Defining the Chemical Space of Public Genomic Data

ClarLynda Williams-DeVane[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)

TOXICOGENOMICS

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) ArrayExpress[1], Environment, Drug, Gene Expression (EDGE)[6] and the National Institute of Environmental Health Sciences CEBS[3] genomic data repositories.

- Create a Structure Index of the three data repositories that can integrate these data with other public data sources (e.g., PubChem[8]).

- 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.
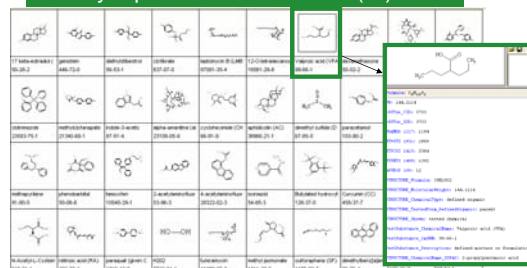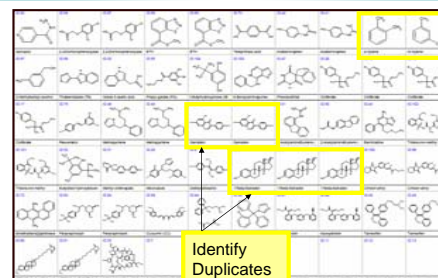
Results

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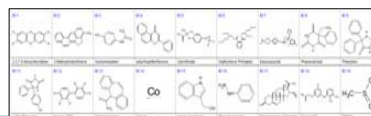
Statistics
- 25 Records
- 18 Chemical Structures
- 8 Chemical Structures overlap GEOMIS (GEO)
- 7 Chemical Structures overlap AREXCH (ArrayExpress)
- 6 Chemical Structures overlap CSTARC (CEBS)
- Chemical Structures overlap 82 Toxicological Records

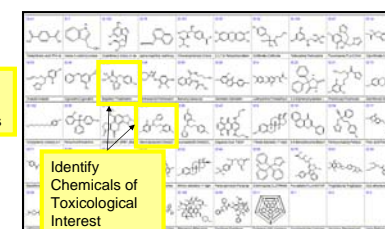### ArrayExpress Structure Index (SI) File

ArrayExpress Database

Identify Similar Chemicals

Identify Duplicates

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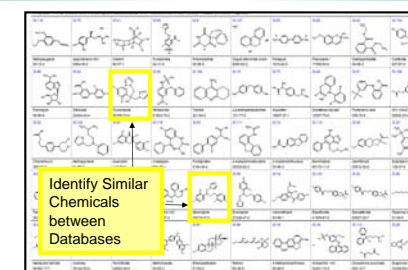
GEO Database

Identify Chemicals of Toxicological Interest

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

CEBS Database

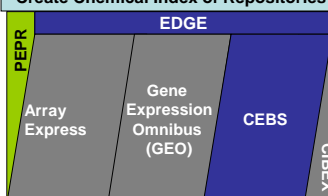Identify Similar Chemicals between Databases

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

Chemical Space

Chemical Overlap Between ArrayExpress and GEO

17beta-Estradiol	50-28-2
Genistein	446-72-0
Dexamethazone	50-02-2
Cycloheximide	66-81-9
Retinoic Acid	302-79-4
Hydrogen peroxide	7722-84-1
Terephthalic acid,TPA	100-21-0
Fenpropimorph	67306-03-0
5-Fluorouracil	51-21-8
Liothyronine	6893-02-3
Mitomycin C	50-07-7
Paclitaxel	33069-62-4

Chemical Overlap between ArrayExpress, GEO, and CEBS

cis-Dichlorodiamine platinum	15663-27-1
Clofibrate	637-07-0
Clotrimazole	23593-75-1
Ethanol	64-17-5
Ethinyl estradiol	57-63-6
Lipopolysaccharide (LPS)	NOCAS
Miconazole	22916-47-8
Valproic acid	99-66-1

Chemical Overlap between ArrayExpress and CEBS

2-Acetylaminofluorene	53-96-3
4-Acetylaminofluorene	28322-02-3
Acetaminophen (4-hydroxyacetanilide)	103-90-2
Bemitrafone	18133-11-3
Butylated hydroxytoluene (BHT)	128-37-0
Doxylamine succinate	562-10-7
Etoposide	33419-42-0
Methapyrilene	91-80-5
Methyl eugenol	1910-42-5
N-Nitrosodimethylamine	62-75-9
phenobarbital	50-06-6
Tamoxifen	10540-29-1

