Systems Toxicology and the Chemical Effects in Biological Systems (CEBS) Knowledge Base

Michael Waters, Gary Boorman, Pierre Bushel, Michael Cunningham, Rick Irwin, Alex Merrick, Kenneth Olden, Richard Paules, James Selkirk, Stanley Stasiewicz, Brenda Weis, Ben Van Houten, Nigel Walker, and Raymond Tennant

National Center for Toxicogenomics, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina USA

The National Center for Toxicogenomics is developing the first public toxicogenomics knowledge base that combines molecular expression data sets from transcriptomics, proteomics, metabonomics, and conventional toxicology with metabolic, toxicological pathway, and gene regulatory network information relevant to environmental toxicology and human disease. It is called the Chemical Effects in Biological Systems (CEBS) knowledge base and is designed to meet the information needs of "systems toxicology," involving the study of perturbation by chemicals and stressors, monitoring changes in molecular expression and conventional toxicological parameters, and iteratively integrating biological response data to describe the functioning organism. Based upon functional genomics approaches used successfully in analyzing yeast gene expression data sets, relational and descriptive compendia will be assembled for toxicologically important genes, groups of genes, single nucleotide polymorphisms (SNPs), and mutant and knockout phenotypes. CEBS data sets will be fully documented in the experimental protocol and therefore searchable by compound, structure, toxicity end point, pathology end point, gene, gene group, SNP, pathway, and network as a function of dose, time, and the phenotype of the target tissue. A knowledge base is being developed by assimilating toxicological, biological, and chemical information from multiple public domain databases and by progressively refining that information about gene, protein, and metabolite expression for classes of chemicals and their biological effects in various species. By analogy to the GenBank database for genome sequences, researchers will globally query (or BLAST) CEBS using a transcriptome of a tissue of interest (or a list of outliers) to have the knowledge base return information on genes, groups of genes, metabolic and toxicological pathways, and contextually associated phenotypic information for compounds that display similar response profiles. With high-quality data content, CEBS will ultimately become a resource to support hypothesis-driven and discovery research that contributes effectively to drug safety and the improvement of risk assessments for chemicals in the environment. The CEBS development effort will span a decade or more. *Key words:* bioinformatics, compendia, database, global query, gene expression, heuristic algorithms, knowledge base, linkage-disequilibrium, metabonomics, microarray, molecular expression, ontologies, phenotype, phenotypic anchoring, proteomics, sequence, single nucleotide polymorphisms, systems biology, systems toxicology, toxicogenomics, transcription factors. *Environ Health Perspect* 111:811–824 (2003). doi:10.1289/txg.5971 available via *http://dx.doi.org/* [Online 7 November 2002]

Toxicogenomics is a new scientific field in which researchers study how the genome responds to environmental stressors or toxicants (Aardema and MacGregor 2002; Afshari 2002; Burchiel et al. 2001; Fielden and Zacharewski 2001; Hamadeh et al. 2002a; Nuwaysir et al. 1999; Olden 2002; Tennant 2002; Thomas et al. 2001; Ulrich and Friend 2002). It combines studies of genetics, genomic-scale mRNA expression (transcriptomics), cell and tissuewide protein expression (proteomics), metabolite profiling (metabonomics), and bioinformatics with conventional toxicology in an effort to understand the role of gene–environment interactions in disease. New molecular technologies such as DNA microarray analysis and protein chips can measure the expression of hundreds to thousands of genes and proteins at a time, providing the potential to accelerate discovery of toxicant pathways and specific chemical and drug targets. The power and potential of these new toxicogenomics methods are capable of revolutionizing the field of toxicology. In recognition of this fact, the National Institute of Environmental Health Sciences (NIEHS) has created the National Center for Toxicogenomics (NCT; *http://www.niehs.nih.gov/nct/concept.htm*). The NCT has five major goals:

1) To facilitate the application of gene and protein expression technology
2) To understand the relationship between environmental exposures and human disease susceptibility
3) To identify useful biomarkers of disease and exposure to toxic substances
4) To improve computational methods for understanding the biological consequences of exposure and responses to exposure
5) To create a public database of environmental effects of toxic substances in biological systems

The NCT was formally established in September 2000 and is working to implement a strategy through which these goals can be achieved. This article is an initial response to goal 5. It delineates the conceptual framework and some major design considerations for the proposed Chemical Effects in Biological Systems (CEBS) knowledge base. The concept is open for discussion and debate.

Ideker et al. (2001) have used the phrase "systems biology" to describe the integrated study of biological systems at the molecular level—involving perturbation of systems, monitoring molecular expression, integrating response data, and modeling the system structure and function. Here we similarly use the phrase "systems toxicology" to describe the toxicogenomics evaluation of biological systems, involving perturbation by toxicants and stressors, monitoring molecular expression and conventional toxicological parameters, and iteratively integrating response data. CEBS

This article was previously published in the inaugural Toxicogenomics Section of *EHP*.

Address correspondence to M.D. Waters, NIEHS, PO Box 12233, MD F1-05, 111 Alexander Drive, Research Triangle Park, NC 27709 USA. Telephone: (919) 316-4589. Fax: (919) 541-1460. E-mail: waters2@niehs.nih.gov

We thank the following scientists for valuable discussions on the proposed development of the CEBS knowledge base: C. Afshari on basic design characteristics, R. DeWoskin on PB/PK modeling, H. Hamadeh on user interfaces and gene annotation, P.H.M. Lohman on dose profiling, dynamic linkage, and P450 metabolism, S. Nadadur on comparative genomics, N. Stegman, J. Nehls, and J. Doherty on database design and information technology, and J. Sluka on literature mining and PDQ_MED. We also thank A. Abu-Shakra, N. Cariello, S. Eastin, J. Grovenstein, P. Nettesheim, L. Tomatis, S. Sansone, H. Wan, and L. Wright for their many helpful ideas and comments on the manuscript. We are indebted to S. Wilson and L. Birnbaumer for their continuing support of the NCT program.

Received 30 August 2002; accepted 25 October 2002.

will incorporate high-quality data sets from each of the new toxicogenomics technologies as well as from contemporary molecular and cellular toxicology.

The goals of CEBS are *a*) to create a reference toxicogenomic information system of studies on environmental chemicals/stressors and their effects; *b*) to develop relational and descriptive compendia on toxicologically important genes, groups of genes, single nucleotide polymorphisms (SNPs), and mutant and knockout phenotypes in animal models relevant to human health and environmental disease; and *c*) to support hypothesis-driven research and discovery research in environmental toxicology. We must approach these goals in an incremental fashion, recognizing that in the face of rapid technological change it is impossible to anticipate all opportunities and problems that can develop.

The conceptual design framework for CEBS is based upon functional genomics approaches that have been used successfully in analyzing yeast gene expression data sets (Hughes et al. 2000). The proposed framework is illustrated in Figure 1.

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. CEBS data sets will be fully documented in experimental protocols and therefore searchable by compound, structure, toxicity end point, pathology end point, gene, gene group, etc., as a function of dose, time, and the condition of the target tissue. Controlled vocabularies, dictionaries, and descriptive explanatory text or metadata (that can be processed by a computer) will guide researchers in understanding toxicogenomics data sets. A knowledge base will be developed by carefully assimilating toxicological, biological, and chemical information from multiple public domain databases and by progressively refining that information about classes of chemicals and their biological effects in various species (Tennant 2002; Zweiger 1999). By analogy to the GenBank database for genome sequences, ultimately it will be possible to query the CEBS globally using a transcriptome of a tissue of interest (or a list of outliers from a gene expression analysis) to BLAST (Altschul et al. 1990) the knowledge base and have it return information on genes, groups of genes, metabolic and toxicological pathways, and associated phenotypic information observed in data sets for hits (i.e., compounds that display similar effects in multiple tissues and species, and the dose, time, and phenotypic severity with which these effects are observed). With the expected high-quality data content, CEBS will rapidly become an important scientific resource that provides users with the suite of tools needed to interpret

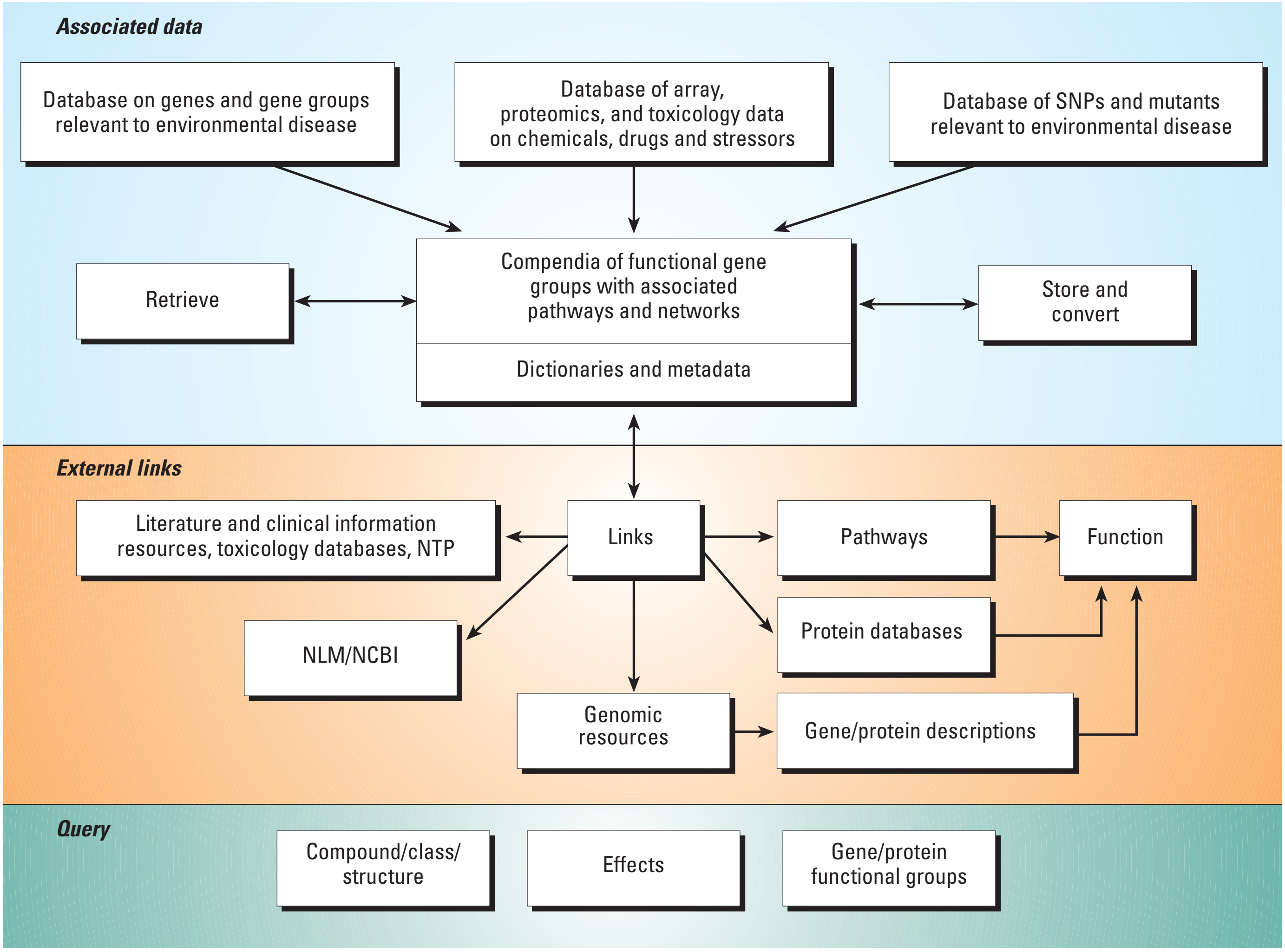

Associated data
Database on genes and gene groups relevant to environmental disease
Database of array, proteomics, and toxicology data on chemicals, drugs and stressors
Database of SNPs and mutants relevant to environmental disease
Retrieve
Compendia of functional gene groups with associated pathways and networks
Dictionaries and metadata
Store and convert
External links
Literature and clinical information resources, toxicology databases, NTP
Links
Pathways
Function
NLM/NCBI
Protein databases
Genomic resources
Gene/protein descriptions
Query
Compound/class/structure
Effects
Gene/protein functional groups

Figure 1. Conceptual framework of the CEBS knowledge base—a cross-species reference toxicogenomic information system on chemicals/stressors and their effects. NLM, National Library of Medicine.

toxicogenomics data and a toxicological reference information system with which to model biological responses across species.

