



Emerging Issues

in Environmental Health Sciences

Issue 4
November 2003

The Newsletter of the
Committee on Emerging Issues and
Data on Environmental Contaminants

Bioinformatics: Getting There from Here

With the advent of techniques to measure a variety of changes in gene, protein, or metabolites on a large scale, bioinformatics and computational biology are becoming more important in extracting knowledge from all types of biological data. Toxicogenomics (TG) data are no exception. The Committee on Emerging Issues and Data on Environmental Contaminants organized an open meeting on September 15, 2003 to learn more about bioinformatics and computational biology challenges in examining changes in gene, protein, and metabolite expression.

Challenges in collecting, storing and analyzing TG data, and “-omics” data in general, were the focus of several presentations. Catherine Ball (Stanford U.) emphasized the importance of ontologies and controlled vocabularies in providing structure for capturing and exchanging information, while Susanna-Assunta Sansone (European Bioinformatics Institute) described the collaborative efforts of the EBI, NIEHS, and ILSI-HESI to engage the TG community in harmonizing collection of gene expression data. In collaboration with the Microarray Gene Expression Data Society, these organizations have developed the MIAME/TOX standards for describing TG experiments, including how to represent data, collect contextual information, and record controls. Sansone said that further issues include agreeing on appropriate minimal descriptors and on controlled vocabularies to describe experiments, especially conventional toxicology vocabularies. Paul Spellman (Lawrence Berkeley National Laboratory) described efforts to develop a mark-up language to address these issues. In addition, Ball advocated funding of existing

“grassroots” projects, communication and synergy between projects, and open access to controlled vocabularies and ontologies.

Beyond harmonizing the collection and storage of data, challenges also lie in developing tools to mine databases for meaningful information. Roger Bumgarner (U. Washington) advocates that databases allow direct access to raw data, but the different raw data formats make this difficult. Analysis software also needs improvement so that different experiments probing the same gene can be compared. As to who develops the analytical software, Bumgarner suggested that market-driven software is more likely to be well supported and driven by users’ needs than free software.

The question of responsibility for database maintenance was also discussed by Bumgarner. He believes it is most practical for large centers to take

Continued on page 3



IN THIS ISSUE

1 *Bioinformatics: Getting There from Here*

2 *A Bioinformatics and Biostatistics Perspective*

4 *Validation of “-OMICS” Biomarkers*

7 *Agenda for December 15-16 Meeting*

The Challenges of Toxicogenomics: A Bioinformatics and Biostatistics Perspective

John Quackenbush, The Institute for Genomic Research



Completion of draft genome sequences from human, mouse, and rat provides a preliminary “parts list” for these organisms and sets the stage for using holistic approaches to answer a wide range of basic biological questions. To exploit the resources provided by genome projects, a variety of functional genomics approaches, including expression profiling, proteomics, and metabonomics, have evolved.

Toxicologists were among the first to see the promise of these genomic approaches. The emerging field of toxicogenomics attempts to understand the basis of toxicity and predict which compounds are likely to trigger specific toxic responses. The reason for the great enthusiasm for toxicogenomics approaches is very simple – they can provide information on responses of tens of thousands of genes, or hundreds of proteins or metabolites, in a single assay. These data can then be used to elucidate the pathways involved in particular toxic responses.

With the rapid proliferation of these technologies, it is natural to ask why fulfilling these goals remain little more than a promise. The answer is that there are a number of issues associated with collecting, managing, and analyzing the data that must first be addressed.

John Quackenbush is a member of the Committee on Emerging Issues and Data on Environmental Contaminants. He is an Investigator at The Institute for Genomic Research in Rockville, MD

Experimental Design

The starting point for any experiment, including those for functional genomics, is a

sound experimental design. Ideally, an experimental design should investigate a wide range of toxicological parameters, including the exposure level, the exposed population and its genotype, and the temporal pattern of response. In reality, however, given the high cost of functional genomics experiments, parameter measurements must be carefully chosen. As a result, there is no standard design for a functional genomics experiment. Rather, each experiment must be tailored to appropriately address the question under investigation while taking into account constraints such as the availability of RNA, the ease of automation and implementation in the laboratory, and resources and budget. Nevertheless, some simple design considerations have emerged, such as the need for sufficient biological replication to allow for sound estimates of the relative expression levels and their changes.

