Defining the Chemical Space of Public Genomic Databases

Authors: ClarLynda Williams[1], Maritja Wolf[2], Ann Richard[3]
[1]North Carolina State University/U.S. EPA COOP/ Bioinformatics, United States
[2]Lockheed Martin (Contractor to U.S. EPA), United States
[3]U.S. EPA/Office of Research and Development (ORD)/National Center for Computational Toxicology (NCCT)

The current project aims to chemically index the genomics content of public genomic databases to make these data accessible in relation to other publicly available, chemically-indexed toxicological information. There are currently more than 20 public genomic data repositories/databases, five of which contain data of chemogenomic interest: the Chemical Effects in Biological Systems (CEBS) knowledgebase; Public Expression Profiling Resources (PEPR) web database; ArrayExpress genomic repository; the Gene Expression Omnibus (GEO) repository; and the Environment, Drugs, and Gene Expression database (EDGE). CEBS and EDGE are currently chemically indexed, but contain information on relatively few chemical exposure experiments. ArrayExpress and GEO contain more chemical exposure experiments but require the development of new methodologies to mine the author-submitted content. These methodologies consist of a series of Perl programs and manual annotations developed to transform these text files into mineable toxicogenomic, chemically-indexed data files. PEPR content is mirrored in GEO and does not require additional methodology. After chemical exposure experiments are identified, the chemical space is defined through structural similarity. By defining the chemical space of public genomic data, it is possible to identify classes of chemicals on which to develop methodologies for the integration of chemogenomic data into predictive toxicology. These methodologies will consist of methods to compare the experimental data across labs, chemicals, platforms and species. The chemical space of public genomic data will be presented as well as the methodologies and tools developed to identify this chemical space. *This work does not necessarily reflect official U.S. EPA policy.*

To be presented:
Conference: 6th Annual International Conference on Computational Systems Bioinformatics CSB2007
Date: August 13-17, 2007
Location: San Diego, California
Presentation Type: Poster