As compendia of expression profiles are indexed and compared to discern diagnostic signatures, it will become increasingly possible to characterize an unknown physical or chemical exposure by comparing its gene or protein expression profile to profiles in the database. Joint research by scientists at the NIEHS Microarray Center (NMC) and Boehringer-Ingelheim Pharmaceuticals has shown that global gene expression profiles for chemicals from different mode-of-action classes can provide gene expression "signatures" of chemical exposures in male rats (Hamadeh et al. 2002b, 2002c). These studies were performed on acutely exposed animals, and the expression patterns appear to be representative of the adaptive or pharmacological activity of the chemicals. Using a small training set, Hamadeh et al. (2002d) were able to correctly ascertain chemical class signatures based on pattern recognition of genes induced acutely. This study, in essence, validated the toxicogenomics hypothesis that knowledge can be gained regarding the nature of blinded samples using an initial training set of chemicals.

NCT Intramural Research

Current NCT research aims to formally discriminate between "chemical signatures" reflecting early adaptive or pharmacological responses with no ensuing pathology and "effects signatures" that entail altered tissue steady state, toxicity, histopathology, or disease (Bartosiewicz et al. 2001). We are therefore developing learning sets of genomic profiling data for various classes of agents, with doses ranging from those that are pharmacologic to those that are toxic. We will also perform comparative studies that address cross-species differences in toxicological responses as well as susceptibility differences in human subgroups.

The combined and integrated data on gene/protein/metabolite changes collected in the context of dose, time, target tissue, and phenotypic severity across species are providing the interpretive information needed to define the molecular basis for chemical toxicity and to model the resulting toxicological and pathological outcomes (Boorman et al. 2002). It will then be feasible to search for evidence of exposure or injury prior to any clinical or pathological manifestation, facilitating identification of early biomarkers of exposure, toxic injury, or susceptibility. We anticipate that toxicogenomics research will lead to the identification, measurement, and evaluation of biomarkers that are more accurate, quantitative, and specific. These biomarkers will be recognized as important factors in a sequence of key events that will help to define the way in which specific chemicals or environmental exposures cause disease. In other words, toxicogenomics will help to delineate the mode of action of various classes of agents and the unique attributes of certain species and population subgroups that render them susceptible to toxicants as an important step in comparatively assessing potential human health risk (Farland 1992).

NCT intramural scientists are now performing additional proof-of-principle experiments designed to establish how effects signatures can be defined and to link the patterns of altered gene expression to specific parameters of well-defined, conventional indices of toxicity. For example, experiments are being designed to correlate gene expression patterns with liver pathologies such as hepatomegaly, hepatocellular necrosis, or inflammation. It is also possible to look for correlative patterns, for example, in enzyme levels, in liver, and in other tissues or cells such as blood. Changes in serum enzymes provide diagnostic markers of organ function that are commonly used in medicine and in toxicology. This "phenotypic anchoring" of gene expression data to conventional indices removes some of the subjectivity of conventional molecular expression analyses and helps to distinguish the toxicological signal from other gene expression changes that may be unrelated to toxicity, such as the varied pharmacological or therapeutic effects of a compound (Tennant 2002).

Future NCT studies will define molecular perturbations caused by environmental chemicals in terms of phenotypic severity, dose, and time (Hamadeh 2002b). We will explore quantitative or absolute gene expression profiling (Dudley et al. 2002) and consider combining such an approach with physiologically based pharmacokinetic (PB/PK) and pharmacodynamic modeling. PB/PK modeling can be used to derive a quantitative estimate of target tissue dose at any time after treatment, thus creating the possibility of anchoring molecular expression profiles in internal dose as well as in time and phenotypic severity. Relationships among gene, protein, and metabolite expression may then be described as a function of the applied dose of an agent and the ensuing kinetic and dynamic dose–response behavior in various tissue compartments. In addition, the species under study and the interspecies interindividual differences must be considered. With the aid of the knowledge systematically generated and assembled (Zweiger 1999) through literature mining, comparative analysis, and iterative biological modeling of molecular expression data sets over time, the adaptive responses of biological systems will be differentiated from those changes associated with or precedent to clinical or visible adverse effects. We anticipate that our understanding of mechanisms of toxicity and disease will improve as these new methods are used more extensively and toxicogenomics databases are developed more fully. The expected result will be the emergence of toxicology as an information science that will enable thorough analysis, iterative modeling, and discovery across biological species and chemical classes. CEBS is being designed to meet the information and modeling requirements of an integrated systems toxicology illustrated conceptually in Figure 2.

A key priority for NCT intramural toxicogenomics studies is the profiling of specific compounds and disease processes that lead to target organ toxicities (e.g., hepatotoxicity and nephrotoxicity). These studies will entertain the following considerations, and emphasis will be on the early steps in the disease processes. Multiple compounds that elicit a particular hepatotoxicity or nephrotoxicity will be studied at multiple sampling times after exposure. Subtoxic as well as toxic doses will be used, and nontoxic isomers and related compounds will be included to assess the specificity of effects observed. Drugs and chemicals will be selected for study on the basis of criteria such as human exposure and recent toxicology studies demonstrating consistent cross-species effects. Ideally, a drug will show a therapeutic effect, and chemicals will display mechanism(s) of toxicity that are prototypical for other agents, including those in our proof-of-principle studies. For example, acetaminophen, or paracetamol, is the first agent to be studied comprehensively by the NCT. Selection was based on an extensive literature (Bessems and Vermeulen 2001) showing that liver toxicity from this agent is a common response in rodents and in humans; its metabolism is similar in rodents and in humans; it displays both therapeutic and toxic effects; and there are opportunities for clinical investigation. Furthermore, acetaminophen has been studied by several laboratories using toxicogenomic methods (Cunningham et al. 2000; Reilly et al. 2001a, 2001b; Ruepp et al. 2002; Yamazaki et al. 2002), which offers the possibility of comparative assessment of observed molecular expression, toxicology, and pathology.

Toxicogenomics Research Partnerships

The magnitude and complexity of the science underlying the broad goals of the NCT is such that no one organization has

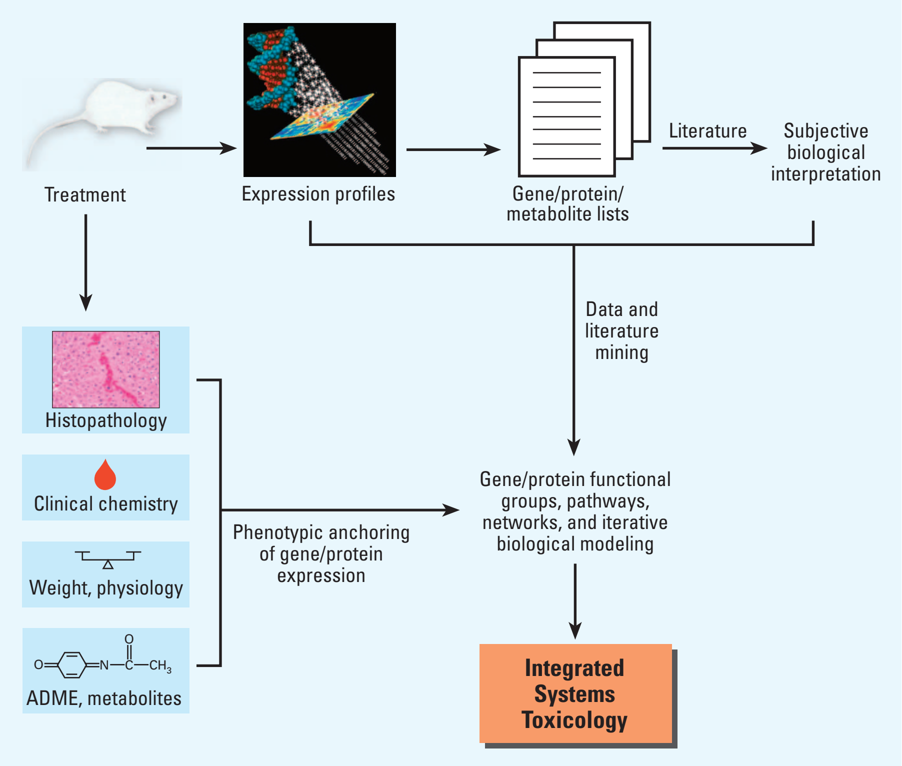

Treatment
Expression profiles
Gene/protein/ metabolite lists
Literature
Subjective biological interpretation
Data and literature mining
Histopathology
Clinical chemistry
Weight, physiology
ADME, metabolites
Phenotypic anchoring of gene/protein expression
Gene/protein functional groups, pathways, networks, and iterative biological modeling
Integrated Systems Toxicology

Figure 2. Interpretation of molecular expression profiles with literature mining, phenotypic anchoring, and iterative biological modeling for systems toxicology. ADME refers to absorption, distribution, metabolism, and excretion.

the technical, fiscal, or intellectual resources with which to solely accomplish them. A central strategy of the NCT, therefore, is the development of partnerships with universities, other federal research and regulatory agencies, and the private sector through the formation of consortia that will address critical scientific challenges in toxicogenomics. The NCT is, in fact, a synergistic collaboration between intramural and extramural scientists based on research partnerships.

