Journal of Glycomics & Lipidomics

Extended Abstract

Current and up-coming analysis tools for the NCI-60 in CellMiner: Resources for data integration and systems pharmacology

William C. Reinhold

Abstract

The CellMiner web-application; a suite of tools that provides rapid access to multiple forms of molecular and pharmacological data available for the NCI-60. These tools also aid in the integration of this high-throughput data in a manner doable by the non-bioinformaticist. The integration tools are accessed in the "NCI-60 Analysis Tools" section. These include the "Cell line signature" tool, which provides i) transcript levels, ii) compound activities, and iii) microRNA levels for the NCI-60. Access to the exome sequencing data will be demonstrated, including use of the "Graphical output for DNA:Exome sequencing". The individual forms of data may in turn be integrated using the "Cross-correlation" tool for up to 150 genes (transcript levels), microRNA levels and drug activities. The data may also be integrated using "Pattern comparison", which allows the input of any pattern (for the NCI-60), and from that pattern correlates all gene transcript and microRNA levels, and compound activities. Several indevelopment tools are also demonstrated. The first of these are "aCGH copy number" for the determination of DNA copy number variations. Second is "Genetic variant summation", which provides a summation of mutations in a pathway for up to 150 genes. Third is "Genetic variant versus drug visualization" which provides rapid visualization of gene variant versus drug relationships. Each of these tools permits the user to search for potential relationships in a manner specific to their area of expertise and interest. Expertise in computer science or bioinformatics is not required. The data is available at no cost to the scientific public.

The NCI-60 cancer cell line panel provides a premier model for data integration and systems pharmacology being the largest publicly available database of anticancer drug activity,, genomic, molecular, and phenotypic data. It comprises gene expression (25,722 transcripts), microRNAs (360 miRNAs), whole genome DNA copy number (23,413 genes), whole exome sequencing (variants for 16,568 genes), protein levels (94 genes), and cytotoxic activity (20,861 compounds). Included are 158 Food and Drug Administration (FDA)-approved drugs and 79 that are in clinical trials. To improve data accessibility to bioinformaticists and non-bioinformaticists alike, we have developed the CellMiner web-based tools. Here we describe the newest CellMiner version, including integration of novel databases and tools associated with whole exome sequencing and protein expression, and review the tools. Included are i) "Cell line signature" for DNA, RNA, protein and drugs, ii) "Cross correlations" for up to 150 input genes, microRNAs, and compounds in a single query, iii) "Pattern comparison" to identify connections among drugs, gene expression, genomic variants, microRNA and protein expressions, iv) "Genetic variation versus drug visualization" to identify potential new drug:gene DNA variant relationships, and v) "Genetic variant summation", designed to provide a synopsis of mutational burden on any pathway or gene group for up to 150 genes. Together, these tools allow users to flexibly query the NCI-60 data for potential relationships between genomic, molecular and pharmacological parameters in a manner specific to the user's area of expertise. Examples for both gain- (RAS) and loss- (PTEN) of-function alterations are provided.

This review provides a synopsis of both the use and novel features of the CellMiner web application. CellMiner is designed specifically for the purpose of facilitating integration of pharmacological with molecular data from the NCI-60 cell lines. Its provision of "Cell line signatures" for both drug activity and multiple forms of molecular data, in which many of the preprocessing steps have already been done, allow a broad segment of the scientific community to make rapid and meaningful explorations into pharmacological-molecular relationships. The cell line models will always form the basis for studies of this type, due to their obvious advantages in providing testable models. Observations and hypotheses with translational importance made with these models will increasingly form the intellectual basis for a more specific and logical application of treatments, based on a patient's disease's specific molecular characteristics.

These two tools give access to the specific data in the absence of the additional quality control assumptions applied within the "NCI-60 Analysis Tools" section. The data in this form may be preferential dependent on the question being asked, and allows users flexibility to apply their own judgment and assumptions. "Query Genomic Data" functions as the unfiltered data query tool for the molecular data sets (Fig. 1A). In Step 1, users select the type of queries that best fit their needs. The query options include i) gene name, ii) RefSeq (mRNA or protein), iii) Entrez identifier, iv) chromosome number, v) chromosome location, vi) cytoband, or vii) four types of platform specific identifier. In Step 2, users input these identifiers either as a list or as an uploaded text (.txt) or Excel (.xls) file. In Step 3, users select from among 17 datasets that provide various types of information at the DNA, RNA,

or protein level described previously (2, 3, 8–13). In Step 4, users provide their E-mail address, and click "Get data". There are currently 25,722 transcripts, including genes, pseudogenes, and open reading frames with data in this format.

"Query Drug Data" functions as the unfiltered data query tool for the compound activity data set (Fig. 1B). In Step 1, users may select the type of query: i) NSC, ii) compound name, iii) molecular formula-exact match, iv) molecular formula-element match, v) molecular weight range, or vi) mechanism of action (introduced here).

The "Molecular formula-element match" allows the used to search based on specific elements in the compound, such as Zn or Se. In Step 2, the user inputs these identifiers either as a list or as an uploaded text (.txt) or Excel (.xls) file. In Step 3, users provide their E-mail address, and click "Get data". There are currently 52,269 compounds with activity data in this format.

.

This work is partly presented at 2nd International Conference on Big Data Analysis and Data Mining 30-December 01, 2015 San Antonio, USA

William C. Reinhold
National Cancer Institute, USA, E-mail: wcr@mail.nih.gov