**République Algérienne Démocratique et Populaire**
**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**
**Université A.MIRA-BEJAIA**


**Faculté des Sciences Exactes**
**Département d'Informatique**

# THÈSE

**Présentée par**

**KAMEL Mohamed**

**Pour l'obtention du grade de**

## DOCTEUR EN SCIENCES

**Filière : Informatique**

**Option : Cloud Computing**

**Thème**

**Contribution to Bioinformatics search algorithms**

Soutenue le : 13/04/2022

Devant le Jury composé de :

| Nom et Prénom | Grade | | |
|---|---|---|---|
| **Mr BELAID Ahror** | Professeur | Univ. de Béjaia | Président |
| **Mr TARI Abdelkamel** | Professeur | Univ. de Béjaia | Rapporteur |
| **Mr HEMMAK Allaoua** | MCA | Univ. de M'sila | Examinateur |
| **Mr SAYAD Lamri** | MCA | Univ. de M'sila | Examinateur |

**Année Universitaire** : 2021/2022

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
University of Abderrahmane MIRA - Bejaia


Faculty of Exact Sciences
Department of Computer Science

# Thesis

Submitted by

**Mohamed Kamel**

In partial fulfillment of the requirements for the degree of

## Doctor of science

Field : Computer Science

Option : Cloud Computing

Subject

# Contribution to Bioinformatics search algorithms

Defended publicly on : 13/04/2022

In front of the jury composed of

| | | | |
|---|---|---|---|
| BELAID Ahror | Professor | University of Bejaia | President |
| TARI Abdelkamel | Professor | University of Bejaia | Supervisor |
| HEMMAK Allaoua | MCA | University of M'sila | Reviewer |
| SAYAD Lamri | MCA | University of M'sila | Reviewer |

Academic Year : 2021/2022

# ACKNOWLEDGMENTS

I would like to give special thanks to everyone who showed and provided their advice, support and motivations at all stages of my thesis;

Prof. Dr. Abdelkamel TARI (my supervisor), Thank you for providing a stress-less supervising and support,

Prof. Dr. Andrade Miguel (CBDM Group Leader), Thank you for your collaboration and support.

Rabah, Makhlouf, Naçereddine, Meftah, Rached, Ali, Bilal, Hichem (co-workers), Thank you for your company during my studies,

Thank you my parents, brothers and sisters for your encouragement and support.

Thank you my wife for your support and patience.

CONTENTS

# LIST OF FIGURES

## LIST OF TABLES

Chapter

# 1

# INTRODUCTION

Bioinformatics is by definition an interdisciplinary field of science. That feature makes the domain extremely attractive. Research and development in areas like biology, computer science, information engineering, mathematics and statistics fields is a challenging task. Huge benefits are possible, However, if these areas are effectively used to solve basic biological problems.

It is difficult to identify the origins of bioinformatics, both as a concept and as a discipline. The expression was used by the Dutch theoretical biologist Paulien Hogeweg as early as 1977 (Mahdavi, 2011) when she described its main research field as bioinformatics and established a bioinformatics group at the University of Utrecht.

Human genome sequencing began in 1994. The release of a draft human DNA took 10 years of collaborative effort by many research groups from different countries. Modern technologies allow a full genome to be sequenced in few days.

Sequenced biological data are flooding in at an unprecedented rate. For example as of December 1982, the GenBank repository of nucleic acid sequences contained 606 sequences, in October 2019 it contained 216,763,706 sequences. From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.

That huge amount of data cannot be processed by humans alone. Computational solutions provided by bioinformatics can be used for various biological problems. Analysis of human genome and sequences in general is one of the most important tasks in bioinformatics.

## 1.1 MOTIVATION

Research areas in bioinformatics are many, one of them is repeats analysis. Several algorithms have been proposed in the literature, however we would like to focus on low complexity regions in a global manner. Low complexity regions (LCRs) are amino acid sequences that contain repeats of single amino acids or short amino acid motifs (Toll-Riera et al., 2011). Our method is intended to emphasis the analysis on low complexity regions of full human proteomes as well as other species' proteomes in order to get more globular view with most deeper findings.

Along with low complexity regions, our focus will be also on short tandem repeats (STR) which are short repetitive and consecutive sequences. There is a generally accepted idea that LCRs are disordered regions. However, some structural analyses suggest that LCRs can acquire structures, which can be flexible and dependent on the context.

The work described in this thesis can be described as *structural proteomics*, where we focus on structure rather than function, while there is a tendency towards structural disorder in LCRs, various examples, and particularly homo-repeats of single amino acids, suggest that very short repeats could adopt structures.

Tandem repeats are known to be found in many proteins with basic functions or related to some diseases (Baxa et al., 2006; Hackman, Vihola, and Udd, 2003; Itoh-Satoh et al., 2002; Machado, Sunkel, and Andrew, 1998; Nelson and Eisenberg, 2006; Siwach and Ganesh, 2008), Therefore, our structural study of protein sequences could be useful for functional studies or diseases research.

## 1.2 RESEARCH PROBLEMS

When working with low complexity regions, usually short repeats are found with different lengths and different number of mutations. this work could answer many questions including:

**Q1 : Does LCRs has any structure or particular pattern globally ?** There is a generally accepted idea that LCRs are considered as disordered regions. However, some structural analyses suggest that LCRs can acquire structures.

**Q2 : Is it possible to find all short tandem repeats ?** Here we have to find out if we could develop an algorithm that has a complexity that finds all short tandem repeats with all lengths and all number of mutations.

**Q3 : How many repeats are there in each proteome for different lengths and different number of mutations ?** This question leads us to first find the all repeats for proteomes under study, those repeats will then be grouped according to their lengths as well as their number of mutations.

**Q4 : Is there any relation between repeats ?** After finding all repeats, categorizing them, repeats will be compared to each other internally to find out if any hidden relations appears.

**Q5 : Is protein repeats in human proteome similar to other species' proteomes ?** This question leads us to do the comparison between repeats externally. We will focus in our study on human proteome as well as some other species proteomes.

Our research is intended to provide a fast algorithm, and a web tool that is simple and easy to use yet productive and helpful to researchers wanting to analyse their protein sequences for any potential tandem repeats found in them.

Exploring protein repeats in more different ways will add more findings to the literature. This study may help understanding, even with a tiny amount, of the huge complexity of processes occurring within living systems, from the genome to the nucleus to the cell to the whole organism.

## 1.3 CONTRIBUTIONS

Our contributions are a repeatability scanner algorithm (RES Algorithm), a web tool accessible at (http://cbdm-01.zdv.uni-mainz.de/~munoz/res/) that allows to find regions with approximate short repeats in protein sequences, and helps to characterize the variable use of LCRs and compositional bias in different organisms, along with a full proteomes analyses of several species including Human, Yeast and many others.

## 1.4 OUTLINE

The next two chapters serve as an introduction to the research field of this thesis, we organized the chapters in a way that we believe will cover most important background information needed to understand this work. In the next chapter, we will present some basic concepts of bioinformatics as well as a brief information on molecular biology. The following chapter is dedicated to give an overview on

tandem repeats research field. In chapter 4, we will explain our approach for short tandem repeats search. In the fifth chapter, the results of our proposed approach (RES algorithm) are compared and analyzed. Lastly, we end this thesis by a set of conclusions.

Chapter

# 2

# BIOINFORMATICS

## 2.1 INTRODUCTION

Bioinformatics is a multidisciplinary research domain that is the intersection of biological and computational sciences. Although the domain is not well defined, we could consider that Bioinformatics involves molecular biology and genetics, computer science, mathematics, and statistics. In the most cases, bioinformatics is concerned by analysing and organizing biological data (Rana and Vaisla, 2012).

Because of the huge amount of data generated in molecular biology field, most of Bioinformatics' studies focus on structures and functions of genes and proteins.

Bioinformatics is defined by (Huerta et al., 2000) as "Research, development, or application of computational tools and approaches for expanding the use of

biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data".

Computational Biology is also defined by (Huerta et al., 2000) as "The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems".

Studying Bioinformatics involves studying many Molecular Biology concepts, including and not limited to Cells, DNA and Proteins. Before getting an overview on those basic concepts in the following four sections, we will see what Biology and Molecular Biology are.

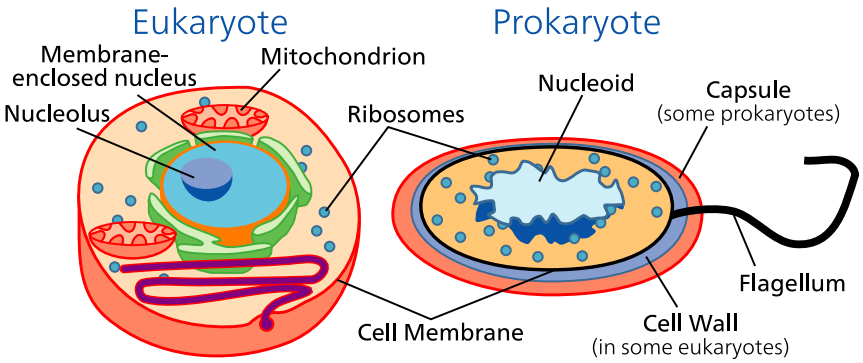
Biology is the natural science that studies structure, chemistry, interactions of molecules, physiological mechanisms, growth and development of life and living organisms. In biology, cells are regarded as the fundamental building blocks of life, genes as the fundamental unit of inheritance, and development of species is considered as driven by evolution process.

Biology is composed of many sub-disciplines, one of them is Molecular biology. In Molecular biology studies are focused on molecular basis of biological activity in and between cells, including mechanisms, molecular synthesis, interactions and modifications.

## 2.2 CELLS

Cells are considered as fundamental units of living organisms. The cell contains a nucleus, mitochondria and chloroplasts, ribosomes, endoplasmatic reticulum, etc, see Figure 2.1. The nucleus is important since it includes chromosomes which include the DNA. The DNA is basically a model for the cell because it encodes the information required to synthesize proteins. Molecular biologists would like to understand how biology works in human with the goal to fight diseases like

cancer. One can study organisms that are considered as simple such as yeasts to understand how human biology works. It is true that unicellular yeasts are very distinct from humans with roughly 1014 cells. However, the DNA is similar among all living creatures. For example, humans share 99% of DNA with chimps. Naturally, we aim to know what information contained in the 1% of DNA that is so critical to cause all those distinguishing characteristics of humans.




(Source: US National Center for Biotechnology Information)

Figure 2.1: Eukaryotic cell (left) and prokaryotic cell (right)

According to the most recent facts, the tree of life has three major branches.

Carl Woese et al., in 1990, classified biological organisms into a Three-domain system (Woese and Fox, 1977; Woese, Kandler, and Wheelis, 1990), with this classification living organisms are divided into archaea, bacteria, and eukaryote domains. particularly, it separates prokaryotes into two groups, the so called Bacteria (before Eubacteria) and Archaea(before Archaebacteria). Woese proposed that these two groupings and the eukaryotes each originated from a different ancestor with poorly developed genetic machinery, usually called progenote. Based on these primary definitions, Woese treated each as a domain, each domain divided into many kingdoms. At first he was referring to the three primary phylogenic grouping by "kingdom", to finally adopt the term "domain" in 1990. (Woese, Kandler, and Wheelis, 1990)

### 2.2.1 *Archaea*

The Archaea are considered as prokaryotic, they have no nuclear membrane, different biochemistry from bacteria (Woese and Fox, 1977; Woese, Kandler, and Wheelis, 1990). The Archaea have a distinct ancient evolutionary history, they are considered to be some of the oldest living organisms on Earth, especially their ability to feed on inorganic matter caused by diverse exotic metabolisms. At first they were classified as exotic bacteria but later the term archaebacteria was adopted, it's possible to distinguish them bacteria is to see the extreme, harsh environment in which they thrive.

Those are some examples of archaeal organisms:

- **Methanogens**, microorganisms that produce methane gas
- **Halophiles**, they are able to live in very salty water
- **Thermoacidophiles**, they are able to thrive in acidic high-temperature water

### 2.2.2 *Bacteria*

The Bacteria is also considered as prokaryotic; mainly consists of cells with bacterial rRNA, with no nuclear membrane, and whose membranes mainly contains diacyl glycerol diester lipids (Woese and Fox, 1977; Woese, Kandler, and Wheelis, 1990). Because many types of bacteria live in the same humans environments, they were the first discovered prokaryotes, for short period of time, they were named Eubacteria ("true" bacteria), Archaea, then, was classified as different clad.

Bacteria are considered easier to grow in laboratories than Archaea, that is one of the reasons that makes studies focused more on Bacteria.

Here are some of bacteria examples:

- **Cyanobacteria**, photosynthetic bacteria, live in water and moist soils.
- **Spirochaetes**, spiral-shaped bacteria, some of them are serious pathogens for humans, causing diseases such as syphilis and Lyme disease.
- **Actinobacteria**, group of bacteria with great benefits to humans.

### 2.2.3 *Eukarya*

Eukarya are unique organisms, their cells are characterized by membrane-bound nucleus (eukaryotes, eukaryotes) (Woese and Fox, 1977; Woese, Kandler, and Wheelis, 1990). Eukarya include many large unicellular organisms.

Some examples of eukaryotic organisms:

- **Kingdom Animalia**, includes animals
- **Kingdom Fungi**, includes fungis
- **Kingdom Plantae**, includes plants
- **Saccharomycotina**, includes true yeasts
- **Basidiomycota**, includes mushrooms
- **Magnoliophyta**, includes flowering plants

## 2.3 DNA

DNA stands for deoxyribonucleic acid. DNA is a very long molecule composed of two polynucleotide chains that twist around each other to form a double-helix. The double-helix is composed of sugars and phosphates, each sugar has an attached base. There are four types of bases: adenine ($\boldsymbol{A}$), thymine ($\boldsymbol{T}$), cytosine ($\boldsymbol{C}$), guanine ($\boldsymbol{G}$). The DNA consists of two strands, each base from the first strand

bonds with a corresponding base on the other strand. A bonds to T and C bonds to G. The order of these bases makes the genetic code, or instructions of DNA. Human DNA has around 3 billion base pairs, and more than 99% of those bases are the same in all people (Crow, 2002). Three bases makes an amino acid. A sequence of amino acids makes a protein. The functional subsequence of DNA that forms a protein is called a gene. A DNA illustration is given in Figure 2.2.

### 2.3.1 *DNA structure*

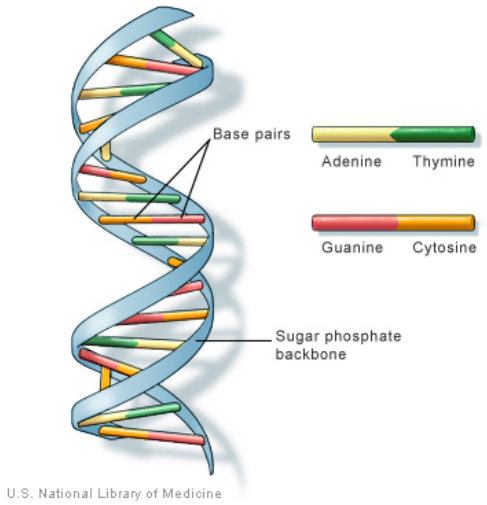



Figure 2.2: DNA structure illustration

DNA is a very long molecule and doesn't fit inside the cell if it's not packed. When the DNA is packed, it makes a specific structure called chromosome, see Figure 2.3. Each DNA molecule packs and forms a single chromosome. Human cell's nucleus contains 23 pairs of chromosomes.

### 2.3.2 *DNA discovery*

The first person to observe the DNA is Frederich Miescher in 1869, a Swiss physician . But for several years, the molecule's importance hasn't been discovered, until it was discovered in 1953 by James Watson, Francis Crick, Maurice Wilkins and Rosalind Franklin, where they found the double helix structure of DNA and the potential importance of the biological data that it may contains. Later in 1962, Watson, Crick and Wilkins where awarded a Nobel Prize in Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material". (Dahm, 2005)

### 2.3.3 *DNA sequencing*

DNA sequencing is a technique that is used by scientists to find out order of bases in a DNA sequence, it can be used on genes, chromosomes and genomes. The complete sequence of the human genome was published in 2001 (Dahm, 2005).

### 2.3.4 *Chromosomes*

The chromosome is the packaged structure form of DNA, it's found inside cell's nucleus. The DNA is coiled tightly around a histones proteins that support its structure.

When the cell is not dividing, it's not possible to see chromosomes, even with microscope. However, when cell starts dividing, chromosome's DNA becomes very packed and that makes it visible under a microscope.

Chromosome has many parts or zones, near the center there is a shrink area called centromere, it divides the chromosome into two arms, a short arm and a long arm, the short arm is called "p arm" while the long arm of the chromosome

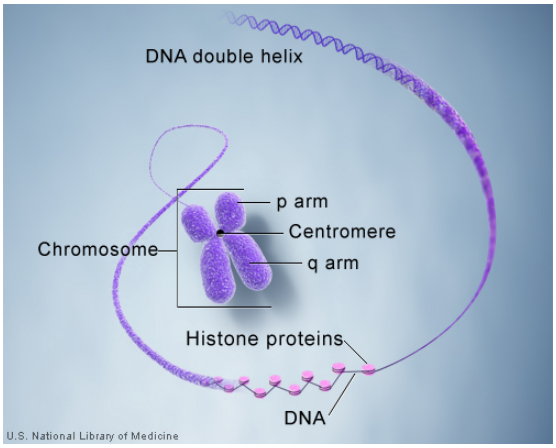



Figure 2.3: Chromosome illustration

is called "q arm", see Figure 2.3. The position of centromere on the chromosome affects how the chromosome's shape is characterized, also, it can be used by researchers to ease locating some specific genes.

## 2.4 RNA

Ribonucleic acid (RNA) is one of the most important biological macro molecules indispensable to all known life forms (besides DNA and proteins). A central principle in molecular biology describes the flow of genetic information in a cell goes from DNA to RNA then to proteins: "DNA Encodes RNA, RNA Encodes Protein". Proteins have important roles in the cell, for example as enzymes, as structural components, signal transmission and many others. DNA is considered as the call's data storage, it contains various genetic information necessary for the cell to perform diverse necessary tasks, such as growth, taking nutrition and reproduction. In this role, RNA is considered to be the first copy of DNA. When a certain protein is needed to be produced, the cell activates a portion of DNA that corresponds to that protein, namely the gene, and produces many copies of that portion of DNA in a form of messenger RNA, or mRNA. After that, those copies of mRNA are used to produce proteins.

### 2.4.1 *mRNA*

Messenger RNA (mRNA) is a family of RNA molecules that takes genetic information from DNA to the ribosome, during this process the amino acid sequence of the protein is specified based on gene expression. The RNA polymerase enzyme transcribes genes into primary transcriptional mRNA, resulting in mRNA, then finally, it is translated into an amino acid sequence that forms the protein.

