Database Development in Toxicogenomics: Issues and Efforts

William B. Mattes,[1] Syril D. Pettit,[2] Susanna-Assunta Sansone,[3] Pierre R. Bushel,[4] and Michael D. Waters[4]

[1]Pfizer Inc, Groton, Connecticut, USA; [2]ILSI Health and Environmental Sciences Institute, Washington, DC, USA; [3]European Molecular Biology Laboratory–European Bioinformatics Institute, Hinxton, United Kingdom; [4]National Center for Toxicogenomics, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina, USA

The marriage of toxicology and genomics has created not only opportunities but also novel informatics challenges. As with the larger field of gene expression analysis, toxicogenomics faces the problems of probe annotation and data comparison across different array platforms. Toxicogenomics studies are generally built on standard toxicology studies generating biological end point data, and as such, one goal of toxicogenomics is to detect relationships between changes in gene expression and in those biological parameters. These challenges are best addressed through data collection into a well-designed toxicogenomics database. A successful publicly accessible toxicogenomics database will serve as a repository for data sharing and as a resource for analysis, data mining, and discussion. It will offer a vehicle for harmonizing nomenclature and analytical approaches and serve as a reference for regulatory organizations to evaluate toxicogenomics data submitted as part of registrations. Such a database would capture the experimental context of *in vivo* studies with great fidelity such that the dynamics of the dose response could be probed statistically with confidence. This review presents the collaborative efforts between the European Molecular Biology Laboratory–European Bioinformatics Institute ArrayExpress, the International Life Sciences Institute Health and Environmental Science Institute, and the National Institute of Environmental Health Sciences National Center for Toxigenomics Chemical Effects in Biological Systems knowledge base. The goal of this collaboration is to establish public infrastructure on an international scale and examine other developments aimed at establishing toxicogenomics databases. In this review we discuss several issues common to such databases: the requirement for identifying minimal descriptors to represent the experiment, the demand for standardizing data storage and exchange formats, the challenge of creating standardized nomenclature and ontologies to describe biological data, the technical problems involved in data upload, the necessity of defining parameters that assess and record data quality, and the development of standardized analytical approaches. *Key words:* **ArrayExpress, bioinformatics, CEBS, database, EBI, HESI, MIAME, NCT, toxicogenomics.** *Environ Health Perspect* **112:495–505 (2004). doi:10.1289/txg.6697 available via *http://dx.doi.org/* [Online 15 January 2004]**

Toxicology, the study of poisons, focuses on substances and treatments that cause adverse effects in living things. A critical part of this study is the characterization of the adverse effects at the level of the organism, the tissue, the cell, and the molecular makeup of the cell. Thus, studies in toxicology measure effects on body weight and food consumption of an organism, on individual organ weights, on microscopic histopathology of tissues, and on cell viability, necrosis, and apoptosis. Recently added to the arsenal of end points that such toxicological studies can use is the measurement of levels of the thousands of proteins and mRNAs present in the cell. The former measurement was made possible with the advent of two-dimensional gel electrophoresis and forms the basis of the field of proteomics. The latter measurement was made possible with the advent of whole genomic sequencing and the subsequent development of microarrays capable of measuring thousands of transcripts at once and is best described as transcript profiling, although it has often been referred to as genomics or transcriptomics. The application of these technologies to toxicology is based on the assumption that the sequelae of events leading to adverse events at the cellular and organism levels will include critical changes in certain mRNAs and proteins. Consequently, these changes may give insight into the molecular mechanisms of toxicity and/or may be diagnostic for a given mode of toxicity. Thus the number of toxicology studies incorporating either proteomics or transcript profiling has been exponentially increasing for several years.

Although both proteomics and transcript profiling measure molecular events at a global and cellular levels, the two are dramatically different in both technology and readout. Proteomics relies on the physical separation of all the proteins of a sample, usually by means of two separate characteristics such as charge and molecular weight, followed by detection of the protein with a dye, and finally, identification by means of mass spectrometry. Transcript profiling with microarrays makes use of hundreds to thousands of defined probes, each of which is intended to detect a single mRNA molecule. The mRNA sample is labeled and hybridized to the microarray such that the signal at a given probe is related to the amount of that particular mRNA in the sample. This readout characteristic makes microarray-based transcript profiling particularly appealing because the identities of the signals are predetermined. In this sense, data generated in transcript profiling experiments are rather straightforward. However, because of the relatively poor annotation of expressed genes and sequence tags, particularly in the dog and rat, the interpretation of transcript profiling experiments is challenging.

The field of toxicogenomics integrates the data-rich science of transcript profiling with traditional toxicological end point evaluation. If successfully implemented, this integration has the potential to serve as a powerful synergistic tool for understanding the relationship between gross toxicology and genome-level effects. From its inception the field of transcript profiling using microarrays has, through the sheer volume

This article is part of the mini-monograph "Application of Genomics to Mechanism-Based Risk Assessment."

Address correspondence to W.B. Mattes, GeneLogic, Inc., 610 Professional Dr., Gaithersburg, MD 20879 USA. Telephone: (240) 364-6238. Fax: (240) 364-6262. E-mail: wmattes@genelogic.com

We thank A. Brazma, Microarray Informatics, (EMBL–EBI);C. Bradfield, McArdle Laboratory for Cancer Research, University of Wisconsin, Madison, WI; W. Tong, National Center for Toxicological Research, Jefferson, AR; and W. Eastin, National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC, for their review of this manuscript prior to submission. We also thank the microarray informatics team at EMBL-EBI, the expression profiler developers, and the ArrayExpress curation and development teams. We especially thank S. Contrino for his contribution to Tox-MIAMExpress. The ArrayExpress project is funded by EMBL, the European Commission [TEMBLOR (The European Molecular Biology Linked Original Resources) grant], the EBI Industry Programme (Biostandards), the CAGE (Compendium of Arabidosis Gene Expression) consortium, and the Health and Environmental Sciences Institute (HESI) Toxicogenomics Database grant.

The authors declare they have no competing financial interests.

Received 25 August 2003; accepted 12 January 2004.

of data involved, required incorporation of resources for bioinformatics, data management, and statistical analysis (Bassett et al. 1999; Eisen et al. 1998; Ermolaeva et al. 1998). The addition of toxicology information to these data poses additional and unique informatics challenges. A typical toxicogenomics study might involve an animal study with three dose groups (one vehicle group, one low-dose group, and one high-dose group), two to three sacrifice times, and four to five animals per group. Even if only one tissue is examined per animal, this represents 36–45 arrays per study, not including replicates. In addition, each animal will have associated data on total body and organ weight measurements, clinical chemistry measurements (often up to 25 parameters), and microscopic histopathology for several tissues. The challenges and opportunities for a rigorous toxicogenomics database are the capture, storage, and integration of a large volume of diverse data. Several commercial ventures, including GeneLogic (Gaithersburg, MD; www.genelogic.com) (Castle et al. 2002), Curagen (New Haven, CT; http://www.curagen.com/) (Rininger et al. 2000), and Iconix Pharmaceuticals (Mountain View, CA; http://www.iconixpharm.com/) have developed proprietary databases of this type. In this article we focus on the development of public toxicogenomics databases and the application of international database standards in that process. The authors acknowledge that the review does not include all public databases of microarray toxicogenomics experiments.

Role of Public Toxicogenomics Databases

Although several reports have described software for managing genomic/transcript profiling data at the local or laboratory level (Bumm et al. 2002; Bushel et al. 2001; Ermolaeva et al. 1998; Liao et al. 2000; Stoeckert et al. 2001), there are compelling reasons for the establishment of public databases that house not only such transcript profiling data but also the associated toxicological end points. First and foremost is that such a public warehouse would provide a means for the scientific community to publish and share the data from such experiments to advance understanding of biological systems. These repositories would also serve as a resource for data mining and discovery of expression patterns common to certain experimental conditions. In addition, a public repository would offer the regulatory community a resource for comparison with toxicogenomics data submitted as part of the compound registration process (Petricoin et al. 2002). Deposition of data into public databases has already been proposed as a requirement for journal publication of standard genomics experiments (Anonymous 2002; Ball et al. 2002), and public databases for microarray data have been established (Anonymous 2002; Brazma et al. 2003; Edgar et al. 2002).

