Challenges of Data Management in Next Generation Sequencing

Manisha Rana*

Department of Forensic Science, Amity University, Uttar Pradesh, India

***Corresponding author:** Manisha Rana, Department of Forensic Sciences, Amity University, Uttar Pradesh, India; E-mail: rana.manisha1493@gmail.com

Data management has been perceived as a key challenge for current molecular biology research ever since the start of the Human Genome Project. Before the end of the 1990s, developments had been developed that adequately supported most ongoing operations, typically focused on frameworks of relational database management. A sensational increase in the amount of information created by running projects that extend in this area has been seen over the years.

Although it took more than ten years, nearly $3 billion, and more than 200 gatherings worldwide to collect the entire human genome, the new sequencing machines produce a comparable amount of crude information in seven days, at a cost of about $2000 and on a single device. Currently a few national and international programs handle a large number of genomes, and trends such as customized medicines call for attempts to sequence the entire population. In this article, we highlight difficulties that emerge out of this data explosion, such as calculation parallelization, genomic sequence compression, and cloud-based execution of complex scientific work processes.

They also point out some potential challenges that lie ahead due to the increasing increase in translational medicine, i.e. the rapid transition in the effects of biomedical research into medical practice (Figure 1). While much research has just been done to expand the performance of mapping devices for reading, scalability remains an open challenge. If the state of the art parallel or disseminated read aligners will process the amount of data generated in vast sequencing projects within a sensible measure of time and space is still unknown. Finding alignments that display large holes requires exceptional calculations as common heuristics normally create unacceptable accuracy when confronted with such data. In transcriptome projects these data especially shows mature mRNA sequencing.

In eukaryotic life forms the majority of genes includes several non-coding stretches of DNA called introns. Transcripts of the genes undergo a process called splicing, where they remove these introns. Introns can be a few hundred thousand nucleotides in length; in this way, changing a sequenced mRNA back to a genome requires incredibly large holes to manage. Another area where large holes occur is prompts notable genomic rearrangements. Both of these matters are highly diverse research areas. Another open question is how to integrate consistency scores into the calculations for read mapping. All sequencing machines produce qualitative scores alongside each base, showing the probability that this particular base is accurate.

Using these quality scores while reading mapping is known to improve mapping accuracy [1], but it is beyond the reach of imagination with current read mapping tools. With a consistently expanding number of bundles of read mapping programming, picking the best one for a specific sequencing venture and checking the nature of the subsequent alignment is a significant challenge. It is further muddled by constant changes to the software packages, which can improve the output both in terms of running time and level of alignment. While a few papers gave the impression that the output of various devices was analyzed, a widely accepted criterion by which read mapping programming could be measured still cannot seem to develop [2].

Scalability and compression speeds are the primary obstacles for series compression. As regards scalability, the question as to how an ideal compression can be obtained in a short time is still open. An ideal referential compression is the one with the minimum space requirements which involves taking care of complex optimization issues in order to change the length of referential matches and

the length of crude sequences in the middle. No solution for this problem is known to the best of our knowledge, nor is it understood how similar the existing approaches lead in these (hypothetical) present circumstances. Regardless, compression rate must be matched with speed of compression. Another open problem occurs as long as a set of thousands of sequences can be stored in compact form, not a singular, but rather a group S. The higher the rate and speed of compression, the more comparable the reference and the sequences to be compressed are.

Now, the problem is to find the one sequence s from S which is most likely to be identical to any other sequence, thereby making s the best possible comparison to be used. K-mer hashing can be used to base heuristics for finding a good reference sequence. High k-mers comparability indicates a high compression capacity as for the comparison. In any case, k should be selected higher than 15 at the genome scale, in order to preserve a strategic buffer from such a large number of irregular matches. Another open issue is compression-reading. Although compression of the genome typically only thinks about the sequence itself, quality scores should also be included in reading compression. Compression of these consistency scores overwhelms the compression rate of peruses as these scores have higher entropy than the base symbols. Future research will explore how the quality scores are actually used, and the scoring resolution is significant.

Eventually, a typically unexplored problem is how compressed sequences can be examined in a clear way, rather than decompressed before any use. On the off chance of having to compare 1000 genomes together, nothing is achieved by compressing them on the off chance that they will all be decompressed again before study. There is a need for string search calculations along these lines, which can essentially allow use of the current list structure of a reference series and compressed files by reference. Scientific research methods have raised interest in computational science over the last few years. A further open challenge is the combination of referential compression and string search into these work processes. The growth of cloud computing innovation has made flexible software of exceptional flexibility easily available and affordable for the end user.

In this way, the use of (open) cloud assets to carry out scientific work processes has become a noteworthy subject of enthusiasm for late years [3,4]. Nevertheless, the efficient use of a cloud of (generally virtual) computers for scientific work processes brings up a few problems that are still largely unexplored. Next, the question of how to get input (and output) data to (and from) the cloud sets up a severe challenge when attempting to use the cloud to analyze BIG data.

One option for NGS data may be compression; another solution is that cloud providers provide pre-designed images that contain important sequence data such as reference genomes. EC2 clients can mount the entire Genbank database from any image, for example. Unmistakably, if the novel series is to be checked, the last option is of no help. Thereafter, a critical but open problem is the consistent integration of compression/decompression measurements into scientific work processes. Second, the problem of skillfully mapping work process failures on heterogeneous distributed computing nodes – virtual machines in a cloud, for example-is not yet acceptably illuminated. Different kinds of parallelism may be misused. Due to the enormous amount of NGS data, data transfer times must be considered when deciding which tasks to execute on which machines.

