shows that two of the 141 features, Cyp3a9 and Ephx1, involves in the pathway for the metabolism of xenobiotics by cytochrome P450. Cytochrome P450s (CYP), comprising a superfamily of heme-thiolate proteins, is the main metabolizing enzyme system for foreign compounds, including drugs, and has a primary role in organism protection against potential harmful assaults from the environment [30]. It is often used as biomarker to determine human exposure to environmental molecules or to predict the susceptibility to certain pathologies [31,32].

The GO category enrichment analysis results (see Table S4, Table S5 and Table S6) show that many of these candidate biomarkers are involved in insulin signaling pathway. The insulin-mediated receptor tyrosine kinase (RTK) signaling pathways [33,34] by downstream effectors such as phosphatidylinositol 3-kinase, mitogen activated protein kinase (MAPK), Akt/protein kinase B (PKB), mammalian target of rapamycin (mTOR), and the p70 ribosomal protein S6 kinase (p70S6 kinase) have been reviewed [35] in the regulation of drug metabolizing enzyme expression in response to insulin and growth factors. The term fatty acid metabolism, comprising genes such as fatty acid synthase, enoyl-CoA hydratase, acyl-CoA synthetase among others is also enriched. The liver is a major site for fatty acid and lipid metabolism, and several major classes of compounds appearing in the database (statins, fibrates, glitazones, estrogen receptor modulators and others) affect the lipid synthesis and degradation. Fatty acids are a major energy source and important constituents of membrane lipids, and they serve as cellular signaling molecules that play an important role in the etiology of the metabolic syndrome [36]. Some liver samples exhibited elevated triglyceride levels that were correlated with changes in the urinary associated with defective metabolism of fatty acids, confirmed by the in vitro experiments [37].

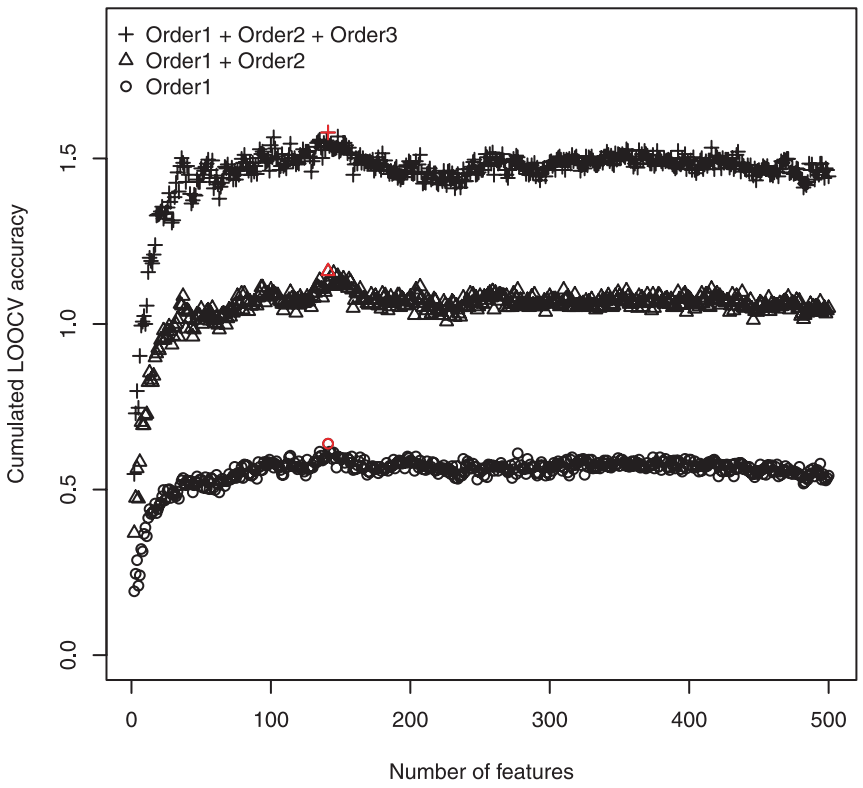

Figure 1. The IFS curve of first three responses prediction. The order-1 response is the most possible response according to the prediction. The highest prediction accuracy of first order response was 63.9% with 141 features. The highest cumulated prediction accuracies of first two responses and first three responses were also achieved with these 141 features. The red color points represent the highest accuracy points of each kind of accuracy.
doi:10.1371/journal.pone.0008126.g001

Discussion

Microarray gene expression profiles has been proved valuable in numerous applications including disease classification, diagnosis, survival analysis, choice of therapy etc [38], but rarely used for drug response prediction. The Connectivity Map [39,40] was a new tool for finding connections among small molecules sharing a mechanism of action, chemicals and physiological processes, and diseases and drugs. But it couldn't systematically research drug response, because the reference collection of gene-expression profiles in Connectivity Map were from cultured human cells treated with bioactive small molecules and most cells were cancer cell lines. The dataset we used were from in vivo rat liver which is closer to clinic. The compound-treated rats had same background. The bias in our research was much smaller. In the dataset of our research, the small molecules were well organized and all the responses were explicit recorded. There were thirty-four distinct pharmacological and toxicological responses. In meta-dataset like Connectivity Map's reference collection, each experiment only provided the phenotype this research group was interested in; other responses were ignored in most time.

