Potentially Valuable Metadata Tool for Medical Diagnoses

**Introduction**

If one keeps track of what were being discussed among physicians, I have been under the impression that doctors have been saying basically this: “The diseases and disorders may have patterns of synonyms, but they are so biologically individualism that medically diagnosing the disease/disorder isn’t boring.”

The new medical journal, “Cold Spring Harbor Molecular Case Studies,” with its first issue released to public on October 1st, 2015 stated that their mission is to strengthen the “precision medicine” approaches. Instead of the hit-or-miss innovation discoveries, this journal wants to focus on evaluating the data collecting and data analyzing technologies and methods.

Historically, since about the 1950s (e.g. test strips for diabetes), medical scientists were aware that evaluating a precision diagnose cannot simply be done by looking at RNA and/or protein expressions. Measuring the expression of RNA and proteins have been the foundations of tracking possible biological function regulations. But we also have essential biological molecules with little to no RNA or protein features, catalogizing them as “metabolites.” For example, lipids, fatty acids, bile acids, amines, cationic, and polar molecules. Processing the expression levels of most metabolites, however, is quite complicated and not feasible. Worse, the biological specimens’ data are prone to environmental impressions, both externally and internally. When handling possibly confusing data, it could be helpful to think, “That is biology.”

The "that is biology" attitude, however, could progressively become less "acceptable" with the growing availably of metadata sharing networks. The network specified for sharing measurements of metabolite expression is at the <https://metabolomicsworkbench.org/> site. I didn’t read enough metabolites-type studies, but it seems like the STEM field is still developing protocols and approaches for the metabolites-type datasets.

The metabolite dataset, containing metabolomic information from 30 knockout mice, was selected for this final project. The "data descriptor" journal article for this 30 knockout mice dataset will be referenced as the metabolomic workbench ID number, ST00114.

Let it be noted that the ST00114 study has multiple goals, the main one being an advocator for benchmarking processes for plugging in a lab's data acquisition parameters in order to identify this lab's metabolites of interest. The ST00114 study used the ChemRICH process in their R programmed files, which are available on the ST001144 author's GitHub, https://github.com/barupal/ChemRICH. The care provider may be please to that some development is made in finding

ways to adjust the parameters for different processing methods could solve the problems of sharing metabolomic data between thousands of labs around the world. But as much as I talk about that, I realized that learning how to apply the chemical compound parameters are beyond my current R programming skills.

For my small final project, I am going to try find any gender-dependent phenotype differences.

Methods

I think the ST00114's Kaggle moderate grade of 7 out of 10 was because the datasets were not user-friendly. The headers could have been labeled better (e.g. are those mice ID numbers or chemical ID numbers?). And the series of data could be organized better (e.g. the tables with mice ID numbers for six animals from 30 knockout lines' were columned in no particular order.) It did take me a while to figure out what I was looking at and find information about seemingly so many columns of ID numbers.

The Metabolomics Workbench's [Statistics Toolbox for Study: ST00114](https://metabolomicsworkbench.org/data/stats_toolbox.php?STUDY_ID=ST001154&ANALYSIS_ID=AN001941) appears to currently be an artificial intelligent machine setup. That process probably will improve later, but for now, at best, this "toolbox for study" could help with preliminary exploration of the chosen dataset. Some results were jamming data from all 30 knockout groups together as one group. Other results look at the knockout lines, but generally ignored specific factors such as genders. The ST00114's raw, processed, and exported datasets are available as supplement tables and Kaggle files for anyone to download copies of.

The Metabolomics Workbench's Statistic Toolbox for Study provides links to external sites such as GitHub folders.

Results

Conclusion

In the days before the data acquitting and processing technologies became sensitive enough to detain differences of multiple molecule types and expressions of the individualism samples, the testing for metabolomic expressions had have been limiting.

The ST00114's article closed with this statement: "We foresee this dataset’s use in developing next generation bioinformatics as well as in teaching courses for metabolomics and as a test case for benchmarking software."

My understanding of the "benchmarking software," which is related to as ChemRICH, is that with today's technology and resources, the various data acquitting processions and devices from thousands of labs around the world would not have exact matching measurable chemical parameters. Thus, tools such as the ChemRICH as benchmarking protocols could help scientists compare data from thousands of labs.