BioChemMed

Technium BioChemMed Vol. 2, Issue 4 pp.96-105 (2021)
ISSN: 2734-7990
www.techniumscience.com

Bioinformatics Storing Databases

Raghad M. M. Abed[1], Yusra A. Y. Al-Najjar[2*]

1 Medical Analysis, Al-Balqa'a Applied University As Salt – Jordan, ORCID ID: https://orcid.org/0000-0001-9390-5332, email: raghad.abed.borini@gmail.com

2 College of Computer Science, Taibah University, Al-Madinah Al-Munawarah – Saudi Arabia, ORCID ID: https://orcid.org/0000-0002-3369-4999, email: yalnajar@taibahu.edu.sa

*Corresponding author: Yusra Al-Najjar, yalnajar@taibahu.edu.sa

Abstract. An exceptional branch of data that requires huge databases has been shown lately from genome sequencing projects which is a field that employs computational approaches to answer biological questions. With this huge sequence of information that is available for researchers, bioinformatics plays a big role in studying basic medical-biological problems. The challenge that faces bioinformatical scientists is to help in discovering genes and designing molecular models, site-directed mutagenesis, and other experiments that reveal the unknown relationships concerning the structure and function of genes and proteins. This become a big challenge especially with the huge amount of data that is generated using the human genome and other systematic sequencing efforts up till now. Bioinformatics solves biological problems depending on available data. It is concerned with creating databases and predicting the outcome of lab experiments.

Keywords. Bioinformatics, chromosomes, databases, gene, genome, protein, DNA sequencing

1. Introduction:

Bioinformatics science is an integration of combining biology with computer science and information technology. It is also the statistical, mathematical, and computing methods used in solving biological problems using DNA and amino acid sequences and their related information. See figure 1. In general, we can say that bioinformatics is a management information system for molecular biology and has many practical applications.

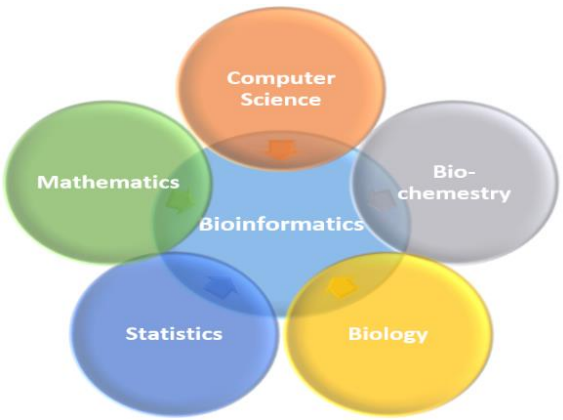

Fig. 1: fields involved in Bioinformatics

Technium BioChemMed Vol. 2, Issue 4 pp.96-105 (2021)
ISSN: 2734-7990
www.techniumscience.com

Bioinformatics differs from computational biology where bioinformatics concentrates on the structure, function, and analysis of genes and genomes and their products. It is an interdisciplinary field concerning developing new methods and software tools for understanding biological data. Bioinformatics is limited to genes sequencing and genomes besides their corresponding products which makes it considered computational molecular biology, whereas computational biology includes all biological fields that require computation and is restricted to the theoretical development of algorithms that are used for bioinformatics. Bioinformatic is a way that allows scientific researchers to access biological databases freely. This paper discusses how computer science and technology are employed for biology.

2. General Review

2.1. Why bioinformatics?

The goal of bioinformatics is to understand living cell in a better way and how it works on the molecular level by analyzing the raw molecular sequences and the structural data. Bioinformatics studies could generate a new vision for the cell. Figuring cell functions is better understood by the flow of generic information which is committed to the central dogma of biology in which the DNA is copied into RNA which is translated into protein [1]. See figure 2.

The functions of the cells are performed mainly by proteins whose capabilities are determined through their sequence. So, solving problems using sequences and structure proved to be helpful [2]. Bioinformatics is not important for biological genomes and basic nuclear only but has a great effect on many other fields of biological technology and biomedical sciences. Bioinformatics has many applications such as drug design which is based on knowledge, forensic DNA analysis, and agriculture biotechnology [2].

