Algerian and Democratic and People's Republic

Ministry of Higher Education and Scientific Research

Mohamed EL Bachir El Ibrahimi University of BBA

Faculty of Mathematics and Computer Science

Computer Science department

MEMOIR

Present for graduation

Computer science's master

Specialty**: Network and Multimedia**

THEME

Identification of DNA-protein relationships using association rules

Presented by:

- CHADI Nadjet **- SLAMA Ratiba**

Supported closed on: 31/09/2020

In front of the jury composed of:

President**: ATTIA Abdelouahab**	MCA à L'U.EBachir El Ibrahimi-BBA.
Examiner **: NAILI Makhlouf**	MAA à L'U.EL Bachir El Ibrahimi-BBA.
Supervisor : **Dr ZAOUACHE Djaafar**	MCB à L'U.EL Bachir El Ibrahimi-BBA.

Promotion: 2019/2020

Dedication

I dedicate this modest work:

To my dear parents, source of life, love and affection

To my dear Brother Ayoub source of joy and happiness

To all my family, a source of hope and motivation

To all my friends, especially BAHLOUL AMINA

In Ratiba, dear friend before being a pair

To you dear reader

Nadjet

Dedication

I dedicate this modest work:

To the person dearest to me, the one who left life, to my mother who sacrificed everything for her children, who ensured our education who without her I would not be there.

To dear father source of life, love and affection

To my dear Brothers and sisters and their children, source of joy and happiness

To all my family, a source of hope and motivation

To all my friends, especially Siham, Sakina and Maria

In Nadjet, dear friend and sister before being a pair

To you dear reader

Ratiba

KNWLDGEMENT

First of all we thank the
only god

We also wish to express our great gratitude to my

supervisor Dr Zaouache Djaafer.

Who support us during all the stages of memory

by showing a great interest in our work.

Our thanks also go to the members of the jury for

agreeing to evaluate my research work.

To all those who have helped us from near or far,

by gesture, word or advice, I say thank you.

Without forgetting all those who contributed to us gaining

knowledge and reaching our level

Summary

Chapter II: Alignment of biological sequences

Chapter III: Association Rules

Chapter IV proposition and realization

Figures lists

Arrays List

Abstract:

In biology, cells perform multiple functional processes related mainly to heredity, and among these processes we mention transcription and translation, which target the links between protein and deoxyribonucleic acid, specifically TF factors and TFBS sites, which play a fundamental role in these processes, and with the quantitative development of informatics in terms of storage capacity, it became possible to store it, so we targeted in this work, we explore the relationship between these factors by using data mining techniques with the use of the algorithm of '**Apriori '** and **'Eclat'.**

Résumé:

Dans la biologie, les cellules exécutent de multiples processus fonctionnels liés principalement à l'hérédité, et parmi ces processus, nous mentionnons la transcription et la traduction, qui ciblent les liens entre les protéines et l'acide désoxyribonucléique, en particulier les facteurs FT et les sites TFBS, qui jouent un rôle fondamental dans ces processus, nous explorons la relation entre ces facteurs en utilisant des techniques de data mining avec l'utilisation de l'algorithme Eclat et Apriori.

تلخيص

في علم الأحياء ، تؤدي الخلايا عمليات وظيفية متعددة تتعلق أساسًا بالوراثة ، ومن بين هذه العمليات نذكر النسخ والترجمة ، والتي تستهدف الروابط بين البروتين وحمض الريبي منقوص الاكسجين ، وتحديدًا عوامل *TF* ومواقع *TFBS* ، والتي تلعب دورًا أساسيًا في هذه ألعمليات و مع التطور الكمي للمعلوماتية من حيث السعة التخزينية ، أصبح من الممكن تخزينها ، لذلك استهدفنا في هذا العمل استكشاف العلاقة بين هذه العوامل باستخدام تقنيات التنقيب عن البيانات باستخدام الخوارزمية *Apriori* و *Eclat*

GENERAL
Introduction

Study Context:

Bioinformatics is a multi-disciplinary field of research which consists of all the concepts and techniques necessary for the computer interpretation of biological information. Several fields of application or sub-disciplines of bioinformatics have emerged such as Sequence bioinformatics, is particularly interested in the identification of similarities between sequences, structural bioinformatics, which deals with the reconstruction, prediction or analysis of 3D structure and Network bioinformatics, which is concerned with the interactions between genes, proteins, cells, organisms, trying to analyze and model the collective behaviors of sets of elementary bricks of the living.

The relationship between biology and computer science begins very early from the design of the first computer networks, biological data are used computer science, whether in storage, management or interpretation. For example the processing of an alignment of biological sequences done by computer science concerns some aspects:

1. **Data compilation and organization:** this aspect mainly revolves around creating databases, organizing as much information as possible.
2. **Systematic processing of sequences:** this aspect is based on Selection or characterizing a feature or an interesting biological element.
3. **Strategy development:** the latter aims to provide additional biological knowledge that can then be integrated into standard treatments.
4. **Evaluation of the different approaches in order to validate them:** it is the evaluation of the approaches mentioned above in order to validate.

Biology today knows many problems as a result of the vaguely applied scientific theories. This was the strongest motivation for researchers to conduct their research and exploit the technological development of computer science.

Among the most prominent problems that biology faces is the development of a relationship between the organ and the protein, which determines several regular processes in biological functions, such as no transcription, or no replication. In particular, we will pay attention to the protein-DNA bond.

Among the proteins binding to DNA, there are in particular transcription factors **TF'** and his binding sites **'TFBS'** which modulate the transcription of DNA into RNA, various polymerases, nucleases, which cleave DNA molecules, as well as his tones, which are involved in the conditioning of chromosomes and in the regulation of transcription within the nucleus of cells. TF and TFBS sequences are very similar in function and importance, so their domains are preserved by preserving their sequence by exploiting sequencing preservation.

Data miming: this term refers to analyzing data from different perspectives and turning that data into useful information, by establishing relationships between data or spotting patterns. This information can then be used by businesses to increase revenue or reduce costs. They can also be used to better understand a customer base in order to establish better marketing strategies.

- **Problematic:**

How to discover the binding motifs between DNA sequences 'TFBS' and protein sequences 'TF' using data mining techniques? To answer the previous question our proposed contributions are summarized in the following points:

✓ We extracted the relevant patterns on the TF and TFBS binding using the 'Apriori' algorithm of association rules.

✓ We will also extracted the hidden patterns between the 'TF' and 'TFBS' sequences using **'Apriori**' help by **'Eclat'**

✓ Finally we are going to transform the TFs and TFBS sequence bases into binary databases.

Our thesis is organized into 4 chapters:

- **The First chapter:** Biology for a computer scientist

- **The Second chapter:** Alignment of biological sequences
- **The Third chapter:** Association Rules
- **The Fourth chapter:** Proposition and realization

Chapter I

Biology for a computer scientist

❖ **Scientific background:**

In this first chapter, the scientific context of this thesis manuscript is to be situated. We introduce the key concepts essential to understanding the different themes addressed in the rest of the work. First, we perform a summary of genomics in eukaryotes. In a second step, we detail some aspects of transcriptomics. Then in a last part, we define the vocabulary of bioinformatics directly related to the processing of sequences.

I.1 Introduction to genomics:

Genome is a fancy word for your entire DNA. All living organisms have their own genome. Each genome contains the information needed to build and maintain that organism throughout its life. Your genome is the operating manual containing all the instructions that helped you develop from a single cell into the person you are today. It guides your growth, helps your organs to do their jobs, and repairs itself when it becomes damaged. In addition, it is unique to you. The more you know about your genome and how it works, the more you'll understand your own health and make informed health decisions. However, how are DNA and RNA linked? How does the protein synthesis? What is a gene? We all asked ourselves these types of questions during our first steps in genomics, so it seems important to fix these few axioms before entering more at the heart of the problem.[1]

I.2 Basic notions in molecular biology:

I.2.1 The cell:

In biology, Cell is the basic membrane-bound unit that contains the fundamental molecules of life and of which all living things are composed.

A single cell is often a complete organism in itself, such as a bacterium or yeast. Other cells acquire specialized functions as they mature. These cells cooperate with other specialized cells and become the building blocks of large multicellular organisms, such as humans and other animals. Although cells are much larger than atoms, they are still very small. The smallest known cells are a group of tiny bacteria called mycoplasmas. So is the basic structural, functional, and biological unit of all known organisms. A cell is the smallest unit of life. Cells are often called the "building blocks of life. [2]

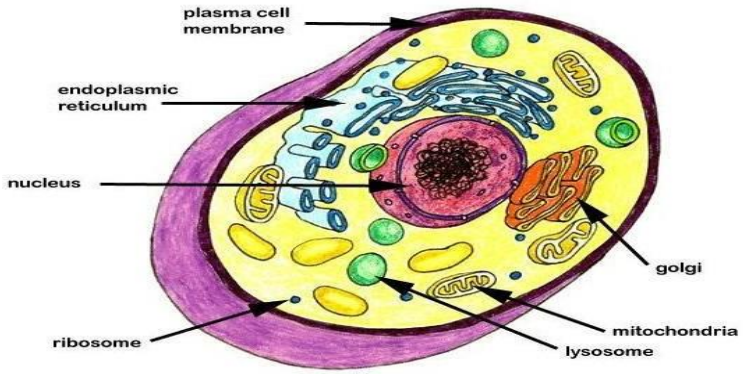


Figure I.1. *A celle*

We can distinguish two organisms according to the number of cells:

1. Prokaryotic: are organisms made up of cells that lack a cell nucleus or any membrane-encased organelles. This means the genetic material DNA in prokaryotes is not bound within a nucleus. In addition, the DNA is less structured in prokaryotes.

2. Eukaryotes: are organisms made up of cells that possess a membrane-bound nucleus (that holds DNA in the form of chromosomes) as well as membrane-bound organelles. Eukaryotic organisms may be multicellular or single-celled organisms. All animals are eukaryotes. Other eukaryotes include plants, fungi, and protests.

I.2.2 the chromosomes:

In each cell of our body, we retrieve the genetic information specific to each individual this information is located in the nucleus of the cell and precisely on the chromosomes.

The chromosomes are element of the cell nucleus, of characteristic shape and in constant number for a given species and considered to be the support of hereditary factors.

Genetic information is distributed across 46 chromosomes (23 pairs). For each pair, there is a chromosome of paternal origin and a chromosome of maternal origin. Thus, for the same pair, the two chromosomes will not be identical. The first 22 pairs are called "**Autosomes**". The 23rd pair is the one that determines the sex of the person. These are the X and Y chromosomes. Women have two X chromosomes, while men have an X chromosome and a Y chromosome. [3]

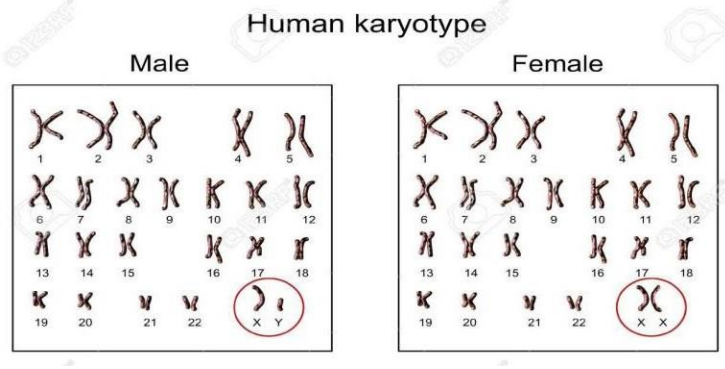


Figure I.2.*Chromosomes*

I.2.3 the Gene:

A gene is the basic physical and functional unit of heredity. Genes are made up of DNA. Some genes act as **instructions** to make molecules called proteins. However, many genes **do not code for proteins**. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. **The Human Genome Project** estimated that humans have between 20,000 and 25,000 genes.

Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people. Alleles are forms of the same gene with small differences in their sequence of DNA bases. These small differences contribute to each person's unique physical features.

Scientists keep track of genes by giving them unique names. Because gene names can be long, genes are also assigned symbols, which are short combinations of letters (and sometimes numbers) that represent an abbreviated version of the gene name.[4]

I.2.4 the Genome:

A genome is an organism's complete set of genetic instructions. Each genome contains all of the information needed to build that organism and allow it to grow and develop. Our bodies are made up of millions of cellseach with their own complete set of instructions for making us, like a recipe book for the body. This set of instructions is known as our genome and is made up of DNA. Each cells in the body for example a skin cell or a liver cell contains this same set of instructions:

- ✓ The instructions in our genome are made up of DNA
- ✓ Within DNA is a unique chemical code that guides our growth, development and health.
- ✓ This code is determined by the order of the four-nucleotide bases that make up DNA.
- ✓ DNA has a twisted structure in the shape of a double helix.
- ✓ Single strands of DNA are coiled up into structures called Chromosomes.
- ✓ Your chromosomes are located in the nucleus within each cell.

✓ Within our chromosomes sections of DNA are "read" together to form genes.

✓ Genes control different characteristics such as eye color and height.

✓ All living things have a unique genome.

