Catherine Mendenhall
December 13, 2002
MB&B 452a

The growing field of toxicology and the importance of databases: good databases for good public policies

What is toxicology and why is it important?

The field of toxicology focuses on safeguarding public health by identifying the effects of chemicals and the levels of exposure at which they may become hazardous to humans. Although relatively few chemicals are thought to pose a significant risk to human health, more than 2,000 new chemicals are introduced annually for use in everyday items such as food, personal care products, prescription and nonprescription drugs, household cleaners, and lawn products. More than 80,000 chemicals are already registered for use in commerce in the United States [1]. Consumers are exposed to these chemicals during their manufacture, distribution, use, and disposal, and as pollutants in the air, water, and soil. Without toxicology, the effects of many of these chemicals on human health would go unknown.

What is the problem?

Toxicology is by design an interdisciplinary science, governed by agencies as diverse as the National Institutes of Health's National Institute of Environmental Health Sciences (NIH/NIEHS), the Centers for Disease Control and Prevention's Agency for Toxic Substances and Disease Registry (CDC/ATSDR), and the Food and Drug Administration's National Center for Toxicological Research (FDA/NCTR).

An unfortunate consequence of having such diverse agencies overseeing the field of toxicology is the development of eight separate major databases for toxicological information:

1. **HDSB** is an acronym for the Hazardous Substances Data Bank. This scientifically peer-reviewed database illustrates the broad scope in human and animal toxicity, safety and handling, and environmental fate.
2. The Environmental Protection Agency's (EPA) Integrated Risk Information System **(IRIS)** consists of EPA's data in support of human health risk assessment, focusing on hazard identification and dose-response assessment.
3. **GENE-TOX** is comprised of peer-reviewed mutagenicity test data from the EPA.
4. **CCRIS**, the Chemical Carcinogenesis Research Information System, organizes carcinogenicity, mutagenicity, tumor promotion, and tumor inhibition data provided by the National Cancer Institute
5. **TOXLINE** is the National Library of Medicine collection of online bibliographic information covering the biochemical, pharmacological, physiological, and toxicological effects of drugs and other chemicals.
6. **DART** stands for Developmental and Reproductive Toxicology and Environmental Teratology Information Center, and organizes current and older literature on developmental and reproductive toxicology.
7. The EPA's Toxics Release Inventory **(TRI)** reports numerous chemical synonyms, structures, and regulatory list information.

8. **ChemIDplus** contains over 367,000 chemical records, and is a database of the structure and nomenclature authority files used for the identification of chemical substances cited in National Library of Medicine databases.

An option to search multiple toxicological databases can be found at http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?Multi [2]. Here, one can search any or all of the following databases: HSDB, IRIS, CCRIS, and GENE-TOX. However, there is no integration of the data. The search simply presents a new page, with four links to the chemical listings in the four databases. This is of little use except to those who do not know of the existence of some of the databases listed. There is no option to compare data, or even to visualize the data easily- long lists of toxicological tests done on rats in the early 1960s are mixed in with useful dose-response modeling tools such as NOAEL (No Observed Adverse Effect Level) and LOAEL (Lowest Observed Adverse Effect Level). Improper data evaluation due to the fractionated and hard to read formats could very well result in incorrect conclusions being made (resulting in sub-optimal public policy), or at the very least, result in conclusions being made without knowledge of the scope of available information.

Though the volumes of unorganized data are inconvenient now, the situation is only going to get worse. The field of toxicology is currently undergoing a molecular revolution. The community is recognizing that integrating molecular techniques with traditional toxicity tests will not only tell investigators what happened after a chemical exposure, but how it happened. Advances in genomics and the associated technologies, such as microarrays, will be applied and integrated to create innovative, sensitive and discriminating technologies for environmental toxicity assessments. For example, we are approaching the age of "individual expression". Scientists are currently able to accurately predict "good prognosis" or "bad prognosis" breast cancer tumors via visualizing gene expression patterns in microarrays [3]. The same principle will be applied to toxicology- genetic background is likely to greatly affect the expression of toxicity in individuals exposed to chemicals. "Good prognosis/bad prognosis" studies alone for the myriad environmental chemicals will greatly increase the amount of information to be evaluated when assessing potential risk.

What is the solution?

Bioinformatics tools are absolutely crucial to keep abreast of the growing mass of information produced by diverse government agencies and private companies alike. Bioinformatics will provide the necessary data analysis, management, and organization for effective use of data. As said before, new chemicals are introduced daily. Being able to compare data, such as visualizing similarities between toxicological properties of similar chemicals, will allow for the development of hypotheses based on chemical relationships- the creation of context [4]. This, in turn, will foster the determination of scientifically relevant questions to ask and answer. These answers will form the basis for future public health policies.

The next best thing to a central brain.

The goal is to be able to access any data across the multiple separate toxicological databases quickly and easily. It is impossible for all the information to be kept in a single

"brain" in a single format, so the alternative is to have a federation of resources [4]. Information on a single chemical would be joined by web hyperlinks. Ideally, there would also be the option of a search page, in which commonly used and well defined risk assessment tools (such as NOAELs and LOAELs) could be presented for each database, side by side.

The major problem with genomic information database federations is database interoperability. Many technologies exist for the creation of virtual meta-databases to circumvent this problem. However, this problem could be avoided altogether with the development of a standardized nomenclature [5]. Here, the field of toxicology has a unique opportunity. Since the field is still on the brink of explosion, the development of a more standardized vocabulary is, at least in theory, possible. Investment in the adoption of this vocabulary would result in a simple and effective search protocol for the future. It would also result in greater ease when annotating database information. And, after all, a database is only as good as the quality of its data.

1. National Toxicology Program on World Wide Web. URL: http://ntp-server.niehs.nih.gov/main_pages/about_NTP.html
2. National Library of Medicine Specialized Information Services on World Wide Web. URL: http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?Multi
3. Veer et al. Nature. 2002 Jan 31;**415**:530-535.
4. Gerstein. Nature Structural Biology. 2000 Nov 7;**11**: 960 — 963.
5. Frishman, D., Heumann, K., Lesk, A. & Mewes, H.W. Bioinformatics 1998 **14**:551-61.