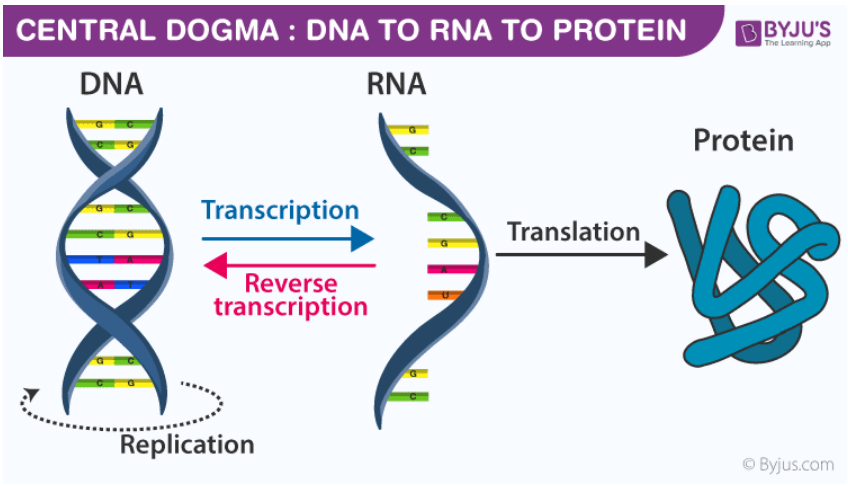

Fig. 2: Central Dogma [2]

2.2 Proteins

Proteins are macromolecules made of a smaller sequence of amino acids that differ in their structure and characteristics. Mainly humans need 20 different types of amino acids which could be classified into 2 groups - based on the ability of the cells to make - into essential amino acids (include “histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine”) and that the body can't produce thus must be obtained from the diet and the non-essential amino acids (include

Alanine, Arginine, Asparagine, Aspartic acid, Cysteine, Glutamic acid, Glutamine, Glycine, Proline, Serine, and Tyrosine) which the cell genome has the genetic recipe to make [3]. See figure 3.

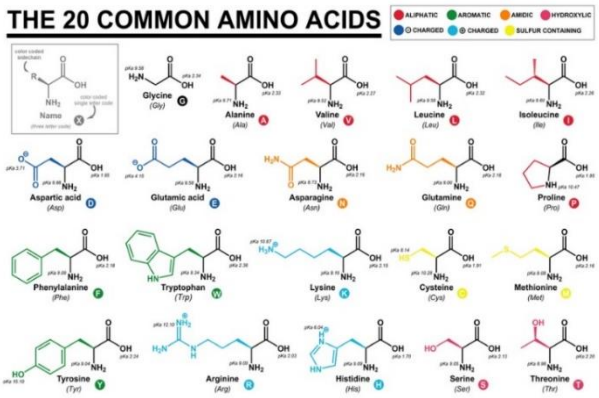

Fig. 3: Types of amino acids [3]

2.3 Amino acids

Amino acids are monomers of the proteins, each amino acid has a central carbon atom bound to a carboxyl group, amine group, hydrogen atom, and the R side chain that give the variation in amino acids. Some of These amino acids are encoded in the human genome meaning that they are produced from cells through transcription and translation. Each amino acid has a specific sequence in the DNA nucleotides code called codon (made of 3 nucleotides). See figure 4.

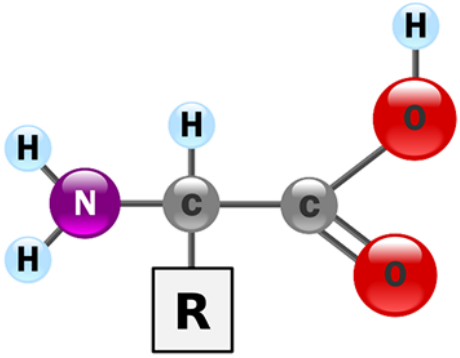

Fig. 4: Amino acid general structure [3]

2.4 Gene, genome, and DNS

The word genome describes the whole genetic information in our cells, this information is coded in the DNA sequence. DNA used 4 basic nucleotides to write the information (four types of chemical building blocks, adenine, thymine, cytosine, and guanine, with abbreviations A, T, C, and G.) the arrangement of these nucleotides provides the information about specific functional proteins, the arrange of specific nucleotides called a gene. During living, the cell enters different stages which are called cell-cycle to develop and multiply, normally the DNA strands are found unpacked within the nuclease but at a certain stage it needs to be packed using specific proteins (histones) and form defined chromosomes (an X shape structure) to ease the separation when cells multiply. Humans have 24 pairs of chromosomes within their cells representing the genome. See figure 5.

