Bioinformation
www.bioinformation.net

open access
Database

GeneComps and ChemComps: a new CTD metric to identify genes and chemicals with shared toxicogenomic profiles

Allan Peter Davis[1*], Cynthia G. Murphy[1], Cynthia A. Saraceni-Richards[1], Michael C. Rosenstein[1], Thomas C. Wiegers[1], Thomas H. Hampton[2] and Carolyn J. Mattingly[1]

[1]Department of Bioinformatics, The Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672, USA; [2]Center for Environmental Health Sciences, Dartmouth Medical School, Hanover NH 03755, USA; Allan Peter Davis – Email: apd@mdibl.org; *Corresponding author

Received September 10, 2009; Accepted October 13, 2009; Published October 15, 2009

Abstract:
The Comparative Toxicogenomics Database is a public resource that promotes understanding about the effects of environmental chemicals on human health. Currently, CTD describes over 184,000 molecular interactions for more than 5,100 chemicals and 16,300 genes/proteins. We have leveraged this dataset of chemical-gene relationships to compute similarity indices following the statistical method of the Jaccard index. These scores are used to produce lists of comparable genes ("GeneComps") or chemicals ("ChemComps") based on shared toxicogenomic profiles. GeneComps and ChemComps are now provided for every curated gene and chemical in CTD. ChemComps are particularly significant because they provide a way to group chemicals based upon their biological effects, instead of their physical or structural properties. These metrics provide a novel way to view and classify genes and chemicals and will help advance testable hypotheses about environmental chemical-gene-disease networks.

Availability: CTD is freely available at http://ctd.mdibl.org/

Keywords: gene, chemical, toxicogenomic, database, curation.

Background:
The Comparative Toxicogenomics Database (CTD) is a public resource that promotes understanding about the effects of environmental chemicals on human health **[1]**. CTD biocurators manually curate interactions from the scientific literature in a structured format using controlled vocabularies and ontologies for chemicals, genes, diseases, molecular interactions, and organisms **[2, 3]**. These datasets can be used to explore relationships and also to generate novel, testable hypotheses about chemical-gene-disease pathways. Finding chemicals and genes with similar interaction profiles could promote alternative methods for classifying chemicals and help identify additional members of interaction networks. There are many ways to discover and organize related genes and proteins: sequence similarity, co-expression analysis, shared protein-protein interactions, and common biomarkers for a disease. In addition, the Gene Ontology (an annotation vocabulary used to define gene products by molecular function, biological process, and cellular localization) can be data-mined to produce lists of comparable genes **[4]**. In contrast, criterion for finding similar chemicals has been largely restricted to the physical properties (*e.g.*, molecular weight, atomic elements, boiling point, molar volume, *etc*.) or atomic structure of compounds **[5]**. A relatively new approach, however, is to classify chemicals based upon their effect on mRNA expression detected by microarrays **[6, 7]**. While this method and the development of extensive chemical vocabularies **[8]** and new ontologies **[9]** may further help organize compounds, comparative analysis among chemicals with similar gene interaction profiles (beyond mRNA expression) is still lacking.

At CTD we developed a simple approach to discover analogous genes and chemicals based upon shared chemical-gene interaction profiles, which we call GeneComps and ChemComps for comparable genes and comparable chemicals, respectively. CTD biocurators manually curate the literature and annotate over 50 different types of chemical-gene molecular interactions, including, among others, effects on mRNA expression, protein expression, phosphorylation, activity, localization, degradation, metabolic processing, transport, and promoter methylation. Every gene in CTD has a profile of chemicals with which it interacts, and likewise every chemical has a profile of genes with which it interacts. These extended, detailed, and more complete interaction profiles essentially define a "footprint" for a gene or chemical that can be leveraged to discover and cluster genes and chemicals.

Methodology:
We used the data available in CTD in September 2009, which included 184,646 chemical-gene interactions for 5,124 chemicals and 16,308 genes. Similarity indices were computed for chemicals (ChemComps) and genes (GeneComps). The degree of similarity was estimated using a modification of the Jaccard index, whose value ranges between 0 and 1 **[10]**. The index is the ratio of the number of elements in the intersection of two sets (number of shared interactions between two chemicals or genes) divided by the number of elements in the union of two sets (number of combined interactions between two chemicals or genes).

Utility:
CTD computes statistics that reflect the degree of similarity between the gene interaction profiles of each curated chemical and generates a list of ChemComps (**Figure 1a)**. Likewise, chemical interaction profiles are compared between each curated gene to produce GeneComps. ChemComps and GeneComps provide a simple approach to view chemicals and genes that share interaction profiles. ChemComps especially provide a novel way to classify and organize chemicals based upon biological effects, which can be considered a molecular signature or footprint. Every curated chemical in CTD now includes a ChemComps data tab that lists the top 20 comparable chemicals based upon their ranked similarity index derived from their interaction profile. For example, the chemical bisphenol A (a plastic additive) has curated interactions to 473 genes in CTD. ChemComps lists comparable chemicals that share the most number of interactions with those 473 genes to produce a ranked list that includes polychlorinated biphenyls, genistein, and estradiol (**Figure 1b**), suggesting that bisphenol A shares many of the networks common to these three chemicals. Similarly, GeneComps are displayed on curated gene pages and lists the genes that share a chemical profile.