✓ The human genome is made of 3.2 billion bases of DNA but other organisms have different genome sizes.[5]

I.2.5 the DNA:

I.2.5.1 Definition:

Alternatively, deoxyribonucleic acid is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria (where it is called **mitochondrial DNA** or mtDNA). Mitochondria are structures within cells that convert the energy from food into a form that cells can use. [6]

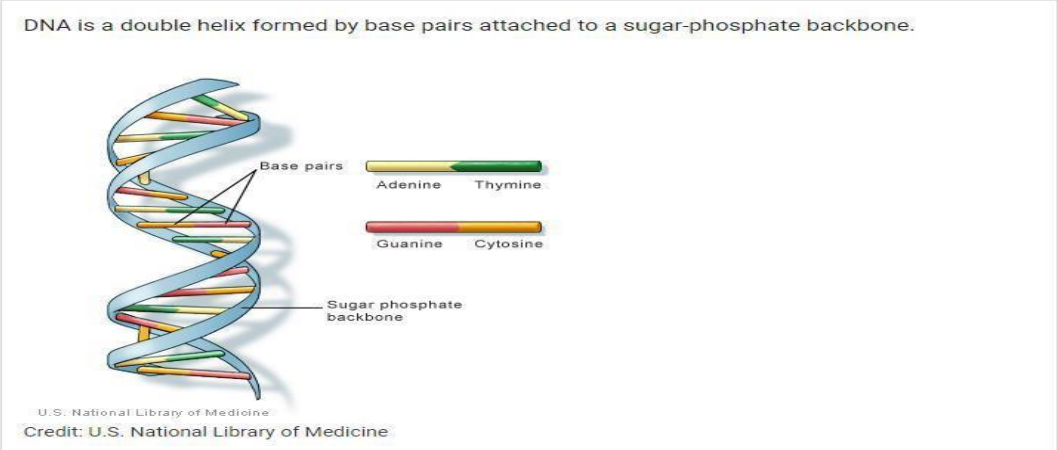


Figure I.3. *DNA*

I.2.5.2 DNA structure:

The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences.

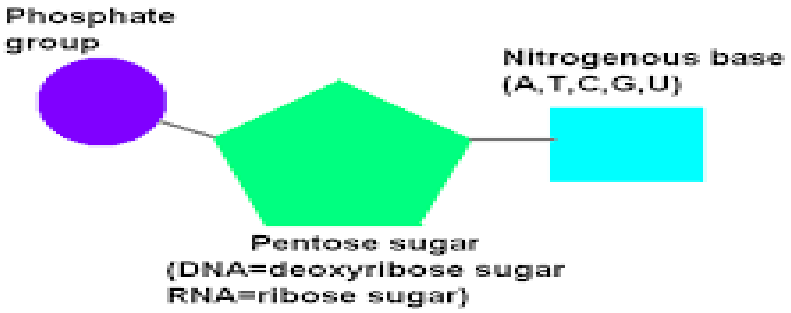


Figure I.4. *DNA structure*

DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together a base sugar and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder.

An important property of DNA is that it can replicate, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell. **Phosphodiester bonds in DNA polymers connect the 5' carbon of one nucleotide to the 3' carbon of another nucleotide**

The nucleotide monomers in a DNA polymer are connected by strong electromagnetic attractions called Phosphodiester bonds. Phosphodiester bonds are part of a larger class of electromagnetic attractions between atoms that chemists refer to as covalent bonds.

In order to keep things organized, biochemists have developed a numbering system for talking about the molecular structure of nucleotides. These numbers are applied to the carbon atoms in the sugar, starting at the carbon immediately to the right of the oxygen in the deoxyribose ring, and continuing in a clockwise fashion: the numbers range from 1‟ (“one prime”), identifying the carbon immediately to the right of the oxygen) all the way to 5‟ (“five prime”), identifying the carbon that sticks off the fourth and final carbon in the deoxyribose ring.

I.2.5.3 the Biological function of DNA:

- ✓ **DNA polymers direct the production of other polymers called proteins:**

A protein is one or more polymers of monomers called amino acids. Proteins are the workhorse molecules in your cells. They act as enzymes, structural support, hormones, and a whole host of other functional molecules. All traits derive from the interactions of proteins with each other and the surrounding environments.

- ✓ **chromosome consists of smaller segments called genes:**

Chromosomes are very long structures consisting of two DNA polymers, joined together by hydrogen bonds connecting complementary base pairs. A chromosome is divided into segments of double-stranded DNA called genes.

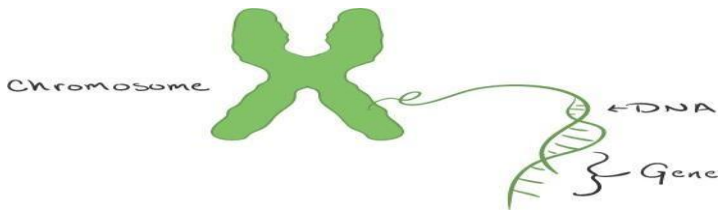


Figure I.5. *The relationship Chromosome, DNA and Gene*

Each gene is further divided into three nucleotide sub segments called cordons:

A codon is a segment (or piece) of double stranded DNA that is three nucleotides long. A gene can be thought of as many three-nucleotide codons strung together.

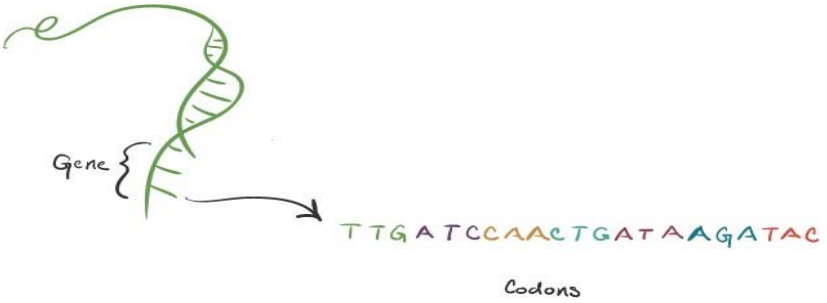


Figure I.6. *The relationship Gene Codons*

I.2.5.4 DNA Function:

DNA stores the information needed to build and control the cell. The transmission of this information from mother to daughter cells is called vertical gene transfer and it occurs through the process of DNA replication. DNA is replicated when a cell makes a duplicate copy of its DNA, then the cell divides, resulting in the correct distribution of one DNA copy to each resulting cell. DNA can also be enzymatic ally degraded and used as a source of nucleosides and nucleotides for the cell. Unlike other macromolecules, DNA does not serve a structural role in cells.

I.2.6 the ARN:

Ribonucleic acid or RNA is one of the three major biological macromolecules that are essential for all known forms of life (along with DNA and proteins). A central tenet of molecular biology states that the flow of genetic information in a cell is from DNA through RNA to proteins: “DNA makes RNA makes protein”. Proteins are the workhorses of the cell; they play leading roles in the cell as enzymes, as structural components, and in cell signaling, to name just a few.

DNA (deoxyribonucleic acid) is considered the “blueprint” of the cell; it carries all of the genetic information required for the cell to grow, to take in nutrients, and to propagate. RNA–in this role–is the “DNA photocopy” of the cell. When the cell needs to produce a certain protein, it activates the protein’s gene–the portion of DNA that codes for that protein–and produces multiple copies of that piece of DNA in the form of messenger RNA, or mRNA. The multiple copies of mRNA are then used to translate the genetic code into protein through the action of the cell’s protein manufacturing machinery, the ribosome’s. Thus, RNA expands the quantity of a given protein that can be made at one time from one given gene, and it provides an important control point for regulating when and how much protein gets made.

For many years RNA was believed to have only three major roles in the cell–as a DNA photocopy (mRNA), as a coupler between the genetic code and the protein building blocks (tRNA), and as a structural component of ribosome’s (rRNA). In recent years, however, we have begun to realize that the roles adopted by RNA are much broader and much more interesting. We now know that RNA can also act as enzymes (called ribozymes) to speed chemical reactions. In a number of clinically important viruses RNA, rather than DNA, carries the viral genetic information. RNA also plays an important role in regulating cellular processes–from cell division, differentiation and growth to cell aging and death. Defects in certain RNAs or the regulation of RNAs have been implicated in a number of important human diseases, including heart disease, some cancers, stroke and many others.

DNA alone cannot account for the expression of genes. RNA is needed to help carry out the instructions in DNA.[7]

Like DNA, RNA is made up of nucleotide consisting of a 5-carbon sugar ribose, a phosphate group, and a nitrogenous base. However, there are three main differences between DNA and RNA:

- RNA uses the sugar ribose instead of deoxyribose.
- RNA is generally single-stranded instead of double-stranded.
- RNA contains uracil in place of thymine

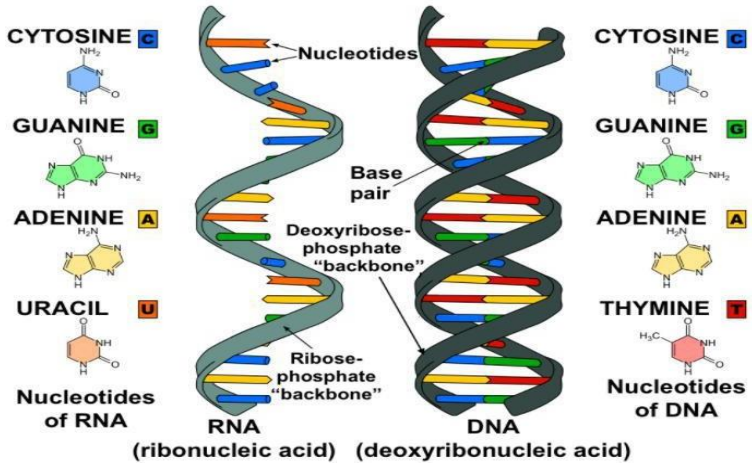


Figure I.7. *Differences help enzymes in the cell to distinguish DNA from RNA*

I.2.7 the Protein:

I.2.7.1 Definition:

Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs. Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein. The sequence of amino acids determines each protein's unique 3-dimensional structure and its specific function.[8]

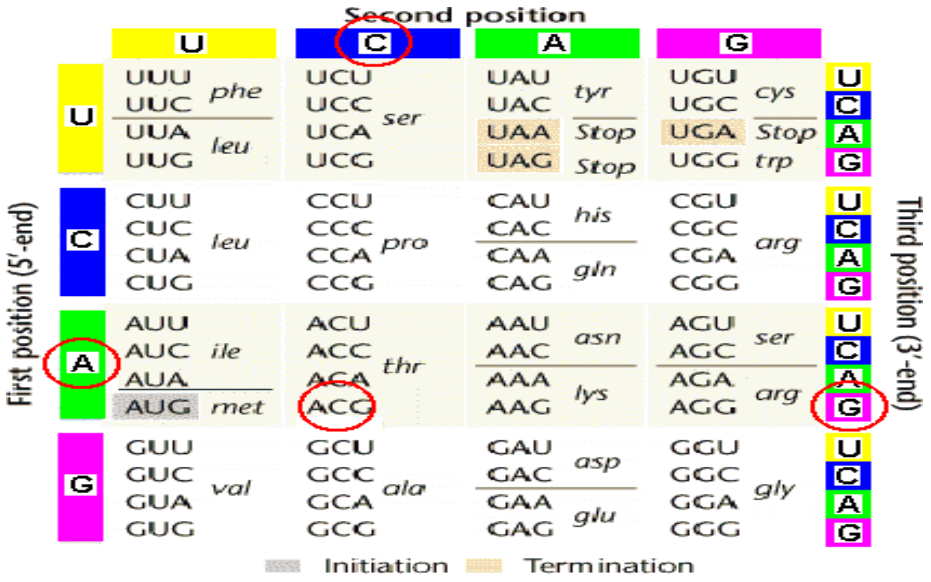


Figure I.8. *The genetic code*

I.2.7.2 production of proteins:

a) DNA replication and repair:

DNA replication and repair are critical processes that ensure the correct genetic material of a biological system is carried on. DNA continually undergoes a process of replication and division and errors can sometimes occur in the process. It is essential for the biological system to have a mechanism in place to detect and repair these errors.

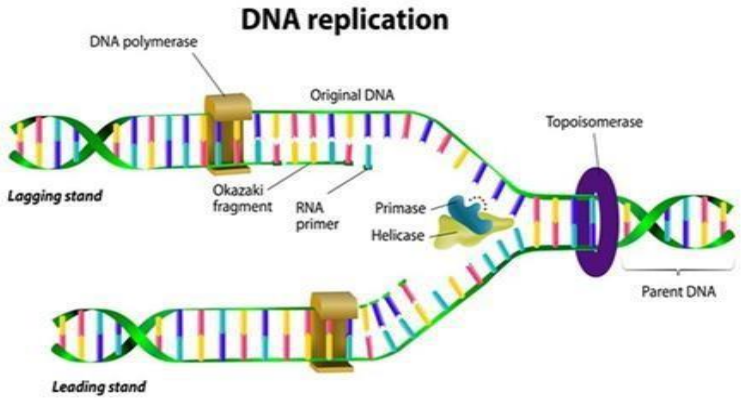


Figure I.9. *DNA replication*

In order for DNA to be replicated correctly, there are several guiding principles that should be present, including:

- DNA in a state ready to begin the process of replication.
- Clear starting to commence the replication.
- Ending point to finish the DNA copy.
- Proofreading and repair mechanism in place to detect any errors.
- Ability to distinguish between the original and copy of DNA.

b) Stages of DNA and Cell Replication:

The cell cycle of eukaryote cells includes four main phases for the replication of DNA and new cells:

1) G1 phase: the first gap phase, when the cell prepares for division with metabolic changes.