The information carried within the genome is varied it holds all the instructions and information that helped you to develop from a single cell into a full person. It guides the growth, helps organs to do their jobs, and repairs itself when damaged. The genome is unique to each organism [4].DNA in the cell is not a continuous long molecule; it is divided into parts of uneven lengths. These parts could be packed bundles at some points of the cell life cycle. These packed bundles are called chromosomes and they look like an X shape. Every creature has a specific number of chromosomes, e.g., a human has 46 chromosomes (23 pairs), plant rice has 24 chromosomes. See figure 6.

Technium BioChemMed Vol. 2, Issue 4 pp.96-105 (2021)
ISSN: 2734-7990
www.techniumscience.com

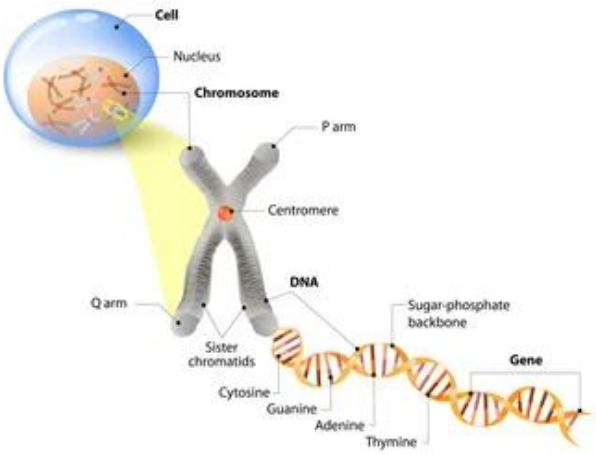

Fig. 5: Genetic arrangement in the cell [4]

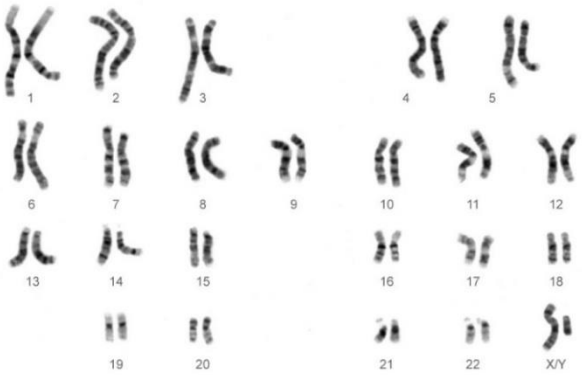

Fig. 6: pairs of chromosomes [4]

3. Computer Science Behind and Sequencing

In 1950, Alan Turing's Automatic Computing Engine (ACE) was introduced, three years before the publication of DNA structure. ACE was an early electronic serial stored-program computer. In 1970, EF Codd introduced a data model that proved to be essential for managing a huge amount of data. Later in 1977 DNA sequencing method by Frederick Sanger was published, and at the same time, computer scientists were about to announce the first draft of the human genome [5].

Proteins are an important issue to people and that is because the changes in proteins are the main reason for disease-causing mutations. Sanger was a chemist who said that biology and chemistry were both needed for Human Genome Project to succeed. And now the third strand is computer science [6]. Many disease-causing mutations occur in proteins which become of great interest. The DNA sequence codes for the amino acids were prepared earlier by reading the DNA sequence from a gel and then translating it to amino acids, but this was slow and boring. The computer started to show by 1976 and started to work closely with chemists and biologists in producing DNA sequence data. The operation of converting triplet DNA code into an amino acid sequence for printing was easily done using computer software [7].

