Developing Predictive Toxicity Signatures Using In Vitro Data from the EPA ToxCast Program

Richard Judson, David Dix, Keith Houck, Matthew Martin, Holly Mortensen, Robert Kavlock

National Center for Computational Toxicology, Office of Research and Development,
United States Environmental Protection Agency, Research Triangle Park, NC.

Abstract: A major focus in toxicology research is the development of *in vitro* methods to predict *in vivo* chemical toxicity. Numerous studies have evaluated the use of targeted biochemical, cell-based and genomic assay approaches. Each of these techniques is potentially helpful, but provides only a partial view of the complex biology that leads to tissue, organ or whole organism toxic effects. Here we present the first results from the ToxCast program that combines multiple types of assays into "toxicity signatures" that are optimally predictive of particular *in vivo* toxicity endpoints. A toxicity signature is in essence a function that takes as input the results of a set of assays run on a chemical, and produces a prediction of the toxicity of that chemical for a specific *in vivo* endpoint. We used the EPA ToxCast program Phase I data, which includes >500 *in vitro* assays run on a total of 320 chemicals, the majority of which are pesticide active compounds. The corresponding *in vivo* toxicity data is taken from the Toxicology Reference Database (ToxRefDB), which is tabulated data from guideline animal studies for chronic, cancer, developmental and reproductive toxicity endpoints. We present a variety of machine learning approaches to mine this complex data set for toxicity signatures with both high sensitivity and specificity. These include linear discriminant analysis, support vector machines and neural networks. In addition to these automated approaches, we also present more hypothesis-driven, mechanism-based signatures including ones probing nuclear receptor-mediated liver toxicities. An important observation from these first analyses is that one often needs to include assays of multiple types, probing multiple mechanisms or pathways, to adequately predict *in vivo* toxicity across a wide range of chemicals

309 unique structures
3 triplicates, 5 duplicates for QC
8 metabolites

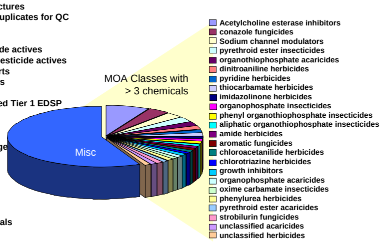
291 total pesticide actives
273 registered pesticide actives
22 pesticide inerts
33 antimicrobials

56 of 73 proposed Tier 1 EDSP

23 IUR
13 HPV
11 HPV Challenge

73 OW PCCL
11 CCL1
10 CCL2
25 CCL3

122 IRIS chemicals



More Information:

ToxCast: http://www.epa.gov/ncct/toxcast

ToxCast Data Analysis Summit: May 14, 15, RTP NC

Public Data Release June 2009 in ACToR: http://actor.epa.gov

Contact: judson.richard@epa.gov

Data Sets Used for Predictions

Name	Description	Assay
ACEA	Real-time Cell Electronic Sensing Measures cell growth and cell shape kinetics	7
Attagene	Transcription factor assays	81
BioSeek	Cell-based protein level assays for 8 human primary cell types	87
Cellumen	Cell imaging assays	11
CellzDirect	Transcription assays for nuclear-receptor mediated liver ADME gene induction	16
Gentronix	GreenScreen Genotoxicity assay	1
NCGC	Cell-based Nuclear receptor assays plus cell viability and p53	24
Novascreen	Receptor binding and enzyme inhibition assays	239
Solidus	Phase I and/or II Enzymes with cytotoxicity readout Measures requirement for metabolic activation / deactivation	4
Gene Assays	Derived – minimum IC50 / LEL for any assay for a given gene	315
Pathway Assays	Derived – minimum IC50 / LEL for any gene in a given GO or KEGG pathway or process	121
ToxRefDB	Toxicology Reference Database - Guideline animal study data (Endpoints for Prediction)	76

