A Collaborative Web-Based Architecture For Sharing ToxCast™ Data

Sean Ekins[1], Ellen L. Berg[2], Keith A. Houck[3], Moses M. Hohman[1] and Barry A. Bunin[1]

[1]Collaborative Drug Discovery, Inc. 1633 Bayshore Highway, Suite 342 Burlingame, CA 94010, USA. [2]BioSeek, Inc 310 Utah Ave, South San Francisco, CA 94080. USA.[3]National Center for Computational Toxicology, Office of Research And Development, United States Environmental Protection Agency, Research Triangle Park, NC. USA

ABSTRACT

Collaborative Drug Discovery (CDD) has created a scalable platform that combines traditional drug discovery informatics with Web2.0 features. Traditional drug discovery capabilities include substructure, similarity searching and export to excel or sdf formats. Web2.0 features include “invitation-only, username/password protected” secure groups, secure intra-group messaging and reply capabilities, dashboards with time/date stamped audit trails per user activity, selective merging of external and internal datasets, the ability to securely share data sets between groups, and selectively edit and mask content. These combined capabilities promote inter-group collaborations. The current project demonstrates selective data sharing among BioSeek and EPA researchers and that CDD technology can handle complex multidimensional toxicology related data. We have archived BioSeek High Content Screening (HCS) data for the ToxCast™ compounds securely into the CDD database to enable sophisticated data mining across this and other datasets. We will describe how searches can be conducted within this data set and across public data in the database. We will highlight how CDD can enable further insights into these compounds, prompting discoveries that would not otherwise occur, facilitating and encouraging participants to collaborate and exchange data. Other ToxCast™ datasets could be incorporated into CDD to enable their integration and analysis as a whole.

INTRODUCTION

ToxCast™ represents a major initiative for prioritizing the toxicity testing of large numbers of chemicals in a short period of time. It uses high throughput screening bioassays to build computational models to forecast the potential human toxicity of chemicals (Dix et al., 2007, Martin et al., 2008). ToxCast™ is profiling over 300 well-characterized chemicals (primarily pesticides) in over 400 HTS endpoints. One of the approaches uses complex cell systems, such as the BioMAP primary human cell-based disease models, to leverage cellular regulatory networks to detect and distinguish chemicals with a broad range of target mechanisms and biological processes (e.g. tissue and inflammatory disease) relevant to human toxicity (Kunkel et al 2004a 2004b and Berg et al 2006).

The ToxCast™ data is creating a unique community of researchers, bringing together those generating the data and the computational analyses, a community that needs a means to share data. The major components of an effective scientific community include: (1) unifying goal, or focus; (2) multiple research areas/expertise; (3) uniform database platform that allows effective data accumulation and management; (4) easy access and sharing of information; (5) potential for unlimited growth.

FIGURE 1. Groups sharing data via CDD.

We have developed a novel web-based database that can be used to securely archive, mine and share chemistry and biology data. The database has been widely adopted by neglected disease researchers e.g. Malaria, Chagas Disease, Tuberculosis, as well as biotechnology and pharmaceutical companies that are moving towards becoming fully integrated pharmaceutical companies (Hohman et al 2009). This database has now been used to securely share and mine HTS data between BioSeek and the EPA. We propose that this use could be expanded to other groups (Figure 1).

THE NEED FOR NEW COMMUNITY-BASED COLLABORATIVE APPROACH

General web collaboration tools have no provision to archive and manage laboratory data and no ability to mine based on chemical structure.

Traditional chemistry and biology data management systems have no collaboration features, do not allow data sharing nor open source public data exchange and have a HIGH COST to maintain and support.

Open, public chemistry and biology data repositories (PubChem, ZINC, eMolecules, ChemSpider, etc.) do not support heterogenous data formats (critical to encourage use) and have HIGH BARRIERS that inhibit routine, comprehensive data archiving. They also do not have the ability to specify private data or limit sharing to specific groups (critical to respect these wishes and IP rather than try to force an all or nothing choice).

