International Journal for Computational Biology (IJCB)
Vol.6, No.1, July 2017, pp. 25-30
ISSN: 2278-8115

Structural Bioinformatics and Big Data Analytics: A mini-review

Pulkit Anupam Srivastava[1], Siddhant Kalra[1], Ragothaman M. Yennamalli[1*]
[1] Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Himachal Pradesh India 173234

Article Info

Article history:

Received Apr 12th, 2017
Revised Jun 6th, 2017
Accepted Jun 6th, 2017

Keyword:

Structural Bioinformatics
Structural Biology
Hadoop
Big Data
MMTF
Data Compression

ABSTRACT

Structural Biology and Structural Bioinformatics are two complementary areas that deal with three dimensional structures of biomolecules. With the advent of high-throughput techniques and automation of identifying structures there is a barrage of data generated currently, which fall under the area of Big Data. In this review, we present examples and current approach to handle massive volume of structural data and some potential applications of Big Data from Structural Bioinformatics perspective.

Corresponding Author:

Ragothaman M. Yennamalli,
Department of Biotechnology and Bioinformatics,
Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India 173234
Email: ym.ragothaman@juit.ac.in
Phone: 01792-239227

How to Cite:

Pulkit Anupam Srivastava *et. al.* Structural Bioinformatics and Big Data Analytics: A mini-review. IJCB.2017; Volume 6 (Issue 1): Page 25-30.

1. INTRODUCTION

Basic characteristics of Big Data are: Volume, Velocity, Variety, Variability, and Complexity. Any data that is enormous in size (i.e. large in volume) must have diverse nature making it difficult to mine and analyze. The speed of generation of these data should ideally be continuous and massive, which further increases the complexity to mange and organize the data [1]. In general, Big Data describes a holistic approach that includes and integrates various new-fangled types of data and data management besides traditional data.

Advent in robotics, automation, and emerging technologies in the field of structural biology and structural bioinformatics in recent years has resulted in increase in the volume, variety and complexity of data generated. More specifically, the X-ray free electron laser (XFEL) would collect thousands of electron diffraction data for each protein, where a stream of intense X-ray can capture the fast biological process using small or micro crystals. It is estimated that ~10 million x-ray images can be generated within 48 hours, and ~150-200 petabytes of data will be generated within a few years [2]. Thus, new technologies of structural biology and the data generated bring them under the category of Big Data.

2. Protein Sequence and Structural data is Big Data

The Protein Data Bank (PDB) was established as a digital data source with open accessibility providing a single gateway to access biomolecules' structures. Deposition of three-dimensional (3D) structural data has tremendously increased in the last five years (2012-2016) resulting in increased variety of data (i.e., source of generation; X-ray, NMR, Electron microscopy), complexity, volume, and speed of submission as reflected by the total and different structures deposited each year (Figure 1). The rate of submission of new structures is in the range of 8000-10000 per year. Having all the basic characteristics as of Big Data, protein 3D information requires the incorporation of multiple approaches required to analyze them. For example, BioMagResBank, Protein Data Bank Japan, Protein Data Bank in Europe, and RCSB PDB in partnership has epitomized how an international association can efficiently deal with Big Data as a freely available public that helps in further advances in fundamental and applied research and learning globally.