### 2.4.2 *tRNA*

Transfer RNA (tRNA) molecule plays an important and necessary role in translation, tRNA consists of a single RNA strand with typically 76 to 90 nucleotides (Sharp et al., 1985), its role is bonding between mRNA and the amino acid sequence of proteins by providing an amino acid to the ribosome based on a codon (3-nucleotide sequence) in a mRNA.

### 2.4.3 *rRNA*

Ribosomal RNA (rRNA) is the main construction unit for ribosomes, it is mandatory for all living forms, it's the part of ribosome that is responsible for protein synthesizing. rRNA is synthesized in the nucleus from ribosomal DNA (rDNA), and with other ribosomal proteins they bound with each other to form ribosomes (Berk et al., 2013). The majority of ribonucleic acid found in most cells are rRNA, it makes around 80% from total cellular RNA.

### 2.4.4 *Flow of Information*

The earliest step is transcription of mRNA starting from DNA regions whether they are coding or noncoding ones, After that the mRNA is first processed by removing the noncoding regions, called introns, leaving only coding regions, also called exons, which will be joined together in a later stage to form mature mRNA, the mature mRNA is then prepared for export outside the nucleus, particularly to cytoplasm where it will be used in protein construction. See Figure 2.4 for an illustration.

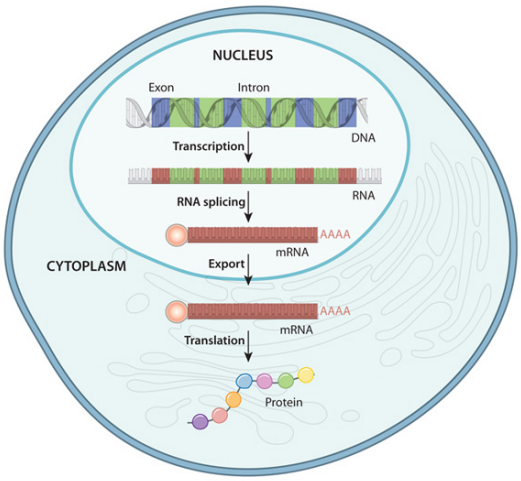



Figure 2.4: An overview of the flow of information from DNA to protein in a eukaryote

## 2.5 PROTEINS

Protein is a complex and large molecule that is vital to living organisms, Proteins are necessary for function, regulation and structure of the body's organs and tissues, and they perform the majority of work inside the cell.

Proteins are long chains composed of tens, hundreds or even thousands of small units called amino acids. The order of those amino acids determines function and spatial structure of the protein.

### 2.5.1 *Amino acids*

There exist 20 different amino acids that can be found in a protein, each amino acid have a different chemical structure (Vickery and Schmidt, 1931). Amino acids names, abbreviation and some characteristic properties are shown in Table 2.1. The Table 2.2 shows different genetic codes for each amino acid.

| Amino acid | Abbreviation | Symbol | Basic properties |
|---|---|---|---|
| Alanine | Ala | A | Nonpolar, hydrophobic |
| Arginine | Arg | R | Polar, hydrophilic |
| Asparagine | Asn | N | Polar, hydrophilic |
| Aspartic acid | Asp | D | Polar, hydrophilic |
| Cysteine | Cys | C | Polar, hydrophilic |
| Glutamine | Gln | Q | Polar, hydrophilic |
| Glutamic acid | Glu | E | Polar, hydrophilic |
| Glycine | Gly | G | Polar, hydrophilic |
| Histidine | His | H | Polar, hydrophilic |
| Isoleucine | Ile | I | Nonpolar, hydrophobic |
| Leucine | Leu | L | Nonpolar, hydrophobic |
| Lysine | Lys | K | Polar, hydrophilic |
| Methionine | Met | M | Nonpolar, hydrophobic |
| Phenylalanin | Phe | F | Nonpolar, hydrophobic |
| Proline | Pro | P | Nonpolar, hydrophobic |
| Serine | Ser | S | Polar, hydrophilic |
| Threonine | Thr | T | Polar, hydrophilic |
| Tryptophan | Trp | W | Nonpolar, hydrophobic |
| Tyrosine | Tyr | Y | Polar, hydrophilic |
| Valine | Val | V | Nonpolar, hydrophobic |

Table 2.1: Amino acids list and their basic properties

### 2.5.2 *DNA/RNA codes for Proteins*

The genetic code is considered as a dictionary for translation process of linear sequences of mRNA bases into a linear sequences of amino acids. this process is carried out by cell organelles, namely ribosomes, which are located in the

cytoplasm and composed of RNA and proteins. The basic principle of genetic code is discovered by Francis Crick, and based on the fact that there exist 20 different amino acids (see Table 2.2). He assumed that combining acid nucleics "A", "C", "G" and "T" with a certain word length give us coding units of protein. The minimum word length that has the capacity of coding all 20 amino acids as well as the transcription/translation signaling sequences in DNA is 3, because: $4\ (number\ of\ acid\ nucleics)^{3\ (word\ length)} = 64$. That basic coding unit, which is of length 3, is called a codon. Nirenberg and Matthaei have discovered the nature of codons and most of the genetic code through a number of experiments by creating artificial RNA strands and observing the resulting amino acid sequences (Nirenberg and Matthaei, 1961). Afterwards, a work by Har Gobind Khorana identified the rest of genetic code. Shortly thereafter, Robert W. Holley determined the structure of transfer RNA (tRNA), the molecule that ease the process of translating RNA into protein, All genetic code is presented in Table 2.2. This work led to a joint Nobel Prize in Physiology or Medicine 1968.

### 2.5.3 *Protein structure*

The shape of protein is important to its function. To understand how the protein gets its final shape or conformation, we need to understand the four levels of protein structure: primary, secondary, tertiary, and quaternary.

#### 2.5.3.1 *Primary structure*

The primary structure is the sequence of amino acids itself. For example the peptide hormone insulin that helps to maintain blood sugar within a healthy range, which is produced inside pancreas, has tow peptide chains, A-chain ($GIVEQCCTSICSLYQLENYCN$) and B-chain ($FVNQHLCGSHLVEALY LVCGERGFFYTPKT$), they have two interchain disulfide bonds, A-chain has

| | | Second base | | | |
|---|---|---|---|---|---|
| | | U | C | A | G |
| First base | U | UUU (Phe/F) Phenylalanine<br>UUC (Phe/F) Phenylalanine<br>UUA (Leu/L) Leucine<br>UUG (Leu/L) Leucine, Start | UCU (Ser/S) Serine<br>UCC (Ser/S) Serine<br>UCA (Ser/S) Serine<br>UCG (Ser/S) Serine | UAU (Tyr/Y) Tyrosine<br>UAC (Tyr/Y) Tyrosine<br>UAA Stop (Ochre)[B]<br>UAG Stop (Amber)[B] | UGU (Cys/C) Cysteine<br>UGC (Cys/C) Cysteine<br>UGA Stop (Opal)[B]<br>UGG (Trp/W) Tryptophan |
| | C | CUU (Leu/L) Leucine<br>CUC (Leu/L) Leucine<br>CUA (Leu/L) Leucine<br>CUG (Leu/L) Leucine, Start | CCU (Pro/P) Proline<br>CCC (Pro/P) Proline<br>CCA (Pro/P) Proline<br>CCG (Pro/P) Proline | CAU (His/H) Histidine<br>CAC (His/H) Histidine<br>CAA (Gln/Q) Glutamine<br>CAG (Gln/Q) Glutamine | CGU (Arg/R) Arginine<br>CGC (Arg/R) Arginine<br>CGA (Arg/R) Arginine<br>CGG (Arg/R) Arginine |
| | A | AUU (Ile/I) Isoleucine<br>AUC (Ile/I) Isoleucine<br>AUA (Ile/I) Isoleucine<br>AUG (Met/M) Methionine, Start | ACU (Thr/T) Threonine<br>ACC (Thr/T) Threonine<br>ACA (Thr/T) Threonine<br>ACG (Thr/T) Threonine | AAU (Asn/N) Asparagine<br>AAC (Asn/N) Asparagine<br>AAA (Lys/K) Lysine<br>AAG (Lys/K) Lysine | AGU (Ser/S) Serine<br>AGC (Ser/S) Serine<br>AGA (Arg/R) Arginine<br>AGG (Arg/R) Arginine |
| | G | GUU (Val/V) Valine<br>GUC (Val/V) Valine<br>GUA (Val/V) Valine<br>GUG (Val/V) Valine | GCU (Ala/A) Alanine<br>GCC (Ala/A) Alanine<br>GCA (Ala/A) Alanine<br>GCG (Ala/A) Alanine | GAU (Asp/D) Aspartic acid<br>GAC (Asp/D) Aspartic acid<br>GAA (Glu/E) Glutamic acid<br>GAG (Glu/E) Glutamic acid | GGU (Gly/G) Glycine<br>GGC (Gly/G) Glycine<br>GGA (Gly/G) Glycine<br>GGG (Gly/G) Glycine |

Table 2.2: Genetic code

an internal disulfide bonds (see Figure 2.5). The sequence, or primary structure, of A-chain and B-chain are unique to insulin.

Every protein has its unique sequence determined by the source gene in DNA sequence, so any change in the gene's nucleotide sequence may cause a change in amino acid sequence, which may finally alter structure and function of produced protein. Sometimes that change, or mutation, makes no difference at all, or very little difference, In some cases mutation cause amino acids to be incorporated, which make the protein more effective. More frequently, it causes the protein to be less effective in doing its function.

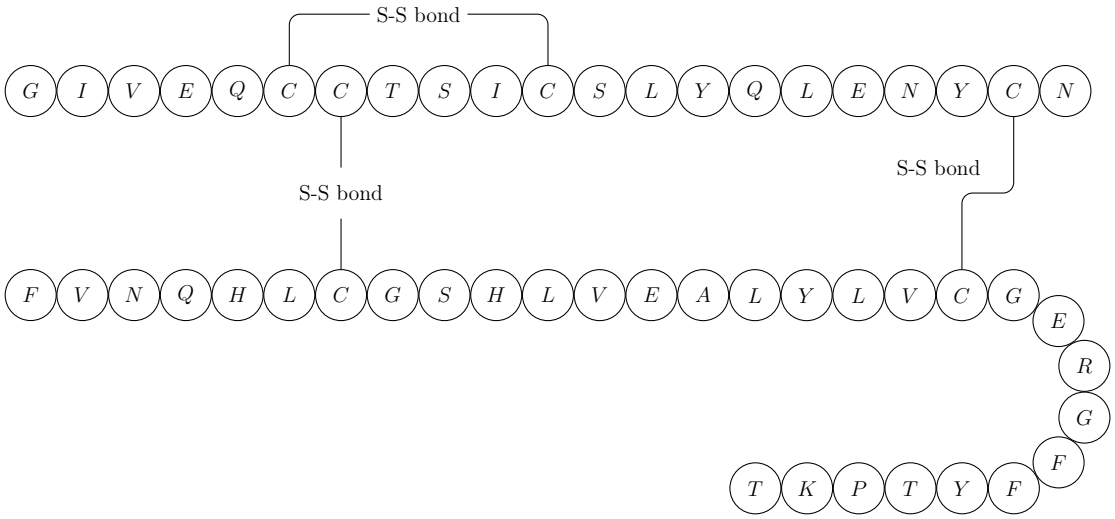



Figure 2.5: Primary structure of insulin

#### 2.5.3.2 *Secondary structure*

The secondary structure of a protein appears when some regions in the sequence are folded. $\alpha$-helix and $\beta$-pleated sheet are the most frequent types of secondary structures (see Figure 2.6).

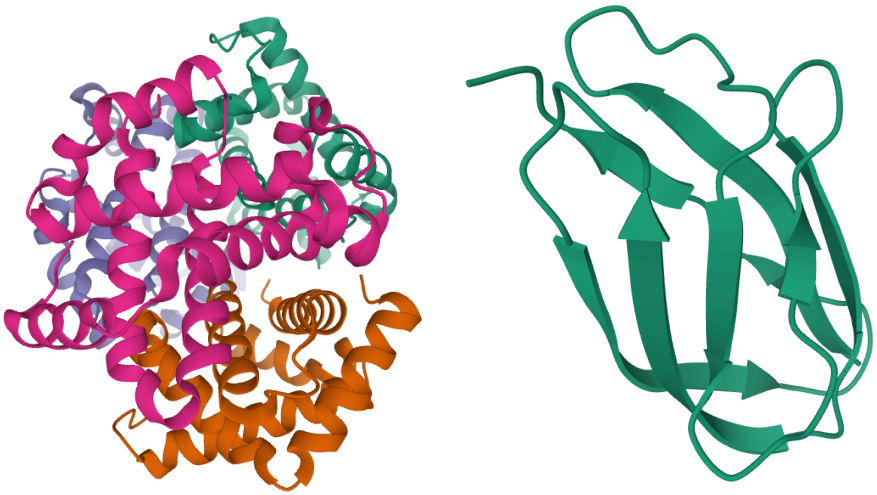


Figure 2.6: $\alpha$-helix (left) and $\beta$-strand (right) structures

In $\alpha$-helix structure, the protein chain is coiled like a spring, the word alpha($\alpha$) means that if you look from one end of the spring to the other you will see turns that look like alpha shapes, the direction of turns is clockwise as you move forward

on the spring, these turns are called helical turns, which are approximately 3.6 amino acids long each, turns are held together by hydrogen bonds.

In $\beta$-pleated sheet structure, the protein chain is folded so that folds lie alongside each other, chain folds are also held to each other by hydrogen bonds.

#### 2.5.3.3 *Tertiary structure*

The tertiary structure of a protein is how the whole chain, including $\alpha$-helices and $\beta$-pleated sheets as well as other secondary structures, is folded into its final 3-dimensional shape. This structure is formed due to interactions between polypeptid chain parts called R groups which are side chains sticking out along the peptide chain.

Figure 2.7: Tertiary structure

R groups interactions are varying from weak to strong, all of these combinations of interactions participate in the protein's final 3D structure, if the protein's 3D structure is lost, the protein might no longer achieve its function.

#### 2.5.3.4 *Quaternary structure*

Quaternary structure is formed when tow or more identical or different protein subunits, called polypeptides, interact with each other, see Figure 2.8. We say

that the protein is a dimer if it's made of two subunits, a trimer if it's made of three subunits, a tetramer if it's made of four subunits, etc. The prefix "homo" is added if subunits are identical, e.g. "homotrimer", we use "hetero" if subunits are different as in "hetertrimer".

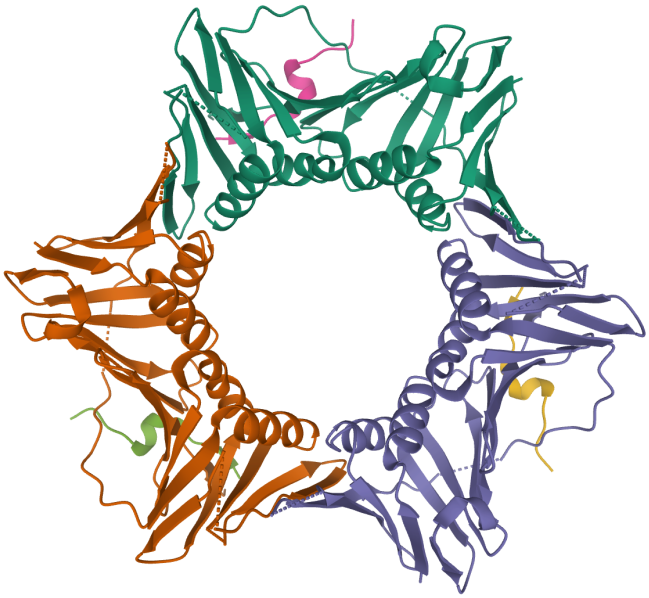


Figure 2.8: Quaternary structure

## 2.6 SEQUENCE ANALYSIS

Since the first DNA decoding in 1977, tens of thousands of DNA have been sequenced and huge amount of data are generated. This sequence data is analyzed in order to separate encoding genes for proteins and RNA, 3D structure, and repetitive sequences. Because of the huge amount of data, it's nearly impossible to manually analyze all DNA sequences.

In sequence analysis there exist several topics:

1. Sequence similarity analysis (a.k.a. homology),
2. Internal features search like active sites, gene-structures and distributions of introns and exons,

3. Evolution analysis,

4. Finding molecular structure replying on sequence alone.

### 2.6.1 *DNA sequencing*

DNA sequencing is the task where the sequence of a DNA is determined. When DNA sequencing is performed, the produced raw data may be affected by weak signals and noise, because of that computational techniques are still needed to process and clean the noisy data.

See Figure 2.9

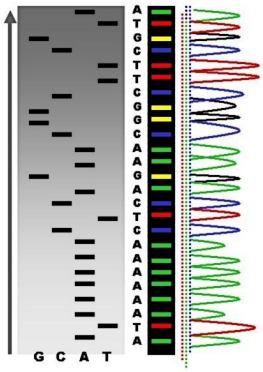



Figure 2.9: Radioactive Fluorescent Sequencing

### 2.6.2 *Sequence assembly*

After DNA sequencing phase, we get small fragments of DNA, see Figure 2.10, here comes the sequence assembly phase where those fragments are aligned and merged to form the full DNA structure. This phase is important because finding the full DNA sequence at once is not possible, and only fragment of around 30000 nucleotides are determined at once.

In DNA assembly, there exits tow techniques that could be used:

1. De novo, this method is used for new genomes that are not sequenced before,
2. Comparative assembly, here the obtained fragments are aligned to the original or similar genome. The alignment process may allow mismatches.

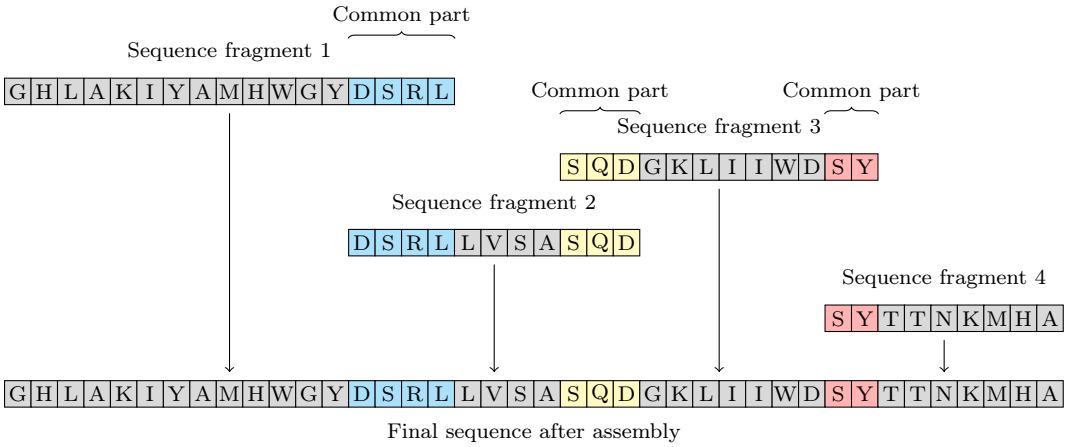



Figure 2.10: Sequence assembly

### 2.6.3 *Annotation*

DNA annotation is the process of assigning functional properties to different regions of the genome sequence Figure 2.11. This task is necessary because the sequence produced by sequencing and assembly doesn't have any functional information. In the last three decades, computational annotation has been used in genome annotation for protein-coding genes on genome (Abril and Castellano, 2019). The first annotation software was designed in 1995 by Dr. Owen White at the Institute for Genomic Research (Fleischmann et al., 1995). The software designed to find protein-coding genes, transfert RNAs and genes functions suggestions.