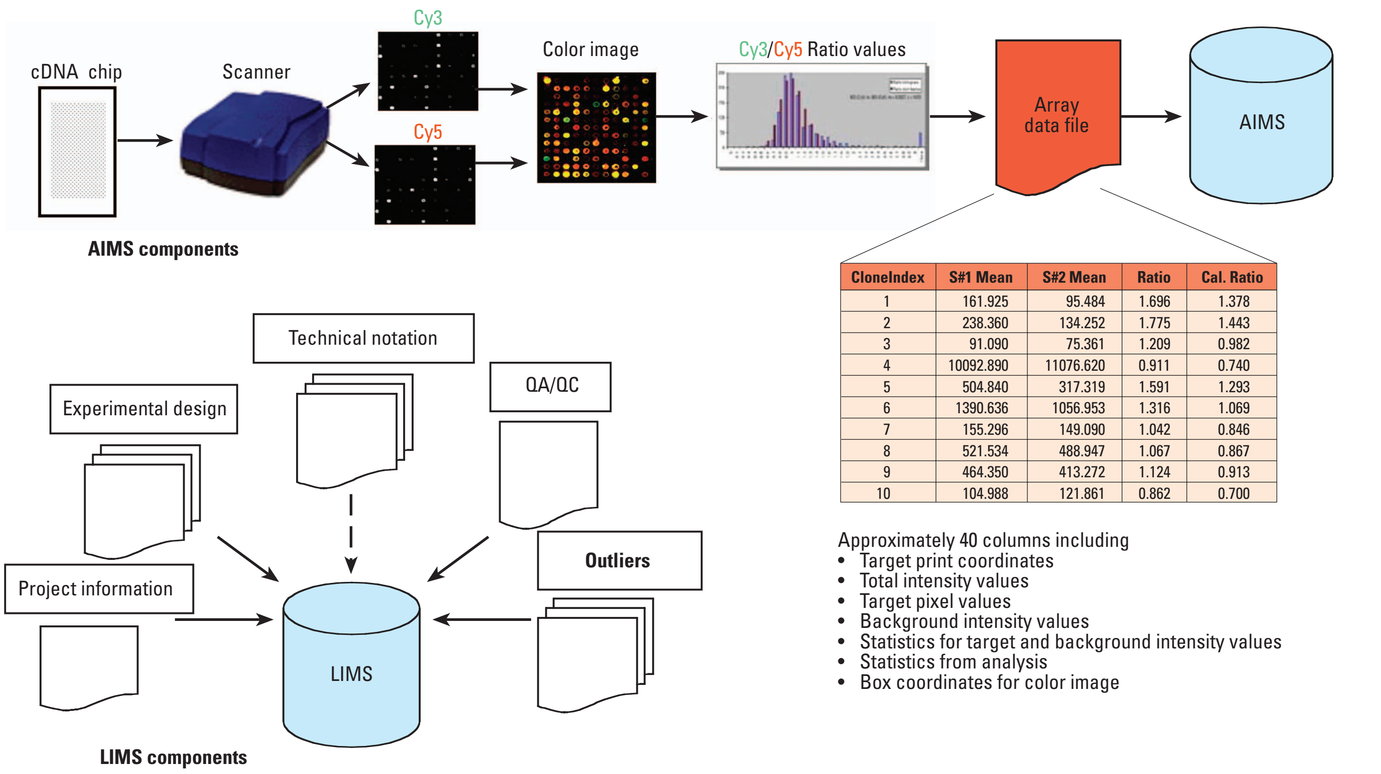
Operating under a National Institutes of Health cooperative agreement mechanism, the Toxicogenomics Research Consortium (TRC) is a key model for achieving the strategic objectives of the NCT. The TRC consists of five academic centers in addition to the NMC: University of North Carolina at Chapel Hill; Fred Hutchinson Cancer Research Center, Seattle, Washington; Oregon Health and Science University, Portland, Oregon; Duke University, Durham, North Carolina; and Massachusetts Institute of Technology, Cambridge, Massachusetts. The consortium members provide specialized expertise in gene expression profiling and bioinformatics; they will perform both independent and cooperative research on various aspects of toxicogenomics. In the current state of gene expression technology, various methodologies for arraying genes and assessing mRNA expression, as well as multiple bioinformatics tools, are being applied in the analysis and management of such data. Therefore, an initial goal of the TRC is to perform a series of "standardization" experiments for gene expression in order to address sources of variation, develop standard practices, and establish data quality criteria and bioinformatics standards. Initial proof-of-principle experiments are being performed to assess the ability of the consortium members to perform standardized toxicogenomics experiments and to exchange and interpret data across multiple microarray platforms. Data generated from such experiments will be incorporated into the CEBS knowledge base and ultimately will be used to design further hypothesis-driven research. The TRC will build on these standardization experiments in performing additional collaborative studies to investigate molecular responses to various environmental stressors. These efforts of the TRC will make a unique contribution to the field of toxicogenomics and to the quality of the CEBS knowledge base.

The NCT participates in a second consortium that addresses many of the same platform and bioinformatics issues as the TRC: the Health and Environmental Sciences Institute of the International Life Sciences Institute (ILSI/HESI) (*http://hesi.ilsi.org/activities/index.cfm?pubentityid=8*). ILSI/HESI is coordinating the efforts of approximately 30 pharmaceutical companies in a worldwide effort to harmonize cross-platform gene expression data and analysis methods. The ILSI Genomics Project is focusing on three categories of toxicants: *in vivo* hepatotoxins, *in vivo* nephrotoxins, and *in vitro* genotoxins. The NCT is involved in the former two categories of study in which animals were dosed, tissues were taken for histopathology and RNA extraction, and RNA samples were then distributed to participating laboratories for microarray analysis using methods chosen by the respective participating laboratories. This type of collaboration will minimize problems associated with RNA extraction and quality control issues and provide a basis for direct comparisons among various microarray platforms.

CEBS will house data from NIEHS intramural and extramural research programs and will accept high-quality data sets from other federal, academic, and industrial partners. For example, through the courtesy of Abbott Laboratories, Rosetta Inpharmatics, Rosetta Biosoftware, and Merck Pharmaceuticals, a set of data from hepatotoxicity experiments on more than 60 chemicals and drugs (52 hepatotoxins) is being made available to the NCT (Waring et al. 2003). By agreement with these private sector partners, this learning data set will be made publicly available via CEBS to the research community.

Microarray Analysis

Microarray data resulting from intramural NCT toxicogenomics experiments are currently captured in the NIEHS MicroArray Project System (MAPS). MAPS is a laboratory management information system developed at NIEHS (Bushel et al. 2001) in which approximately 40 data fields are defined to *a*) manage microarray project information; *b*) detail experimental design; *c*) track clones, sample preparation, labeling and hybridization; and *d*) survey the quality control and assurance of processed microarray chips. The NMC currently produces Yeast Chip v. 1 (6.2 K clones) and four mammalian chips: the Human ToxChip v. 3. (2.2 K clones), the Rat ToxChip v. 2. (6.8 K clones), and human and mouse oligonucleotide discovery chips (17.0 K and 16.0 K oligonucleotides, respectively). Gene accession numbers for each gene or expressed sequence tag (EST) on each chip are automatically updated biweekly from *http://www.ncbi.nlm.nih.gov/UniGene/* to reflect the current National Center for Biotechnology Information (NCBI)



CloneIndex	S#1 Mean	S#2 Mean	Ratio	Cal. Ratio
1	161.925	95.484	1.696	1.378
2	238.360	134.252	1.775	1.443
3	91.090	75.361	1.209	0.982
4	10092.890	11076.620	0.911	0.740
5	504.840	317.319	1.591	1.293
6	1390.636	1056.953	1.316	1.069
7	155.296	149.090	1.042	0.846
8	521.534	488.947	1.067	0.867
9	464.350	413.272	1.124	0.913
10	104.988	121.861	0.862	0.700

Approximately 40 columns including
- Target print coordinates
- Total intensity values
- Target pixel values
- Background intensity values
- Statistics for target and background intensity values
- Statistics from analysis
- Box coordinates for color image

Figure 3. Components of the microarray image and data analysis process. Abbreviations: AIMS, Analysis Information Management System; cal., calibrated; LIMS, Laboratory Information Management System; QA/QC, quality assurance/quality control.

UniGene build. Hundreds of thousands of novel EST sequences have been included in NCBI's UniGene. NMC cDNA chips include a substantial proportion of ESTs, thus offering the potential to discover novel genes involved in important biological or toxicological outcomes and disease processes. To provide some perspective on the information management requirements of gene expression analysis, we have illustrated in Figure 3 the image and data analysis processes for microarray experiments.

Implementation of a CEBS Prototype

With the assistance of the NIEHS Computer Technology Branch, the NCT is currently implementing a prototype version of the CEBS database through the application and integration of software developed for the NMC and the National Toxicology Program (NTP). Toxicology and pathology data from intramural NCT toxicogenomics experiments are currently being captured in the NTP's Toxicology Database Management System (TDMS) in an Oracle database and are being integrated with microarray gene expression data (Figure 4).

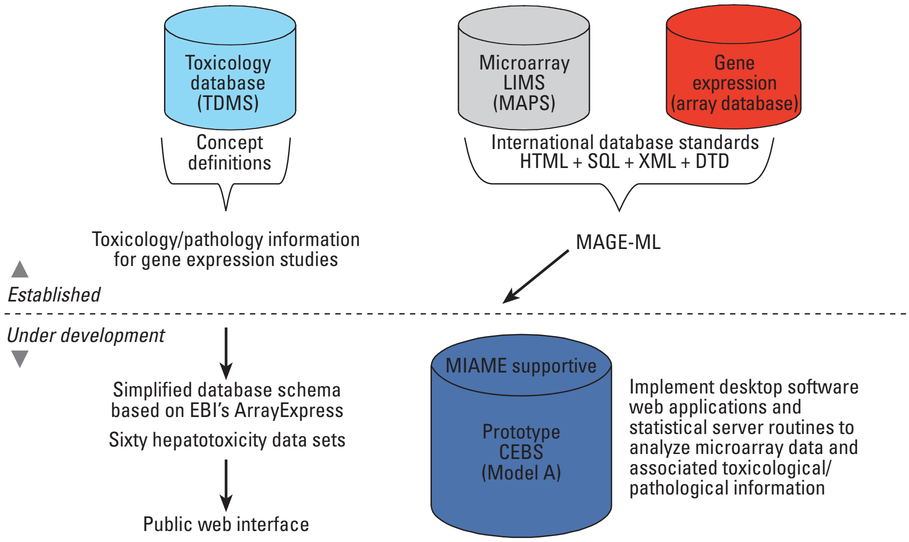

Toxicology database (TDMS)
Microarray LIMS (MAPS)
Gene expression (array database)
Concept definitions
International database standards HTML + SQL + XML + DTD
Toxicology/pathology information for gene expression studies
MAGE-ML
Established
Under development
Simplified database schema based on EBI's ArrayExpress
Sixty hepatotoxicity data sets
Public web interface
MIAME supportive
Prototype CEBS (Model A)
Implement desktop software web applications and statistical server routines to analyze microarray data and associated toxicological/ pathological information

Figure 4. Prototype CEBS (Model A). Abbreviations: DTD, document-type definition; EBI, European Bioinformatics Institute; HTML, hypertext markup language; MAGE-ML, microarray gene expression-markup language; MAPS, MicroArray Project System; MIAME, minimum information about a microarray experiment; SQL, structured query language; TDMS, Toxicology Database Management System; XML, extensible markup language.

Prototype CEBS (Model A) will be a temporary workbench for concept definition and systems integration in the development of CEBS. Nevertheless, this model will provide early public web access to NCT data sets and will implement software applications and statistical server routines required to analyze microarray data and associated toxicological information. It will provide MIAME (minimal information about a microarray experiment) (Brazma et al. 2001), supporting the MIAME standard of the Microarray Gene Expression Database (MGED) Society (*http://www.mged.org/*). The underlying motivation for MIAME is to enable the establishment of public repositories for microarray data and to serve as a basis for designing a microarray data exchange format or markup language (microarray gene expression markup language, MAGE-ML).