The statistic basis of Connectivity Map wasn't solid [40]. The methods we used like mRMR and NNA have solid statistic basis and have been widely used in machine learning studies for a long time. The results were proved effective strictly using Jackknife Cross-Validation.

This paper presents a multi-target prediction for pharmacological and xenobiotic responses from drugs, i.e. allocating a drug treatment to several responses. Microarray data from liver xenobiotic and pharmacological responses are adopted for the prediction. Each drug treatment is coded by the genes of the treated subjects, derived from the microarray profile, resulting in thousands of features. Then mRMR method and IFS are used to select a compact feature set (141 features) for the reduction of feature dimension and improvement of prediction performance. Finally, the features in the compact set, considered to be most important for the prediction, are analyzed through GO category enrichment analysis.

Supporting Information

Table S1 IFS prediction accuracy using different number of features. The first column is the number of features used in prediction. The following columns gave the prediction accuracies from order-1 (the most possible response) to order-34 (the most impossible response). The highest prediction accuracy of first order response was 63.9% with 141 features.
Found at: doi:10.1371/journal.pone.0008126.s001 (0.32 MB XLS)

Table S2 The detailed information of 141 features. The first column is the feature name (probe name with time point). There are 3 time points: day 1, 3 and 5 after treatment start. The third column is the mRMR score.
Found at: doi:10.1371/journal.pone.0008126.s002 (0.12 MB XLS)

Table S3 The KEGG enrichment of 141 features.
Found at: doi:10.1371/journal.pone.0008126.s003 (0.01 MB XLS)

Table S4 The Gene Ontology Biological Process enrichment of 141 features.
Found at: doi:10.1371/journal.pone.0008126.s004 (0.07 MB XLS)

Table S5 The Gene Ontology Molecular Function enrichment of 141 features.
Found at: doi:10.1371/journal.pone.0008126.s005 (0.03 MB XLS)

Table S6 The Gene Ontology Cellular Component enrichment of 141 features.
Found at: doi:10.1371/journal.pone.0008126.s006 (0.02 MB XLS)

Acknowledgments

We would like to thank Iconix Biosciences for storing its data into public repositories to support not-for-profit research efforts and the progress of science.

Author Contributions

Conceived and designed the experiments: YXL YDC. Performed the experiments: TH. Analyzed the data: TH WC. Contributed reagents/materials/analysis tools: WC LH. Wrote the paper: TH WC KF.

References

1. Blomme EA, Yang Y, Waring JF (2009) Use of toxicogenomics to understand mechanisms of drug-induced hepatotoxicity during drug discovery and development. Toxicol Lett 186: 22–31.
2. Boorman GA, Anderson SP, Casey WM, Brown RH, Crosby LM, et al. (2002) Toxicogenomics, drug discovery, and the pathologist. Toxicol Pathol 30: 15–27.
3. Butte A (2002) The use and analysis of microarray data. Nat Rev Drug Discov 1: 951–960.
4. Ganter B, Snyder RD, Halbert DN, Lee MD (2006) Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. Pharmacogenomics 7: 1025–1044.
5. Ulrich R, Friend SH (2002) Toxicogenomics and drug discovery: will new technologies help us produce better drugs? Nat Rev Drug Discov 1: 84–88.
6. Parker GA, Picut CA (2005) Liver immunobiology. Toxicol Pathol 33: 52–62.
7. Morgan KT, Pino M, Crosby LM, Wang M, Elston TC, et al. (2004) Complementary roles for toxicologic pathology and mathematics in toxicogenomics, with special reference to data interpretation and oscillatory dynamics. Toxicol Pathol 32 Suppl 1: 13–25.
8. Davis AP, Murphy CG, Rosenstein MC, Wiegers TC, Mattingly CJ (2008) The Comparative Toxicogenomics Database facilitates identification and understanding of chemical-gene-disease associations: arsenic as a case study. BMC Med Genomics 1: 48.
9. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, et al. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res 37: D786–792.
10. Mattingly CJ, Rosenstein MC, Davis AP, Colby GT, Forrest JN Jr, et al. (2006) The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. Toxicol Sci 92: 587–595.
11. Waring JF, Ciurlionis R, Jolly RA, Heindel M, Ulrich RG (2001) Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity. Toxicol Lett 120: 359–368.
12. Hamadeh HK, Bushel PR, Jayadev S, DiSorbo O, Bennett L, et al. (2002) Prediction of compound signature using high density gene expression profiling. Toxicol Sci 67: 232–240.
13. Waring JF, Jolly RA, Ciurlionis R, Lum PY, Praestgaard JT, et al. (2001) Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. Toxicol Appl Pharmacol 175: 28–42.
14. Cai Y, He J, Li X, Lu L, Yang X, et al. (2009) A Novel Computational Approach To Predict Transcription Factor DNA Binding Preference. J Proteome Res 8: 999–1003.
15. Cai YD, Qian Z, Lu L, Feng KY, Meng X, et al. (2008) Prediction of compounds' biological function (metabolic pathways) based on functional group composition. Mol Divers 12: 131–137.
16. Jia P, Qian Z, Feng K, Lu W, Li Y, et al. (2008) Prediction of membrane protein types in a hybrid space. J Proteome Res 7: 1131–1137.
17. Jia P, Qian Z, Zeng Z, Cai Y, Li Y (2007) Prediction of subcellular protein localization based on functional domain composition. Biochem Biophys Res Commun 357: 366–370.
18. Li S, Liu B, Cai Y, Li Y (2007) Predicting protein N-glycosylation by combining functional domain and secretion information. J Biomol Struct Dyn 25: 49–54.

