Big Data Technologies for DNA Sequencing

Lena Wiese, Armin O. Schmitt, and Mehmet Gültas

Definition

DNA sequencing is a modern technique for the precise determination of the order of nucleotides within a DNA molecule. Using this technique a huge amount of raw data is generated in Life Sciences.

Synonyms

Next-generation sequencing

1 Overview

Genome analyses play an important role in different applications in the Life Sciences ranging from animal breeding to personalized medicine. The technological advancements in DNA sequencing lead to vast amounts of genome data being produced and processed on a daily basis. This chapter provides an overview of the big data challenges in the area of DNA sequencing and discusses several data management solutions.

Lena Wiese
Institute of Computer Science, Georg-August University, Goldschmidtstraße 7, 37077 Goöttingen, Germany, e-mail: wiese@cs.uni-goettingen.de

Armin O. Schmitt and Mehmet Gültas
Department of Breeding Informatics, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany, e-mail: armin.schmitt@uni-goettingen.de, gueltas@cs.uni-goettingen.de

Next Generation Sequencing (NGS) technologies make it possible for life scientists to produce huge amounts of DNA sequence data in a short period of time [19]. Using these technologies, in recent years thousands of genomes and short DNA sequence reads for humans, plants, animals, and microbes have been collected and explored, which enables us to develop a deeper understanding and gain new insights into the molecular mechanisms of different diseases including many types of cancer, allergies or other disorders.

There is no doubt that NGS technologies bring considerable advantages in productivity or significant reduction in cost and time. The complexity and sheer amount of the resulting biological data sets are, however, more intricate than expected making their analysis and handling a real challenge.

Pedersen and Bongo [16] state that off-the-shelf big data management systems might not be appropriate for an efficient and effective management of biological data: "Biological data differs in that it has more dimensions and noise, it is heterogeneous both with regards to biological content and data formats, and the statistical analysis methods are often more complex." This generally also applies to the specific case of DNA sequencing data. In particular, DNA data are often processed in a data analysis pipeline (like the META-pipe [16]) where not only the raw data but also several additional contextual metadata and provenance data are generated and have to be maintained. The raw data produced during DNA sequencing are image data generated by the sequencing hardware. Further DNA processing steps comprise

primary analysis: producing short DNA sequences (called reads) out of the raw data and assigning quality scores to them;
secondary analysis: assembling several short reads guided by a reference DNA sequence – this process is called read mapping – as well as analyzing the reads with respect to the reference sequence by, for example, identifying single nucleotide variants or deletions;
tertiary analysis: processing the genome data to achieve advanced analyses by integrating several data sources (like multiple DNA samples, metadata, annotations etc.).

1.1 Big Data Challenges

Current life sciences are more and more data-driven. In practice, this means that data are recorded and generated in an essentially automatic way. A large portion of life science data is taken up by DNA sequences that are for example used to identify the genetic basis of a disease in the medical sciences or to identify (wanted or unwanted) traits of cultivated plants or animals which are supposed to be used in breeding programs. It is obvious that such large bodies of data can no longer be stored and analyzed in the traditional way but can only be harnessed by advanced data management and analysis systems.

For the years around the millennium the growth rate of computing power kept pace with the growth rate of data output of sequencing facilities, such that then

data analysis was feasible even with ordinary PCs. This remarkable parallelism has ended in 2008 with the emergence of a new sequencing technology—next generation sequencing. While computing power is predicted to follow Moore's law, that is, roughly doubling every 24 months, the years after the advent of NGS witnessed a decline of sequencing cost that could only be measured in orders of magnitudes as can be seen in so-called Carlson curves [4].

The challenges that arise by this gap between accumulation of data on the one hand and the capacities to store, analyze and interpret them in a meaningful way is asking for new and creative, perhaps also radical, ways to deal with data. We can distinguish two related, but distinct, aspects of this problem: (1) data storage and accessibility and (2) data analysis in acceptable times.

2 Key Research Findings

Several ways to address the big data challenges when processing genome sequencing data have been proposed in the last decade. We briefly survey several streams of research in the following subsections.

2.1 From data storage to data disposal

Thus far, the generally acknowledged, albeit in most cases tacit, agreement in the life sciences was to keep all raw data for the purpose of reproducibility or also for secondary analyses. This attitude has begun to erode in DNA sequencing projects. First, it should be clearly defined what raw data are in sequencing projects. In its most primitive form the sequence raw data are image files which are processed in several steps to yield short sequence reads which in turn are assembled to transcripts, genes or genomes. It has become generally adopted practice to consider the short sequence reads, together with quality scores, as raw data which can be archived in the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra). Probably unknowingly data scientists and analysts have adopted the plant breeders' wisdom: "Breeding is the art of throwing away" transforming it into: "Managing data is the art of throwing away" [1].