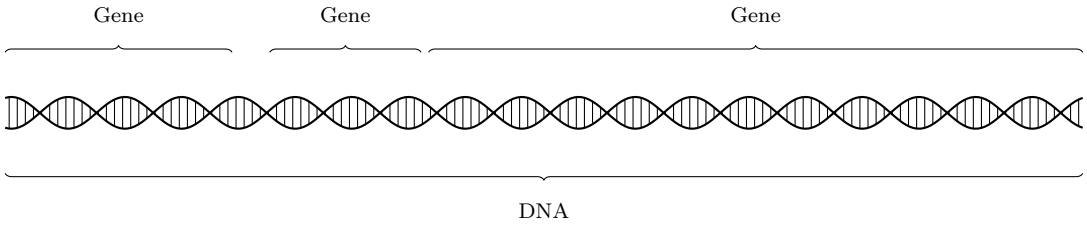



Figure 2.11: DNA Annotation

### 2.6.4 *Comparative genomics*

Comparative genomics is a biological research field where organisms are compared based on their genomic characteristics. Those genomic characteristics include the sequence itself, genes, gene order and other genomic structural features. In that comparison, entire or large part of genomes produced during genome sequencing are compared to each other to learn basic resemblance and differences between living organisms. The most important principle of comparative genomics is that common organisms features will often lead to similar DNA regions.

### 2.6.5 *Analysis of gene and protein expression*

In gene and protein expression analysis the process of encoding a gene product is studied. Gene products are often proteins, but in non-protein-coding genes like transfer RNA genes, the product is a functional RNA. The most common application of gene expression analysis is to compare expression levels of one or more genes from different species.

## 2.7 STRUCTURAL BIOINFORMATICS

Among the applications of bioinformatics there is protein structure prediction. After determining primary structure of a protein it's possible then, in the majority

of cases, to find the structure of this protein in its native environment. Knowing the structure of a protein is important to know and understand its function. Generally, the structure of a protein is either secondary, tertiary or quaternary structure.

## 2.8 NETWORKS AND SYSTEMS ANALYSIS

Networks and systems biology is a field where the aim is to understand relationships and interactions in biological networks and the effects of them on a global scale, involving various molecules simultaneously. An example is Protein-protein interactions.

Protein-protein interaction analysis is a field in bioinformatics where interactions between proteins is studied. The importance of understanding protein-protein interactions is found in the investigation of signaling pathways inside the cell, finding protein structure and understanding different biochemical processes.

## 2.9 ALIGNMENT-FREE SEQUENCE ANALYSIS

In alignment-free sequence analysis, molecular sequences and structures are analyzed without relying on sequence alignment techniques. Alignment-free sequence methods provide an alternative to those based on sequence alignment (Vinga and Almeida, 2003).

Alignment-free methods have significantly lower complexity when compared to methods which are based on multiple sequence alignment (Fan et al., 2015). Alignment-free methods are said to be 140 times faster than alignment based methods (Chan et al., 2014). Several studies showed that with alignment-free methods results of phylogenetic relationship can be accurate even with low sequencing coverage (Ren et al., 2018)

## 2.10 BIOLOGICAL DATABASES

Due to the huge amount of data produced in biology, biochemistry and other clinical data, bioinformatics is always concerned with the necessity of creating and maintaining databases. Bioinformatics databases are fast growing in volume as well as in number, and often are heavily linked to each other, because they often contain information about the same subject or target like, for example, a protein sequence. Also, several internet services provide access, search and processing to bioinformatics databases. In this section we will see many databases with a brief description of the bioinformatics data they contain.

### 2.10.1 *Nucleic acid databases*

#### 2.10.1.1 *DNA databases*

An example of DNA database is GenBank database, created and maintained by National Center for Biotechnology Information in United States of America. It contains annotated nucleic acid and amino acid sequences.

#### 2.10.1.2 *Gene databases*

Gene databases are databases that contain data on expression level. Examples of gene databases are:

- Gene Ontology Consortium http://geneontology.org
- OBO, open biomedical ontologies http://www.obofoundry.org
- ONCOMINE,cancer profiling database https://www.oncomine.org

#### 2.10.1.3 *RNA databases*

RNA databases usually contain information about coding and non coding RNA sequences, structure of RNA molecules and their functions. Some examples are in the following list:

- UniProt, Universal Protein resource http://uniprot.org
- Rfam, RNA families https://rfam.xfam.org

### 2.10.2 *Amino acid databases*

Many databases of amino acids are available, containing information on sequences, functions, structure, families and domains, here we will summarize some of them.

#### 2.10.2.1 *Protein sequence databases*

Due to the links between amino acids and codon sequences, there is a strong link between nucleotide and protein databases. Therefore, amino acid sequences of proteins are also available at some nucleic acid databases like for example GenBank database, we list here UniProt and GenBank

- UniProt, Universal Protein resource http://uniprot.org
- GenBank, Genetic sequence database https://www.ncbi.nlm.nih.gov/genbank

#### 2.10.2.2 *Protein structure databases*

The purpose of protein structure databases is to annotate and organize the structures of different proteins and present them in a way that helps community access information. Data presented by protein structure databases usually include three-dimensional information about proteins as well as other experimental data. Examples of protein structure databases are:

- PDB, Protein Data Bank, it was created in 1971 `http://uniprot.org`
- SWISS-MODEL Repository, contains annotated 3D protein structure models `https://swissmodel.expasy.org/repository`

#### 2.10.2.3 *PPI databases*

PPI (Protein to Protein Interactions) databases contain data about interactions between proteins. Information stored is related to binding, interaction features, energy, annotations, domains and pathways.

### 2.10.3 *Data formats*

In bioinformatics many file formats can be used to store DNA and protein sequence or structure information. There is no format that is perfect for every use, but several format can be used in different situations. Formats are usually convertible from one to another.

#### 2.10.3.1 *GenBank*

GenBank format is composed of tow sections, an annotation section that starts with a line beginning with "LOCUS", and a sequence section that stats with line beginning with "ORIGIN" and ends with the line "//". See Figure 2.12.

```
LOCUS       JF909299                285 bp    mRNA    linear   PRI 25-JUL-2016
DEFINITION  Homo sapiens insulin (INS) mRNA, partial cds.
ACCESSION   JF909299
VERSION     JF909299.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
ORIGIN
        1 ctggggacct gacccagccg cagcctttgt gaaccaacac ctgtgcggct cacacctggt
       61 ggaagctctc tacctagtgt gcggggaacg aggcttcttc tacacaccca agacccgccg
      121 ggaggcagag gacctgcagg tggggcaggt ggagctgggc gggggccctg gtgcaggcag
      181 cctgcagccc ttggccctgg aggggtccct gcagaagcgt ggcattgtgg aacaatgctg
      241 taccagcatc tgctccctct accagctgga gaactactgc aacta
//
```

Figure 2.12: GenBank format

### 2.10.3.2 *FASTA*

FASTA is a common format used for DNA and protein sequence storing. The first line contains a description for the sequence, it must start with ">" symbol, next lines represents sequence data. One file can hold multiple sequences, with the condition that each sequence starts with one description line. See Figure 2.13.

```
>sp|P00738|HPT_HUMAN Haptoglobin OS=Homo sapiens OX=9606 GN=HP PE=1 SV=1
MSALGAVIALLLWGQLFAVDSGNDVTDIADDGCPKPPEIAHGYVEHSVRYQCKNYYKLRT
EGDGVYTLNDKKQWINKAVGDKLPECEADDGCPKPPEIAHGYVEHSVRYQCKNYYKLRTE
GDGVYTLNNEKQWINKAVGDKLPECEAVCGKPKNPANPVQRILGGHLDAKGSFPWQAKMV
SHHNLTTGATLINEQWLLTTAKNLFLNHSENATAKDIAPTLTLYVGKKQLVEIEKVVLHP
NYSQVDIGLIKLKQKVSVNERVMPICLPSKDYAEVGRVGYVSGWGRNANFKFTDHLKYVM
LPVADQDQCIRHYEGSTVPEKKTPKSPVGVQPILNEHTFCAGMSKYQEDTCYGDAGSAFA
VHDLEEDTWYATGILSFDKSCAVAEYGVYVKVTSIQDWVQKTIAEN
```

Figure 2.13: FASTA format

#### 2.10.3.3 *EMBL*

EMBL file can also hold multiple sequences. Every sequence should stat with an identifier line "ID", followed by a set of descriptive lines. The sequence data start is marked by a line starting with "SQ", the end is marked by line "//". See Figure 2.14.

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```

Figure 2.14: EMBL format

## 2.11 CONCLUSION

In this chapter we have discussed the basic concepts and definitions related to Bioinformatics and Molecular Biology. We started with Cells, DNA, RNA and proteins. Then, we presented some research fields in Bioinformatics, finally we gave an overview on some biological databases and data formats. In the next chapter we will see an overview on tandem repeats basic concepts and some methods for tandem repeats detection.

Chapter

# 3

# TANDEM REPEATS

## 3.1 INTRODUCTION

Tandem repeats represent an important feature in genomic sequences. Because of their functional meanings, many algorithms have been developed during the last two decades. Even with the continues TR algorithms development, tandem repeats detection problem still captures researchers' interest. In this chapter we will present an overview on tandem repeats and their basic concepts.

Tandem repeats, denoted TRs, are repetitive and consecutive sequences found in either genomic or proteomic sequences. Usually TRs are characterized by repeat length and repeat number, and classified into microsatellites, minisatellites and satellites.

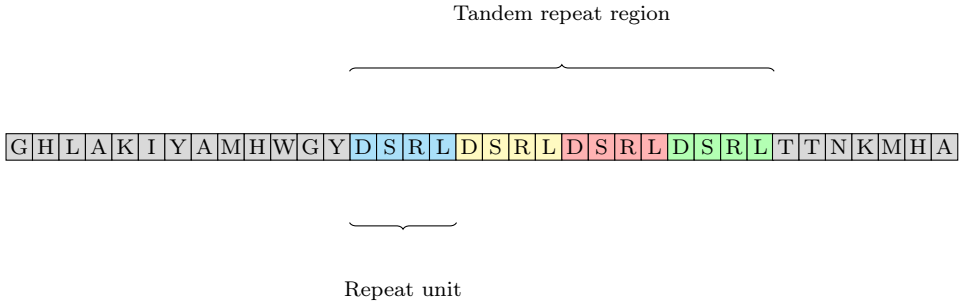



Figure 3.1: Tandem repeats inside a protein sequence

TRs are classified based on their length into many different classes, Microsatellites are tandem repeats where repeats length varies from 1 to 10 base pairs or amino acids, the first discovered microsatellite was discovered by chance in human samples (Wyman and White, 1980). Minisatellites are tandem repeats where repeats length varies from 10 to 100 base pairs or amino acids, they are found to be useful as markers in genetic profiling (Jeffreys, Wilson, and Thein, 1985). Satellites are tandem repeats with longer repeats, where length varies from 100 to 1000. Usually tandem repeats in coding areas in DNA sequences would become also tandem repeats in proteins sequences.

## 3.2 STRUCTURE AND FUNCTION OF TANDEM REPEATS

Tandem repeats are often observed in proteins with basic functions or related to some diseases (Baxa et al., 2006; Hackman, Vihola, and Udd, 2003; Itoh-Satoh et al., 2002; Machado, Sunkel, and Andrew, 1998; Nelson and Eisenberg, 2006; Siwach and Ganesh, 2008). Proteins with tandem repeats are often linked to binding sites and protein to protein interactions.

It is generally considered that regions with low complexity (including and not limited to tandem repeats) are disordered regions. However, recent studies on structure suggest that those regions can be structured (Petrakis et al., 2013; Regad et al., 2017; Schaefer, Wanker, and Andrade-Navarro, 2012; Totzeck, Andrade-Navarro, and Mier, 2017).

Protein structure is not affected directly by repeating motifs. Generally, short repeats are intrinsically considered as disordered, and not taking part of folded domains. Long repeats (more than 30 to 40 amino acids) has higher possibility to be part of a folded protein domains.

There are very popular examples of proteins with tandem repeats like collagen, leucine-rich repeat proteins and zinc-finger proteins.

Proteins containing tandem repeats often have a role of protein-protein interaction modules. An ideal example of this function is the WD40 repeat. (Stirnimann et al., 2010)

## 3.3 SEQUENCE BASED ALGORITHMS

Many approaches have been proposed since the late 1990s. Recently, there is a tendency to making use of sequence data into tandem repeats detection algorithms. The table Table 3.1 shows a list of sequence based algorithms for protein tandem repeats detection.

| Algorithm | Reference | Website |
|---|---|---|
| REP | Andrade et al., 2000 | http://www.bork.embl.de/~andrade/papers/rep/search.html |
| RADAR | Heger and Holm, 2000 | https://github.com/AndreasHeger/radar/ |
| REPRO | Heringa and Argos, 1993 | http://www.ibi.vu.nl/programs/reprowww/ |
| TRUST | Szklarczyk and Heringa, 2004 | http://www.ibi.vu.nl/programs/trustwww/ |
| HHrep | Söding, Remmert, and Biegert, 2006 | http://toolkit.tuebingen.mpg.de/hhrep |
| XSTREAM | Newman and Cooper, 2007 | http://jimcooperlab.mcdb.ucsb.edu/xstream/ |
| HHRepID | Biegert and Söding, 2008 | http://toolkit.tuebingen.mpg.de/hhrepid/ |
| T-REKS | Jorda and Kajava, 2009 | http://bioinfo.montp.cnrs.fr/ |
| REPETITA | Marsella et al., 2009 | http://protein.bio.unipd.it/repetita/ |
| PTRStalker | Pellegrini, Renda, and Vecchio, 2012 | http://bioalgo.iit.cnr.it/ |
| TRDistiller | Richard and Kajava, 2014 | Available upon request |
| TRAL | Schaper et al., 2015 | https://www.vital-it.ch/software/tral |
| MSHDTR | Rudenko and Korotkov, 2021 | http://victoria.biengi.ac.ru/aarep/ |

Table 3.1: Sequence-based algorithms

Algorithms shown in Table 3.1 used different methods and techniques, we summarise them in Table 3.2

| Algorithm | Approach |
|---|---|
| REP | Statistical method based on homology |
| RADAR | Detecting sub-optimal alignments in the self-alignment matrix |
| REPRO | Detecting sub-optimal alignments in the self-alignment matrix |
| TRUST | Detecting sub-optimal alignments in the self-alignment matrix |
| HHrep | HMM comparison |
| XSTREAM | Seed expansion approach |
| HHRepID | HMM comparison |
| T-REKS | Clustering approach based on K-means |
| REPETITA | Amino acids biochemical properties |
| PTRStalker | Heuristic method based on using a normalized BLOSUM-weighted edit distance |
| TRDistiller | Statistical method based on tandem repeats regions filter |
| TRAL | Circular profile hidden Markov model |
| MSHDTR | Statistical approach |

Table 3.2: Approach used in each algorithm

In the next sections, we will see some details on two well known tandem repeats search algorithms (Jorda and Kajava, 2009; Newman and Cooper, 2007). We have chosen them because of their performance, degree of similarity to our method (by modifying algorithm parameters) as well as their availability on the web. Comparing those two algorithms with our method will be detailed in the following chapter.

### 3.3.1 *Early algorithms*

Some of the early tandem repeat detection algorithms are (Andrade et al., 2000; Marcotte et al., 1999; Pellegrini, Marcotte, and Yeates, 1999), they were contributory to the first protein tandem repeats algorithms. recent studies had tendencies to providing web servers tools or producing databases. REP in (Andrade et al.,

2000) is one of the first tandem repeats algorithms, which is based on an homology method to detect tandem repeats.

Other early tandem repeat detection algorithms was based on sub-optimal alignments. Some of those methods are Internal Repeat Finder (Marcotte et al., 1999; Pellegrini, Marcotte, and Yeates, 1999), prospero (Mott, 1999), RADAR (Heger and Holm, 2000), REPRO (George and Heringa, 2000; Heringa and Argos, 1993) and TRUST (Szklarczyk and Heringa, 2004). Usually those algorithms find tandem and interspersed repeats.

### 3.3.2 *XSTREAM*

XSTREAM is an algorithm developed by (Newman and Cooper, 2007) for tandem repeats finding, it uses a seed expansion approach to identify perfect and degenerated tandem repeats in protein sequence.

The main tasks performed by XSTREAM Algorithm, like shown in Figure 3.2, has been divided by (Newman and Cooper, 2007) into five main tasks: Pre-Processing, TR Detection, TR Characterization, Post-Processing, and Output. We will summarize them here briefly.

#### 3.3.2.1 *Pre-processing*

XSTREAM accepts input as FASTA format. Only valid FASTA input is sent to seed finding module where XSTREAM finds short exact substring repeats (seeds) of two or three sizes, depending on the input length. Seed pairs are then considered as starting points to detect tandem repeats. This technique allows XTREAM to quickly find tandem repeats candidates. adjacent pairs of seeds make an indication of tandem repeats period candidate.

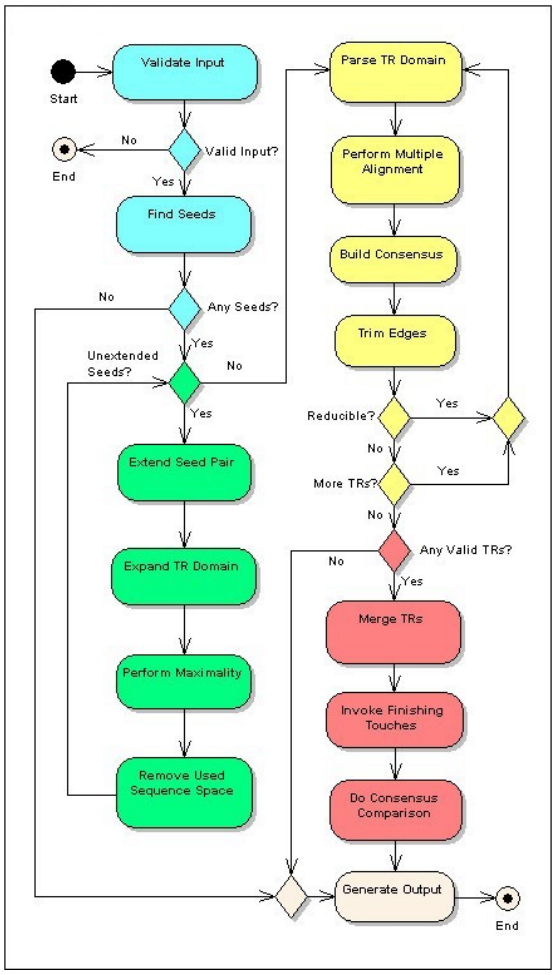



Figure 3.2: XSTREAM Algorithm Flow Chart

#### 3.3.2.2 *Tandem repeats detection*

After seed detection phase, XSTREAM tries to extend each pair of seeds in an iterative manner until certain conditions are met. More details on this step are found in (Newman and Cooper, 2007).

