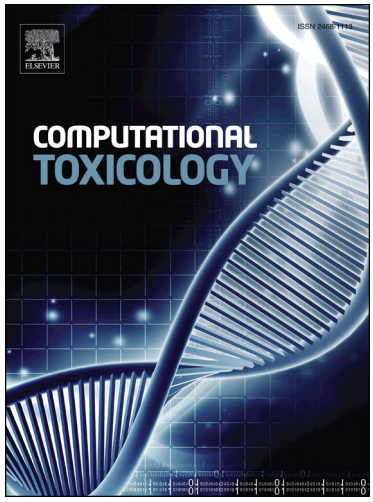

Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis

Sean M. Watford, Rachel G. Grashow, Vanessa Y. De La Rosa, Ruthann A. Rudel, Katie Paul Friedman, Matthew T. Martin

PII: S2468-1113(18)30029-X
DOI: https://doi.org/10.1016/j.comtox.2018.06.003
Reference: COMTOX 42

To appear in: *Computational Toxicology*

Received Date: 3 April 2018
Revised Date: 8 June 2018
Accepted Date: 18 June 2018

Please cite this article as: S.M. Watford, R.G. Grashow, V.Y. De La Rosa, R.A. Rudel, K.P. Friedman, M.T. Martin, Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis, *Computational Toxicology* (2018), doi: https://doi.org/10.1016/j.comtox.2018.06.003

Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis

Authors: Sean M. Watford[1,2], Rachel G. Grashow[3,4], Vanessa Y. De La Rosa[3,5], Ruthann A. Rudel[3], Katie Paul Friedman[7,] Matthew T. Martin[6,7]

1 ORAU, contractor to U.S. Environmental Protection Agency through the National Student Services Contract, Oak Ridge, TN.
2 Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, UNC-Chapel Hill, Chapel Hill, North Carolina, United States.
3 Silent Spring Institute, Newton, MA.
4 Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA
5 Social Science Environmental Health Research Institute, Northeastern University, Boston, MA
6 U.S. Environmental Protection Agency, National Center for Computational Toxicology, Research Triangle Park, NC, USA.
7 Currently at Pfizer Worldwide Research & Development, Groton, CT, USA.

Corresponding author:
Matthew Martin
Previously at US EPA
Now at Pfizer Worldwide Research and Development
445 Easter Point Rd, MS 8274-1224
Groton, CT 06340
Phone: 860-441-4510
Email: Matthew.Martin@pfizer.com

Keywords:
Biomedical Literature; Data Integration; Genes; Breast Carcinogenesis; Chemical Exposures; Literature Mining

Highlights (3-5 highlights, 85 characters including spaces maximum):

- Biomedical text resources were integrated to identify gene-disease associations
- Normalized pointwise mutual information was used to identify and rank genes linked to key carcinogenic characteristics
- A relevant breast cancer gene set ranked higher than random gene sets
- Methods scale to include other biological and chemical concepts

***Disclaimer**: The United States Environmental Protection Agency (U.S. EPA) through its Office of Research and Development has subjected this article to Agency administrative review and approved it for publication. Mention of trade names or commercial products does not constitute endorsement for use. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the US EPA.*