A work process scheduler will ideally have the ability to continuously adjust the execution of an provided work process to a dynamic domain where transfer speed, memory accessibility, and size of appointed nodes change with high recurrence, as this is currently the case in most accessible cloud situations [5]. A great scheduler will likewise be able to make use of the elasticity offered by public clouds to finish anything. Successfully using elasticity in the correct implementation of the work cycle is a task that none of the present structures seems to address adequately.

In any case, the use of an explosion of (ordinarily virtual) computers to gain coherent modes of work poses a few problems that are still usually unexplored. For example, the question of how to get data (and yield) to (and from) the cloud involves a genuine check when attempting to use fogs for BIG data analysis. Pressure could be one solution for NGS data; another course of action is that cloud providers provide pre-planned images that contain simple progression data such as reference genomes. For

example, EC2 clients can mount the entire Genbank database from any image. Undeniably, if novel progressions are to be penniless down the last decision does not help. Pressure may be one solution for NGS data; another course of action is that cloud providers provide pre-planned images containing simple progression data such as reference genomes. For example, EC2 clients can mount the entire Genbank database from any image. Undeniably, if novel progressions are to be penniless down the last decision does not help. Remember that the situation will soon turn out to be much more detestable (or much harder): First, the scale of sequencing undertakings must develop and grow due to the dropping sequencing costs. Second, it is common for the third era of sequencing machines to end up accessible within the following few years [6].

A few development routes are being pursued; they all share that the speed of sequencing and the length of reads will increase dramatically for all intents and purpose. There's no question that the "100 Million Genome" is only a few years ahead. There are also some issues we didn't mention in this article. Metadata management for a large number of genomes, for example, must be carefully designed so as not to lose critical genome-related data. Another issue is reconciling big genomic data sets with various kinds of information, similar to gene ability or interaction. One especially difficult topic is data security. The genomic data is profoundly personal and fragile. Also, sequencing data anonymization or pseudonymization is not just a matter of removing the identity of the sender from the data, as the data itself would probably identify the donor. Pro-bands of genomic studies may need to be assured, in a research sense, that they have some form of authority over this sensitive individual data.

Genomic data may be regarded as personal health information in a clinical environment, making its safety essential and even required by law [7]. It highly limits the use of free cloud-based read mapping administrations, and also brings into question sequencing for commercial administrations. Conceivable arrangements integrate the basis of non-open "walled" cloud-based arrangements with strict and secure access control, or the development of cloud-based read mapping that does not entail the transfer of the genuine read sequence to the general public cloud [8].

References

1. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves the accuracy of solexa read mapping. BMC Bioinform. 2008;9.

2. Holtgrewe M, Emde AK, Weese D, Reinert K. A novel and well-defined benchmarking method for the second generation read mapping. BMC Bioinform. 2011;12.

3. Hoffa C, Mehta G, Freeman T, Deelman E, Keahey K, et al. On the Use of Cloud Computing for Scientific Workflows. In: Proceedings of the 2008 Fourth IEEE International Conference on EScience. IEEE Computer Society, USA. 2008; 640-645.

4. Juve G, Deelman E, Vahi K, Mehta G, Berriman B, et al. Data Sharing Options for Scientific Workflows on Amazon EC2. In: SC'10: Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis. New Orleans, LA. 2010;1-9.

5. Zaharia M, Konwinski A, Joseph AD, Katz RH, Stoica I. Improving MapReduce Performance in Heterogeneous Environments. In: Proceedings of the 8th USENIX conference on Operating systems design and implementation. USENIX Association. Berkeley, California, USA. 2008;29-42.

6. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227-R240.

7. OCR privacy brief: Summary of the HIPAA privacy rule; HIPAA Compliance Assistance. Department of Health and Human Services, USA. 2003.

8. Chen Y, Peng B, Wang X, Tang H. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. In: Proceeding of the 19th Network & Distributed System Security Symposium. 2012.

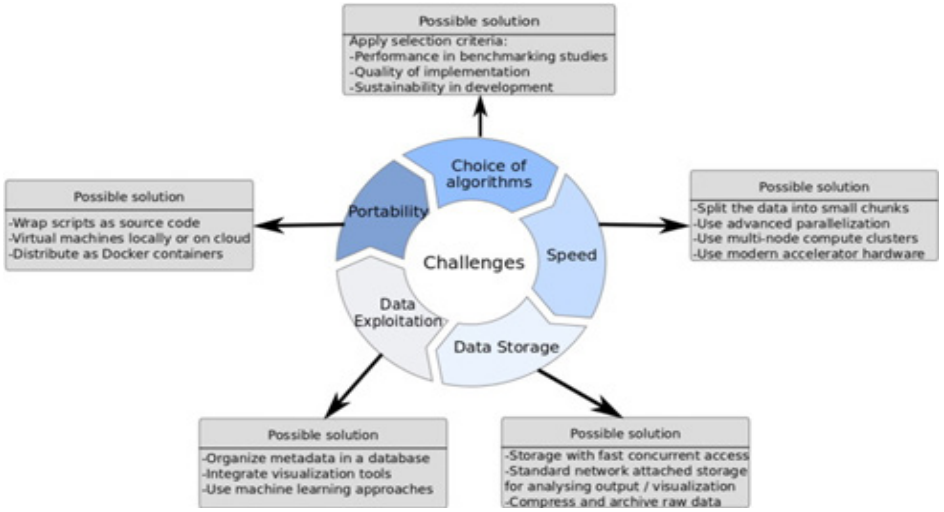

Figure 1 Possible challenges with possible solution.