#### 3.3.2.3 *Tandem repeats characterization*

In this step, XSTREAM will refine tandem repeats candidates using optimal or heuristic methods.

#### 3.3.2.4 *Post-processing*

This final step merges similar tandem repeats that are overlapping or are close enough with a certain distance.

#### 3.3.2.5 *Output*

XSTREAM outputs results in three HTML files, the first contains a summary and some statistics, the second contains tandem repeats with their alignments, the third file reports a list of all input sequences containing reported tandem repeats, an optional output is graphical representation of tandem repeats over the sequence.

### 3.3.3 *T-REKS*

T-REKS is an algorithm developed by (Jorda and Kajava, 2009), it uses a clustering approach. The clustering is made on lengths between identical short strings using k-means clustering algorithm.

## 3.4 STRUCTURE BASED ALGORITHMS

Regarding functional features of a protein, structural information is generally more meaningful than its primary structure, because of that, structural information could be used in detecting protein repetitive motifs. However, this technique is only applicable to proteins with determinable 3d shape (Goodsell and Olson, 2000).

This spatial structural information as well as rotations and translations make the algorithmic even more challenging when matching substructures with each

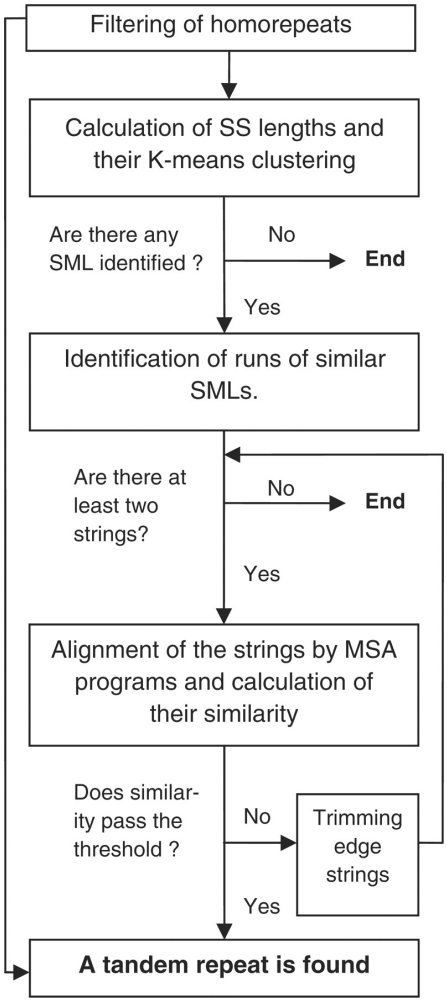



Figure 3.3: T-reqs Algorithm Flow Chart

other in order to determine protein tandem repeats. Table 3.3 shows a set of structure based algorithms.

| Algorithm | Reference | Website |
|---|---|---|
| DAVROS | Murray, Taylor, and Thornton, 2004 | http://www.ebi.ac.uk/~murray/davros/ |
| OPAAS | Shih and Hwang, 2004 | http://www.ibms.sinica.edu.tw/ |
| Swelfe | Abraham, Rocha, and Pothier, 2008 | http://wwwabi.snv.jussieu.fr/public/Swelfe/ |
| RQA | Chen, Huang, and Xiao, 2009 | |
| Gplus | Guerler, Wang, and Knapp, 2009 | http://agknapp.chemie.fu-berlin.de/gplus/ |
| SymD | Kim, Basner, and Lee, 2010 | http://symd.nci.nih.gov/ |
| ProSTRIP | Sabarinathan, Basu, and Sekar, 2010 | http://cluster.physics.iisc.ernet.in/prostrip/ |
| RAPHAEL | Walsh et al., 2012 | http://protein.bio.unipd.it/raphael/ |
| Frustratometer | Parra et al., 2013 | http://www.proteinphysiologylab.tk/ |
| ConSole | Hrabe and Godzik, 2014 | http://console.sanfordburnham.org/ |
| PRIGSA | Chakrabarty and Parekh, 2014 | http://bioinf.iiit.ac.in/PRIGSA/ |
| TAPO | Do Viet, Roche, and Kajava, 2015 | https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=2 |
| RepeatsDB-lite | Hirsh et al., 2018 | http://old.protein.bio.unipd.it/repeatsdb-lite/ |

Table 3.3: Structure-based algorithms

Algorithms shown in Table 3.3 used different approaches, we summarise them in Table 3.4

| Algorithm | Approach |
|---|---|
| DAVROS | Fourier Transform approach based on structural alignment |
| OPAAS | Symmetry detection approach |
| Swelfe | Statistical based on dynamic programming |
| RQA | Symmetry detection approach |
| Gplus | Symmetry detection approach |
| SymD | Symmetry detection approach |
| ProSTRIP | Statistical based on dynamic programming |
| RAPHAEL | Extracting periodicity and distance from 3d structure |
| Frustratometer | TopMatch based approach |
| ConSole | Contact matrix based approach |
| PRIGSA | Graph based approach |
| TAPO | SVM approach |
| RepeatsDB-lite | TR library structural search |

Table 3.4: Approach used in each algorithm

## 3.5 PROTEIN TANDEM REPEATS DATABASES

After the improvement of tandem repeats search algorithms and the massive amount of sequenced proteins a major problem is naturally faced, which is understanding this amount of data. To accomplish this, we need a large analyses operations which can shed more light on sequence motifs, evolution, functions and structures of tandem repeats. Popular databases like UniProt, InterPro or UniRef are helpful for this purpose (Hunter et al., 2009; Schultz et al., 1998). Also those databases are helpful when searching for related structure and function information (Kajava, 2012).

However, protein tandem repeats analyses also requires special software and databases. Over the last years, Several databases focusing on tandem repeats has been developed. For instance, TRIPS database collects tandem repeats from Uniprot's database "SwissProt" that show amino acids substitutions but not insertions or deletions (Katti et al., 2000). Another database which is "Homopeptide Repeat Database" that exists on (http://repeats.med.monash.edu.au/php/index.php) has repeats that are selected using "Regex" applied on GENPEPT database (Faux et al., 2005). ProtRepeatsDB database (Kalita et al., 2006) and PRDB database (Jorda, Baudrand, and Kajava, 2012) are extended to cover longer protein tandem repeats. These databases are available along with tools which are useful in further analyses of protein tandem repeats like structure, function, distribution over the full proteome, repeats size, number of repeats and composing residues.

| Name | Reference | Website |
|---|---|---|
| RepSeq | Depledge, Lower, and Smith, 2007 | http://www.repseq.org |
| PRDB | Jorda, Baudrand, and Kajava, 2012 | http://bioinfo.montp.cnrs.fr/?r=repeatDB |
| ProRepeat | Luo et al., 2012 | http://prorepeat.bioinformatics.nl |
| PTRStalkerDB | Pellegrini, Renda, and Vecchio, 2012 | http://bioalgo.iit.cnr.it |
| RepeatsDB | Di Domenico et al., 2013 | http://repeatsdb.bio.unipd.it |

Table 3.5: Common protein tandem repeats databases

Many protein repeats databases where made available during the last two decades. Unfortunately, not all of these online databases are still available and functional like RepSeq (Depledge, Lower, and Smith, 2007) and ProtRepeatDB (Kalita et al., 2006).

## 3.6 CONCLUSION

The present chapter on tandem repeats covers several aspects of tandem repeats research field, including functions, related TR detection algorithms, and finally databases. In the following chapter we will present our method for tandem repeats detection and analysis.

Chapter

# 4

# PROPOSED APPROACH

## 4.1 INTRODUCTION

Protein sequences often contain repeats arranged in tandem patterns. Despite the existence of many algorithms and tools that are designed to find protein tandem repeats, useful tools for finding tandem repeats are still needed for speed and effectiveness reasons. In this chapter we will describe our proposed approach.

## 4.2 BACKGROUND

Studies on tandem repeats have been always focusing on DNA. DNA tandem repeats are classified into microsatellites, minisatellites, and large-scale repeats.

Repeats in nucleotides found in coding genes can cause protein repeats in the resulting protein. Amino acid repeats are known to be found in proteins (Gatherer and McEwan, 2005; Marcotte et al., 1999).

Various algorithms have been developed for repeats detection in DNA and protein sequences. One class of algorithms is alignment-free class. Alignment-free methods has significantly lower complexity when compared to methods which are based on multiple sequence alignment (Fan et al., 2015).

To improve and complement current algorithms for protein repeats detection, we implemented an algorithm to find tandem repeats in proteins without using sequence-self alignment (with substitution and without indels). Our new algorithm, called RES for REpeatability Scanner, is designed to efficiently search for tandem repeats.

### 4.2.1 *Mathematical representation*

Strings can represent many biological objects. In biology, a protein is a sequence of amino acids, every amino acid is represented by a letter. Consequently, a protein can be represented by a sequence of letters that corresponds to the 20 amino acids.

**Definition 1.** Protein alphabet is composed of twenty amino acids, these amino acids are presented by an alphabet made of letters. Therefore the alphabet set denoted by $\sum$ is defined as:

$$\sum = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

And thus a protein can be represented by a finite string over the alphabet $\sum$, which is often called the primary structure of the protein.

**Definition 2.** Let k be an integer with $k \geq 1$, we call protein k-word a string of length k over the alphabet $\sum$. Specifically, let $\sum^k$ be the set of all possible k-words formed using the alphabet $\sum$, the size of $\sum^k$ being $20^k$. Moreover, for k = 1 each 1-words denotes one of the twenty amino acids.

$$\sum^k = \{a_1 \cdots a_k \mid a_i \in \sum \wedge 1 \leq \ i \leq k\}$$

**Definition 3.** We denote $\sum^*$ all finite strings (regardless of their length), that is the union of all $\sum^k$ where k $\geq 0$.

$$\sum^* = \bigcup_{i \in \mathbb{N}} \sum^i$$

### 4.2.2 *Repeatability*

**Definition 4.** Let $s \in \sum^*$ a sequence of amino acids, $w$ a sub-sequence of $s$, we denote $Rep(w)$ the repeatability of $w$ and is defined as the Hamming distance of $w$ from begin a perfect repeat.

## 4.3 IMPLEMENTATION

Our method, RES Algorithm, is implemented using C programming language due to its high performance. Different aspects of the algorithm is explained in this section. All calculations were performed in Macbook Pro laptop with a i5 CPU (520M) and a 8GB of RAM.

### 4.3.1 *Data/Amino acids representation*

We have implemented RES Algorithm based on Table 2.1 representation. Each amino acid could be coded with 5 bits, but because of the typical size used in data structure which is 8 bit, we are going to represent amino acids by 8 bit character to avoid unnecessary calculations.

### 4.3.2 *Indels and depth of analysis*

Our method's main focus is to help researchers in their studies for individual repeats search, because of that we are not taking indels into consideration for the sake of speed and performance. So in the implementation section we will consider analyzing only repeats with mutations, we will see more details on performance later in this chapter.

### 4.3.3 *Algorithm*

We devised an algorithm that finds the minimum number of mutations required for a sequence to be a perfect repeat and provides that perfect repeat. The algorithm for the first window is defined as follows:

**Algorithm 1: Window processing with Matrix initialization**

**Input:** W: sequence window, RL: repeat length

**Output:** R: a repeat, MM: The minimum number of mutations

**begin**

```
let L be the length of W                                          /* O(1) */
let AC be the count of all amino acids                            /* O(1) */
create matrix M with RL columns and AC rows                       /* O(1) */
initialize M with zeros                                           /* O(1) */
foreach amino acid A in W                                         /* O(n) */
 do
    let Pos be the position of A in W                             /* O(1) */
    let C equals Pos modulo RL                                    /* O(1) */
    let RN be row number in matrix M that corresponds to A        /* O(1) */
    add 1 to the position with row RN and column C                /* O(1) */
end
let T be a table with the maximum of each column of M             /* O(n) */
let R be a pattern of amino acids corresponding to rows of the maximum
 of each column of M                                              /* O(n) */
let MM be the difference between L and the sum of elements of T
 /* O(n) */
output (R)                                                        /* O(1) */
output (MM)                                                       /* O(1) */
```

**end**

After executing the algorithm Algorithm 1 for the first time, the matrix M is calculated and saved for later use, after we slide the window by one amino acid, the calculation of matrix M is no longer necessary, we only need to do two modifications:

- Subtract 1 from the corresponding cell to the *falling out* amino acid
- Add 1 to the corresponding cell to the *falling in* amino acid

And then consider the second column of matrix M as the starting column. So the optimized RES Algorithm becomes as follows:

**Algorithm 2: Window processing without Matrix initialization**

**Input:** W: sequence window, RL: repeat length

**Output:** R: a repeat, MM: minimum number of mutations

**begin**

| | |
|---|---|
| use M, T, R, MM from previous window | /* $\mathcal{O}(1)$ */ |
| let $AA_{prev}$ the first amino acid of the previous window | /* $\mathcal{O}(1)$ */ |
| let $AA_{last}$ the last amino acid of W | /* $\mathcal{O}(1)$ */ |
| subtract 1 from M at postion that correspods to $AA_{prev}$ | /* $\mathcal{O}(1)$ */ |
| check for a new column maximum | /* $\mathcal{O}(1)$ */ |
| add 1 to M at position that correspods to $AA_{last}$ | /* $\mathcal{O}(1)$ */ |
| check for a new column maximum | /* $\mathcal{O}(1)$ */ |
| consider the second column of M as the starting column | /* $\mathcal{O}(1)$ */ |
| according to new maximums modify R | /* $\mathcal{O}(1)$ */ |
| consider the second aa of R as the starting aa | /* $\mathcal{O}(1)$ */ |
| according to new maximums modify MM | /* $\mathcal{O}(1)$ */ |
| output (R) | /* $\mathcal{O}(1)$ */ |
| output (MM) | /* $\mathcal{O}(1)$ */ |

**end**

The algorithm Algorithm 2 is then used for the rest of windows. We can write the final RES Algorithm as follows:

**Algorithm 3: RES Algorithm**

**Input:** S: protein sequence, RL: repeat length

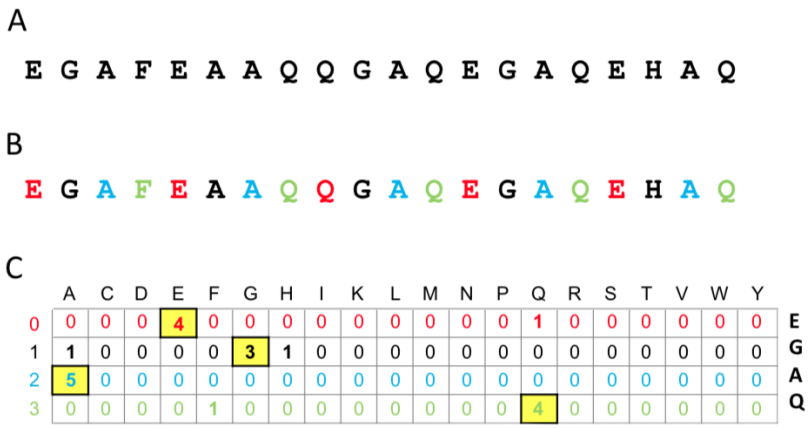
**Output:** list of (R: repeat, MM: minimum number of mutations)

**begin**

```
let W_0 be the first window ;                          /* O(1) */
call Algorithm 1 for W_0 and RL ;                      /* O(n) */
foreach window W in the rest of windows of S ;         /* O(n) */
 do
    call Algorithm 2 for W and RL ;                    /* O(1) */
end
```

**end**

The repeatability of the sequence is 1 for a perfect repeat, or, if the sequence needs MM mutations to be converted to a perfect repeat, the fraction of residues that do not need to be changed. MM is actually the Hamming distance from the sequence to the perfect repeat.




A : Sequence of length 20

B : Colour frame indicating the amino acids that will be counted to detect tandem repeats of length 4

C : Count matrix for amino acids at each position

Figure 4.1: Illustrative example of the repeatability RES Algorithm at work

Given the example shown in Figure 4.1, $E$ is present 4 times at position 0, $G$ 3 times at position 1, $A$ 5 times at position 2 and $Q$ 4 times at position 3. This gives the perfect nearest repeat (EGAQ) and the number of mutations needed to convert the sequence above to the perfect repeat, 20 - (4 + 3 + 5 + 4) = 4.

The execution of the algorithm on the sequence of length 20 shown in (Figure 4.1-A) with the parameter RL = 4 (Figure 4.1-B) generates the counts matrix M (shown, transposed, in (Figure 4.1-C). The maximum count of each row is an array T of values (4, 3, 5, 4), that corresponds to the pattern EGAQ. The minimum number of mutations required is calculated by MM = L - Sum(T), where MM = minimum number of mutations, L = Length of the window and T is an array representing the maximum of each row of matrix M. In this case, the value is 20–16 = 4. Thus, the result of the algorithm is that this sequence of length 20 can be converted to the perfect repeat EGAQ x 5 with 4 mutations.

The algorithm has linear complexity, which makes it very fast even for large sequences. Speed and running time for different types of input is detailed in the next sections.

### 4.3.4 *Algorithm flow chart*

The main functionalities of RES Algorithm, see Figure 4.2, is divided into three modules: Pre-Processing, Tandem Repeats Detection, and Output.

