Chemical-Gene Interactions from ToxCast Bioactivity Data Expands Universe of Literature Network-Based Associations

Sean Watford, Imran Shah, Richard Judson, Matt Martin

National Center for Computational Toxicology, U.S. Environmental Protection Agency, Office of Research and Development

Sean Watford | watford.sean@epa.gov | 919-541-7655

Abstract

Characterizing the effects of chemicals in biological systems is often summarized by chemical-gene interactions, which have sparse coverage in literature. The ToxCast chemical screening program has produced bioactivity data for nearly 2000 chemicals and over 450 gene targets. To evaluate the information gained from the ToxCast project, a ToxCast bioactivity network was created comprising ToxCast chemical-gene interactions based on assay data and compared to a chemical-gene association network from literature. The literature network was compiled from PubMed articles (excluding ToxCast publications) mapped to genes and chemicals. Genes were identified by curated associations available from NCBI while chemicals were identified by PubChem submissions. The frequencies of chemical-gene associations from the literature network were log-scaled then compared to the ToxCast bioactivity network. In total, 140 times more chemical-gene associations were present in the ToxCast network, highlighting that many chemical-gene associations in the ToxCast network had no previously existing association in the literature. There were 165 associations found in the literature network that were reproduced by ToxCast bioactivity data, and 336 associations in the literature network did not correlate with the ToxCast bioactivity network. These findings suggest ToxCast can greatly increase the number of defined chemical-gene associations The literature network relies on publication bias such that chemical-gene associations are assumed to represent bioactivity. Without manual curation or natural language processing methods these literature-based associations cannot be specifically qualified. Meanwhile, the ToxCast bioactivity network establishes chemical-gene associations based directly on *in vitro* data providing broader coverage and reliable data that does not require manual curation. This approach can contribute to the characterization of chemical-gene associations and help identify gaps in data to inform future planning of chemical screening.*This abstract does not necessarily reflect U.S. EPA policy.*

Objectives

- Develop a comprehensive resource that maps chemical-gene bioactivities from literature
- Use measures of co-occurrence to evaluate existing chemical-gene bioactivity resources
- Integrate ToxCast chemical-gene bioactivities to evaluate information gained from the project

Genes and Chemicals in Literature

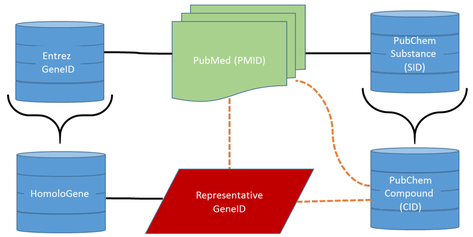


Figure 1: Integration of publicly available biomedical resources
Publicly available resources from NCBI are integrated via direct ( ) and indirect ( ) connections. The direct connections are provided as standalone resources provided in Table 1. Figure 3 provides an example of an indirect connection between a chemical and gene.

Table 1: Summary of Resources for mapping

Name	Source Name (Link)
Entrez GeneID - PubMed	gene2pubmed (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA)
HomoloGene	HomoloGene (ftp://ftp.ncbi.nlm.nih.gov/pub/HomoloGene)
PubChem Substance	PubChem (ftp://ftp.ncbi.nlm.nih.gov/pubchem/)

Table 2: Summary of Chemical-Gene Co-occurrence (CGco) Frequencies

	CIDs	Rep Genes	CGco
Baseline	20,179	692,499	5,577,613
χ^2 p << 0.05	20,068	672,957	4,916,354
PMID Freq > 1	5,455	34,223	222,965
χ^2 p << 0.05 and PMID Freq > 1	5,418	27,135	118,047

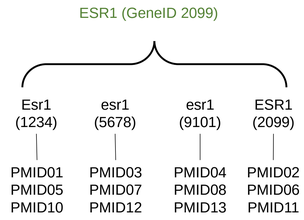


Figure 2: Gene grouping for representative gene
HomoloGene was used to group genes for species non-specific mappings to literature. The HomoloGene ID (HID) is replaced with a human GeneID if it exists in the HID group. If a human gene is not present, a random GeneID is chosen. If the GeneID is not in HomoloGene, then the GeneID represents itself.

Table 1: A summary of the publicly available NCBI resources from which all data for this study were obtained. Entrez Gene provides gene curated GeneID-PMID mappings. HomoloGene contains genes grouped together based on sequence similarity. PubChem provides curated PubChem SID-PMID mappings.

Table 2: A summary of the genes and chemicals along with the co-occurrence frequency available in the integrated NCBI resources. As constraints for significance are placed on the dataset, the number of chemicals and genes decrease with the co-occurrence frequency. A majority of the co-occurrences are supported by only one PMID.

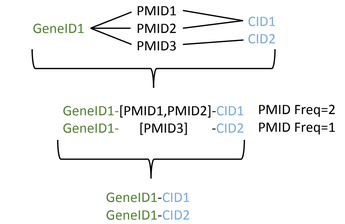


Figure 3: Representative GeneID-CID indirect relationships via PMIDs GeneIDs and CIDs co-occurrences are formed via direct mappings to PMIDs. Under the assumption of publication bias, GeneID-CID co-occurrences with a higher PMID frequency have a stronger association that could be inferred as bioactivity.

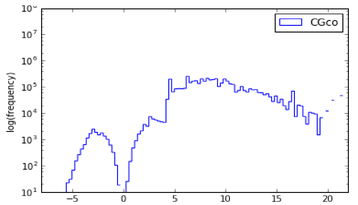


Figure 4: PMI distribution for GeneID-CID co-occurrences
Pointwise Mutual Information (PMI) is a measure of co-occurrence typically used for collocation extraction in natural language processing. PMI is a rank measure, not a significance measure, that provides a general idea of which co-occurrences provide more information than the others.