Chronic / Cancer 2-Year Rat Bioassay Endpoints

Endpoint	Endpoint Description
Cholinesterase Inhibition	Cholinesterase inhibition measurement
Kidney Nephropathy	Kidney nephropathy
Kidney Proliferative Lesions	All neoplastic and non-neoplastic proliferative kidney lesions
Liver Hypertrophy	Liver hypertrophy
Liver Necrosis	All effect related to apoptosis and necrosis
Liver Proliferative Lesions	All neoplastic and non-neoplastic proliferative liver lesions
Liver Tumors	All neoplastic liver lesions
Spleen Pathology	All non-neoplastic and neoplastic spleen pathology
Testicular Atrophy	Testis atrophy
Testicular Tumors	All neoplastic testis tumors
Thyroid Hyperplasia	Thyroid hyperplasia
Thyroid Proliferative Lesions	All neoplastic and non-neoplastic proliferative thyroid lesions
Thyroid Tumors	Chronic Rat Endpoint for All neoplastic thyroid lesions
Tumorigen	All neoplastic lesions (any target)

Variable Scaling

- All assays run in concentration-response format
- Curve fitting was performed to derive either:
  - IC50
  - LEL – Lowest Effective Level (lowest concentration at which significant difference from controls was seen)
- Endpoint (in vivo toxicity) values are LEL – lowest dose at which a phenotype was observed
- Concentrations given in µM for assays, mg/kg/day for endpoints
- All chemical-assay or chemical-endpoint combinations with no observed activity given default value of 1000000

Single Variable Correlations

Calculate correlations between each in vitro or derived assay and each endpoint

- Treat both as continuous variables
  - Perform log-10 scaling on both
  - Calculate linear regression statistics
- Treat both as discrete
  - Treat Endpoint LEL < 100mg/kg/day as a "hit"
  - Treat Assay IC50 / LEL < 100 µM as a "hit"
  - Calculate Fisher's exact test statistic

Multi-variable predictions - Continuous Input

<u>Use Machine Learning methods</u>
- SVM – Support Vector Machines
- LDA – Linear Discriminant Analysis
- SLR – Stepwise Logistic Regression (linear and quadratic forms)
- NB – Naïve Bayes
- NNet – Artificial Neural Networks

<u>Treat endpoints as binary</u> – Activity at any observed dose is a "hit"

<u>Variable / Feature Selection</u>
- Use T-test with features selected with p-value <0.1
- Keep no more than 1 variable per 20 cases

Multi-variable predictions - Dichotomous Input

<u>Use Association Rules method - RuleFinder</u>
Creates ruled of the form
IF
(Chemical A is active in assay A and B and ...) or
(Chemical A is active in assay C and D and ...)
THEN
Chemical A causes Endpoint E

Treat input and endpoint as discrete (1 / 0)

For "AND" clause – allow up to 3 assays, Allow up to 3 OR clauses

Use tree-trimming approach to reduce size of search space – only allow rules that have a minimum of 5 "hits"

Genes in KEGG Thyroid Cancer Pathway

NTRK1	neurotrophic tyrosine kinase, receptor, type 1
TP53	tumor protein p53
RET	ret proto-oncogene
TCF7L1	transcription factor 7-like 1 (T-cell specific, HMG-box)
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
TPR	translocated promoter region (to activated MET oncogene)
RXRB	retinoid X receptor, beta
CDH1	cadherin 1, type 1, E-cadherin (epithelial)
MAPK1	mitogen-activated protein kinase 1
MAP2K1	mitogen-activated protein kinase kinase 1
LEF1	lymphoid enhancer-binding factor 1
CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa
MAP2K2	mitogen-activated protein kinase kinase 2
RXRG	retinoid X receptor, gamma
CCDC6	coiled-coil domain containing 6
PPARG	peroxisome proliferator-activated receptor gamma
MAPK3	mitogen-activated protein kinase 3
TCF7	transcription factor 7 (T-cell specific, HMG-box)
RXRA	retinoid X receptor, alpha
TPM3	tropomyosin 3
PAX8	paired box 8
HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homolog
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog
TCF7L2	transcription factor 7-like 2 (T-cell specific, HMG-box)
NCOA4	nuclear receptor coactivator 4
TFG	TRK-fused gene
CCND1	cyclin D1
BRAF	v-raf murine sarcoma viral oncogene homolog B1