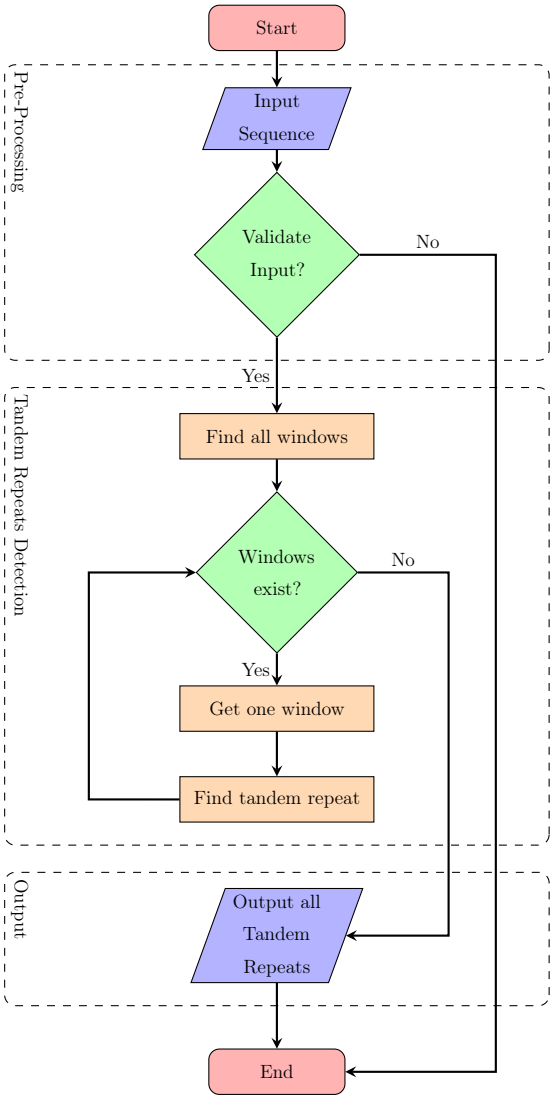



Figure 4.2: RES Algorithm Flowchart

RES Algorithm requires input to be in FASTA format. RES sends valid sequences to the next module "Tandem Repeats Detection".

Following validating input, the sequence is then parsed to get all windows, for each window the RES Algorithms is applied to find the tandem repeat for the required length. In this step, the generated matrix M, see Algorithm 3, will be kept to the next detection cycle so the processing time will be the lowest possible.

In the last module, RES automatically generates CSV files containing tandem repeats found during analysis. CSV output contains several columns holding information about repeats.

### 4.3.5 *Complexity*

Algorithmic complexity is a measure used to know how long an algorithm would require to process a data input of size n. When increasing data that the algorithm has to process, the processing time should be finite and practical even for big values of n. Complexity is usually analysed in terms of time, but sometimes it is analysed in terms of space, which represents the memory allocated by the algorithm.

To determine time complexity of RES Algorithm, we analysed the program's statements of the algorithms Algorithm 1 and Algorithm 2.

For Algorithm 1, the algorithm consists of initializing the matrix M, M has a fixed number of columns and a number of rows equals to the desired repeat length, this makes analysing M has a time complexity of $\mathcal{O}(n)$, detailed time complexity is show in Algorithm 1. Algorithm 1 is applied only to the first window.

Starting from the second sliding window, Algorithm 2 is applied, the algorithm modifies M and other variables accordingly without parsing the whole structures, time complexity drops then to $\mathcal{O}(1)$ as shown in Algorithm 2.

The resulting main algorithm, namely RES Algorithm, have a time complexity $\mathcal{O}(n)$ as detailed in Algorithm 3.

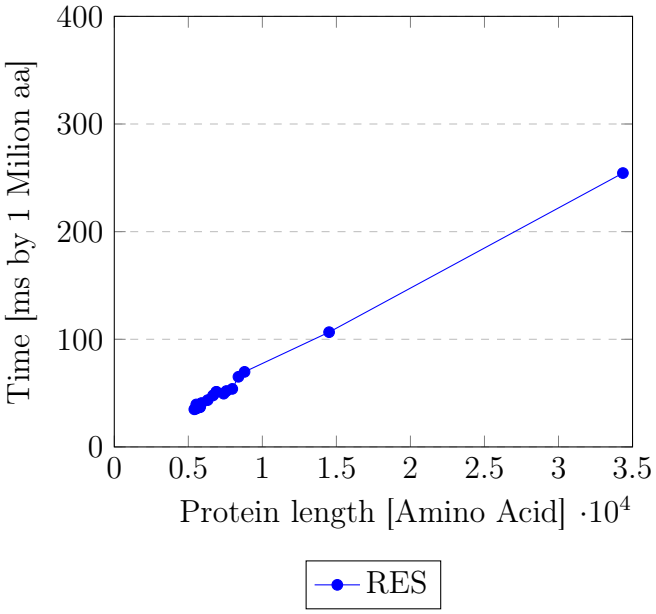



Figure 4.3: RES processing time by protein length

We implemented RES Algorithm so that it will have a $\mathcal{O}(n)$ time complexity for short tandem repeats detection.

In Figure 4.3 we see that our method (RES Algorithm) has indeed a linear complexity when increasing protein length.

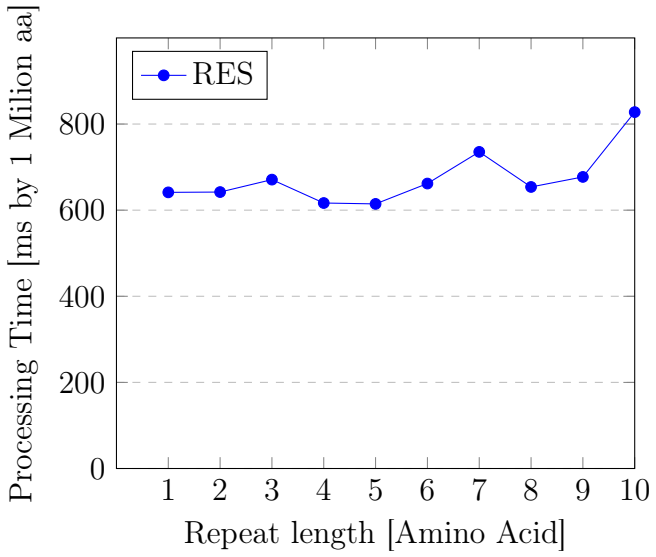



Figure 4.4: RES processing time by repeat length

A benefit of the linear complexity of RES Algorithm is that the algorithm has a static processing time when increasing repeat length to be found in the sequence, see Figure 4.4 as an example.

### 4.3.6 *Output*

RES automatically generates CSV files containing tandem repeats found during analysis. CSV output contains several columns holding information as described below:

1. **Protein order**, This is the first column, it contains an internal protein number withing the set of input sequences,
2. **Protein identifier**, Here we put UniProt identifier of the protein under analysis,
3. **Repeat length**, This column represents the size of suggested repeat,
4. **Minimum mutations**, The minimum mutations required for the current window to become a perfect repeat,
5. **Windows length**, This number represents the size of current window,
6. **Window start**, Windows start is the zero based position of the first amino acid of the current window within the input sequence,
7. **Window**, The current window under analysis is put in this column,

## 4.4 EVALUATION AND VALIDATION

To evaluate RES Algorithm, we have used a set of different input datasets. Processing time analysis is shown in Table 4.2 for every tandem repeat length.

To compare our method with competitors, we studied sequence based methods, and then selected two algorithms, T-REKS and XSTREAM. We have selected those two because of their performance, degree of similarity of the tool by modifying some parameters, also for their availability on the web, as not all the tools are still available for use and download. The algorithms HHrep, ARD2, REPETITA, PTRStalker, TRDistiller have been checked, they are either not available or doesn't provide performance statistics.

After more investigation in those tow tools, we found that T-REKS shows processing times in seconds only, with no option to show milliseconds, with the minimum time is being one second, so we are going to compare only with XSTREAM algorithm.

### 4.4.1 *Data*

As we are comparing speed and processing time, its wise to use the longest proteins possible, the following table, Table 4.1, shows the longest 20 human proteins obtained from the reviewed database SwissProt, proteins are ordered from longest to shortest.

| Protein names | Entry code | Entry name | Length |
|---|---|---|---|
| Titin | Q8WZ42 | TITIN_HUMAN | 34350 |
| Mucin-16 | Q8WXI7 | MUC16_HUMAN | 14507 |
| Nesprin-1 | Q8NF91 | SYNE1_HUMAN | 8797 |
| Mucin-19 | Q7Z5P9 | MUC19_HUMAN | 8384 |
| Obscurin | Q5VST9 | OBSCN_HUMAN | 7968 |
| Dystonin | Q03001 | DYST_HUMAN | 7570 |
| Microtubule-actin cross-linking factor 1 | Q9UPN3 | MACF1_HUMAN | 7388 |
| Fibrous sheath-interacting protein 2 | Q5CZC0 | FSIP2_HUMAN | 6907 |
| Nesprin-2 | Q8WXH0 | SYNE2_HUMAN | 6885 |
| Nebulin | P20929 | NEBU_HUMAN | 6669 |
| Adhesion G-protein coupled receptor V1 | Q8WXG9 | AGRV1_HUMAN | 6306 |
| Neuroblast differentiation-associated protein AHNAK | Q09666 | AHNK_HUMAN | 5890 |
| Protein AHNAK2 | Q8IVF2 | AHNK2_HUMAN | 5795 |
| Mucin-5B | Q9HC84 | MUC5B_HUMAN | 5762 |
| Mucin-5AC | P98088 | MUC5A_HUMAN | 5654 |
| Hemicentin-1 | Q96RW7 | HMCN1_HUMAN | 5635 |
| Midasin | Q9NU22 | MDN1_HUMAN | 5596 |
| Histone-lysine N-methyltransferase 2D | O14686 | KMT2D_HUMAN | 5537 |
| Mucin-12 | Q9UKN1 | MUC12_HUMAN | 5478 |
| IgGFc-binding protein | Q9Y6R7 | FCGBP_HUMAN | 5405 |

Table 4.1: Longest Twenty Human proteins

Using proteins presented in Table 4.1, we computed RES running time for each protein for every repeat length, starting from homo repeats and going up to repeats with length 10. Results are shown in Table 4.2.

| | Repeat length | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Protein | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Q8WZ42 | 17.87 | 21.79 | 22.00 | 22.77 | 23.22 | 21.31 | 26.62 | 18.84 | 23.50 | 27.15 |
| Q8WXI7 | 8.75 | 9.33 | 9.28 | 9.39 | 7.40 | 9.12 | 10.24 | 9.67 | 8.69 | 11.83 |
| Q8NF91 | 4.76 | 5.73 | 6.70 | 5.96 | 5.12 | 5.80 | 6.64 | 5.14 | 5.18 | 7.34 |
| Q7Z5P9 | 5.63 | 5.36 | 5.53 | 5.82 | 5.85 | 6.15 | 6.30 | 6.72 | 6.79 | 7.08 |
| Q5VST9 | 6.00 | 5.37 | 5.41 | 4.36 | 5.89 | 6.25 | 5.96 | 5.15 | 4.65 | 6.63 |
| Q03001 | 4.13 | 5.00 | 5.02 | 5.25 | 5.21 | 6.16 | 5.52 | 6.18 | 5.56 | 6.46 |
| Q9UPN3 | 5.26 | 4.94 | 5.02 | 3.69 | 3.91 | 4.79 | 5.42 | 5.58 | 4.27 | 6.10 |
| Q5CZC0 | 4.08 | 4.62 | 4.60 | 4.62 | 5.02 | 4.23 | 5.42 | 3.40 | 5.65 | 5.67 |
| Q8WXH0 | 3.79 | 4.62 | 4.67 | 5.01 | 3.45 | 5.23 | 4.78 | 5.67 | 5.65 | 5.44 |
| P20929 | 3.94 | 3.90 | 4.40 | 3.79 | 4.68 | 3.91 | 4.35 | 5.40 | 5.47 | 5.69 |
| Q8WXG9 | 4.92 | 4.33 | 4.13 | 3.04 | 3.49 | 4.24 | 3.96 | 5.04 | 3.81 | 6.03 |
| Q09666 | 2.98 | 3.62 | 3.97 | 2.68 | 2.99 | 3.42 | 3.29 | 4.13 | 3.84 | 5.02 |
| Q8IVF2 | 4.04 | 3.85 | 3.79 | 2.99 | 3.22 | 3.14 | 4.57 | 2.91 | 4.93 | 5.03 |
| Q9HC84 | 5.05 | 4.05 | 4.10 | 2.69 | 3.22 | 3.43 | 4.34 | 2.91 | 4.40 | 5.09 |
| P98088 | 5.00 | 3.85 | 4.15 | 3.18 | 2.87 | 4.31 | 4.35 | 4.06 | 3.21 | 4.43 |
| Q96RW7 | 3.66 | 3.62 | 3.85 | 3.99 | 2.55 | 3.95 | 4.13 | 2.80 | 3.13 | 3.97 |
| Q9NU22 | 4.78 | 3.84 | 3.66 | 2.96 | 2.66 | 4.07 | 4.00 | 4.43 | 3.57 | 4.67 |
| O14686 | 5.07 | 3.76 | 4.01 | 2.58 | 4.10 | 4.00 | 4.42 | 4.17 | 2.94 | 4.85 |
| Q9UKN1 | 3.76 | 2.58 | 3.68 | 4.06 | 4.41 | 3.53 | 4.06 | 3.40 | 4.56 | 4.56 |
| Q9Y6R7 | 3.30 | 2.73 | 3.73 | 3.82 | 3.05 | 3.14 | 4.05 | 3.26 | 2.94 | 4.76 |

Table 4.2: RES performance for 20 longest proteins (in ms)

Same tandem repeats searches are done using XSTREAM algorithm, results are shown in Table 4.3. searches are executed individually for each repeat length to show differences against RES algorithm.

| | Repeat length | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Protein | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Q8WZ42 | 197.0 | 219.0 | 349.0 | 234.0 | 208.0 | 244.0 | 270.0 | 230.0 | 194.0 | 202.0 |
| Q8WXI7 | 162.0 | 143.0 | 182.0 | 146.0 | 144.0 | 136.0 | 151.0 | 123.0 | 127.0 | 134.0 |
| Q8NF91 | 116.0 | 138.0 | 126.0 | 132.0 | 92.0 | 105.0 | 114.0 | 116.0 | 97.0 | 99.0 |
| Q7Z5P9 | 140.0 | 107.0 | 143.0 | 119.0 | 105.0 | 93.0 | 101.0 | 97.0 | 101.0 | 182.0 |
| Q5VST9 | 146.0 | 108.0 | 98.0 | 185.0 | 97.0 | 105.0 | 108.0 | 88.0 | 138.0 | 108.0 |
| Q03001 | 121.0 | 153.0 | 110.0 | 97.0 | 81.0 | 90.0 | 85.0 | 103.0 | 169.0 | 100.0 |
| Q9UPN3 | 100.0 | 100.0 | 158.0 | 122.0 | 106.0 | 93.0 | 86.0 | 84.0 | 89.0 | 105.0 |
| Q5CZC0 | 103.0 | 92.0 | 104.0 | 108.0 | 102.0 | 105.0 | 81.0 | 89.0 | 85.0 | 115.0 |
| Q8WXH0 | 128.0 | 88.0 | 97.0 | 128.0 | 101.0 | 80.0 | 92.0 | 86.0 | 86.0 | 86.0 |
| P20929 | 97.0 | 114.0 | 115.0 | 86.0 | 98.0 | 86.0 | 89.0 | 89.0 | 101.0 | 83.0 |
| Q8WXG9 | 109.0 | 118.0 | 130.0 | 89.0 | 95.0 | 94.0 | 104.0 | 91.0 | 81.0 | 102.0 |
| Q09666 | 99.0 | 112.0 | 111.0 | 92.0 | 102.0 | 90.0 | 104.0 | 90.0 | 108.0 | 88.0 |
| Q8IVF2 | 107.0 | 95.0 | 81.0 | 84.0 | 149.0 | 76.0 | 98.0 | 86.0 | 143.0 | 104.0 |
| Q9HC84 | 109.0 | 128.0 | 92.0 | 124.0 | 127.0 | 86.0 | 88.0 | 85.0 | 119.0 | 87.0 |
| P98088 | 116.0 | 121.0 | 298.0 | 103.0 | 164.0 | 88.0 | 104.0 | 237.0 | 102.0 | 132.0 |
| Q96RW7 | 123.0 | 106.0 | 103.0 | 88.0 | 90.0 | 80.0 | 93.0 | 91.0 | 80.0 | 86.0 |
| Q9NU22 | 111.0 | 92.0 | 108.0 | 84.0 | 95.0 | 93.0 | 98.0 | 90.0 | 92.0 | 81.0 |
| O14686 | 120.0 | 150.0 | 120.0 | 121.0 | 109.0 | 98.0 | 121.0 | 121.0 | 137.0 | 80.0 |
| Q9UKN1 | 107.0 | 93.0 | 107.0 | 79.0 | 97.0 | 91.0 | 101.0 | 99.0 | 115.0 | 99.0 |
| Q9Y6R7 | 89.0 | 91.0 | 118.0 | 90.0 | 85.0 | 85.0 | 82.0 | 93.0 | 86.0 | 78.0 |

Table 4.3: XSTREAM performance for 20 longest proteins (in ms)

Results clearly shows that RES algorithm outperforms XSTREAM algorithm when targeting individual tandem repeats lengths for perfect and non perfect repeats with no indels inside.

When comparing complexity, we have used the proteins provided in Table 4.1, Figure 4.5 show that RES Algorithm has a grater performance over XSTREAM.

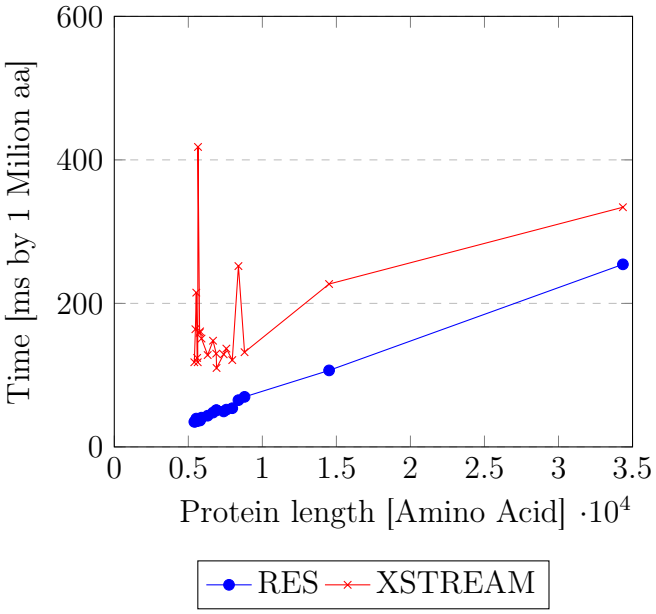



Figure 4.5: Time complexity comparison

Finally, to get an overall idea on RES processing times over individual proteins, we provide Figure 4.6 that clearly shows performance of the algorithm.

