Towards a Chemogenomic Future

ClarLynda Williams[1], Ann Richard[2], Maritja Wolf[3], Jennifer Fostel[4]

[1]U.S EPA-NC State University Student COOP, Bioinformatics Graduate Program; [2]U.S EPA, ORD, Computational Toxicology Research Program; [3]Lockheed Martin, Contractor to the US EPA; [4]National Center for Toxicogenomics, National Institute of Health Sciences

research&development

IIB-1

Science Question

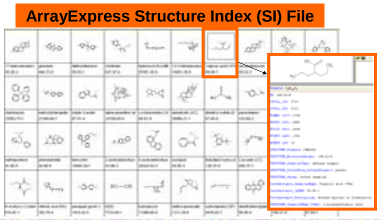
Chemogenomics refers to the integrated use of genomic and chemical information for furthering understanding of biology. Chemogenomics has been applied in pharmacology and toxicology for enhancing target identification/validation, efficacy evaluation, and toxicity assessment. Whether approaching the toxicity prediction problem from a biological or chemical perspective, the extent of available data for training models is severely limiting. Neither approach applied in isolation from the other makes optimal use of potentially relevant data from both biological and chemical domains. To fully realize the potential of chemogenomics, chemical and biological data must be fully integrated into relational data mining applications. Although limited chemogenomics successes have been realized in the commercial pharmaceutical domain, it has yet to be shown if these can be extended to the regulatory domain of environmental chemical hazard assessment. Assessing the feasibility of success will require creation of a mineable chemogenomic database. Major hurdles to be overcome include: lack of publicly available, standardized, reference genomic/bioactivity databases; lack of chemical structure indexing for public genomic and toxicity databases; and scattered, fragmented sources of toxicity and genomic data.

Question: Can limited successes in applying chemogenomics to predictive toxicology in the commercial pharmaceutical domain be extended to the environmental regulatory domain through the development and adoption of tools and methodologies that will further the integration and leveraging of public data?

Research 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], 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.
- Create a mineable Chemogenomic Database
- Incorporate PASS (Prediction of activity spectra for biologically active substances)[7] predictions into the Chemogenomic Database as a surrogate for bioassay profile data.
- 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.

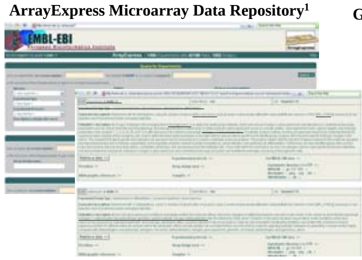
ArrayExpress Structure Index (SI) File



Methods/Approach

Genomic Data Sources

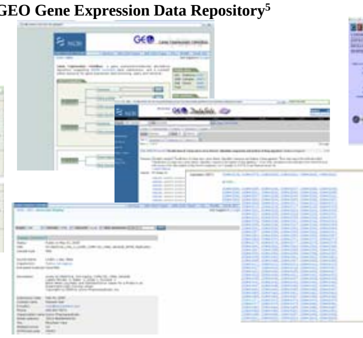
ArrayExpress Microarray Data Repository[1]



1

- **The Chemical space is embedded in the free-text description of the database provided by the depositor.**
- **Required a manual chemical annotation of the entire repository to identify chemical exposure experiments or records, where keywords and descriptors were noted for the development of automated free-text mining update software.**
- **CTD's[4] edited MeSH supplementary chemical vocabulary was proposed as a text mining tool. However, several chemicals were not identified using this method.**
- **Currently Batch Boolean queries are used to identify new chemical exposure records.**

GEO Gene Expression Data Repository[5]



2

- **The chemical space is buried in a hierarchical description of experimental series and arrays.**
- **Required manual extraction and annotation of chemical information, where keywords and descriptors were identified for use in an automated update software utilizing NCBI's Eutils to be developed in the near future.**
- **The MeSH vocabulary, as edited by CTD[4], will also be incorporated into the automation software.**

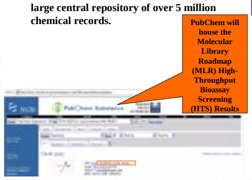
CEBS Toxicogenomic Database[3]

3

The Chemical Index for CEBS was provided.