2) S phase: the synthesis phase, when the DNA and genetic material of the cell are duplicated, forming two sister chromatids.

3) G2 phase: the second gap phase, when cytoplasmic materials needed for division are gathered with metabolic changes.

4) M phase: the mitosis phase, when the genetic material and the cell divide.

The actual DNA replication occurs during the synthesis (S) phase, when two copies of the original cell DNA are produced, each containing one original and one new strand of DNA.[9]

***c) Transcription:* DNA into RNA**

Transcription is the first step of gene expression. During this process, the DNA sequence of a gene is copied into RNA. Before transcription can take place, the DNA double helix must unwind near the gene that is getting transcribed. The region of opened-up DNA is called a transcription bubble.

RNA polymerases are enzymes that transcribe DNA into RNA. Using a DNA template, RNA polymerase builds a new RNA molecule through base pairing. For instance, if there is a G in the DNA template, RNA polymerase will add a C to the new, growing RNA strand.

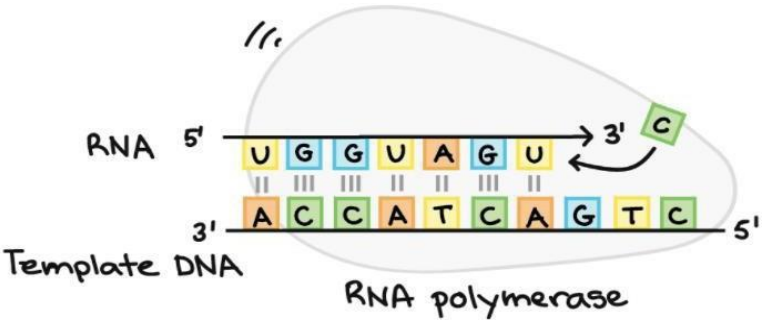


Figure I.10. *Transcription DNA into ARN*

- **Mutations:** Sometimes cells make mistakes in copying their genetic information, causing mutations. Mutations can be irrelevant, or they affect the way proteins are made and genes are expressed.
- **Substitutions:** A substitution changes a single base pair by replacing one base for another.
- **Insertions and deletions:** An insertion occurs when one or more bases are added to a DNA sequence. A deletion occurs when one or more bases are removed from a DNA sequence. Because the genetic code is read in Codons (three bases at a time), inserting or deleting bases may change the "reading frame" of the sequence. These types of mutations are calledframeshift mutations.

	Original reading frame	Frameshift mutation (deletion of 4th T nucleotide)	New reading frame
DNA:	3'-TACTATGCCTTA-5'	→	3'-TACATGCCTTA-5'
mRNA:	5'-AUGAUACGAAU-3'		5'-AUGUACGGAAU-3'
Codons:	5'-AUG-AUA-CGG-AAU-3'		5'-AUG-UAC-GGA-AU-3'
Polypeptide:	Met-Ile-Arg-Asn		Met-Tyr-Gly

Figure I.11. *Insertions and deletion's Example*

d) Translation: RNA into Protein

Translation involves “decoding” a messenger RNA (mRNA) and using its information to build a polypeptide, or chain of amino acids. For most purposes, a polypeptide is basically just a protein.

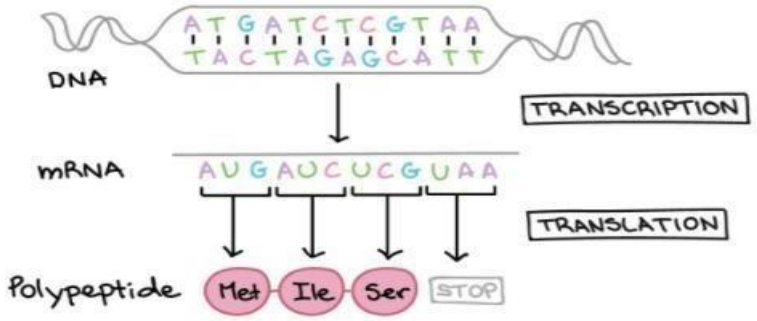


Figure I.12. *Protein production operations*

I.2.8. Promoter:

Promoters are DNA sequences whose purpose is not to encode information about the organism itself, but rather they serve as a kind of "On" switch to initiate the biological process of transcription for the genes which follow the promoter DNA sequence. The enzyme, RNA polymerase, which performs the transcription process, binds to the promoter sequence and then beings to work its way down the DNA segment, constructing RNA to match the DNA nucleotides over which the enzyme passes.[10]

I.2.9 Transcription factor and their binding sites:

The transcription factors and their binding sites are called "Transcription factor" and "Transcription factor binding site" TFBS. They are jointly involved in the initiation of gene expression and / or in its regulation. TFBS are short DNA sequences on which TF can bind.

A TF is a protein made up of a DNA or other TF binding domain and a transcription regulation domain. There are two categories of TF:

- The general transcription factors or GTF (General Transcription Factors).
- Specific transcription factors.

a. Transcription Factor:

Transcription factor: A protein that controls when genes are switched on or off-whether genes are transcribed or not. Transcription factors bind to regulatory regions in the genome and help control gene expression.

b. transcription factor binding sites : The region of the gene to which TF binds is called a **transcription factor binding site**. These sites are a subset of DNA binding sites. Overall, these sites can be defined as short segments of DNA that are specifically bound by one or more proteins with various functions. If you think about a parking lot with assigned spaces, the one labeled 'Transcription Factors Only' would be a transcription factor binding site.

I.3 Sequence banks:

The origin of biological databases goes back to the use of the first computers by crystallographers or biochemists. Among these, Margaret Dayhoff, American biochemist, was the first to see the interest in gathering all the data on protein sequences in order to study their evolutionary relationships and to classify them into families. It published the first atlas of proteins containing the sequence and structure of 65 of them (Atlas of Protein Sequence and Structure) in 1965. This atlas was periodically updated and published on paper until 1978. Distributed on magnetic support from 1978, it is now available online since 1981, via the Internet (Margaret Dayhoff). This increasingly large atlas became, in 1984, the P.I.R. (Protein Information Resource) from the National Biomedical Research Foundation (N.B.R.F.), the first concerning proteins and which remains a reference for their analysis. In 2004, it contained some 283,000 protein sequences which totaled 96 million amino acids. At the same time and in a concerted manner, two nucleic acid sequence databases took off on each side of the Atlantic in 1982. In the United States, GenBank took shape at L.A.N.L. (Los Alamos Nuclear Laboratory) with Doug Brut lag and

Temple Smith; since 1987, it has been managed and distributed by the N.C.B.I. (National Center for Biotechnology Information). The European nucleic bank, for its part, took the name of the laboratory within which it was developed in Heidelberg (Germany): E.M.B.L. (European Molecular Biology Laboratory). Since 1997, a special antenna for IT has been created in Cambridge: the E.B.I. (European Bioinformatics Institute), to further develop the E.M.B.L.

Given the considerable work involved in maintaining these two banks, the organizations decided to join their efforts, starting in 1986, to create standard formats. Founded in 1974, EMBL is Europe's flagship laboratory for the life sciences – an intergovernmental organization with more than 80 independent research groups covering the spectrum of molecular biology. It operates across six sites: Heidelberg, Barcelona, Hamburg, Grenoble, Rome and EMBL-EBI Hinxton.

I.3.1 Sequence banks nucleic:

- **EMBL:** Founded in 1974, EMBL is Europe's flagship laboratory for the life sciences an intergovernmental organization with more than 80 independent research groups covering the spectrum of molecular biology. It operates across six sites: Heidelberg, Barcelona, Hamburg, Grenoble, Rome and EMBL-EBI Hinxton.[11]
- **GenBank:** GenBank is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. Built and distributed by the National Center for Biotechnology Information NCBI, a division of the National Library of Medicine NLM, located on the campus of the US National Institutes of Health NIH in Bethesda, MD.
- **DDBJ (DNA Data Bank):** The DNA Data Bank of Japan (DDBJ) is a biological database that collects DNA sequences. It is located at the National Institute of Genetics (NIG) in the Shizuoka prefecture of Japan. It is also a

member of the International Nucleotide Sequence Database Collaboration or INSDC. It exchanges its data with European Molecular Biology Laboratory at the European Bioinformatics Institute and with GenBank at the National Center for Biotechnology Information on a daily basis. Thus these three databanks contain the same data at any given time.

Not only that, but there are other banks, for example:

- DBSTS (Sequence Tagged Sites).
- GBCONTIG

I.3.2 Sequence banks protein:

- **PIR-NBRF:** The Protein Information Resource (or PIR), is a bioinformatics database of protein sequences and analysis tools freely accessible to the scientific community. PIR was established in 1984 by the National Biomedical Research Foundation to assist researchers in the identification and interpretation of their protein sequences, it is located at the University of Georgetown in the United States.
- **Swiss Port:** The SWISS-PROT Protein Knowledgebase is an annotated protein sequence database established in 1986.is a curate protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases. The Swiss Institute maintains it collaboratively for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI).

I.3.3 the pattern banks:

- **PROSITE:** The PROSITE database can be considered as a dictionary, which lists protein motifs with biological significance. (Bairoch 1993, Bairoch and Bucher 1994, Falquet 2002)It is established by grouping where possible the proteins contained in Swissprot by family such as for example kinases or

proteases. We then seek within these groups consensus motives likely to characterize them specifically.

The design of the base is based on four essential criteria:

- Collect as many significant reasons as possible.
- Have highly specific motifs to best characterize a family of proteins.
- Provide complete documentation on each of the grounds listed.
- Periodically review the reasons to ensure their validity in relation to the last experiments.

I.4 Conclusion:

The number of data in the field of biology continues to increase in particular with the sequencing of the genomes of different organisms but we are also witnessing a great diversification of the information produced (primary sequences, molecular structures, mapping, and collection of strains or clones ...). All these data are gathered in databases very varied in their volume and nature. We can now imagine their consultation without the help of IT. This contribution has become considerable in recent years, in particular with the extension of broadband networks. It allows scientists to use new tools ranging from simple textual interrogation to the graphical presentation of data through the use of multi-windowing or audio or video documents.

It is therefore obvious that the organization and the interrogation of the data will be radically changed. This transformation is already apparent in the development of certain software that offers more and more interactions between databases, this by further exploiting the links that exist between them. SRS software, which is installed on many WWW servers, is an example of this development by offering multibase consultation with the same graphical interface.

Chapter II

Alignment of biological sequences

II .1 Introduction:

Over the past thirty years, the process of data collection in biology has increased quantitatively thanks to the development of new technical methods used to understand other components of organisms, including DNA.

To analyze this data, which is also more numerous and more complex, researchers turned to new information technologies, the huge data storage and analysis capacity offered by computers has enabled them to gain power for their studies. The combination of biology and computer science is what we call bioinformatics. This includes specializations in life sciences such as genomics, proteins and systems biology. [12]

Given the importance of the problems, it is necessary to be able to obtain results good quality so that they can be used by biologists. The objective of bioinformatics is therefore to bring together the skills available in each scientific discipline in order to contribute to the improvement of resolution methods. For example for the problem of multiple sequence alignment, computer and mathematical solutions have been provided given in the form of a algorithm [12] In this chapter we give the definition of bioinformatics followed by its process, we will present in parallel the problem of alignment of two sequences as a whole, that is to say we treat the problem in a global way according to the principle and the uses that can be made of them, even if aligning two sequences appears to be a big problem to solve.

So, as a solution to this problem there is an exact algorithm of polynomial complexity. However, when it is necessary to calculate many alignments, this complexity may prove to be too great for the algorithm to be usable.

II .2 what is bioinformatics?

Bioinformatics (La bioinformatique in French) English word composed of two parts: “Bio” means life or all that is natural and “Informatics or Computing” means the science of information processing by the computer.

Several definitions associated to this term:

- **Definition 1:** Bioinformatics is the study of biological information when it passes from its storage site in the genome to the various products of genes in the cell. It includes the creation and development of advanced computer technologies for the problems of molecular biology. [13]
- **Definition 2:** Bioinformatics refers specifically to the research and use of models and structures in biological data and the development of new methods of accessing databases. [13]
- **Definition 3:** An interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. [14] So in short, bioinformatics is a young field in turmoil, innovative, hybrid and dynamic!!

II .3 Basics:

II.3.1 General:

Aligning two sequences defined on an alphabet consists in cutting these sequences in order to highlight common areas to bring out the similarities. This fractionation makes it possible to superimpose the identical zones between the two sequences. To make it easier to read, the cuts within the sequences are materialized by the symbol ‘-‘[2]

For example to align the following two sequences:

S1: GACTGAG –

S2: GCTGGAAG

The sequences S1 and S2 are initially defined on an alphabet Σ = {A, C, G, T}. A possible alignment is represented by the pair of sequences S′**1** and S′**2** defined on an extended alphabet Σ′ = Σ ∪ {-}: [2]

S′1: GACTG--AG

S′2: G-CTGGAAG

II.3.2 Distance of Hamming:

In information theory, the Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different. In other words, it measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other. [15]

In a more general context, the Hamming distance between two code words is simply the number of bit positions in which they differ. If the Hamming distance between two code words c1 and c2 is d, and c1 is transmitted, then d errors would have to occur for codeword c2 to be received.[16]

Example:

Let the sequences of the same languor S1, S2, S3:

S1 ACATGCCATGAC

S2 ACGTGCGATGAT

S3 AACTCGGATGTC

The Hamming distance of the three sequences noted as follows:

- dH (S1 / S2) = 3
- dH (S1 / S3) = 6
- dH (S2 / S3) = 6

II.3.3 Editing Operations:

The alignment of two sequences consists in comparing all the characters making up these two sequences. The editing operations define the different modifications necessary to explain the evolution of the sequence S up to the sequence T. Thus, for each pair of residues, four cases are possible, each corresponding to an editing operation:

- The pairing or match which corresponds to two characters which correspond (a, a).
- The substitution or mismatch which corresponds to two characters which do not correspond (a, b) with $a \neq b$.
- Adding a breach in S or inserting (-, b).
- Adding a break in T or deletion (a, -).