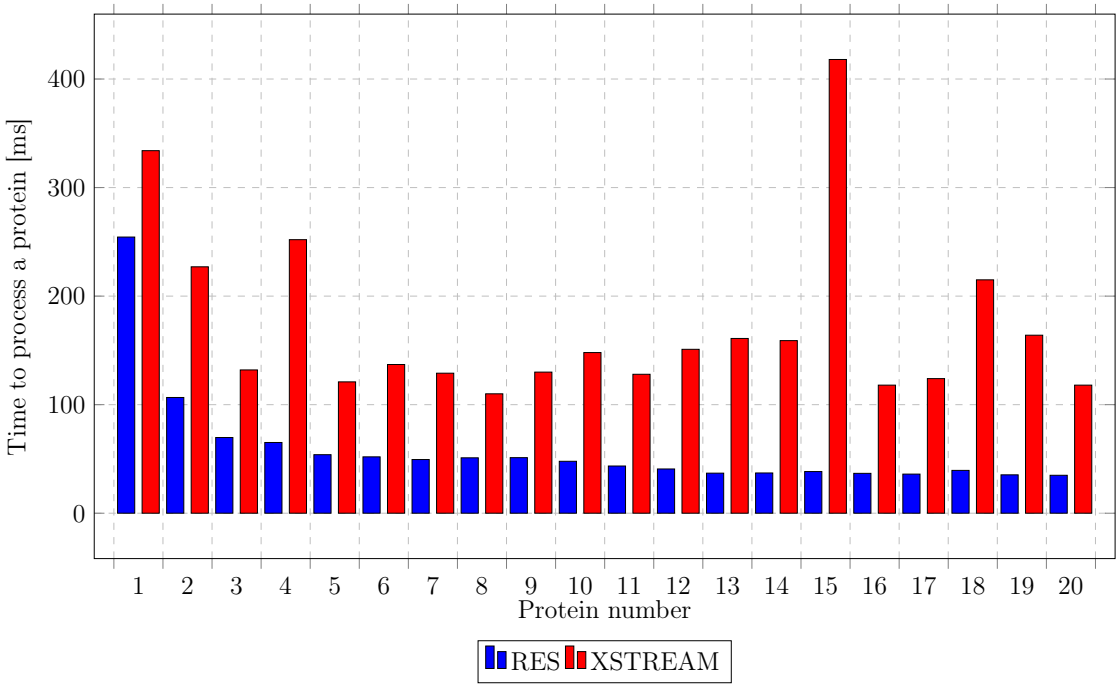



Figure 4.6: Processing Time per Protein Comparison

## 4.5 CONCLUSION

In this chapter, we presented our method, RES Algorithm, we compared our method to one of the well know algorithms in the literature (Newman and Cooper, 2007). RES Algorithm shows strong point compared to XSTREAM in individual repeat length search point of view. In the next chapter we will see more details and analyses about RES algorithm.

Chapter

# 5

# RESULTS AND DISCUSSION

## 5.1 INTRODUCTION

Low complexity regions (LCRs) in protein sequences are abundant (see e.g. (Peng et al., 2015)) but lack the functional, structural, and evolutionary properties of globular domains (Chavali et al., 2017). Their abundance, particularly in eukaryotic complex organisms, and increasing evidence for particular types of compositionally biased LCRs, are bringing interest in them as having possible functions in protein interactions and holding motifs for post-translational modifications with regulatory functions (Harrison, 2006; Toro Acevedo et al., 2017). One particular type of compositionally biased LCRs are those that show periodicity, one extreme case being homorepeats (or polyX; (Jorda and Kajava, 2010; Mier, Alanis-Lobato, and Andrade-Navarro, 2017), (Mier, Alanis-Lobato, and Andrade-Navarro, 2017)).

Other repeats exist of length 2, 3, etc. that have been noted to be frequent in proteins with certain functions (e.g. RG tandem repeats in ribosomal proteins (Suzuki, Olvera, and Wool, 1991)).

There is a generally accepted idea that LCRs are disordered regions. However, some structural analyses suggest that LCRs can acquire structures, which can be flexible and dependent on the context regarding both (i) the sequence holding the LCR (see e.g. (Totzeck, Andrade-Navarro, and Mier, 2017)) and (ii) the interacting molecules (see e.g. in general (Regad et al., 2017)), or specifically for polyQ (Petrakis et al., 2013; Schaefer, Wanker, and Andrade-Navarro, 2012).

The subjacent idea is that the (non-perfect or approximate) repetitions in short tandem repeats (hereafter repeatability) might as well generate flexible, context-dependent, very ordered structures (Mier et al., 2019). These would be very difficult to predict because the folding rules that apply to the formation of secondary structure elements (alpha helix and beta strands) in globular domains do not apply to these short repeats (Vlassi, Brauns, and Andrade-Navarro, 2013).

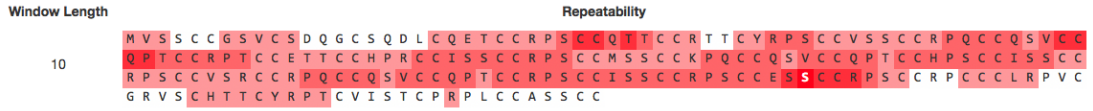
As the general conservation of these repeats is low because their evolution is very fast, we aim here to provide a tool to easily find very short tandem repeats, which, if obscured by large numbers of mutations, can be very difficult to find. Particularly, we wanted to offer the exploratory search for short repeats with mutations, in a fast and interactive fashion for individual sequences. Additionally, we will show that the application of this algorithm to complete proteomes can point to stark differences between organisms that may bring insight into general rules defining the relation between repeatability and function.

## 5.2 WEB TOOL

The algorithm described in Algorithm 3 can be applied to protein sequences through an easy-to-use dedicated web tool developed in PHP (RES = REpeatability Scanner). After analysis, RES generates a coloured protein sequence, where each

amino acid is given a colour shade depending on how many mutations are required for a window starting at that particular amino acid to reach a perfect repeat (see Figure 5.1). The repeat, number of mutations and other details are provided for each window when hovering the mouse cursor on an amino acid. Using the graphical representation of the results from RES, one can focus on LCRs with short approximate tandem repeats by easily spotting regions of interest defined as high repeatability regions.

**A**




**B**

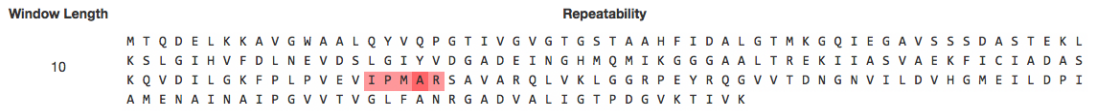



Figure 5.1: Repeatability web tool

A : Result for human Keratin-associated protein 4–7 (UniProt: Q9BYR0). Parameters: window length 10, size of repeats 5, mismatches 3 or less. This protein contains CCxxx tandem repeats and it is accordingly largely covered by hits. The best match is the perfect tandem repeat SCCRP x 2 (black boxes).

B : For comparison, see the result for a globular protein of similar size (Ribose-5-phosphate isomerase A, from Escherichia coli; UniProt: P0A7Z0) run with the same parameters. Only a small region is detected as containing repeats, This is neither a structural repeat nor it is conserved as tandem repeat in orthologs, suggesting a false positive (data not shown).

Note that if the detection of multiple repeat sizes is requested, priority of overlapping repeats is given to the repeat with the highest repeatability, or to the shorter one in case of ties.

## 5.3 TARGETED PROTEIN SEARCH

The linear complexity of this algorithm allows us to examine full proteomes. Thus, the method can be used for a targeted search of proteins with very specific properties in their repeatability. For example, we can find that 106 human proteins have at least one perfect tandem repeat of length 10 amino acids (in 39 proteins, strictly not composed of shorter repeats; Supplementary File S2). These repeats occur often in series containing more than two repeats, hinting at their significance.

The shortest series is found in the Heterogeneous nuclear ribonucleoprotein U-like protein 1 (hnRNPUL1; UniProt: HNRL1_HUMAN), with only one hit to two tandem repeats covering 20 amino acids (starting at position 695). The longest series found is in the Ribosome-binding protein 1 (RRBP1; UniProt: RRBP1_HUMAN), with 141 amino acids covered in four regions: 232–253, 322–380, 392–460 and 472–498 (Suppl. Fig. S1).

The tandem repeat in hnRNPUL1, "QPPPQQPPPP", contains just two types of amino acids. Both hnRNPUL1 and its paralog, hnRNPUL2, are involved in building a complex during DNA damage repair (Polo et al., 2012). However, hnRNPUL2 lacks any QP rich regions or repeats (data not shown). Thus, the tandem repeat in hnRNPUL1 might be responsible of specific differences in function between these two proteins and it becomes one example of how fast these LCRs can evolve.

The multiple tandem repeats in RRBP1 are less compositionally biased and show small variations: "EGTPNQGKKA" or "EGAQNQGKKA". For RRBP1, two different large scale analyses of protein phosphorylation point to the phosphorylation of threonines at position three in four of the repeats (225, 235, 245 and 255) (Olsen et al., 2006, 2010). Except for multiple studies that report the expression of this protein in various cancers, there is no specific functional information for this protein. Post-translational modifications of repeats have been observed for other repeats (see e.g. the mineralocorticoid receptor; (Vlassi, Brauns, and Andrade-Navarro, 2013)) and could be associated with cooperative changes from structured

to unstructured domains of tandem repeats triggered by multiple phosphorylation. These repeats in RRBP1 could then have a phosphorylation-dependent function.

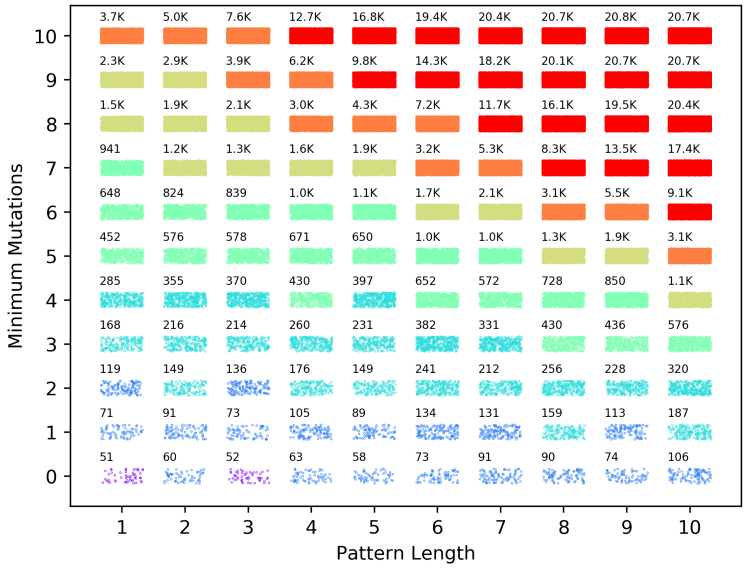
## 5.4 PROTEINS WITH REPEATABILITY IN HUMAN AND YEAST

Comparing the properties of short tandem repeats between different organisms can give us insight into the relation between repeat sequences and their structure and function. To illustrate how this can be done, we start with a comparison of the results obtained using the complete dataset of human and yeast (S. cerevisiae) Swiss-Prot proteins in release 2018_08; 21,057 and 6721 proteins, respectively. To focus on short tandem repeats we first count the proteins with at least one tandem repeat in a window of 20 allowing a maximum of 10 mutations from the perfect repeat.

The numbers of proteins detected to contain at least one tandem repeat in a window of 20 amino acids with the given required length (pattern length, x-axis) and allowing a minimum of mutations (y-axis) are displayed in Figure 5.2. For both species, given a repeat size, the more mutations allowed, the more proteins are found to have the repeat.

Numbers shown in Figure 5.2 are number of proteins detected to contain at least one tandem repeat in a window of 20 amino acids with the given required length (repeat size, x-axis) and allowing a minimum of mutations (y-axis). For example, as discussed in the text, there are 106 proteins in human with a perfect (minimum mutations = 0) tandem repeat of length 10 (pattern length = 10) in a window of 20 amino acids. For yeast, this number is 28 proteins. There are 51 human proteins with a perfect (minimum mutations = 0) homorepeat (pattern length = 1) filling the analysed window of 20 amino acids. For yeast, this number is 18 proteins. If we allow one mismatch (minimum mutations = 1; pattern length = 1) the number raises (71 and 26 proteins in human and yeast, respectively). A total of 21,057 and 6721 proteins were analysed for human and yeast, respectively.

**Human**




**Yeast**

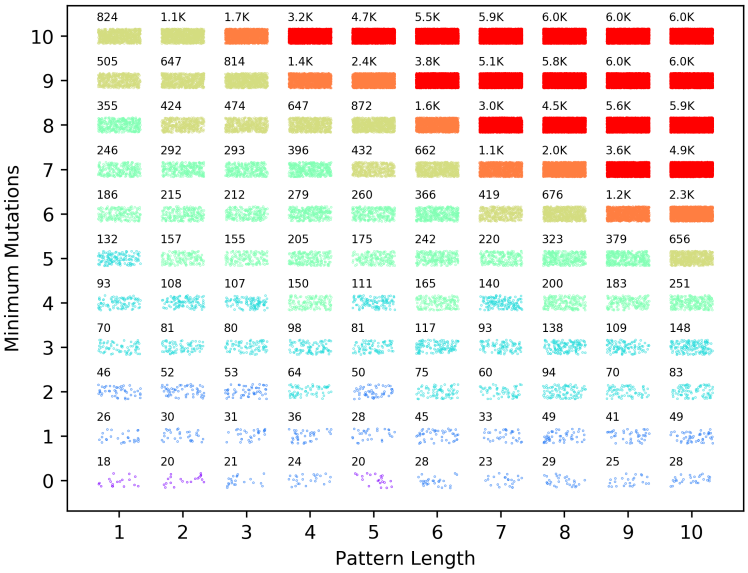



Figure 5.2: Proteins with repeatability

The colour correlates to the number of selected proteins in each condition and allows comparing positions in each graph.

In general, for an allowed number of mutations, we find an increase in the number of proteins when we look for repeats of larger size. But, if we focus on the repeats allowing four or less mutations, for yeast the values of the repeats

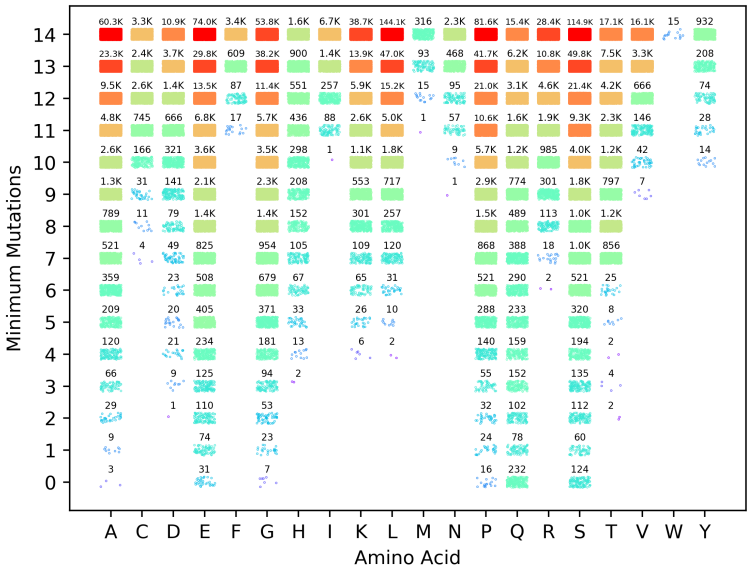
with odd-length 3, 5, 7 and 9 are lower or close to the value of the preceding even-length repeat 2, 4, 6 and 8, respectively. This tendency is not so strong for human. We will investigate further this observation using more species in a later section.

## 5.5 COMPOSITION OF HOMOREPEATS IN HUMAN AND YEAST

With our method, we can evaluate the repeat composition (amino acid choice) in complete datasets. For simplicity, we illustrate this with repeats of length one (homorepeats or polyX) in human and yeast. To account for the fact that the same protein sequence can contain several non-overlapping regions of different homorepeat types (e.g., both polyQ and polyP), instead of reporting the number of proteins containing at least one hit (as in Figure 5.2), we report the number of instances (amino acids) where there was a hit (Figure 5.4). While this is close to counting how many instances of amino acid X were found in a window, the algorithm imposes the additional condition that the homorepeat reported corresponds to the amino acid that was the most frequent in the window. This definition of polyX is much looser than the usual when considering large numbers of mutations but approaches the classical one for low mutations.

One major difference between human and yeast that becomes apparent is the lower use of polyA in yeast compared to human (Figure 5.4). In a window of 20, the most perfect yeast polyA is 5 mutations away to perfect polyA (one single case). In contrast, human proteins contain hundreds of windows equally close or closer to perfect polyA (with 3 for perfect). For polyN the situation is reversed. While the human polyN closest to length 20 is 9 mutations away (one single case), the yeast proteome has hundreds of more perfect polyN (with 18 positions giving the perfect 20 N repeat). These results agree with previous evaluations that found higher frequencies of polyN in fungi and S. cerevisiae and higher frequencies of polyA in human and Metazoa (Karlin et al., 2002; Mier, Alanis-Lobato, and Andrade-Navarro, 2017).

**Human**




**Yeast**

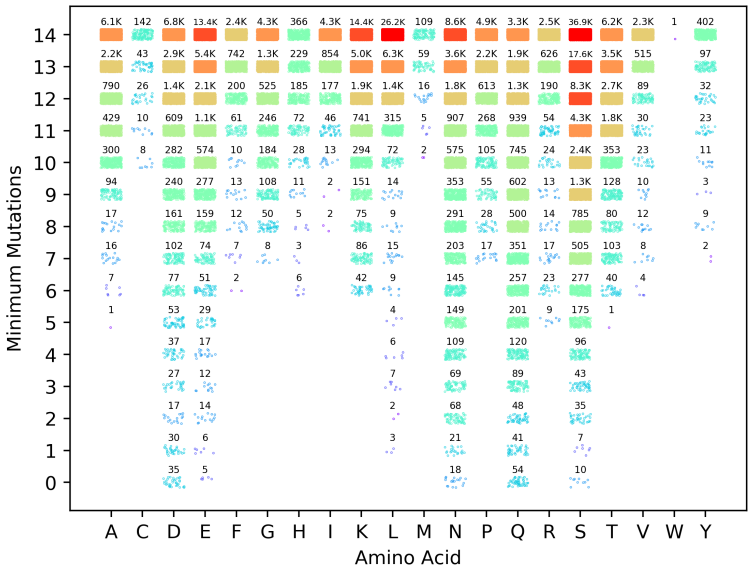



Figure 5.3: Composition of homorepeats

## 5.6 COMPARATIVE STUDY OF THE DISTRIBUTION OF REPEAT LENGTHS IN FULL PROTEOMES

We next addressed the question of which types of repeat lengths are found in different organisms. Following the observations mentioned in the section above

comparing the human and yeast length distributions, we decided to focus on the numbers of proteins presenting repeats of four or less mutations (window length 20). The distributions presented significant variation, which decreased with taxonomical closeness (compare Homo sapiens and Bos Taurus; Figure 5.4).