2.2 Sophisticated algorithms for data analysis

A PUBMED search has revealed a strong increase in scientific articles dealing with ultrafast algorithms suggesting that this is the response to the emergence of big data (Fig. 1). The speed of data processing in DNA sequence analysis can be increased by a combination of several factors, like storing sequence data in data structures

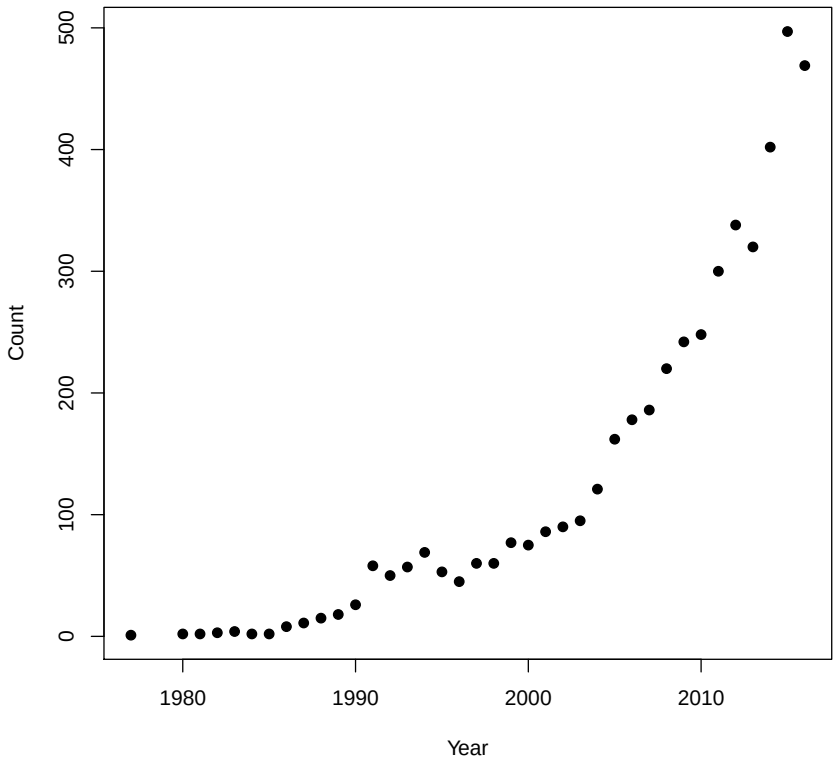

Fig. 1 Articles on ultrafast algorithms in pubmed

that permit rapid access (e. g. hash tables or suffix arrays) and the development of lightweight algorithms that confine themselves to steps that are absolutely necessary. For instance, the quantification of gene expression based upon transcriptome data could be accelerated by an order of magnitude thanks to so called quasi-alignments. In short, it is determined via a quasi-alignment if a short sequence read and a given gene sequence match without the explicit calculation of a nucleotide-by-nucleotide alignment extending over the full length of the short read sequence. This is simply not needed for the purpose of expression quantification [15].

2.3 DNA-specific compression

Due to its string nature, specialized encodings of genomic data can significantly reduce the storage consumption – either with or without loss of accuracy. In particular, the mappings of several short DNA reads to a longer reference genome string offers benefits in terms of compressed data representation. Various lossless and lossy compression methods were devised and implemented to boil down the disk requirements. Accepting limited data losses, a compression by two orders of magnitude has turned out to be feasible [17]. An even higher compression could be reached by so-called delta encoding [5].

Such kinds of compression have been incorporated into standardized file formats for genome representation. The Sequence Alignment/Map (SAM) format stores the start position of a short read with respect to the reference genome, the read's actual sequence and a CIGAR string denoting the differences between the short read sequence and the reference genome, thus representing a significant reduction in disk space requirement. The lossy CRAM file format applies a heavy-weight compression scheme and thus reduces the storage consumption even more [2].

Departing from the idea of processing data stored in flat files, integrating genome-specific compression into a database system which holds data in the main memory was developed in [7]. In this so-called base-centric encoding each base is stored in a separate row of a database table such that database-specific operations can be used to analyse the genome sequence. This technique requires that the entire genome dataset fit into the main memory.

2.4 Parallel processing and modern hardware support

The accurate and fast detection of genomic variants like DNA insertions/deletions or single nucleotide polymorphisms (SNPs) based on NGS data play an essential role in clinical researches. For this aim, several analysis pipelines combining short read aligners (e.g., BWA-MEM, Bowtie2, SOAP3 and Novoalign (http://novocraft.com/)) with variant callers (e.g., the Genome Analysis Tool Kit HaplotypeCaller (GATK-HC), Samtools mpileup, and Torrent Variant Caller (v4.0 Life Technologies)) have been developed. However, the analysis of the raw sequence data is computationally intensive and requires significant memory consumption. In order to deal with this problem, most aligners and variant callers have been implemented in multi-threaded mode (e.g., BWA-MEM, Bowtie2, and Novoalign) or using GPU-based software (SOAP3) to ensure a feasible computation time. For details like memory usage or multi-threading of the different approaches see the review [13]; also see the references therein for details about the different tools mentioned above.

