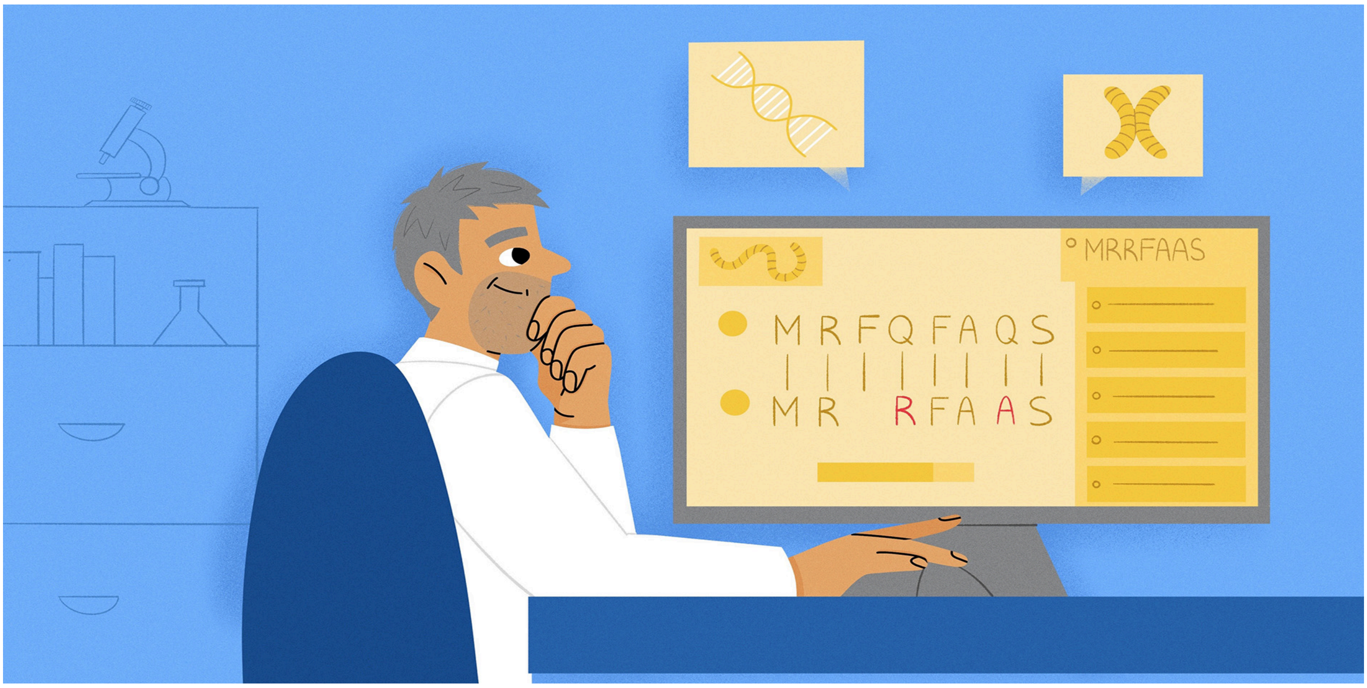
HUMAN HEALTH
Published: 31 January 2024
doi: 10.3389/frym.2024.1266091



BIOINFORMATICS: ANALYZING DATA FROM LIVING THINGS

***Carmine Fruggiero*[†], *Gaetano Aufiero*[†] *and Nunzio D'Agostino*[*]**

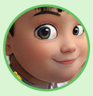
Department of Agricultural Sciences, University of Naples Federico II, Portici, Italy

[†] *These authors have contributed equally to this work*

YOUNG REVIEWERS:

ADI
AGE: 12

DIYA
AGE: 12


JOHNSON
AGE: 14

Do you know how much information is hidden inside your own cells? It is a huge amount. Are you curious about how scientists decipher and handle this large volume of data? Computer science and mathematics have helped biologists to create a new science to analyze, organize, and understand biological data: bioinformatics. Bioinformatics allows us to manage huge amounts of biological data and make sense of it. In other words, bioinformatics allows us to explore the mysteries of life and find answers to complex questions about the functioning of living beings. In this article, we are happy to tell you what this fascinating new field is and what it is used for. We think you will like it!