Another important function of some public repositories is the promotion of international standards in data organization and nomenclature (Anonymous 2002; Bassett et al. 1999; Brazma et al. 2001; Stoeckert et al. 2002). Particularly in the case of biological data, the establishment of standard ontologies allows uniform analysis of diverse data (Ashburner et al. 2000). Finally, public toxicogenomics databases would also offer the larger toxicology community common resources for comparing analytical tools and discussing experimental approaches. Thus a database that organizes results from diverse laboratories and platforms would allow the identification of experimental practices that introduce undesirable variability into toxicogenomics data. Although there are challenges for developing public databases that combine genomic and toxicological data, the example of an international infrastructure for nucleotide sequence data such as the GenBank/EMBL/DDBJ (European Molecular Biology Laboratory/DNA Data Bank of Japan) collaboration points to the vast benefit that the larger scientific community would reap from them.

Challenges

Despite the obvious scientific benefits of public toxicogenomics resources, as described below, many technical and logistical issues challenge their implementation. These problems may be broadly classified as *a*) approaches addressing accuracy and specificity, *b*) standardization of data inputs, *c*) methods assuring data quality and comparability, and *d*) development and design of standardization experiments.

Accuracy and Specificity

The use of advanced data referencing and analysis tools is valuable only to the extent that the data employed by these tools have a high degree of internal accuracy. However, the inherent dynamics of hybridization coupled with the incomplete nature of genomic sequence information create the potential for imprecision and/or error in a transcript profiling experiment even before the assay is run.

Hybridization specificity. At the design stage, each element of a microarray, be it a cDNA clone or oligonucleotide sequence, must be selected from the thousands of entries in sequence databases on the basis of several characteristics (Lockhart et al. 1996; Schena et al. 1995). The first of these is specificity: for example, the mRNA for cytochrome P450 (Cyp) 3A4 (GenBank accession no. NM_017460; http://www.ncbi.nih.gov/GenBank/) is 92% identical to the mRNA for Cyp3A7 (GenBank accession no. NM_000765), and thus a microarray element consisting of a cDNA sequence for Cyp3A7 would be expected to detect Cyp3A4 as well. Similarly, a microarray element may lack specificity because it corresponds to a sequence (e.g., a 3′ untranslated region) common to several alternatively spliced transcripts, for example, the UDP-glucuronyltransferase 1A family, where seven transcripts (UGT1A1, UGT1A3, UGT1A4, UGT1A6, UGT1A7, UGT1A8, and UGT1A9) all share the same 3′ sequences (Burchell et al. 1991). Commercial chip manufacturers have gradually recognized some of these problems and have refined their probe sets accordingly.

Accuracy of gene sequences in public databases. Many early sequence entries deposited in public sequence databases were the product of less advanced and less accurate sequencing techniques than are currently available. Thus, when multiple sequence entries for a single gene are available, they should be cross-checked against each other to determine the best consensus sequence to use. The formerly common practice of relating a given clone sequence to only one of several possible GenBank accession numbers must be avoided. Finally, sequences must be examined for hybridization characteristics, as abnormally high or low G + C content may skew the signal for that target relative to other targets.

Accuracy of annotation on a microarray platform. Commercial array manufacturers and custom array designers use GenBank, EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/index.html), or DDBJ (http://www.ddbj.nig.ac.jp/) gene sequence accession numbers of either cDNA expressed sequence tags (ESTs) or mRNAs, IMAGE clone or RefSeq identifiers (IDs), UniGene cluster numbers, or proprietary/internal accessioning to identify gene features. As such, annotation of the elements present on any given microarray can be a potential confounder for microarray use in toxicology. This problem can arise because entries to sequence databases often predate the standardized gene names and descriptions found in curated resources such as LocusLink (http://www.ncbi.nih.gov/LocusLink/; Pruitt and Maglott 2001) or may not be represented in such resources. Such inconsistency in annotation hampers toxicogenomics in two ways. First, it complicates mechanistic interpretation of transcript changes, as nonstandard

annotation of a sequence element (i.e., gene) may limit the effectiveness of a literature search. Second, differences in annotation both within and between different microarray platforms hamper the ability to compare results obtained with their use. One approach would be to adopt an automated client server–based system that incorporates annotation from several sources and allows those sources to contribute equally in real time to annotation content (Dowell et al. 2001). Another approach (described in this volume) cross-references the GenBank accession number for a given array sequence with UniGene (http://www.ncbi.nih.gov/UniGene) and LocusLink (Mattes 2004). This process creates a single identifier number and annotation, making the assumption that LocusLink information (if available) represents the best (i.e., curated) annotation, with UniGene information as an alternative. The single identifier then allows intra- and interplatform comparisons, with the caveat that different probe sequences annotating to the same gene may still give different results based on the hybridization specificity noted above.

Standardization of Data Inputs

Critical to the utility of a database is the breadth, depth, and uniformity of the information it contains. To address the last issue, the same standard nomenclature and numerical units must be used for different data sets so data may be compared across experiments. To address the first two issues, guidelines must be developed detailing what information must be included in a data set. Minimum Information About a Microarray Experiment (MIAME) guidelines (Brazma et al. 2001) allow sufficient and structured information to be recorded to correctly interpret and replicate the experiments or retrieve and analyze the data. Accordingly, guidelines for journal publication of microarray experiments have been proposed (Ball et al. 2002), along with submission of the data to either of the two existing public repositories: ArrayExpress (http://www.ebi.ac.uk/arrayexpress/; Brazma et al. 2003) or Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/) (Edgar et al. 2002). Several journals now require an accession number (indicating that a data set has been submitted successfully to one of these two public repositories) to be supplied at or before acceptance of publication.

Although current MIAME guidelines address the information content for a variety of microarray experiments, a need for comparable guidelines for the toxicology component of some microarray experiments was identified. To address the additional information content in toxicology studies, the National Institute of Environmental Health Sciences National Center for Toxicogenomics (NIEHS NCT; Research Triangle Park, NC) has partnered with EMBL–European Bioinformatics Institute (EBI) (Toxicogenomics at EBI; http://www.ebi.ac.uk/microarray/Projects/ilsi/index.html; Hinxton, U.K.); the International Life Sciences Institute (ILSI) Health and Environmental Sciences Institute (HESI) Technical Committee on the Application of Genomics to Mechanism-Based Risk Assessment; (http://hesi.ilsi.org/ and http://hesi.ilsi.org/index.cfm?pubentityid=120; Washington, DC), and more recently, the National Center for Toxicological Research (NCTR), Center for Toxicoinformatics, U.S. Food and Drug Administration (U.S. FDA) (Jefferson, AR) to initiate the development of guidelines for describing toxicogenomics experiments—MIAME/Tox [see Microarray Gene Expression Data (MGED); http://www.mged.org/]. MIAME/Tox extends MIAME to provide a structured annotation and framework for capturing information associated with the toxicology component of toxicogenomic experiments. MIAME/Tox includes some free-text fields along with controlled vocabularies or external ontologies, specifically regarding species taxonomy, cell types, anatomy terms, histopathology, clinical chemistry, toxicology, and chemical compound nomenclature. An additional objective of MIAME/Tox is to guide the development of toxicogenomics databases and data management software.

One challenge that has arisen as part of the ongoing formulation of the MIAME/Tox structure is the harmonization of diverse ontologies. Pathology observations, both macro- and microscopic, are critical components of toxicological data. To this end pathologists have developed (and are continuing to develop) controlled vocabularies for both human clinical pathology [e.g., systemized nomenclature of medicine (SNOMED); http://www.snomed.org] and veterinary pathology (e.g., Society of Toxicologic Pathology (STP)/ILSI; http://www.toxpath.org/); National Toxicology Program Pathology Code Table (NTP PCT; http://hazel.niehs.nih.gov/user_spt/pct_terms.htm). These efforts obviously do not include gene expression terms or genomic and postgenomic information and may be contrasted with those efforts in the bioinformatics community to develop ontologies (MGED; http://www.mged.org/), Gene Ontology (GO; http://www.geneontology.org), Human Genome Organisation/Proteomics Standards Initiative (HUGO/PSI; http://psidev.sourceforge.net/) where the clinical annotation of the samples (anatomy, pathology, clinical pathology) is pending. It is unlikely that a single terminology will cover all domains, and recent effort has been placed on the semantic mapping and the interoperability among terminologies [e.g., Standards and Ontologies for Functional Genomics (SOFG) mouse anatomy effort; http://www.sofg.org/]. However, semantic mapping/interoperability can only be achieved among true ontologies. Ontology is a key step to integration—a system of coding knowledge. An ontology is a formal and declarative representation that includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or cannot be related to each other. An ontology has a definition for each term it includes and is built in a tool that allows export in a standard and machine-readable format. Developing pathology-controlled vocabularies as an ontology will facilitate data exchange with databases that use a different ontology, subject to a semantic mapping but will require close collaboration between the pathology and bioinformatics community.