After some years, labs all over the world started to produce more and more data sequences. And scientists started to get sequences from other labs to do comparisons. For example, sequences for beetles from different labs were compared to see how closely the beetles were related. Sequenced records were printed in a journal, but with the existence of computers, people started to share the data using networks. Michael Ashburner - who was a geneticist at Cambridge University - tried in 1980 to compare his sequence data with the data of Stanford University. He used the internet to show the problems he faced. But at that time there was a problem in communicating since network protocols used in the UK were different from the protocols used in the US [8]. The existence of a shared warehouse of data was a good

solution for sharing data problems. So, in 1981 the "European Molecular Biology Laboratory (EMBL) electronic library" for the nucleotide sequence data was established in Heidelberg. And because there was rapid growth for this warehouse, there was a need for a database management system [9].

Now, these databases are available, and the sequence data are freely available over the internet. Other databases were founded in the US and Japan, so people can share data from their personal computers. By completing the project of the Human Genome in the year 2003, for the first time, the sequencing for the human genome was accomplished. This sequencing was very expensive since it costs around one billion dollars and needed 13 years to finish. Nowadays, human genome sequencing could cost around one thousand dollars and could be accomplished in less than two days. The major factor in DNA sequencing technology besides scientists' knowledge was the advancement in computer science and engineering [10].

The amount of storage that computer engineering accomplished had a great effect in affording a space for storing such a huge data of DNA sequencing, in addition to a processing speed that came a long way since 2003. The methods that are used for DNA sequencing become more advanced than before. Previous processing for Human Genome Project – Sanger Sequencing – is implemented in reading a small fragment of DNA. Then these small pieces are put together to assemble the full genome [6]. The technique that is used nowadays is the Next-Generation Sequencing (NGS), it works in parallel. NGS process many micro-scale reactions simultaneously, and the result is 15,000 times more generated data per day more as a Sanger Sequencer [10]. See figure 7.

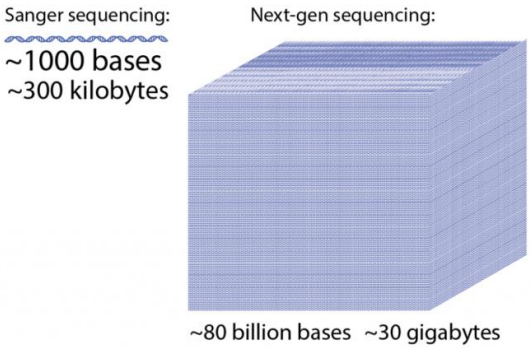

Fig. 7: Sanger versus next-generation sequencing [10]

Due to the progress in processing, sequencing the entire genome using an NGS machine per day became easy, whereas it required years using Sanger Sequencer. Increment in processing requires increment in storing capacity since the production of NGS machines could reach 1 terabyte of data per day. Memory upgrade RAM (Random Access Memory) also played a big role in storing generated information. Without this fast computational advancement, we cannot have this low cost of DNA sequencing.

3.1 Computational Biology

After sequencing the genome, the processing operation starts. In this operation, millions of data points will be analyzed seeking variations (mutations) within the genome. Analyzing the genome and detecting mutations could help in identifying the cause of several diseases.

To do this task manually would be impossible while using computational biology, doctors and scientists become able to detect mutations easily. To derive the meaning from the sequenced data, computational biologists use different techniques including pattern-matching algorithms, mathematical modeling, and image enhancement. They might also use simulation processes to find out how some biological systems will behave in different environments. The simulation shows the reaction of cancer cells according to different drug treatments which could lead to finding a cure. These simulations and models might one day lead to finding new treatments for several diseases [11].

3.2 Algorithms used in Bioinformatics

- Comparing sequences: a comparison between long sequences allowing for insertion, deletion, and mutation of symbols

- Constructing evolutionary (phylogenetic) trees: a comparison between sequence of different organisms, and building trees depending on similarities
- Detecting patterns in sequences: a making searches for genes in DNA or sequences of amino acids
- Inferring cell regulation: modelling the cell from the data
- Determining 3D structures: like inferring RNA shape from sequences, and protein shape from amino acids sequences
- Using scripts language: using a script on the internet to analyze the data [12].

