Importance of databases of nucleic acids for bioinformatic analysis focused to genomics

L R Jimenez-Gutierrez1, C J Barrios-Hernández2, G R Pedraza-Ferreira2, L Vera-Cala2 and F Martinez-Perez2
1 Universidad Autónoma de Sinaloa, Mazatlán, Sinaloa, México.
2 Universidad Industrial de Santander, Bucaramanga, Colombia.

E-mail: ljimenez@uas.edu.mx , fjmartin@uis.edu.co

Abstract. Recently, bioinformatics has become a new field of science, indispensable in the analysis of millions of nucleic acids sequences, which are currently deposited in international databases (public or private); these databases contain information of genes, RNA, ORF, proteins, intergenic regions, including entire genomes from some species. The analysis of this information requires computer programs; which were renewed in the use of new mathematical methods, and the introduction of the use of artificial intelligence. In addition to the constant creation of supercomputing units trained to withstand the heavy workload of sequence analysis. However, it is still necessary the innovation on platforms that allow genomic analyses, faster and more effectively, with a technological understanding of all biological processes.

1. Introduction
Since the last two decades, bioinformatics has evolved from a field that was not considered one of the branches of science; but now is a key element in the development of basic and applied sciences. On the other hand, its definitions change in function of its execution area; for example, Hogeweg [1] defines bioinformatics as the study of processes in biotic systems. Whereas, Tenenbaum [2] shows the translational bioinformatics definition from the American Medical Informatics Association (AMIA): the development of methods of analysis, storage, and interpretation, which optimize the transformation of increasingly genomic and biomedical data, into predictive, and participatory health. Besides, the Dr. Russ Altman from Stanford University, in his 2014 presentation, provided the following definition for bioinformatics: “informatics methods that link biological entities (genes, RNA, proteins, and small molecules) to clinical entities (diseases, symptoms, and drugs) and vice versa”.

We consider bioinformatics as a science which final goal is to generate knowledge from biological data, which are analysed from computational systems by mathematic algorithms and/or artificial intelligence. From this perspective, the generation of data with biological importance are placed in codes in a database in computer systems. This is considered the starting point; meaning the success or failure of any bioinformatic analysis.

This principle is critical, because, if the data not accord to the parameters of experts in the field you want to describe and understand, the study model will not explain the reality and accuracy of the system. That is why that the experimental strategy and technological elements that allow data generation are the key elements for achieving results, which when placed in a code database, describing any type of system; specially a biological system.
2. Biological aspects and experimental developments which has allowed obtain the nucleic acids and protein sequences; to form the basis of current data

Gregor Johann Mendel could be considered the originator and systematizer of the biological data obtained related to inheritance. He assigned a numerical value to his results in plants hybridization to apply their mathematical models and explain their results [3]. Subsequently, a number of discoveries were instrumental in the ongoing development of bioinformatics of nucleic acids. The first, corresponds to the purification of deoxyribonucleic acid by Friedrich Miescher with 5 papers among 1892-1897 [4], he discover the nucleus as the DNA carrier. Phoebus Levine subsequently discovered that the nucleic acids are composed of a sugar: ribose or deoxyribose, and 5 types of bases: adenine, thymine, guanine, cytosine and uracil [5]. Edwin Chargaff [6] describes that they are associated in proportions 1:1 adenine - thymine, adenine - uracil and guanine - cytosine. After that, Avery [7]; Hershey and Chase [8], demonstrated that nucleic acids are molecules containing genetic information in bacterial and virus, respectively, and Watson and Crick [9] proposed the double helix model of DNA.

With all these biochemical and chemical background, in 1977 were published two methods to establish the nucleotide sequence of a DNA segment: one proposed by Allan M. Maxam and Walter Gilbert [10]; which was based on the hydrolysis of DNA. Meanwhile, the second proposed by Frederick Sanger et al [11] with a nucleic acid sequencing performed by incorporating radioactive nucleotides into DNA chains and then separated by electrophoresis. The relative simplicity of implementation of either sequencing method allows obtaining nucleic acid segments, quickly and easily, and with thus, the implementation of codes for its analysis. But with the growing number of nucleotide sequences, it was necessary to provide a solution to a seemingly inconsequential problem, but fundamental; the development of a communication system of the nucleotide sequences for study of any research group.

