**Data Mining Gene Expression Databases –**

**Software for Routine Clinical Use**

Michael Fetick

Coleman University

Abstract

Healthcare can be greatly advanced with improved access to the Whole-Genome Sequencing (WGS) of a patient. The health of each person is based on factors of environment, events, and hereditary traits. To improve healthcare for any person, the diagnosis and treatment should be explicitly tailored to that person. “Clinicians can identify increases in disease risk for specific patients, as well as their responsiveness to certain drugs, by determining the sequence of the billions of building blocks, called **nucleotides** that make up their DNA.” Scientist have studied the DNA of animals and humans over decades and human DNA exists in a huge network of genes called the human genome. Whole-genome sequencing is the complete identification of a person’s DNA and it has been touted as a game-changer in personalized medicine.” However, significant challenges must be overcome before whole-genome sequencing can be routinely clinically useful. Recent developments in remarkable, computer software applications may advance the access and understanding of using whole-genome sequencing for all levels and needs of healthcare.

**Data Mining Gene Expression Databases –**

**Software for Routine Clinical Use**

Over the past decade, the results of thousands of biological experiments are documented in mathematical gene expressions, and stored in massive databases. Laboratory studies of these genes expressions have resulted many common occurrences that are studied to analyze the similarities and differences. The goal is to better diagnose and specifically treat any disease for any person.

Background

Manage all the Data

The complexity of so much data with so many variables, requires computer processing to manage the storage of the data, to categorize and organize the data, to present all the variations of the studies, and to compare whole-genome sequencing at the molecular level.

How Much Data to Manage? Computer software operates with ‘written’ program code that executes software commands to process (manage) data. The complexity and size of a computer software program is measured and described by the count of the number of ‘lines of code written’ to perform the processing. This measurement does not include any reference to the amount of data that is processed. The following analogy illustrates the comparison of several software projects, including the mapping of a human genome:

* the average iPhone application 10,000 lines of code
* Space Shuttle 400,000 lines of code
* Bacteria (Syphilis) 1,000,000 lines of code
* Microsoft Office 2013 45,000,000 lines of code
* a Mouse (DNA base-pairs in genome) 120,000,000 lines of code
* Healthcare.gov website 500,000,000 lines of code
* a Human Genome 3,300,000,000,000 lines of code

Recent Developments

New Computational Tool, ExpressionBlast

“A new computational tool developed by US 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 tool, ExpressionBlast is available and being used extensively, <http://www.expression.cs.cmu.edu/index.html>.

References

Bar-Joseph, Ziv. (13 Oct. 2013) Gene Expression Databases Could Uncover Therapeutic Targets, Biological Processes. *ScienceDaily / Carnegie Mellon University.* Retrieved from <http://www.sciencedaily.com/releases/2013/10/131001192153.htm>.

Scientist have studied the DNA of animals and humans over decades and human DNA exists in a huge network of genes called the human genome. Whole-genome sequencing is the complete identification of a person’s DNA. Over the past decade, the results of thousands of biological experiments are documented in mathematical gene expressions, and stored in massive databases. A new computational tool “**ExpressionBlast**” exploits these massive databases with algorithms to help scientist find new links between diseases and treatments.

Bar-Joseph, Ziv. (30 Sept. 2013) Press Release: New Search Tool for Large Gene Expression Databases Could Uncover Therapeutic Targets, Biological Processes. *Carnegie Mellon News - Carnegie Mellon University.* Retrieved from <http://www.cmu.edu/news/stories/archives/2013/september/sept30\_expressionblast.html>.

The team, headed by Ziv Bar-Joseph of Carnegie Mellon University, reports that the tool, called ExpressionBlast, enables searches based directly on experimental values, rather than keywords. The most exciting example of success assists in extending the lifespan in mice. Scientists already have used ExpressionBlast to uncover 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."