Challenges with Bioactivity Identification

Table 3: Exact NCBI CID and GeneID mapping overlap with CTD

INPUT: only CTD PMIDs (50K)	CTD	
NCBI	count	%
CID and GeneID	190	0.02
CID and !GeneID	765,238	98.2
!CID and GeneID	19	0.002
!CID and !GeneID	13,707	1.8
Total CID-GeneID	779,154	100.0

Table 3: Comparative Toxicogenomics Database (CTD) is used for comparison to NCBI to identify bioactivities. Very few true positives (CID and GeneID) exists when comparing NCBI Chemical-Gene co-occurrences (CGco) with CTD Chemical-Gene Interactions (CGixn). The chemical mappings from NCBI overlap really well with the Chemicals identified in the CTD articles. NCBI has almost no exact overlap of GeneIDs, which leads to the small number of true positives.

Table 4: Summary of CTD Data

GeneID	39K
Rep GeneID	38K
CGixn	1.1M
Rep. CGixn	610K
PMID Freq > 1	101K

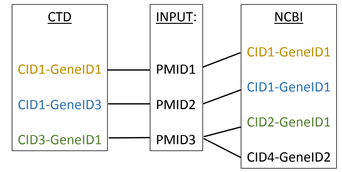
Table 4: Data from CTD was downloaded on 29 December 2015 from http://ctdbase.org/downloads. Data summarized is only from the CTD_chem_gene_ixns.tsv file. Gene grouping via HomoloGene was implemented in order to compare with the NCBI dataset. As in the NCBI dataset, The number of Chemical-Gene interactions decrease with PMID Freq constraint.

Table 5: PMID non-specific overlap of CGixn and CGco

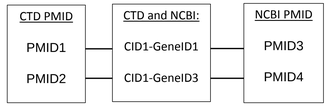
		CTD	NCBI	CGixn and CGco
Baseline	PMIDs	50,474	223,342	**74,649**
	CIDs	9,153	20,179	**3,639**
	GeneIDs	38,108	692,499	**4,202**
	CGixn/CGco	610,760	692,499	**19,587**
PMID Freq > 1	PMIDs	42,859	166,052	**61,144**
	CIDs	3,622	5,418	**1,374**
	GeneIDs	15,375	27,135	**1,772**
	CGixn/CGco	101,908	118,047	**5,812**

Table 5: The PMID non-specific overlap between NCBI and CTD (Figure 6) is used when comparing the two datasets due to the small exact overlap. Out of the 19,587 Chemical-Gene mappings that overlap, only 5,812 were used to calculate PMI because those are the most likely CGco with true bioactivites.

Figure 5 A. Examples of exact matches between NCBI and CTD



B. Examples of PMID non-specific mappings between NCBI and CTD



C. PMI distributions for CGco and (CGco and CGixn)

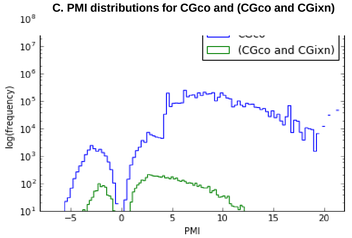


Figure 5:
A. Examples of exact NCBI CID and GeneID mapping overlap with CTD from Table 3.
B. Examples of PMID non-specific mappings between NCBI and CT
C. PMI distributions of both NCBI and the PMID non-specific overlap. The distributions are the same, which means either better GeneID-PMID mappings need to be established or a different measure for comparison needs to be used.

ToxCast Bioactivities

Figure 6: ToxCast Chemical-Gene mappings via Assays

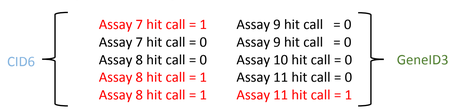
Figure 6: ToxCast chemicals are uniquely identified with a DSSToxID (Distributed Structure-Searchable Toxicity Database). DSSTox chemicals are mapped to PubChem CIDs via exact inchi match. CIDs are then mapped to GeneIDs by aggregating assays with a common GeneID as the target.

Table 6: Summary of ToxCast Chemical-Gene Activity

CID-GeneID actives (CGact)	47,423 (165)
CIDs	5,011
GeneIDs	321

Table 6: A summary of ToxCast Chemical-Gene activity dataset. When compared to the baseline NCBI dataset, only 165 ToxCast activities are supported with literature.

Figure 7: Calculating ToxCast Chemical-Gene activities



Average of all hit calls = (1+0+0+1+1+0+0+0+0+1)/10 = 0.4 = Active

Figure 7: Chemical-Gene activities are calculated by taking the average binary hit call of all aggregated assays.

Future Work

- Improve Gene-PMID mappings for larger recapturing of CTD GeneIDs
  - Other Resources: GeneRif, UniProt, OMIM, etc
  - MeSH Overrepresentation
- Group PubChem Compounds via PubChem related Compounds
- Use different measures of significance or ranking for NCBI bioactivity identification
  - nPMI, tf-idf, F-score, etc
- Improve DSSTox to PubChem Compound mapping for more ToxCast coverage
  - "fuzzy" InChI match over exact
- Improve methods for identifying ToxCast Chemical-Gene bioactivity
  - Incorporate cytotoxicity results

Image source: http://www.inchi-trust.org/technical-faq/

Conclusions

- No Gold Standard exists for mapping or associating Chemical-Gene bioactivities. Existing manually curated resources have low throughput resulting in extremely few replicates. Improvements can be made by incorporating systematic approaches to increase PMID frequencies.
- PubChem is a reliable resource for chemical identifiers mapped to articles.
- Currently available gene mappings are lacking within relevant toxicological publications.