Search tool for gene expression databases could uncover therapeutic targets, biological processes

October 1 2013

A new computational tool developed by U.S. and Israeli scientists will help scientists exploit the massive databases of gene expression experimental results that have been created over the past decade. Researchers say it could uncover new links between diseases and treatments and provide new insights into biological processes.

The team, headed by Ziv Bar-Joseph of Carnegie Mellon University, reports in the October issue of the journal *Nature Methods* that the tool, called ExpressionBlast, enables searches based directly on experimental values, rather than keywords.

The researchers already have used ExpressionBlast to uncover intriguing clues about SIRT6, the first gene shown to extend lifespan in mice and thus a potentially important drug target. By mining experimental data stored in a public repository called the Gene Expression Omnibus (GEO) maintained by the National Center for Biotechnology Information, they found that SIRT6 may be involved with functions that include immune response, metabolism and the regulation of gender-specific genes.

"Because so little is known about SIRT6, it would be difficult to search the hundreds of thousands of GEO datasets using keywords and, without other guidance, it would be practically impossible to find other experiments with gene expression patterns similar to SIRT6," said Bar-Joseph, an associate professor of computational biology and machine

learning. "ExpressionBlast enabled us to take SIRT6 gene expression data from just two mouse experiments and find other experimental data in GEO with similar expression patterns."

The tool is available online, http://www.expression.cs.cmu.edu/. The search engine enables researchers to search for expression patterns that are similar or opposite to their own results and can search within and across species. Guy Zinman, Shoshana Naiman, Yariv Kanfi and Haim Cohen of Bar-Ilan University worked with Bar-Joseph to develop ExpressionBlast and are co-authors of the journal report. Their intention was to develop a tool for gene expression queries that would be the equivalent of Blast, a two-decade-old tool for searching gene sequence databases that remains one of the most widely used tools in bioinformatics.

Genes encode the information necessary for life, while gene expression is the process by which that genetic information is transformed into proteins and by which genes are regulated. Understanding gene expression thus is critical for understanding biological and disease processes. This information is so important that, for the past decade or so, most leading journals have required researchers who publish papers on gene expression to submit their experimental data to public repositories such as GEO.

GEO alone holds data from more than 1 million microarrays. Each of these microarrays might contain up to 40,000 numerical values – which indicate which genes are over- or underexpressed, and by how much. GEO and the European Bioinformatics Institute's ArrayExpress thus represent a treasure trove of potential discoveries. But existing searches are often dependent on keyword summaries submitted by each researcher, or require manual comparisons of microarrays.

ExpressionBlast uses novel, automated and scalable text analysis

algorithms to transform the unstructured data in GEO so that it can be systematically searched. The researchers have thus far processed tens of thousands of expression series representing hundreds of thousands of individual arrays across several species. Once processed in this way, the data can be accessed easily via a graphical interface. This work was supported by a grant from the National Institutes of Health and a National Science Foundation Innovation Corps (I-Corps) award.

Provided by Carnegie Mellon University