Literature-mining and Transcriptomic Stress Response Annotation of a Large Chemical Database

Bryant Chambers[1], Laura Taylor[1], Nancy Baker[2], Richard Judson[1], and Imran Shah[1]

[1]Center for Computational Toxicology and Exposure, US EPA; [2]Leidos

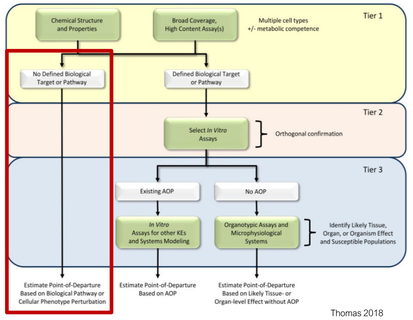
www.epa.gov

Bryant A. Chambers I chambers.bryant@epa.gov I 919.541.4268

Background and Hypothesis

Rationale

- Many environmental chemicals act via non-specific mechanisms:
  - Do not activate molecular initiating events (MIEs)
  - Cannot be related to adverse outcomes (Ankley 2010)
- Overlap between responses obscure discrete reference chemical assignment.
- Currently no SRP knowledge base exists for training SRP classifiers.
- Literature and information retrieval approaches can support SRP annotation
- Existing knowledge bases are:
  - Limited by predefined conceptual space with insufficient SRP annotation
  - Clouded by uneven coverage of SRP context
  - Hand curated requiring extensive person investment



Thomas 2018

Hypotheses

- Information retrieval based cooccurrence coupled to statistic (Pointwise mutual information; PMI) can support SRP annotation
  - Scaled representation
  - Unrestricted conceptual space supporting free text
- Coupling transcriptomic analysis to a well annotated data can improve signature design and inform cell line dependent effects.

Linking a literature database to a transcriptomic database

1

General Methodology

- Couple a large publicly accessible literature database with a large publicly accessible transcriptomic database

Lit Database

PubMed
- 31M Articles
- 3.3B Searchers/y
- Query API / Programmatic Access

2 *Subset a searchable list of chemical*

TRx Database

Library of Integrated Network-based Cellular Signatures (LINCS)
- 20K Chemicals
- 300K Transcriptomic Profiles
- Multi - Time/Conc/Cell
- 12K Genes

ConnectivityMap

3 Information and Text Mining Approaches
Pointwise mutual information
Abstract gene extraction GSEA

4 Transcriptomic Approaches
GSEA profiles
Similarity scoring

5 *Evaluate Accuracy Validated reference set*

6
1. Expanded automated profile clustering
2. model explaining assignments and cell dependency

Cooccurrence frequency table

Entity 1 \ Entity 2	X	Y	Z
A	200	0	123
B	0	2	0
C	523	0	0

$$PMI(A, X) = \log \frac{F(A, X)}{F(A)F(X)}$$

- Annotate chemical activity using PMI scaled retrieval cooccurrence
- Use annotation to assess transcriptomic evaluation using set enrichment (GSEA)

Literature Scoring and Exemplar Chemical Clustering

Approach

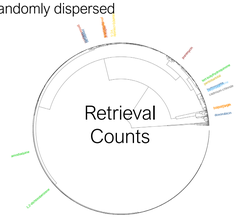
- Manually curated exemplar chemicals
  - Preidentified in key SRP studies
  - Strongly associated with specific pathway
- Source retrieval cooccurrence frequency
- Transform with PMI
- Cluster chemicals by cooccurrence features

Stress Response Pathway		Chemical
DNA Damage Response	DDR	benzo(a)pyrene, etoposide, mitomycin-c
Heat Shock Response	HSR	radicicol, geldanamycin, bortezomib
Hypoxia	HPX	cobalt II chloride, YC-1
Metals Stress	MSR	cadmium chloride
Oxidative Stress Response	OSR	tert butylhydroquinone, 1,2, dichlorobenzene, amodiaquine
Unfolded Protein Response	UPR	brefeldin-a, thapsigargin, tunicamycin

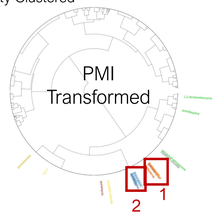
Outcomes

- Exemplar chemicals cluster by SRP
- PMI scores cluster better than search/cooccurrence frequency only
- Short information better than longer information vector
- Chemicals near exemplar chemicals evaluated and found to have similar activity in literature

