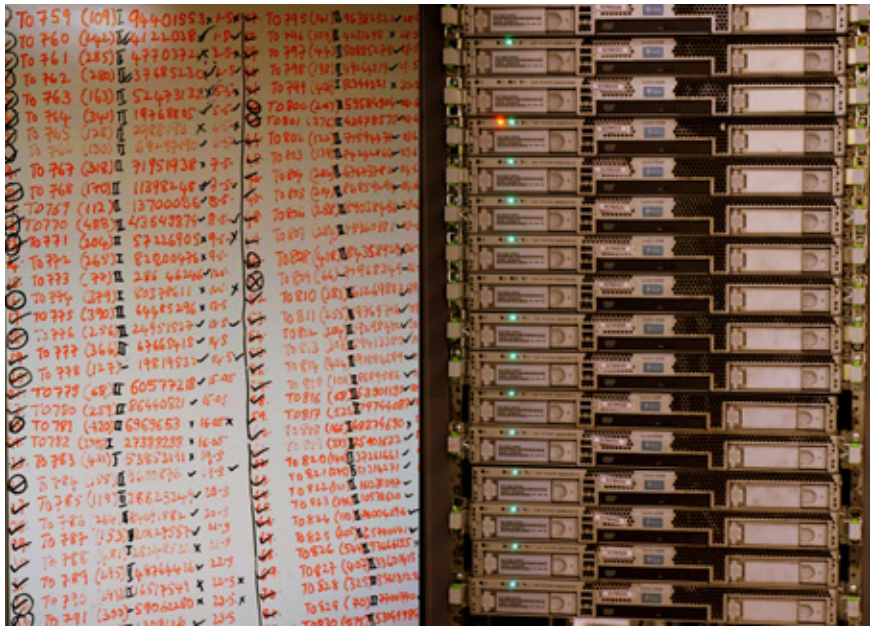
doi: 10.1038/nindia.2015.117

# Analysing large datasets in the omics era

Sarath Chandra Janga[1,2] & Debasis Dash[3]

*Sarath Chandra Janga*

*Debasis Dash*



Advances in high-throughput technologies in a number of omics-related fields in the post-genomic era have revolutionised the approach to understanding biomedicine. Reductionism, which has been the standard paradigm in biological research for more than a century, has armed researchers with immense knowledge of individual cellular components, their functions and mechanisms.

Despite its huge success over the years, post-omics biology has increasingly made it clear that discrete biological functions can only rarely be attributed to an individual molecule. Instead, most biological outcomes in a cell arise from a complex interplay of different cellular entities such as proteins, DNA, RNA and metabolites. This has brought forth the notion of using multi-dimensional data-driven approaches in a number of biomedical settings. Genomics and proteomics have been particularly influenced by an avalanche of datasets originating from a number of laboratories and large-scale consortium funded projects. For instance, the rate of growth of Genbank has been exponential, doubling every 18 months[1] with specific genomic surveys such as the Global Ocean Sampling (GOS) expedition alone contributing to more than 6 million proteins[2]. In fact, the more recent next-generation sequencing technologies show a declining trend in the cost of sequencing as prices go down by half every 5 months[3]. These massive increases in omics data, often referred to as big data, naturally bring in a new set of challenges for the scientific community.

While these big datasets hold great promise for discovering patterns despite heterogeneities in the data, their massive sample sizes and high dimensionality introduce unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation as well as measurement errors[4]. Broadly, any analyses of large-scale omics datasets can be divided into three major steps – data acquisition and pre-processing, data analysis and interpretation or visualisation.

## Challenges in big data analysis

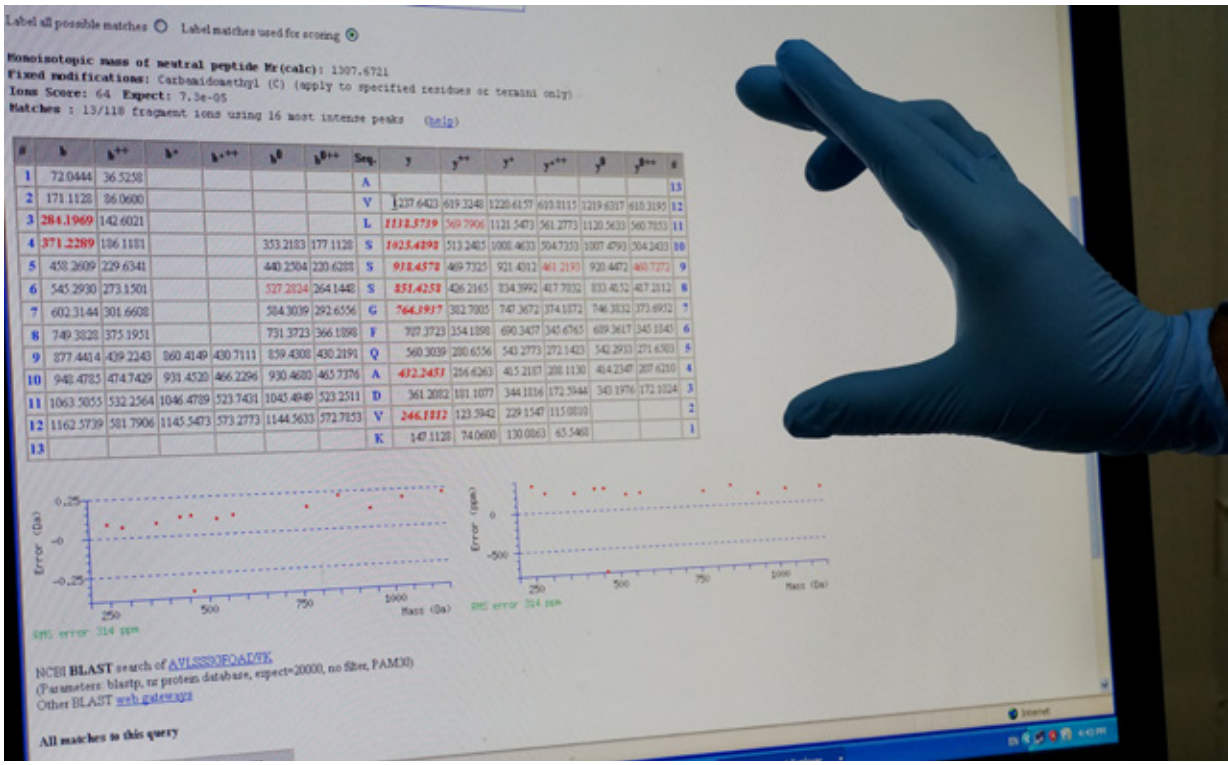
One of the most crucial challenges in analysing big data is acquisition and pre-processing. Some data sources, such as mass spectrometers and DNA sequencing facilities can produce staggering amounts of raw data. Much of this data is of no interest, and can be filtered and compressed by many orders of magnitude.