19. Li S, Liu B, Zeng R, Cai Y, Li Y (2006) Predicting O-glycosylation sites in mammalian proteins by using SVMs. Comput Biol Chem 30: 203–208.
20. Li W, Lin K, Feng K, Cai Y (2008) Prediction of protein structural classes using hybrid properties. Mol Divers 12: 171–179.
21. Lu L, Qian Z, Shi X, Li H, Cai YD, et al. (2009) A knowledge-based method to predict the cooperative relationship between transcription factors. Mol Divers.
22. Lu L, Shi XH, Li SJ, Xie ZQ, Feng YL, et al. (2009) Protein sumoylation sites prediction based on two-stage feature selection. Mol Divers.
23. Niu B, Jin Y, Lu L, Fen K, Gu L, et al. (2009) Prediction of interaction between small molecule and enzyme using AdaBoost. Mol Divers 13: 313–320.
24. Niu B, Jin YH, Feng KY, Lu WC, Cai YD, et al. (2008) Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol Divers 12: 41–45.
25. Qian Z, Cai YD, Li Y (2006) A novel computational method to predict transcription factor DNA binding preference. Biochem Biophys Res Commun 348: 1034–1037.
26. Yu X, Cao J, Cai Y, Shi T, Li Y (2006) Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. J Theor Biol 240: 175–184.
27. Natsoulis G, Pearson CI, Gollub J, B PE, Ferng J, et al. (2008) The liver pharmacological and xenobiotic gene response repertoire. Mol Syst Biol 4: 175.
28. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27: 1226–1238.
29. Friedman JH, Baskett F, Shustek LJ (1975) An algorithm for finding nearest neighbors. IEEE Trans Comput C-24: 1000–1006.
30. Nebert DW, Russell DW (2002) Clinical importance of the cytochromes P450. Lancet 360: 1155–1162.
31. Anzenbacher P, Anzenbacherova E (2001) Cytochromes P450 and metabolism of xenobiotics. Cell Mol Life Sci 58: 737–747.
32. Gueguen Y, Mouzat K, Ferrari L, Tissandie E, Lobaccaro JM, et al. (2006) [Cytochromes P450: xenobiotic metabolism, regulation and clinical importance]. Ann Biol Clin (Paris) 64: 535–548.
33. Ullrich A, Schlessinger J (1990) Signal transduction by receptors with tyrosine kinase activity. Cell 61: 203–212.
34. Porter AC, Vaillancourt RR (1998) Tyrosine kinase receptor-activated signal transduction pathways which lead to oncogenesis. Oncogene 17: 1343–1352.
35. Kim SK, Novak RF (2007) The role of intracellular signaling in insulin-mediated regulation of drug metabolizing enzyme gene and protein expression. Pharmacol Ther 113: 88–120.
36. Wakil SJ, Abu-Elheiga LA (2009) Fatty acid metabolism: target for metabolic syndrome. J Lipid Res 50 Suppl: S138–143.
37. Mortishire-Smith RJ, Skiles GL, Lawrence JW, Spence S, Nicholls AW, et al. (2004) Use of metabonomics to identify impaired fatty acid metabolism as the mechanism of a drug-induced toxicity. Chem Res Toxicol 17: 165–173.
38. Huang T, Tu K, Shyr Y, Wei CC, Xie L, et al. (2008) The prediction of interferon treatment effects based on time series microarray gene expression profiles. J Transl Med 6: 44.
39. Lamb J (2007) The Connectivity Map: a new tool for biomedical research. Nat Rev Cancer 7: 54–60.
40. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313: 1929–1935.