4. Bioinformatics Databases and Cloud Storage

Due to the increment in genetic information, different cloud storage platforms and online databases and cloud storage have appeared that gave the chance for these data to be analysed and accessed. These platforms provided required storage space for about 1 in 25 Americans' DNA to be stored and referenced when needed. Storing data using cloud-based databases allowed global teams to work at the same time and on the same data and find solutions for problems that face them.

In the field of law, DNA databases play a big role. In April 2018 in the USA, the Golden State Killer was arrested due to genetic information that a relative of him shared on the GEDMatch genealogy website [13]. Many USA states besides other countries collect DNA information from different crime scenes and arrest people for different crimes depending on the criminological database. Despite the DNA database's role in catching many criminals and sending them behind bars, there are still many questions concerning privacy [14]. The database is an archive that is used to store and organize data making it easy to be retrieved using search criteria. Databases are developed to classify data made of records that are structured in a way that makes it easy for retrieval. The software that manages databases is called a database management system.

The database that contains biological science such as molecular biology and bioinformatics is called a biological database. These databases work like other databases, they allow indexing data, as well as removing redundancy. They are the central component of bioinformatics [15]. Biological data differs from any other data, they are complicated, have many exceptions, vast and incomplete. So, many databases were established and interpreted to make sure that the results are clear and unambiguous. A good database should contain up-to-date information. Biological databases allow scientists to retrieve biological sequences, structure, metabolic interactions, functional relationships, molecular actions, motifs homologous, and protein families [2]. PubMed is one of the most known databases in biomedical literature. It contains the abstracts and the text of articles for about 4000 journals [2]. Modern biological research especially genomic studies require databases. There are two types of biological databases, primary database, and secondary database. The primary database holds the sequence and structural information, whereas the secondary database is derivative from analysing the primary data. The secondary database is important for controlling protein functions [16].

Examples for some primary biological databases are [17]:

- 'GeneBank (Generic Sequence Databank)': this bank contains nucleotide sequences data. It is an ASCII text file that could be read by humans as well as computers. It is an open-access database.
- 'EMBL (European Molecular Biology Laboratory)': is a DNA and RNA sequences database that is collected from patient offices and submitted by researchers. This database is maintained by EBI (European Bioinformatics Institute).
- Swiss-Port: it is a curated protein sequence database.

- And examples for secondary biological databases are:
- Motif Databases:
  - A protein sequence motif is a set of conserved amino acid residues. These are vital for protein functions.
  - PRINT: is a database for protein fingerprint
- Domain Database:
  - A protein domain is a compact unit that forms a three-dimensional structure
  - SMART: a sensitive tool for domain identification
- 3D Structure Databases:

- ‘PDB (Protein Data Bank)’: a database for the 3D structure of biological macromolecules determined by X-ray
- ‘SCOP (Structure Classification of Protein database)’: classifies protein 3D structure

- Gene Expression Databases:
  - ‘GEO (Gene Expression Omnibus)’: selected online resource and gene expression molecular huge store to browse, query, and retrieve the gene.
  - ‘GXD (Gene Expression Database)’: is a community resource for expressing gene information

This is just a selected group of the available databases that keep on growing over days.

4.1 Turning DNA sequences into protein sequences

The process of turning DNA into mRNA to Protein could be figured out as decoding the instructions concerning proteins, including mRNA as well as tRNA. Once the sequence of amino acids is known, it could be translated into the corresponding protein sequence using genetic codes. This is the same way that cells generate protein sequences and is called translating DNA into Protein. The genetic code in the Table shows how the 4-nucleotide sequence is linked to give a set of 20 amino acids. The table describes the roles by which the coded data is translated to proteins. The diagram shows the DNA codon table as a chart [18]. See figure 8.