The evolution of the sequence S to the sequence T is obtained by performing deletions instead of insertions, and vice versa. [17]

Example:

Let the sequences S, T:

S: TCATGTGAAT S′: TC-ATGTGAAT

T: TCGATGCACT T′: TCGATG-CACT

The sequence of editing operations used to transform S′ into T′ is:

1. Two pairings**:** (T, T) and (C, C).
2. A insertion **:**(-, G)
3. Three pairings**:** (A, A), (T, T) and (G, G).
4. A deletion **:**(T,-).
5. A substitution **:**(G, C).
6. A pairing**:** (A, A).
7. A substitution **:**(A, C).

8. A pairing**:** (T, T).

II.3.3.1 Edit distance:

Different definitions of an edit distance use different sets of string operations. Levenshtein distance operations are the removal, insertion, or substitution of a character in the string. Being the most common metric, the term Levenshtein distance is often used interchangeably with edit distance. [18]

In bioinformatics, it can be used to quantify the similarity of DNA sequences, which can be viewed as strings of the letters A, C, G and T. [19]
According to Levenshtein the editing distance is as follows:

Let S and T be two sequences, and let SOM be the sum of the editing operations used. The editing distance, between the two sequences S and T is the minimum value that can take SOM. [17]

Example :

Let the sequences S and T, a minimal edit script that transforms the former into the latter is:

S: TCATGTAT T: TCGATGCACT

1. TCA TCGA (insert "G").
2. TCGA TCGA (A pairing "A").
3. TCGAT TCGAT (A pairing "T").
4. TCGATG TCGATG (A pairing "G").
5. TCGATGTA TCGATGCA (substitute "C" for "T").
6. TCGATGTA TCGATGCACT (insert "T" at the end).

The Levenshtein distance between S and T is 3.

II.3.3.2 Score function:

Editing operations transform the S and T sequences into S′ and T′ sequences of the same length. To evaluate the alignment of the sequence with the editing distance, we have to assign costs to each pair of elements $\Sigma \cup \{-\}$.

To determine if the alignment is better than the other, we use the score function.

Let Σ be an alphabet, and let $\Sigma' = \Sigma \cup \{-\}$. A score function for an alphabet Σ is an application f defined by F: $\Sigma'^* \times \Sigma'^* - \rightarrow R$, and which has an alignment associates the sum of the values of its editing operations.

The use of a score function requires having two elements:

- A substitution matrix allowing associating a value to the matching and replacement operations.
- An evaluation model for breccias associating a value with the operations of insertion and deletion.

The use of a score function requires having two elements:

- A substitution matrix allowing associating a value to the matching and replacement operations.
- An evaluation model for breccias associating a value with the operations of insertion and deletion. [17]

II.3.4 The substitution matrix:

A substitution matrix is a collection of scores for aligning nucleotides or amino acids with one another which describe the rate at which one character in a sequence changes to other character states over time, where the similarity between sequences depends on their divergence time and the substitution rates as represented in the matrix. These scores generally represent the relative ease with which one nucleotide or amino acid may mutate into or substitute for another. [20][21]

There are two types of substitution matrix depending on the nature of the nucleic or protein sequence.

II.3.4.1 Score matrices for DNA:

Unlike proteins, the templates used for DNA are generally very simple. They have the role of assigning a value to the different possible configurations. The evaluation that we can give to a couple of nucleotides (x, y) is the same as the evaluation of the couple (y, x), all the matrices are therefore symmetrical. They can be divided into two categories [17]

II.3.4.1.1 The identity matrix:

A first possible approximation is to calculate a basic editing distance with an identity measure. In this case, we favor the substitution of an amino acid by itself. We therefore use an identity matrix in which the substitution of an amino acid by itself is at 1 and all the other substitutions are at 0. Applied to the nucleic bases, this gives the following 4x4 matrix. [23]

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

Figure II.1. *A identify matrix*

II.3.4.1.2 The transition / Transervation matrix:

In the case of DNA, the substitutions A-G and T-C (transition) appear more often than A-C, A-T, G-C and G-T (transfusions). Identify =3, transition=1, Transervation= 0.

To represent this information, we can use a transition-transversion matrix like the following [23]

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4

Figure II.2 *the transition / Transervation matrix*

II.3.4.1.3 the BLAST matrix:

The BLAST matrix is a matrix which has the same principle of an identified matrix but the assigned values of substitutions are different from 0 and 1.

	A	T	C	G
A	3	0	1	0
T	0	3	0	1
C	1	0	3	0
G	0	1	0	3

Figure II.3 *The BLAST matrix*

II.3.4.2 Score matrices for Protein:

These substitution matrices are constructed from large sets of protein sequence alignments, these sequences having amino acid frequencies that can be described as "standard". Local sequence alignments require score matrices which generate on average negative values in the case of comparison of random sequences.[24]

For these matrices one can mainly find three large families, but nothing prevents the use of another matrix:

II.3.4.2.1 PAM matrices: A substitution matrix describes the rate at which one character in a sequence changes to other character states over time.

PAM is one of the first amino acid substitution matrices .This matrix estimates what rate of substitution would be expected if 1% of the amino acids had changed. This matrix is calculated by observing the differences in closely related proteins. [25]

Let **S1** and **S2** be two protein sequences such as:

S1: MMLSATQPLSEKLPAHGCRHVAIIMDGNGRWAKKKGKIKAGAKSV RRAVSFAANN

S2: MMLSATQPLSEKLPAHGCRHVAIIMDGNGRWAKKKGKIKAGAKSV RKAVSFAANN

We say that the two sequences are at evolutionary distance of PAM1 (point Accepted Mutation), if S1 has converted to S2 with an average of one amino acid substitution per 100 amino acids.

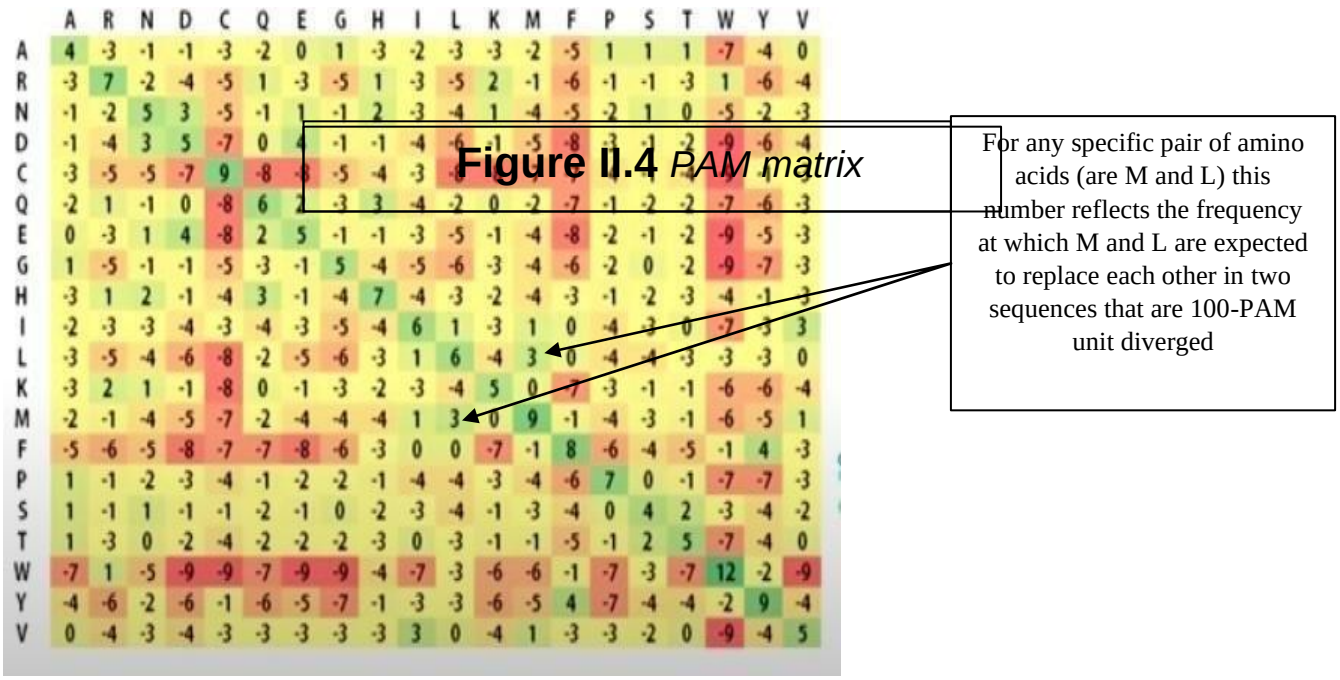


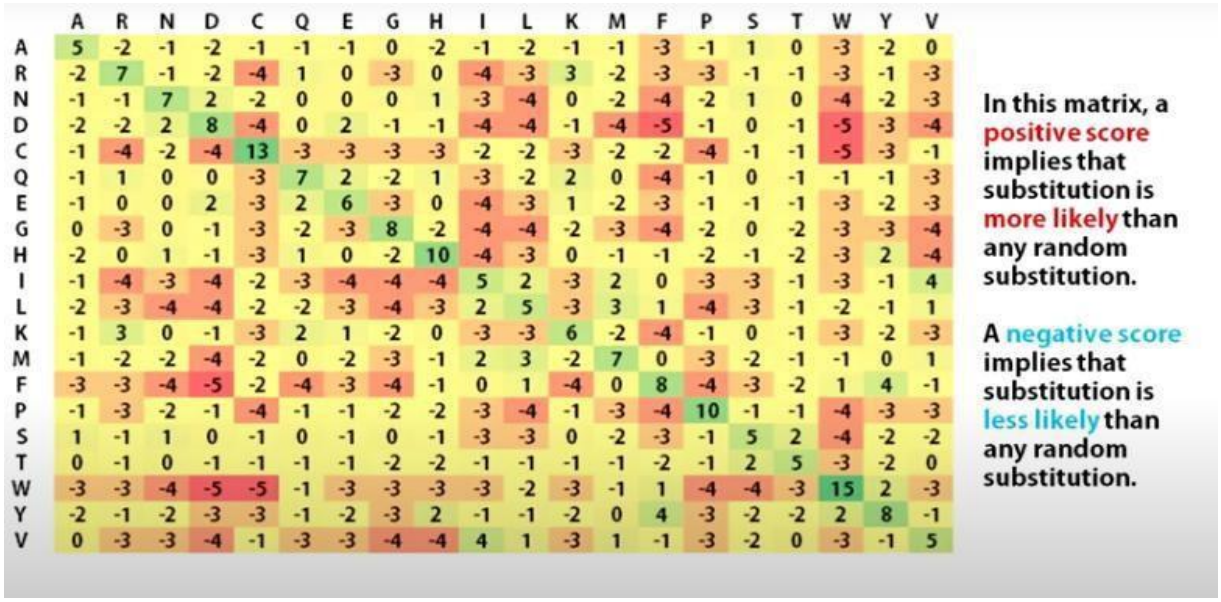
Figure II.4 *PAM matrix*

II.3.4.2.2 BLOSUM matrices:

PAM matrix species turned out not to work very well for aligning evolutionarily divergent sequences. So Henikoff constructed the BLOSUM (Block Substitution Matrix) series of matrices rectifies this problem. These matrices using multiple alignments of evolutionarily divergent proteins:

- Scanned Blocks databases
- Looked for very conserved protein family region (no gaps)
- Counted the relatives frequencies of amino acids and there substution probabilities

Most common matrices BLOSUM 45, 62, 80.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

Figure II.5 *BLOSUM matrix*

II.3.4.2.3 Gonnet matrix:

This type of matrix was built in 1992 by Gonnet, Cohen and Benner.It is an iterative method, based on 16,300 protein sequences corresponding to 2,600 families. Each sequence was compared with all the sequences in the library and the alignments were obtained using an arbitrary chosen initial matrix. A new

matrix has been constructed and the alignments have been recalculated from this new matrix. This procedure was repeated until the matrix remained unchanged. Different Gonnet matrices: Gonnet 40, Gonnet 120 …, Gonnet 250, Gonnet 350. [24]

II.3.4.2.4 Choice of protein matrix:

Given the diversity of types of protein matrices, the choice of matrix depends on the types of analysis we want to perform. There is no ideal matrix and a large number of comparative studies on matrices (schematically) have shown this:

- ✓ For similar and short sequences, it is preferable to use a high BLOSUM or low PAM matrix.
- ✓ For divergent and long sequences, it is preferable to use a low BLOSUM or high PAM matrix.
- ✓ The BLOSUM 62 matrix seems to be the most used matrix for a large number of sequence alignment software; it seems to be the default matrix.