BIOINFORMATICS: INTERPRETING BIOLOGICAL SYMBOLS

You might be surprised to learn that each cell of your body contains a "handbook" with the instructions to build and maintain your entire body—all kept in a space of 4–6 µm (a µm is one millionth of a meter)! This information is stored in the DNA, which is found in almost all living organisms and is the "universal language" of life. In human cells, this language consists of more than six *billion* "letters". How is it possible for scientists to study and interpret the enormous amount of information written in the DNA? This difficult task led to the birth of a new science, called **bioinformatics** [1].

BIOINFORMATICS

The science of storing large amounts of complex biological data and analyzing it to find new insights.

THE HANDBOOK OF LIFE

Scientists call the DNA handbook, or the entire set of DNA instructions found within each cell, the **genome** (Figure 1). The handbook is divided into "chapters" known as chromosomes. Each species has a specific number of chapters, or, in other words, each species has a set number of chromosomes. For example, humans have 23 pairs of chromosomes and tomatoes have 12 pairs. DNA is a long molecule made up of a sequence of four "letters" corresponding to four chemical compounds: adenine (A), guanine (G), cytosine (C), and thymine (T). Here is an example of what a DNA sequence looks like: ATGGTCCCATGCTAGCTAGCTATCGATGCTACGTACGTAG CATAAATCGCGATAGCTA.

GENOME

The set of DNA instructions found in a cell.

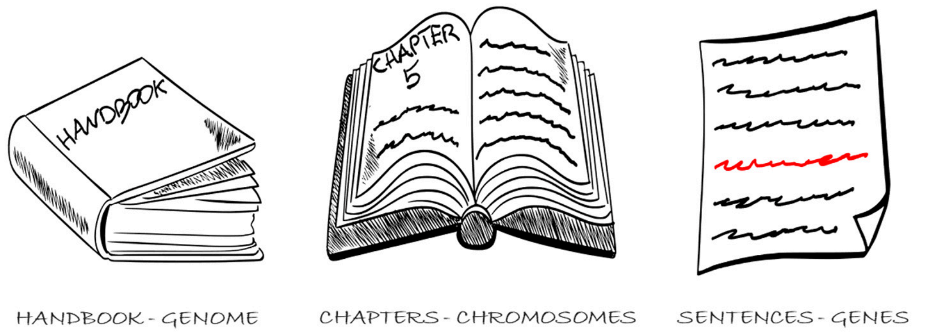


Figure 1

Figure 1

You can think of your genome as a handbook containing all the instructions needed to create your body and allow it to function. Each "chapter" of this handbook is a chromosome, and humans have 23 pairs of them. Within each chapter, the sentences containing the actual instructions are called genes (red).

Combinations of those four letters are used by cells to know how to behave. Everyone's genome is a little bit different, and the sum of these differences leads to the unique way that you look and can even influence how you think and act.

So, how can scientists read and understand the genome? "Reading" a genome means knowing the sequence of letters that make it up. For this purpose, scientists use a technique called **genome sequencing** (you can read more about genome sequencing in this Frontiers for Young Minds article). But knowing just the sequence is like having a book written in an unknown language that must be deciphered. First, bioinformaticians seek to identify specific sentences within the genome, referred to as **genes**. Each gene contains the information needed to make a specific **protein**. In fact, every time the cell

GENOME SEQUENCING

The process that determines the order of the chemical building blocks, called bases, that make up the DNA molecule.

GENES

Functional units of heredity that carry the information that specifies the characteristics transmitted from generation to generation.