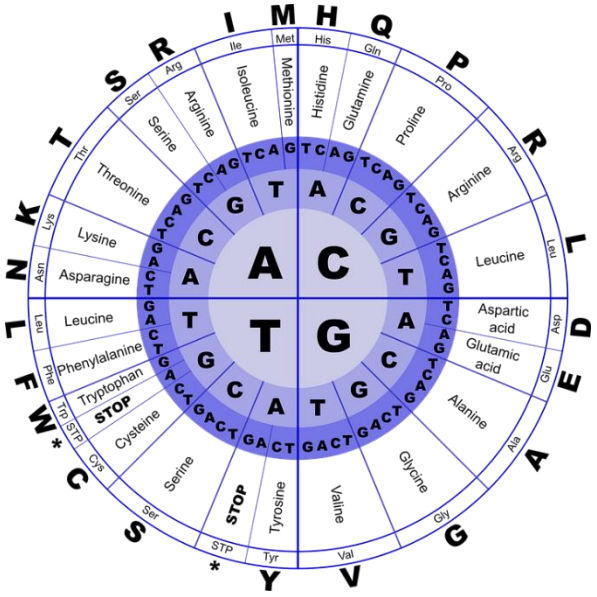

Fig. 8: Genetic Code Chart [22]

4.1.1How to translate DNA to Protein using Genetic Code Chart

- Get the DNA string, e.g. :
  ‘ATGGAAGTATTTAAAGCGCCACCTATTGGGATATAAG’
- Read the sequence of 3 nucleotides (triple) at a time, e.g.:
  ‘ATG GAA GTA TTT AAA GCG CCA CCT ATT GGG ATA TAA G...’
- Use a genetic code chart for reading the amino acid that corresponds to the triplet (codons).
  - Start from the center that represents the first character of the triblet
  - The second circle represents the second character
  - The third circuit represents the last character
- Now, the protein sequence that corresponds to the DNA sequence is ready where:
  - ATG = M, GAA = E, GTA = V, and so on…, so, the results will be like:
    ‘M E V F K A P P I G I STOP’

TAA, TAG, and TGA stop the translation process, so they are called the termination signals. This process of translation could be done using any of the programming languages, the following is a simple python program that could be used to translate DNA into protein.

4.1.2 DNA to Protein Translation using Python

- Starting with a text file containing the sequence of DNA as mentioned above,

- Open the text file
- Read the data stored in it
- Start translating according to the given algorithm

Once the start of protein coding is determined in the DNA sequence, the software starts generating the corresponding protein sequence. DNA could be processed as a virtual protein sequence using a simple software. See the following code:

```
# open the file called dna.txt and read the DNA sequence
file = open('dna.txt', 'r')
dna = file.read()
print "The DNA Sequence is: ", dna
# DNA translation table
protein = {"TTT" : "F", "CTT" : "L", "ATT" : "I", "GTT" : "V",
"TTC" : "F", "CTC" : "L", "ATC" : "I", "GTC" : "V",
"TTA" : "L", "CTA" : "L", "ATA" : "I", "GTA" : "V",
"TTG" : "L", "CTG" : "L", "ATG" : "M", "GTG" : "V",
"TCT" : "S", "CCT" : "P", "ACT" : "T", "GCT" : "A",
"TCC" : "S", "CCC" : "P", "ACC" : "T", "GCC" : "A",
"TCA" : "S", "CCA" : "P", "ACA" : "T", "GCA" : "A",
"TCG" : "S", "CCG" : "P", "ACG" : "T", "GCG" : "A",
"TAT" : "Y", "CAT" : "H", "AAT" : "N", "GAT" : "D",
"TAC" : "Y", "CAC" : "H", "AAC" : "N", "GAC" : "D",
"TAA" : "STOP", "CAA" : "Q", "AAA" : "K", "GAA" : "E",
"TAG" : "STOP", "CAG" : "Q", "AAG" : "K", "GAG" : "E",
"TGT" : "C", "CGT" : "R", "AGT" : "S", "GGT" : "G",
"TGC" : "C", "CGC" : "R", "AGC" : "S", "GGC" : "G",
"TGA" : "STOP", "CGA" : "R", "AGA" : "R", "GGA" : "G",
"TGG" : "W", "CGG" : "R", "AGG" : "R", "GGG" : "G" }
protein_sequence = ""
# Generate protein sequence
for i in range(0, len(dna)-(3+len(dna)%3), 3):
    if protein[dna[i:i+3]] == "STOP" :
        break
    protein_sequence += protein[dna[i:i+3]]
# Print the protein sequence
print "Protein Sequence: ", is the protein_sequence
# end of the program
```

The output of the program is like the following:

DNA Sequence: ATGGAAGTATTTAAAGCGCCACCTATTGGGATATAAG
Protein Sequence: MEVFKAPPIGI

Programming languages that are used for bioinformatics are mostly PHP, Java, Pearl, C, and C++ for developers, and R, Python for analysis [19].