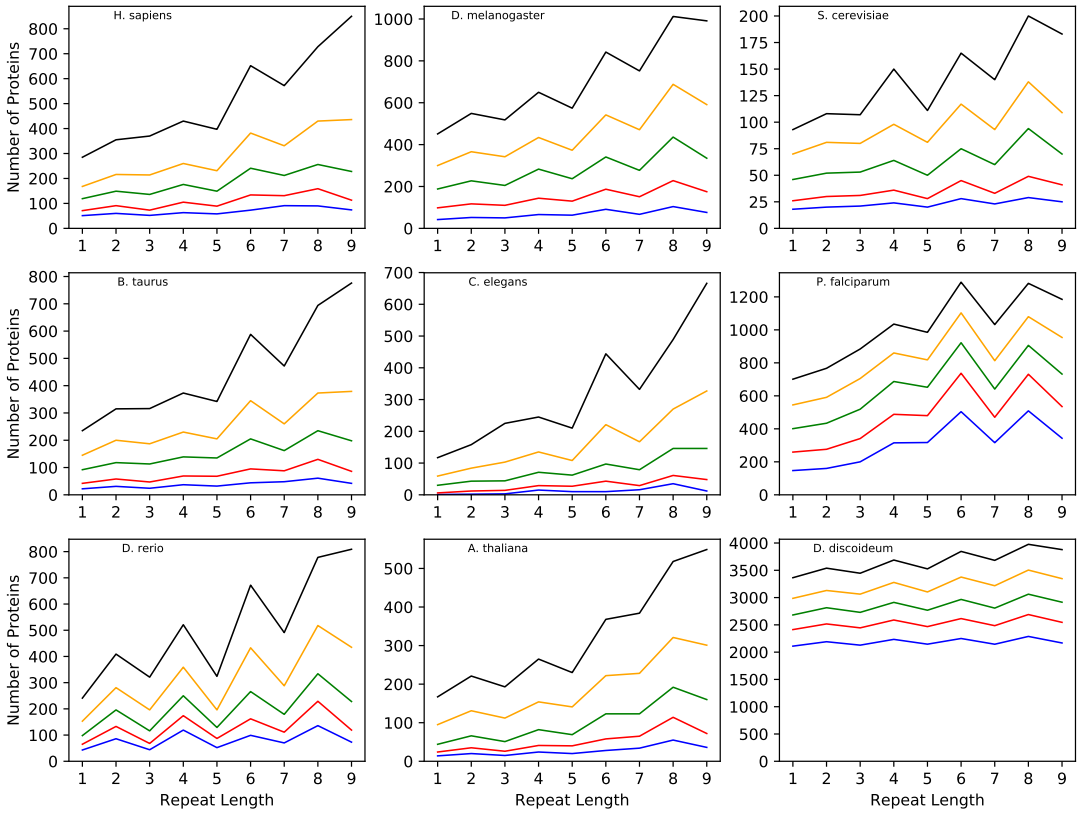



Figure 5.4: Distributions of proteins with repeatability by length in nine species

In Figure 5.4 we computed the number of proteins with repeats of length 1 to 9 (in a window of 20) considering those with perfect repeats (blue line), or allowing mutations to perfect repeats (1, 2, 3 or 4; lines from bottom, red, to top, black). For the complete reference proteomes of: Homo sapiens (73928 proteins), Bos taurus (23965 proteins), Danio rerio (46926 proteins), Drosophila melanogaster (21923 proteins), Caenorhabditis elegans (26898 proteins), Arabidopsis thaliana (39380 proteins), Saccharomyces cerevisiae (6049 proteins), Plasmodium falciparum (5449 proteins) and Dictyostelium discoideum (12746 proteins). The complete reference proteomes were obtained from UniProtKB > Proteomes.

Using human as reference, the results for yeast (S. cerevisiae) stand out by their depletion in the number of proteins with repeats of odd-length compared to the surrounding even-length repeats. While this is not a feature of all the proteomes observed here, traces of this periodicity can be observed in Danio rerio, Drosophila melanogaster and Plasmodium falciparum.

Another highlight when comparing the nine distributions is the high ratio of proteins with repeats of length 2 versus length 3 in D. rerio. This is more obvious when focusing on proteins with perfect repeats (blue line) or allowing one mutation (red line). This could be related to high levels of dinucleotide repeats in these species, which translated into amino acids result in repeats of length two.

Finally, we remark the high ratio of repeats of length 7 respect to length 6 in A. thaliana both allowing for 3 (orange) or 4 (black) mutations. Heptad repeats conform coiled coils but we could not appreciate general differences in coiled coil content explaining this observation (data not shown).

In agreement with the conservation of length profile between taxonomically related species (Figure 5.4), many of the repeats we find are evolutionarily conserved over long evolutionary distances, indicating that, rather than sequencing mistakes due to, for example, DNA tandem repeats, they are conserved functional features. We illustrate this with two examples taken from our previous discussion: length 2 repeats in D. rerio GSE1 protein (Figure 5.5-A), and length 7 repeats in A. thaliana At5g47430 protein (Figure 5.5-C).

Investigation of the homologs of these sequences suggests that the repeats in GSE1 were gained after the emergence of Osteichtyes from Chordata (present in Latimeria and in species closer to human, but not in Branchiostoma, or in sequences from other more distant species to Danio; Figure 5.5-B). A similar strategy indicated that the repeats in At5g47430 were gained after the emergence of Brassicaceae from Malvids (present in Brassica oleracea and species closer to Arabidopsis, but not in Theobroma cacao, or in sequences from other more distant species to Arabidopsis; Figure 5.5-D). It is remarkable that in both cases, the examples lacking the repeats do have sequence material corresponding to

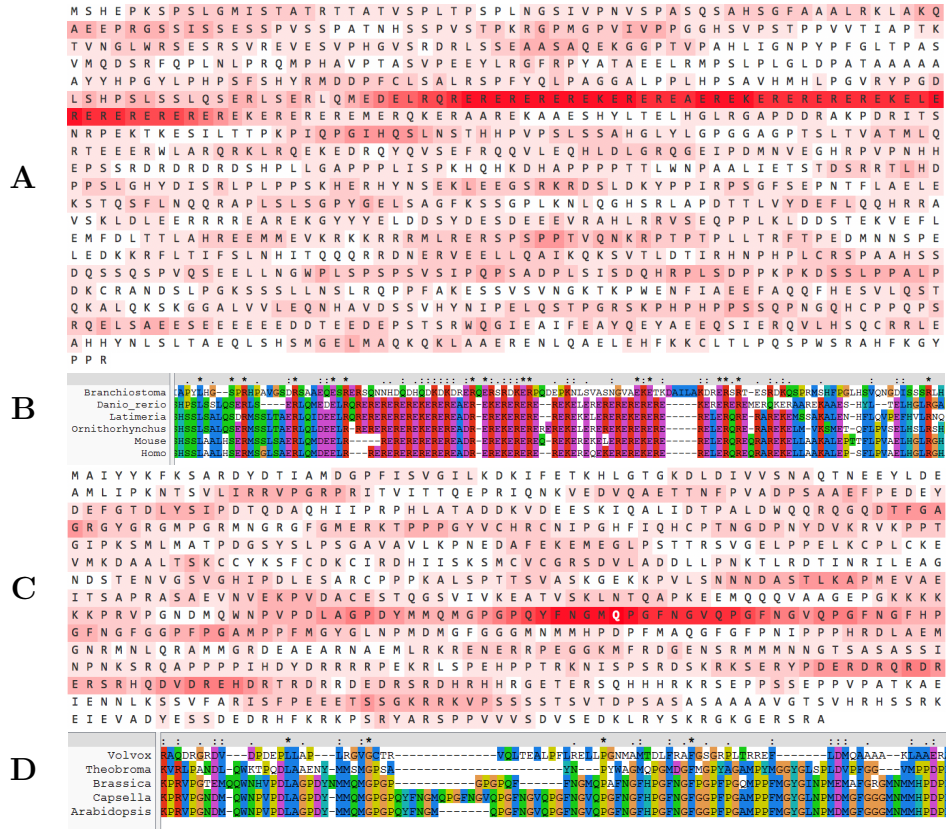



Figure 5.5: Repeats detected conserved over long evolutionary distances

A : Repeatability in D. rerio GSE1 (Genetic suppressor element 1; UniProt: Q1LWL6).

B : Part of the full length multiple sequence alignment of GSE1 with orthologs, including the repeats region. Branchiostoma floridae has sequence in this region but lacks repeatability.

C : Repeatability in A. thaliana At5g47430 (E3 ubiquitin ligase PQT3-like; UniProt: B9DFV2). Coiled coil prediction using COILS ((Lupas et al., 1991)) reported a region (amino acid positions 613-626), which does not overlap with the repeats; this suggests that this repeats of length 7 are not heptad coiled coil repeats.

D : Part of the full length multiple sequence alignment of protein At5g47430 with orthologs, including the repeats region.

the region, and they even have a few residues that resemble the repeats (some scattered "ER" motifs in the Branchiostoma sequence, Figure 5.5-B; one instance

of "QPGMDGF" in the Theobroma sequence, which matches and aligns to the tandem repeats, Figure 5.5-D).

Cursory analysis with disorder predictors in these sequences was not conclusive about the character of these regions (data not shown); further extensive and systematic analysis would be needed to clarify the evolution of such short tandem repeats in these and other taxonomic contexts.

Evolutionary paths and corresponding alignments with orthologs were investigated with the assistance of ProteinPathTracker (Mier, Alanis-Lobato, and Andrade-Navarro, 2017) using as input each of the D. rerio and A. thaliana sequences, with the path "cellular organisms to homo" or "viridiplantae to arabidopsis", respectively. Multiple sequence alignments were produced using the MUSCLE server at EBI (Edgar, 2004) and represented with ClustalX (Larkin et al., 2007). Multiple sequence alignments and sequence FASTA files are available as Supplementary Files S3–S6.

Taken together, our results suggest that the algorithm proposed and implemented provides informative results for individual proteins and at the level of proteomes. In combination with functional and evolutionary studies, the concept and computation of repeatability should aid the investigation of LCRs and their characterization as structure-prone.

## 5.7 CONCLUSION

In this chapter, we have seen in-depth analyses on RES algorithm applied to different proteomes especially Human and Yeast.

Chapter

# 6

# CONCLUSIONS

Here, we have presented a simple and linearly complex, thus fast, algorithm to detect and identify positions in protein sequences approximate to very short tandem repeats. We introduce the related concept of repeatability in protein sequences, which we define as the fraction of amino acid changes necessary to convert a sequence into perfect tandem repeats. Repeatability is a property that allows to classify compositionally biased low complexity regions.

Note that we considered only identical amino acids when computing the repeatability of a region. While amino acid similarity is helpful in defining homology between globular domains (Dayhoff, 1972), accumulating evidence suggests that in dealing with LCRs and homorepeats, amino acid similarity is a concept that does not help very much. For example, while leucine and isoleucine are amino acids with very similar amino acid chains, in the human proteome many polyL

regions can be found whereas there are no polyI regions ((Mier, Alanis-Lobato, and Andrade-Navarro, 2017); see also Figure 5.4).

Our concept of repeatability uses the Hamming distance from the sequence to the perfect repeat. Other scoring functions have been used to evaluate repeat perfection. The Shannon entropy was used to identify low complexity regions (LCRs) (Li and Kahveci, 2006). We believe that our scoring function is more specific as it relates directly to mutations from perfect repeats, while Li and Kahveci were more interested in defining the complexity of the sequence. Most importantly, our measure does not take into account the frequency of amino acids.

There are a number of online tools to detect repeats in protein sequences, which have a different focus and provide different outputs than RES. XSTREAM (Newman and Cooper, 2007), T-REKs (Jorda and Kajava, 2009) and RADAR (Heger and Holm, 2000) provide multiple sequence alignments of groups of repeats and run in a few seconds. REPRO (George and Heringa, 2000) provides pairwise alignments and has the longest running times (10–30 s) for the longer sequences with many repeats (e.g. RRBP1_HUMAN). While the outputs of the first three methods are easier to interpret, REPRO produces a list of pairwise alignments; of the four, this is the only method that reports the alignment of the 10 amino acid perfect tandem repeat in hnRNPUL1 (Suppl. Fig. S1A: QPPPQQPPPP x2), but this is ranked 19 among a total of 50. Our method is different from those above in that RES computes and represents the value of a parameter at each position of the sequence that represents the repeatability in the region. Repeats and number of mutations needed to reach them are provided at each position, which helps to understand the context and extension of the repeat regions in relation to other properties of the sequence. The possibility to restricting the search to a range of repeat lengths is unique to RES; this option can be helpful when testing the hypothesis of whether a particular repeat exists. Our tool is therefore more specific to the characterization of regions with very short tandem repeats. Our proteomics analyses indicate that many proteins have them; while such regions probably look like compositionally biased regions, our method can help in assessing their repeatability.

We recommend to use our algorithm specifically on regions in between domains to pinpoint possible functional regions in otherwise apparently disordered regions. For longer repeats (longer than 10 amino acids) there are other tools that use HMMs and are more appropriate for the detection of such repeats ab initio (e.g. with HHrepID (Zimmermann et al., 2018)). Contrasting the results with the repeat detectors mentioned above might as well help to confirm or aid in the interpretations of results from RES.

Our algorithm is very sensitive to insertions and deletions that break the frame. While this could be solved by allowing indels, we believe that the cost of highest computational time and increased numbers of positives would diminish the efficiency of the method. We also note that if the web tool is used to search for overlapping repeats of different sizes, impurities in frame with the repeats can cause detection of spurious longer repeats. Regardless, the application of the method to different proteomes provided interesting differences and results (e.g. the perfect repeats of length 10, or the shorter repeats used as examples in Figure 5.5).

The results presented indicate that our algorithm opens many possibilities to the study and characterization of LCRs. Their high presence in Eukaryotic species has been attributed to a possible role in the formation of new protein sequences (Toll-Riera et al., 2011). We suggest that many LCRs contain actually conserved approximate short tandem repeats and that these make them prone to adopt structure and function. The few examples investigated in detail here suggest that these regions of short tandem repeats are conserved over relatively long evolutionary distances and that they evolve from regions of low conservation sharing some features with the composition of the repeats. Future work should evaluate the evolution of LCRs with repeatability, which will help us with their characterization and in understanding the rules that govern the emergence of new protein sequences and function.

# BIBLIOGRAPHY


Abraham, Anne-Laure, Eduardo PC Rocha, and Joël Pothier (2008). "Swelfe: a detector of internal repeats in sequences and structures." In: *Bioinformatics* 24.13, pp. 1536–1537.

Abril, Josep F. and Sergi Castellano (2019). "Genome Annotation." In: *Encyclopedia of Bioinformatics and Computational Biology*. Ed. by Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach. Oxford: Academic Press, pp. 195–209.

Andrade, Miguel A, Chris P Ponting, Toby J Gibson, and Peer Bork (2000). "Homology-based method for identification of protein repeats using statistical significance estimates." In: *Journal of molecular biology* 298.3, pp. 521–537.

Baxa, Ulrich, Todd Cassese, Andrey V Kajava, and Alasdair C Steven (2006). "Structure, function, and amyloidogenesis of fungal prions: filament polymorphism and prion variants." In: *Advances in protein chemistry* 73, pp. 125–180.

Berk, Arnold, David Baltimore, Harvey Lodish, James Darnell, Paul Matsudaira, and S Lawrence Zipursky (2013). *Molekulare Zellbiologie*. Walter de Gruyter.

Biegert, Andreas and Johannes Söding (2008). "De novo identification of highly diverged protein repeats by probabilistic consistency." In: *Bioinformatics* 24.6, pp. 807–814.

Chakrabarty, Broto and Nita Parekh (2014). "PRIGSA: protein repeat identification by graph spectral analysis." In: *Journal of bioinformatics and computational biology* 12.06, p. 1442009.

Chan, Cheong Xin, Guillaume Bernard, Olivier Poirion, James M Hogan, and Mark A Ragan (2014). "Inferring phylogenies of evolving sequences without multiple sequence alignment." In: *Scientific reports* 4.1, pp. 1–9.

Chavali, Sreenivas, Pavithra L Chavali, Guilhem Chalancon, Natalia Sanchez de Groot, Rita Gemayel, Natasha S Latysheva, Elizabeth Ing-Simmons, Kevin J Verstrepen, Santhanam Balaji, and M Madan Babu (2017). "Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins." In: *Nature structural & molecular biology* 24.9, p. 765.

Chen, Hanlin, Yanzhao Huang, and Yi Xiao (2009). "A simple method of identifying symmetric substructures of proteins." In: *Computational biology and chemistry* 33.1, pp. 100–107.

Crow, James F (2002). "Unequal by nature: A geneticist's perspective on human differences." In: *Daedalus* 131.1, pp. 81–88.

Dahm, Ralf (2005). "Friedrich Miescher and the discovery of DNA." In: *Developmental biology* 278.2, pp. 274–288.

Dayhoff, Margaret O (1972). "A model of evolutionary change in proteins." In: *Atlas of protein sequence and structure* 5, pp. 89–99.

Depledge, Daniel P, Ryan PJ Lower, and Deborah F Smith (2007). "RepSeq–a database of amino acid repeats present in lower eukaryotic pathogens." In: *BMC bioinformatics* 8.1, p. 122.

Di Domenico, T, E Potenza, I Walsh, RG Parra, M Giollo, G Minervini, D Piovesan, A Ihsan, C Ferrari, AV Kajava, et al. (2013). "RepeatsDB: a database of tandem repeat protein structures." In: *Nucleic Acids Research* 42.Database issue, pp. D352–7.

Do Viet, Phuong, Daniel B Roche, and Andrey V Kajava (2015). "TAPO: A combined method for the identification of tandem repeats in protein structures." In: *FEBS letters* 589.19, pp. 2611–2619.

Edgar, Robert C (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." In: *Nucleic acids research* 32.5, pp. 1792–1797.

Fan, Huan, Anthony R Ives, Yann Surget-Groba, and Charles H Cannon (2015). "An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data." In: *BMC genomics* 16.1, pp. 1–18.

Faux, Noel G, Stephen P Bottomley, Arthur M Lesk, James A Irving, John R Morrison, Maria Garcia De La Banda, and James C Whisstock (2005). "Functional insights from the distribution and role of homopeptide repeat-containing proteins." In: *Genome research* 15.4, pp. 537–551.

Fleischmann, Robert D, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." In: *Science* 269.5223, pp. 496–512.

Gatherer, Derek and Neil R McEwan (2005). "Phylogenetic differences in content and intensity of periodic proteins." In: *Journal of molecular evolution* 60.4, pp. 447–461.

George, Richard A and Jaap Heringa (2000). "The REPRO server: finding protein internal sequence repeats through the Web." In: *Trends in biochemical sciences* 25.10, pp. 515–517.

Goodsell, David S and Arthur J Olson (2000). "Structural symmetry and protein function." In: *Annual review of biophysics and biomolecular structure* 29.1, pp. 105–153.