PROTEINS

Large and complex molecules that play many important roles in living organisms.

wants to do something, it goes back to the handbook, selects the necessary lines of text (i.e., the genes) and uses them to build proteins. Proteins are the workers of the body, each with its own special job to do. Proteins can help to move things around inside cells, can help to build cells, and more. Proteins can interact with each other or with other chemical compounds to build muscles, hair, and nails, for example.

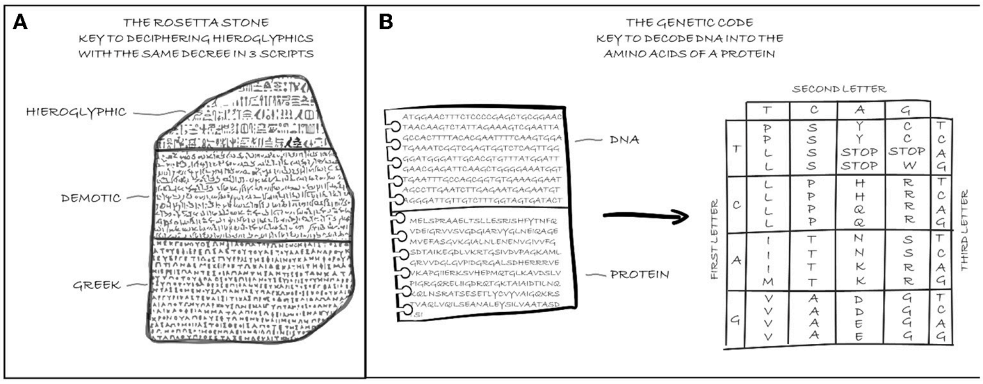
AMINO ACIDS

Molecules used by all living things to make proteins.

Proteins are molecules made up of various assortments of 20 chemical compounds, the **amino acids**. An amino acid can be thought of as a building block, such as a LEGO brick, and a protein as a particular LEGO masterpiece. Depending on how you assemble the blocks, you can get specific creations. Similarly, a cell assembles amino acids to obtain a particular protein (you can read more about proteins in this Frontiers for Young Minds article). Every time the cell needs a protein, it translates the instructions encoded in the corresponding gene into a particular sequence of amino acids. To perform this translation, the cell applies a set of rules known as the **genetic code**. The genetic code allows scientists to decode the DNA sequence just as the three translations on the Rosetta Stone helped people to understand Egyptian hieroglyphics (Figure 2) [2]. Your genome contains around 20,000 genes that can encode at least 80,000 proteins that, together, regulate your body's function.

GENETIC CODE

The set of rules used by cells to translate information encoded within DNA into proteins.


A
THE ROSETTA STONE
KEY TO DECIPHERING HIEROGLYPHICS
WITH THE SAME DECREE IN 3 SCRIPTS
HIEROGLYPHIC
DEMOTIC
GREEK
B
THE GENETIC CODE
KEY TO DECODE DNA INTO THE
AMINO ACIDS OF A PROTEIN
ATGGAACTTTCTCCCCGAGCTGCGGAAC
TAACAAGTCTATTAGAAAGTCGAATTA
GCCACTTTTACACGAATTTTCAAGTGGA
TGAAATCGGTCGAGTGGTCTCAGTTGG
GGATGGGATTGCACGTGTTTATGGATT
GAACGAGATTCAAGCTGGGGAAATGGT
TGAATTTGCCAGCGGTGTGAAAGGAAT
AGCCTTGAATCTTGAGAATGAGAATGT
AGGGATTGTTGTCTTTGGTAGTGATACT
MELSPRAAELTSLLESRISHFYTNFQ
VDEIGRVVSVGDGIARVYGLNEIQAGE
MVEFASGVKGIALNLENENVGIVVFG
SDTAIKEGDLVKRTGSIVDVPAGKAML
GRVVDGLGVPIDGRGALSDHERRRVE
VKAPGIIERKSVHEPMQTGLKAVDSLV
PIGRGQRELIIGDRQTGKTAIAIDTILNQ
KQLNSRATSESETLYCVYVAIGQKRS
TVAQLVQILSEANALEYSILVAATASD
SI
DNA
PROTEIN
SECOND LETTER
FIRST LETTER
THIRD LETTER