Many additional database standards are under review for use in the development of CEBS, but perhaps the most important ones are those under the purview of MGED. MGED has expert working groups on *a*) experimental description and data representation standards; *b*) microarray data extensible markup language (XML) exchange format (CEBS will use XML for data exchange); *c*) ontology (Karp 2000) for sample description (CEBS will follow the gene ontologies of the Gene Ontology (GO) Consortium at *http://www.geneontology.org/* for biological process, molecular function, and cellular component); *d*) normalization, quality markup control, and cross-platform comparison; and *e*) future user group queries, query language, data mining, all of which will provide important input for the development of CEBS. In addition, MGED has developed the microarray gene expression-object model (MAGE-OM), which will be used to develop proteomics and toxicology object models for CEBS. The NIEHS Scientific Computing Laboratory is currently pursuing requirements definition and object modeling for proteomics and toxicology to facilitate a seamless future integration of gene, protein, and toxicology/pathology databases. It should be noted also that the TRC is fully operational and is presently receiving database and bioinformatics support through Srinivasa Nagalla at the Oregon Health and Science University (OHSU) *http://medir.ohsu.edu/~geneview/*. OHSU has implemented an Oracle version of GeneX developed at the National Center for Genome Resources. The resource contractors who will support the TRC will come on line early in 2003. They will begin to receive samples from the TRC and to provide data sets to the CEBS prototype in 2004. With the simultaneous development of the NCT proteomics resource contract, the metabonomics research effort, and further expansion of NCT programs, data, and information resources, the CEBS prototype will begin to evolve into the CEBS knowledge base.

Systems Toxicology—Bioinformatics and Interpretive Challenges

To develop a toxicogenomics knowledge base that will support the requirements of systems toxicology, we must address bioinformatics and interpretive challenges at multiple levels of biological organization and phenotypic severity. Figure 5 illustrates some of these challenges as molecular expression analysis is used to monitor the sequential adaptive, pharmacological, toxicological, and pathological events observable in biological systems after exposure to a chemical.

The lower levels of complexity (genes, gene groups, functional pathways) reflect our current levels of understanding and our ability to describe and package that knowledge using what might be termed "linear bioinformatics." In fact, risk assessors seek to define a sequence of key events and common (linear) modes of action for environmental chemicals and drugs (Farland 1992, 1996; Larsen et al. 2000). The networks and systems level of biological organization reflects global bioinformatics challenges, wherein the cell expresses global change constantly in response to environmental stimuli. This is a systems biology reality that can only be addressed using fully context-documented toxicogenomics data sets properly assembled with appropriate statistical and mathematical modeling to develop an integrated systems toxicology. However, a substantial amount of data entry, data processing, and knowledge building must be performed before such advanced bioinformatics approaches can be applied. It should be recognized that the development of a knowledge base to accurately reflect global molecular expression and to aid systems biological interpretation is a complex issue dealt with only superficially in the present discussion. Keeping these challenges and concepts in mind, we now present some conceptual arguments regarding the phased development of the CEBS knowledge base—a process that undoubtedly will require a decade or more to complete. Progress in the development of CEBS can be monitored at *http://www.niehs.nih.gov/nct/*.

Phased Development of the CEBS Knowledge Base

The CEBS knowledge base will be developed in four substantially overlapping phases: Phase I involves the gathering of microarray gene expression, toxicology and pathology data, and development of gene and protein annotation and bioinformatics tools. Phase II incorporates corresponding proteomics data sets with similar annotation and bioinformatics tools and develops a temporary proteomics database. Phase III integrates gene, protein, and (ideally) metabolite databases and links them with numerous internet resources for metabolic and functional pathway discovery. Phase IV adds two additional databases, one on gene and protein groups and one on SNPs to what has been described above. The three databases then are integrated with a series of bioinformatics tools (data and literature mining) and computational algorithms designed to generate new knowledge.

CEBS Phase I: Microarray Gene Expression Data, Toxicology/Pathology Data, and Associated Analysis Tools

CEBS Phase I will be a public toxicogenomics database containing data sets from the TRC, the intramural NCT research program, and from industrial and

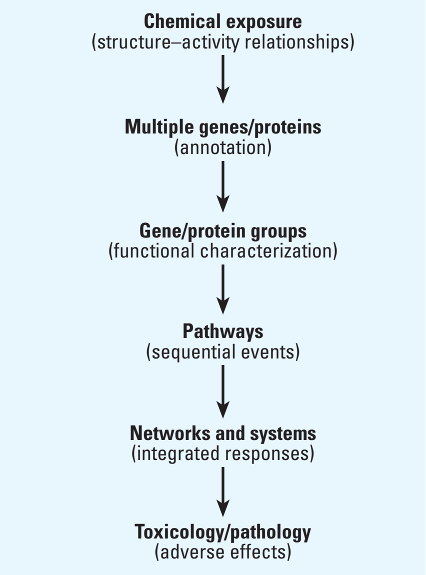
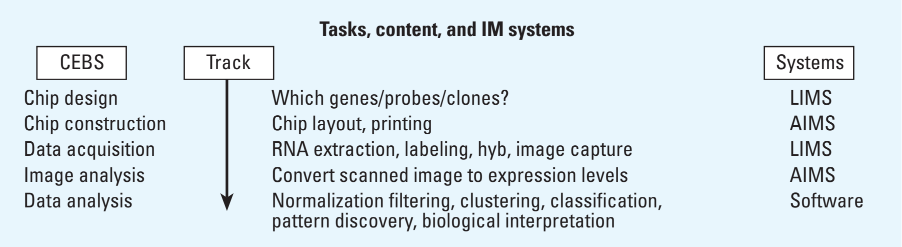


Figure 5. Interpretive bioinformatics challenges at levels of increasing biological complexity in a paradigm leading from chemical exposure to adverse outcomes.



Tasks, content, and IM systems

CEBS	Track		Systems
Chip design		Which genes/probes/clones?	LIMS
Chip construction		Chip layout, printing	AIMS
Data acquisition		RNA extraction, labeling, hyb, image capture	LIMS
Image analysis		Convert scanned image to expression levels	AIMS
Data analysis		Normalization filtering, clustering, classification, pattern discovery, biological interpretation	Software

Figure 6. Microarray experimental components—information management (IM) systems for data acquisition and biological interpretation.

governmental partners. It will comprise mainly microarray and toxicology data and information. To assist the TRC in populating CEBS with microarray data, the NCT awarded a resource contract to provide access to high-throughput microarray gene expression analysis. As illustrated in Figure 6, CEBS Phase I will track all microarray technical and experimental components relating to chip design, construction, data acquisition, image analysis, and data analysis. It will also track clone set gene sequences, descriptors and other genomic annotations, and associated toxicological/pathological end points, and will provide basic bioinformatics tools for data analysis and biological interpretation.

CEBS Phase I will be protocol driven. All data sets within CEBS will be linked by reference to an experimental protocol number and metadata that will specify standard operating procedures, observations, and measurements to be recorded. CEBS Phase I will include complete sample annotation (e.g., sample name, organism, biosource provider, sample source, developmental stage, age and units, time points, organ/tissue, growth conditions, medium, culture temperature, genetic variation, individual name or ID, disease state, additional clinical information and units, target cell type, cell line, treatment application, treatment type, separation technique, sample extraction method, amplification method, label, etc. All the data types (numbers, graphs, observations, images, etc.) will be related by the experimental protocol. The data to be stored and their location will be similarly identified in the process of defining the experimental protocol, as will reports to be generated and analyses to be performed. The purpose of this high degree of context documentation is to facilitate extensive query and biological interpretation. Domain-specific metadata will introduce experimental data sets in each analytical domain: transcriptomics, toxicology, pathology, etc. CEBS Phase I will incorporate raw microarray image files as well as fully processed outlier gene lists together with appropriate visualization tools. Results will be displayed or juxtaposed in various "views," or graphic user interfaces, that will provide insights, facilitate further analysis, and suggest new hypotheses to test.

CEBS also will access biological, chemical, and toxicological resources in public domain databases, as well as pathway information such as that available in the Kyoto Encyclopedia of Genes and Genomes (KEGG) at *http://www.genome.ad.jp/kegg/* (Ogata et al. 1999) and What Is There? (WIT) at *http://wit.mcs.anl.gov/WIT2/* (Selkov et al. 1998). Links will be built to other databases such as the European Bioinformatics Institute (EBI) ArrayExpress database (*http://www.ebi.ac.uk/microarray/ArrayExpress/arrayexpress.html*), the National Library of Medicine's Gene Expression Omnibus (GEO) Database at *http://www.ncbi.nlm.nih.gov/geo/* (Edgar et al. 2002), and the NTP's new Oracle toxicology information bank.

To address the first of the bioinformatics and interpretive challenges mentioned above, basic gene annotation in CEBS Phase I will be largely automated; annotation resources will be routinely consulted to provide a complete range of updated gene/protein information. The process of gene annotation is illustrated in Figure 7, and some major biological data and information resources for gene annotation are shown in Table 1. The links for these annotation resources were operational at the time of publication of this article. However, please consult the NCT website for a current list of links (*http://www.niehs.nih.gov/nct/*).

Continuous refinement of gene annotation and sequence definition will improve the interoperability of cross-platform data sets (Zweiger 1999). Steps for keeping sequence data current can be as follows: *a*) sequence all cDNA clones sets and refer to the known sequences of oligonucleotide sets, *b*) reference GenBank accession numbers and UniGene ID numbers for genes, and GenBank accession numbers and dbEST cluster ID numbers for ESTs, *c*) reference TIGR Gene Indices (*http://www.tigr.org/tdb/tgi.shtml*) for EST or oligonucleotide consensus sequence (Quackenbush et al. 2001) and MegaBLAST against Trace Archives for genomes of interest. MegaBLAST against Trace Archives compares nucleotide sequence data against the current raw data underlying first-pass sequence generated by various genome sequencing centers. This is particularly important for the

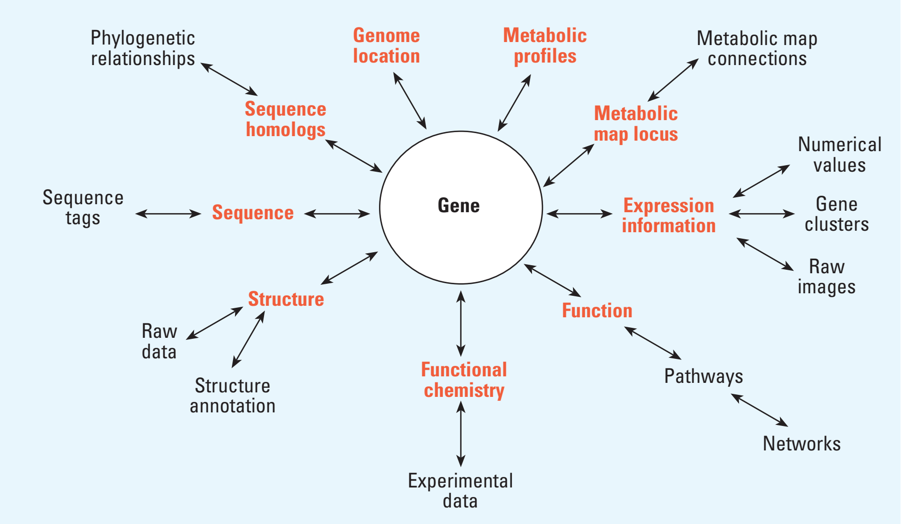


Figure 7. Information (annotation) associated with a single gene. Adapted from Gibas and Jambeck 2001.