Zinman, Guy E., Shoshana Naiman, Yariv Kanfi, Haim Cohen, and Ziv Bar-Joseph. (27 Sept. 2013) ExpressionBlast: Mining Large, Unstructured Expression Databases. *Nature Methods*, (10, 925–926 (2013) doi:10.1038/ nmeth.2630): Nature Publishing Group.

ExpressionBlast is a search engine for gene expression data. It allows you to compare your own expression results to over thousands of studies and close to a million samples currently available at GEO, and find other studies (even across species) that have similar (or opposite) expression patterns to your results.  The amount of gene expression data deposited in public repositories has grown exponentially over the last decade. Specifically, Gene Expression Omnibus (GEO) is one of the largest expression-data repositories. It contains hundreds of thousands of microarray and RNA-seq experiment results grouped into tens of thousands of variations.

Snyder, Michael, PhD, and Thomas Quertermous, MD. (11 Mar. 2014) Whole-genome Sequencing for Clinical Use Faces Many Challenges, Study Finds. *ScienceDaily / Stanford University Medical Center.* Retrieved from <http://www.sciencedaily.com/releases/2014/03/140311163034.htm>.

Whole-genome sequencing (WGS) has been touted as a game-changer in personalized medicine. Clinicians can identify increases in disease risk for specific patients, as well as their responsiveness to certain drugs, by determining the sequence of the billions of building blocks, called nucleotides that make up their DNA. However, significant challenges must be overcome before whole-genome sequencing can be routinely clinically useful. Research is labor-intensive because there is no standard way to assess the potential health impact of each change to the genes. It remains significantly harder to use whole-genome sequencing for disease prediction rather than for disease diagnosis. The authors are confident that the field of whole-genome sequencing is worth pursuing. Sequencing technology is evolving, and the National Institutes of Health-sponsored Clinical Genome Resource, or ClinGen, is meant to help speed the identification of clinically important variants and reduce the burden of manual investigation.

Frederick E. Dewey, MD, Megan E. Grove, Cuiping Pan, MS, Benjamin A. Goldstein, PhD, Jonathan A. Bernstein, MD, PhD, Hassan Chaib, PhD, Jason D. Merker, MD, PhD, Rachel L. Goldfeder, BS, Gregory M. Enns, MB, ChB, Sean P. David, MD, DPhil, Neda Pakdaman, MD, Kelly E. Ormond, MS, Colleen Caleshu, MS, Kerry Kingham, MS, Teri E. Klein, PhD, Michelle Whirl-Carrillo, PhD, Kenneth Sakamoto, MD, Matthew T. Wheeler, MD, PhD, Atul J. Butte, MD, PhD, James M. Ford, MD, PhD, Linda Boxer, MD, John P. A. Ioannidis, MD, PhD, Alan C. Yeung, MD, Russ B. Altman, MD, PhD, Themistocles L. Assimes, MD, PhD, Michael Snyder, PhD, Euan A. Ashley, MRCP, DPhil, Thomas Quertermous, MD. (12 Mar. 2014) Clinical Interpretation and Implications of Whole-Genome Sequencing. *JAMA.* (2014;311(10):1035-1045. Doi:10.1001/ jama.2014.1717). Retrieved from <"http://jama.jamanetwork.com/article.aspx?articleid=1840236>.

This was an exploratory study of 12 volunteer adults and it involved a lot of scientists. The use of whole-genome sequencing (WGS) was associated with incomplete understanding of some inherited disease genes. The study’s results were not good; low reproducibility of detecting genetic variations while using the highest potential clinical effects resulted uncertainty about clinically reportable findings. In certain cases, WGS will identify clinically actionable genetic variants warranting early medical intervention. These issues should be considered when determining the role of WGS in clinical medicine. As our civilization gets better at expressing, sharing, and understanding the genes of our bodies, then scientist may find new links between diseases and treatments.

David McCandless, Concept & Design, Pearl Doughty-White, Miriam Quick, Research. (v 0.71// Oct 30, 2013) Codebases, Millions of Lines of Code. *Knowledge is Beautiful.* Retrieved from <http://www.informationisbeautiful.net/visualizations/million-lines-of-code/>