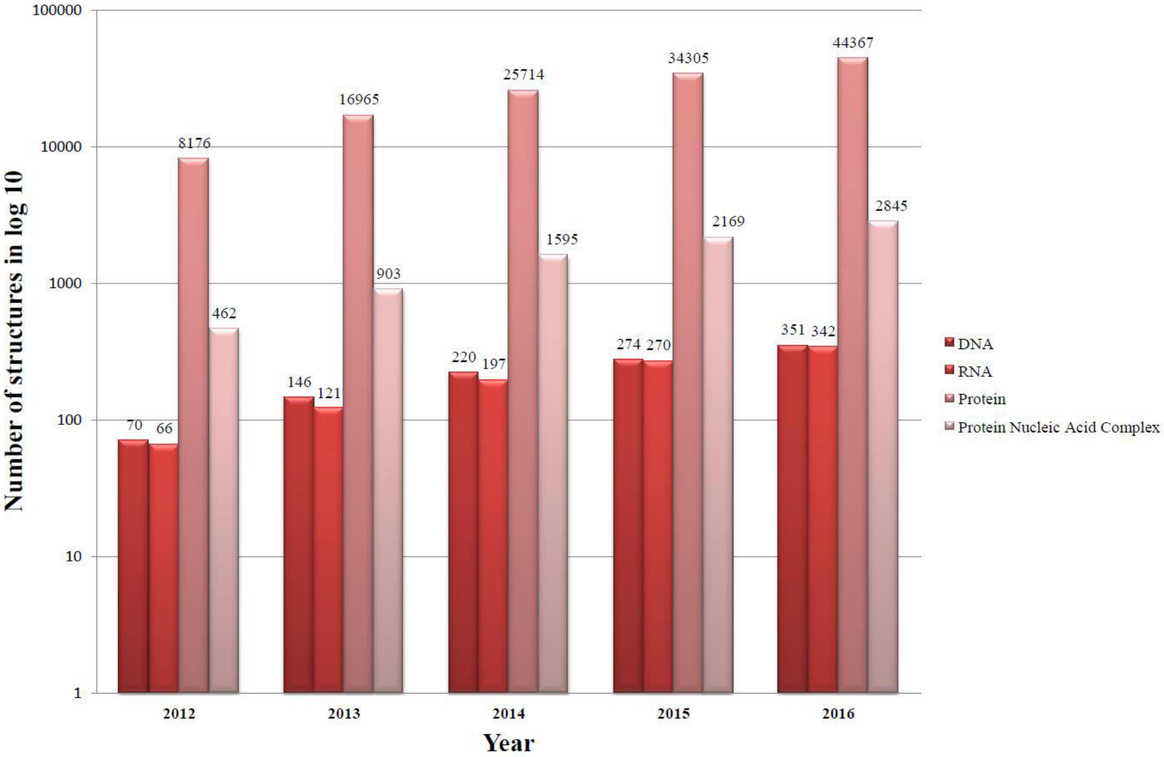

Figure 1: Five year cumulative growth of structures deposited in PDB. A cumulative plot of structures deposited in PDB is indicated in terms of type of molecule (DNA, RNA, Protein, and Protein Nucleic Acid Complex), where ~8000-10000 structures are deposited every year.

Since 2016, a modified format of PDB files (PDBx/mmCIF) has been incorporated to include structures that have more than 62 chains and more than 99,999 atom records. For example, the logistical problem came in forefront while submitting the structure of ribosome and for large viruses. With cryo-EM techniques solving high resolution structures (below 2Å) depositing them in PDB creates problems in accessing the data in developing countries that do not have high bandwidth internet access. In May 2013, two complete HIV capsid structures (PDB id: 3J3Q and 3J3Y) solved by cryo-EM method were deposited containing more than 100 chains, comprising two million atoms, each. The PDBx/mmCIF format, while following the already existing mmCIF format, does not restrict limitations on number of atoms, residues, or chains thereby removing the requirement of submitting split entries, as in the case of ribosome structure. Each file consists of categories or entities that in turn have sub categories that describe the various features. Due to its presentation in key-value form (i.e., tabular form) and use of context-free grammar parsing data from this format is easier. Also, the availability of meta-data makes it versatile format for programmers [3].

On the other hand, structural bioinformatics deals with analysis, prediction, and manipulation of data generated from structural biology projects. There has been significant gain in filling the diverse knowledge gaps with the available structures, using structural bioinformatics tools and algorithms [4]. Currently, PDB holds more than 100,000 structures. However, the protein sequence data is 810 times larger, accessed via Uniprot and NCBI. The one-dimensional information of protein sequence is also complex, varied, voluminous, and with recent advent in sequencing methods more sequences are being deposited (~7 x 10^6 to ~19 x 10^6 sequences in last five years).