Databases

Effective use of genome scale data requires the development of databases that can effectively capture the experimental information and make it available to other researchers, as well as integrating it with other biologically relevant data. The many experimental variables in functional genomics pose significant challenges in managing these experimental data. Consequently, there has been an effort to establish standards for storing and communicating functional genomics data. The Microarray Gene Expression Data Society (MGED) has developed a widely accepted standard known as MIAME (Minimal Information About a Microarray Experiment) that outlines the information necessary for anyone to understand and analyze the data from a microarray study, and MAGE-ML, a data exchange standard. MIAME-

Continued on page 5

Getting There from Here

Continued from page 1

responsibility for the expensive and long term tasks of data repository maintenance and data sharing. Current efforts along these lines are seen as limited because centers focus on data relevant to their own interests.

Access to large databases also leads to new challenges. Sandrine Dudoit (UC-Berkeley) described the need for new statistical methods to analyze multiple datasets. New methods, for example, are being developed for making predictions on the basis of observable variables. Assessing the performance of these methodologies will depend on understanding the underlying biology (i.e., which models are plausible); access to benchmark datasets; and the use of better experimental controls.


Kathleen Kerr (U. Washington) and Bumgarner noted that the ability to analyze large datasets would not in itself create useful scientific findings. They stressed the importance of adequately designing gene expression (microarray) experiments to answer clearly defined questions of interest. Kerr described three factors important to consider when designing microarray experiments: 1) the choice of statistical method to control for technical error can impact experiment conclusions; 2) to control for biological variability, it is more important to sample more biological variables (e.g., more animals) than to sample each animal more often; and 3) randomization is an important classic facet of experimental design often overlooked in gene expression experiments, but it is important for preventing introduction of unidentified artifacts into experiments. Bumgarner suggested that more training of biologists in the use of statistical/analytical tools would help improve experimental design.

Moving from gene expression to proteomics and metabolomics, Richard Smith (Pacific Northwest Research Laboratory) discussed challenges in proteomics. "Proteomics" describes a number of different approaches to assaying protein expression. Just the variety of analytical methods and of

platforms creates a challenge in data sharing and portability, because of the proliferation of resulting data formats and databases. Development of accepted standards for data exchange and open-source software should help overcome this issue. Technical challenges include the need for more confidence in identifying proteins/peptides and increased precision of quantitative measurements. Better understanding of sample/data variation and quality control is needed to increase precision in quantification. Smith is hopeful that an increase in the number of laboratories working in this area will help meet these challenges.

"Metabolomics" is the comprehensive measurement of metabolites in biological samples. According to Pedro Mendes (Virginia Bioinformatics Institute), comparing changes in metabolite levels caused by genetic mutations can increase our understanding of gene functions. Mendes described the challenge of identifying metabolites represented by chromatographic peaks, because the libraries describing chromatographic peaks of known substances are limited in the number of metabolites characterized and the variability in the instruments used. Given the difficulties in identifying metabolites, a standard for classifying and retaining data for unidentified metabolites would be helpful. Similarly, standards for describing metabolomic experiments in general would be useful but have not been developed.

John Weinstein (National Cancer Institute) outlined how "-omics" research may be used for cancer treatment and prognosis. His emphasis on use of "-omics" information for public health provided a different perspective than the description of challenges, prompting some audience and committee members to ask whether enough emphasis is being placed on the analysis and collection of biologically meaningful data.

This open meeting accomplished the goal of learning about some of the technical challenges in bioinformatics. At the next open meeting the committee hopes to focus on knowledge gaps in risk assessment and how "-omics" technologies may address them. 