Table 1. Some major biological data and information resources for gene annotation.[a]

Subject	Source	Link
Biomedical literature	PubMed	*http://www.ncbi.nlm.nih.gov/entrez/query.fcgil*
Nucleic acid sequence	GenBank	*http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=nucleotide*
(e.g., for the rat)	RGD	*http://rgd.mcw.edu/ ebEST*; *http://rgd.mcw.edu/EBEST/*
Annotation (mouse)	MGI	*http://www.informatics.jax.org/*
Genome sequence	GenBank	*http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=genome*
		http://www.ncbi.nih.gov/PMGifs/Genomes/euk_g.html
	TIGR	*http://www.tigr.org/tdb/*; *http://www.tigr.org/tdb/tgi/*
Protein sequence	GenBank	*http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=protein*
	Swiss-Prot	*http://www.expasy.ch/sprot/*
Protein structure	Protein DB	*http://www.rcsb.org/pdb/*
	PIR	*http://www-nbrf.georgetown.edu*
Protein mass spectra	PROWL	*http://prowl.rockefeller.edu*
Posttranslational mods	RESID	*http://www-nbrf.georgetown.edu/pirwww/search/textresid.html*
Biochemical pathways	KEGG	*http://www.genome.ad.jp/kegg//*
	WIT	*http://wit.mcs.anl.gov/WIT2/*; *http://emp.mcs.anl.gov*
	PathDB	*http://www.ncgr.org/software/pathdb/*

[a] Adapted from Gibas and Jambeck (2001).

rat genome, which is presently very incomplete. This effort to derive new information about incomplete genomes will substantially enhance the discovery value of ESTs on cDNA chips and will facilitate cross-species investigation of gene/protein functional analogies, which we will discuss further.

Functional characterization presents a second bioinformatics and interpretive challenge. Functional characterization can involve the grouping of similar genes and gene products. A number of conventional means can accomplish this, including supervised and unsupervised classification/prediction, artificial intelligence, and various genetic algorithms as well as a number of annotation resources just discussed. We propose to use these methods and resources in concert with query of the scientific literature to develop knowledge of the function of genes and gene products.

Literature queries can facilitate gene annotation as well as biological interpretation of microarray expression results. The challenge is to deal not only with accepted microarray gene annotation names but also with legacy data in the earlier scientific literature, with the ultimate objective of making linkages of gene and protein annotations with literature on the basis of sequence information. MEDLINE, the most widely accessible repository of the biomedical literature, currently contains over 11 million abstracts and is growing rapidly. Unfortunately, it is difficult to use the gene name found in a nucleotide sequence database record (or as presented on a list of outliers) to search the biomedical literature effectively.

The generation of names for genes and gene products based on sequence information is a significant challenge. Ultimately, genes and gene products must be linked by sequence data. Sequence-based synonym naming requires expertise in both data extraction and bioinformatics. Expertise in bioinformatics is required, as much of the searching will need to be done using BLAST (*http://www.ncbi.nlm.nih.gov/BLAST/*; Altschul et al. 1990). Genomic BLAST pages are available for human, mouse, rat, zebrafish, and other eukaryotic and microbial genomes at the NCBI's BLAST website just mentioned.

Nucleotide sequence databases, e.g., GenBank or UniGene, do not contain a "gene product" name field. Instead, the name is imbedded in other information. For example, the GenBank nucleotide definition for "estrogen receptor 1" (the HUGO recognized name for this receptor) is "Homo sapiens estrogen receptor 1 (ESR1), mRNA." Extraction of the appropriate search terms estrogen receptor 1 and ESR1 from the GenBank definition is a trivial task that becomes intractable when a large number of genes or protein products are being searched in the literature, or when the process is being automated, as is being contemplated in the development of the CEBS knowledge base.

To improve the interoperability between microarray gene annotation and the scientific literature, all genes in the clone lists are being provided with vetted name lists. By vetting, we mean that each gene name is searched in MEDLINE, and the way in which MEDLINE parses the name is examined to ensure that it is being searched in the desired manner. For example, searching MEDLINE via Entrez (*http://www.ncbi.nlm.nih.gov/Entrez/*) with the query phrase "estrogen receptor 1" does not return any abstracts. Closer inspection of the search results indicates this is because this phrase does not occur in the MEDLINE phrase index. The vast literature (more than 10,000 abstracts) concerning this receptor is only accessible with the legacy names of "estrogen receptor" and "estrogen receptor alpha."

Once name lists suitable for searching MEDLINE are available, we have two tools to help mine the literature data, OmniViz and PDQ_MED. OmniViz (Battelle Memorial Laboratory, Columbus, Ohio) is a global literature search and visualization software package that can help greatly in obtaining an overview of relevant biomedical publications. The proximity-of-data query software, InPharmix's PDQ_MED (Sluka 2002), can facilitate rapid access to relevant abstracts in MEDLINE for multiple genes (as from a list of outliers).

In CEBS Phase I, a database of gene identifiers, gene sequence, and synonym names suitable for searching the scientific literature will be available; such a database is currently in beta test at NIEHS for human, mouse, rat, and yeast chips printed at the NMC. A web interface to the database will be provided allowing CEBS users to enter a chip name and a list of gene IDs or GenBank accession numbers. The output from the interface will be the list of names suitable for searching in MEDLINE or for use with a literature mining tools such as PDQ_MED or OmniViz. This is an important step toward improving the interoperability between microarray gene annotations and the scientific literature and ultimately toward building knowledge in CEBS.

CEBS Phase II: Protein Expression Database and Metabonomics Data Sets

The proteomics efforts within the NCT consist of an intramural research program, a proteomics resource contract, and extramural and innovative research grant awards in proteomics. The close association of the NCT microarray and proteomics research groups and the NTP provides a unique opportunity for integrating genomics, proteomics, and toxicology data sets. The proteomics group and mass spectrometry group perform hypothesis-driven research on differentially expressed proteins in key tissues and biological fluids of interest to toxicogenomics. A primary platform to separate and identify proteins used by NCT proteomics research groups is two-dimensional (2D) gel protein separation and mass spectrometry (MS), or 2D–MS. Analysis by 2D–MS creates protein maps where proteins for a specific tissue are organized by isoelectric point (pI) and molecular weight (MW). To assist the NCT in populating CEBS with proteomics data, the NCT has awarded a proteomics resource contract that will allow access to high-throughput 2D–MS capabilities on an industrial scale. Critical target tissue and serum from toxicology studies is being analyzed for differential protein expression. As discussed earlier, a primary goal of NCT intramural and contract proteomic studies is biomarker discovery for proteins (including serum/plasma proteins) indicative of chemical exposure and/or to provide mechanistic insight into chemical toxicity. Therefore, concurrent analysis of serum/plasma is being performed in addition to specific target organs for each study.

In addition to 2D-MS proteomics, a new platform called surface enhanced laser desorption ionization (SELDI) is being developed intramurally to screen serum from experimental animals and clinical sources to find new biomarkers (Issaq et al. 2002). Serum proteins are selectively bound to chemically active surfaces on SELDI biochips and are rapidly scanned with high mass accuracy. The normalized serum mass spectra from chemical treatment or disease groups can be compared for differences in specific proteins or in key clusters of protein masses to serve as biomarkers of chemical exposure or disease process. Two other important aspects of NCT proteomics are the extramural proteomics granting activities through the Division of Extramural Research (DERT) and Small Business Innovation Research (SBIR) awards, which will engage promising academic research projects in proteomics and also harness new innovative proteomics technologies for toxicology.

An interim protein expression database (PED) will support the intramural proteomics group and the extramural proteomics resource and resource contract. PED will be developed based on microarray standards and proteomics best practices. PED will develop in parallel with the prototypic version of CEBS, and the analytical integration of transcriptomics and proteomics data will be studied. Many of the standards and practices applied in the interpretation of microarray and gene expression are also applicable to the interpretation of protein expression data sets. Thus, we anticipate that the object models built by MGED in the microarray gene expression database arena also will be applicable to proteomics and metabonomics. As mentioned previously, object modeling for proteomics is currently being pursued. Proteomics objects that may be linked in a linear chain by one-to-many relationships might include the biological sample, raw 2D-stained gel image, enzyme digest, feature number (protein spot), MW, pI, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-MS), m/z ions for protein fingerprint identification, sequence tag from tandem MS analysis, MS search data results, and protein identification search results. The derived objects in the database might include the study parameters, including experimental, biological, and toxicological details; processed gel images; annotated master gel images for each specific tissue or biological fluid; differentially expressed protein list determined from image analysis; feature (protein spot) table of estimated pI; MW; accession numbers; and protein functional groups.

CEBS Phase III: Integrate Microarray Gene Expression and Protein Expression Databases Using a Gene/Protein Group Strategy

The integration of microarray/gene expression and protein expression data is a critical step that will require development of knowledge of gene/protein functional relationships, gene/protein groups, and the development of algorithms that will increase our knowledge of the functions of these groups through actual experimentation. To build knowledge, we are mining the published literature for genes and groups of functionally related genes or protein products relevant to known end points in toxicology, pathology, cell regulatory processes, metabolism, and the like. This literature mining and analysis process is using vetted gene names, and the output will be groups of genes/proteins that represent putative functional groups based on the literature. We will then develop algorithms to test these putative functional gene groups derived from the literature against treatment-related expression profiles and against clustered genes (and co-regulated ESTs) to confirm gene grouping on the basis of phenotype (Figure 8).

This literature-based functional classification of gene groups and their association with known toxicant-responsive pathways will begin to define the relationships between gene and protein expression and our conventional understanding of metabolism, toxicology/pathology, modulation and homeostasis, cell regulation, and cell signaling. It will also offer an opportunity for discovery of yet unidentified genes (ESTs) that are co-regulated with known genes.

To the extent possible, we will confirm gene group membership by sequence analysis and develop statistical procedures and algorithms (Wolfinger et al. 2001) to continually refine our knowledge of gene/protein groups and their relationship to functional pathways. With sequence definition of genes, proteins, and gene/protein group members, it will be possible to begin to BLAST outlier genes and proteins from new experimental data sets against data sets already contained in the CEBS database. This will begin to facilitate and inform the integration of transcriptomics and proteomics data sets across treatment, dose, time, tissue type, and phenotypic severity. We also propose to integrate metabonomics data sets into CEBS Phase III because of the pivotal role that metabolism plays in experimental and clinical toxicology as well as in hazard identification and risk assessment (Bundy et al. 2002; Holmes et al. 2000, 2001; Nicholson et al. 1999, 2002).