Thus parallelization of analysis tasks is as important as the sequencing process itself and promises a huge potential for DNA processing. Several frameworks aimed at parallelization for specific applications. The map reduce paradigm has hence received a lot of attraction. [12] describe a distributed framework for read mapping

based on the message passing interface (MPI) in a cluster of nodes. They discuss the issues of splitting and distributing the inputs as well as merging the results. Already in the year 2010 an overview [20] surveyed applications of the Hadoop framework; more recently, other approaches based on the MapReduce framework have been developed [6]. Other distributed processing systems like Spark have also received attention for applications in genome sequencing and processing [14].

To benefit from advances of modern hardware technology, some DNA analysis processes have been ported to run on graphical processing units (GPUs) [18]. However, these approaches incur an overhead for preprocessing the data and require highly specialized algorithms in order to take advantage of the GPU platforms.

2.5 Integration of Heterogeneous Data

Considering the growth rate of DNA sequencing data, Stephens et al. have demonstrated in their study [19] that the sequencing technologies are one of the most important generators of big data. However, these massive datasets are often neither well-structured nor organized in any consistent manner which makes their direct usage difficult. Consequently, the researchers have, first of all, to deal with the handling of big data to perform any analysis or comparison studies between different genomes. Thus, an effort for these big data challenges in life sciences is needed today to store the data in hierarchical systems with more efficient ways that make them available at different levels of analysis.

Management of biological data often requires the connection and integration of different data sources for a combined analysis. By default, biomedical text formats use identifiers (for example for genes or genomic variants) and different text files can only be combined by ID-based linking: data combination heavily relies on string equality matching over these IDs. To exacerbate the situation, these IDs are often hidden in larger strings such that these IDs have to be extracted first before data can be joined. A conjecture shared by many researches in the field is that, on the high-level of data management, explicitly linking data items in a graph structure by navigational access will enable more efficient data combination than join operations over string IDs. Hence, graph databases are deemed to be most suitable for such a data integration layer [9] because different data sources can be combined by a link (edge) in the data graph.

One prototypical framework that integrates genome data sources and other biological data sets in a common graph-based framework is the BioGraphDB presented by Fiannaca et al [8] who use the OrientDB multi-model database to interconnect several text-based external data sources. The BioGraphDB system is able to process queries in the Gremlin graph query language. Textual input data is processed in this framework as follows [8]: “As general rule, each biological entity and its properties have been mapped respectively into a vertex and its attributes, and each relationship between two biological entities has been mapped into an edge. If a relationship

has some properties, they are also saved as edge attributes. Vertices and edges are grouped into classes, according to the nature of the entities."

3 Examples of Application

The raw sequencing reads are often stored in public genome repositories like:

- National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov),
- ENCODE (https://www.encodeproject.org/),
- Genome 10K Project (https://genome10k.soe.ucsc.edu),
- TCGA (https://cancergenome.nih.gov),
- Human Microbiome Project (https://hmpdacc.org/)

in order to make these large bodies of research data easily accessible. We survey some of these repositories with a focus on bioinformatics applications.

3.1 Bioinformatics for Sequencing Data

Bioinformatics skills play an essential role in the exploitation of the full potential of NGS data. Until now, different bioinformatics tools and algorithms have been published for the storage and computational analysis of DNA sequences. Such applications are important for the detection as well as the understanding of relevant biological processes. Currently, in the OmicTools directory (https://omictools.com/) for NGS data analysis there are 47 categories with 9089 applications which are developed, for example, for data processing, quality control, genomic research, data visualization or for the identification of robust genomic associations/variants with complex diseases.

In addition to computational analysis of sequencing data, the field of bioinformatics is crucial for storage of large-scale sequencing data in databases as well as repositories. The Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) is one of the most relevant public-domain repositories in which raw sequence data generated using next generation sequencing technologies are stored for free and continuous access to the data is possible. There are further big data projects that are based on databases/repositories collecting sequencing data in life sciences:

The Encyclopedia of DNA elements (ENCODE) stores more than 15 terabytes of raw data and is thus one of the most general databases for basic biology research.

The Cancer Genome Atlas (TCGA) is an extensive effort to map the key genomic changes in 33 types of cancer. This database contains to date 2.5 petabytes of data collected from more than 11,000 patients.

The Genome 10K Project is a collection for storage and analysis of the sequencing data corresponding to 10,000 vertebrate species. Scientists expect more than 1 petabyte by completion.
Human Microbiome Project is a collection of several databases to provide quick and easy access of all publicly available microbiome data. Currently, the database contains over 14 terabytes of publicly available data in total.