Data Quality and Comparability

The old adage "garbage in, garbage out" is constantly reiterated in the world of database development regarding population of data and information. Toxicogenomics databases are not immune to the pitfalls of a poorly guarded data storage system and may contain data of subpar quality and/or insufficient or incorrect biological information to describe or annotate experiments. Data quality metrics for microarray data have been extensively investigated in recent years without any clear consensus in the toxicological community as to which universal standard to adopt (Finkelstein et al. 2002; Gollub et al. 2003; Hessner et al. 2003; Model et al. 2002; Tseng et al. 2001). The difficulty in reaching consensus is due in part to the diversity of existing (and pending) array platforms, data acquisition methods, and normalization procedures. This systematic complexity makes the distillation of consensus data quality standards a significant challenge.

Some of this complexity can be circumvented by storing pixel intensity images (rawest form of the microarray data) from the array scanners in a data repository. Image processing software could be archived and made available on request to permit reanalysis using a common data acquisition method. In addition, several microarray data analysis tools use the mean, median, or other measure of

central tendency of acquired data. As such, the unnormalized or unadjusted data could also be captured in a database. By starting with unadjusted data, the same background subtraction procedure and common normalization or transformation method could be applied to the data to make assessments of gene expression changes across different microarray platforms more comparable.

Although they do not represent consensus standards, the statistical measures and data quality metrics associated with an individual technical platform can be useful tools. This information may be used to compare data quality between two samples run on the same type of technical platform. Alternatively, these measures can be used as qualitative tools for comparison across different technical platforms.

A major impediment to comparing gene expression data and subsequent data quality across platforms is resolving the annotation of the gene features arrayed on the chips. As discussed above (see "Accuracy of Annotation on a Microarray Platform"), when comparing gene expression of features on separate arrays, the problem is ascertaining whether the probes for the genes with equivalent accessioning are actually derived from the same sequence region (start and end base position for oligos or cDNA fragment in the case of ESTs) of the gene. At worst the features with the same gene ID actually may be probing different alternative splice variants of the same gene (Murphy 2002; Wolfinger et al. 2001). Data integrity and usability within toxicogenomic databases can thus be improved by *a*) maintaining the DNA sequence of gene features on the arrays, *b*) regularly updating gene annotation and description by BLAST sequence analysis, and *c*) clustering similar gene sequences to reduce or identify redundancy.

In addition to the challenges associated with microarray data, the challenge in development of a toxicogenomics database is also to effectively capture clinical chemistry parameters and histopathological observations in a manner not only practical for relational or object-oriented database structuring but also intuitive for extracting informative association rules. Customary laboratory quality control measures and routine calibration parameters for clinical chemistry profiles must be stored in a toxicogenomics database to effectively assess the quality of the data when modeling gene expression data in conjunction with clinical pathology evaluations. For instance, checking the linearity of a response variable and collecting a control standard curve measurement are extremely useful in assessing the quality of clinical chemistry data and will prove to be imperative for correlating biochemical changes in biosamples with gene expression–level alterations in cell populations.

Histopathology data result largely from discrete observations drawn from standard nomenclatures or tables of gross and micropathology observations and thus lend themselves well to constructing indicator (class) variables in statistical modeling and data mining algorithms. Pathologists use these descriptors and conventions to describe the essential components of a specific target organ response to a toxicant. Historically, there has not been much agreement with, or standardization of, a common system for pathologists to describe conventional pathological interpretations. Therefore, compatibility among histopathology observations coded by pathologists is a challenge to resolve in a toxicogenomics database and makes the integration of the microarray, clinical chemistry, and histopathology data domains difficult to stage. The use of a sophisticated indexing system to reconcile differing pathological evaluations stored in the toxicogenomics database will theoretically improve the possibility of merging good-quality microarray data with equally precise toxicological information.

Standardization Experiments

In mid-1999 the membership of the HESI formed a project committee to develop a collaborative scientific program to address issues, challenges, and opportunities afforded by the emerging field of toxicogenomics. This committee, comprising corporate members from the pharmaceutical, agrochemical, chemical, and consumer products industries as well as advisors from academia and government, conducted a program in which common pools of RNA were analyzed in more than 30 different laboratories on both similar and different technical platforms. As reported in this volume, the considerable data set generated by the HESI Genomics Committee has been useful in increasing the understanding of sources of biological and technical variability, the alignment of toxicant-induced transcription changes with the accepted mechanism of action of these agents, and the challenges in the consistent analysis and sharing of the voluminous data sets generated by these approaches (Pennie et al. 2004). The experimental programs have shown that patterns of gene expression relating to biological pathways are robust enough to allow insight into mechanisms even across different platforms and analysis sites. Thus, toxicogenomics experiments within the broad fields of hepatotoxicity, nephrotoxicity, and genotoxicity have determined that known mechanisms and pathways of toxicity can be associated with characteristic gene expression profiles. This data set, including both genomic and toxicology data, is currently being deposited in EBI's ArrayExpress (see discussion below).

In a parallel effort the NIEHS Division of Extramural Research and Training (DERT), under the auspices of the NIEHS NCT, initiated the Toxicogenomics Research Consortium (TRC; http://www.niehs.nih.gov/dert/trc/intro.htm) in November 2000 to serve as the extramural research arm of the NIEHS NCT. The TRC consists of several academic institutions, and its primary goal is to perform investigator-initiated molecular toxicology research using current gene expression technologies. In addition to these independent research projects, all centers participate along with the NIEHS Microarray Group in two types of collaborative research projects, standardization experiments and collaborative toxicology, or Science To Achieve Results (STAR), projects. The STAR projects involve investigators from two or more centers conducting collaborative toxicology research using gene expression profiling. Through a series of experiments, researchers are evaluating sources of technical variation in gene expression experiments, with an eye toward establishing standards for evaluating competency and quality of gene expression data across multiple technology platforms and research centers. Findings from the first standardization experiment indicate that individual centers can identify differentially expressed genes in standard RNA samples with moderate to high correlation across a variety of microarray platforms. Conversely, the greatest variation is observed when gene expression experiments are conducted across multiple centers using one or more microarray platforms (Unpublished data). Gene expression data generated by the TRC will support the field of toxicogenomics as a whole as well as assist the NIEHS NCT in developing the Chemical Effects in Biological Systems (CEBS) knowledge base described below.

Path Forward

The many informatics challenges encountered during the course of toxicogenomics projects (as described above) are all surmountable but are made far more tractable if the data required for and generated by the project flow seamlessly in and out of a well-designed database. Thus, array design information such as clone or

sequence ID, if properly stored, can be verified, updated, and linked with current annotation automatically. Similarly, such information is more readily indexed across platforms if stored in a single database. Analysis of data from both a single experiment and across experiments is simplified if the data already have a structure that can be readily accessed with standard query tools and statistical routines. Integration and correlation of microarray data with biological data are made easier and are often possible only if both are housed within the same database. Certainly the utility of such a database for toxicogenomics is greatly enhanced if data from microarray and other relevant global technologies and biological or toxicological phases of the experiment are captured electronically and/or automatically loaded into the database.

