

Defining the Chemical Space of Public Genomic Data

ClarLynda Williams-DeVane¹, Maritja Wolf², Ann Richard³. ¹North 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)

TOXICOGENOMICS

Results

Project Goals

- Chemically annotate the National Center for Biotechnology Information's (NCBI's) Gene Expression Omnibus (GEO)⁵, the European Bioinformatics Institute's (EBI's) ArrayExpress¹, Environment, Drug, Gene Expression (EDGE)⁶ and the National Institute of Environmental Health Sciences CEBS³ genomic data repositories.
- Create a Structure Index of the 25 data repositories that can integrate these data with other public data sources (e.g., PubChem⁸).
- Populate Structure Index files with historical toxicological and chemical data from DSSTox Data Sources through cross referencing.
- Explore methodologies to address concerns of public genomic data use, including: across laboratory and platform comparisons, as well as extrapolation between species, doses, chemicals, and endpoints.
- Restructure Genomic Data for insertion into a Chemogenomic Database.
- Use Chemogenomic Database in conjunction with both developed and adopted methodologies, as well as data mining tools, to guide future experiments, discern new patterns of toxicological interest, and explore new hypotheses related to toxicological potential.

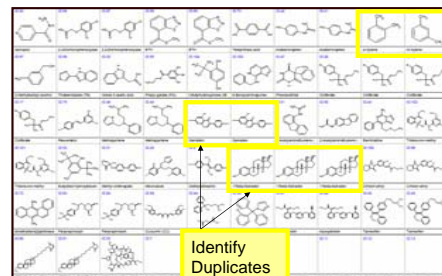
EDGE Database



Statistics

- 25 Records
- 18 Chemical Structures
- 8 Chemical Structures overlap GEOMIS (GEO)
- 7 Chemical Structures overlap AREXCH (ArrayExpress)
- 6 Chemical Structures overlap CSTARC (CEBS)
- Chemical Structures overlap 82 Toxicological Records

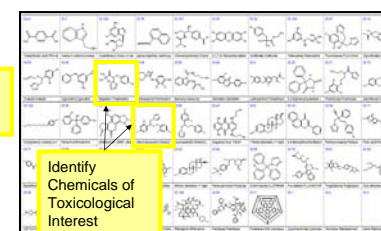
ArrayExpress Database



Statistics

- Initially 106 Chemical Exposure Records
- Currently 351 Chemical Exposure Records
- 106 Chemical Structures overlap 273 Toxicological Records

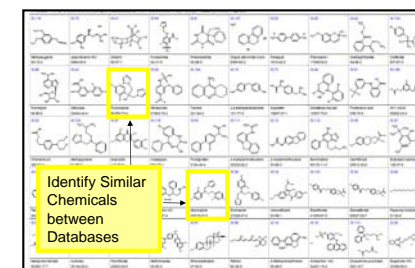
GEO Database



Statistics

- Initially 129 Chemical Exposure Records
- Currently Over 500 Chemical Exposure Records
- 129 Chemical Structures overlap 246 Toxicological Records

CEBS Database

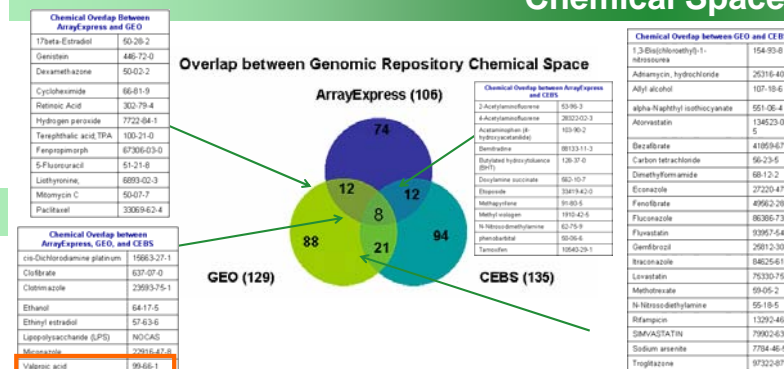


Statistics

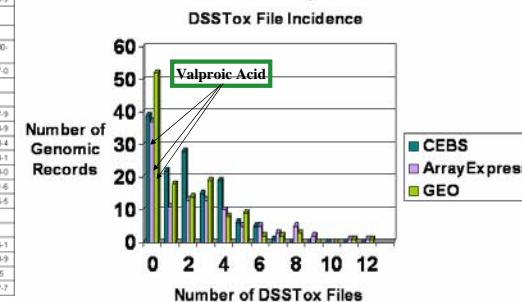
- Initially 135 Chemical Exposure Records
- Currently All Records contain Toxicology Information
- Chemical Structures overlap 538 Toxicological Records

ArrayExpress Structure Index (SI) File

Chemical Space



Overlap between Genomic Chemical Space and DSSTox Chemical Space

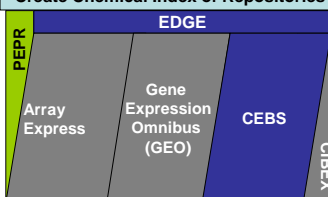


Methods/Approaches

Identification of Genomic Repositories and Databases

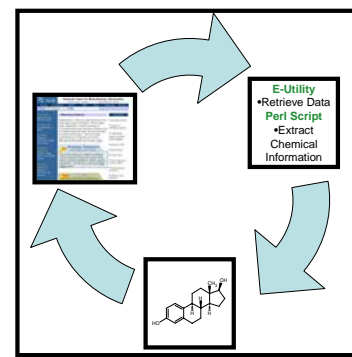
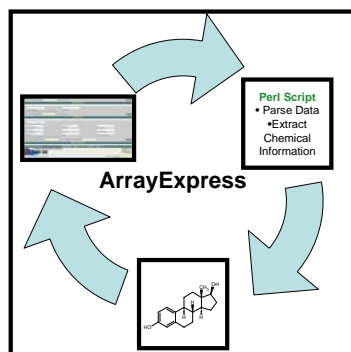
Identification of Genomic Repositories and Databases of Possible Toxicogenomic Interest

Create Chemical Index of Repositories



Legend:

- Indexed Databases
- MGED Databases
- Small Database



Future Analysis

1. Analysis of Chemical Space

2. Coverage of Species

- Human
- Rat
- Mouse
- Zebra Fish

3. Coverage of Platforms

Chemical Replicates???

Bioassay Results ???

Toxicological Results ???

References

- ArrayExpress <http://www.ebi.ac.uk/ArrayExpress>
- ArrayTrack <http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack>
- CEBS <http://cebs.niehs.nih.gov>
- Comparative Toxicogenomics Database (CTD) <http://www.ctd.mdibl.org>, Mattingly et al, *The Comparative Toxicogenomics Database (CTD): A Cross-Species Resource for Building Chemical-Gene Interaction Networks*. ToxSci. 2006, in press
- Gene Expression Omnibus (GEO) <http://www.ncbi.nlm.nih.gov/geo>
- Environment, Drug, and Gene Expression (EDGE) <http://edge.oncology.wisc.edu/>
- Iconix B. Ganter et al, *Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action*, Journal of Biotechnology (2005) 119: 219-44.
- PubChem <http://pubchem.ncbi.nlm.nih.gov>

This work was reviewed by U.S. EPA and approved for publication but does not necessarily reflect official Agency policy.