4 Future Directions for Research

DNA sequencing is one of the major producers of Big Data. Novel sequencing technology will even increase the amount of DNA sequencing data produced. The introduction of mobile sequencing devices (like nanopore-based technologies [11]) will turn DNA-sequencing into an everyday diagnosis and monitoring tool. Though initially the available devices suffered from high error rates, ongoing technological advances will lead to an improved data quality [10].

It might be worthwhile to closely inspect genome data processing pipelines in order to identify bottlenecks. [12] report on the use of high-performance profiling that revealed idling hardware resources. By improving the task scheduling, a better runtime could be achieved. This kind of low-level performance optimization is one field of future research to improve performance of DNA processing.

Extrapolating the current drop in sequencing cost, the most radical option could be so-called streaming algorithms for sequence analysis. With streaming algorithms, DNA sequences are analyzed in real-time and are not stored at all, as shown in [3].

References

1. Becker, H.: Pflanzenzüchtung. UTB basics. UTB GmbH (2011)
2. Bonfield, J.K., Mahoney, M.V.: Compression of FASTQ and SAM format sequencing data. PLoS ONE **8**(3), e59,190 (2013)
3. Cao, M.D., Ganesamoorthy, D., Elliott, A.G., Zhang, H., Cooper, M.A., Coin, L.J.: Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time minion tm sequencing. GigaScience **5**(1), 32 (2016)
4. Carlson, R.: The pace and proliferation of biological technologies. Biosecurity and bioterrorism: biodefense strategy, practice, and science **1**(3), 203–214 (2003)
5. Christley, S., Lu, Y., Li, C., Xie, X.: Human genomes as email attachments. Bioinformatics **25**(2), 274–275 (2008)
6. Chung, W.C., Chen, C.C., Ho, J.M., Lin, C.Y., Hsu, W.L., Wang, Y.C., Lee, D.T., Lai, F., Huang, C.W., Chang, Y.J.: Clouddoe: a user-friendly tool for deploying hadoop clouds and analyzing high-throughput sequencing data with mapreduce. PloS one **9**(6), e98,146 (2014)
7. Dorok, S., Breß, S., Teubner, J., Läpple, H., Saake, G., Markl, V.: Efficiently storing and analyzing genome data in database systems. Datenbank-Spektrum pp. 1–16 (2017)
8. Fiannaca, A., La Rosa, M., La Paglia, L., Messina, A., Urso, A.: Biographdb: a new graphdb collecting heterogeneous data for bioinformatics analysis. Proceedings of BIOTECHNO (2016)

9. Have, C.T., Jensen, L.J.: Are graph databases ready for bioinformatics? Bioinformatics **29**(24), 3107 (2013)
10. Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., Akeson, M.: Improved data analysis for the minion nanopore sequencer. Nature methods **12**(4), 351–356 (2015)
11. Loman, N.J., Watson, M.: Successful test launch for nanopore sequencing. Nature methods **12**(4), 303 (2015)
12. Martínez, H., Barrachina, S., Castillo, M., Tárraga, J., Medina, I., Dopazo, J., Quintana-Ortí, E.S.: A framework for genomic sequencing on clusters of multicore and manycore processors. The International Journal of High Performance Computing Applications p. 1094342016653243 (2016)
13. Mielczarek, M., Szyda, J.: Review of alignment and snp calling algorithms for next-generation sequencing data. Journal of Applied Genetics **57**(1), 71–79 (2016). DOI 10.1007/s13353-015-0292-7. URL https://doi.org/10.1007/s13353-015-0292-7
14. Mushtaq, H., Liu, F., Costa, C., Liu, G., Hofstee, P., Al-Ars, Z.: Sparkga: A spark framework for cost effective, fast and accurate dna analysis at scale. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 148–157. ACM (2017)
15. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C.: Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods **14**(4), 417–419 (2017)
16. Pedersen, E., Bongo, L.A.: Big biological data management. In: Resource Management for Big Data Platforms, pp. 265–277. Springer (2016)
17. Popitsch, N., von Haeseler, A.: Ngc: lossless and lossy compression of aligned high-throughput sequencing data. Nucleic acids research **41**(1), e27–e27 (2012)
18. Salavert Torres, J., Blanquer Espert, I., Tomas Dominguez, A., Hernendez, V., Medina, I., Terraga, J., Dopazo, J.: Using gpus for the exact alignment of short-read genetic sequences by means of the burrows-wheeler transform. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **9**(4), 1245–1256 (2012)
19. Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E.: Big data: astronomical or genomical? PLoS biology **13**(7), e1002,195 (2015)
20. Taylor, R.C.: An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. BMC bioinformatics **11**(12), S1 (2010)