Learn about me!
Scan to view my profile and learn more about my work.

Searching for LINCS to Stress

Using text-mining to automate reference chemical curation

www.epa.gov

Bryant A. Chambers[1] I chambers.bryant@epa.gov I 919.541.4268
Danilo Basili[2], Laura Word[1], Nancy Baker[1,3], Alistair Middleton[2], Richard Judson[1], Imran Shah[1*]
[1]Center for Computational Toxicology and Exposure, US EPA ; [2]Unilever ; [3]Leidos

View Presentation!
Scan this QR code to watch a recorded presentation of this poster!

Overview

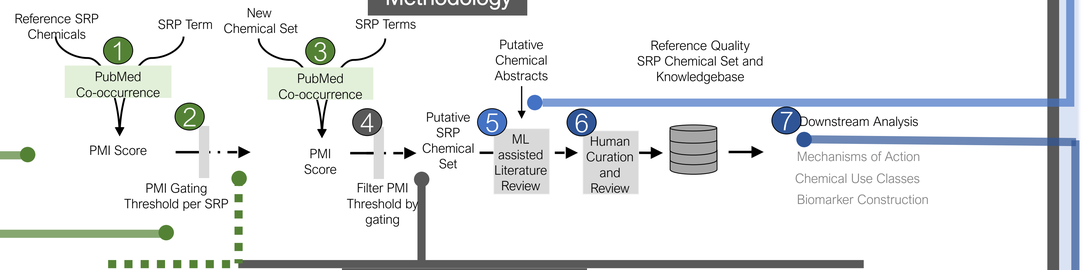
Problem

- Stress Response Pathways (SRPs) are hypothesized to be a key part of non-specific toxicity routes and play a role in disease.
- Chemical annotations for Stress Response Pathways (SRPs) activity are unevenly represented and no formal database exists.
- A database of SRP-chemical associations is needed to inform and support SRP assay development for non-specific toxicity.

Solution

- We developed an approach to automate the curation SRP active chemicals from literature linked to bioinformatics data
- Using an information statistic and machine learning we reduced human burden during curation and identified chemical use classes with significant SRP activity

Methodology

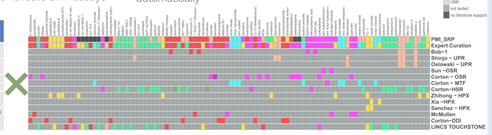


1a Reference Chemicals and Queries were taken from key studies

Established reference chemicals were searched against stress response pathways

- 98 Chemical identified in 12 published works
- Chemicals used previously to validate SRP assays
- Queries linked to SRP medical subject headings (MeSH terms)
- Term expansion includes additional phrases to extend queries automatically

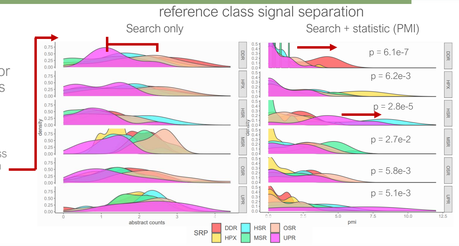
Stress Response Pathway	Query
DNA Damage Response	'dna damage'
Heat Shock Response	'heat shock'
Hypoxia	'hypoxia'
Metals Stress	'metal stress'
Oxidative Stress Response	'oxidative stress'
Unfolded Protein Response	'(endoplasmic reticulum stress OR unfolded protein response)'



1b Mutual information identifies reference class better than search

Only Pairwise Mutual Information (PMI) transformed abstract counts had significance for Chemical-Reference class assignment

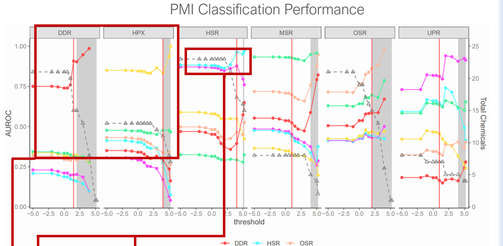
- All significant; p values < 0.05
- Kernel density of reference class separated from all other search terms
- PMI transformed data used for analysis



2 Maximizing performance by chemical gating defines inclusion criteria

Gating criteria was defined by balancing performance and recall with test set thresholds and size

- The highest performance (AUROC) for PMI greater than 1 was selected
- Additionally, we required 70% of original reference chemicals for each SRP class
- PMI annotations for DDR, HPX, and MSR have almost 3x better performance than the
- UPR and HSR are the most difficult to disentangle but share mechanism
- Mechanist similarity is evident (e.g., OSR v. MSR) hinting at a multiclass annotation as well



SRP	Gated PMI Score Threshold	AUROC	Reference Chemical Cases	Fraction of Reference Cases
DDR	1.5	0.91	15	71%
HPX	4	0.83	11	85%
HSR	4	0.96	17	74%
MSR	1	0.93	8	100%
OSR	2	0.86	15	71%
UPR	1	0.81	7	88%

Key Outcome

3 & 4 Applying PMI thresholds to LINCS dataset increases chemical set for down stream analysis

