Integrating Biological and Chemical Data for Hepatotoxicity Prediction

Jie Liu,[†,‡,#] Kamel Mansouri,[†,#] Richard Judson,[†] Matthew T. Martin,[†] Huixiao Hong,[§] Minjun Chen,[§] Xiaowei Xu,[‡,§] Russell Thomas,[†] and Imran Shah[†]

[†]National Center for Computational Toxicology, US EPA, Research Triangle Park, NC 27711

[‡]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR 72204

[#]Oak Ridge Institute for Science Education

[§]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration,, Jefferson, AR, USA

Jie Liu | liu.jie@epa.gov | 919-541-5369

Abstract

The U.S. EPA ToxCast™ program is screening thousands of environmental chemicals for bioactivity using hundreds of high-throughput *in vitro* assays to build predictive models of toxicity. A set of 677 chemicals were represented by 711 bioactivity descriptors (from ToxCast assays), 4,376 chemical structure descriptors, and three hepatotoxicity categories (from animal studies), then used supervised machine learning to predict their hepatotoxic effects. Hepatotoxicants were defined by rat liver histopathology observed after chronic chemical testing and grouped into hypertrophy (161), injury (101) and proliferative lesions (99). Classifiers were built using six machine learning algorithms: linear discriminant analysis (LDA), Naïve Bayes (NB), support vector machines (SVM), classification and regression trees (CART), k-nearest neighbors (KNN) and an ensemble of classifiers (ENSMB). Classifiers of hepatotoxicity were built using chemical structure, ToxCast bioactivity, and a hybrid representation. Predictive performance was evaluated using 10-fold cross-validation testing and in-loop, filter-based, feature subset selection. Hybrid classifiers had the best balanced accuracy for predicting hypertrophy (0.78±0.08), injury (0.73±0.10) and proliferative lesions (0.72±0.09). CART, ENSMB and SVM classifiers performed the best, and nuclear receptor activation and mitochondrial functions were frequently found in highly predictive classifiers of hepatotoxicity. ToxCast provides the largest and richest data set for mining linkages between the *in vitro* bioactivity of environmental chemicals and their adverse histopathological outcomes. Our findings demonstrate the utility of high-throughput assays for characterizing rodent hepatotoxicants, the benefit of using hybrid representations that integrate bioactivity and chemical structure, and the need for objective evaluation of classification performance.

(LDA: Linear discriminant analysis ; SVM: Support vector machines; NB:Naïve Bayes; CART: classification and regression trees; KNN: k-nearest neighbors; ENSMB, ensemble classifier.)

Data Sources

Table 1. Data Sets of Chemicals Used for Classification

Data sets	Total chemicals	Hypertrophy lesions	Injury lesions	Proliferative lesions	Negative set	Descriptors
Bioactivity	677	161	–	–	463	125 ToxCast HTS assay endpoints
		–	101	–	463	
		–	–	99	463	
Chemical	677	161	–	–	463	726 chemical structure descriptors
		–	101	–	463	
		–	–	99	463	
Bioactivity & Chemical	677	161	–	–	463	125 ToxCast HTS assay endpoints & 726 chemical structure descriptors
		–	101	–	463	
		–	–	99	463	

Supervised Machine Learning

We used 677 chemicals represented by 125 ToxCast bioactivity assays, 726 chemical structure descriptors and three hepatotoxicity categories (Hypertrophy, Injury, and Proliferative lesions) for supervised machine learning. Performance was evaluated by 10-fold cross-validation. In-loop filter-based feature selection chose different number of top features to build the models. Hepatotoxicity predictive models were built using ToxCast bioactivity assay only, chemical structure descriptors only, or combined data (bioactivity and chemical structure descriptors) by six machine learning algorithms: linear discriminant analysis (LDA), Naïve Bayes (NB), support vector classification (SVCL, SVCR), classification and regression trees (CART), k-nearest neighbors (KNN) and an ensemble of all classifiers (ENSMB).

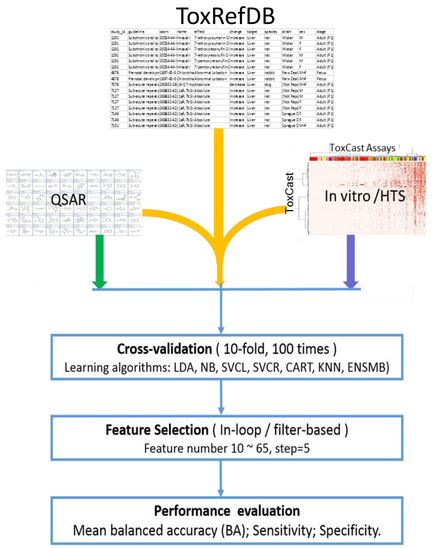


Figure 1. The Workflow for the whole classification process.