Assay	Gene	Endpoint	TP	FP	FN	TN	Sens	Spec	BA	Rel Risk	p(dichot)
GENE_Ache_Rattus_norvegicus	Ache	CHR_Rat_CholinesteraseInhibition	14	4	31	208	0.31	0.98	0.65	6	7.00E-09
GENE_HTR2C_Homo_sapiens	HTR2C	CHR_Rat_KidneyProliferativeLesions	6	3	25	223	0.19	0.99	0.59	6.61	1.30E-04
logP_TPSA		CHR_Rat_LiverHypertrophy	57	109	9	82	0.86	0.43	0.65	3.47	1.10E-05
PATH_KEGG_hsa00980_Metabolism_of_xenobiotics_by_cytochrome_P450		CHR_Rat_LiverNecrosis	20	101	3	133	0.87	0.57	0.72	7.49	4.70E-05
PATH_KEGG_hsa00591_Linoleic_acid_metabolism		CHR_Rat_LiverNecrosis	20	102	3	132	0.87	0.56	0.72	7.38	9.50E-05
GENE_AR_Homo_sapiens	AR	CHR_Rat_LiverNecrosis	16	65	7	169	0.7	0.72	0.71	4.97	9.70E-05
ATG_PPARa_TRANS	PPARA	CHR_Rat_LiverNecrosis	9	19	14	215	0.39	0.92	0.66	5.26	1.70E-04
NCGC_ERalpha_Antagonist	ESR1	CHR_Rat_LiverNecrosis	11	32	12	202	0.48	0.86	0.67	4.56	2.50E-04
BSK_SM3C_Thrombomodulin	THBD	CHR_Rat_LiverNecrosis	8	18	15	216	0.35	0.92	0.64	4.74	6.60E-04
GENE_PLAT_Homo_sapiens	PLAT	CHR_Rat_LiverProliferativeLesions	14	8	52	183	0.21	0.96	0.59	2.88	9.30E-05
ATG_PPARa_TRANS	PPARA	CHR_Rat_LiverProliferativeLesions	16	12	50	179	0.24	0.94	0.59	2.62	1.70E-04
OxidativeStress_24hr	H2AFX	CHR_Rat_LiverProliferativeLesions	13	8	53	183	0.2	0.96	0.58	2.76	2.70E-04
GENE_CYP3A4_Homo_sapiens	CYP3A4	CHR_Rat_LiverProliferativeLesions	10	4	56	187	0.15	0.98	0.57	3.1	2.90E-04
ATG_PPARa_TRANS	PPARA	CHR_Rat_LiverTumors	10	18	13	216	0.43	0.92	0.68	6.29	2.10E-05
CellLoss_72hr		CHR_Rat_ThyroidProliferativeLesions	34	109	9	105	0.79	0.49	0.64	3.01	6.80E-04

AChE / Cholinesterase Inhibition is a "true" association – gives estimate of maximum significance of single-variable effects

Endpoint	Learner	BA(Train)	BA(Test)
CHR_Rat_CholinesteraseInhibition	slr	0.76	0.77
CHR_Rat_LiverNecrosis	nnet	0.96	0.66
CHR_Rat_LiverProliferativeLesions	svm	0.89	0.67
CHR_Rat_TesticularTumors	nnet	0.71	0.55
CHR_Rat_ThyroidProliferativeLesions	nnet	0.92	0.61
CHR_Rat_Tumorigen	nnet	0.77	0.55

(1) Cytotox, **Ache(Rat),** Cyp2c6(Rat), SLC6A3 (Human), RXRa(Human), VCAM1 (Human), Tspo(Rat)

(2) KEGG(Calcium signaling), KEGG(Linoleic acid metabolism), KEGG(Hematopoietic cell lineage), Ache(Rat), AR(Human), NR1H4 / FXR (Human), %absorption, Cytotox

(3) KEGG(Focal adhesion),KEGG(Metabolism of xenobiotics by cytochrome P450) KEGG(Linoleic acid metabolism), KEGG(Epithelial cell signaling / H.pylori infection) KEGG(Cell adhesion molecules), KEGG(Prostate cancer) AR(Human), IP10(Human), Cytotox, HTR2C(Human), 5HT2C(Human), LXR(Human)

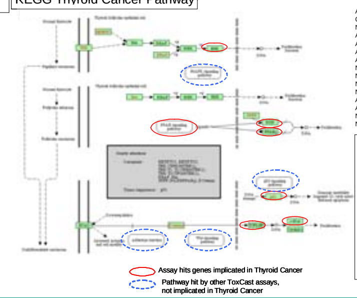
(4) KEGG(Prostate cancer), p53(Human), CAR(Human), CYP3A5(Human), CYP3A4(Human),VCAM1(Human), ERa(Human), Microtubule Destabilization, Cytotox