VALIDATION OF “-OMICS” BIOMARKERS: IT’S BACK TO BASICS

John D. Groopman, Johns Hopkins University



People are exposed to chemical, physical, or biological agents through contaminated air, water, soil, or food [1]. Each person is believed to have a unique response to exposure to an environmental agent in terms of dose received and time to disease onset. This response may trigger the presence of one or more biomarkers. Biomarkers are indicators of an event in a biological system; they may be used to indicate exposure to an environmental agent, an effect in a biological system as a result of an exposure, or to identify individuals that are likely to be highly sensitive to an exposure. The development of “-omic” technologies to determine changes in gene (genomics) and protein (proteomics) expression, and in metabolism (metabonomics) has led to their use in identifying biomarkers of human health status. Genomic, proteomic, or metabonomic-based biomarkers can provide important information for regulatory, clinical, and public health problems [2,3].

Exposure Markers

Human exposures result from contact with an environmental chemical but they may be modified by many factors, both intrinsic (genetic) or extrinsic (e.g., dietary). Biomarkers of exposure may be the presence of the chemical itself in a person’s tissues, such as lead in the blood of children who have lead-based paint in their homes. More frequently, however, it is the metabolic products of the chemical

in the body that serve as the markers of exposure. Other biomarkers of exposure include complexes (adducts) of the chemical with DNA, proteins, or other biological molecules. For example, biomarkers for aflatoxin adducts may be used as early indicators of liver cancer. Finally, altered protein structures or functions can also serve as the exposure markers when the changes in gene expression (genomics) can be directly linked to changes in the protein (proteomics). For example, viral protein markers in blood are used to relate specific stages of exposure to disease risk.

Risk Markers

Not every measure of exposure will accurately reflect an individual’s risk of disease. Most biomarkers will only indicate risk if they are involved in the mechanism of disease. For example, a chemical-DNA adduct, if shown to be linked to a specific gene mutation that might cause cancer, could be considered a validated risk biomarker. Given the complex biological processes and long latency period for the development of cancer and other chronic human diseases, relatively few chemical-specific biomarkers are expected to be validated as risk

markers for these diseases. Furthermore, many experimental studies show that the presence of individual chemical-specific biomarkers do not always indicate the development of an adverse effect in that individual. Consequently, multiple chemical-specific biomarkers may be necessary to estimate risk from an environmental chemical. Indeed, most validated risk markers may turn out to be groups of biomarkers, each of which contributes quantifiably to an individual’s overall risk. For example, the presence of biomarkers for aflatoxin adducts, combined with aflatoxin-induced gene mutations and

John Groopman is a member of the Committee on Emerging Issues and Data on Environmental Contaminants. He is the Anna M. Baetjer Professor of Environmental Health and Chair of the Department of Environmental Health Sciences at the Johns Hopkins Bloomberg School of Public Health.

Continued on page 6

Bioinformatics and Biostatistics Perspective

Continued from page 2

TOX represents the first step in incorporating descriptions of the toxicological portions of functional genomics experiments. Efforts are underway to create similar standards and vocabularies for proteomics and metabonomics as well.

Data Analysis

Toxicogenomics data analysis also presents a number of challenges—a significant problem is multiple testing, which results from looking at far more variables (typically tens of thousands of genes or thousands of proteins or metabolites) than biological samples. Developing better tools and techniques for analyzing microarray data will help avoid the problem of false positives and negatives. Additional biostatistical and bioinformatics research will be necessary to provide greater confidence in experimental results.

Placing differentially regulated genes, proteins, or metabolites into a biological context also remains difficult. Although these entities may be useful for predicting whether a compound is toxic, identifying the underlying mechanisms requires intensive data mining and integration of functional genomics experiments with knowledge about the individuals and compounds being studied and toxicological and biological data from other experimental techniques and systems. While software and techniques are being developed to address these issues, the results must be ultimately filtered by their biological relevance and validated through further experimentation.