ChemComps and GeneComps are datasets that allow researchers to view and cluster chemicals and genes with similar biological activities. This new metric at CTD provides researchers with additional predictive information that will help construct novel, testable hypotheses about chemical-gene-disease pathways.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4 (4):173-174 (2009)

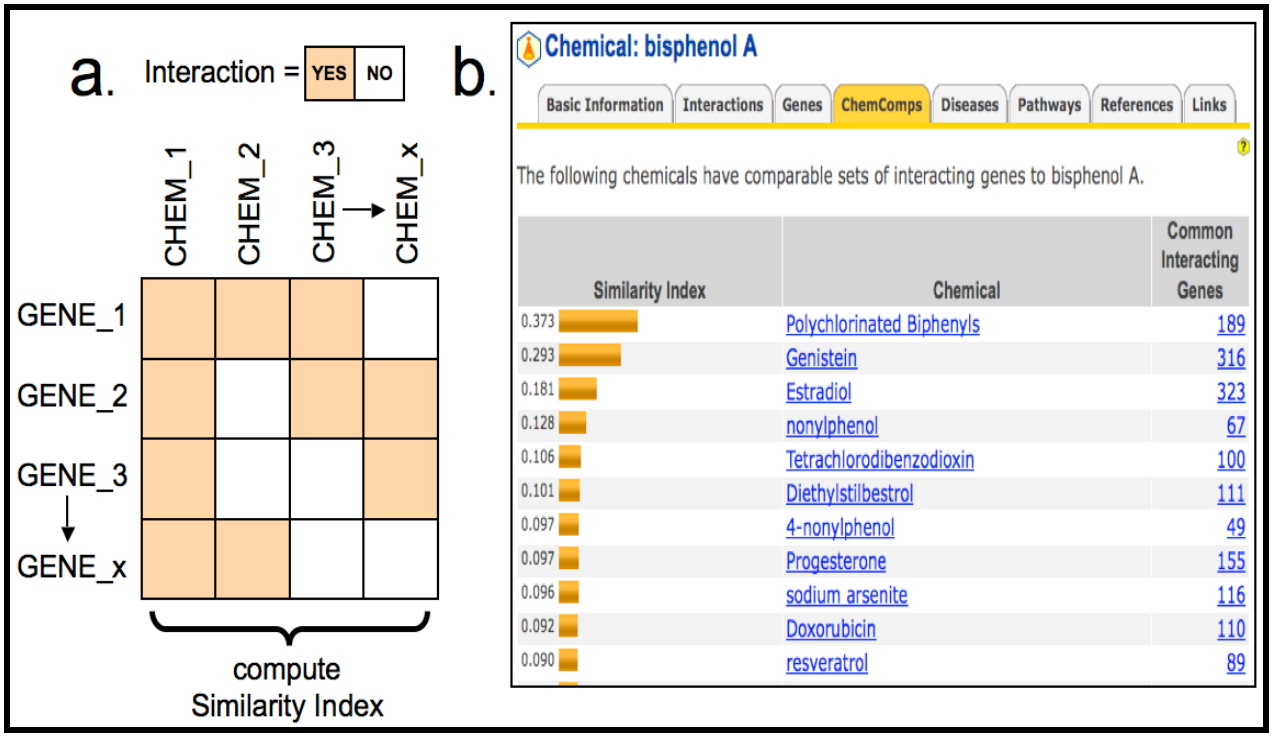

Figure 1: A contingency table of chemical-gene interactions computes the similarity indices (a) to rank comparable chemicals for a user's molecule of interest, such as bisphenol A, which are listed under the ChemComps data tab in CTD (b).

Future development:
Currently, the similarity indices used to generate GeneComps and ChemComps are derived exclusively by looking for either a "yes" or "no" binary interaction relationship between a gene and chemical (**Figure 1a**). However, CTD biocurators capture many details about the interactions between these two types of molecules, such as "chemical Y decreases the phosphorylation of protein B" or "protein C results in chemical resistance to chemical Z", *etc*. **[2, 3]**. These specific details might be leveraged to derive more qualitative similarity indices, such as finding comparable chemicals that increase *vs*. decrease the phosphorylation of a protein, or increase *vs*. decrease the methylation of a gene's promoter.

Acknowledgement:
This work was supported by the National Institute of Environmental Health Sciences [R01ES015065 and P42ES007373] and the Idea Network of Biomedical Research Excellence program of the National Center for Research Resources [P2RR016463] of the National Institutes of Health.

Reference:
[1] CJ Mattingly *et al*., *Toxicol Sci.* **92**: 587 (2006) [PMID: 16675512]
[2] AP Davis *et al*., *BMC Med Genomics.* **1**: 48 (2008) [PMID: 18845002]
[3] AP Davis *et al*., *Nucleic Acids Res.* **37**: D786 (2009) [PMID: 18782832]
[4] M Ashburner *et al*., *Nat Genet.* **25**: 25 (2000) [PMID: 10802651]
[5] RP Sheridan, SK Kearsley, *Drug Discov Today*. **7**: 903 (2002) [PMID: 12546933]
[6] J Lamb *Nat Rev Cancer*. **7**: 54 (2007) [PMID: 17186018]
[7] CR Williams-Devane *et al*., *Toxicol Sci.* **109**: 358 (2009) [PMID: 19332651]
[8] http://www.nlm.nih.gov/mesh/
[9] K Degtyarenko *et al*., *Nucleic Acids Res.* **36**: D344 (2008) [PMID: 17932057]
[10] http://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=303769606

Edited by P. Kangueane
Citation: Davis *et al.*, Bioinformation 4(4): 173-174 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4 (4):173-174 (2009)