3. Development of computing and informatics applied to building databases

Bioinformatics is a multidisciplinary domain with a traditional and large use of high performance computing capabilities. In fact, the idea to improve knowledge with the innovative application of computer science and computer resources allows the possibility of treatment of data in life sciences. High Throughput data, sequencing and data visualization are today bio-informatics technologies linked with laboratory experiences and theory.

Recently, EEUU presidents Bush and Obama, in 2004 and 2010 respectively, proposed improvement of computational resources to understand human biology addressed to personalized healthcare. Then, projects as human genome [12], human brain [13], Parkinson Modelling Project, human heart model, among others, are developed faster in EEUU, with a worldwide participation; mainly of the European Union and China. These projects demand an important use of high performance computing centres to use de genomic database results.

If we observe the Top500.org site [14], which contains the 500 supercomputers with the most important performance of the world, it is possible to distinguish that the 48% of these supercomputer platforms are used for research, of which 50% is related to life sciences. Bioinformatics with corresponding databases for any field puts enormous challenges to computer sciences, this challenge implies five important subjects: Complexity of the System, Size of the Data, Processing and Data Flow, Heterogeneity of the Data and Computing Power. The complexity of the systems must be treated with the corresponding algorithms and implementation mechanisms. study is necessary to propose solutions, and of course, a posterior implementation in a computing algorithm.

The size of the data has two points of views: technology for storage and security requirements, and mathematical (related to complexity). In the first point of view, the capacity and safety of the technological solutions are important to store the data. The second, same than the complexity, algorithms and mathematic process are important, for example, using computational techniques; such as big data or deep learning for large scale databases.
Processing and Data Flow are treated using software engineering techniques, and of course, the processing is transversal to the algorithm implementation and coherence of the results. At the same time, it is important to observe the data heterogeneity. This involving the nature of the algorithms, the use of non-structured formalization of the data and its implementation, and at same time, technological computer architecture. Finally, the computing power improves the performance, and of course, it allows to obtain results in predictable short times.

For genomics, databases involve sequences, maps, annotations and chromosomes. Different available resources have different sizes and formats, then for the analysis and computer treatment is important to observe the format and provide “simple” standard results. The large size of the datasets and the heterogeneity requires specific techniques to make an implementation of algorithms and codes. On the other hand, the visualization of the large scale data and results is important; however, in this paper this subject is not shown.

The particularity of the datasets and the high-granularity of the data and related tasks to genomic studies allow implementations in heterogeneous computer platforms, based in GPUs (Graphical Processing Units); for example, the GUANE platform (GpUs AdvaNced computing Environment), a 128 Fermi TESLA GPUs supercomputer based in the High Performance and Scientific Computing of the Universidad Industrial de Santander, Colombia. The implementation of the algorithms is developed in NVIDIA(r) CUDA(r) [15]; a programing model proposed by NVIDIA to implement mechanisms of exploitation and implementation of granularity over GPUs. This architecture allows to treat complexity, large scale and heterogeneity of data.

4. Nucleic acids and genomic database

Due to the high number of reports of sequences of nucleic acids and proteins generated by the old methodologies and new technological advances, in 1979, 35 scientists meeting at Rockefeller University, they concluded the urgency of establishing a nucleic acids International database, which could correlate, as much as possible, the biological information; concluding with the creation of the GenBank. After this meeting the international scientific community continued with similar events [16, 17]. However, with the increase of nucleotide sequences, in 1987 was created The International Nucleotide Sequence Database Collaboration (INSDC), which currently consists of the DNA Data Bank of Japan (DDBJ) at the National Institute for Genetics in Mishima, Japan; the European Nucleotide Archive at the European Bioinformatics Institute (EMBL) in Cambridge, UK; and GenBank at the National Centre for Biotechnology Information (NCBI) in Bethesda, MD, USA [18, 19].