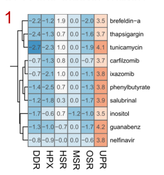
Randomly dispersed



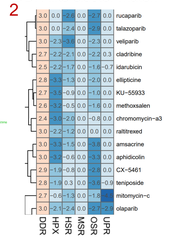
Activity Clustered



UPR-like cluster match



DDR-like cluster match



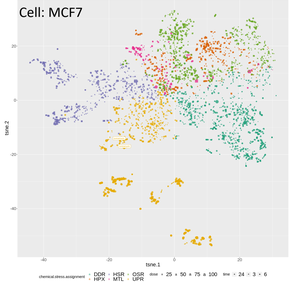
Predicted chemical activity transcriptomic clustering

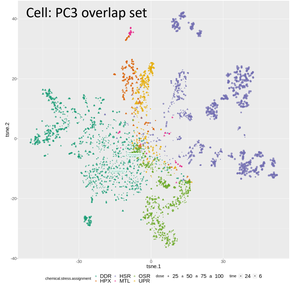
Approach

- Chemical TRx profiles annotated with PMI predicted
- All chemical > PMI 1-1.5 selected
- Two cell types with most abundant profiles selected
- t-sne clustering of transcriptomes

Outcome

- Profiles generally cluster by PMI assignment
- MCF7 HPX present but absent in PC3
  - Potential role of ERS1 increase in basal increase of HPX genes
- OSR and DDR cluster together more than protein misfolding SRPs
- UPR and HSR overlap
- Lower doses are more generally shared between all SRPs





Accuracy against a hand curated validation set

Approach

- Curated 93 chemical set
  - Seeded using literature search results
- Hand validated:
  - 5 reference per chemical
  - 2 reviewers per chemical
  - 68 surviving after review
  - Presence of positives and negatives in set
    - Pathway activating
    - Pathway protective (*e.g.*, chelators)
- Activity scored by PMI annotation and GSEA
- Accuracy evaluated by matching assignments within top n ranked scores
  - GSEA scores aggregated as median across complete set

Outcome

- Good matching between PMI and Validated Annotation
  - 70% top ranked, 80% by top two
- Poor matching between Signatures and Validated Annotation
  - 35% top ranked

Role of cell line in GSEA activity assignment

Approach

- Aggregate concertation and time by finding 5th, 50th, and 95th percentile scores for each chemical and cell type
- Evaluate performance as accuracy and find AUROC as each cell model and signature

Outcome

- Overall accuracy improves when considering cell type
- Specific cell models are more accurate for a given SRP
  - PC3 is generally the best model
  - Adaquate in PC3, MCF7 and HEPG2

PMI Activity Scoring

SRP	Top Ranked	Top 2	Top 3
DDR	100%	100%	100%
HSR	63%	82%	90%
HPX	100%	100%	100%
MSR	0%	0%	0%
OSR	56%	100%	100%
UPR	100%	100%	100%

Cell independent GSEA Activity Scoring

SRP	Top Ranked	Top 2	Top 3
DDR	7%	7%	14%
HSR	64%	82%	90%
HPX	0%	0%	0%
OSR	0%	17%	33%
UPR	100%	100%	100%

Cell and scoring dependent GSEA Activity

SRP	Top Ranked	Top 2	Top 3
DDR – PC3	43%	71%	90%
HSR – NPC	38%	38%	50%
HPX - HEPG2	0%	100%	100%
OSR - PC3 p5	50%	100%	100%
UPR – HCC515 50p	25%	50%	50%

Conclusions and future directions

Key Conclusions

- Information retrieval approaches adequately support activity annotation
- Data cluster better by PMI transformed data
- Transcriptomic profiles show a base level of clustering using automated assignment
- GSEA scoring is cell type dependent
- Transcriptomic profile clustering indicate some native profile similarity that is lost in signatures

Future Directions

- Boot strap signature development with fully automated PMI assignment
- Expression of stress response systems is partially dependent on cell and tissue type; as such, a deeper understanding of tissue dependency must be achieved.

References

Ankley (2010) Environmental Toxicology and Chemistry, 29: 730-741
Simmons (2009)*Toxicological Sciences* 111(2,) 202–225
Judson (2016) *Toxicological Sciences* 152(2):323-339
Shah (2016) *Environ Health Perspect.* 124(7):910-9
Thomas (2019) *Toxicological Sciences* 169(2):317–332
Stathias (2020) *Nuc. Acids Res.* 48(D1):431-439