Learning Sequence Motifs Using Expectation Maximization (EM) algorithm and the hyperthermophilic archaeon Pyrococcus Furiosus genome

Physical Methods of Biology

Author: Manuela Carriero

Dated: September 24, 2021

Contents

Abstract

This project provides an introspection into what are *motifs* in molecular biology and their biological importance together with methods to detect DNA and protein sequence motifs. The application of this work is focused on the hyperthermophilic archaeon Pyrococcus Furiosus and a genetically tractable Pyrococcus Furiosus strain called COM1 that has the particular characteristics to resist to gamma irradiation despite that a number of DNA repair proteins are actually or potentially disrupted.
The known diversity of metabolic strategies and physiological adaptations of archaeal species to extreme environments is extraordinary and in this project, some possible mechanisms of metabolic adaptation of COM1 are studied by searching for overrepresented DNA sequence motifs related to radioresistance. The results show that the DNA sequence motif GARGADRHTGHGGMAGAGDTY, which is detected by the software MEME using a set of 8 Ferritin-like DNA sequence genes involved in the protection against oxidative damage, is overrepresented in COM1 with respect to the reference P. Furiosus genome. This can suggest that the radioresistance of COM1 is due to amplification of genes encoding proteins which belong to Ferritin family, or of genes with similar functions, following the destruction or inactivation of DNA repair genes.

1 Introduction

Archaea and the Tree of Life

For the longest time we have had this belief in biology that there is these two different and extremely distinct forms of life on our planet. One is everything we can see, right from the insects to the trees, human beings, other mammals, fishes, and so on and so forth. You know, the eukaryotes. And there is everything that is invisible and we call them prokaryotes. So prokaryotes was this all-encopassing term that was used for unicellular microorganisms that did not have organelles within the cells. So their chromosome, essentially, just floats around in their cytosol. In contrast Eukaryotes are organisms that have evolved these complex organelles one of which is the nucleus which encapsulates all of their chromosomal material. And so that distinction was what was used to make evolutionary relationships between everything that does have a nucleus and everything that does not. Turn's out that it is not as simple as that anymore. In 1977 Carl Woese at the University of Illinois wanted to use not just how organisms look, their morphology, as a way to understand how they are related to each but looking at parts of their genetic

material, their DNA, to make this evolutionary connections. The evolutions of new species begins with mutations in DNA. These mutations can change genes which in turn can change the physical traits of organisms. Sometimes when an organism passes these genetic changes to their offspring, it can cause a new species to emerge. In the 1960's scientists began to wonder if they could determine the evolutionary relationship between different species by comparing their DNA sequences. The more similar the DNA sequences of two organisms, the more closely related they must be. Less shared DNA suggests a more distant relationship. To build the Tree of Life using this approach, Carl Woese neeeded to find a gene that he could compare across all life forms. He chose one that was needed for cells to form one of their most essential functions, making proteins. By comparing the sequence of this gene between different types of organisms, he was able to infer the Tree of Life. This particular molecule that is narrowed in on is called the 16S ribosomal RNA. By looking at just this one particular molecule he was able to glean evolutionary relationships of these organisms without even having to look at them. The 16S ribosomal RNA was in fact different. So if you look them at the microscope they might look similar, but if you look deeper into the cell, into the DNA, you see that they are actually different group of organisms. Microbes that we all thought that there was just this big pool of prokaryotes were not that. There were the bacteria that are well enstablished and studied, but then there were these archaea, this enigmatic third domain of life. Archaea initially when they were discovered were found in extreme environments. So, one of the first archaea that was sequenced was isolated from hydrothermal vent at the bottom of the ocean. Now we have essentially found Archaea everywhere, in the environment around us, as well as within us. There is Archaea on our skin, in our oral cavity, in our guts and every other place you can think of there are Archaea.

Archaea in our investigation

The genome sequence of the model archaeon Pyrococcus Furiosus was determined 21 years ago. A variant, designated COM1, was discovered within the wild-type population that is naturally and efficiently competent for exogenous DNA uptake in both circular and linear forms [3]. The genome sequence comparison of the two strains was performed [2] and this shows a high plasticity of the P. furiosus genome: despite many differences (such as many inversions) between the two genome sequences, phenotypic properties are preserved. In particular, despite a number of genes involved in DNA repair are actually or potentially inactive in COM1 (because of nonsynonymous mutations in these genes that frequently lead to nonfunctional

proteins), COM1 showed no significant differences in its ability to recover from exposure to UV or gamma irradiation.

