Bioinformaticians develop new data mining tools

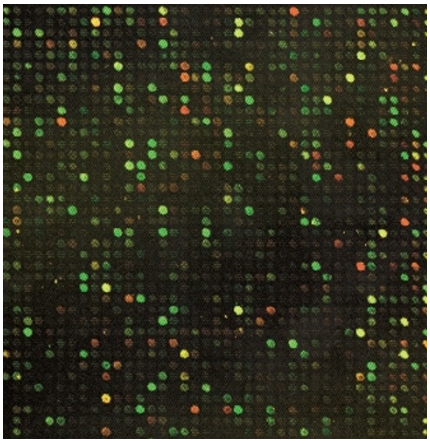
Driven by the massive amounts of data flowing from the sequencing of human and other genomes, scientists doing functional genomics research are developing new statistical and computational methods that will allow them to better mine this information. Two presentations showcased at the Intelligent Systems for Molecular Biology Conference (http://ismb2000.sdsc.edu) at the University of California at San Diego illustrate the first steps in this brave new world of bioinformatics.

Both programs are powerful applications that throw a wider net over documents on Medline or PubMed than a manual, generic search engine. One links clusters of expressed genes on microarrays to published literature, the second establishes a database of literature on each gene as well as related ones.

Daniel Masys, director of biomedical informatics and professor of medicine at the University of California at San Diego, has developed a microarray search engine that has now been made available to the public http://array.sdsc.edu. The web-based program, which is funded by the National Institutes of Health and is a joint project of UCSD, the San Diego Supercomputer Center and the Veteran's Medical Research Foundation, is free to users, much like the publicly-funded GenBank repository of gene sequencing data.

The web-based program allows users to browse, visualize, and analyze data obtained from gene expression experiments. Data are uploaded into a database that can be queried using various analytical tools. In addition, it uses indexing terms for published literature to link documents on a cluster of expressed genes.

One important drawback is that less than half of the genes available on commercial microarrays have one or more literature citations linked to them. The UCSD group tested their program using $120,000 of gene chips donated chips from Affymetrix that contained the expressed genes of cells responding to HIV



infection over a given time period.

By linking the clusters to published literature, one researcher involved in the project found a new insight into the origin of HIV-initiated cell death. Jacques Corbeil, a UCSD infectious disease specialist, discovered that the expressed genes all were involved in apoptosis within the cell, rather than death caused by signaling from the virus to the cell surface. "But this path is only a suggestion," says Corbeil, "...microarrays are the beginning of the process;" the proposed mechansim must now be assessed experimentally.

The second program, developed at the National Center for Biotechnology Information (NCBI), searches for functional relationships among genes in PubMed abstracts. Users must input reference numbers of the genes from a microarray experiment, for example. In response, they get a database of documents related to that gene, to other related genes and to a set of keywords that characterize the gene's function.

The program's goal is to establish common functions between different genes that would otherwise be missed because of the vast amount of published material available on PubMed, according to Hagit Shatkay, a postdoctoral researcher at NCBI.

"We can find hints to what these genes are doing and why they are co-expressed," Shatkay said.

However, Greg Colello, software programmer at the National Center for Genome Resources (NCGR) in Santa Fe, New Mexico, cautions that these search engines are no substitute for laboratory experimentation. "You still have to test things," he said. "Nobody is going to sit on a literature search as way of deciding what to do."

Eric Niiler, San Francisco

Can Italy catch up in genomics research?

After failing to make any notable investment in genomics research for the past five years, the Italian government is proposing a 'catch-up' plan. IL270 billion (US$135 million) is to be dedicated to genomics through a project called IPERGEN, which will fund 6 leading centers and 70 individual research groups focusing on bioinformatics, gene expression analysis, animal models, genotyping and proteomics. The government is also setting aside IL37.5 billion to create five new biotech start-up companies.

So far, the country's only national genomics project has been an effort to map a portion of the X chromosome, which was started in 1988 and ran until 1995, when it was abandoned because of lack of funding. Arturo Falaschi, president of the International Center of Genetics, Engineering and Biotechnology in Trieste, complains that information gleaned from this national project and written up as a report in 1997 has not been taken into consideration by the government in planning the IPERGEN project. "This document never obtained proper government attention even though it focused on areas—including local population genetics and gene expression of mouse mutants—where Italy really can be internationally competitive," he says. Other critics of the IPERGEN project cite the lack of a precise definition of targets tailored to the country's genomic strengths as a major flaw. They say it will be a poor duplication of research efforts already advanced in other countries.

Luca Cavalli-Sforza, a human population geneticist at Stanford University, warns that the coordinating committee could also undermine the project. "Most names will be proposed by politicians, except for one from a list nominated by the US NIH, and one by EMBO," he says, adding, "very few politicians have shown that they can pick competent scientists, or, even able administrators of science."

Another issue is the government's apparently misguided optimism that five biotech companies can be created so simply. "I doubt that the Italian government will become such an efficient developer of small private companies starting from scratch, in a culture where venture capitalism does not exist, and where banks lend money only to the truly rich who are reliable debtors," says Cavalli-Sforza.

Martina Ballmaier, Milan