Global Journal of Technology & Optimization

Sarkar, Global J Technol Optim 2016, 7:$ {issueNumber}
DOI: 10.4172/2229-8711.S1e103

Editorial Open Access

The Big Data Deluge in Biology: Challenges and Solutions

Sarkar RR[1,2]

[1]*Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune 411008, Maharashtra, India*

[2]*Academy of Scientific and Innovative Research (AcSIR), New Delhi, 110 001, India*

Corresponding author: Sarkar RR, Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune 411008, Maharashtra, India, Tel: +91 2590 3040; Fax: +91202590 2621; E-mail: rr.sarkar@ncl.res.in

Received date: June 24, 2016; **Accepted date:** June 25, 2016; **Published date:** Jun 27, 2016

Copyright: © 2016 Sarkar RR. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Editorial

Over the years, continuous efforts in understanding the complex multi-step processes at different levels have transformed biology from a qualitative to a more quantitative subject, resulting an enormous generation of information at diverse scales from molecular/genome to ecological as well as epidemiological/clinical. Biological data is highly overrepresented with respect to its quantity, diversity and analysis. Advances in high throughput experimental techniques have expanded the lengths and scales of biology by providing a large amount of diverse biological data with relatively less input costs. Its heterogenous and complex nature poses a great challenge for both scientists and technicians in high quality data generation, its management, accessibility, handling and integration.

The term 'big data' not only talks about the quantity/volume of data generated but also the rate of increment, processing, diversity as well as reliability [1]. Biological big data ranges from laboratory scale 'omics' type experiments to geographical distribution of human populations around the world. The advent of cost-effective high throughput techniques like next generation sequencing for rapid genomic and RNA sequencing, mass spectrometry of identifying proteomes and metabolomes, microscopy-based image generation of cells, microarray and RNA sequencing for mRNA expression, Chip-Seq for binding site identification, yeast two-hybrid assay for protein-protein interactions, X-ray crystallography and NMR for protein structures [2] has benefitted biologists to generate a diverse set of information even with small scale lab setups although the size of data generated is huge.

If we only consider genomics, even a single newly sequenced human genome is around 140 gigabytes (GBs) in size. After the Human Genome Project, which showcased the first generation of big data in biology, a large number of collaborative projects globally have further added to the quantity of biological big data. ENCODE, HapMap and 1000 Genomes project are some of the important projects that have revolutionized the generation of a catalogue of information, largely using standardized protocols, reagents and analysis schemes. The sequence read archive (SRA) of the National Centre of Biotechnology Information (NCBI) now hosts around 3.6 petabases of such raw sequence data. It is predicted that by 2025, around 2-40 exabytes (1 exabyte = 109 gigabytes) of only human genomic data accounting for 1 zettabases per year, would be generated [3]. Similarly, if we consider only the gene expression data, EBI-ArrayExpress database alone hosts around 42.67 terabytes (TB) of archived data accounting for 65849 transcriptional profiling experiments. **Table 1** gives an idea about the volume of data stored within a few popular databases. This further raises the concern for storage and management of large-scale datasets efficiently.

Data type	Database	Datasets	References
Gene expression	ArrayExpress	65849 experiments, 2016701 assays	[4]
Protein expression	PaxDB 4.0	419 datasets, >300000 proteins covered	[5]
Protein 3D structure	RCSB-PDB	114643 released structures	[6]
Nucleotide sequence	GenBank	189232925 sequences, 203939111071 bases	[7]

Table 1: Number of datasets stored in few popular databases.

To give a flavour of the increasing focus towards generation and analysis of diverse biological datasets, year-wise data (between the years 1970 to 2015) from few popular databases hosting diverse biological datasets is compared **(Figures 1 and 2)**. Historically, the protein-protein interaction and protein structure data were first made available **(Figures 2A and 2B)**. There were 3 protein-protein interaction pairs experimentally identified in the year 1970 and 13 protein 3D structures made available in the PDB database in 1976. Genome sequencing became popular in the early 2000s after the Human Genome Project with around 64729 sequences from 2124 species available in 2003 **(Figure 1A)**. Gene sequence data was made publicly available in 1982 with 606 gene sequences deposited in GenBank **(Figure 1B)**. RNA-seq **(Figure 1C)**, Microarray **(Figure 1D)**, Proteomics **(Figure 2C)**, ChIP-Seq **(Figure 2D)** datasets are considerably new technologies that originated after 2003, hence the quantity of data is low as compared to protein interactions or gene sequences.