3. Conventional methods to access Big Data

The difficulties in storing and accessing the data can be taken care of by the use of client server architecture where data is distributed among several devices and is accessible via local network. Currently, there are tools available to perform the same using different file systems approach namely: cluster, parallel, and distributed file systems. In parallel and cluster file systems, several processes can concurrently read and write a single file. While, distributed file systems comprises disk arrays connected to I/O servers through fast networks followed by sharing it to the further nodes of a cluster.

Lustre (http://lustre.org/) and General Parallel File System (GPFS) are two of the highest performance parallel file systems that can dynamically store data (least accessed data to cost effective storage) thereby making space for important data in expensive storage. Hadoop Distributed File System (HDFS) (http://hadoop.apache.org/) can achieve the same by spanning through large clusters of commodity servers and also making use of MapReduce approach to move the code to the data on a large, distributed system than to send all the data to the code.

In order to develop user specific solutions, many open source frameworks like, Hadoop (http://hadoop.apache.org/), Disco (http://discoproject.org/), Strom (http://storm.apache.org/), and Apache Spark were developed. However, Hadoop (has become synonymous with Big Data) and the programming model of MapReduce can be tedious to master for few tasks. Overcoming this limitation, Apache Spark does not depend on MapReduce and is comparatively 10 times faster on disk and 100 times faster in memory than Hadoop [5].

In past few years, HPC clusters facilitate the Big Data analysis on traditional platforms by accessing grid computing. However, the limitation of providing computational customization as per the user is one of the foremost disadvantages. This has led to increased importance of cloud computing services to perform large-scale analysis. Cloud computing, Xeon Phi, GPU computing, and Cluster computing are chief bioinformatics applications that makes use of these platforms. To handle the vast amount of data being generated by high throughput technologies, a new approach of moving in-house data to cloud computing is done to analyze the data in more efficient manner. Xeon Phi, on the other hand, is based on Intel’s Many Integrated Core (MIC) to perform the analysis but its dependency on parallelization paradigm and lower clock frequency affecting the long sequential part of an application makes it not fit for native mode. In contrast, GPU computing represents a relative inexpensive and powerful solution with high floating performance and parallelism and therefore motivated by the stipulation of game industries. In present, the comparison between best devices of GPU computing and Xeon Phi shows on average 30% more performance of the early one [6]. Cluster computing uses the data parallel approach, subdivision of data for analysis through independent process, which makes it highly scalable for many kind of Big Data analysis.

4. Current Advances in accessing Big Data: Application to Structural Bioinformatics

The PDB archive is technically a Big Data, and it become tedious to perform large-scale structural calculations such as geometric queries or structural comparisons, transmit and visualize 3D structure of biological macromolecules and store it efficiently. To overcome these drawbacks a new compact binary format, Macromolecular Transmission Format (MMTF), has been developed to store and transmit biomolecular structural data quickly (https://mmtf.rcsb.org/). For example, the PDB archive comprising of ~29 GB in mmCIF format can be stored fewer than ~7 GB using the MMTF format. Besides, it makes the analysis easier and simple as it contains pre-calculated information and thus it occupies less space and time. Specifically, the parsing time for the HIV viral capsid (PDB id: 3J3Q) was estimated to take 400 minutes and the same entry in the binary MMTF format was parsed within seconds (https://mmtf.rcsb.org/). It contains various fields such as format data, structure data, model data, chain data, group data and atom data. MMTF uses Java, Python and JavaScript for its implementation. The implementation of MMTF format from a normal data is schematically shown in Figure 2.