Is this survival strategy based on the fact that for this organism DNA repair genes are not important in the survival mechanisms following radiation exposure (against all the actual knowledges) or is there a "metabolic compensation", so that once one metabolic pathway is denied an other one is activated ?

We can represent a system ready to deal with ionizing radiation as a system with two main machineries: one devoted to *repair* of DNA demages and an other one devoted to *protection* against oxidative damage. If the system addicted to repairing is compromised, could it be possible that the part dedicated to the protection of DNA is enriched?

In *Microarray analysis of the hyperthermophilic archaeon Pyrococcus furiosus exposed to gamma irradiation* by Jocelyne et al. [4], P. furiosus cultures were exposed to 2,500 Gy of gamma radiation and the results of microarray analysis show a very large increase in mRNA of ferritin/Dps-like proteins following gamma irradiation.

Ferritin, an iron storage protein, is the primary iron storage mechanism and is critical to iron homeostasis. Ferritin makes iron available for critical cellular processes while protecting lipids, DNA, and proteins from the potentially toxic effects of iron [5]. Iron is an essential trace element for all living organisms, as it is vitally involved in a variety of cellular functions, such as oxygen transport and storage, energy production, cell cycle, and DNA synthesis. However, the excess of iron is toxic in the cell, and free Fe^{2+} donates electrons to produce hydroxyl radicals $OH^{\cdot}$ via the Fenton reaction

$$Fe^{2+} + H_2O_2 \longrightarrow Fe^{3+} + OH^- + OH^{\cdot} \tag{1}$$

Free hydroxyl radicals can catalyze the oxidative damage of biomolecules; nature's answer to this dual problem of availability and toxicity is the ferritin and ferritin-like proteins which are major non-heme iron storage proteins found in the three domains of life. In particular, the archaea *Sulfolobus solfataricus* and *Pyrococcus furiosus* and bacteria *Bacteroides fragilis* have been shown to possess DPS-like (DPSL) proteins that are members of 12-subunit ferritins, which might be an intermediate of the ferritin family evolution [6]. Therefore, the very large increase in mRNA of ferritin and ferritin-like proteins encoding genes indicates its critical role in removing free iron from solution and thereby limiting the production of hydroxyl radicals by Fenton

chemistry.

In this work, firstly we check if there is amplification of ferritin genes for protection of DNA in COM1 with respect to the reference genome; secondly we search for known motifs and de novo motifs associated to radioresistance in order to understand if they are overrepresented in COM1. The genome DNA sequences of COM1 and Reference are in GenBank (accession number CP003685 and NC_003413).

1.1 Motifs in molecular biology

In molecular biology, motif is a region of protein or DNA that has a specific structure [7]. The french word *motif* means *pattern*, that indicates repetition, but also *emblem* suggesting a means of identifying the group to which something belongs. There are two types of motifs: motifs in DNA and in proteins both in their nucleotides or amino acids sequences but also in their three dimensional structure. In the following subsections, the most common motifs will be explored before deeping in motifs associated to radiation resistance.

1.2 DNA sequence motif

The DNA is a linear chain of four **nucleotides** (adenine (A), thymine (T), guanine (G), and cytosine (C)) that are arranged in defined ways which we refer to as the **sequence** of the macromolecule (see **Appendix A** for insights on DNA structure). Within the overall sequence there can be sub-sequences, which, if they repeat in the genome, represent patterns, that are the so called DNA sequence *motifs*. They are assumed to be related to biological function of the macromolecule and often they indicate sequence-specific binding sites for proteins such as transcription factors (TFs).

Encoded in the structure of DNA is the information that programs all the cell's activities through the production of proteins. The central dogma of molecular biology states that genes specify the sequence of mRNA molecules, which in turn specify the sequence of proteins (DNA $\rightarrow$ RNA $\rightarrow$ Protein).