Classification Performance Results

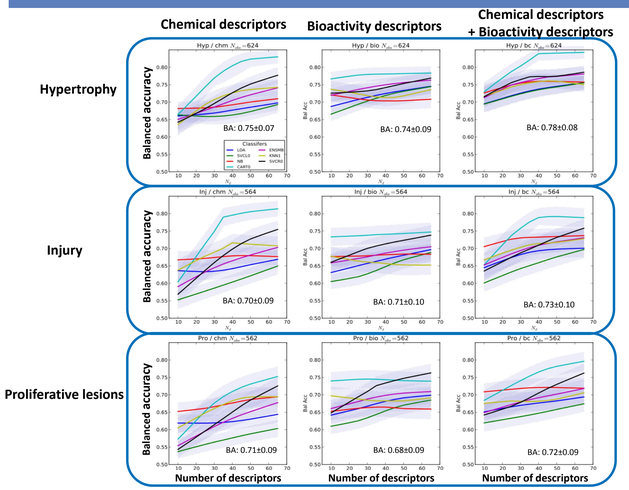


Figure 2. Classification performance

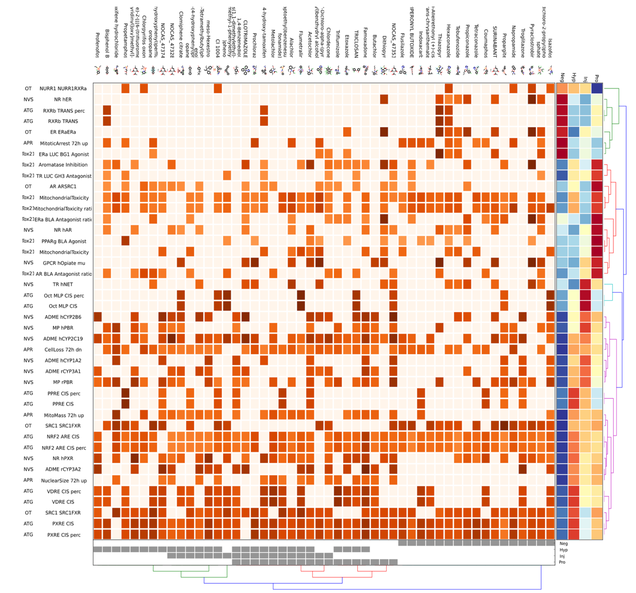
Table 2. The maximum predictive performance of different classification methods.

Toxicity	Classifier	# desc. BIO	CHM	BC	BA BIO	CHM	BC
Hyp	CART0	60	60	65	0.79(0.06)	0.83(0.09)	0.84(0.08)
	ENSMB	55	65	65	0.76(0.06)	0.74(0.08)	0.78(0.08)
	KNN1	15	65	35	0.74(0.08)	0.74(0.10)	0.77(0.09)
	LDA	65	65	65	0.75(0.07)	0.70(0.08)	0.76(0.08)
	NB	25	65	40	0.72(0.08)	0.71(0.08)	0.76(0.08)
	SVCL0	65	65	65	0.74(0.07)	0.69(0.08)	0.76(0.08)
	SVCR0	65	65	30	0.77(0.07)	0.77(0.07)	0.80(0.08)
				mean(sd)	0.75(0.07)	0.74(0.09)	0.78(0.08)
Inj	CART0	65	60	40	0.75(0.09)	0.81(0.11)	0.80(0.11)
	ENSMB	65	65	65	0.70(0.08)	0.70(0.08)	0.73(0.10)
	KNN1	10	40	30	0.68(0.10)	0.72(0.10)	0.73(0.10)
	LDA	65	65	50	0.70(0.09)	0.67(0.08)	0.70(0.08)
	NB	40	45	55	0.69(0.08)	0.68(0.08)	0.74(0.08)
	SVCL0	60	65	65	0.69(0.08)	0.64(0.08)	0.69(0.08)
	SVCR0	65	65	65	0.74(0.09)	0.75(0.09)	0.75(0.10)
				mean(sd)	0.70(0.09)	0.71(0.10)	0.73(0.10)
Pro	CART0	40	65	55	0.75(0.08)	0.75(0.11)	0.79(0.09)
	ENSMB	65	65	65	0.71(0.08)	0.68(0.08)	0.72(0.08)
	KNN1	20	50	65	0.70(0.09)	0.70(0.10)	0.70(0.09)
	LDA	65	65	65	0.70(0.09)	0.64(0.08)	0.70(0.08)
	NB	35	60	55	0.67(0.09)	0.69(0.08)	0.72(0.09)
	SVCL0	65	65	65	0.68(0.08)	0.60(0.08)	0.67(0.08)
	SVCR0	65	65	65	0.76(0.08)	0.72(0.08)	0.76(0.09)
				mean(sd)	0.71(0.09)	0.68(0.09)	0.72(0.09)

Table 3. The most frequently selected bioactivity descriptors for classification.