Further, an increasing exponential trend can be observed in all forms of biological datasets. After 2014, there is a considerable reduction in microarray datasets and a corresponding increase in RNA-seq experiments respectively, suggesting the choice of a better technology to achieve the same goal **(Figures 1C and 1D)**. The amount of curated nucleotide sequences of different species is also increasing considerably due to better understanding of gene function as an incentive from recent phenotype describing technologies **(Figure 1A)**. In contrast, the number of gene sequences added yearly to GenBank has reduced after 2010, a trend that is rather confusing while

considering the availability of cost-effective next generation sequencing technologies **(Figure 1B)**. The amount of proteomics **(Figure 2C)** and RNA-seq **(Figure 1C)** datasets are also exponentially increasing suggesting the collective focus towards technologies that explain the phenotype of any organism. The advent of high throughput proteomics has promoted the extensive purification of proteins that are further probed for their 3D structure determination. Hence, there is also an exponential increase in the number of protein structures **(Figure 2B)** being solved. The ChIP-Seq datasets have recently gained high importance for high-throughput identification of transcription factor binding sites and are also increasing in number rapidly **(Figure 2D)**.

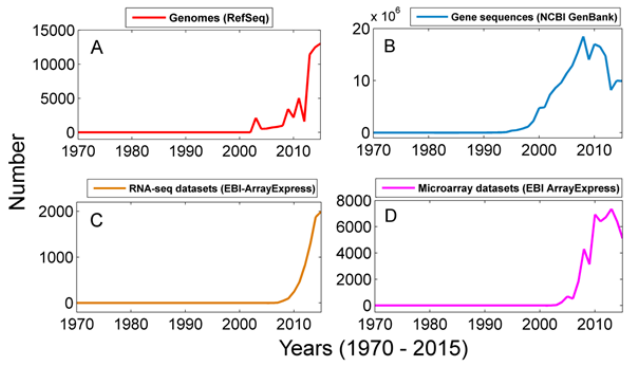

Figure 1: Plots displaying the trends of different nucleotide sequence-related biological datasets as reported in corresponding databases for the period 1970-2015. Each subplot represents the number of datasets generated from a distinct type of experimental technology and deposited in the respective database. A) Number of curated genomes in RefSeq database; B) Number of nucleotide sequences in the GenBank database; C) Number of RNA-seq datasets reported in EBI-ArrayExpress database; D) Number of Microarray datasets in the EBI-ArrayExpress database.

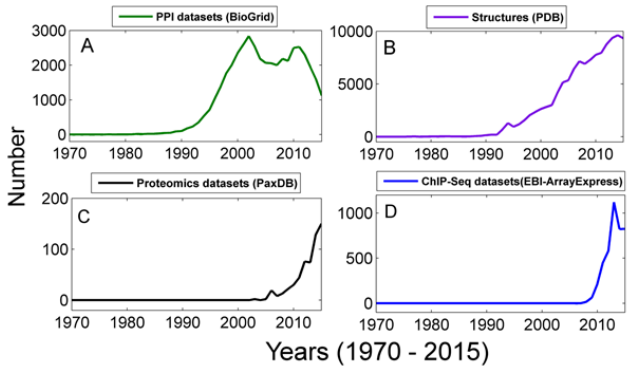

Figure 2: Plots displaying the trends of different protein-related biological datasets as reported in corresponding databases for the period 1970-2015. A) Number of protein-protein interacting pairs reported in BioGRID database; B) Number of structures released by PDB; C) Number of proteomic datasets present in PaxDB; and D) Number of ChIP-Seq datasets in EBI-ArrayExpress.

The continuous advancements in cellular and molecular biology experiments, genomics or proteomics studies have not only generated a plethora of data but also helped to identify the sub cellular localization of the pathway components and to annotate the biochemical pathway diagrams, eventually leading to development of various databases with reconstructed pathway maps, interactive user friendly interfaces to facilitate several operations, such as, pathway data retrieval, sharing and storing [8]. At the same time, these databases also face several challenges, such as, automated data curation and annotation, automated pathway image generation, requirement of pathway nomenclatures, lack of specific boundary conditions for pathway reconstruction, inability to show protein complexes, etc., which further leads to requirement of proper computation tools and methods to deal such problems. Similarly, at ecological scales a tremendous amount of data is generated to deal with global-scale environmental issues, climate change, food security, spread of disease, availability of clean water and most importantly species survival [9].

