Abstract submitted for International Science Forum on Computational Toxicology
RTP, NC May 21-23, 2007

Defining the Chemical Space of Public Genomic Data

ClarLynda Williams[1], Maritja Wolf[2], Ann Richard[3]

[1] NC State University Bioinformatics Graduate Program, EPA Student COOP, Raleigh, NC, USA
[2] Lockheed Martin, Contractor to the US EPA, RTP, NC, USA
[3] National Center for Computational Toxicology, US EPA, RTP, NC, USA

The pharmaceutical industry has demonstrated success in integrating of 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. There are currently more than 20 public genomic data repositories/databases, but only five of the 20 contain data of potential chemogenomic interest: 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.