GEO: world of diagnostic potential

By Michael J. Haas, Senior Writer

Researchers from California and Taiwan have developed a way to transform DNA-based microarray data and phenotype information found in public repositories into a searchable database.¹ They think their approach will allow patients with genetically based diseases to be diagnosed faster and more accurately from microarray data than from existing gene-based methods.

Despite an abundance of DNA-based microarray data—such as those from SNP studies—translating that information into a disease-diagnosis product is challenging. The biggest roadblock is the dearth of information on the precise roles that changes in individual gene expression play in a disease pathway. Also, differences between platforms or laboratories often preclude the direct comparison of results obtained in different DNA microarray studies.

A team co-led by Xianghong Jasmine Zhou, associate professor in molecular and computational biology at the **University of Southern California (USC)**, and Haiyan Huang, assistant professor of statistics at the **University of California, Berkeley**, set out to bypass these limitations by developing a database that could be used to evaluate the results of a patient's genetic test to provide a disease diagnosis. They expect that the results obtained from their database would be more accurate than any one single test and faster because the database compares information available from different platforms and different labs.

To achieve this, the group devised a method for transforming the vast amounts of data in the Gene Expression Omnibus (GEO) repository into an automated, searchable database. GEO is a public repository of genetic data at the **NIH's National Center for Biotechnology Information (NCBI)**.

Given GEO's size, the team first applied its method to a small portion of the data to generate a test database.

The researchers started by standardizing about 9,200 microarray datasets found in GEO to remove variation between data acquired on different platforms or in different labs. The standardization process required pairing each disease microarray dataset with at least one healthy control dataset.

The team then annotated the standardized datasets with 110 disease-related terms that described the microarray's phenotype. The terms included diseases grouped by site and biological system, such as gastrointestinal diseases; disease categories, such as inflammatory bowel disease (IBD); and specific indications, such as ulcerative colitis (UC).

Next, Zhou and her colleagues developed an automated program for searching the database. When a set of microarray results was used to query the database, the program searched for matching gene expression data and returned the corresponding set of disease terms that collectively provided a diagnosis.

The program used Bayesian methods that allow a model to be changed or modified as new data—such as additions to GEO—become available.

The researchers evaluated the capability of the test database using standardized microarray data based on DNA derived from Duchenne muscular dystrophy (DMD) patients as search parameters. The automated program returned a list of 10 disease terms that corresponded to the DMD phenotype. Additionally, using standardized DNA microarray data from patients with melanoma, the team also correctly diagnosed the cancer.

Finally, Zhou's team showed that the approach could provide useful information about specific diseases that were not represented in the small portion of GEO microarray data used to construct the database. For example, the test database did not include any HIV-related microarray data, but a search of it using microarray results from HIV patients returned the disease terms "RNA virus infection" and "genital disease."

This finding demonstrated that the approach could "identify important features of understudied or unknown diseases" for which public microarray data might be lacking, the team wrote in its report in the *Proceedings of the National Academy of Sciences*.¹

In addition to USC and Berkeley, the team included Chun-Chi Liu, now an assistant professor of bioinformatics at Taiwan's **National Chung Hsing University**.

The use of disease terms to represent biological functions and disease-related pathways produced "a very interesting and promising new method," said Trey Ideker, associate professor of medicine and bioengineering at the **University of California, San Diego**.

"The use of pathway information in disease diagnosis is the wave of the future," he said. "When your car breaks down, you don't try to diagnose the problem in a vacuum—you open the manual, which provides the wiring diagram of the car. The approach proposed by Jasmine Zhou tries to do this for disease diagnosis. We need more approaches like this one."

Zhou told *SciBX* her team's approach can be tailored to different purposes. "For instance, queries could be crafted to reveal the subtle distinctions between related phenotypes such as type 1 and type 2

"When your car breaks down, you don't try to diagnose the problem in a vacuum—you open the manual, which provides the wiring diagram of the car. The approach proposed by Jasmine Zhou tries to do this for disease diagnosis."

**—Trey Ideker,
University of California,
San Diego**

diabetes, metastatic and nonmetastatic cancer or two differentiation stages of embryonic stem cells,” she said. “Our diagnostic database can also be extended by adding phenotype classes beyond disease.”

Extended classes could include side-effect profiles induced by drugs, assuming a drug-treatment phenotype is reflected in altered gene expression patterns, said Haiyan Huang, first author on the *PNAS* paper and assistant professor of statistics at UC Berkeley.

RNA routes

Zhou said her team is applying to the NIH for funding to acquire the personnel and computing power needed to transform the entire GEO repository into a diagnostic database. “We also plan to build a diagnostic web server to benefit the broader scientific community,” she told *SciBX*.

In the longer term, the team wants to expand its diagnostic method to include transcriptome sequencing (RNA-seq) data, which would sidestep the current need for control datasets.

“Microarrays measure relative gene expression, while RNA-seq provides absolute expression,” Huang noted. “RNA-seq data are thus expected to be generally comparable and reproducible across labs and platforms” and would not require standardization before adding them to the database.

However, Huang said, more work is needed to identify the relative advantages of using RNA-seq data instead of DNA-based microarrays to diagnose disease.

“Whether the absolute expression levels found in RNA-seq data really reflect disease characteristics and how they could be used for disease comparison” is not fully understood, she said.

In contrast, Huang said, the team’s method for standardizing microarray data helps distinguish disease-specific genes from non-specific genes, thereby facilitating diagnoses.

Zhou said the findings reported in *PNAS* have not been patented. She declined to disclose whether they had been licensed.

Haas, M.J. *SciBX* 3(14); doi:10.1038/scibx.2010.426
Published online April 8, 2010

REFERENCES

1. Huang, H. *et al. Proc. Natl. Acad. Sci. USA*; published online March 22, 2010; doi:10.1073/pnas.0912043107
Contact: Xianghong Jasmine Zhou, University of Southern California, Los Angeles, Calif.
e-mail: xjzhou@usc.edu
Contact: Haiyan Huang, University of California, Berkeley, Calif.
e-mail: hhuang@stat.berkeley.edu

COMPANIES AND INSTITUTIONS MENTIONED

National Center for Biotechnology Information, Bethesda, Md.
National Chung Hsing University, Taichung, Taiwan
National Institutes of Health, Bethesda, Md.
University of California, Berkeley, Calif.
University of California, San Diego, La Jolla, Calif.
University of Southern California, Los Angeles, Calif.