Existing and Emerging Databases

As noted above, several public databases for microarray data have been established (Anonymous 2002; Brazma et al. 2003; Edgar et al. 2002; Thomas et al. 2002). The extension of efforts like these to incorporate toxicological and biological end points critically defines and distinguishes toxicogenomic databases, and several key initiatives will be discussed here. It should be noted that creation of an internationally compatible informatics platform for toxicogenomics data will enhance the impact of the individual data sets and provide the scientific community with easy access to integrated data in a structured standard format that will facilitate data comparison and data analysis. Coordination of database structure development and acceptance of common guidelines will result in robust databases valuable to many scientific communities.

One effort to build a public toxicogenomics database focuses on structuring the existing ArrayExpress database to include toxicogenomics data and is under way at the EMBL-EBI, in collaboration with the HESI Technical Committee on the Application of Genomics to Mechanism-Based Risk Assessment. A parallel effort is the CEBS knowledge base (Waters et al. 2003) under development by the NIEHS NCT (http://www.niehs.nih.gov/nct/cebs). Both the EBI ArrayExpress and NIEHS NCT CEBS database models are based on the international standards developed by the MGED Society (http://www.mged.org/), including common minimal descriptors, standard data storage and exchange format, and harmonized nomenclature. These similarities position both CEBS and ArrayExpress as highly collaborative public repositories for scientists internationally. Other efforts include the public Comparative Toxicogenomics Database (CTD; http://www.niehs.nih.gov/oc/factsheets/ctd.htm) at the Mount Desert Island Biological Laboratory, (Mount Desert Island, Salsbury Cove, ME), the dbZach System (http://dbzach.fst.msu.edu/) at the Molecular and Genomic Toxicology Laboratory at Michigan State University (East Lansing, MI) and the Toxicoinformatics Integrated System (TIS; http://www.fda.gov/nctr/science/centers/toxicoinformatics/) at the NCTR.

European Bioinformatics Institute: ArrayExpress and Tox-MIAMExpress

The ArrayExpress infrastructure for microarray-based data (ArrayExpress; http://www.ebi.ac.uk/arrayexpress) (Brazma et al. 2003) has been accepting submissions since February 2002 and sees a rapidly growing volume of data deposited. The goals of the ArrayExpress infrastructure are to *a*) provide the community with easy access to high-quality data in a well-structured standard format; *b*) serve as a repository for gene expression data and any biological metadata correlated with the experiments (e.g., toxicological or pharmacological end points) that support publications; *c*) allow data mining, data comparison, and data analysis across different technology platforms, associating gene expression patterns with the biological metadata; and *d*) facilitate the sharing and reuse of array designs and experimental protocols. The meaningful exchange of information is supported by the use of standard contextual information, MIAME (Brazma et al. 2001) (MIAME1.1;http://www.mged.org/Workgroups/MIAME/miame_1.1.html) and MIAME/Tox (http://www.mged.org/), and a common data exchange format, MAGE Markup Language (MAGE-ML) (Spellman et al. 2002), developed by MGED Society (http://www.mged.org/). MAGE-ML is an extensible markup language (XML)-based data exchange format adopted as a standard by the Object Management Group (OMG; http://www.omg.org/technology/documents/formal/gene_expression.htm). The ability to compare data obtained across different platforms is facilitated by a set of procedures for updating the array annotation and formatting the design into a standard referencing system. A high level of data annotation is ensured by a team of curators assisting the data producers in providing the appropriate information. An MAGE-ML pipeline for direct data submission has been established or is under testing and construction with a number of companies and many academic and governmental laboratories, including Affymetrix, Agilent, Sanger Institute, Stanford University, The Institute for Genomics Research, NIEHS NCT, NCTR, and National Environmental Research Council-United Kingdom. Currently, others (Xybion, Rosetta Biosoftware, Silicon Genetics, National Cancer Institute, Lund University) are adopting MAGE-ML format, including recently developed analysis tools [J-Express (http://www.ii.uib.no/~bjarted/jexpress/), Imagene (http://www.biodiscovery.com/imagene.asp), BioConductor (http://www.bioconductor.org/)].

New collaborations have been established to populate ArrayExpress with high-quality reference data sets, such as human and mouse expression atlases (e.g., in collaboration with Human Genome Mapping Project-Medical Research Council, Cambridge, U.K.), gene expression time courses of basic biological processes in model organisms, and expression profiles of toxic substances (HESI and NIEHS NCT). In the long term, ArrayExpress aims to build a gene expression atlas characterizing gene expression in different tissue and cell types by systematic added value annotation from the team of curators. As of 17 December 2003, the database contains more than 4,000 hybridizations (excluding the approximately 1,000 hybridizations from the HESI genotoxicity, hepatotoxicity, and nephrotoxicity studies in curation phase).

The ArrayExpress infrastructure consists of two data submission routes, a core repository, an online query interface, a query-optimized data warehouse (under development), and an online analysis tool, Expression Profiler (http://ep.ebi.ac.uk/EP/). The first data submission route (via an ftp site) allows batch submission in MAGE-ML format. The creation of MIAME-compliant MAGE-ML files is a demanding but necessary exercise for high-throughput data transfer. For smaller data sets, users can take advantage of a simpler submission process via MIAMExpress. MIAMExpress (http://www.ebi.ac.uk/miamexpress/) is an online annotation and submission tool presented in the form of a MIAME-based questionnaire, where MGED Ontology is used to structure inputs and provide controlled vocabularies for entry.

As part of a collaborative undertaking with the HESI Committee on Genomics (http://hesi.ilsi.org/ and http://hesi.ilsi.org/index.cfm?pubentityid=120), MIAMExpress has undergone further development. These modifications allow the incorporation of standard microarray data in conjunction with conventional toxicological end points (e.g., clinical observations, histopathology

evaluation, and clinical pathology) (Figure 1). The new annotation and submission tool Tox-MIAMExpress (http://www.ebi.ac.uk/tox-miamexpress/) is tailored to accommodate input of the results from the HESI Committee on Genomics toxicogenomics experimental program. Tox-MIAMExpress allows three type of submission: experiment, protocol, and array design. The experiment (a set of related hybridizations) contains the information related to samples, descriptions, treatments, and toxicological assessments data. The integration of toxicology and microarray end points in this database will serve as a prototype for toxicogenomic database design and execution and will allow for a more powerful analysis of the experimental data than is possible in the absence of such a resource.

To ensure harmonization of the toxicological end points, allowing successful data mining, data evaluation, and data comparison, Tox-MIAMExpress is designed according to proposed MIAME/Tox standards. Tox-MIAMExpress uses MGED Ontology concepts that point to established controlled vocabularies for toxicological end points: the International Union of Pure and Applied Chemistry (http://www.iupac.org/dhtml_home.html) for clinical pathology and the NTP PCT (http://hazel.niehs.nih.gov/user_spt/pct_terms.htm) for clinical observations and pathological and histopathological evaluations.

Tox-MIAMExpress is an open-source project, consisting of a perl-CGI interface, MySQL database, and MAGE-ML export component implemented using MAGE Software ToolKit. The system can be also installed locally and used as an electronic notebook for toxicogenomics experiments, potentially allowing one-click submissions to ArrayExpress or to any other toxicogenomics database or tool that accepts MAGE-ML–formatted data. The first beta version of the Tox-MIAMExpress was launched in January 2003, and the public online version has been accepting submissions since September 2003.

The ArrayExpress core repository itself accepts experiments, protocols, and array designs submissions. An accession number is assigned to each completed and curated submission. Upon submission of array designs, a set of procedures is provided to the users to format the array into a standard referencing system. This format will unambiguously locate each element on the array and provide a consistent biological annotation for data mining, data evaluation, and data comparison across different arrays and technology platforms. Another set of tools allows the user to access the latest gene annotation, to reannotate or update their array, by the link provided to another EBI database, EnsMart (http://www.ensembl.org/EnsMart/). Although the array annotation is created on the basis of the sequence information available at the time of its release, drafts of the genomes are continuously updated and subsequent array annotation can be always improved and harmonized. EnsMart is built on the data in EnsEMBL, the genome database at EBI containing consistent species-specific and interspecies annotation (including *Homo sapiens, Mus musculus,* and *Rattus norvegicus*). EnsMart is the recipient of the latest updated drafts of the genomes, with cross-references between identifiers from a wide variety of the public sequence repositories and internal EnsEMBL identifiers (LocusLink, RefSeq, Swiss-Prot, Interpro, GO, VEGA). The output of the query to the EnsMart system can be downloaded in various formats also for direct submission to Tox-MIAMExpress.