In all the three domains of life, there are two main stages from DNA sequence of a gene to protein: *transcription*, that is the process of making a strand of RNA that is complementary to a single strand DNA, and *translation*, that is the process by which a protein is synthesized from the information contained in a molecule of messenger RNA (mRNA). In the transcription

process, proteins called *transcription factors* play a central role and we can group them into different types according to the transcription step in which they act:

- the process of transcription starts with *initiation* and in archaea this step is governed by **TATA-binding protein (TBP)**, **Archaeal transcription factor B (TFB)**, and **Archaeal transcription factor E (TFE)**. These constitute the **basal transcription factors** and transcription initiation is regulated by **DNA elements** that are recognized by basal transcription factors that recruit the enzyme RNA Polymerase (RNAP) that will copy DNA into RNA. Indeed, the RNA Polymerase can not start initiation of transcription by its own but it requires the help of the mentioned transcription factors which allow the specific binding of the RNAP with the promoter. There are four DNA elements currently known to regulate archaeal transcription initiation: the **TATA-box** located approximately 25 bp upstream of the site of transcription initiation; the **TFB recognition element (BRE)** located immediately upstream of the TATA box; the **initiator element (INR)** located within the initially transcribed region, and the **promoter proximal element (PPE)** located between the TATA box and the site of transcription initiation (they are sketched in figure 1). Of these four, only the TATA box and the BRE are required for transcription initiation, although alterations to all four elements can influence the total output of a promoter. The INR is not a required DNA element for transcription initiation; however, it is a regulatory element that can increase the strength of the promoter (that means it can increase the rate of transcription) in a TATA- and BRE-dependent manner. PPEs, centered approximately 10 bps upstream of the site of initiation, have been shown to increase transcription output through recruitment of TFB.

  After initiation, there are other two transcription steps: *elongation*, in which the RNA polymerase transcribes the gene and release the RNA product, and *termination*, that is the process of stopping the release of RNA product.

- As transcription transitions from initiation to elongation, RNAP undergoes a conformational change accompanied by the replacement of initiation factors with elongation factors. Very few transcription elongation factors have been bioinformatically identified within archaeal genomes, and it is probable that archaeon-specific factors await discovery. Transcription elongation factors have various roles, including

increasing processivity and fidelity of RNAP and/or increasing genome stability. Only two archaeal elongation factors have been experimentally studied: the **elongation factor Spt5**, often with a conserved binding partner Spt4, and **transcription factor S (TFS)**.

- ◇ Finally, transcription termination factors have been characterized only in Bacteria and Eukarya. Bioinformatic analyses reveal some potential targets that remain to be more fully evaluated, but there are no easily identified homologues of known eukaryotic or bacterial termination factors [8].

Thus one distinct feature of transcription factors is that they have DNA-binding *domains* that give them the ability to bind to specific sequences of DNA [9]. Binding might occur only if the fit is exactly right, or sometimes if the fit is about right. We can use sequence logos to understand where the transcription factor's behavior falls along this spectrum (see subsection 1.7).

As said, some transcription factors bind to a DNA promoter sequence near the transcription start site and help form the transcription initiation complex. In eukaryotic cells, other transcription factors bind to regulatory sequences, such as *enhancers* and *silencers* that can respectively stimulate and repress transcription of the related gene and for this reason they are called also *transcriptional activators* or *transcriptional repressors*.

Thus, we can summarize the "anatomy" of transcriptional regulation with figure 1:

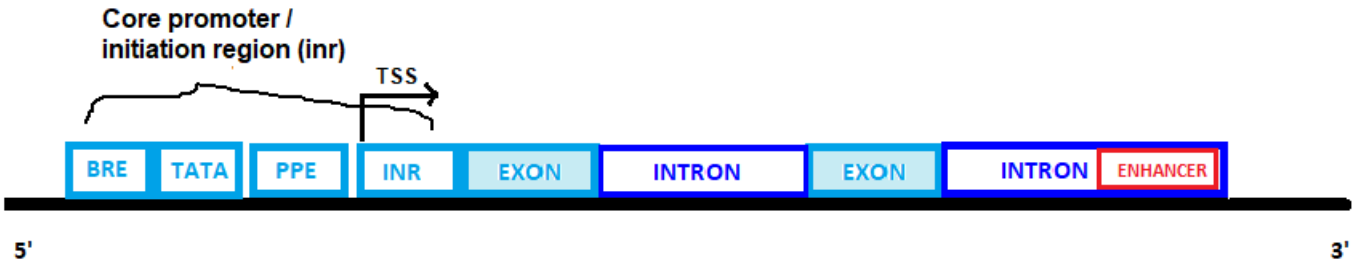

Figure 1: Sketch of "anatomy" of transcriptional regulation in Archaea.