	T	C	A	G	
T	P P L L	S S S S	Y Y STOP STOP	C C STOP W	T C A G
C	L L L L	P P P P	H H Q Q	R R R R	T C A G
A	I I I M	T T T T	N N K K	S S R R	T C A G
G	V V V V	A A A A	D D E E	G G G G	T C A G

Figure 2

Figure 2

(A) The Rosetta Stone was the key to deciphering hieroglyphics. **(B)** The genetic code is the key to decoding the DNA sequence to know which protein it codes for.

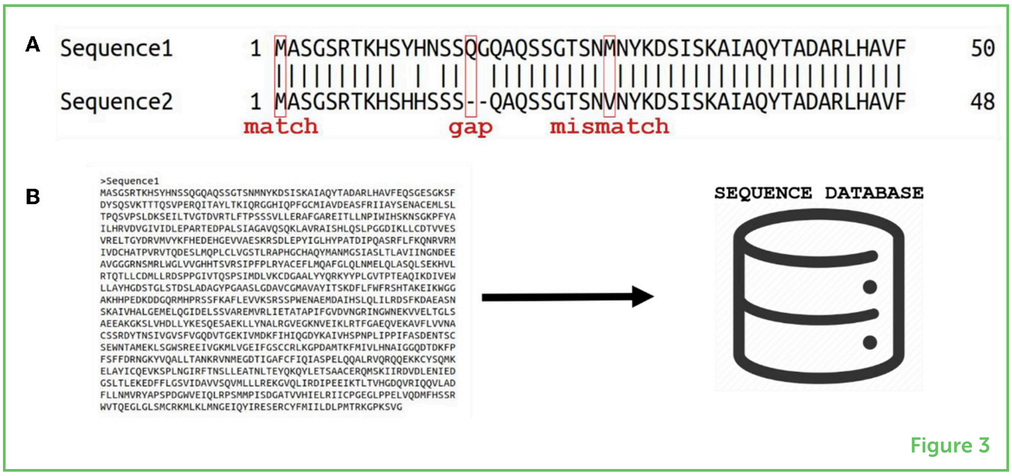
In summary, both genes and proteins can be represented by a string of characters, and the exciting role of bioinformaticians is to study these sequences to decipher their secrets. If this interests you and you cannot wait to become a bioinformatician, do not worry! All you need is a laptop to get started! The first thing bioinformaticians do is to get familiar with text files on computers where DNA or protein sequences are stored. Then, with the help of some software, they can navigate, decipher, and interpret this world made of letters.

THE BASIC TASK IN BIOINFORMATICS: SEQUENCE COMPARISON

BIOSEQUENCE

A series of biological molecules, such as DNA and proteins.