Thus the increasing amount of data generated at each level raises important issues in big data management including storage, accessibility, processing of data on the run, and security [10]. One of the most sought after solutions to these problems is the use of High Performance Computing clusters (HPCs), which not only provide storage solutions but also parallelize processing of computational tasks over the stored data. Specific applications can be catered towards distribution of data into all the CPUs of the cluster to obtain the processed information simultaneously. Towards the usage of HPCs, technologies like the MapReduce model have been developed within open-source software frameworks like the Hadoop project, where major computational tasks can be distributed as multiple small tasks and their outputs re-integrated for your final solution. Further, cyber-security protocols and web proxies installed in the HPCs secure the usage of the stored data. But, the flipside of using HPCs is the substantial cost involved in buying the required components of the cluster, their maintenance and support.

A relatively new solution to handle big data is the usage of "cloud computing" [10]. Cloud computing refers to a virtualization technology where the user on demand can store, process or retrieve data distributed over many virtual machines hosted on remote servers. These remote servers, the technology and services associated with it are provided by companies, like Amazon, Microsoft and Google, where the user only has to pay for the virtual system and the service that is required. Such methods provide a flexible, cost-effective way of using high-level computation power for the end-user to analyze petabytes of data at a time. Cloud computing can also be combined with the MapReduce model to further enhance the computational power to manipulate exabytes of data. As compared to storage and on the fly use of applications, data sharing and retrieval is the biggest concern for scientists. Users are limited by the usage of internet bandwidth and hence, face a problem to retrieve this large-scale processed data at high speed. A lot needs to be achieved in this direction as cloud computing largely relies on the uninterrupted, strong network connections between the host and the remote server.

With the advancements of science and technology, there will be a continuous generation of biological big data so as to understand more about the complex processes and it will also evolve continuously at different scales. Eventually this will pose several challenges demanding more time, space, large amount of monetary investment and at the same time require development of new computational tools, methods and technologies. The best way to tackle this data explosion may be to

follow a hypothesis-driven research, where a specific question is asked and the data generated or acquired aids to answer the preferred question [1]. A focused approach would reduce the extraneous generation and storage of data, as experiments would be directed towards the specific hypothesis needed to be proven. Statistical and mathematical techniques of data integration need to be applied on the already existing data to build a phenomenological or data-driven hypothesis that can explain core fundamental observations, which can be tested or proven on newly generated data. Predictions from such techniques would further help to reduce the search space and would narrow down the requirement of appropriate data generation. Decreasing the data redundancy at each level may also be a probable solution to deal this huge amount of information, hence development of novel techniques to scale-down the data is an utmost requirement.

The recently developed field of systems biology aims towards a similar integration of known hetergeneous data, their analysis and generation of experimentally testable hypotheses. It further explains the underlying design principles of different elements of biological systems and the association between different phenomena. A systemic view can help to narrow down the focus towards fundamental questions that can help to generate specific data thereby reducing the intensity of data explosion. This demands the collaboration of mathematicians, statisticians, experimental and computational biologists to propose specific projects aimed towards relevant implementation of hypothesis-based experiments that can control the data flood for meaningful information. Finally, development of new data mining techniques, faster and efficient search algorithms, novel computational strategies for integration and interpretation of information from different scales can be some of the future challenges to the researchers working in this direction.

Acknowledgment

The author is thankful to Mr. Abhishek Subramanian, CSIR-National Chemical Laboratory, India, for collating the information and plotting the figures. This work is supported by the grant from Council of Scientific and Industrial Research, XII Five Year Plan Project "GENESIS" (BSC0121).

References

1. Yixue Li, Chen L (2014) "Big biological data: challenges and opportunities." Genomics, Proteomics Bioinformatics 12: 187-189.
2. Marcotte EM, Date SV (2001) "Exploiting big biology: integrating large-scale biological data for function inference." Briefings in Bioinformatics 2: 363-374.
3. Stephens ZD, Skylar YL, Faraz F, Campbell RH, Zhai C, et al. (2015) "Big data: astronomical or genomical?" PLoS Biology 13: e1002195.
4. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, et al. (2014) "ArrayExpress update—simplifying data submissions." Nucleic Acids Res.
5. Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, et al. (2012) "PaxDb, a database of protein abundance averages across all three domains of life." Mol Cell Proteomics 11: 492-500.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) "The protein data bank." Nucleic Acids Research 28: 235-242.
7. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, et al. (2013) "GenBank." Nucleic Acids Research 41: D36-D42.
8. Chowdhury S, Sarkar RR (2015) "Comparison of Human Cell Signaling Pathway Databases - Evolution, Drawbacks and Challenges." Database (Oxford) 2015: bau126.
9. Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, et al. (2013) "Big data and the future of ecology." Front Ecol Environ 11: 156-162.
10. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) "Computational solutions to large-scale data management and analysis." Nature Reviews Genetics 11: 647-657.

This article was originally published in a special issue, entitled: **"${splIssueContent}"**, Edited by S1