II.3.5 gap penalty:

A Gap penalty is a method of scoring alignments of two or more sequences. When aligning sequences, introducing gaps in the sequences can allow an alignment algorithm to match more terms than a gap-less alignment can. The editing process poses a cost that must be estimated in order to represent as close as possible to biological reality, and several weighting systems have been proposed.

II.3.5.1 Fixed penalty per gap:

In this case we assign a gap to a fixed value without taking into account its position in the sequence or its length: p = k

Example: if we consider the following example with

- **Score (identity) = 2**

- **Score (substitution) = 1**
- **Score (gap) = -1**

The Score of this alignment is then:

SEQ1: G A R F I E V H E L - - T F A T T C A T

SEQ2: G A R F I E L T H E V A S Y F- - C A T

Total score: 2+2+2+2+2+2+1+1+1+1-1-1+1+1+1-1-1+2+2+2=+21

II.3.5.2 variable penalty or affine function:

The most widely used gap penalty function is the affine gap penalty. The affine gap penalty combines the components in both the constant and linear gap penalty, we talk about: **P = A + B * L…… (II.1)**

Such as:

A: is known as the gap opening penalty **($GOP_{(x)}$).**

B: is the gap extension penalty **($GEP_{(x)}$).**

L: is the length of the gap.

And **P: is the overall cost of a length gap.**

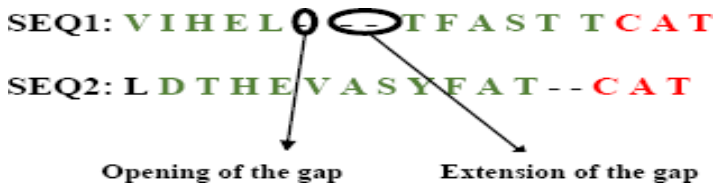
Gap opening refers to the cost required to open a gap of any length, and gap extension the cost to extend the length of an existing gap by 1 Often it is unclear as to what the values **A** and **B** should be as it differs according to purpose.

In general, if the interest is to find closely related matches, a higher gap penalty should be used to reduce gap openings.

The relationship between **A** and **B** also have an effect on gap size. If the size of the gap is important, a small **A** and large **B** is used and vice versa. Only the ratio **A/B** is important.

Example:

We will take the following alignment in this example



When we use a refinement function the score of this alignment becomes with:

- **Score (identity) = 2**
- **Score (substitution) = 1**
- **Score (GOP) = - 2**
- **Score (GEP) = 1**

Then the alignment score would be:

SEQ1: G A R F I E V H E L - - T F A T T C A T

SEQ2: G A R F I E L T H E V A S Y F - - C A T

Total score: 2+2+2+2+2+2+1+1+1+1+ (-2+1)1 +1+1+1+ (-2+1)1+2+2+2= +23

Another version of Affine's function is as follows:

$$P = A + B * \log (L) \quad \text{.....(II.2)}$$

Such as:

A: is known as the gap opening penalty **($GOP_{(x)}$).**

B: is the gap extension penalty **($GEP_{(x)}$).**

L: is the length of the gap.

And **P: is the overall cost of a length gap.**

Furthermore, we find formula **II.1** among several form of affine function is currently the most used by alignment methods for reasons of complexity of aligned sequences.

II .4 Comparison of sequences:

Alignment-free sequence analyses have been applied to problems ranging from whole-genome phylogeny, and detection of recombined sequences. The strength of these methods makes them particularly useful for next-generation sequencing data processing and analysis. However, many researchers are unclear about how these methods work, how they compare to alignment-based methods, and what their potential is for use for their research.

In general, we can compare two sequences by placing them above each other in rows and comparing them character by character. This way we could align two different sequences.

But in bioinformatics, we have a different purpose, and that is why we will look at the DNA, RNA and protein sequences. Deoxyribonucleic acid (DNA) consists of four bases adenosine, guanine, thymine, and cytosine that are represented by single letters A, G, T, and C respectively. In ribonucleic acid (RNA) thymine (T) is replaced by uracil, represented by the letter U. Protein sequences consist of 20 different amino acids. The single-letter codes for each amino acid are: I, L, V, F, M, C, A, G, P, T, S, Y, W, Q, N, H, E, D, K, and R.

II .4.1 why compare?

Alignments allow the comparison of biological sequences such comparisons are necessary for different studies:

- Identification of homologous genes: if two known functional molecules come together, we can conclude that part of their mechanism of action must be common.
- Search for functional constraints in a set of genes or protein.
- Function prediction.
- Structure prediction.

II .5 Pair wise Sequence Alignment:

II .5.1 definition:

Pair wise Sequence Alignment is a mapping between the residuals with a possible insertion of gaps in order to obtain sequences of equal length. All matches are allowed provided that the order of the residuals is respected.

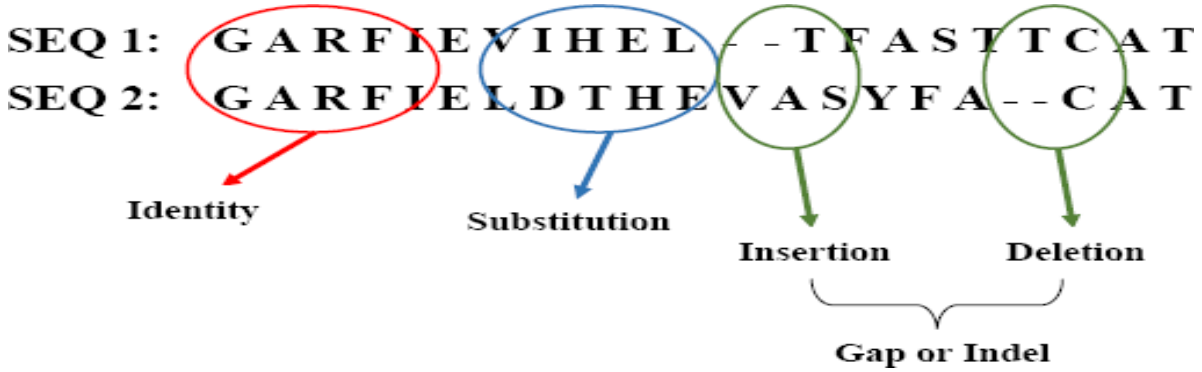
Pair wise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

Three situations are possible for a given position of the alignment:

- ✓ Case of the same characters **(Identity).**
- ✓ Case of the different characters **(Substitution).**
- ✓ Case of one of the position is a gap **(Insertion/deletion).**

Example:

Let **SEQ1** and **SEQ2** be two protein sequences, the pairwise of these two sequences are:



II .5.2 Pair wise Alignment Evaluation:

It became clear that for each sequence of protein or DNA data there are several possible alignments, so it became necessary to know the best alignment or the optimal if possible and to evaluate must be measured:

- ✓ **Quality:** by determining the distance between the two sequences.
- ✓ **Score:** by determining the score sum of all the base positions (residues) taken two by two

- **Formal definition (alignment Evaluation):**

Let A be an alignment of length n of the two sequences S'1 and S'2 defined on an alphabet Σ, and let M be a substitution matrix for **Σ**. We define the evaluation of A by the following value: $Eval = \sum_1^n Val(S'1[i], S'2[i])$…. *(II.3)*

Where **Val (S′1 [i], S′2 [i])** is equal to:

- **M (S′1 [i], S′2 [i])** if **S′1[i]** and **S′2 [i]** are not breaches
- The cost associated with a breach otherwise.

The cost associated with a breach depends on the method used to model that cost. If it is a breach at the start or end of the lineout, it may possibly not be counted. [17]

Example:

If we consider the previous example:

- **Score (identity) = 2**
- **Score (substitution) = 1**
- **Score (gap) = -1**

The Score of this alignment is then:

SEQ 1: G A R F I E V H E L- -T F A T T C A T
SEQ 2: G A R F I E L T H E V A S Y F- - C A T

2+2+2+2+2+2-1-1-1-1-2-2-1-1-1-2-2+2+2+2= 3

To evaluate this alignment we calculate its score: $Score\ (A, B) = \sum_{i=0}^{1} SC\ (A_i, B_i)$.....*(II.4)*

- Such us A and B two sequences of equivalent length L

II .5.3 Similarity percentage:

First simple method to get a good idea of the quality of the alignment of two sequences. This method consists in traversing all the pairs of residuals alignment. When the two amino acids are identical we assign the value 1 and 0 in all others cases. This gives the number of identities in the alignment. The breaches are counted in the other cases that are to say that there are no special penalties as explained in the previous part.

For this value to be representative, it is necessary to relate it to the length alignment. Let **V** be the identity function, which associates 1 with two identical residues, and 0 if not. The similarity Sim between two aligned sequences S'1 and S'2 is given by:

$$Sim\ (A) = \sum_1^1 V(S'1[i], S'2[i])/L \quad \text{..... (II.5)}$$

This method is very easy to implement, as it only involves the alignment itself. It is mainly used to compute distance matrices between several sequences. Indeed, if we assume that A is the best alignment for two sequences S"1 and S"2, then 1 - Sim (A) allows defining a distance between the sequences. [17]

II .5.4 Use of a substitution matrix:

Percent similarity is an easy method to use, but it is not always very representative. Indeed, assigning the same value to all the nucleotides is not in conformity with what is observed in reality. Moreover the value 0 in the case where there is no identity is even less acceptable. We saw in the section on substitution matrices that it is not necessarily bothersome that two different amino acids are facing each other.

Another notable difference with the percentage of similarity, the gaps are here also taken into account in a more realistic way. That is, it is necessary to use a more relevant evaluation, as discussed above. [17]

II .5.5 Pair wise alignment methods:

Pair wise alignment methods are used to compare pairs of sequences. They are used to find a homology between a test sequence and a reference sequence, often extracted from a database.

They are the simplest to implement, and they are the only ones for which there are optimal algorithmic solutions, based on dynamic programming. There are also rapid heuristic methods, which allow systematic searches to be carried out in the sequence banks. In this case, we compare an unknown sequence with all the sequences of the base, by testing them successively one by one. We present the most used algorithms in pair wise alignment. [17]

II .5.5.1 Global alignment:

Alignment methods can either try to align the sequences over their entire length, this is referred to as global alignment, or restrict themselves to limited regions in which the similarity is strong, to the exclusion of the rest of the sequences , we then speak of local alignment.

Example: Let **Seq1** and **Seq2** be two protein sequences such as:

Seq1: F T F T A L I L L A V A V

Seq2: F T A L L L A A V

The overall alignment of Seq1 and Seq2 is:

GLOBAL F T F T A L I L L A V A V

F - - T A L - - L L A – A V

Several methods have been developed in order to achieve a global alignment of two sequences as correct as possible.

II .5.5.1.1 DOT matrix:

A dot matrix analysis is primarily a method for comparing two sequences to look for possible alignment of characters between the sequences. The method is also used for finding direct or inverted repeats in protein and DNA sequences, and for predicting regions in RNA that are self-complementary and that,

therefore, have the potential of forming secondary structure through base-pairing.

The construction of a dot matrix is based on two sequences of length m and n. across or a dot if $\mathbf{X_i = X_j}$ where $\mathbf{X_i}$ is an element of the first sequences and $\mathbf{Y_j}$ an element of the second sequences, otherwise nothing

II .5.5.1.2 Dynamic programming:

Dynamic programming is a principle that is often simple to implement to solve complex problems. However, it only applies to a certain category of problems, and it is necessary to verify certain conditions before it can be applied.

Dynamic programming is a means which makes it possible to limit the increase in a computation time by a factor of **2*L** (L length of two sequences) and to conserve a computation time of ∂ **(L_2).**

II .5.6.1 local Alignment

The global alignment method that we have just presented makes it possible to determine the best alignment of the two complete sequences S and T. However, this optimality on all the sequences corresponds to the best possible compromise. It does not guarantee that it is not possible to better align two sub-sequences of S and T.

II .5.6.1.1 Definition:

Let **S** and **T** be two sequences. We define the local alignment of **S** and **T** as being the alignment **A** of the sequences **S′** and **T′** verifying:

- **S′** is a sub-sequence of **S**
- **T′** is a sub-sequence of **T**

There are no other sub-sequences of S and T whose alignment B verifies **Eval (B)> Eval (A) (**where **Eval (B) <Eval (A)** in the case of minimization).

The local alignment therefore represents the alignment whose evaluation is the best among the alignments of all the subsequences of **S** and **T**. [17]

Example:

In this example we align two DNA sequences that are longer. We can thus see that the local alignment is different from the global alignment, and only takes into account part of the two sequences. The sequences used are :

GCAGAGCACT

GCTGGAAGGCAT

The values of the substitution matrix used to determine the alignment are the following:

- Match: 5,
- Mismatch: -4,
- Insertion or deletion: -7.