Toxicology and Cross-species Comparisons

Although functional genomics provides tools and unprecedented quantities of data to address toxicological questions, the value of the data relies on the validity of the underlying toxicological experimental system. Questions about any

compound must be carefully considered and addressed, including the relevant dose, time of exposure, mode of delivery, and the presence or absence of additional factors that might modify response. Equally important are questions associated with the validity of cross-species comparisons. While rodent and other models are used as surrogates for humans, comparing specific pathways and mechanisms of response requires a careful understanding not only of the analogous gene products, but more importantly, determining which pathways and systems can be compared across species.

Delivering on the Promise

While many challenges remain in toxicogenomics, none is insurmountable. Work is already underway to develop standards for expression studies and their data, databases have been established and are becoming increasingly utilitarian, software is increasing in sophistication, and methods for better statistical analysis and data integration are rapidly evolving. Testing and validation of these systems will, however, require a large body of high quality data from carefully controlled studies. While many of the studies that have been conducted to date are valuable, they only represent a fraction of what is needed to fully understand the limitations of our current approaches and to identify areas that will require further development. Although we may not be ready just yet to deliver on the promise of toxicogenomics, it is clearly time to start building the experimental, computational, and intellectual infrastructure that will allow us to do so. ●

Presentations from the September 15, 2003 open meeting on challenges in bioinformatics are now posted on the Committee's website at <http://dels.nas.edu/emergingissues>.

VALIDATION OF “-OMICS” BIOMARKERS

Continued from page 4

exposure to hepatitis B virus, has been found to be a good predictor of liver cancer.

Need for Biomarker Validation Strategies

The major goals of “-omic”-based biomarker research are to develop and validate biomarkers that reflect specific exposures and to predict an individual’s risk of disease. To be useful in molecular epidemiology studies, biomarkers should be reliable, accurate, and precise. Biomarker studies should help identify the molecular processes of chemically induced disease and underlying susceptibility factors. The analytical methods for measuring the biomarkers must be sufficiently sensitive and specific to quantify them in a limited biological sample (e.g., blood, urine) from a single person. As a result, biomarkers are interpreted on an individual basis and not on a population basis. For biomarkers to be truly valid and valuable, they should be indicative of the range of an individual’s response from initial exposure to disease outcome. Biomarkers that reflect the mechanism of action (for example, gene mutation) of an environmental chemical may be strong predictors of an individual’s risk of disease. Biomarkers may also help determine whether individuals, communities, or larger populations have been exposed to an environmental agent and the magnitude of exposure. The practical development of specific biomarkers to assess exposure to and risk from environmental agents, will require integrating multiple routes of exposure and fluctuating exposures with time, relating time of exposure to the internal and biologically effective dose, and examining molecular mechanisms in biological targets. Accurate biomarkers will reduce misclassification of exposure in individuals and across populations; exposure misclassification is often the greatest source of error in environmental epidemiology studies. Finally, biomarkers should provide an objective measure for determining the effectiveness of interventions to lower exposure and risk.

The Future of Biomarkers

Biomarker strategies for measuring exposure in individuals could completely change how risk assessments and environmental regulations are used to improve public health. Biomarkers can be used to address questions about exposure and risk, such as: What is the exposure of an individual living along the fence-line of a hazardous waste area? Are there individuals in the population at greater risk for disease? Using chemical-specific biomarkers to identify high-exposure and high-risk individuals will reduce some of the current uncertainties in the risk assessment process.

Although progress has been made in applying “-omic” technologies to specific health risks, the initial excitement at being able to measure changes in genes, proteins, and metabolites has been replaced by the challenge of how to interpret the results. This problem is exacerbated by the complex interactions between genes, proteins and other molecules, and environmental factors that underlie most human disease. Biomarkers cannot be effectively used until they can be shown to link exposure and specific changes in a gene, protein, or other molecule. The overwhelming majority of biomarkers have not been validated and cannot currently be used in health analyses. To use biomarkers to guide public health issues, molecular epidemiologists must devise and follow careful strategies for the validation, application, and dissemination of information about biomarkers to the public. Assembling and validating this collection of biomarkers will be a major challenge over the next decade. ●