CEBS Phase IV: Knowledge Technology

The development of a knowledge base for systems toxicology will require merging several different knowledge-building strategies. In addition to mining the literature for chemical-specific functionally characterized gene/protein groups, testing putative functional gene/protein groups against

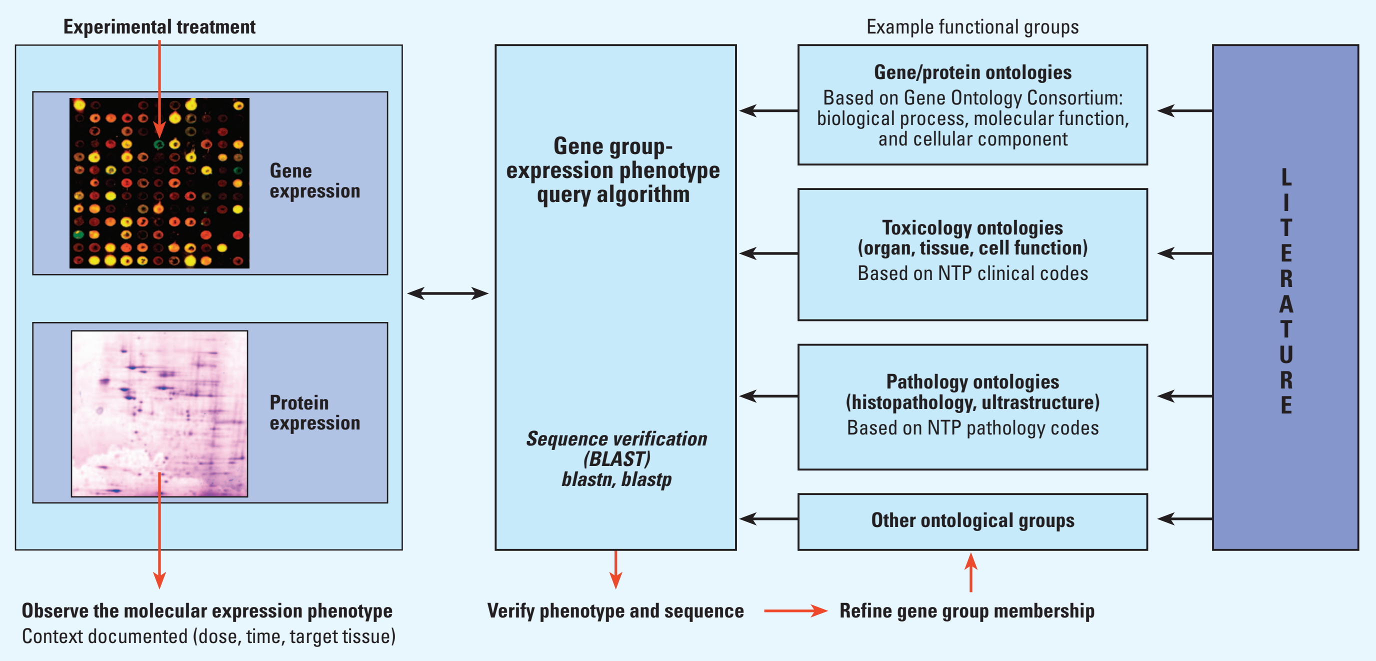

Experimental treatment
Gene expression
Protein expression
Gene group-expression phenotype query algorithm
Sequence verification (BLAST) blastn, blastp
Example functional groups
Gene/protein ontologies
Based on Gene Ontology Consortium: biological process, molecular function, and cellular component
Toxicology ontologies (organ, tissue, cell function)
Based on NTP clinical codes
Pathology ontologies (histopathology, ultrastructure)
Based on NTP pathology codes
Other ontological groups
LITERATURE
Observe the molecular expression phenotype
Context documented (dose, time, target tissue)
Verify phenotype and sequence
Refine gene group membership

Figure 8. Literature-derived putative functional gene groups validated against actual expression profiles of known toxicant-responsive pathways.

treatment-related gene and protein expression profiles, and determining the relationships of these gene/protein groups to functional pathways, we will consult gene ontology from the GO Consortium, *http://www.geneontology.org/*, and attempt to verify the accuracy of the ontologies in terms of biological process, molecular function, and cellular component. This standard gene ontology reflects broad biological goals accomplished by ordered assemblies of molecular functions, tasks performed by individual gene products, and subcellular structures, locations, and macromolecular complexes, respectively.

Standardized gene and protein ontological relationships are significant in that they can help to define functional relationships among genes and groups of genes and proteins. Therefore, we will attempt to confirm the putative functional relationships across multiple molecular expression data sets in the evolving knowledge base. Gene/protein ontology is an important corollary to the gene/protein group strategy and may prove to be an effective approach to the integration of gene and protein expression data sets, especially if it can effectively be converted to a heuristic process. As a further adjunct to the knowledge building, a more complete and heuristic data compendium strategy will be devised based on statistical classification and clustering algorithms (to look for co-regulation) of genes and proteins as a function of dose, time, and target site (Figure 9). Here the experimental protocol defines the doses and the time course as well as the bioassays and biological measurements that will be made. The bioinformatics protocol specifies the various statistical and clustering algorithms that will be applied to look for correlated and co-regulated genes. Ontologies will be used as described above. Note that an ontology lists similar elements, whereas a pathway describes an interaction among diverse elements. Using literature-derived putative gene groups (ideally vetted in appropriate gene ontologies), an iterative and heuristic gene/protein group phenotype analysis is expected to yield validated gene/protein groups that map to known functional pathways and, in terms of toxicology, to define the sequence of key events and common modes of action for environmental chemicals and drugs. Compendia of data will be assembled within each toxicogenomic and toxicological/pathological domain.

Thus, CEBS Phase IV will enable query by compound, structure and class, toxic or pathologic effects, gene annotation, gene/protein groups, and functional (e.g., metabolic and toxicological) pathways that lead to toxicity and disease. To facilitate integration of compound-specific data sets, all genes, proteins, and gene/protein groups will be linked to the gene/protein name and sequence database that was described earlier. This will facilitate query using any of the query categories listed above. Ultimately, one will globally query (or BLAST) the CEBS knowledge base using a transcriptome of a tissue of interest (or a list of outliers from gene expression, or proteins from proteomics analysis) and have the knowledge base return all similar toxicogenomics data and data sets as well as contextually associated phenotypic information for compounds tested in various species and tissues represented in the knowledge base. This will be possible because of the derivation and maintenance of up-to-date sequence information on all genes and proteins represented in the knowledge base. In a sequence-driven knowledge base, a global query can return comparative genomic information (discussed below) based on BLAST cutoff values selected by the user. For example, a BLAST-$\log_{10}$ (E-value) cutoff for human-to-human comparisons might be 250, whereas rat-to-human may be 150, and yeast-to-human may be as low as 100 or less, i.e., the cutoff values are significantly organism related and may not be related to the assigned names of genes. The actual cutoffs used must also consider the nature of the query sequence; in particular, 3′ tails (poly-A containing) are more difficult to match across species than are full-length coding sequences.

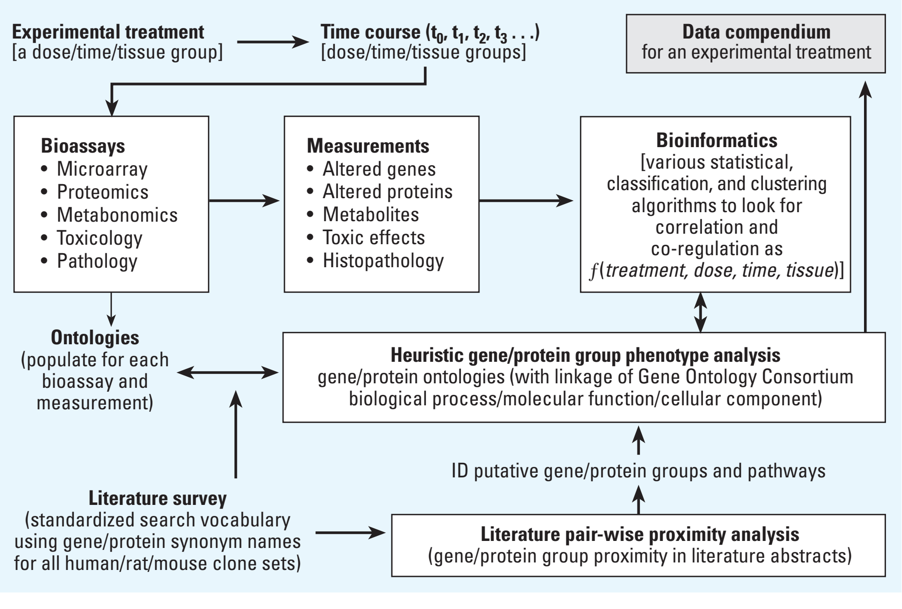

Experimental treatment
[a dose/time/tissue group]
Time course (t_0, t_1, t_2, t_3 . . .)
[dose/time/tissue groups]
Data compendium
for an experimental treatment
Bioassays
• Microarray
• Proteomics
• Metabonomics
• Toxicology
• Pathology
Measurements
• Altered genes
• Altered proteins
• Metabolites
• Toxic effects
• Histopathology
Bioinformatics
[various statistical, classification, and clustering algorithms to look for correlation and co-regulation as
f(*treatment, dose, time, tissue*)]
Ontologies
(populate for each bioassay and measurement)
Heuristic gene/protein group phenotype analysis
gene/protein ontologies (with linkage of Gene Ontology Consortium biological process/molecular function/cellular component)
ID putative gene/protein groups and pathways
Literature survey
(standardized search vocabulary using gene/protein synonym names for all human/rat/mouse clone sets)
Literature pair-wise proximity analysis
(gene/protein group proximity in literature abstracts)

Figure 9. Heuristic gene/protein functional analysis, including ontological and literature orientation to describe biological process, molecular function, and cellular component.

A Dose/Phenotype Strategy

Another strategy to be carefully considered in the development of the CEBS knowledge base is one based on the lowest effective dose required to produce a particular molecular expression phenotype or phenotype severity. We believe that quantitative structure–activity relationship (QSARs) can be developed only for discrete toxicogenomic events and outcomes that can be anchored in effective dose and a particular toxicological/pathological response or outcome. Precise phenotypic anchoring of discrete toxicogenomic events (derivation of unique gene/protein group signatures) at their lowest effective dose will be possible only if the internal dose can be established or modeled for the particular agent or its metabolites in the target tissue. This lowest effective dose/toxicant signature strategy has been employed successfully in the development of the U.S. Environmental Protection Agency/International Agency for Research on Cancer genetic activity profile database (Waters et al. 1991). Graphic profiles and corresponding data listings of lowest effective/highest ineffective doses for genotoxic agents in various cell types and organisms and for various end points are available in this database of approximately 700 compounds. To develop a similar database for toxicogenomic end points, one annotates and organizes gene expression data sets as a function of compound, organism, end point, dose, and time for select verified gene groups and co-regulated ESTs. One then plots, for example, as a histogram, outlier upregulated and downregulated genes for any appropriate toxicological or pathological end point as a function of lowest effective dose. Note that unidentified but co-regulated ESTs (i.e., ESTs associated with other genes seen to be upregulated or downregulated in response to an environmental toxicant) can contribute to the histogram and potentially to the generation of new knowledge about the mechanism of action of the compound. It should be noted that there will be primary, secondary, and tertiary effects of the same toxicant that will be distinguished from one another on the basis of the molecular and toxicological/pathological phenotypes described and documented in the knowledge base.