Scientists call DNA and protein sequences **biosequences**. Your own biosequences can be compared with all other known biosequences to infer a common evolutionary descent or a common structural function. But what does it mean to compare biosequences? It means using computer programs to line them up in pairs, to match as many letters as possible (Figure 3). The score of this alignment is given by the sum of the matches (matching symbols), the mismatches (non-matching symbols), and the gaps (additions or deletions relative to the other sequence). In other words, aligning two sequences means to identify regions of similarity [3]. Protein sequences can be compared, too. There are bioinformatics tools that allow scientists to translate a DNA sequence into a protein sequence, and then proceed with the alignment and analysis.



```
Sequence1    1 MASGSRTKHSYHNSSQGQAQSSGTSNMNYKDSISKAIAQYTADARLHAVF    50
               |||||||||| | ||  ||||||||| |||||||||||||||||||||||
Sequence2    1 MASGSRTKHSHHSSS--QAQSSGTSNVNYKDSISKAIAQYTADARLHAVF    48
```

```
>Sequence1
MASGSRTKHSYHNSSQGQAQSSGTSNMNYKDSISKAIAQYTADARLHAVFEQSGESGKSF
DYSQSVKTTTQSVPERQITAYLTKIQRGGHIQPFGCMIAVDEASFRIIAYSENACEMLSL
TPQSVPSLDKSEILTVGTDVRTLFTPSSSVLLERAFGAREITLLNPIWIHSKNSGKPFYA
ILHRVDVGIVIDLEPARTEDPALSIAGAVQSQKLAVRAISHLQSLPGGDIKLLCDTVVES
VRELTGYDRVMVYKFHEDEHGEVVAESKRSDLEPYIGLHYPATDIPQASRFLFKQNRVRM
IVDCHATPVRVTQDESLMQPLCLVGSTLRAPHGCHAQYMANMGSIASLTLAVIINGNDEE
AVGGGRNSMRLWGLVVGHHTSVRSIPFPLRYACEFLMQAFGLQLNMELQLASQLSEKHVL
RTQTLLCDMLLRDSPPGIVTQSPSIMDLVKCDGAALYYQRKYYPLGVTPTEAQIKDIVEW
LLAYHGDSTGLSTDSLADAGYPGAASLGDAVCGMAVAYITSKDFLFWFRSHTAKEIKWGG
AKHHPEDKDDGQRMHPRSSFKAFLEVVKSRSSPWENAEMDAIHSLQLILRDSFKDAEASN
SKAIVHALGEMELQGIDELSSVAREMVRLIETATAPIFGVDVNGRINGWNEKVVELTGLS
AEEAKGKSLVHDLLYKESQESAEKLLYNALRGVEGKNVEIKLRTFGAEQVEKAVFLVVNA
CSSRDYTNSIVGVSFVGQDVTGEKIVMDKFIHIQGDYKAIVHSPNPLIPPIFASDENTSC
SEWNTAMEKLSGWSREEIVGKMLVGEIFGSCCRLKGPDAMTKFMIVLHNAIGGQDTDKFP
FSFFDRNGKYVQALLTANKRVNMEGDTIGAFCFIQIASPELQQALRVQRQQEKKCYSQMK
ELAYICQEVKSPLNGIRFTNSLLEATNLTEYQKQYLETSAACERQMSKIIRDVDLENIED
GSLTLEKEDFFLGSVIDAVVSQVMLLLREKGVQLIRDIPEEIKTLTVHGDQVRIQQVLAD
FLLNMVRYAPSPDGWVEIQLRPSMMPISDGATVVHIELRIICPGEGLPPELVQDMFHSSR
WVTQEGLGLSMCRKMLKLMNGEIQYIRESERCYFMIILDLPMTRKGPKSVG
```

Figure 3

Figure 3

(A) Pairwise alignment between two protein sequence. The identical (matches) and different (mismatches) amino acids are highlighted in the two sequences as well as the gaps (no matches). **(B)** A new sequence, defined as a "query", can be compared against a whole database of known sequences. This operation, known as database searching, returns a list of pairwise alignments and can tell scientists whether the sequence is already known or how similar it is to known sequences.

Instead of doing pairwise comparison, bioinformaticians can also use tools that allow them to search a database for a specific sequence. For example, if they have a sequence and want to know what type of protein it codes for or what its function might be, specific software can compare that new biosequence with all known biosequences stored in specific databases. The software will return pairwise alignments of all known sequences that are similar to the unknown sequence. The greater the similarity, the more likely the genes or proteins perform the same functions inside the cell.

Imagine that you have just obtained a new sequence from an experiment. You know nothing about its properties or what its role might be in a living cell. So, the first thing you can do is compare your unknown sequence (sequence X) with a large database of sequences with known properties. If sequence X is similar to sequence Y, which is known to move chemical compounds in or out of the cell, for example, then you can assume that sequence X might also be involved in the transport of specific molecules. Overall, sequence comparison is the first and most informative step in the analysis of newly determined sequences.