(5) **KEGG(Thyroid cancer)**, KEGG(PPAR signaling), KEGG(Hematopoietic cell lineage) PLAUR(Human), TA(Human), CXCL10(Human), Mitochondrial Memrane Potential, Cytotox, Cell Cycle Arrest

(6) KEGG(Thyroid cancer), KEGG(Endometrial cancer), KEGG(PPAR_signaling), KEGG(Acute myeloid leukemia), KEGG(Melanoma), KEGG(Glioma), KEGG(Prostate cancer), KEGG(Wnt signaling pathway), p53(Human), PPARA(Human), PPARG(Human), NR1I2/PXR(Human)

Endpoint	TP	FP	FN	TN	Sens	Spec	Relative Risk	p-value	Rule 1	Rule 2	Rule 3
Rat Thyroid Proliferative Lesions	6	24	1	226	0.86	0.90	45.4	1.0E-05	SM3C Proliferation AND KEGG(Focal adhesion) AND KEGG(Thyroid_cancer)	ACEA(Cell growth) AND KEGG(Complement and coagulation cascades)	
Rat Liver Proliferative Lesions	10	74	0	173	1	0.7		9.5E-06	NRF2 AND GABPA AND Bioavailability	PPARG AND KEGG(Limonene and pinene degradation) AND KEGG(Chronic myeloid leukemia)	MCSF AND KEGG(PPAR signaling) AND KEGG(Cell adhesion molecules)

KEGG Thyroid Cancer Pathway



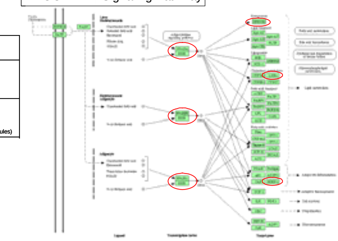
Assays in KEGG Thyroid Cancer Pathway

Attagene Factorial cis p53
Cellumen p53
NCGC p53 assay
Attagene Factorial cis Myc
Attagene Factorial trans RXRb
Attagene Factorial cis PPRE
Attagene Factorial trans PPARg
NCGC Reporter Gene Assay PPARg Agonist
NCGC Reporter Gene Assay PPARg Antagonist
Novascreen Human PPARg
Novascreen Human MAPK3
Attagene Factorial cis TCF/b-cat
Attagene Factorial trans RXRa
NCGC Reporter Gene Assay RXRa Agonist
NCGC Reporter Gene Assay RXRa Antagonist

Assays in KEGG PPAR Signaling Pathway

Attagene Factorial cis DR4/LXRE
Attagene Factorial cis PPRE
Attagene Factorial cis PPRE
Attagene Factorial cis PPRE
Attagene Factorial trans LXRa
Attagene Factorial trans PPARa
Attagene Factorial trans PPARd
Attagene Factorial trans PPARg
Attagene Factorial trans RXRa
Attagene Factorial trans RXRb
BioSeek BrEP IL1b / TNFa / IFNg / MMP1
BioSeek HDFn / IL1b / TNFa / IFNg / EGF / FGF /PDGFbb / MMP1
CellzDirect HMGCS2
NCGC Reporter Gene Assay PPARd Agonist
NCGC Reporter Gene Assay PPARd Antagonist
NCGC Reporter Gene Assay PPARg Agonist
NCGC Reporter Gene Assay PPARg Antagonist
NCGC Reporter Gene Assay RXRa Agonist
NCGC Reporter Gene Assay RXRa Antagonist
Novascreen Human MMP1
Novascreen Human PPARa
Novascreen Human PPARg

KEGG PPAR Signaling Pathway



Conclusions

- The EPA ToxCast program has produced >500 assays on 320 chemicals
- A range of standard statistical techniques have been used to correlate in vitro assays with rat chronic in vivo endpoints
- In vitro interaction with the thyroid cancer pathway is associated with observed thyroid tumors
- In vitro interaction with the PPAR pathway is associated with multiple types of in vivo tumors
- This is a first step towards building predictive in vitro in vivo models for chemical screening and prioritization

This work was reviewed by EPA and approved for publication but does not necessarily reflect official Agency policy.