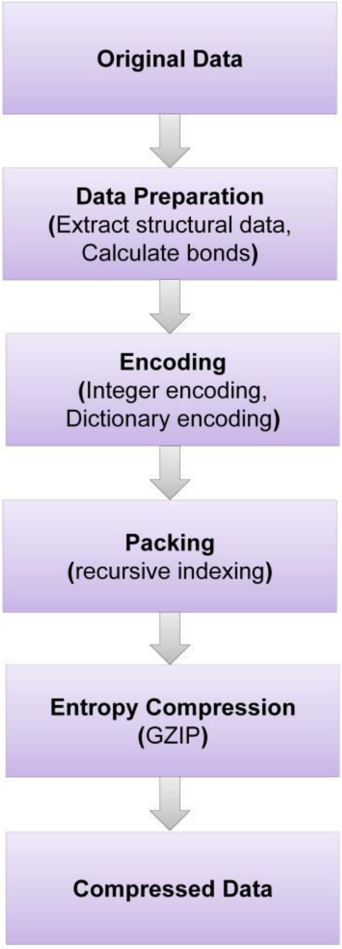

Figure 2: Converting normal data into MMTF format. Flowchart demonstrating the conversion methodology used to convert large data into MMTF format to store and parse it in an easier way (Adapted from https://mmtf.rcsb.org).

In the recent past, many applications of structural bioinformatics have dealt with large scale data analysis. For example, metagenomics data was used for *de novo* prediction of protein structures. However, methods required in extraction of high quality protein sequences and to distinguish between intra and inter chain contacts still presents cumbersome challenges in the path of *de novo* structure prediction of proteins [7, 8].

In order to analyze Big Data generated via experiments and complement the existing data numerous tools have been developed recently. Availability of such tools, although specific to their applications, will help in transcending the barrier from user point of view. For example, msBiodat is a web based tool that merges the data available in public databases with that of mass spectrometry results generated via high-throughput methods, resulting in much more precise list of proteins based on their annotations and conditions and help in pin pointing their biological processes [9].

Similarly, TFBSbank is a platform where transcription factor and DNA interactions are compiled. It consists of 1870 ChIP data from Drosophila, *C. elegans*, Human, *M. musculus*, and *S. cerevisiae* has been integrated along with analysis of co-binding, motif enrichment, and cofactor prediction. To counter the data-increasing day-by-day, *de novo* prediction of motifs present in transcription factors and their respective functional annotation is performed using chip-set data. This helped in analysis of transcription factors reported and left unannotated, since they are responsible for controlling gene expression [10].

SBGrid, a consortium that provides a curated suite of crystallographic software and packages, has an active community across the world. The SBGrid community allows storage, handling, and analysis of huge datasets collected from the primary source of crystallographic data collection. They have established a centralized data publication service called Structural Biology Data Grid (SBDG) where structural coordinates, diffraction image data sets can be published and disseminated. It is estimated that 48 Petabytes of storage are required to store crystallographic images for 100,000 structures [11]. Such consortium based efforts will aid the scientific community to not limit with data management problems; rather the presence of global infrastructure will demonstrate a cost-effective strategy to manage crystallographic Big Data.

5. Big Data for Drug Discovery

Drug discovery pipeline is an important application for Big Data research in structural bioinformatics. The most common approach of one target screened against millions of drugs *in silico* by molecular docking simulations has resulted in many successful lead and drug development stories. However, pharmacogenomic approaches for drug development require integrating multiple sources of information (genomic, proteomic, and other omics) data to drive a systems level approach towards drug development. Currently, the number of FDA approved drugs that include the pharacogenomic biomarkers in their labels are more than 138 [12]. The large number of samples with gene expression profile can be treated as Big Data and can be efficiently managed by organizing them in clusters with co-related expression patterns. Incorporating other associated data for the same set of genes with the pre-existing expression profiles will aid in accelerating the goal of personalized medicine [13].

Developing new methods that change the paradigm of data handling is the need of the hour for drug discovery. For example, Yabuuchi et al aimed to use the 10^{60} compound chemical space along with multiple large scale compound-protein interaction databases that are available to visualize and characterize the full complexity of interactions [14]. Thus, effective use of big data along with development of new methods to integrate and incorporate other resources should be done in order to evaluate and visualize the potential lead molecules for drug development.

