Project Creates Repository for Microarray Datasets

As the number of microarray studies increases and more cancer researchers use gene expression technology to search for cancer biomarkers or possible therapeutic targets, those who want to compare datasets face a daunting task: obtaining and compiling all this information from researchers scattered around the world. A new online tool called Oncomine assembles these datasets in one easily searchable site to help researchers take advantage of the vast amount of data available.

Arul Chinnaiyan, M.D., who spearheaded the effort to create the new database, had realized there was a need for a central repository of microarray information when he was asked whether the genes that he found to be differentially expressed in prostate cancer, his field of study, were expressed in other cancer types.

“We wanted to compare data from different platforms and do meta-analyses,” but there were no tools to do this, said Chinnaiyan, associate professor of pathology at the University of Michigan in Ann Arbor. He developed Oncomine “so the average cancer biologist could take advantage of the wealth of information that’s out there,” he said.

With some pilot funds from his university, Chinnaiyan worked with colleagues at the Institute of Bioinformatics in India and Johns Hopkins University in Baltimore for a year to develop the site (http://www.oncomine.org). The site went online in late November 2003 with 65 datasets that included 4,702 microarrays and more than 47 million data points and covered 18 cancer types, and it has already gathered more than 1,000 users in 18 countries. Registration is free for academic and nonprofit users.

The information is amassed from publicly available datasets. Because there are no requirements that microarray datasets be stored in a central location, Chinnaiyan’s staff searches the literature for new studies and contacts these researchers for their data.

Users can search the results of a single microarray dataset, look for a gene’s activity across multiple datasets, or search for multiple genes across datasets. (An animated online tutorial provides instructions.) Chinnaiyan’s group has integrated data analysis with other resources, such as gene ontology annotations and a database of therapeutic targets. “One of the challenges of this has been to develop new data mining tools so people can take advantage of the data,” Chinnaiyan said.

Oncomine 2.0, currently in beta version but scheduled to be released in its final version within weeks, has more datasets (90, including more than 6,000 microarrays and nearly 71 million data points) and new features, such as pathway analysis.

Mark A. Rubin, M.D., associate professor of pathology at Harvard Medical School in Boston, an Oncomine user, called the site “a convenient resource” that “offers a way of [study comparison] that couldn’t be done in the past.”

There are numerous programs available to comb through data in any one laboratory, but to analyze data from across laboratories, a researcher must contact each laboratory separately or download the data from each laboratory’s Web site. It is not impossible, Rubin said, but it is inconvenient.

In addition, even after researchers receive all of the datasets, they often have problems comparing them because of the different platforms that the various groups use to collect their data. Oncomine normalizes the data. “We try to keep everything consistent within each dataset and then make qualitative comparisons across datasets,” said Chinnaiyan.

For example, if a researcher finds that gene X is upregulated in prostate cancer, he can then go into Oncomine and look at the status of gene X in studies of other cancers.

One of the values that Rubin has found in Oncomine is the ability to do virtual validation studies. In one of the first studies to use the site, published in June in Cancer Research, Rubin and his colleagues looked at the status of the gene TPD52 in prostate cancer. With Oncomine, they were able to confirm previous work that found that TPD52 was overexpressed in breast cancer as well as find that the gene was overexpressed in a number of other tumor types.

Oncomine can also act as a “discovery engine,” by generating discoveries or hypotheses that then need to be followed up with other research, Chinnaiyan said. In the first of these discoveries, published in the June 22 issue of the Proceedings of the National Academy of Sciences, Chinnaiyan and his colleagues reported finding 67 genes that were universally activated in most types of cancers. Another set of 69 genes was commonly activated only in aggressive undifferentiated cancers, the type that often results in poorer patient outcomes.

Chinnaiyan’s group plans to continue adding datasets to Oncomine in addition to developing new data mining tools to do gene correlation and anti-correlation studies. They also hope to be able to take a broader approach to the data in the future and look at the activation and repression of entire pathways within cancers.

—Sarah L. Zielinski