BIOINFORMATICS: AN EVER-EVOLVING SCIENCE

Since 1978, when the term "bioinformatics" was introduced, the field has seen explosive growth— especially in the last 20 years. Technology has evolved to make bioinformatics even more powerful. Data are securely stored in supercomputers and organized into databases. Data are also analyzed by increasingly efficient computer programs developed to cope with a growing volume of information, often called "big data". The use of supercomputers and some forms of artificial intelligence can help to extract even more meaningful information from biological data. As bioinformatics continues to develop, this science will help us to better understand how living organisms work and to plan all kinds of exciting ways to help people, from defending humans against diseases to helping plants adapt to climate change.

ACKNOWLEDGMENTS

The authors thank all members of the GenoPom Lab for stimulating discussions during the preparation of this article. Special thanks go to our little friend (13 years old) Elizabeth Villmer for giving us feedback from a younger audience's perspective.

REFERENCES

1. Luscombe, N. M., Greenbaum, D., and Gerstein, M. 2001. What is bioinformatics? An introduction and overview. *Yearbook Med. Informat.* 10:83–100. doi: 10.1055/s-0038-1638103
2. Caskey, C. T., and Leder, P. 2014. The RNA code: Nature's Rosetta Stone. *Proc. Natl. Acad. Sci. U. S. A.* 111:5758–9. doi: 10.1073/pnas.1404819111
3. Rosenberg, M. S. 2009. *Sequence Alignment: Methods, Models, Concepts, and Strategies*. Berkeley, CA: University of California Press.

SUBMITTED: 24 July 2023; **ACCEPTED:** 15 January 2024;
PUBLISHED ONLINE: 31 January 2024.

EDITOR: Ornella Cominetti, Nestlé Research Center, Switzerland

SCIENCE MENTORS: Anita Singh and Renee W. Y. Chan

CITATION: Fruggiero C, Aufiero G and D'Agostino N (2024) Bioinformatics: Analyzing Data From Living Things. Front. Young Minds 12:1266091. doi: 10.3389/frym.2024.1266091

CONFLICT OF INTEREST: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

YOUNG REVIEWERS

ADI, AGE: 12

I am a Lego enthusiast who loves science and math. I also enjoy making 3D prints of airplanes and gears. The most fun thing for me to do when not coding or 3D printing is to read about new topics in science and technology. I enjoy immersion in projects and videos of these topics that allow him to learn more about them in a fun way. I play percussion instruments and am just learning a Tenor Saxophone.

DIYA, AGE: 12

I am a proud twin who loves to listen to music and enjoy art. I am passionate about helping people who do not have access to resources that help enhance STEM learning. I help run an engineering summer camp every year and visit countries where I can share my learning. I am also a dancer and love to spend time with my fluffy dog.

JOHNSON, AGE: 14

I began serving as a Young Reviewer for this journal when I was 10 years old. Currently, I am a Grade 10 student studying Science Stream and also a YouTuber. My experience as a young reviewer of scientific articles has broadened my perspective. It is amazing to see how science in fiction translates into real-life experimental outcomes.

AUTHORS

CARMINE FRUGGIERO

Carmine Fruggiero graduated with a degree in agro-environmental and food biotechnology from the University of Naples Federico II with full marks. At present, he is a Ph.D. student in bioinformatics.

GAETANO AUFIERO

Gaetano Aufiero graduated with a degree in agro-environmental and food biotechnology from the University of Naples Federico II with full marks. At present, he is a Ph.D. student working on the analysis of biosequences.

NUNZIO D’AGOSTINO

Nunzio D’Agostino is an associate professor of bioinformatics and genomics in the Department of Agricultural Sciences, University of Naples Federico II. His main research interest is bioinformatics applied to the investigation of plant genomes and genetic improvement of plant species. In particular, his research activity is focused on the analysis of next-generation sequencing data and on the development of strategies, methods, and tools for the management of -omics data. *nunzio.dagostino@unina.it