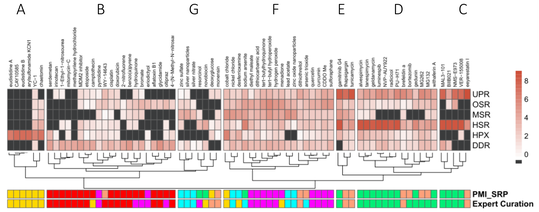
Hierarchical clustering of PMI score vectors exhibit seven activity group in literature data

Single activity

- B:DDR, A:HPX, D:HSR, E:UPR, F:OSR, G:MSR

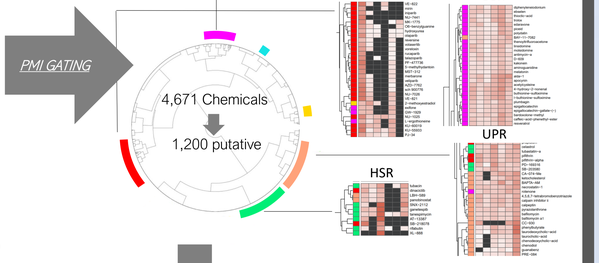
Dual activity and overlapping system

- C: Mixed UPR/HSR G:Mixed MSR/OSR
- Some mixed OSR and DDR
- Mimics mechanistic induction observed in cross-activated and co-induced systems



Using gating criteria to identified 1,200 chemicals in the bioinformatics database with SRP literature support

- DDR, OSR, and HSR were the most represented classes
- MSR and HPX were the least represented classes
- IR identified almost 10x as many SRP active chemicals as are annotated in the database



Key Result

7b SRP biological activity matches from literature to expression

Among the same transcriptomic profiles targets and use classes were enriched by SRP

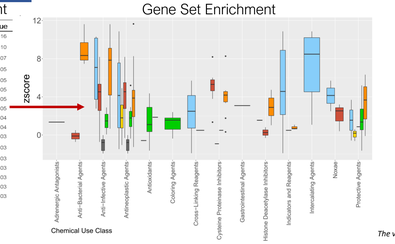
Most surprisingly, scoring with an orthogonal set of SRP biomarkers indicated use-pathway matching

Process identified predicted activities in chemical and automated activity annotation

Literature Enrichment

Use Class	p value
Neurotransmitter Agents	8.08E-16
Peripheral Nervous System Agents	5.44E-10
Central Nervous System Agents	3.94E-07
Central Nervous System Depressants	1.01E-05
Anti-Infective Agents	1.05E-05
Antioxidants	2.89E-05
Psychotropic Drugs	4.26E-05
Anti-Bacterial Agents	5.95E-05
Adrenergic Agents	3.71E-04
Sensory System Agents	4.52E-04
Gastrointestinal Agents	2.28E-03
Poly(ADP-ribose) Polymerase Inhibitors	2.39E-03
Cysteine Proteinase Inhibitors	3.70E-03
Chelating Agents	3.76E-03
Sequestering Agents	3.76E-03
Indicators and Reagents	4.10E-03
Antineoplastic Agents	4.26E-03

Gene Set Enrichment



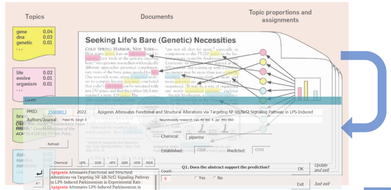
Conclusion

- We annotated 480 additional SRP active chemicals in LINCS
- Literature mining reduces effort in hazard assay development
- Transcriptomic space mirrored literature
- New chemical use group were associated with SRP activity and that activity was validated orthogonally with transcriptomic data
- The process is generalizable beyond SRPs and can build validation and test sets for a variety of problems

The views expressed in this presentation are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency

5&6 ML assisted systematic review prioritizes literature and identifies 480 new chemicals

Topic modeling – human curation overview



SRP	total chemicals	passed	failed	success rate
DDR	131	117	14	89.3
HPX	3	2	1	66.7
HSR	12	11	1	91.7
MSR	3	1	2	33.3
OSR	220	203	17	92.3
UPR	115	91	24	79.1

ML assisted review reduced human burden and prioritized evidence to confirm chemical-SRP relationships

- Multiclass-classifier used to predict topic and relationship between chemical and text
- Negative examples were confirmed using up to 5,000 abstracts per chemical negating SRP language
- 268,000 abstract were reduced to the 2,053 most essential
- Process save almost 240% of person-hours
- UPR and HSR best likely hood with high support rates down to 53% match probability
- OSR had greatest drop of support rate after 68% match probability

7a PMI annotations exhibit unique transcriptomic signals

Transcriptomic profiles of SRP annotated chemicals classified by PMI only cluster by PMI SRP annotation which indicates that literature annotation maps to unique biological spaces

- Similar mechanistic annotations like UPR and HSR clustered together
- Clustering was cell specific indicating some potential differences in cell type SRP response

Transcriptomic profile clustering