

[HOME](#)[CURRENT ISSUE](#)[CHEMJOBS](#)[JOIN ACS](#)[EMAIL ALERTS](#)[ADVANCED SEARCH](#)

- Latest News
- Business
- Government & Policy
- Science/Technology
- Career & Employment
- ACS News



August 2, 2004
Vol. 82, Iss. 31

[View Current Issue](#)

[Back Issues](#)

SUPPORT

- How to log in
- Contact Us
- Site Map

ABOUT C&EN

- About the Magazine
- How to Subscribe
- How to Advertise



[Join ACS](#)

Science & Technology

August 2, 2004

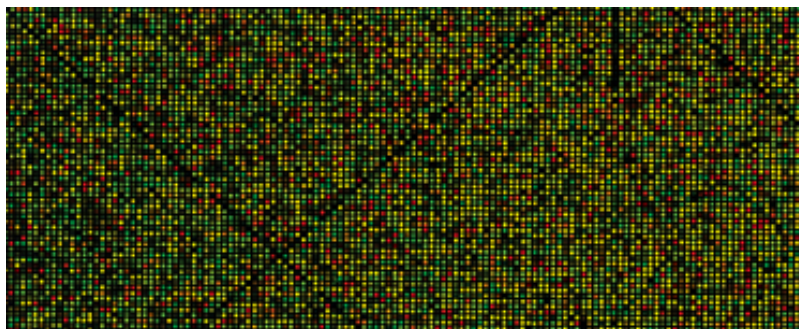
Volume 82, Number 31

pp. 36-39

STANDARDIZING DNA MICROARRAYS

Efforts to develop standards for microarray data are aimed at making data sharing easier

[CELIA M. HENRY, C&EN WASHINGTON](#)



SEEING SPOTS Standards for the information provided with DNA microarrays make it easier to interpret experiments done by others.

COURTESY OF KATJA SCHWARTZ AND GAVIN SHERLOCK/STANFORD UNIVERSITY SCHOOL OF MEDICINE

Chips containing orderly arrays of thousands of DNA sequences--so-called DNA microarrays--have become an indispensable part of the biologist's tool kit for studying gene expression and gene mapping. But replicating experiments from different labs and understanding the results have been difficult because, until recently, standards didn't exist to govern the types of information, or annotation, that researchers provide about the samples and experimental conditions. The [Microarray Gene Expression Data Society](#), also known as MGED, has taken the lead in establishing standards, which are starting to make data sharing easier.

"We need standards so we can understand experiments done by someone else," says [Catherine A. Ball](#), director of the [Stanford Microarray Database](#) and president of MGED. "When trying to verify someone else's work or see why a competitor got conflicting results or [to use] their analysis methods on your data, you need to have a common vocabulary."

Dietrich Stephan, director and senior investigator in the Neurogenomics Division of the [Translational Genomics Research Institute \(TGen\)](#) in

to a friend



[Print this article](#)



[E-mail the editor](#)

Phoenix, believes that standardization needs to be about providing the appropriate "background information" rather than about how the experiment is run. For example, when microarrays are used in a clinical setting, the "metadata set" should include an entire clinical record for the patient, he says.

"The more information you provide, the more of a correct analysis you can do on the data, and the more those data become useful to other investigators downstream," Stephan says.

For example, standards can allow data from one set of experiments to be combined with data from another experiment to increase the statistical power. "If you think about a group in Massachusetts doing 100 leukemia samples, and we're doing 100 in Phoenix, and a group in California is doing 100, none of those data sets individually has enough power to figure anything out. There just aren't enough samples," Stephan says. Combining those data sets produces what Stephan calls a "real data set."

MGED developed a set of guidelines called the MIAME standards, which stands for minimum information about a microarray experiment [*Nat. Genet.*, **29**, 365 (2001)]. The standards include all the information that scientists should provide so that others can interpret and understand their microarray results.

The MIAME standards are just a guideline. One possible way of implementing the guidelines is by using an object model and a computer language called MAGE-ML, which stands for microarray gene expression markup language. The object model breaks the microarray experiment into its components and shows how those parts are related to one another. MAGE-ML, which is based on the computer language XML, flows directly from the object model. MGED is also working on developing an ontology--a list of agreed-upon terms and definitions with references to their derivations--along with the object model, so that all the terms used will have strict definitions.

Stephan says one of the challenges is that MAGE-ML can be difficult to use. "People have been really hesitant about using the MAGE markup language just because it's so hard to get the data in the right format," he says.



MAKING OLDER data fit the format is a challenge, Ball says. "I've run a database that is now faced with having to shoehorn our database and our data into the right format to meet these standards. That's not necessarily easy, and it does take a certain amount of dedication and expertise and education," she says. "There is a certain activation energy that is required." The two main public repositories for microarray data are the Gene Expression Omnibus, or GEO, run by the [National Center for Biotechnology Information](http://www.ncbi.nlm.nih.gov/geo/) at the National Institutes of

PUBLIC PLACE Gene Expression Omnibus is one of the public repositories where scientists can put their microarray data.