Descriptor	Technology	Target/Gene	Target/Family
APR_CellLoss_72h_dn	Apredica	NA	cell cycle
APR_MitoMass_72h_up	Apredica	NA	cell morphology
APR_NuclearSize_72h_up	Apredica	NA	cell morphology
ATG_NRF2_ARE_CIS	Attagene	NFE2L2	dna binding
ATG_PPRE_CIS	Attagene	PPARD; PPARG; PPARA	nuclear receptor
ATG_PXRE_CIS	Attagene	NR1I2	nuclear receptor
ATG_VDRE_CIS	Attagene	NR1I1	nuclear receptor
NVS_ADME_hCYP1A2	Novascreen	CYP1A2	cyp
NVS_ADME_hCYP2C19	Novascreen	CYP2C19	cyp
NVS_MP_rPBR	Novascreen	Tspo	transporter
NVS_NR_hPXR	Novascreen	NR1I2	nuclear receptor
OT_SRC1_SRC1FXR_1440	Odyssey Thera	FXR	nuclear receptor
ATG_BRE_CIS	Attagene	SMAD1	dna binding
ATG_Oct_MLP_CIS	Attagene	POU2F1	dna binding
NVS_ADME_hCYP2B6	Novascreen	CYP2B6	cyp
NVS_MP_hPBR	Novascreen	TSPO	transporter
NVS_TR_hNET	Novascreen	SLC6A2	transporter
ATG_RARa_TRANS	Attagene	RARA	nuclear receptor
NVS_GPCR_hOpiate_mu	Novascreen	OPRM1	gpcr
NVS_NR_hAR	Novascreen	AR	nuclear receptor
OT_AR_ARSRC1_0960	Odyssey Thera	AR	nuclear receptor
Tox21_AR_BLA_Antagonist_ratio	Tox21/NCGC	AR	cyp
Tox21_Aromatase_Inhibition	Tox21/NCGC	CYP19A1	cyp
Tox21_ERa_BLA_Antagonist_ratio	Tox21/NCGC	ESR1	nuclear receptor
Tox21_MitochondrialToxicity_ratio	Tox21/NCGC	NA	cell morphology
Tox21_PPARg_BLA_Agonist_ch1	Tox21/NCGC	PPARG	nuclear receptor
Tox21_TR_LUC_GH3_Antagonist	Tox21/NCGC	THRB	nuclear receptor
APR_MitoticArrest_72h_up	Apredica	NA	cell cycle
ATG_ERE_CIS_perc	Attagene	ESR1	nuclear receptor
ATG_RXRb_TRANS	Attagene	RXRB	nuclear receptor
NVS_NR_hER	Novascreen	ESR1	nuclear receptor
NVS_NR_mERa	Novascreen	Esr1	nuclear receptor
OT_ER_ERaERa_0480	Odyssey Thera	ESR1	nuclear receptor
Tox21_AR_BLA_Agonist_ch1	Tox21/NCGC	AR	nuclear receptor
Tox21_ERa_LUC_BG1_Agonist	Tox21/NCGC	ESR1	nuclear receptor
OT_NURR1_NURR1RXRa_0480	Odyssey Thera	RXRA	nuclear receptor

Figure 3. Bioactivity descriptors most frequently selected in classifying hepatotoxicity and representative chemicals.



Conclusions

- High-throughput bioactivity assays are useful for characterizing hepatotoxic liability of chemicals in rodents.
- Hybrid representations that integrate bioactivity and chemical structure descriptors can improve predictive accuracy.
- Machine learning techniques can provide linkages between the *in vitro* bioactivity and chemical structure of environmental chemicals to adverse histopathological outcomes.

References

1. ToxRefDB http://www.epa.gov/ncct/toxrefdb/
2. ToxCast http://www.epa.gov/ncct/toxcast/
3. Liu J, Mansouri K, Judson RS, Martin MT, Hong H, Chen M, Xu X, Thomas RS, Shah I. Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. *Chem. Res. Toxicol.* 2015 Mar 9. (available online with QR code)