Editorial

Open-source Tool for Data Analysis in Medicine—Are We Ready?

The vital role played by the statistical analysis in the field of healthcare and pharmaceutical research is well known.

Whether it is the randomization of clinical trials, sample size estimates, statistical testing or modeling the data, every step of medical research extensively uses statistical algorithms. While the professional statisticians can decide to use a correct algorithm for the problem in hand, its implementation is always realized through established commercial tools like statistical package for social studies (SPSS) and Microsoft Excel.

This happy going is being disturbed by the recent advances in medical research mostly contributed by the 'omics' studies on the human genome, namely genomics, transcriptomics, proteomics, and metabolomics. These new fields, generally termed as "computational biology," generate an enormous amount of data at the level of gene and protein sequences of individual patients that fuel the study of biomarkers of diseases, genetic disorders and disease networks in the body. Thus, the ultimate era of "personalized medicine" has just begun!

Will the commercial tools currently used be able to meet the enormity and the complexity of data from these new fields? It is unlikely. Since a proprietary tool is always created for an existing market, it may not be able to meet the rapidly changing requirements of new areas of research. Thus the medical researchers need a tool that can handle these ever-expanding discoveries while retaining all the earlier statistical methods and features already being used.

Such a tool exists, and it is existing for a long time! It is called "R statistical package," a freely distributed statistical analysis tool. R is not just a tool; it is an environment for statistical computation, data analysis, and graphics. It is made available as free software under the Free Software Foundation's General Public License (GNU) and can run on Windows, Linux and Mac operating systems. R was initially created in 1995 by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. The "R Foundation" controls the design and development of the tool, with contributions from academicians in various countries.

What does R contain? Well, it contains almost everything a data analyst needs: facility to read tabled data, algorithms for statistical tests, regression, modeling, and data analytics methods like clustering, classification, Principal Components Analysis, etc. More than one (many many) flavors of algorithms for each task are available. A bunch of well-developed graphics packages using which we can plot almost anything we want! (Not impressed?. Google with the term "R + plots" to see Google images of R plots created by various researchers). Those who are coding freaks can do programming in R as they do with any other language like C, C++, and Python.

As such, the statistical algorithms of R are not created with life sciences data in mind. It always exists as a generalized statistical package. However, in 2001, a new open-source, open development software project called "Bioconductor" was initiated for the analysis of genomic data generated by wet-lab experiments. It is primarily based on the R programming language and utilizes all the algorithms and programming constructs of R. Its development and maintenance is overseen by a "Bioconductor core team" based primarily at the Fred Hutchinson Cancer Research Center, USA.

Bioconductor has more than 1500 libraries that can be run under R package. They cover almost all aspects of data analysis in genomics, including data from microarrays, DNA/RNA sequencing, metagenomics, genetic analysis like Linkage Disequilibrium Studies, Genomewide Association Studies (GWAS), etc. Once R package is installed, any of these algorithms can be downloaded from the Bioconductor website and can be used as an R extension. The use of Bioconductor as an R extension grants the medical researcher the power of programming with very minimal knowledge and experience of writing codes!

New analysis methods and libraries are added to R and Bioconductor almost every day. Each one of these libraries is published and thrown open to the academic community for scrutiny. Not only the formulas of the algorithms, but even the source codes are also available to the users! Through well-established user forums, bugs are reported regularly, and bug fixing is done by the next (once in 6 months) release.

For those researchers from life sciences who use commercial tools, the programmatic R syntax can be difficult to comprehend in the beginning. This can be overcome by a few hours of a formal training course on R. This is a

small price to pay for the vast world that R and Bioconductor can open to us. It is the future tool for open-ended research in the field of biology, medicine and genetics. If you are a young researcher or a project student, do not postpone the pleasure of learning R!

Dr. R Srivatsan
Faculty Scientist
Institute of Bioinformatics and Applied Biotechnology
Biotech Park, Electronics City
Bengaluru, Karnataka, India
Tel: 9901949917, e-mail: Srivat_in@yahoo.com