4.2 Privacy and Security

There is a risk on any data stored in the cloud including the sequenced DNA. Cloud stored data are open and can be accessed freely, which makes this huge data - about a million profiles for DNA on GEDMatch - available for anyone. These DNA profiles concern half of the USA population.

Since there is a great number of hackers, even private databases such as 23andMe could be hacked nowadays. People are afraid that their genetic information could be used for a crime or even cloned if stolen. There is a big challenge provided for securing DNA due to its nature. To accomplish this, this requires the need from computer science experts [20].

4.3 DNA Future Storage

Technological advances facilitated the understanding of DNA, but the opposite might also be true. The data is growing fast, over 2.5 exabytes per day, where the exabyte is equal to one billion gigabytes. The

work is still going on to find new techniques for storing this huge amount of data, some computer scientists even turned to a weird place: the DNA. DNA is capable of storing and encoding information using DNA sequences. This encourages big companies such as Microsoft's to provide DNA storage for storing data. In the year 2016, scientific researchers from Microsoft and the University of Washington managed to store 200 megabytes of data using synthetic DNA. This medium for storing DNA sequences are artificial genes which scientists created in the laboratory [21].
Nowadays researchers are looking to store more data and find ways for quick access in a time where synthetic DNA is important field that could make data storage using synthetic DNA available in the coming future. Figure 9 shows synthetic DNA as storing media that could store data much more efficiently in a small area. This technique can store around 1250 terabytes of data in one cubic millimeter, whereas the flash drive or the hard disk stores around 1.25 gigabytes per cubic millimeter. This means that DNA can store up to 1,000,000 times than other storage devices [21].

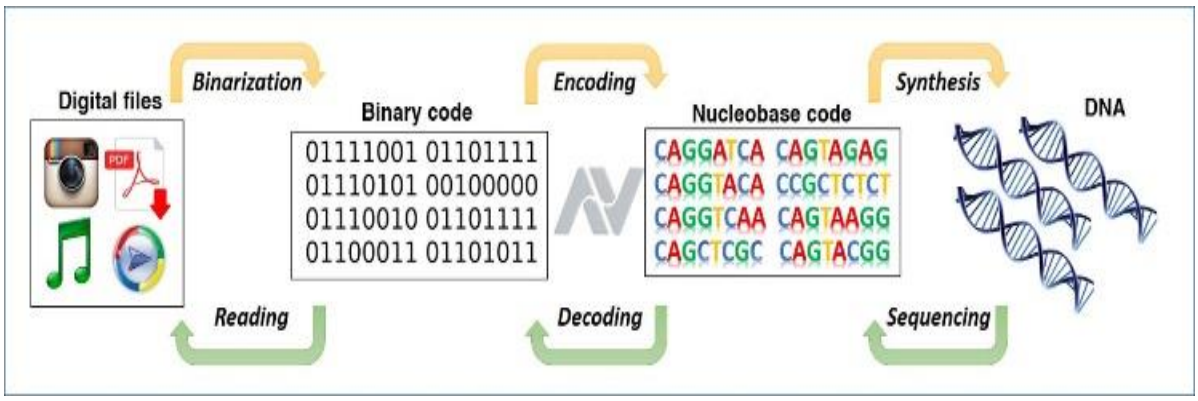

Fig. 9: Storing data using synthetic DNA [23]

5. Conclusion

With the existence of a huge amount of DNA data showing every day, typical storing devices became insufficient. Bioinformatics is an essential field for biological sciences that provides storing techniques for DNA sequences. Information could not be stored or retrieved so quickly and easily without the interference of technology, which is still developing new storage techniques. Bioinformatics continues to grow to include other topics and areas in biology, chemistry, etc., using computations, statistics, algorithms, and most important: databases.