The ArrayExpress online query interface allows simple queries and data retrieval. The query parameters can be either general experiment properties (e.g., accession number, author name, condition tested) or sample properties (e.g., species used). The query results are provided as lists of experiments, array designs, and protocols with associated gene expression data and toxicological end points. Users can also receive the data as a MAGE-ML download for easy export into any MAGE-standard supportive tool (MGED; http://www.mged.org/).

The data warehouse, under development, will allow gene- and datacentric queries where queries can be expressed in terms of gene properties (e.g., GO categories, accession numbers) and expression value restrictions. The query results can span multiple experiments, combining MAGE patterns with toxicological end points, and provide cross-array platform analysis facilities. To facilitate gene-based queries, a gene index will be established. This index will link standard gene identifiers, where they exist, to the elements on the array. The gene index will be developed in collaboration with other groups at the EBI and the established model organism

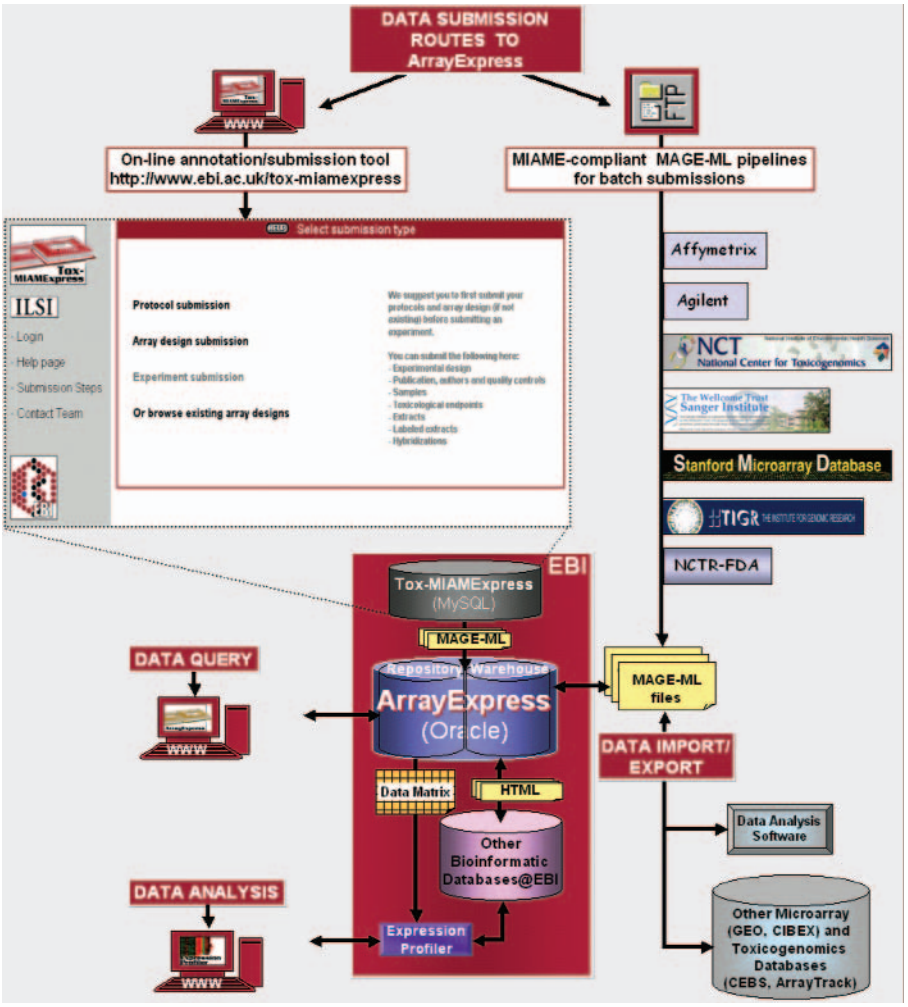

Figure 1. EBI ArrayExpress toxicogenomics infrastructure.

databases, forming the basis of database integration.

The gene expression data from ArrayExpress can also be exported into Expression Profiler (http://ep.ebi.ac.uk/EP/), a set of online tools for the analysis and interpretation of gene expression and other functional genomics data. It incorporates data subselection and transformation components as well as hierarchical and K-means clustering, principal component analysis (PCA), between group analysis (BGA), and several others, together with facilities for the visualization of the data and the analysis results. The data cross-linking module augments the analysis by linking to other tools and databases, for example, metabolic pathway databases. Expression Profiler allows the user to benefit from the latest annotations of the genomes via EnsMart. Third-party components and algorithms, both those installed locally and remote web services, can also be integrated into the workflow mechanism, enabling the platform to expand and develop with the needs of the microarray data analysis community.

The ArrayExpress database will be also fully integrated with other relevant databases at EBI, also as part of Integr8, a new data integration project coordinated by the EBI and funded by the European Union as part of the TEMBLOR (The European Molecular Biology Linked Original Resources) project (http://www.ebi.ac.uk/msd/Temblor/Temblor1.html). It aims to provide a new integrated layer for the exploitation of genomic and proteomic data by drawing on databases maintained at major bioinformatics centers throughout Europe.

The ArrayExpress infrastructure for toxicogenomics provides the community with easy access to highly curated, quality integrated data in a structured standard format, supporting publications, guiding the harmonization process, and facilitating data comparison and analysis.

The Chemical Effects in Biological Systems Knowledge Base

The CEBS knowledge base (http://www.niehs.nih.gov/nct/cebs) is under development (Waters et al. 2003) by the NIEHS NCT as a public toxicogenomics information resource combining data sets from transcriptomics, proteomics, metabonomics, and conventional toxicology with pathway and network information relevant to environmental toxicology and human disease. The overall goal of CEBS is to support hypothesis-driven and discovery research in environmental toxicology and the research needs of risk assessment. Specific objectives are *a*) to compare toxicogenomic effects of chemicals/stressors across species—yielding signatures of altered molecular expression; *b*) to phenotypically anchor these changes with conventional toxicology data—classifying biological effects as well as disease phenotypes; and *c*) to delineate global changes as adaptive, pharmacologic, or toxic outcomes—defining early biomarkers, the sequence of key events, and mechanisms of toxicant action. CEBS is designed to meet the information needs of systems toxicology (Waters et al. 2003), and involves study of chemical or stressor perturbations, monitoring changes in molecular expression, and iteratively integrating biological response data to describe the functioning organism (Ideker et al. 2001). CEBS is a dynamic concept for integrating large volumes of transcriptomic, proteomic, metabonomic, and toxicological knowledge in a framework that serves as a continually changing heuristic engine. The international data-capture guideline, MIAME, and draft MIAME/Tox for toxicogenomics experiments, MAGE-ML data exchange format, is used to assemble and exchange high-quality data sets with the goal of creating a system of predictive toxicology. Toxicogenomics experiments performed using validated NIEHS NCT and NTP methodologies will be captured in their entirety via a unique microarray, proteomics, metabonomics object model that has been extended from MAGE Object Model (MAGE-OM) OMG.