The calculated maximum value is 19. It is therefore the starting point for the construction of the local alignment. Going up to the last strictly positive value we get the following result:

GAAG--GCA

GCAGAGCA

	-	G	C	T	G	G	A	A	G	G	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	0	0	5	**5**	0	0	5	5	0	0	0
C	0	0	10	3	0	1	**1**	0	0	1	10	3	0
A	0	0	3	6	0	0	6	**6**	0	0	3	15	8
G	0	5	0	0	11	5	0	2	**11**	5	0	8	11
A	0	0	1	0	4	7	10	5	4	**7**	1	5	4
G	0	5	0	0	5	9	3	6	10	**9**	3	0	1
C	0	0	10	3	0	2	5	0	3	6	**14**	7	0
A	0	0	3	6	0	0	7	10	3	0	7	**19**	12
C	0	0	5	0	2	0	0	3	6	0	5	12	15
T	0	0	0	10	3	0	0	0	0	2	0	5	17

Figure II.6 *Example of local Alignment*

II .5 Multiple Alignments:

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment as in the image at right illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.

Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. MSAs require more sophisticated methodologies than pair wise alignment because they are more computationally complex. Most multiple sequence alignment programs use heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally expensive. On the other hand, heuristic methods generally fail to give guarantees on the solution quality, with heuristic solutions shown to be often far below the optimal solution on benchmark instances. [26]

II.6 Conclusion:

In this chapter we have presented an important problem of molecular biology: the problem of alignment of two sequences. This is a very common problem in bioinformatics because it has practical uses but also because it will be used for other operations. Indeed, as we will see in the next chapter, it is often a necessary step for many multiple alignment algorithms.

The number of possible alignments of two sequences is very large. It is therefore necessary to define evaluation criteria in order to determine which the best alignment according to these criteria is. We use for this a substitution matrix which has each pair of residuals associate a value, as well as a function allowing evaluating the cost of inserting breaches

Chapter III

Association Rule

III.1 Introduction:

Today all companies store large databases. These gigantic databases, which keep growing day after day, are little exploited, while they hide decisive knowledge in the face of the market and the competition. To meet this need, a new discipline was born: data mining which we would call (Fouille de Donnée in French). Data mining is a process of discovering useful, new and understandable information and knowledge from a large database, a data warehouse or other databases. Generally, data mining methods are classified into two main categories: descriptive techniques and predictive techniques.

Data Mining is an essential component of Big Data technologies and big data analysis techniques. This is the source of Big Data Analytics, predictive analytics and data mining.

III.2 Association Rules:

III.2.1 Definition:

Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases. Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

An association rule can be formally defined like this:

Is **I= {i_1, i_2, i_3...i_m}** a set of indices (items) and **T= {t_1, t_2, t_3...t_n}** a set of transactions, such as t_i, either a subset of **I ($t_i \subseteq$ I)**

An association rule, which may be true or false, is expressed in the form:

$X \rightarrow Y$, or XϵT, YϵT, and X$\cap$Y $\neq \emptyset$.

Transaction	Item set
t_1	i_3 i_4 i_5
t_2	i_1 i_2
t_3	i_1 i_3 i_4 i_5 i_6 i_7
t_4	I_3 i_5
t_5	i_2 i_6 i_1 i_7

Array III.1 *Example of a transaction table*

III.2.2 Support:

We notice **X** an item set and a transaction. we say that the transaction t contains **X** if and only if **X** $\subseteq$ **t**. for example, in the previous table, my transaction t_2 contains the item set $\{i_1, i_2\}$.

The relationship between the numbers of transactions containing X and the total number of transactions D is called: **support** for the item X noted **sup(X)**.

$$Sup(X) = \frac{|\{t\ D, X\ t\}|}{|D|} \quad \text{..... (III.1)}$$

For X, Y two non-empties, disjoint($X \neq \emptyset, Y \neq \emptyset, X \cap Y = \emptyset$ item sets we have the couple (X, Y) form a noted association rule $X \to Y$ .and the support of $X \to Y$ is defined as follows:

$$Sup(X \to Y) = Sup(X \cup Y) = \frac{|t \epsilon D, X \subset t, Y \subset t|}{|D|} \quad \text{..... (III.2)}$$

III.2.3 frequent item set:

We say that an item set is frequent if and only if its support is greater than or equal α to a given threshold.

If Sup (X) ≥ α So X is common, else X uncommon.

If X is frequent then all subset of item set X is frequent, moreover, if X is infrequent then all subset of item set X is infrequent.

III.2.4 The trust: the confidence of the rules $X \to Y$ is defined as the relationship between the support of X∪Y and the support of X

$$\textbf{Conf}\ (X \to Y) = \frac{Sup(X \cup Y)}{Sup(X)} \quad \text{..... (III.3)}$$

III.3 Extraction of association rules: the a priori algorithm:

III.3.1 Apriori algorithm:

III.3.1.1 what is the Apriori Algorithm:

Apriori algorithm, a classic algorithm, is useful in mining frequent itemsets and relevant association rules. Usually, you operate this algorithm on a database containing a large number of transactions. One such example is the items customers buy at a supermarket.

It helps the customers buy their items with ease, and enhances the sales performance of the departmental store.

This algorithm has utility in the field of healthcare as it can help in detecting adverse drug reactions (ADR) by producing association rules to indicate the combination of medications and patient characteristics that could lead to ADRs.

III.3.1.2 the principle of this algorithm:

The Apriori algorithm is executed in two stages:

- Let minsupp be the given minimum support index, and minconf the given confidence index.
- Generation of all frequent item sets.
- Generation of all trust association rules from frequent item sets.

The Apriori algorithm uses an iterative approach, where k - Item sets are used to explore the (k + 1) – Item sets. First, the 1– Item sets are found by scanning the database to calculate the support of each item, and the collecting these Itemsets, which have minsupp$\geq$ support. The resulting set is denoted L1 'then used to find Lz, the 2-itemsets, which is used to find L3, and so continue until no k-Itemsets can be found. Obtaining each Lk requires a complete

analysis of the database. The full description of the Apriori algorithm is summarized in the following steps:

Input: minimum support and a transaction database.

Output: generation of frequent Itemsets

1. $M_i = \emptyset$, i=0
2. C_1=all l-Itemsets in the database
3. L_1= all frequent items of C_1

While (M_i is not empty) do

1. C_{i+1} = Candidate-gen (L_i)
2. L_{i+1} =all frequent Itemsets from C_{i+1}
3. I++
4. Return the union of M_i

End While

III.3.1.3 Generate association rules from frequent itemsets:

A strong association rule satisfies both minsupp and minconf. Association rules can be generated as follows:

- For each frequent itemset, generate all the non-empty subsets ofL.
- For any non-empty subset S of L,the rule "S $\Rightarrow (L - S)$"is generated if the support of (L-S)divided by the support of s is greater than or equal to minsupp.Where(L-S)is the set of elements that belong to L but not as S.

III.3.1.4 Benefits:

The advantages of association rules can be summarized in:

- The possibility of discovering useful knowledge hidden in databases.
- Their ease of understanding, efficiency and simplicity.
- Their unsupervised and general formalism.
- The drilling of association rules is a great success in various fields, whether in commercial, social or human activities.

III.3.1.5 Disadvantages:

Some disadvantages of association rules:

- The discovery of a large number of association rules, most of which are not interesting.
- The search time for frequent Itemsets is enormous.
- The algorithms used have too many parameters, therefore data extraction, for non-experts, becomes complicated.
- A security issue could arise: confidential information can be easily disclosed, using this technique.
- Using just one minsupp could create a rare item dilemma; this means that all the elements of the database are of the same nature. This is not always true.

III.3.1.6 Eclat:

Eclat Is an algorithme for discovering frequent itemsets in a transaction database. It was proposed by Zaki (2001). Contrarily to algorithms such as Apriori, **Eclat** uses a depth-first search for discovering frequent itemsets instead of a breath-first search..**Eclat** is a variation of the

Eclat algorithm that is implemented using a structure called "diffsets" rather than "tidsets".[32]

III.3.1.7 How the algorithms work?

The basic idea is to use Transaction Id Sets(tidsets) intersections to compute the support value of a candidate and avoiding the generation of subsets which do not exist in the prefix tree. In the first call of the function, all single items are used along with their tidsets. Then the function is called recursively and in each recursive call, each item-tidset pair is verified and combined with other item-tidset pairs. This process is continued until no candidate item-tidset pairs can be combined.[33]

example: Consider the following transactions record:

Transaction Id	Bread	Butter	Milk	Coke	Jam
T1	1	1	0	0	1
T2	0	1	0	1	0
T3	0	1	1	0	0
T4	1	1	0	1	0
T5	1	0	1	0	0
T6	0	1	1	0	0
T7	1	0	1	0	0
T8	1	1	1	0	1
T9	1	1	1	0	0

Array III.2 *transactions record*

The above-given data is a boolean matrix where for each cell (i, j), the value denotes whether the j'th item is included in the i'th transaction or not. 1 means true while 0 means false.We now call the function for the first time and arrange each item with it's tidset in a tabular fashion:

k = 1, minimum support = 2

ITEM	TIDSET
Bread	{T1, T4, T5, T7, T8, T9}
Butter	{T1, T2, T3, T4, T6, T8, T9}
Milk	{T3, T5, T6, T7, T8, T9}
Coke	{T2, T4}
Jam	{T1, T8}

Array III.3 *horizontal transaction*

We now recursively call the function till no more item-tidset pairs can be combined:-

k = 2

ITEM	TIDSET
{Bread, Butter}	{T1, T4, T8, T9}
{Bread, Milk}	{T5, T7, T8, T9}
{Bread, Coke}	{T4}
{Bread, Jam}	{T1, T8}
{Butter, Milk}	{T3, T6, T8, T9}
{Butter, Coke}	{T2, T4}
{Butter, Jam}	{T1, T8}
{Milk, Jam}	{T8}

Array III.4 *frequent itemset for two combination*

k = 3

ITEM	TIDSET
{Bread, Butter, Milk}	{T8, T9}
{Bread, Butter, Jam}	{T1, T8}

Array III.5 *frequent itemset for three combination*

k = 4

ITEM	TIDSET
{Bread, Butter, Milk, Jam}	{T8}

Array III.6 *frequent itemset for four combination*

We stop at k = 4 because there are no more item-tidset pairs to combine. Since minimum support = 2, we conclude the following rules from the given dataset:

ITEMS BOUGHT	RECOMMENDED PRODUCTS
Bread	Butter
Bread	Milk
Bread	Jam
Butter	Milk
Butter	Coke
Butter	Jam
Bread and Butter	Milk
Bread and Butter	Jam

Arrav III.7 *final result*

III.4 Advantages over Apriori algorithm:

- Memory Requirements: Since the ECLAT algorithm uses a Depth-First Search approach, it uses less memory than Apriori algorithm.
- Speed: The ECLAT algorithm is typically faster than the Apriori algorithm.

- Number of Computations**:** The ECLAT algorithm does not involve the repeated scanning of the data to compute the individual support values[31]

III.5 Conclusion:

In this chapter, we have presented the association rules, their basic concepts, the Apriori algorithm and its development. In addition, we have discussed the disadvantages and advantages of this approach.

Chapter IV

Proposition and Realization

IV. 1 Introduction:

Bioinformatics is the study of biological information as it passes from its storage site in the genome to the various gene products in the cell. [14] It includes the creation and development of advanced computer technologies for the problems of molecular biology. it has become a solution for known problems of biology in order to model biological elements based on deferent computer tools

One of the most important scientific problems that biology faces today is how to identify the relationship between biological chains, that is, between lean and protein.TF and TFBS to better understand biological activities.

We try in this chapter to present our work on the DNA-Protein liaison between TF and TFBS

IV.2 General architecture :