Resulting histogram plots are phenotypically anchored in dose and condition of target tissue and facilitate ready development of global QSARs for compounds and specific end points under consideration. Such a quantitative end point profiling approach can readily be combined with PB/PK and pharmacodynamic modeling. (In fact, such modeling can be used to derive an estimate of internal dose in the target tissue.) One then has the possibility to develop quantitative descriptions of the relationships among gene, protein, and metabolite expression profiles as a function of applied dose of the agent under consideration and to model ensuing kinetic and dynamic dose–response parameters in various tissue compartments. This is an important strategy for CEBS, as it will contribute directly to future advancements in PB/PK and pharmacodynamic modeling and support a formal quantitative risk-assessment process (Simmons and Portier 2002).

Cross-Species Gene/Protein Comparative Expression Profiling

With the availability of full genome sequences for several model organisms, there is intensive research toward the prediction, annotation, and mapping of genes across species. Of particular interest are the protein-coding genes and the intracellular signaling networks and their interactions. Similarities among novel protein sequences in model organisms have become an important and extremely useful source for hypotheses concerning protein function. *Drosophila melanogaster* and *Caenorhabditis elegans* are attractive animal model systems for studying human genes because of their genetic tractability and their phenotypically well-characterized genes (Chervitz et al. 1998; Culetto and Sattelle 2000; Nelson 1999a; Rubin and Merchant 2000).

The genome database at the NCBI has assembled Clusters of Orthologous Groups (*http://www.ncbi.nlm.nih.gov/COG/*) for homologous nucleotide sequences in more than 40 species, mainly microbial but including *D. melanogaster, C. elegans, and Saccharomyces cerevisiae*. The functional analysis of homologous genes in diverse genetic models is particularly relevant for proteins involved in human diseases to gain rapid understanding of human disease mechanisms and to enhance the probability for development of novel therapies (Rubin et al. 2000).

A number of cell functions are regulated by similar gene families across organisms (e.g., genes for the regulation of the cell cycle, cytoskeleton, cell adhesion, cell signaling, and apoptosis). This conservation of essential genes is also observed for transcription factors and many downstream signaling processes. It is believed that the completion of mouse, rat, and zebrafish genome sequencing efforts will provide information not only for the characterization of novel genes but also for the existence of homologous genes involved in every aspect of cell growth and functional differentiation. Gaining an understanding of the evolution and function of stress-response genes from yeasts to humans, for example, could be extremely valuable. Thus, we will provide within CEBS links appropriate genome information resources and eventually develop a comprehensive inventory of homologous genes/proteins across species from yeast to humans that may be important in toxicology and human disease. We anticipate that many of these homologous genes may be expressed similarly in response to environmental exposures that display similar modes of action. Strategically, these stressor-responsive genes and gene clusters could be crucial for the interpretation of cross-genome expression profiles in an integrated health and ecological risk assessment. A core set of homologous genes should include genes involved in xenobiotic activation/detoxification mechanisms, perturbations in cell homeostasis mechanisms, oxidative damage, cell injury, death, and regeneration, and genes controlling critical signaling mediator molecules for these biological processes. Phase I and Phase II enzymes metabolize most environmental xenobiotic chemicals, and much is known about their chemical substrates, inducers, and inhibitors. Phase I enzymes, the cytochromes P450 (CYPs), bioactivate as well as detoxify xenobiotics. The primary step involved in the activation process mediated by CYP proteins is oxidation, or bioactivation of xenobiotics to electrophiles. Phase II enzymes conjugate some of these oxidized metabolites to water-soluble excretable substances. We will begin our compilation of cross-species gene/protein comparative expression analysis by focusing on xenobiotic metabolic enzymes, the CYPs. Approximately 2,500 *CYP* genes have been characterized from many organisms (*http://drnelson.utmem.edu/CytochromeP450.html*), including bacteria and mammalian systems (Nebert and McKinnon 1994; Nelson 1999b; Nelson et al. 1996) and their substrate, inducer, and inhibitor specificities must be studied in relation to alterations in molecular expression across species and across classes of xenobiotics.

We anticipate that as homologous genes are identified, as compendia of gene/protein expression profiles are developed, and as functional pathways are derived and studied across species, we will be able to begin defining the networks and systems level of biological organization, wherein the cell expresses global change in response to environmental stimuli. Again, we believe that fully context-documented toxicogenomics data sets and mathematical modeling will enable development of an integrated systems toxicology and bioinformatics. In summary, CEBS Phase IV will create the capability to assess the global toxicogenomic responses of biological systems to environmental stressors and to relationally link toxicogenomics data to conventional effects data. Because CEBS Phase IV will include data sets on multiple experimental organisms, cross-species comparisons and extrapolations will be possible at molecular, subcellular, cellular, organ, and systems levels.

Further Development of the Phase IV CEBS Knowledge Base

On the basis of the foregoing discussion and advances in the field, we have attempted to describe the basic strategies for the development of the core of the CEBS knowledge base as it is conceptualized at the present time. Two additional CEBS Phase IV modules are envisioned for the future. One is a transcription module that may be used to predict the expression of genes a priori, and the other is a haplotype linkage-disequilibrium module that may be used to predict the differential expression of genes in human haplotypes and to estimate the relative sensitivity of population subgroups. The transcription module will build upon rapidly developing knowledge of transcription factors and their pivotal importance in gene regulation. Because the number of transcription factors appears limited (around 2000 for humans), their study to include sequence definition and binding sites can be developed into a predictive science as related to gene and protein expression (Forde et al. 2002; Schrem et al. 2002; Wingender et al. 2000, 2001). The haplotype linkage-disequilibrium module, on the other hand, will take advantage of our evolving knowledge of human haplotypes and associated SNPs that confer differential responses within human population subgroups to various classes of environmental toxicants and stressors (Li 2001). This module will require the addition of a SNPs database. NIEHS has for some time been engaged in the development of the GeneSNPs Database (*http://www.genome.utah.edu/genesnps/*). It should be noted that SNPs represent only approximately 90% of all DNA sequence variants. The remainder includes insertions, deletions, inversions, and duplications (1 base or many bases or kilobases). Any or all of these can be important in any gene being studied. We anticipate that the addition of a SNPs database will enable an understanding of the relationship between environmental exposures and human disease susceptibility (Li 2001). This module is important, therefore, both in a toxicological and in a risk-assessment context. Field and clinical research applications of toxicogenomics methods are anticipated by the NCT. It is well known that a single nucleotide polymorphism—a single base change in the message of a gene—can cause a protein to malfunction. Experimentally, it is possible to construct panels of mutants that enable discovery of the impacts of malfunctions in transcription and translation.

Preliminary data indicate that gene expression profiles will be useful as diagnostic tools for identifying early stages of various pathologies, including cancer (Alaiya et al. 2002; Alizadeh et al. 2001; Golub et al. 1999; Perou et al. 2000). If this approach enables earlier detection of disease than is currently possible through other approaches, it may allow earlier initiation of therapeutic interventions. Additionally, gene expression profiling may become an important tool for predicting therapeutic outcome and may be particularly useful in addressing the significant variability that has been observed in how well patients respond to different types of drug therapy. Such patterns of variability have been studied using expression profiling and, in some cases, expression signatures have been associated with individuals who are responders or nonresponders for a particular type of drug therapy. Once this kind of result is validated, it may be possible to use expression profiling to optimize the therapeutic regimen for individual patients, thus increasing the chance of a good treatment outcome. It may also be possible to identify susceptible subpopulations for purposes of quantitative risk assessment.

Conclusions

The NCT and other organizations (Castle et al. 2002; Pennie and Kimber 2002; Waring et al. 2001) are performing experiments to validate the concept of gene expression profiles as signatures of toxicant classes, disease subtypes, or other biological end points. Initial studies indicate that classes of toxicants and toxic responses can be recognized as gene expression signatures using microarray technology. Such experiments have begun to correlate gene expression profiles with other well-defined parameters, including toxicant class, chemical structure, pathological or physiological response, or other validated indices of toxicity. For example, experiments have been designed to correlate gene expression patterns with liver pathologies such as necrosis, apoptosis, fibrosis, or inflammation. It is also possible to look for correlative patterns in surrogate tissues such as blood. Changes in serum enzymes provide diagnostic markers of organ function that are routinely used in medicine and in toxicology. Such phenotypic anchoring of gene expression data using conventional indices will distinguish the toxicological signal from other gene expression changes that may be unrelated to toxicity, such as the adaptive, pharmacological, or therapeutic effects of a compound.

By constructing and populating the CEBS knowledge base, the NCT is assisting the field of environmental health research to evolve into an information science in which experimental gene and protein expression data sets are compiled and made readily available to the scientific community. Analysis of these expression profiles for different chemicals from different classes over dose and time can be used to identify expression profiles consistently and mechanistically linked to specific exposures and disease outcomes. Once sufficient high-quality data have been accumulated and assimilated, it will be possible to characterize an unknown biological or physical sample by comparing its gene and/or protein expression profile to compendia of expression profiles in the database (Hughes et al. 2000). The NCT will develop the capacity to use gene expression signatures to facilitate toxicological characterization of toxicants and their biological effects. As the field of toxicogenomics evolves, toxicogenomics databases will begin to support predictive toxicology and hazard assessment. This will help scientists predict the toxicological impact of suspected toxicants and calculate how much of a hazard these toxicants actually represent to human and environmental health.

Infrastructure development is essential to facilitate the integration of the existing public toxicology and structure–activity databases with those under development in toxicogenomics (Richard and Williams 2002). In this way, conventional toxicology and structure–activity databases and the CEBS public knowledge base can realize their full potential in supporting mechanistic interpretations and risk assessment (Simmons and Portier 2002) in the future. Development of the databases must be concomitant with the evolution of bioinformatics and data mining tools and the individuals trained to apply them.

The NIEHS is committed to the development of the CEBS knowledge base with which to initiate this evolutionary process. This article attempts to provide a vision of what the CEBS knowledge base will offer and, in general terms, how it will be constructed. The magnitude of the effort required to develop and populate such a knowledge base requires a collective will and collaborative efforts. Therefore, we will pursue the interoperability of CEBS with other databases elsewhere (e.g., those on cell signaling, protein–protein interactions, and biological and metabolic pathways) to enhance our ability to interpret toxicogenomic data sets. We will seek to develop additional mechanisms through which partnerships with scientists in academic, private sector, and other governmental organizations can be created, and we welcome advice, criticism, and participation in this enterprise.

As the CEBS knowledge base expands to include structurally or functionally related agents and as gene identity and annotation progresses, it will be possible to search in a comprehensive way for common, critical, or causal relationships. It will then be possible to create pathway maps of common cellular processes, to map partial genome arrays to pathways, and to link such changes to known phenotypic markers of toxicity. The proposed knowledge base and its relational linkages must grow incrementally, and the developers and users must have the patience and dedication to stay the course. Such incremental growth will eventually become exponential growth and the field of toxicology will be profoundly changed.