To phase the development of CEBS as well as test the design and implementation of the knowledge base components and system information technology architecture, a prototypic database system has been constructed at the NIEHS to explore the management, integration, mining, and analysis of microarray, histopathology, and clinical chemistry data (Figure 2). Microarray data assessed for distinct gene expression signatures (Bushel et al. 2002) are formatted in an abridged version of MAGE-ML and loaded into a custom-designed Oracle database table with a version of the EBI ArrayExpress database installed. Clinical chemistry and histopathology data on chemically exposed biological samples are obtained from the NTP Clinical Pathology and Toxicology Database Management System (TDMS). Data from several domains can be extracted by means of structured queries and formatted for SAS MicroArray Solution software (SAS Institute Inc., Cary, NC), which can be used to conduct several analyses, including a mixed linear model gene selection method (Wolfinger et al. 2001) to identify genes that are significantly differentially expressed after chemical exposure. Data can also be explored using SAS JMP client software (SAS Institute Inc.) to assess the quality of the MAGE data as well as visualize patterns of gene expression associated with toxicity phenotypes. CEBS itself will provide "scripted analysis tool workflows" whereby a series of statistical approaches will be linked to a detailed description explaining in clear terms how each approach works and in what situation(s) each should be applied. The CEBS online analysis tool suite will enable researchers to save an analysis workflow and the corresponding parameters to CEBS. Colleagues will then be able to apply the same statistical analysis routine by searching for the workflow by name, specifying the data set, and submitting the request. CEBS has leveraged the bioinformatics infrastructure and annotation engine cancer Bioinformatics Infrastructure Objects (caBIO), developed by the National Cancer Institute Center for Bioinformatics (NCICB, Bethdesda, MD), to provide automated full-array annotation (Dowell et al. 2001) for CEBS. A standards-based set of genomic components, caBIO objects, simulate genomic components—genes, chromosomes, sequences, libraries, clones, ontologies, etc. caBIO provides access to a variety of genomic data sources including GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), Unigene (http://www.ncbi.nlm.nih.gov/UniGene/), LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), Homologene (http://www.ncbi.nlm.nih.gov/HomoloGene/), Ensembl (http://www.ensembl.org/), GoldenPath (http://genome.ucsc.edu/cgi-bin/hgGateway?org=human), BioCarta (http://www.biocarta.com/), dbSNP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp), and the NCICB Cancer Genome Anatomy Project data repositories. caBIO is open source and provides access to genomic information using a standardized tool set. CEBS will feature automated pathway projection of expressed genes onto BioCarta (Figure 3) and Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.ad.jp/kegg/) pathways. Pathway visualization is linked to gene annotation, enabling point-and-click annotation of genes on these pathways. It will be possible to navigate the dose–time relationships within a toxicogenomics experiment and to retrieve clinical chemistry profiles and histopathology images at will to phenotypically anchor molecular expression profiles. Links to the supporting literature also will be provided.

The NIEHS NCT has released the CEBS Systems Biology object model (CEBS SysBio-OM) to capture MAGE, proteomics, and metabolomics domain information (http://cebs.niehs.nih.gov/protein/). The model is comprehensive and leverages other open-source efforts, namely, the

MAGE-OM and the PEDRo (Proteomics Experiment Data Repository) object model. The NIEHS NCT contractor, Science Applications International Corporation (SAIC; San Diego, CA), in consultation with the NIEHS NCT and The Scripps Research Institute (La Jolla, CA), has designed the CEBS SysBio-OM by extending MAGE-OM to represent protein expression data elements (including those from PEDRo), protein–protein interaction data, and metabonomic data. CEBS SysBio-OM promotes the standardization of data representation as well as the standardization of the data quality by facilitating the capture of the minimum annotation required for an experiment so that the resulting data can be interpreted accurately. The CEBS SysBio-OM is open source and can be implemented on varied computing platforms.

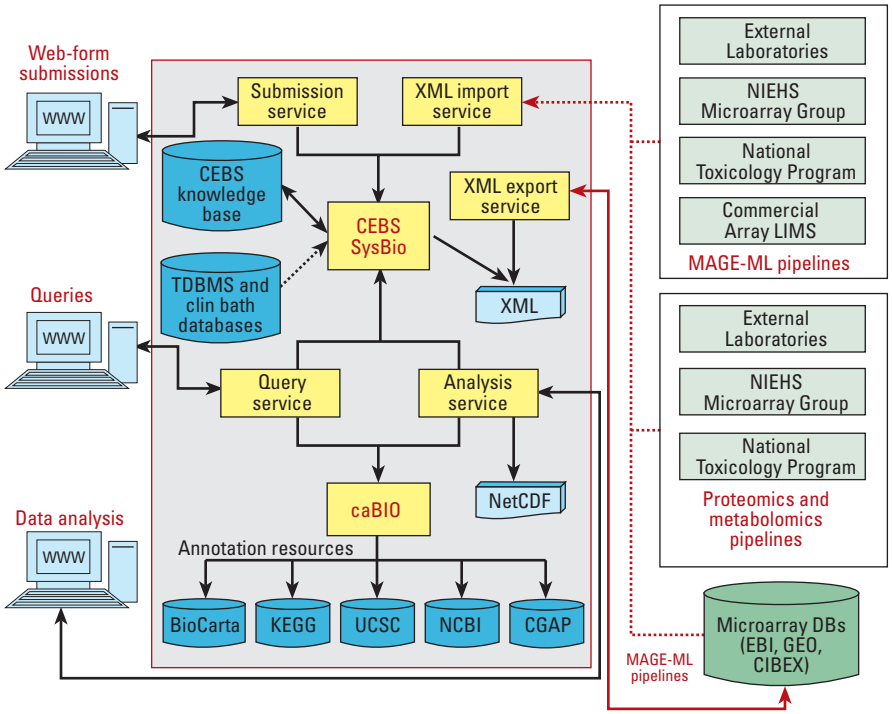

Figure 2. CEBS knowledge base infrastructure as of 18 December 2003. CGAP, Cancer Genome Anatomy Project (http://cgap.nci.nih.gov/); LIMS, Laboratory Information Management System; NCBI, National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/); USCS, USC Genome Bioinformatics (http://genome.ucsc.edu/index.html).

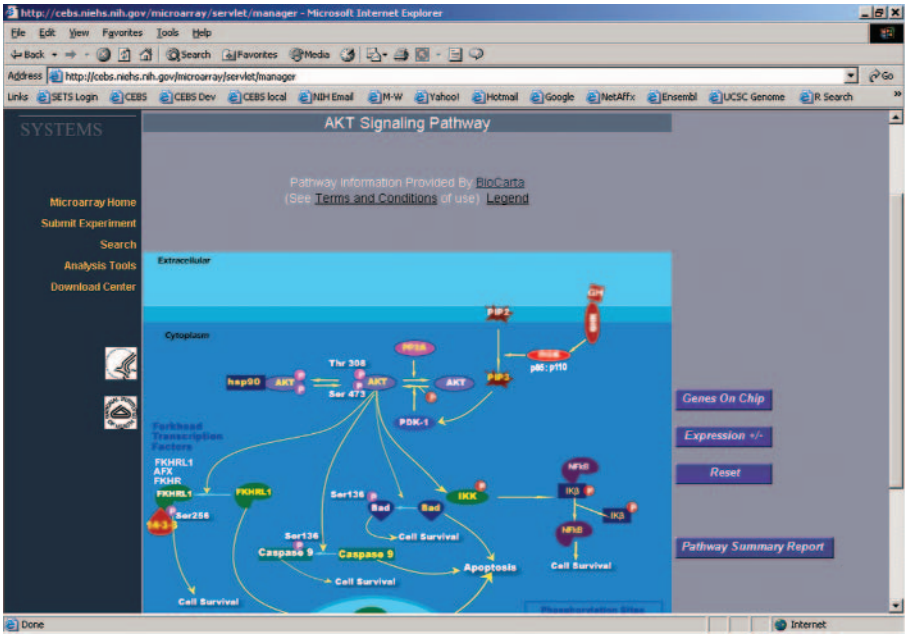

Figure 3. CEBS display of differentially expressed genes on a BioCarta pathway.

As a multigenome knowledge base, CEBS allows characterization of the effects of chemicals or stressors across species as a function of dose, time, and phenotype severity. This permits classifying toxicological effects and disease phenotypes, as well as ultimately delineating biomarkers, sequences of key molecular events responsible for biological response, and mechanisms of action of a chemical or stressor on a biological system. By analogy to GenBank, CEBS will support global sequence-based query using, for example, probe sequence of differentially expressed genes or analytically determined proteins. This will be possible because all probe sets and analytically determined proteins represented in the knowledge base will be sequence-aligned to gene models for all genes known to the knowledge base. As a consequence of this design, reverse query of phenotypic severity attributes (e.g., specific histopathology) can provide entry into molecular expression profiles and associated sequelae. Molecular expression profiles that match a query data set of nucleic acid or amino acid sequence for an experimentally determined gene or protein expression profile can be presented in rank order by quality of match for all significant matches, together with all contextually associated (e.g., dose, time, phenotypic severity) experimental data. In situations involving proprietary chemicals or drugs, a sequence-based (DNA, RNA, or amino acid sequence) query can be performed without divulging the name or chemical structure. CEBS also will support conventional simple and complex query for compound/structure/class, toxic/pathologic effects, gene annotation, gene groups, pathways and phenotypes, etc. Because CEBS will contain data from multiple species of organisms, as the understanding of genetic and biochemical pathways builds toward congruence over time, the sequence-based system facilitates more precise definition of biological pathways and networks as well as genetic variability and susceptibility to, for example, environmental, chemical, or biological insult among species.