References

[1] J. J. Li and M. D. Biggin, "Statistics requantitates," Science (80-.)., vol. 347, no. 6226, pp. 1066–1067, 2015.
[2] J. Xiong, Essential Bioinformatics. United Kingdom: Cambridge University Press, 2006.
[3] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and P. W., Molecular Biology of the Cell, 4th editio. New York, 2002.
[4] Evelyn Fox Keller, "Genes, Genomes, and Genomics," Biol. Theory, vol. 6, pp. 132–140, 2012.
[5] B. J. Copeland, Alan Turing's Automatic Computing Engine: The Master Codebreaker's Struggle to build the Modern Computer. 2008.
[6] A. O. Stretton, "The First Sequence: Fred Sanger and Insulin," Genetics, vol. 162, no. 2, pp. 527–532, 2002.
[7] H. P. Langtangen and G. K. Sandve, "Illustrating Python via Bioinformatics Examples Basic Bioinformatics Examples in Python," Online, pp. 1–46, 2012.
[8] J. Mullins and B. J. M. McKay, "International society for computational biology honors Michael Shburner and Olga Troyanskaya with top Bioinformatics/Computational biology awards for 2011," PLoS Comput. Biol., vol. 7, no. 6, 2011, doi: 10.1371/journal.pcbi.1002081.
[9] G. Stoesser, M. A. Moseley, J. Sleep, M. McGowran, M. Garcia-Pastor, and P. Sterk, "The EMBL nucleotide sequence database," Nucleic Acids Res., vol. 26, no. 1, pp. 8–15, 1998, doi: 10.1093/nar/26.1.8.
[10] J. K. Kulski, Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. 2015.

[11] Gautam B. Singh, Fundamentals of Bioinformatics and Computational Biology. 2015.
[12] C. Jou, "Introduction to Bioinformatics," 2014.
[13] G. C. Michael (Doc) Edge, "How lucky was the genetic investigation in the Golden State Killer case?," 2019, doi: 10.1101/531384.
[14] L. S. Lewis, "THE ROLE GENETIC INFORMATION PLAYS IN THE CRIMINAL JUSTICE SYSTEM," 2005.
[15] B. B. Noor Ahmad Shaik, Ramu Elango, Muhummadh Khan, "Introduction to Biological Databases," Springer, Cham, vol. 1, pp. 19–27, 2019, doi: https://doi.org/10.1007/978-3-030-02634-9_2.
[16] A. Tyagi, "Biological Databases," 2020.
[17] Pascale Anderle, Pascale Anderle, Manuel Duval, Manuel Duval, Sorin Draghici, S. D. et al, "Gene Expression Databases and Data Mining" Biotech. Suppl, vol. 3, pp. 36–44, 2003, doi: 10.2144/mar03anderle.
[18] W. B. Suzanne Clancy, "Translation: DNA to mRNA to Protein," Scitable by Nat. Educ., vol. 1, no. 1, 2018, [Online]. Available: https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/.
[19] M. R. G. Mathieu Fourment, "A comparison of common programming languages used in bioinformatics," BMC Bioinformatics, vol. 9, no. 82, 2008, doi: 10.1186/1471-2105-9-82.
[20] G. M. Nina F de Groot, Britta C van Beers, "Commercial DNA tests and police investigations: a broad bioethical perspective," J. Med. Ethics, vol. 0, pp. 1–8, 2021, doi: 10.1136/medethics-2021-107568.
[21] Potomac Institute for Policy Studies, "The Future of DNA Data Storage," 2018. [Online]. Available: https://potomacinstitute.org/images/studies/Future_of_DNA_Data_Storage.pdf.
[22] M. Mitra, "Elements of RNA , its Techniques and Applications," no. January, 2019, doi: 10.5281/zenodo.2552818.
[23] Ashutosh Viramgama, "DNA Data Storage – Synthetic DNA – Future Of Storage," 2018. https://ashutoshviramgama.com/dna-data-storage-synthetic-dna-future-of-storage/.