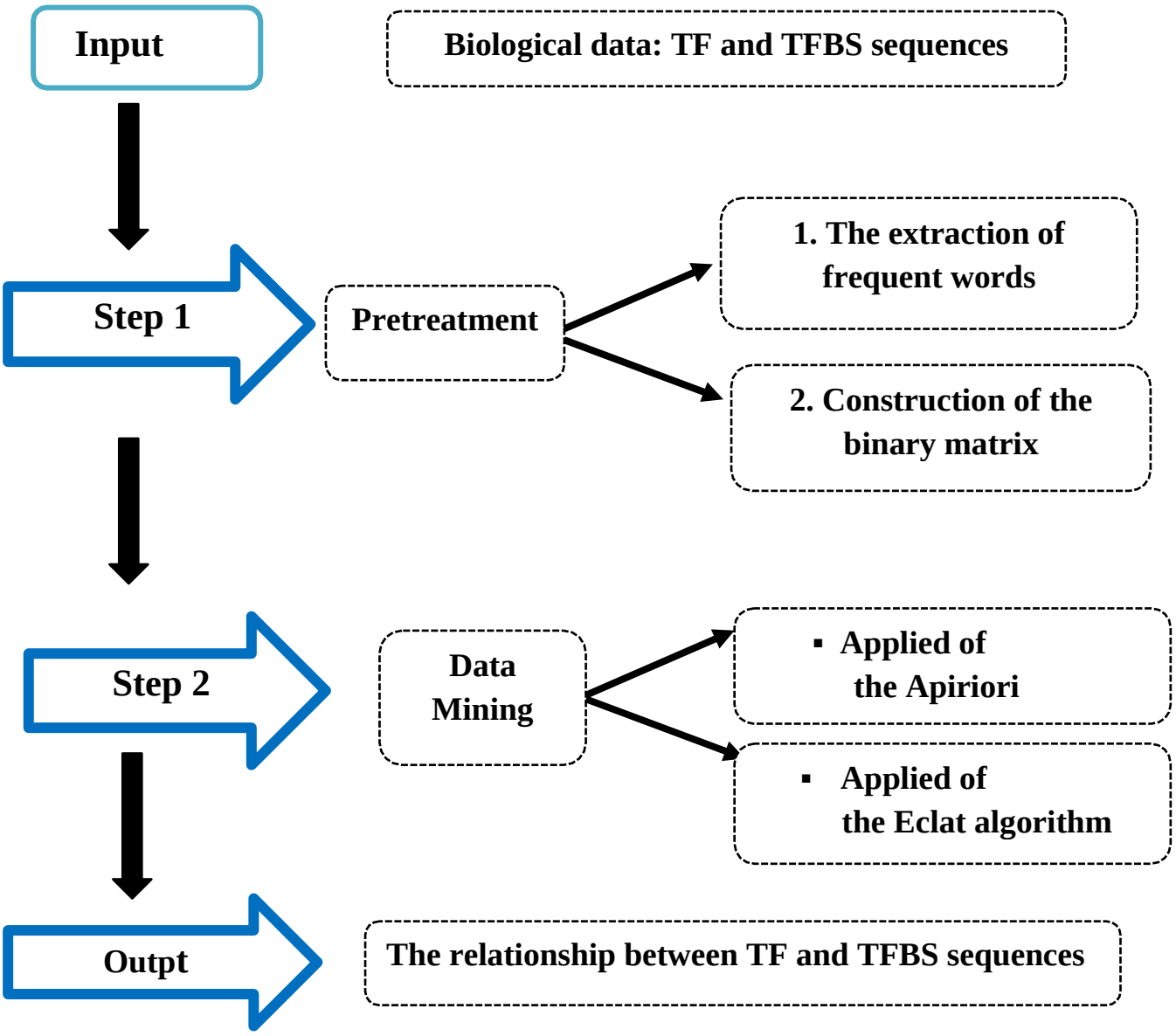


Figure I.1: overall scheme of the implementation

The diagram indicated here explains the steps of our application, as it appeared there are two steps. The first is the preprocessing it is a question of building a base of sequences TF thus all its TFBS then making the extraction of the frequent words then build the binary matrix. The second step apply one of the data mining algorithm here we choose both of Apriori and Eclat algorithm.

IV.3 Python and Bioinformatics:

Biology has become an information science". The emergence and evolution of sequencing technologies shave accompanied new needs for processing and analyzing data, which are massive, complex, interdependent and distributed. Our story begins like many with a mystery ... a DNA sequence of a few hundred base pairs that we will have to translate, identify, place in its context and analyze by seeking the associated scientific literature.

Bioinformatics is an activity where biology and experimentation are the basis for the prediction of biological properties. If programming comes into play in bioinformatics, it is only in a minority way because the data and knowledge manipulated are of such volume that their management requires a computer. What makes a good bioinformatics engineer is not programming; it is a method of reasoning with a lot of hindsight and adapted to the living to produce good predictions.

To avoid the frequent confusion between programming and bioinformatics, it seems useful to inventory how bioinformatics is involved in all areas of biology.

Bio python is a set of tools written in python for computational biology and bioinformatics. There are similar initiatives for most classic languages1: Bio Perl, Bio Java, Bio Ruby, etc. The main functionalities relate to the manipulation of sequences, the recovery and themanipulation of data from classic sources such as ExPASy for genomics or proteomics data, or PubMed for scientific articles. The main strength of Bio python are its 'parsers', modules capable of reading and manipulating the most common standard biological data formats. Widespread the ability to automatically access and use online databases makes it easy to integrate processing into automated analysis workflows.

IV.4 1st implementation: the preprocessing phase:

In this step, we built a sequence base, which contains TF sequences as well as all of its TFBS. Then we cut these sequences into items where each item represents a sub-sequence of 5 letters "4 amino acids for the protein sequences or 4 purine and pyrimidic bases for the DNA sequences.

a) Decomposition step:

For each sub-sequence s of 5 characters in a sequence S **do**

Word-array[i]

Fin

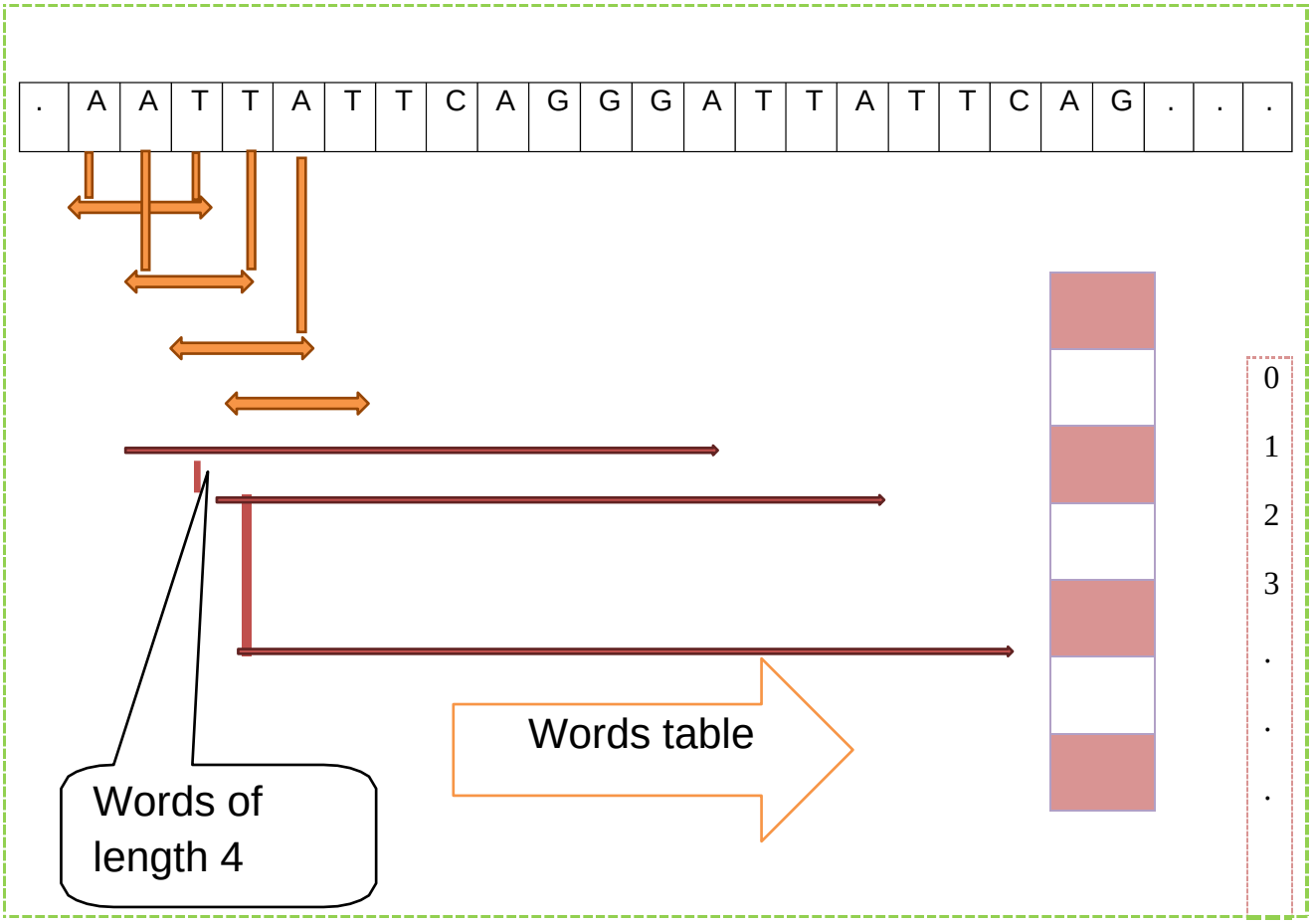


Figure IV.2 : visual algorithm

The TF matrix is a binary matrix that contains a set of TFs and their frequent items. Lines our TF and as columns the items of these TF, each cell contains a 1 if the item exists in the TF or a 0 if it does not exist.

b) Binary algorithm:

For each **TF** in the base **do**

For each item in a sequence **TF** **do**

IF exist in **TF** **So**

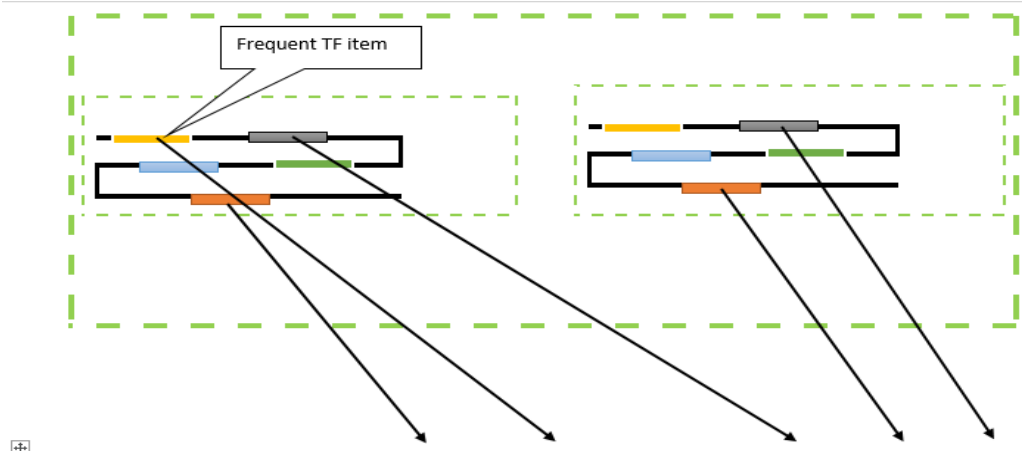
Matrix [i, j] =1

Else

Matrix [i, j] =0

End

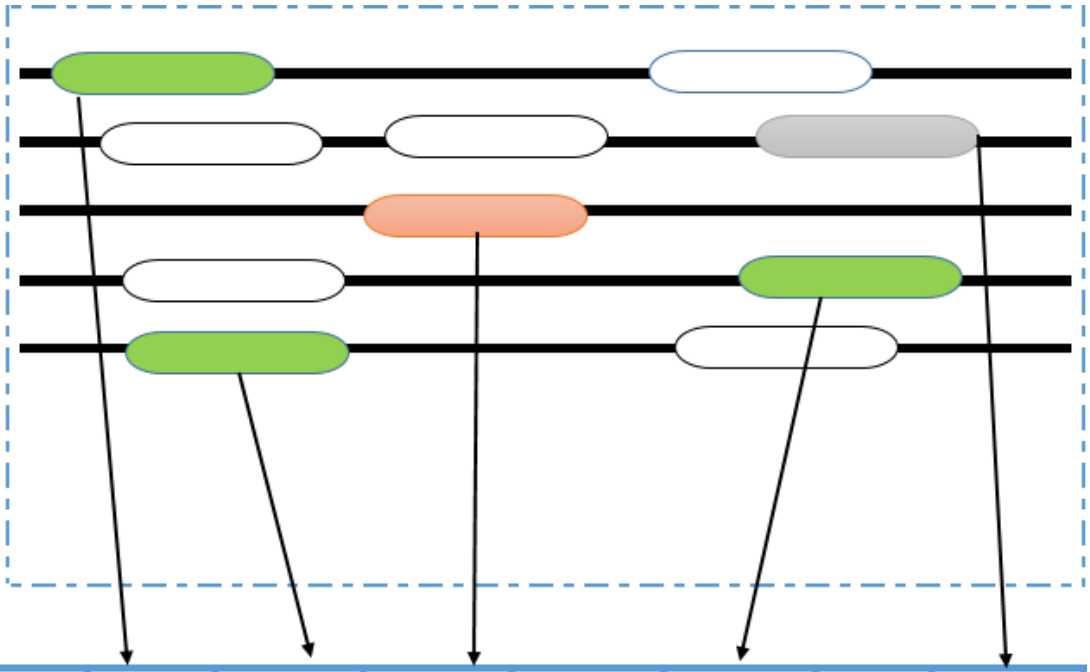
End



Item	PEMVR	EMVRG	LALRV	GRGSG	AAAAG	……..	AAAGA
TF1	0	0	0	1	1	…….	1
TF2	1	1	0	0	1	……..	0
TF3	1	1	0	0	1	…….	1
TF4	1	0	0	0	0	…….	0
TF5	1	1	0	0	1	…….	1

Array IV 1. : frequent TF items

For the TFBS we first concatenate all the TFBS of the same TF, then we build the binary matrix in the way that if the item exists more than 50% we put a 1 otherwise we put a 0.



Item	GATT	ATTTC	TTTCT	CTATA	TAAAT	……..	AAATT
TF1	1	1	0	0	0	……..	0
TF2	1	1	0	0	0	……..	0
TF3	1	0	0	0	0	……..	0
TF4	0	1	0	0	0	……..	0
TF5	1	1	1	0	0	…….	0

Array IV 2. : frequent items of TFBS

Filling algorithm:

For each length-TFBS in the database **do**

For each item I in a length-TFBS sequence **do**

If I exists in TF **Then**

Matrix [i, j] =1

Else

Matrix [i, j] =0

End

End

We continue we build the final matrix, which is the concatenation of the two matrices.