Since its creation, the INSDC guideline and policy is keeping nucleotide sequences obtained by experimental methods in public databases, and allow unrestricted access to any person or institution. In this sense, in 2015 this database was constituted by 2.400 trillion bases [19]. This amount will increase in the next years thanks to new genome sequencing technologies developed by Pacific Biosciences, Illumina, Roche or Ion Torrent [20, 21]; besides to the new Computer platforms and computational programs that will allow the completion of projects such as personalized genomic medicine [21], food, and basic research projects.

But even with this genomic revolution, that will increase the nucleotide databases, the challenge remains finding genomes regions whose composition and base order has not allowed their obtaining. For this purpose, the scientific community is developing technologies to increase capacity and fidelity of polymerization performs by DNA polymerases and reduce assembly times of genomes and their products. The solution to these and other technological processes will allow increase the genomic databases, but the biggest challenge is the proper use and application of the nucleotide sequences, which will be built for humanity benefit.
References

[1] Hogeweg P 2011 The roots of bioinformatics in theoretical biology *PLoS. Comput. Biol.* 7(3) e1002021
[2] Tenenbaum J D 2016 Translational Bioinformatics: Past, present, and future *Genomics Proteomics Bioinformatics* 14(1) 31-41
[3] Mendel G 1866 Versuche über Pflanzen-Hybriden *Verhandlungen der naturforschung Vereins* 4 3–47
[4] Dahm R 2008 Discovering DNA: Friedrich Miescher and the early years of nucleic acid research *Hum. Genet.* 122(6) 565-581
[5] Levene P A 1903 On the chemistry of the chromatin substance of the nerve cell *J. Med. Res.* 10(2) 204-211
[6] Chargaff E 1951 Some recent studies on the composition and structure of nucleic acids *J. Cell. Physiol. Suppl.* 38(1.1) 41-59
[7] Avery O T Macleod C M McCarty M 1944 Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: induction of transformation by a deoxyribonucleic acid fraction isolated from Pneumococcus type III *J. Exp. Med.* 1(79-2) 137-58
[8] Hershey A D Chase M 1952 Independent functions of viral protein and nucleic acid in growth of bacteriophage *J. Gen. Physiol.* 36(1) 39-56
[9] Watson J D Crick F H 1953 Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid *Nature.* 171(4356) 737-738
[10] Maxam A M Gilbert W 1977 A new method for sequencing DNA *Proc. Natl. Acad. Sci. USA* 74(2) 560-564
[11] Sanger F Nicklen S Coulson A R 1977 DNA sequencing with chain-terminating inhibitors *Proc. Natl. Acad. Sci. USA.* 74(12) 5463-5467
[12] National Human Genome Research Institute 2016 https://www.genome.gov/
[13] The Human Brain Project (THBP) 2016 https://www.humanbrainproject.eu/
[14] Top 500 List 2016 http://www.top500.org
[15] NVIDIA, CUDA 2016 http://www.nvidia.com/object/cuda_home_new.html
[16] Smith T F 1990 The history of the genetic sequence databases *Genomics.* 6(4) 701-707
[17] Strasser B J 2008 Genetics. GenBank—Natural history in the 21st Century? *Science* 322(5901) 537-53
[18] Cochrane G Karsch-Mizrachi I Nakamura Y 2011 International Nucleotide Sequence Database Collaboration in The International Nucleotide Sequence Database Collaboration *Nucleic. Acids. Res.* 39 D15-D18
[19] Cochrane G Karsch-Mizrachi I Takagi T 2016 Sequence Database Collaboration in The International Nucleotide Sequence Database Collaboration *Nucleic. Acids. Res.* 44(4) D48-50
[20] Liu L Li Y Li S Hu N He Y Pong R Lin D Lu L Law M 2012 Comparison of next-generation sequencing systems *J. Biomed. Biotechnol.* 2012 251364
[21] Quail M A Smith M Coupland P Otto T D Harris S R Connor T R Bertoni A Swerdlow H P Gu Y 2012 A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers *BMC Genomics.* 13 341
[22] Linderman M D Nielsen D E Green R C 2016 Personal genome sequencing in ostensibly healthy individuals and the peopleSeq consortium *J. Pers. Med.* 6(2) E14