For instance, quality scores of reads from next generation sequencing data may not be of much use to most downstream analytical pipelines once high quality reads are identified. Likewise, spectra once mapped to peptides and their abundance estimated may not be of much use for downstream analysis. So a natural challenge is to define filters in the pipelines in such a way that they do not discard useful information.

A related issue is to automatically generate the right metadata to describe what data should be measured and stored. This metadata may be crucial to downstream analysis. For example, we may need

[1]Department of Biohealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Indianapolis, Indiana, USA (scjanga@iupui.edu). [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Health Information and Translational Sciences (HITS), Indianapolis, Indiana, USA. [3]Institute of Genomics and Integrative Biology, New Delhi (ddash@igib.res.in).

to know the software and corresponding parameters used to generate a particular file format, which encodes the compressed information, so that the end user knows how the data is organised and stored based on the raw data. Frequently, the information collected is not in a format ready for analysis. Therefore, these two steps involve an information extraction process that pulls out required information from the underlying sources and expresses it in a structured form suitable for analysis.




Data analysis is also becoming considerably more challenging not only because of the volume of data but also due to the heterogeneity of the embedded datatypes which need to be integrated into computing frameworks. For instance, when computing on large datasets from diverse sources, data analytic frameworks need to consider the available computing infrastructure, scalability of the algorithmic implementation and level of automation that is desirable and possible for the problem being addressed. The last of these factors requires differences in data structure to be expressed in formats that can be automatically resolved for building efficient workflows for high-throughput data processing.

Mining data also assumes that the data is clean and in a format ready for analysis and that there are algorithms available to process the data on computing clusters. Both of these assumptions may not always be true. For instance, most current algorithms in omics cannot be readily deployed in Hadoop clusters and hence new code needs to be written to make them usable in cloud environments which can significantly speed up running times as compared to traditional multi-node computing clusters where there is no interaction between the compute nodes. Also since data stored in Hadoop clusters is typically replicated, computing resources and infrastructure have to be taken into consideration.

Likewise, if a noSQL framework is used as the underlying database, data needs to be imported to the data server to facilitate such efforts. These challenges also provide unique opportunities for exciting inter-disciplinary collaborations with experts from biomedical data science and engineering.

The most important step from an end user's perspective is data interpretation. Unless the results of an analysis and its process are well documented and visualised, the analysis is of little value to the user. So it is essential to document the various steps of the implemented framework along with user-friendly visualisations which can enhance usability of the software. Providing a workflow would allow users to not only vary the choice of the parameters to study the impact on their results but also help understand the causes of noise in the dataset. Such an effort from the developer can also help in iteratively improving the software in the long run, especially if a feedback system or a listserve is maintained. Such level of information would also enable the user to realise the potential and utility of the processed data in making interpretations or in using it to integrate with other in-house datasets.

Scientific research has been revolutionised by big data. Several resources such as Gene Expression Omnibus (GEO) from National Center for Biotechnology Information (NCBI) and the PRoteomics IDEntifications (PRIDE) database from European Bioinformatics Institute (EBI) have become the central resource for omics researchers. Omics fields are being transformed from one where investigators measured individual genes or proteins of interest to one where the levels of all genes or proteins across a number of conditions and contexts are already in a database and the investigator's task is to mine for interesting genes and phenomena. In most omics settings, there is a well-established tradition of depositing scientific data into a public repository to create public databases that can be used by all.

Data sharing in proteomics (e. g. PRIDE, Peptide Atlas, Massive and ProteomeXchange consortium) have led to free availability of data in the public domain. This has enabled researchers to develop new algorithms, reannotate and reinterpret, thereby providing deeper insight. Many Indian laboratories have contributed to and benefitted from this global sharing of high quality proteomics data and added value to the field through their participation.

## References


1. Benson, D. *et al.* "GenBank". *Nucleic Acids Res.* **36**, D25–D30 (2008).
2. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Expanding the universe of protein families. *PLoS Biol.* **5**(3):e16 (2007).
3. Stein, L. D.The case for cloud computing in genome informatics. *Genome Biol.* **11**, 207 (2010).
4. Labrinidis, A. & Jagadish, H. V. Challenges and opportunities with big data *Proc. VLDB Endowment.* **5**, 2032-2033 (2012).