A *promoter* is a sequence of DNA to which proteins bind that initiate gene transcription of a single RNA from the DNA downstream of it. Pro-

moters are located upstream on the DNA (towards the 5' region of the sense strand). *Core promoter* is the minimal portion of the promoter required to properly initiate transcription and includes the transcription start site (TSS) and elements directly upstream, that are general transcription factor binding sites, e.g. TATA-box and TFB recognition element but also many other elements/motifs may be present. There is no such thing as a set of "universal elements" found in every core promoter.

Eukaryotic genes are usually made of exons, which are regions coding for proteins, and introns, that are non-coding regions that are eliminated by splicing before translation. Regulatory elements, such as enhancers and silencers, may be located up to 50 kilobases upstream or downstream from the cap site (the site on a DNA template where transcription begins) or within an intron [10]. However, we have to point out that archaea and bacteria are typically characterized by *intron-less genes* [12] and that enhancer-like sites are usually found in the upstream promoter region [11]. But, there are some studies that show the presence of DNA elements that are binding sites for regulatory proteins and function at large distances from promoter elements in bacteria [13] and a few archaeal genes that are interrupted by microintrons [50]. Moreover, Archaea possess a eukaryotic-type basal transcription apparatus (as we can notice from the description of transcription steps given above) that is regulated by bacteria-like transcription regulators [48]. However, it is not yet well known how regulation of archaeal transcription is achieved [49], so we can consider the sketch shown in figure 1 as valid also for archaea. In general, it is for these reasons (the lack of knowledges about the archaea domain and the strong similarities to eukaryotic-type transcription machinery and to bacteria transcription regulators) that, even though the organism of our investigation is the Pyrococcus Furiosus belonging to archaea domain, we consider in this work the characteristics of eukaryotic and bacteria organisms as probably valid also for archaea organisms and, therefore, some aspects of eukaryotic and bacteria organisms are described as well.

Our discussion leads to the importance of regulation of transcription, that is the most common form of gene control. The action of transcription factors allows for *unique* expression of each gene. These kind of DNA sequences, recognized by activators or repressors, represent motifs that can be called "regulatory motifs" since they are used to control the expression of genes, dictating under which conditions a gene will be turned on or off [15].

Thus, DNA sequences recognized by TFs can somehow be considered as gene

"fingerprints" since transcription factors have evolved different ways to contact the DNA double helix, and these are reflected in different DNA sequence motifs. One such example in eukaryotic organisms is the Mbp1 transcription factor involved in the timing of events such as DNA replication during cell division and recognizes the motif ACGCGT, also known as MCB elements that are found in the promoter of most DNA synthesis genes. The DNA-binding domains of other factors are made of two identical parts (and hence called homodimers), contacting each other and each contacting the DNA helix. The two parts recognize identical sequences, but on opposite strands. One such example is the Gal4 factor involved in galactose metabolism since it is a positive regulator for the gene expression of the galactose-induced genes such as GAL1, GAL2, GAL7, GAL10, and MEL1 which code for the enzymes used to convert galactose to glucose. It recognizes CGGNNNNNNNNNNCCG, namely CGG on one strand spaced by 11 nucleotides (one full turn of the double helix) from its reverse complement, CCG.

Therefore, there are transcription factors (general or basal TFs needed for initiation of transcription and specific TFs for enhancers and repressors) that recognize specific DNA sequences and that are "specific for specific genes". Thus, these transcription factor binding sites can be considered as gene "fingerprints".

In the past, binding sites were typically determined through DNase footprinting, and gel-shift or reporter construct assays, whereas binding affinities to artificial sequences were explored using SELEX. Nowadays, computational methods (one of which we will study in section 2) are generating a flood of putative regulatory sequence motifs by searching for overrepresented (and/or conserved) DNA patterns of functionally related genes (for example, genes with similar expression patterns or similar functional annotation) [14].