REFERENCES

1. Groopman, J.D., Jackson, P.E., Turner, P., Wild, C.P. and Kensler, T.W. (2002) Validation of Exposure and Risk Biomarkers: Aflatoxin As A Case Study. *Biomarkers of Environmentally Associated Disease*, S.H. Wilson and W.A. Suk, Editors, pp. 307-318
2. Anonymous (1987) Biological Markers in Environmental Health Research. *Environ. Health Perspect.*, **74**, 3-9.
3. Wogan, G.N. (1992) Molecular epidemiology in cancer risk assessment and prevention: Recent progress and avenues for future research. *Environ. Health Perspect.*, **98**, 167-178.

CRITICAL ISSUES IN CARCINOGENIC RISK ASSESSMENT AND TOXICOGENOMICS TECHNOLOGIES

December 15-16, 2003

The National Academies
500 5th Street, NW
Room 201
Washington, DC 20001

Monday (Dec. 15) Morning (8:00 AM - 12:00 PM)

Key issues (critical gaps) in cancer determinations

Overview: The use of intermediate endpoint data in cancer risk assessment, using a pre-existing case study as an example

Case Study Discussion of Chemical I: Identify specific situations in risk assessment where toxicogenomics information could be very useful and where it would be less useful

Monday Afternoon (1:00 PM- 5:15 PM)

Case Study Discussion of Chemical II

Summary discussion of issues presented for the two chemicals, and extrapolation from specific cases to opportunities for toxicogenomics in general.

Tuesday (Dec. 16) Morning (8:30 AM-11:00AM)

Discussion of Monday session

Review of Risk Communication workshop plans
Mark Rothstein

Brief description of future workshop ideas
NIEHS and Federal Liaison Group input on future workshop topic selection

ADJOURN OPEN SESSION (11:00 AM)

For more information,
please contact
Jennifer Saunders
by phone at
202-334-2616,
or by e-mail at jsaunder@nas.edu.

Regulatory agencies (e.g., Environmental Protection Agency, Food and Drug Administration) are often told that new “-omic” technologies will improve chemical risk assessment, but specifics on how this may occur are not always clear. At the other end of the spectrum, scientists using the new “-omic” technologies could have a greater impact on risk assessment and toxicology by asking questions critically important to the assessment of chemical carcinogenicity. By illustrating critical gaps and discussing how the technologies may be most helpful, this meeting will help stimulate a dialog among risk assessors, toxicologists, and genomics researchers.

STAFF

Board Directors

James Reisa (BEST)

Fran Sharples (BLS)

Project Officers

Marilee Shelton

Roberta Wedge

Research Assistant

Jennifer Saunders

Project Assistants

Lucy Fusco

Robert Policelli

Newsletter

Robert Policelli

This newsletter as well as additional information about the committee and its activities can be found at <http://dels.nas.edu/emergingissues>.

The newsletter of the Committee on Emerging Issues and Data on Environmental Contaminants, “Emerging Issues in Environmental Health Sciences,” is published to keep you informed of committee activities. This is a joint project of the National Research Council’s Board on Environmental Studies and Toxicology and Board on Life Sciences. The views expressed in the articles in this Newsletter are those of the individual authors and do not reflect the findings or conclusions of The National Academies.

Committee on Emerging Issues and Data on Environmental Contaminants
Board on Environmental Studies and Toxicology
Board on Life Sciences
THE NATIONAL ACADEMIES
500 Fifth Street NW
Washington, DC 20001



An Invitation to Critical Issues in Carcinogenic Risk Assessment and Toxicogenomics Technologies

DECEMBER 15-16, 2003

The National Academies
500 5th Street, NW.
Room 201
Washington, DC 20001

THE NATIONAL ACADEMIES™
Advisers to the Nation on Science, Engineering, and Medicine

The nation turns to the National Academies—National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council—for independent, objective advice on issues that affect people's lives worldwide.

www.national-academies.org