Guerler, Aysam, Connie Wang, and Ernst-Walter Knapp (2009). "Symmetric structures in the universe of protein folds." In: *Journal of chemical information and modeling* 49.9, pp. 2147–2151.

Hackman, J Peter V, Anna K Vihola, and A Bjarne Udd (2003). "The role of titin in muscular disorders." In: *Annals of medicine* 35.6, pp. 434–441.

Harrison, Paul M (2006). "Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila." In: *BMC bioinformatics* 7.1, p. 441.

Heger, Andreas and Liisa Holm (2000). "Rapid automatic detection and alignment of repeats in protein sequences." In: *Proteins: Structure, Function, and Bioinformatics* 41.2, pp. 224–237.

Heringa, Jaap and Patrick Argos (1993). "A method to recognize distant repeats in protein sequences." In: *Proteins: Structure, Function, and Bioinformatics* 17.4, pp. 391–411.

Hirsh, Layla, Lisanna Paladin, Damiano Piovesan, and Silvio C E Tosatto (2018). "RepeatsDB-lite: a web server for unit annotation of tandem repeat proteins." In: *Nucleic acids research* 46.W1, W402–W407.

Hrabe, Thomas and Adam Godzik (2014). "ConSole: using modularity of Contact maps to locate Solenoid domains in protein structures." In: *BMC bioinformatics* 15.1, pp. 1–12.

Huerta, Michael, Gregory Downing, Florence Haseltine, Belinda Seto, and Yuan Liu (2000). "NIH working definition of bioinformatics and computational biology." In: *US National Institute of Health.*

Hunter, Sarah, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, et al. (2009). "InterPro: the integrative protein signature database." In: *Nucleic acids research* 37.suppl_1, pp. D211–D215.

Itoh-Satoh, Manatsu, Takeharu Hayashi, Hirofumi Nishi, Yoshinori Koga, Takuro Arimura, Takeshi Koyanagi, Megumi Takahashi, Shigeru Hohda, Kazuo Ueda, Tatsuhito Nouchi, et al. (2002). "Titin mutations as the molecular basis for dilated cardiomyopathy." In: *Biochemical and biophysical research communications* 291.2, pp. 385–393.

Jeffreys, Alec J, Victoria Wilson, and Swee Lay Thein (1985). "Individual-specific 'fingerprints' of human DNA." In: *Nature* 316.6023, pp. 76–79.

Jorda, Julien, Thierry Baudrand, and Andrey V Kajava (2012). "PRDB: Protein Repeat Data Base." In: *Proteomics* 12.9, pp. 1333–1336.

Jorda, Julien and Andrey V Kajava (2009). "T-REKS: identification of Tandem REpeats in sequences with a K-means based algorithm." In: *Bioinformatics* 25.20, pp. 2632–2638.

Jorda, Julien and Andrey V Kajava (2010). "Protein homorepeats: sequences, structures, evolution, and functions." In: *Advances in protein chemistry and structural biology* 79, pp. 59–88.

Kajava, Andrey V (2012). "Tandem repeats in proteins: from sequence to structure." In: *Journal of structural biology* 179.3, pp. 279–288.

Kalita, Mridul K, Gowthaman Ramasamy, Sekhar Duraisamy, Virander S Chauhan, and Dinesh Gupta (2006). "ProtRepeatsDB: a database of amino acid repeats in genomes." In: *BMC bioinformatics* 7.1, pp. 1–11.

Karlin, Samuel, Luciano Brocchieri, Aviv Bergman, Jan Mrázek, and Andrew J Gentles (2002). "Amino acid runs in eukaryotic proteomes and disease associations." In: *Proceedings of the National Academy of Sciences* 99.1, pp. 333–338.

Katti, Mukund V, R Sami-Subbu, Prabhakar K Ranjekar, and Vidya S Gupta (2000). "Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications." In: *Protein Science* 9.6, pp. 1203–1209.

Kim, Changhoon, Jodi Basner, and Byungkook Lee (2010). "Detecting internally symmetric protein structures." In: *BMC bioinformatics* 11.1, pp. 1–16.

Larkin, Mark A, Gordon Blackshields, NP Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. (2007). "Clustal W and Clustal X version 2.0." In: *bioinformatics* 23.21, pp. 2947–2948.

Li, Xuehui and Tamer Kahveci (2006). "A Novel algorithm for identifying low-complexity regions in a protein sequence." In: *Bioinformatics* 22.24, pp. 2980–2987.

Luo, Hong, Ke Lin, Audrey David, Harm Nijveen, and Jack AM Leunissen (2012). "ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins." In: *Nucleic acids research* 40.D1, pp. D394–D399.

Machado, Cristina, Claudio E Sunkel, and Deborah J Andrew (1998). "Human autoantibodies reveal titin as a chromosomal protein." In: *The Journal of cell biology* 141.2, pp. 321–333.

Mahdavi, Mahmood A (2011). *Bioinformatics: Trends and Methodologies*. BoD–Books on Demand.

Marcotte, Edward M, Matteo Pellegrini, Todd O Yeates, and David Eisenberg (1999). "A census of protein repeats." In: *Journal of molecular biology* 293.1, pp. 151–160.

Marsella, Luca, Francesco Sirocco, Antonio Trovato, Flavio Seno, and Silvio CE Tosatto (2009). "REPETITA: detection and discrimination of the periodicity

of protein solenoid repeats by discrete Fourier transform." In: *Bioinformatics* 25.12, pp. i289–i295.

Mier, Pablo, Gregorio Alanis-Lobato, and Miguel A Andrade-Navarro (2017). "Context characterization of amino acid homorepeats using evolution, position, and order." In: *Proteins: Structure, Function, and Bioinformatics* 85.4, pp. 709–719.

Mier, Pablo, Lisanna Paladin, Stella Tamana, Sophia Petrosian, Borbála Hajdu-Soltész, Annika Urbanek, Aleksandra Gruca, Dariusz Plewczynski, Marcin Grynberg, Pau Bernadó, et al. (2019). "Disentangling the complexity of low complexity proteins." In: *Briefings in bioinformatics*.

Mott, Richard (1999). "Local sequence alignments with monotonic gap penalties." In: *Bioinformatics (Oxford, England)* 15.6, pp. 455–462.

Murray, Kevin B, William R Taylor, and Janet M Thornton (2004). "Toward the detection and validation of repeats in protein structure." In: *Proteins: Structure, Function, and Bioinformatics* 57.2, pp. 365–380.

Nelson, Rebecca and David Eisenberg (2006). "Structural models of amyloid-like fibrils." In: *Advances in protein chemistry* 73, pp. 235–282.

Newman, Aaron M and James B Cooper (2007). "XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences." In: *BMC bioinformatics* 8.1, p. 382.

Nirenberg, Marshall W and J Heinrich Matthaei (1961). "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides." In: *Proceedings of the National Academy of Sciences* 47.10, pp. 1588–1602.

Olsen, Jesper V, Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann (2006). "Global, in vivo, and site-specific phosphorylation dynamics in signaling networks." In: *Cell* 127.3, pp. 635–648.

Olsen, Jesper V, Michiel Vermeulen, Anna Santamaria, Chanchal Kumar, Martin L Miller, Lars J Jensen, Florian Gnad, Jürgen Cox, Thomas S Jensen, Erich A Nigg, et al. (2010). "Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis." In: *Sci. Signal.* 3.104, ra3–ra3.

Parra, R Gonzalo, Rocío Espada, Ignacio E Sánchez, Manfred J Sippl, and Diego U Ferreiro (2013). "Detecting repetitions and periodicities in proteins by tiling the structural space." In: *The Journal of Physical Chemistry B* 117.42, pp. 12887–12897.

Pellegrini, Marco, Maria Elena Renda, and Alessio Vecchio (2012). "Ab initio detection of fuzzy amino acid tandem repeats in protein sequences." In: *Bmc Bioinformatics*. Vol. 13. 3. BioMed Central, pp. 1–13.

Pellegrini, Matteo, Edward M Marcotte, and Todd O Yeates (1999). "A fast algorithm for genome-wide analysis of proteins with repeated sequences." In: *Proteins: Structure, Function, and Bioinformatics* 35.4, pp. 440–446.

Peng, Zhenling, Jing Yan, Xiao Fan, Marcin J Mizianty, Bin Xue, Kui Wang, Gang Hu, Vladimir N Uversky, and Lukasz Kurgan (2015). "Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life." In: *Cellular and Molecular Life Sciences* 72.1, pp. 137–151.

Petrakis, Spyros, Martin H Schaefer, Erich E Wanker, and Miguel A Andrade-Navarro (2013). "Aggregation of polyQ-extended proteins is promoted by interaction with their natural coiled-coil partners." In: *BioEssays* 35.6, pp. 503–507.

Polo, Sophie E, Andrew N Blackford, J Ross Chapman, Linda Baskcomb, Serge Gravel, Andre Rusch, Anoushka Thomas, Rachel Blundred, Philippa Smith, Julia Kzhyshkowska, et al. (2012). "Regulation of DNA-end resection by hnRNPU-like proteins promotes DNA double-strand break signaling and repair." In: *Molecular cell* 45.4, pp. 505–516.

Rana, Jagmohan and Dr. Kunwar Vaisla (Dec. 2012). "Introduction To Bioinformatics." In: pp. 11–18. ISBN: 978-81-923296-3-5.

Regad, Leslie, Jean-Baptiste Chéron, Dhoha Triki, Caroline Senac, Delphine Flatters, and Anne-Claude Camproux (2017). "Exploring the potential of a structural alphabet-based tool for mining multiple target conformations and target flexibility insight." In: *PloS one* 12.8, e0182972.

Ren, Jie, Xin Bai, Yang Young Lu, Kujin Tang, Ying Wang, Gesine Reinert, and Fengzhu Sun (2018). "Alignment-free sequence analysis and applications." In: *Annual Review of Biomedical Data Science* 1, pp. 93–114.

Richard, Francois D and Andrey V Kajava (2014). "TRDistiller: A rapid filter for enrichment of sequence datasets with proteins containing tandem repeats." In: *Journal of structural biology* 186.3, pp. 386–391.

Rudenko, Valentina and Eugene Korotkov (2021). "Search for Highly Divergent Tandem Repeats in Amino Acid Sequences." In: *International journal of molecular sciences* 22.13, p. 7096.

Sabarinathan, R, Raunak Basu, and Krishna Sekar (2010). "ProSTRIP: A method to find similar structural repeats in three-dimensional protein structures." In: *Computational biology and chemistry* 34.2, pp. 126–130.

Schaefer, Martin H, Erich E Wanker, and Miguel A Andrade-Navarro (2012). "Evolution and function of CAG/polyglutamine repeats in protein–protein interaction networks." In: *Nucleic acids research* 40.10, pp. 4273–4287.

Schaper, Elke, Alexander Korsunsky, Jūlija Pečerska, Antonio Messina, Riccardo Murri, Heinz Stockinger, Stefan Zoller, Ioannis Xenarios, and Maria Anisimova (2015). "TRAL: tandem repeat annotation library." In: *Bioinformatics* 31.18, pp. 3051–3053.

Schultz, Jörg, Frank Milpetz, Peer Bork, and Chris P Ponting (1998). "SMART, a simple modular architecture research tool: identification of signaling domains." In: *Proceedings of the National Academy of Sciences* 95.11, pp. 5857–5864.

Sharp, Stephen Jefferson, Jerone Schaack, Lyan Cooley, Debroh Johnson Burke, and Dieter Soil (1985). "Structure and transcription of eukaryotic tRNA gene." In: *Critical Reviews in Biochemistry* 19.2, pp. 107–144.

Shih, Edward SC and Ming-Jing Hwang (2004). "Alternative alignments from comparison of protein structures." In: *Proteins: Structure, Function, and Bioinformatics* 56.3, pp. 519–527.

Siwach, Pratibha and Subramaniam Ganesh (2008). "Tandem repeats in human disorders: mechanisms and evolution." In: *Front. Biosci* 13, pp. 4467–4484.

Söding, Johannes, Michael Remmert, and Andreas Biegert (2006). "HHrep: de novo protein repeat detection and the origin of TIM barrels." In: *Nucleic acids research* 34.suppl_2, W137–W142.

Stirnimann, Christian U, Evangelia Petsalaki, Robert B Russell, and Christoph W Müller (2010). "WD40 proteins propel cellular networks." In: *Trends in biochemical sciences* 35.10, pp. 565–574.

Suzuki, Katsuyuki, Joe Olvera, and Ira G Wool (1991). "Primary structure of rat ribosomal protein S2. A ribosomal protein with arginine-glycine tandem repeats and RGGF motifs that are associated with nucleolar localization and binding to ribonucleic acids." In: *Journal of Biological Chemistry* 266.30, pp. 20007–20010.

Szklarczyk, Radek and Jaap Heringa (2004). "Tracking repeats using significance and transitivity." In: *Bioinformatics* 20.suppl_1, pp. i311–i317.

Toll-Riera, Macarena, Núria Radó-Trilla, Florian Martys, and M Mar Alba (2011). "Role of low-complexity sequences in the formation of novel protein coding sequences." In: *Molecular biology and evolution* 29.3, pp. 883–886.

Toro Acevedo, Carlos A, Bruna M Valente, Gabriela A Burle-Caldas, Bruno Galvão-Filho, Helton da C Santiago, Rosa M Esteves Arantes, Caroline Junqueira, Ricardo T Gazzinelli, Ester Roffê, and Santuza MR Teixeira (2017). "Down modulation of host immune response by amino acid repeats present in a Trypanosoma cruzi ribosomal antigen." In: *Frontiers in microbiology* 8, p. 2188.

Totzeck, Franziska, Miguel A Andrade-Navarro, and Pablo Mier (2017). "The protein structure context of polyQ regions." In: *PloS one* 12.1, e0170801.

Vickery, Hubert Bradford and Carl LA Schmidt (1931). "The history of the discovery of the amino acids." In: *Chemical Reviews* 9.2, pp. 169–318.

Vinga, Susana and Jonas Almeida (2003). "Alignment-free sequence comparison—a review." In: *Bioinformatics* 19.4, pp. 513–523.

Vlassi, Metaxia, Katharina Brauns, and Miguel A Andrade-Navarro (2013). "Short tandem repeats in the inhibitory domain of the mineralocorticoid receptor: prediction of a $\beta$-solenoid structure." In: *BMC structural biology* 13.1, p. 17.

Walsh, Ian, Francesco G Sirocco, Giovanni Minervini, Tomás Di Domenico, Carlo Ferrari, and Silvio CE Tosatto (2012). "RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures." In: *Bioinformatics* 28.24, pp. 3257–3264.

Woese, Carl R and George E Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." In: *Proceedings of the National Academy of Sciences* 74.11, pp. 5088–5090.

Woese, Carl R, Otto Kandler, and Mark L Wheelis (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." In: *Proceedings of the National Academy of Sciences* 87.12, pp. 4576–4579.

Wyman, Arlene R and Ray White (1980). "A highly polymorphic locus in human DNA." In: *Proceedings of the National Academy of Sciences* 77.11, pp. 6754–6758.

Zimmermann, Lukas, Andrew Stephens, Seung-Zin Nam, David Rau, Jonas Kübler, Marko Lozajic, Felix Gabler, Johannes Söding, Andrei N Lupas, and Vikram Alva (2018). "A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core." In: *Journal of molecular biology* 430.15, pp. 2237–2243.

**Abstract**

Short tandem repeats and homorepeats are considered as low complexity regions (LCRs) which have special properties that are very different from those of globular proteins. The rules that define secondary structure elements do not apply when the distribution of amino acids becomes biased. While there is a tendency towards structural disorder in LCRs, various examples, and particularly homorepeats of single amino acids, suggest that very short repeats could adopt structures very difficult to predict.

For these reasons, we have developed an algorithm to quickly analyze local repeatability along protein sequences, that is, how close a protein fragment is from a perfect repeat.

Our method (REpeatability Scanner, RES, accessible at http://cbdm-01.zdv.uni-mainz.de/~munoz/res/) allows to find regions with approximate short repeats in protein sequences, and helps to characterize the variable use of LCRs and compositional bias.

**Keywords:** Amino acid short tandem repeats, Low complexity regions, Computational detection of sequence repeats, Homorepeats, Repeatability

**Résumé**

Les répétitions courtes en série et le homorépétitions sont considérées comme des régions de faible complexité, ils ont des propriétés particulières qui sont très différentes de celles des protéines globulaires. Les règles qui définissent les éléments de structure secondaire ne s'appliquent pas lorsque la distribution des acides aminés devient biaisée. Alors qu'il existe une tendance au désordre structurel dans les régions de faible complexité, divers exemples, et notamment les homorépétitions d'acides aminés simples, suggèrent que des répétitions très courtes pourraient adopter des structures très difficiles à prédire.

Pour ces raisons, nous avons développé un algorithme permettant d'analyser rapidement la répétabilité locale le long des séquences protéiques, c'est-à-dire la proximité d'un fragment protéique par rapport à une répétition parfaite.

Notre méthode (REpeatability Scanner, RES, accessible à l'adresse http://cbdm-01.zdv.uni-mainz.de/~munoz/res/) permet de trouver des régions présentant des répétitions courtes approximatives dans les séquences protéiques, et aide à caractériser l'utilisation variable des LCR et le biais de composition.

**Mots clés:** Répétitions courtes en série d'acides aminés, régions de faible complexité, détection informatique des répétitions de séquences, homorépétitions, répétabilité.

**ملخص**

**تعتبر التكرارات المتجانسة من أحماض أمينية متعددة أو منفردة مناطق منخفضة التعقيد وتمتلك خواص فضائية مختلفة كثيرا عن خواص البروتينات الكروية. القواعد التي تحدد عناصر البنية الثانوية لا تنطبق عندما يصبح توزيع الأحماض الأمينية متحيز. بالرغم من أن هناك ميلان الآراء نحو وجود اضطراب في المناطق المنخفظة التعقيد، إلا أنه توجد أمثلة متعددة، خاصة التكرارات المتجانسة من أحماض أمينية ذات نوع واحد، تشير إلى أن التكرارات القصيرة للغاية قد تتخذ بنيات جد صعبة التوقع.**

**لهذه الأسباب، طورنا خوارزم لتحليل، وبشكل سريع، القابلية التكرارية المحلية على طول سلاسل البروتين، أي مدى قرب قطعة بروتينية معينة من تكرار مثالي أو تام.**

**إن طريقتنا** REpeatability Scanner أو RES **المتوفرة على الرابط:**

http://cbdm-01.zdv.uni-mainz.de/~munoz/res/

**تسمح بإيجاد مناطق ذات تكرارات قصيرة بالسلاسل البروتينية، وتساعد على تمييز الاستعمالات المتعددة للمناطق المنخفظة التعقيد المتحيزة تركيبيا.**

**كلمات مفتاحية:** **تكرارات أحماض أمينية قصيرة ومترادفة، مناطق منخفظة التعقيد، كشف التكرارات الآلي، القابلية التكرارية.**