6. CONCLUSION

As with any data that is sourced from others, there are some caveats or rules that apply. Recently, Zook et al came up with 10 simple rules for handling and conducting responsible research with Big Data. Specifically, they caution about ethics, privacy, weaknesses, and accountability with respect to Big Data research [15]. Big Data and its analytics with respect to structural biology and structural bioinformatics has been gaining attention and efforts are being made to develop tools/methods and algorithms that reduce the burden of data management, archival, and accessibility. We have highlighted some of the key examples that tackle these issues, specifically for handling protein sequence and structure information.

Due to lack of experimental evidences, many protein structures are yet to be solved and a method that can fill this gap is using homology modeling. While ModBase and other databases are progressing to generate models automatically, the accumulation of predicted models and experimental structures will have effect in the way these are accessed. Also, many groups publish coarse-grained and fine-grained simulation data for single and for huge complexes. With the push from many journals for data storage and reproducibility, quantitative and qualitative data also needs attention and practical solutions for their accessibility.

While, there is an enormous quantity of protein structures, the presence of mutated protein structures in equal measures is lacking in PDB. In developing countries, such as India, Big Data analytics in the area of Structural Biology and Structural Bioinformatics are in its nascent stages. However, with time it is estimated that the concurrent use of cloud-based networks with high-level machine learning approaches and computational methods for feature identification and validation can improve efficiency and reduce the cost of storage and managing Big Data.

ACKNOWLEDGEMENTS

PAS, SK, and RMY acknowledge Jaypee University of Information Technology for providing infrastructure for writing this manuscript.

REFERENCES

1. [Internet]. 2017 [cited 10 April 2017]. Available from: 1. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html
2. [Internet]. 2017 [cited 10 April 2017]. Available from: https://phys.org/news/2013-06-slac-x-ray-laser-explores-big.html /
3. PDBx/mmCIF General FAQ [Internet]. mmcif.wwpdb.org. 2017 [cited 10 April 2017]. Available from: http://mmcif.wwpdb.org/docs/faqs/pdbx-mmcif-faq-general.html
4. Samish I, Bourne P, Najmanovich R. Achievements and challenges in structural bioinformatics and computational biophysics. Bioinformatics. 2014;31(1):146-150.
5. Bell J. Machine Learning for Big Data: Hands-On for Developers and Technical Professionals. 1st ed. Indianapolis, Indiana: John Wiley & Sons, Inc.; 2015
6. Fang J, Sips H, Zhang L, Xu C, Che Y, Varbanescu A. Test-driving Intel Xeon Phi. Proceedings of the 5th ACM/SPEC international conference on Performance engineering - ICPE '14. 2014.
7. Söding J. Big-data approaches to protein structure prediction. Science. 2017;355(6322):248-249.
8. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim D et al. Large-scale determination of previously unsolved protein structures using evolutionary information. eLife. 2015;4.
9. Muñoz-Torres P, Rokć F, Belužic R, Grbeša I, Vugrek O. msBiodat analysis tool, big data analysis for high-throughput experiments. BioData Mining. 2016;9(1).
10. Chen D, Jiang S, Ma X, Li F. TFBSbank: a platform to dissect the big data of protein–DNA interaction in human and model species. Nucleic Acids Research. 2016;45(D1):D151-D157.
11. Meyer P, Socias S, Key J, Ransey E, Tjon E, Buschiazzo A et al. Data publication with the structural biology data grid supports live analysis. Nature Communications. 2016;7:10882.
12. Elsevier R&D Solutions. Big Data, Wider Mindset, The Netherlands: Elsevier Publications; 2015.
13. Zhang B, Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology. 2005;4(1).
14. Yabuuchi H, Niijima S, Takematsu H, Ida T, Hirokawa T, Hara T et al. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. Molecular Systems Biology. 2014;7(1):472-472.
15. Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan S et al. Ten simple rules for responsible big data research. PLOS Computational Biology. 2017;13(3):e1005399.