Once got a conserved nucleic acid pattern, how do you understand if it has a relevant biological meaning and, in other words, it is really a gene "fingerprint"?

We can consider two main characteristics of DNA sequence motifs: one is the *position* in the gene sequence and the other one is the *number of nucleotides* of the motif. As regards the first one, previously we have described the anatomy of transcriptional regulation and so if we find a motif upstream of several genes with similar expression patterns or similar functional annotation, we can say that this motif represents a general transcription binding site; if it is located at the bottom of the gene sequence or in other positions,

it can represent a regulatory sequence (enhancer or silencer). The length of binding sites, instead, range from 5 nt to 30 nt, in both eukaryotes (see figure 2 left, 454 curated transcription factor motifs) and prokaryotes (figure 2 right, 79 motifs).

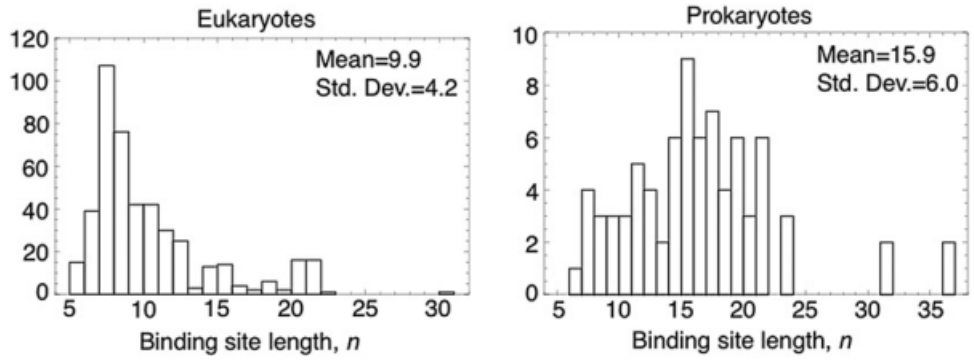

Figure 2: Typical length of binding sites for Eukaryotes and Prokaryotes (image from [37]).

Thus, these are at least two characteristics to take into account in order to understand if our DNA pattern has a biological meaning and so it can represent a transcription factor binding site.

1.3 DNA structural motif

In a chain-like biological molecule, such as a nucleic acid, we have to consider also the three-dimensional structure of the molecule because motif can be composed also of structural components.

First of all, whether or not transcription factors bind depends on **the shape of the DNA**, which in turn depends on the sequence.

However, we can also define a motif as **a common three-dimensional structure which is thought to have biological significance**. In order to understand what this means, we can describe the following interesting structural DNA motifs:

- ◇ *cruciform DNA*: a secondary structure on a helical double-stranded DNA molecule that comprises a four-way junction and two closed hairpin-shaped points. Cruciform DNA is found in both prokaryotes and eukaryotes and has a role in DNA transcription and DNA replication,

double strand repair, DNA translocation and recombination. They also serve a function in epigenetic regulation along with biological implications such as DNA supercoiling, double strand breaks, and targets for cruciform-binding proteins [21].

- ◇ *G-quadruplex* (G4): four guanine bases can associate through Hoogsteen hydrogen bonding to form a square planar structure called a guanine tetrad (G-tetrad or G-quartet), and two or more guanine tetrads (from G-tracts, continuous runs of guanine) can stack on top of each other to form a G-quadruplex. The placement and bonding to form G-quadruplexes is not random and serve very unusual functional purposes. G-quadruplex structures can be computationally predicted from DNA or RNA sequence motifs, but their actual structures can be quite varied within and between the motifs, which can number over 100,000 per genome. G4 DNA is thought to play an important role in transcriptional and translational regulation of genes, DNA replication, genome stability, and oncogene expression in eukaryotic genomes. In other organisms, including archaea, several bacterial pathogens and some plant species, the biological roles of G4 DNA and G4 RNA are starting to be explored [21]. For example, a genome-wide study predicts that promoter-G4 DNA motifs regulate selective functions in bacteria. In particular, it is shown that genes imparting resistance to radiation have enriched promoter-PG4 motifs in radioresistant bacteria and so that radioresistance in these organisms such as *D. Radiodurans* involves G4 DNA-mediated regulation [31].