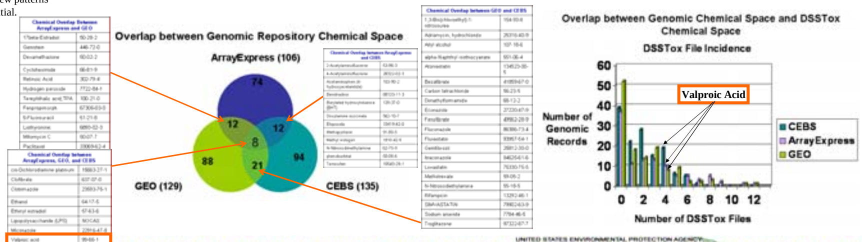
4

- **The resulting chemical inventory listings were converted to DSSTox Structure-Index (SI) files and incorporated into the DSSTox Master SI file (a consolidated central inventory of all DSSTox structure lists).**
- **The Chemogenomics Structure Index files will include links to the source database (e.g., ArrayExpress, GEO), as well as reference indices to the toxicological databases DSSTox.**
- **The resulting toxicological database inventories will be linked to the genomic data from the genomic repositories through the chemical structure indexing.**
- **These Chemogenomics structure-index files with direct links to the chemical experimental records will also be incorporated into the NCBI's PubChem project, allowing structure-searchable access to these on-line genomics data resources from within PubChem's very large central repository of over 5 million chemical records.**



Results

Chemical Space



Results/Conclusions

- The initial chemical annotation of ArrayExpress, GEO and CEBS identified **106**, **129** and **135** Chemical Records, respectively.
- The Chemical Space of ArrayExpress, GEO, and CEBS is approximately **70% Unique** for each database.
- **60-70%** of the Chemical Space of each genomic dataset overlaps the chemical space of DSSTox.
- The chemical annotation of the public genomic repositories proved to be an unexpected challenge. Chemical names, often ambiguous and unreliable in nature, accounted for most of the challenges during the annotation process due to the necessity of free-text mining within the genomic repositories.
- The chemical space covered by the genomic repositories, although limited, proved to be fairly structurally diverse, significantly overlap with existing DSSTox datasets, and be well studied across multiple endpoints.
- By creating a DSSTox Structure Index File for each database, theses public genomic repositories become chemically and structure searchable through PubChem, which is not currently possible.
- Three Chemogenomic Structure Index Files have been developed and will be published in the near future.

Impact and Outcomes

- Prior to the completion of the initial chemical annotation of the public MGED (Microarray Gene Expression Data Society) there was little to no knowledge of the chemical space of public genomic data repositories.
- Although the chemical space covered by these datasets is not large, there is the potential for knowledge discovery because this chemical space has been well studied throughout legacy studies, these chemicals are currently being studied by multiple sources, and this data has not been studied with this combination of data.
- With the success of this project and the aforementioned research goals, public data can be more fully utilized for hypothesis-driven research to support environmental regulatory objectives.
- This project has the potential to enhance the potential of other projects within the EPA and other regulatory agencies by adopting and/or developing methodologies and tools to use both legacy and new public data with modern data mining tools for the chemical, biological, and/or toxicological knowledge discovery.
- This project, lastly, provides the potential for enhancement of other public toxicological and genomic datasets with genomic and toxicological information respectively in addition to bioassay data.

Future Directions

- Through a collaboration with the European Bioinformatics Institute (EBI), for each previously unidentified chemical located in ArrayExpress, a Chemicals of Biological Importance (CheBI) Record will be created. This will be available on EBI's CheBI webpage, http://www.ebi.ac.uk/CheBI. This initiative will link chemical exposure ArrayExpress records to a chemical page.
- The current chemogenomic structure index files will be integrated into a Chemogenomic Database by utilizing the Food and Drug Administration's (FDA's) ArrayTrack tool.
- Methodologies will be developed and/or adopted based on four reference datasets to explore data mining considerations for public chemogenomic data from different laboratories and platforms as well as methods of comparison between different species, doses, chemicals, and endpoints. The four reference datasets will include the Interagency Conzole/Triazole Project results and collaborations with Iconix and GeneLogic pharmacogenomic companies.
- Standardized methodologies will be used in the analysis and interpretation of gene expression data from the genomic repositories for deposit into a minable chemogenomic database.
- PASS bioassay profiles will be computed and used as surrogates for bioassay data to be loaded into the chemogenomic database. Pass predictions will serve as a test case for future bioassay results such as the results of the MLR HTS initiative.
- The chemogenomic database will be mined using predictive data mining technologies with the hope of new knowledge discovery.

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. 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.
7. Prediction of Bioactivity Spectra for Substances **http://www.ibmc.msk.ru/PASS**
8. PubChem **http://pubchem.ncbi.nlm.nih.gov**

Acknowledgements: David Dix (USEPA), Carolyn Mattingly (CTD), Chihae Yang (Leadscope), Philippe Roca-Serra (ArrayExpress), Helen Parkinson (ArrayExpress), Weida Tong (FDA), and Susan Hester (USEPA).