Item	PEMVR	EMVRG	LALRV	GRGSG	AAAAG	……..	AAAGA	GATT	ATTTC	TTTCT	CTATA	TAAAT	……..	AAATT
TF1	0	0	0	1	1	…….	1	1	1	0	0	0	……..	0
TF2	1	1	0	0	1	……..	0	1	1	0	0	0	……..	0
TF3	1	1	0	0	1	…….	1	1	0	0	0	0	……..	0
TF4	1	0	0	0	0	…….	0	0	1	0	0	0	……..	0
TF5	1	1	0	0	1	…….	1	1	1	1	0	0	…….	0

Array IV.3: the final matrix

IV.5 2nd Implementation: the Apriori algorithm

Now we apply the a priori algorithm on the binary matrix, we calculate the support of each TF item and of each frequent TFBS item, we use the following formula :**(formula IV.1)**

The support of an item j= $\sum_i^N MatBin[i,j]/N$ **where N is the number of TF.**

According to the Apriori algorithm we select only frequent items, i.e. items which are a support more than a threshold defined by use here we determine a minimum threshold of 50%.

Here is this table shows that the frequent items where the support≥ min threshold..

Iteration	Items	Support
1- iteration	AAAAG	3
	AAAGA	3
	ATTTC	4
	GATTT	4
	PEMVR	6
2- iteration	AAAAG AAAGA	2
	AAAGA ATTTC	2
	ATTTC GATTT	3
	GATTT PEMVR	3
3- iteration	AAAAG AAAGA ATTTC	2
	AAAGA ATTTC GATTT	2
	ATTTC GATTT PEMVR	3
4- iteration	AAAAG AAAGA ATTTC GATTT	2

Array IV.4: frequent item support

Example of execution of Apriori algorithm with python:

- Confidence of the association rule {'PEMVR'}--> ('GATTT', 'TTTCT') =100.00%
- Confidence of the association rule {'GAAGG'}--> ('GATTT', 'GRGSG') = 100.00%
- Confidence of the association rule{'GATTT'} --> ('GAAGG', 'GRGSG') = 100.00%
- Confidence of the association rule {'AAGAA'} --> ('AAAAG', 'AAAGA') = 50.00%
- Confidence of the association rule {'GAAGG'} --> ('AAAAG', 'AAAGA') = 50.00%
- Confidence of the association rule {'GRGSG'} --> ('AAAAG', 'AAAGA') = 50.00%
- Confidence of the association rule {'PEMVR'}--> ('AAAAG', 'AAAGA') = 50.00%

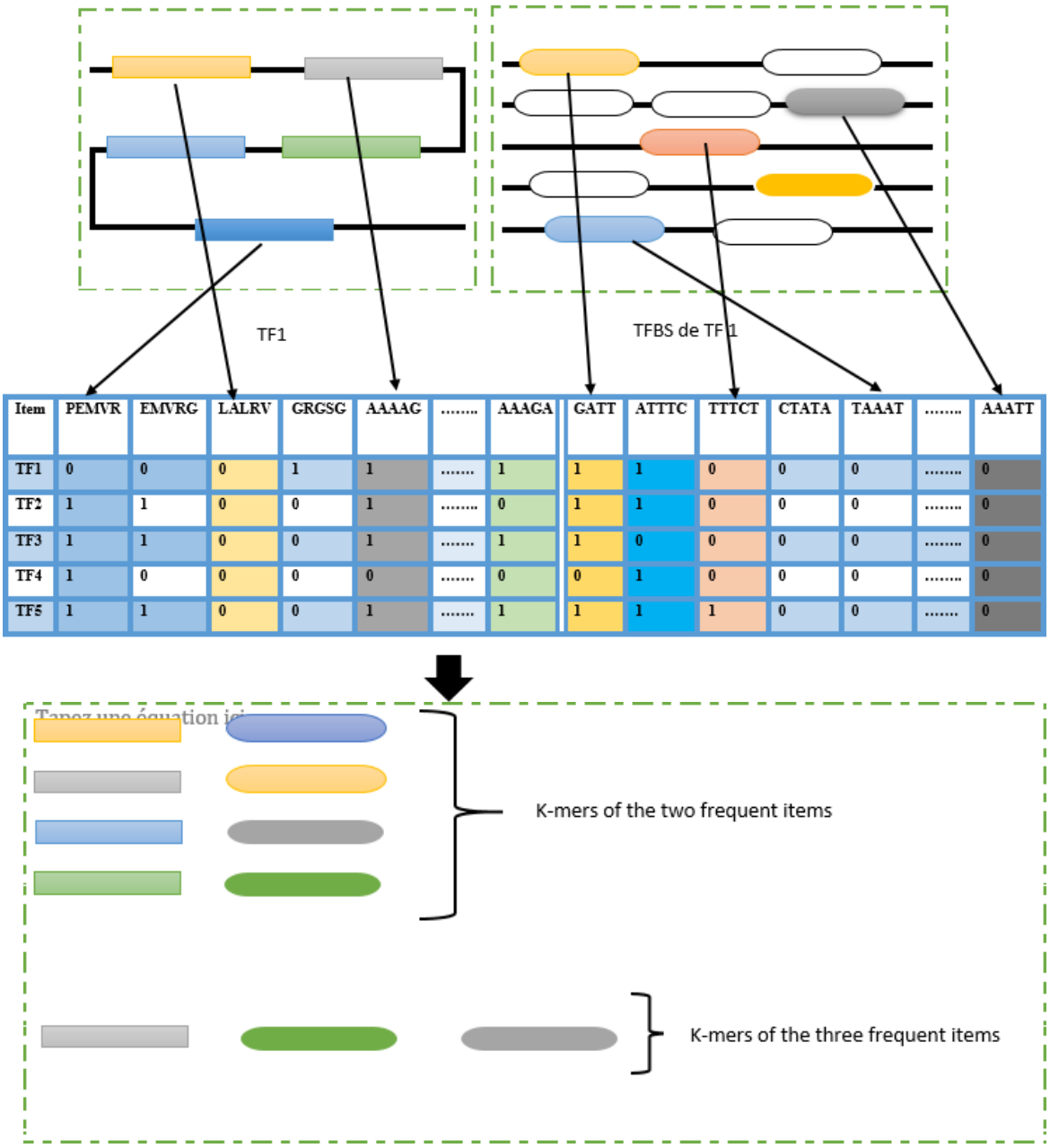


Figure IV.3 : global diagram of Apriori

IV.6 3rd Implementation: the Eclat algorithm

Eclat algorithm finds the elements from bottom like depth first search. Eclat algorithm is very simple algorithm to find the frequent item sets. This algorithm uses vertical database. It cannot use horizontal database. If there is any horizontal database, then we need to convert into vertical database. There is no need to scan the database repeatedly. Eclat algorithm scans the database only once. Support is counted in this algorithm. Confidence is not calculated in this algorithm.

Items	Support
AAAAG PEMVR PEMVR GATTT ATTTC	**4**
PEMVR ATTTC	**3**
AAAGA PEMVR EMVRG GATTT	**2**

Array IV.5: frequent item

Example of execution of Eclat algorithm with python:

- 'AAAAG,PEMVR,PEMVR,GATTT,ATTTC': [1]
- 'AAAGA,PEMVR,EMVRG,GATTT': [2].
- 'PEMVR, ATTTC': [3].
- 'AAAAG,AAAGA,PEMVR,PEMVR,GATTT,ATTTC,TTTCT': [4].
- 'AAAGA, PEMVR, EMVRG, GATTT': [0.2]
- 'AAAAG,AAAGA,AAGAA,AGAAG,GAAGG,AAGGR,GRGSG,GATTT,ATTTC': 0.2.
- 'AAAAG, PEMVR, PEMVR, GATTT, ATTTC': [0.2]
- 'PEMVR,ATTTC': [0.2]
- 'AAAAG, AAAGA, PEMVR, PEMVR, GATTT, ATTTC, TTTCT': [0.2]

IV.7 Conclusion:

We have implemented two algorithms Apriori and Eclat. First take sequences that are in sequence libraries and then cut them into 5-mer length items that contain TF and TFBS. Then we construct a final binary matrix which is the concatenation of two matrices one for the TF and the other for the TFBS.

At the end and after the construction of the binary matrix we apply the Apriori algorithm which determines all the frequent items of TF and TFBS using the support musure, we also apply the Eclat algorithm That scans the data base only once.

GENERAL
Conclusion

❖ General conclusion:

With the current deluge of biological data, computer methods have become essential for biological investigation. Originally developed for the analysis of biological sequences, bioinformatics now covers a wide range of fields including structural biology, genomics and the study of gene expression, while biology and bioinformatics are two very broad fields.

This thesis deals with a problem in bioinformatics "identifying DNA-protein bonds using data mining techniques"; this problem requires the use of data mining in biology. On choose to apply the Apriori algorithm which performs a horizontal count on the database (BFS: Breadth First Search) and the Eclat algorithm which performs a vertical count on the database (DFS: Depth First Search).

The choice of a data mining technique or algorithm to solve a problem strongly depends on the context of the application. the nature of the data and the resources available. An analysis of the data helps to choose the best algorithm

To arrive at a solution for the problem we built a base of the TF sequences and all its TFBS using one of the sequence libraries published in the net TRANSFAC then we follow two steps, which are the preprocessing, and the use of Data mining and towards the end, we arrive at association rules.

The preprocessing phase aims to extract the frequent words after splitting the sequences into items using the K-mer and then constructing the binary matrix. The use phase of the data mining we applied the two methods related to the extraction of the association rules which are the A priori and Eclat algorithm. In the end, very necessary relationships are arrived in several biological activities.

List of abbreviations:

DNA: Disoxyribose Nucleic Acid

RNA: RiboNucleic Acid

RNAm: RiboNucleic Acid messenjer

RNAt: RiboNucleic Acid transfer

TF: Transcription factor

TFBS: Transcription factor binding sites

ECLAT: Equivalence Class Clustering and bottom-up Lattice Traversal

Bibliography

[1]https://www.genome.gov/About-Genomics/Introduction-to-Genomics

[2] https://www.britannica.com/science/cell-biology

[3]https://fr.wikipedia.org/wiki/Chromosomes_humains

[4]https://medlinesplus.gov/genetics/undrestinding/basics/gene

[5] https://www.genome.gov/About-Genomics/Introduction-to-Genomics

[6]https://ghr.nlm.nih.gov/primer/basics/dna

[8] protein-synthesis/a/hs-ran-and-protein-synthesis

[9]https://www.news-medical.net/life-sciences/DNA-Replication-and-Repair.aspx

[10]https://www.nature.com/scitable/definition/promoter-259

[11]https://www.ncb.nlm.nih.gov/bubmed/28623591

[12] JRM, " qu'est ce que la bioinformatique ...",
sur https://www.rts.ch/decouverte/sciences-et-environnement/4637771-qu-est-ce-que-la-bioinformatique-.html, consulted le 03 / 07 / 2020 à 13h 11.).

[13] Jacques van Helden Jacques.van-Helden@univ-amu.fr Aix-Marseille Université, France Technological Advances for Genomics and Clinics (TAGC, INSERM Unit U1090).

[14] JRM, " Bioinformatics...", https://en.wikipedia.org/wiki/Bioinformatics consulted le 04 / 07 / 2020 à 09 h 35).

[15] JRM, " Introduction et définition de Bioinformatique ...", http://biochimej.univ-angers.fr/Page2/BIOINFORMATIQUE/7ModuleBioInfoJMGE/4IntroDefBioInfo/1IntroDefBioInfo.htm, consulted le 04 / 07 / 2020 à 11 h 09).

[16] JRM, " Distance de Hamming ...",
https://en.wikipedia.org/wiki/Hamming_distance,consulted le 01 / 08 / 2020 à 09 h 20

[17] Vincent Derrien ; thèse Heuristiques pour la résolution du problem d'alignement multiple 2009

[18] DrM D Macleod MA PhD MIEEE, in Telecommunications Engineer's Reference Book, 1993

[19] GONZALO NAVARRO University of Chile ACM Computing Surveys, Vol. 33,

No. 1, March 2001.
[19] JRM, " Edit_distance...", https://en.wikipedia.org/wiki/Edit_distance#cite_note-navarro-1,consulted le 01 / 08 / 2020 à 10 h 20).
[20] JRM, "Substitutions matrices...", https://www.researchgate.net/publication/228028830_Substitution_Matrices, consulted le 01 / 08 / 2020 à 10 h 35)
[21] JRM, "Substitutions matrices +Identify matrix ...", https://en.wikipedia.org/wiki/Substitution_matrix#Identity_matrix,consulted e 01 / 08 / 2020 à 10 h 42)
[22] JRM, " Identify matrix ...", https://en.wikipedia.org/wiki/Identity_matrix , l consulted e 01 / 08 / 2020 à 10 h 55)
[23] Bioinfo-FR .net
[24] Biochimie.univ
[26] Alignements multiples Celine Brochier celine.brochier@up.univ-mrs.fr http://194.57.197.233:800
[27] JRM, " Multiple sequence alignment...", https://en.wikipedia.org/wiki/Multiple_sequence_alignment, consulted le 12 / 08 / 2020 à 14h 32)
[29] Fabrice Muhlenbach. Méthode de regroupement par graphe de voisinage. Extraction et gestion des connaissances (EGC'2009)
[31] D. Chessel, J. Thioulouse & A.B. Dufour, Introduction à la classification hiérarchique 2007
[32] https://www.philippe-fournier-viger.com/spmf/Eclat_dEclat.php
[33] https://www.geeksforgeeks.org/ml-eclat-algorithm/