Chemical Overlap between GEO and CEBS

1,3-Bis(chloromethyl)-1-nitrosourea	154-93-8
Adriamycin, hydrochloride	25316-40-9
Allyl alcohol	107-18-6
alpha-Naphthyl isothiocyanate	551-06-4
Atorvastatin	134523-00-5
Bezafibrate	41859-67-0
Carbon tetrachloride	56-23-5
Dimethylformamide	68-12-2
Econazole	27220-47-9
Fenofibrate	49562-28-9
Fluconazole	86386-73-4
Fluvastatin	93957-54-1
Gemfibrozil	25812-30-0
Itraconazole	84625-61-6
Lovastatin	76330-76-5
Methotrexate	59-05-2
N-Nitrosodiethylamine	55-18-5
Rifampicin	13292-46-1
SIMVASTATIN	79902-63-9
Sodium arsenite	7784-46-5
Troglitazone	97322-87-7

Overlap between Genomic Repository Chemical Space

ArrayExpress (106) — 74
GEO (129) — 88
CEBS (135) — 94
12, 12, 8, 21

Overlap between Genomic Chemical Space and DSSTox Chemical Space

DSSTox File Incidence

Valproic Acid

Number of Genomic Records — Number of DSSTox Files

Legend: CEBS, ArrayExpress, GEO

Methods/Approaches
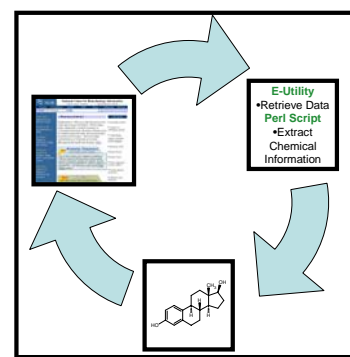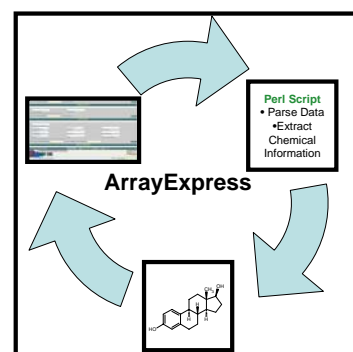
Identification of Genomic Repositories and Databases

Identification of Genomic Repositories and Databases of Possible Toxicogenomic Interest
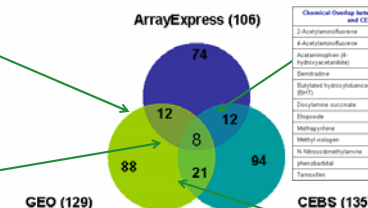
Create Chemical Index of Repositories

EDGE

PEPR | Array Express | Gene Expression Omnibus (GEO) | CEBS | CIBEX

Legend:
- Indexed Databases
- MGED Databases
- Small Database

ArrayExpress

Perl Script
- Parse Data
- Extract Chemical Information

E-Utility
- Retrieve Data
Perl Script
- Extract Chemical Information

Future Analysis

1. Analysis of Chemical Space
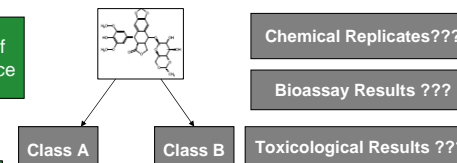
2. Coverage of Species
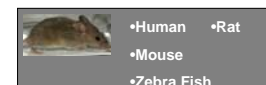
3. Coverage of Platforms

Class A — Class B

Chemical Replicates???

Bioassay Results ???

Toxicological Results ???

- Human
- Rat
- Mouse
- Zebra Fish

References

1. ArrayExpress http://www.ebi.ac.uk/ArrayExpress
2. ArrayTrack http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack
3. CEBS http://cebs.niehs.nih.gov
4. 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
5. Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo
6. Environment, Drug, and Gene Expression (EDGE) http://**edge**.oncology.wisc.edu/
7. 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.
8. PubChem http://pubchem.ncbi.nlm.nih.gov