

## Defining the Chemical Space of Public Genomic Data

**Authors:** ClarLynda Williams<sup>1</sup>, Maritja Wolf<sup>2</sup>, Ann Richard<sup>3</sup>

<sup>1</sup>North Carolina State University/U.S. EPA COOP/ Bioinformatics, United States

<sup>2</sup>Lockheed Martin (Contractor to U.S. EPA), United States

<sup>3</sup>U.S. EPA/Office of Research and Development (ORD)/National Center for Computational Toxicology (NCCT)

**Keywords:** chemogenomics, database, predictive toxicology, chemical index, public genomic data

The pharmaceutical industry has demonstrated success in integrating chemogenomic knowledge into predictive toxicological models, due in part to industry's access to large amounts of proprietary and commercial reference genomic data sets. The environmental regulatory domain heretofore has lagged behind in such efforts due to reliance on relatively small amounts of in-house data and publicly available genomic databases. We have surveyed over 20 public genomic data repositories/databases, the majority of which contain functional and physiological genomic data. Of these, only five contain data pertaining to chemical exposure: National Institutes of Environmental Health Science's Chemical Effects in Biological Systems (CEBS) knowledgebase, Public Expression Profiling Resources (PEPR) web database, European Bioinformatics Institute's ArrayExpress genomic repository, the National Center for Biotechnology Information's GEO repository, and the Environment, Drugs, and Gene Expression database (EDGE). The current project aims to chemically index the genomics content of these databases to make these data accessible in relation to other publicly available, chemically-indexed, toxicological information. CEBS and EDGE are currently chemically indexed, but contain information on relatively few chemical exposure experiments. The other genomic resources, ArrayExpress, GEO, and PEPR, are based on three different data structures. Hence, it was necessary to develop three different methodologies for mining the author-submitted content to support the chemical indexing process. These methodologies consist of a series of Perl programs that transform these text files into mineable toxicogenomic, chemically-indexed data files. By defining the chemical space of public genomic data, it becomes possible, for the first time, to assess the scope of chemical coverage of these data, and to identify classes of chemicals, or neighborhoods of similar chemicals, having sufficient data to support methodologies for the integration of chemogenomic data into predictive toxicology. These methodologies will also have to deal with the problems of comparing experimental data across diverse sources (i.e., labs, chemicals, and species). The chemical space of public genomic data will be presented along with the methodologies and tools used to identify this chemical space. Progress towards developing methods to deal with the problems of integrating public data from diverse sources will also be reported.

### Point of Contact:

ClarLynda Williams-DeVane

Graduate Student

North Carolina State University/U.S. EPA COOP

MD343-03

Research Triangle Park, NC 27711

919-541-0814

williams.clarlynda@epa.gov