COURTESY OF RON EDGAR

Health, and ArrayExpress, run by the European Bioinformatics Institute. A third repository in Japan, called CIBEX, is not yet up and

running.

Currently, there is little overlap between the repositories, although Ron Edgar, team leader at GEO, thinks that will change for the better. "I think in the future you will find more and more data shared, but it's never going to be as transparent as the sequence databases" such as GenBank.

For the transfer of data to become more transparent, Edgar believes that people need to come to an agreement about how to translate the MIAME guidelines into an exchange format. "People have to understand--and I see this confusion all the time--that there is a huge difference between MIAME as defining the required content and MAGE-ML, which is the mechanism that ArrayExpress and MGED chose to provide a syntax for data transfer." He points out that MAGE-ML is an "elaborate form" that "includes much more than what MIAME talks about."

Edgar believes that there's a fine line between asking people to do too little and too much. In fact, he believes that asking people to do too much can lead to unintentional misinformation because people take shortcuts to filling in the information.

"Let's say you have an ontology list that you need to pick something from. I think statistically you will find that most of the pickings will be from the top of the list. It's better to have free text, if the list is very long, than to have many layers of lists that you have to go through," Edgar says. "From a computational informatics point of view, nothing is better than having a very restricted vocabulary or ontology, but I'm not sure it works in the real world because the process of entering required information is often subjective and error prone."

Stephan is the chairman of a consortium that provides microarray services to about 3,000 researchers funded by two NIH institutes--the National Institute of Neurological Disorders & Stroke (NINDS) and the National Institute of Mental Health (NIMH). The [NINDS/NIMH Microarray Consortium](#) is made up of a genome center at TGen; one at Duke University; and one at the University of California, Los Angeles (<http://arrayconsortium.tgen.org>).

Working with the consortium, the Phoenix-based software company [5AM Solutions](#) has created a Web-based solution to help researchers get their data into a format that "can be disseminated throughout the community," Stephan says.

"Our company was contracted to build a solution for this consortium," says Brent Gendleman, chief executive officer at 5AM. That solution needed to be able to handle a variety of tasks, such as managing workflow, collaborations, data publication, and online analysis, but two aspects in particular drove them into the world of microarray standards.

"There was a requirement that they be able to publicly exchange the data in a meaningful way," Gendleman says. "It's more than saying, 'Here on my FTP site is a bunch of array data with a project name.' That's not enough.

"The other requirement was they wanted the project data to be MIAME compliant," he says. They created a custom mechanism for indicating MIAME compliance using the MAGE object model and MGED guidelines and significantly improved the performance of the system. Gentleman described the current state of the software project at an American Chemical Society ProSpectives-sponsored meeting on microarrays held in Boston this past June.

Stephan says that the software makes annotation painless. "If it's going to take you five hours to sit and try to figure out how to put something in MAGE-ML format, or if it's going to take you an hour to annotate each array, you're not going to do it," he notes. In contrast, "the website asks you a plain-English question, and you type in [the answer]. You basically push a button, and it will parse it all into the standard format. Then we can dump it out into all of the public databases."



SHARING The Web-based solution developed by 5AM Solutions for the Microarray Consortium makes it easier to put information about microarray experiments in the proper format.

COURTESY OF TGEN

The software is available to scientists within the NINDS/NIMH consortium.

It is also available to others by contacting 5AM. The source code itself is in the public domain, but the company charges for its time and maintenance if other institutions wish to install it at their own sites.

JOURNALS WILL play a role in making the standards, which are otherwise strictly voluntary, more mainstream. "Just like in the olden days when we were sequencing genes [and] would have to have a GenBank accession number before we could publish our papers, there are some journals that are requesting GEO or ArrayExpress accession numbers before they publish," Ball says. In her role at MGED, Ball sometimes worries about individual bench scientists who may not have access to resources as extensive as those at Stanford. "The users I interact with most are obviously the users here at Stanford, who, I have to say, have it pretty easy. We handle most of the things for them behind the scenes with software," she says. "There are a lot of people who don't really have the same access. Those are the people who are going to be faced with a lot more work putting their data in the right format and annotating them to the right level of accuracy."

Ball initially got involved with MGED to keep it from "evolving into something that was overly stringent and overly rigid," she says. "I wanted to make sure that somebody who was being pragmatic and who could represent a real bench biologist was involved. When you start talking about standards, it's very easy to start thinking about how things should be and how they could be if everything were perfect. There needs to be room for people who are doing something unexpected and creative. You don't want to prevent those people from being able to publish their data because the object model doesn't have room for that."

The current standards help ensure that there's enough background information about the microarrays. In the future, MGED would like to standardize the way people manipulate the data as well. "We need to work pretty hard on coming up with ways to record and understand data transformations, analysis, normalization--all those things we do to

the data so people can get the results out," Ball says.

Chemical & Engineering News
ISSN 0009-2347
Copyright © 2004

[Home](#) | [Latest News](#) | [Current Issue](#) | [ChemJobs](#)

[Pubs Page](#) / [chemistry.org](#) / [ChemPort](#) / [CAS](#)

[Copyright © 2004 American Chemical Society](#)