FIGURE 2. THE CDD DATABASE

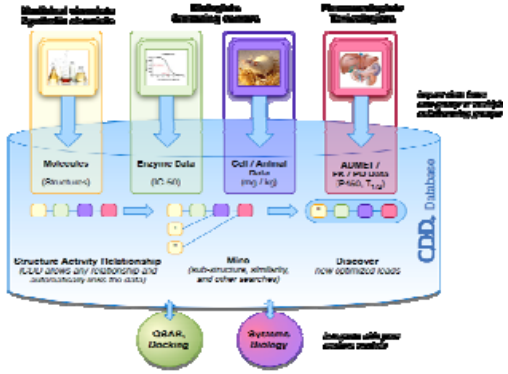
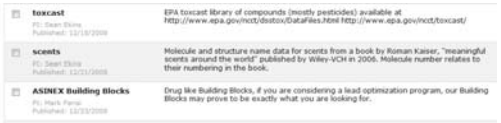


FIGURE 3. COMMUNITY OPEN ACCESS DATA
To date there are 27 datasets that are open access in the CDD database (FDA approved drugs, toxicity data, ADME data, compound libraries etc.). These datasets are structure searchable and can be readily mined.

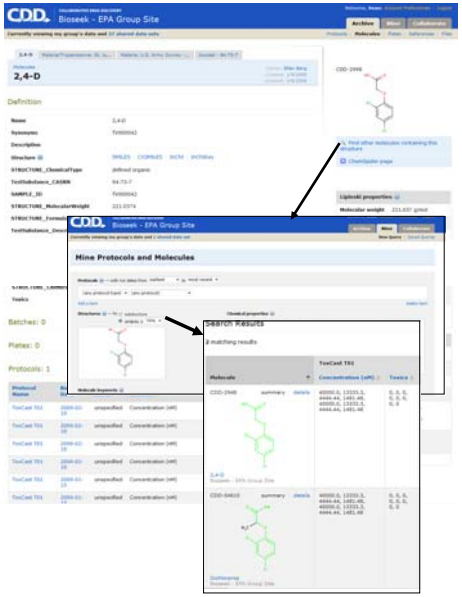


METHODS

The Collaborative Drug Discovery (CDD) database (Figure 2) is built by utilizing community-based web technologies and currently provides a platform that allows scientists to archive, mine, and share both unpublished and published data (Hohman et al 2009). The CDD database:

- Provides an easy to use, searchable format to effectively archive and mine research data.
- Allows the creation of virtual drug discovery networks within scientific community to speed up research and discovery.
- Affords web sharing of unpublished data could save time/resources and prevent duplication of results.
- Data can be selectively shared with other scientists or openly shared within the whole scientific community (Figure 3).

FIGURE 4. MINING CDD.
This illustrates mining the BioSeek data as well as across public databases, illustrating substructure searching as well as links to other external databases such as ChemSpider.



RESULTS AND DISCUSSION

Using the ToxCast compounds one can mine to find similar compounds or other related data on the same structure. For example mining with 2,4-D indicates that it has been published in 2 malaria screening datasets (Figure 4) published to the CDD community (Figure 3) which are available in CDD. Multidimensional structure activity data such as that generated by the BioMAP approach was securely shared (Figure 1) and mined (Figure 4) between BioSeek and the EPA using the CDD database. This data can be readily mined, as well as exported to .xl or .sdf formats for evaluation with other modeling or analysis software. We suggest that this database could be useful to other members of the ToxCast™ research community to securely share their data, facilitate collaborations and gain exposure to other researchers, in the same way that it is being used by those in the neglected disease areas as well as Tuberculosis research laboratories (funded by the Bill and Melinda Gates foundation).

REFERENCES

Berg EL, et. al., Characterization of compound mechanisms and secondary activities by BioMAP analysis. J Pharmacol Toxicol Methods. 2006, 53: 67-74.

Dix, D.J., et al., The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci. 95:5-12, 2007.

Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S and Bunin B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery, Drug Disc Today, 14: 261-270, 2009.

Kunkel, E.J., et al., Rapid Structure-Activity and Selectivity Analysis of Kinase Inhibitors by BioMAP analysis in Complex Human Primary Cell-Based Models. Assay and Drug Development Technologies, 2: 431-41, 2004a.

Kunkel, E.J., et al., An Integrative Biology Approach for Analysis of Drug Action in Models of Human Vascular Inflammation. FASEB Journal, 18: 1279-1281, 2004b.

Martin M., et al., Profiling Chemicals Based on Chronic Toxicity Results from the U.S. EPA ToxRef Database. Environmental Health Perspectives, Environ Health Perspect. 117:392-399, 2009.