CEBS will leverage the NTP Clinical Pathology and TDMS Oracle databases in experimental design and interpretation of phenotypes (Figure 3). In addition, the NIEHS NCT CEBS is taking steps to extend MAGE-ML to load toxicology and pathology data sets compatible with NTP and Xybion toxicology databases (Cedar Knolls, NJ; http://www.xybion.com/). This will facilitate pipelining of toxicogenomics data sets from several

NTP contract laboratories that use NTP databases and from pharmaceutical laboratories that use the Xybion toxicology database. The NIEHS NCT CEBS is currently in the process of mapping the NTP and Xybion toxicology databases to the MIAME/Tox minimal toxicogenomics data guidelines. The availability of NTP toxicogenomics data sets in CEBS will be announced in *Environmental Health Perspectives*.

Because CEBS will contain data on global gene expression, protein expression, metabolite profiles, and associated chemical/stressor-induced effects in multiple species (e.g., from yeast to humans), it will be possible to derive functional pathway and network information based on cross-species homology and pathway conservation. CEBS will ultimately become a knowledge base for both discovery and hypothesis-driven research. CEBS version 1.0 (microarray) was available for internal evaluation in August 2003 and will be available for public use by the end of 2004. Completion of the knowledge base is scheduled for 2012.

Comparative Toxicogenomics Database

The NIEHS DERT supports an international public database devoted primarily to comparative toxicogenomics in aquatic and mammalian species, the CTD (http://www.mdibl.org/). The Mount Desert Island Biological Laboratory is developing CTD as a community-supported genomic resource devoted to genes of human toxicological significance. CTD will be the first publicly available database to *a*) provide annotated associations between genes, references, and toxic agents; *b*) include nucleotide and protein sequences from diverse species with a focus on aquatic and mammalian organisms; *c*) offer a range of analytical tools for customized comparative studies; and *d*) provide information to investigators on available molecular reagents. This combination of features will facilitate cross-species comparisons of toxicologically significant genes and proteins, providing unique insights into the significance of conserved sequences and polymorphisms, the genetic basis of variable sensitivity, molecular evolution, and adaptation. The CTD was developed through a collaboration of five NIEHS-funded marine and freshwater biomedical sciences centers. These centers include Mount Desert Island Biological Laboratory, Oregon State University (Corvallis, OR), University of Wisconsin–Milwaukee (Milwaukee, WI), University of Miami (Miami, FL), and The Jackson Laboratory (Bar Harbor, ME). The goal of the CTD is to develop a comparative database that links sequence information for genes relevant to toxicology to information about gene expression, toxicology, and biological processes. The primary focus of the CTD is on marine and aquatic organisms as model systems for human diseases. The initial focus is also on genes that have been identified through the NIEHS Environmental Genome Project as important for toxicology in model systems. However, the database will eventually merge all gene sequence information generated on all vertebrates and invertebrates, including aquatic organisms, worms, flies, rodents, and people. The CTD provides information about genes and annotation (gene synonyms, sets, and functions) and links between gene sequence and toxicity data published in the scientific literature. These aspects of the database represent an important advancement for comparative toxicogenomics. Such information will include all the synonyms by which a gene is known in different organisms, the toxicant responses identified for specific genes in different species, and a platform that promotes comparisons of gene sequences and toxicant activity among diverse organisms. This data structure will provide comprehensive information about the mechanism of action of toxicants. Understanding these mechanisms will allow more informed assessment of human risk by extrapolating toxicity data from animal models to people and will provide a mechanism by which members of the research community can share their data and promote fruitful avenues for future toxicological research.

dbZach

The Molecular and Genomic Toxicology Laboratory, Michigan State University (East Lansing, MI) has developed the dbZach System (http://dbzach.fst.msu.edu/), a multifaceted toxicogenomics bioinformatics infrastructure. The goal of the dbZach System is to provide *a*) facilities for the modeling of toxicogenomics data; *b*) a centralized source of biological knowledge to facilitate data mining and allow full knowledge-based understanding of the toxicological mechanisms; and *c*) an environment for bioinformatic algorithmics and analysis tools development. dbZach, designed in a modular structure to handle multispecies array-based toxicogenomics information, is the core database implemented in Oracle. This includes several subsystems:

- Clone Subsystem, containing information concerning the cDNA/EST clones
- Microarray Subsystem, containing information concerning custom array designs and the microarray data files
- Gene Function Subsystem, cataloging all the genes represented on the arrays and their annotations
- Sample Annotation Subsystem, collecting MIAME-compliant information about the samples and their treatments
- Toxicology Subsystem, indexing end-point toxicity measures to facilitate correlation with gene expression analyses. The Toxicology Subsystem stores clinical chemistry parameters and manages histopathological data, allowing the pathologist to annotate the sample using the controlled vocabulary from the Pathology Ontology developed by Pathbase database (http://eulep.anat.cam.ac.uk/).

dbZach also stores the actual histopathological images in the database, allowing users to mine large numbers of images, such as a tissue microarray, with ease. By storing the results of several toxicology assessments in the database, dbZach facilitates comparisons of chemical mechanisms of action and supports functional toxicogenomic and chemoinformatic investigations of structure–activity relationships. In contrast to ArrayExpress, CEBS, and CTD, dbZach is not designed to be a public repository for data sets contributed from diverse groups but rather to serve as a standalone database for a laboratory or institution.

ArrayTrack

The NCTR (http://www.fda.gov/nctr/science/centers/toxicoinformatics/tools.htm) is developing TIS to integrate genomics, proteomics, and metabonomics data with conventional *in vivo* and *in vitro* toxicology data (Tong et al. 2003). TIS is designed to meet the challenge of data management, analysis, and interpretation through the integration of toxicogenomics data, gene function, and pathways to enable hypothesis generation. To achieve this, TIS will provide or interface with a large collection of tools for data analysis and knowledge mining.

A first prototype of TIS has been developed. ArrayTrack is a microarray data management and analysis system comprising three integrated components: a database (Microarray DB) storing microarray experiment information; a library (LIB) mirroring critical data in public databases; and a tool module (TOOL) providing analysis capability on experimental and public data for knowledge discovery. ArrayTrack allows users to select an analysis method from the TOOL, apply it to the data stored in Microarray DB, and link the result to gene information stored in the LIB module. The Microarray DB component is designed to be a rich resource for cross-experiment and platform comparison, storing information according to the

MIAME requirements and deriving toxicity-specific signatures from data analysis. The LIB component contains information on gene annotation, protein function, and pathways from a diverse collection of public biological databases (GenBank, Swiss-Prot (http://www.us.expasy.org/sprot/), LocusLink, KEGG, GO), facilitating the annotation and the interpretation of MAGE data. The TOOL component provides a spectrum of algorithmic tools for microarray data visualization, quality control, normalization, pattern discovery, and class prediction. TOOL is also designed to provide interoperability between ArrayTrack system and other analysis software.

TIS will serve as repository for genomics, proteomics, metabonomics, and conventional toxicology data management, supporting data mining and analysis activities. Through cross-linking with a diverse collection of public biological data, TIS will serve as a robust system for exploring toxicological mechanisms.

EDGE

The EDGE database (http://genome.oncology.wisc.edu/edge2/edge.php) was developed at the McArdle Laboratory for Cancer Research, University of Wisconsin (Madison, WI) as a resource for toxicology-related gene expression information (Thomas et al. 2002). It is based on experiments conducted using custom cDNA microarrays that include unique ESTs identified as regulated under conditions of toxicity. To a large extent the experiments were conducted in mice using a variety of agents known to produce toxicity. The database is not designed as a public repository of submitted data; rather, it serves as a query reference and the basis for an algorithm predictive of toxicological potential for a chemical treatment. The ultimate goal of the EDGE is to map transcriptional changes from chemical exposure to predict toxicity and provide valuable insights into the basic molecular changes responsible.