In the realm of molecular epidemiology, our growing understanding of genomic anatomy (gene sequence and polymorphisms) will form the basis for characterizing person-to-person and ethnogeographic sequence variations in genes that affect responses to drugs and chemicals that affect human susceptibility/vulnerability. Eventually, gene and protein expression profiles from exposed humans (and from organisms in the environment) will be compared with reference expression profiles based on national or international gene expression databases (Ermolaeva et al. 1998). Studying and analyzing patterns of gene expression across species will help us understand the relationship between DNA sequence variation and the phenotype, which in turn will help us understand and integrate the assessment of human and ecological risk.

Given the vast numbers and diversity of drugs, chemicals, and environmental agents, the diversity of species in which they act, the time and dose factors critical to the induction of beneficial and adverse effects, the diversity of phenotypic consequences of exposures, etc., it is only through the development of a profound knowledge base that toxicology and environmental health can rapidly advance. Toxicogenomics has the potential to change how environmental toxicology is performed. It will contribute new methods, new data, and new interpretation to the field. The ultimate goal of the NCT is to create a knowledge base that allows environmental health scientists and practitioners to understand and prevent adverse environmental exposures in the 21st century.

> Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"—T. S. Elliott

References

Aardema MJ, MacGregor JT. 2002. Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies. Mutat Res 499(1):13–25.

Afshari CA. 2002. Perspective: microarray technology, seeing more than spots. Endocrinology 143(6):1983–1989.

Alaiya AA, Franzen B, Hagman A, Dysvik B, Roblick UJ, Becker S, et al. 2002. Molecular classification of borderline ovarian tumors using hierarchical cluster analysis of protein expression profiles. Int J Cancer 98(6):895–899.

Alizadeh AA, Ross DT, Perou CM, van de Rijn M. 2001. Towards a novel classification of human malignancies based on gene expression patterns. J Pathol 195(1):41–52.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403–410.

Bartosiewicz MJ, Jenkins D, Penn S, Emery J, Buckpitt A. 2001. Unique gene expression patterns in liver and kidney associated with exposure to chemical toxicants. J Pharmacol Exp Ther 297(3):895–905.

Bessems JG, Vermeulen NP. 2001. Paracetamol (acetaminophen)-induced toxicity: molecular and biochemical mechanisms, analogues and protective approaches. Crit Rev Toxicol 31(1):55–138.

Boorman GA, Anderson SP, Casey WM, Brown RH, Crosby LM, Gottschalk K, et al. 2002. Toxicogenomics, drug discovery, and the pathologist. Toxicol Pathol 30(1):15–27.

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(4):365–371.

Bundy JG, Spurgeon DJ, Svendsen C, Hankard PK, Osborn D, Lindon JC, et al. 2002. Earthworm species of the genus *Eisenia* can be phenotypically differentiated by metabolic profiling. FEBS Lett 521(1–3):115–120.

Burchiel SW, Knall CM, Davis JW II, Paules RS, Boggs SE, Afshari CA. 2001. Analysis of genetic and epigenetic mechanisms of toxicity: potential roles of toxicogenomics and proteomics in toxicology. Toxicol Sci 59(2):193–195.

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(6):564–565.

Castle AL, Carver MP, Mendrick DL. 2002. Toxicogenomics: a new revolution in drug safety. Drug Discov Today 7(13):728–736.

Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, et al. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 282(5396):2022–2028.

Culetto E, Sattelle DB. 2000. A role for *Caenorhabditis elegans* in understanding the function and interactions of human disease genes. Hum Mol Genet 9(6):869–877.

Cunningham MJ, Liang S, Fuhrman S, Seilhamer JJ, Somogyi R. 2000. Gene expression microarray data analysis for toxicology profiling. Ann N Y Acad Sci 919:52–67.

Dudley AM, Aach J, Steffen MA, Church GM. 2002. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc Natl Acad Sci U S A 99(11):7554–7559.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30(1):207–210.

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(1):19–23.

Farland WH. 1992. The U.S. Environmental Protection Agency's Risk Assessment Guidelines: current status and future directions. Toxicol Ind Health 8(3):205–212.

———. 1996. Cancer risk assessment: evolution of the process. Prev Med 25(1):24–25.

Fielden MR, Zacharewski TR. 2001. Challenges and limitations of gene expression profiling in mechanistic and predictive toxicology. Toxicol Sci 60(1):6–10.

Forde CE, Gonzales AD, Smessaert JM, Murphy GA, Shields SJ, Fitch JP, et al. 2002. A rapid method to capture and screen for transcription factors by SELDI mass spectrometry. Biochem Biophys Res Commun 290(4):1328–1335.

Gibas C, Jambeck P. 2001 Developing Bioinformatics Computer Skills. Sebastopol, CA:O'Reilly & Associates, Inc.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537.

Hamadeh HK, Amin RP, Paules RS, Afshari CA. 2002a. An overview of toxicogenomics. Curr Issues Mol Biol 4(2):45–56.

Hamadeh HK, Bushel PR, Jayadev S, DiSorbo O, Bennett L, Li L, et al. 2002b. Prediction of compound signature using high density gene expression profiling. Toxicol Sci 67(2):232–240.

Hamadeh HK, Bushel PR, Jayadev S, Martin K, DiSorbo O, Sieber S, et al. 2002c. Gene expression analysis reveals chemical-specific profiles. Toxicol Sci 67(2):219–231.

Hamadeh HK, Knight BL, Haugen AC, Sieber S, Amin RP, Bushel PR, et al. 2002d. Methapyrilene toxicity: anchorage of pathologic observations to gene expression alterations. Toxicol Pathol 30:470–482.

Holmes E, Nicholls AW, Lindon JC, Connor SC, Connelly JC, Haselden JN, et al. 2000. Chemometric models for toxicity classification based on NMR spectra of biofluids. Chem Res Toxicol 13(6):471–478.

Holmes E, Nicholson JK, Tranter G. 2001. Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. Chem Res Toxicol 14(2):182–191.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. 2000. Functional discovery via a compendium of expression profiles. Cell 102(1):109–126.

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

Issaq HJ, Veenstra TD, Conrads TP, Felschow D. 2002. The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. Biochem Biophys Res Commun 292(3):587–592.

Karp PD. 2000. An ontology for biological function based on molecular interactions. Bioinformatics 16(3):269–285.

Larsen JC, Farland W, Winters D. 2000. Current risk assessment approaches in different countries. Food Addit Contam 17(4):359–369.

Li H. 2001. A permutation procedure for the haplotype method for identification of disease-predisposing variants. Ann Hum Genet 65:189–196.

Nebert DW, McKinnon RA. 1994. Cytochrome P450: evolution and functional diversity. Prog Liver Dis 12:63–97.

Nelson DR. 1999a. Cytochrome P450 and the individuality of species. Arch Biochem Biophys 369(1):1–10.

———. 1999b. A second CYP26 P450 in humans and zebrafish: CYP26B1. Arch Biochem Biophys 371(2):345–347.

Nelson DR, Koymans L, Kamataki T, Stegeman JJ, Feyereisen R, Waxman DJ, et al. 1996. P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics 6(1):1–42.

Nicholson JK, Connelly J, Lindon JC, Holmes E. 2002. Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 1(2):153–161.

Nicholson JK, Lindon JC, Holmes E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29(11):1181–1189.

Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. 1999. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24(3):153–159.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27(1):29–34.

Olden K. 2002. New opportunities in toxicology in the post-genomic era. Drug Discov Today 7(5):273–276.

Pennie WD, Kimber I. 2002. Toxicogenomics; transcript profiling and potential application to chemical allergy. Toxicol In Vitro 16(3):319–326.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. 2000. Molecular portraits of human breast tumours. Nature 406(6797):747–752.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, et al. 2001. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 29(1):159–164.

Reilly TP, Bourdi M, Brady JN, Pise-Masison CA, Radonovich MF, George JW, et al. 2001a. Expression profiling of acetaminophen liver toxicity in mice using microarray technology. Biochem Biophys Res Commun 282(1):321–328.

Reilly TP, Brady JN, Marchick MR, Bourdi M, George JW, Radonovich MF, et al. 2001b. A protective role for cyclooxygenase-2 in drug-induced liver injury in mice. Chem Res Toxicol 14(12):1620–1628.

Richard AM, Williams CR. 2002. Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. Mutat Res 499(1):27–52.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, et al. 2000. Comparative genomics of the eukaryotes. Science 287(5461):2204–2215.

Rubin RB, Merchant M. 2000. A rapid protein profiling system that speeds study of cancer and other diseases. Am Clin Lab 19(8):28–29.

Ruepp SU, Tonge RP, Shaw J, Wallis N, Pognan F. 2002. Genomics and proteomics analysis of acetaminophen toxicity in mouse liver. Toxicol Sci 65(1):135–150.

Schrem H, Klempnauer J, Borlak J. 2002. Liver-enriched transcription factors in liver function and development. Part I: The hepatocyte nuclear factor network and liver-specific gene expression. Pharmacol Rev 54(1):129–158.

Selkov E Jr., Grechkin Y, Mikhailova N, Selkov E. 1998. MPW: the Metabolic Pathways Database. Nucleic Acids Res 26(1):43–45.

Simmons PT, Portier CJ. 2002. Toxicogenomics: the new frontier in risk analysis. Carcinogenesis 23(6):903–905.

Sluka JP. 2002. Extracting knowledge from genomic experiments by incorporating the biomedical literature. In: Methods of Microarray Data Analysis, II (Lin SM, Johnson KF, eds). Boston:Kluwer.

Tennant RW. 2002. The National Center for Toxicogenomics: using new technologies to inform mechanistic toxicology. Environ Health Perspect 110(1):A8–A10.

Thomas RS, Rank DR, Penn SG, Zastrow GM, Hayes KR, Pande K, et al. 2001. Identification of toxicologically predictive gene sets using cDNA microarrays. Mol Pharmacol 60(6):1189–1194.

Ulrich R, Friend SH. 2002. Toxicogenomics and drug discovery: will new technologies help us produce better drugs? Nat Rev Drug Discov 1(1):84–88.

Waring JF, Cavet G, Jolly RA, McDowell J, Dai H, Ciurlionis R, Zhang C, Stoughton R, Lum P, Ferguson A, et al. Development of a DNA microarray for oxicology based on hepatotoxin-regulated sequences. Environ Health Perspect 111:863–870 (2003).

Waring JF, Jolly RA, Ciurlionis R, Lum PY, Praestgaard JT, Morfitt DC, et al. 2001. Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. Toxicol Appl Pharmacol 175(1):28–42.

Waters MD, Stack HF, Garrett NE, Jackson MA. 1991. The Genetic Activity Profile database. Environ Health Perspect 96:41–45.

Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, et al. 2001. The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29(1):281–283.

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, et al. 2000. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28(1):316–319.

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(6):625–637.

Yamazaki K, Kuromitsu J, Tanaka I. 2002. Microarray analysis of gene expression changes in mouse liver induced by peroxisome proliferator-activated receptor alpha agonists. Biochem Biophys Res Commun 290(3):1114–1122.

Zweiger G. 1999. Knowledge discovery in gene-expression-microarray data: mining the information output of the genome. Trends Biotechnol 17(11):429–436.