Conclusions

"Ocean in view! O! the joy." So wrote William Clark in November 1805 on sighting the Pacific Ocean after traveling 4,142 miles in his expedition across the western North American continent (Ambrose 1996). Yet it would be several days before the expedition actually reached the ocean, and a year filled with many hardships before it returned home. So it is with toxicogenomics—we can see the goal more clearly than before, but a thorough understanding of this landscape is yet to come. All the sources of experimental and analytical variability are not yet known. The nature of the data demands machine handling and analysis. The path forward must involve a rigorous meshing of biology and toxicology with the computer science of bioinformatics and statistics. To allow the field to develop and to allow scientists to view, assess, and mine each other's data demands the development of public databases and standards for data storage and transmission.

The challenges in the path forward may be seen as falling into three main categories: information quality, information standardization, and analytical conventionalization. Information quality challenges include not only the need for codified, standard metrics for assessing data quality, but also encompass the accuracy, specificity, and annotation of microarray probe elements. Ideally, each element would have associated with it a wealth of information regarding exactly what transcript it detects and how that transcript fits into the overall program of a particular cell type. Furthermore, not only must this information be easily accessible to the casual user (e.g., through a browser interface) but it must also be in a machine-readable format suitable for data mining and statistical analysis. Approaches to either automated or controlled data upload can also be considered issues of information quality. Information standardization, on the other hand, focuses on issues covered in the MIAME and MIAME/Tox efforts. Standardization of pathology terms, which is an ambitious effort given the spectrum of conceivable gross and microscopic observations and the diversity of opinions as to how these observations may best be described, is critical for toxicology. Although a few independent groups have attempted to solve this issue, only true international harmonization of terms and ontologies, resulting in uniform, machine-readable data, will allow the field of toxicogenomics to realize its potential. Information standardization also encompasses the creation of standard data sets that can be used as reference points; the studies of the HESI Committee on the Application of Genomics to Mechanism-Based Risk Assessment can be cited as efforts along these lines. Finally, the challenge of defining best practice analytical approaches to toxicogenomic data may be referred to as analytical conventionalization. Simply put, there are not enough public examples where the same data set was analyzed by multiple approaches and the results compared. Such comparisons would allow the field of toxicogenomics to move forward with confidence in specific approaches. Together, the outstanding challenges in the field of toxicogenomics provide evidence that despite its rapid growth, toxicogenomics remains a nascent field.

Clearly the key to many of these challenges lies in creating well-structured databases to house the results of toxicogenomic experiments. The contents of these databases allow data quality parameters to be explored and analytical approaches compared; in addition, the very effort of creating these databases forces a focus in the community on information standardization. Only in such a database can microarray annotation be kept current and at a high level of sophistication. The collaborative efforts described in this review represent major steps along this path and will represent an additional bonus in information standardization and knowledge sharing. Success in these efforts will be determined by the acceptance of and support for these databases in the toxicogenomics community. Such acceptance will also spur input from the broad scientific community and lead to increasingly diverse and extensive inputs to and use of these public resources.

Toxicogenomics is an information- and informatics-intensive field. To share experimental results successfully between labs and to address larger issues of data handling and analysis, public databases designed to warehouse toxicogenomic data must be developed and populated. The databases mentioned in this review represent the promise and future of toxicogenomics.

References

Ambrose SE 1996. Undaunted Courage: Meriwether Lewis, Thomas Jefferson, and the Opening of the American West. New York:Simon and Schuster, 310.

Anonymous. 2002. Microarray standards at last [Editorial]. Nature 419:323.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, et al. 2002. Standards for microarray data. Science 298:539.

Bassett DE, Jr, Eisen MB, Boguski MS. 1999. Gene expression informatics—it's all in your mine. Nat Genet 21:51–55.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31:68–71.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 29:365–371.

Bumm K, Zheng M, Bailey C, Zhan F, Chiriva-Internati

M, Eddlemon P, et al. 2002. CGO: utilizing and integrating gene expression microarray data in clinical research and data management. Bioinformatics 18:327–328.

Burchell B, Nebert DW, Nelson DR, Bock KW, Iyanagi T, Jansen PL, et al. 1991. The UDP glucuronosyltransferase gene superfamily: suggested nomenclature based on evolutionary divergence. DNA Cell Biol 10:487–494.

Bushel PR, Hamadeh HK, Bennett L, Green J, Ableson A, Misener S, et al. 2002. Computational selection of distinct class- and subclass-specific gene expression signatures. J Biomed Inform 35:160–170.

Bushel PR, Hamadeh H, Bennett L, Sieber S, Martin K, Nuwaysir EF et al. 2001. MAPS: a microarray project system for gene expression experiment information and data validation. Bioinformatics 17:564–565.

Castle AL, Carver MP, Mendrick DL. 2002. Toxicogenomics: a new revolution in drug safety. Drug Discov Today 7:728–736.

Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L. 2001. The distributed annotation system. BMC Bioinformatics 2:7.

Edgar R, Domrachev M, Lash AE. 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210.

Eisen M, Spellman P, Brown P, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868.

Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, Chen Y, et al. 1998. Data management and analysis for gene expression arrays. Nat Genet 20:19–23.

Finkelstein D, Ewing R, Gollub J, Sterky F, Cherry JM, Somerville S. 2002. Microarray data quality analysis: lessons from the AFGC project. Arabidopsis Functional Genomics Consortium. Plant Mol Biol 48:119–131.

Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, et al. 2003. The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res 31:94–96.

Hessner MJ, Wang X, Hulse K, Meyer L, Wu Y, Nye S, et al. 2003. Three color cDNA microarrays: quantitative assessment through the use of fluorescein-labeled probes. Nucleic Acids Res 31:e14.

Ideker T, Galitski T, Hood L. 2001. A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–372.

Liao B, Hale W, Epstein CB, Butow RA, Garner HR. 2000. MAD: a suite of tools for microarray data management and processing. Bioinformatics 16:946–947.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680.

Mattes WB. 2004. Annotation and cross-indexing of array elements on multiple platforms. Environ Health Perspect 112:506–510.

Model F, Konig T, Piepenbrock C, Adorjan P. 2002. Statistical process control for large scale microarray experiments. Bioinformatics 18(suppl 1):S155–S163.

Murphy D. 2002. Gene expression studies using microarrays: principles, problems, and prospects. Adv Physiol Educ 26:256–270.

Pennie WD, Pettit SD, Lord PG. 2004. Toxicogenomics in risk assessment: an overview of an HESI collaborative research program. Environ Health Perspect 112:417–419.

Petricoin EF III, Hackett JL, Lesko LJ, Puri RK, Gutman SI, Chumakov K, et al. 2002. Medical applications of microarray technologies: a regulatory science perspective. Nat Genet 32(suppl):474–479.

Pruitt KD, Maglott DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137–140.

Rininger JA, DiPippo VA, Gould-Rothberg BE. 2000. Differential gene expression technologies for identifying surrogate markers of drug efficacy and toxicity. Drug Discov Today 5:560–568.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470.

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, et al. 2002. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3:RESEARCH0046.

Stoeckert C, Pizarro A, Manduchi E, Gibson M, Brunk B, Crabtree J, et al. 2001. A relational schema for both array-based and sage gene expression experiments. Bioinformatics 17:300–308.

Stoeckert CJ Jr., Causton HC, Ball CA. 2002. Microarray databases: standards and ontologies. Nat Genet 32(suppl):469–473.

Thomas RS, Rank DR, Penn SG, Zastrow GM, Hayes KR, Tianhua H, et al. 2002. Application of genomics to toxicology research. Environ Health Perspect 110:919–923.

Tong W, Cao X, Harris S, Sun H, Fang H et al. 2003. ArrayTrack—supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research. Environ Health Perspect. 111:1819–1826.

Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 29:2549–2557.

Waters MD, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A, et al. 2003. Systems toxicology and the chemical effects in biological systems knowledge base. Environ Health Perspect 111:811–824.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, et al. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8:625–637.