1.4 Protein sequence motifs

A protein **sequence** motif, or pattern, can be defined as a region (a subsequence) of amino acid sequence that has a specific structure and that is important for protein function. Protein sequence motifs are signatures of protein families and can often be used as tools for the prediction of protein function and as a base of protein classification [19]. Therefore a protein sequence motif is found in similar proteins and change of a motif changes the corresponding biological function.

One of the first sequence motifs reported were the so called *Walker motifs*, which later were shown to correspond to ATP- or GTP- binding and therefore are characteristic to a very broad range of proteins. For example, *Walker motif A* has the pattern **GXXXXGK(T/S)**, where G, K, T and S are glycine, lysine, threonine and serine residues, X any other amino acid

[41].

1.5 Protein structural motifs

As in the case of DNA, protein motifs can indicate also regions of **protein three-dimensional structure** shared among different proteins. They are recognizable regions of protein structure that may (or may not) be defined by a unique chemical or biological function [20].
One of the most common structural motifs in proteins is *helix-turn-helix* (HTH) that is two α helices joined by a short strand of amino acids and found in many proteins that regulate gene expression and is a common motif in basal and specific transcription factors (proteins that bind DNA as explained before) in the three kingdoms of life [22]. Moreover, in the previous subsection we have mentioned the Walker motifs that are protein sequence motifs and they are known to have highly conserved three-dimensional structures [42]. An examination of the immediate neighborhood of the Walker sequence indicates that this region is preceded by a β-strand and followed by an α-helix, resulting in the motif β-W-α, an invariant feature amongst nucleotide-binding proteins [43]. Thus Walker motif is a fingerprint sequence that characterizes ATP- or GTP-binding proteins and is part of the nucleotide-binding motif .

The helix-turn-helix is also refered to as *domain*. A **motif** is similar 3-D structure conserved among different proteins that serves a similar function. An example is the helix-turn-helix motif. This is a structure that is seen in unrelated proteins that bind DNA, so the presence of a helix-turn-helix motif is an indication of a protein's function. **Domains**, on the other hand, are regions of a protein that has a specific function and can (usually) function independently of the rest of the protein. A protein can have multiple domains. It can have a DNA binding domain located towards the N-terminus of the protein, and a catalytic domain that is located closer to the C-terminus. Theoretically you can separate the domains from each other and the DNA binding domain will still bind DNA and the catalytic domain will still perform catalysis. There is some overlap with the definitions of domain and motif. Some motifs are also considered domains, and vice versa [30] as the case of helix-turn-helix motif/domain.

1.6 Motifs associated to radioresistance

We have already discussed that motifs (in particular DNA sequence motifs) can be considered as "fingerprints" of genes or, in case of protein motifs, can

be used for protein classification. Therefore, it is natural to ask if there are motifs related to radioresistance. We have already mentioned the case of enriched promoter-PG4 motifs in radioresistant bacteria. Below we examine some motifs associated to the radioresistance of an organism and that we will consider in our investigation:

- **TATA-box**: as said in section 1.2, it is located approximately 25 bp upstream of the site of transcription initiation and generally contains the consensus sequence **5'-TATA(A/T)A(A/T)-3'**. In yeast, for example, one study found that various Saccharomyces genomes had the consensus sequence 5'-TATA(A/T)A(A/T)(A/G)-3', but only about 20% of yeast genes contained even the TATA sequence. Similarly, in humans only 24% of genes have promoter regions containing the TATA-box.

  Genes containing the TATA-box tend to be involved in stress-responses and certain types of metabolism and are more highly regulated when compared to TATA-less genes. Generally, TATA-containing genes are not involved in essential cellular functions such as cell growth, DNA replication, transcription, and translation because of their highly regulated nature [26]. A variety of studies have underscored, instead, that the TATA box is the major basal promoter element throughout the Archaea, and that it is important for transcription initiation at stable RNA (rRNA and tRNA) genes as well as at protein-encoding genes [27].

- **The ferritin-like di-iron domain**: this is a less well known motif but it could be associated to radioresistance. It is a structural protein motif made of a four-helix bundle sorrounding a non-heme, non-sulphur, oxo-bridged diiron site. The diiron site is contained within a twisted left-handed four-helix-bundle constituted of two anti-parallel helix pairs connected through a left-handed crossover connection [28] (see figure 3).

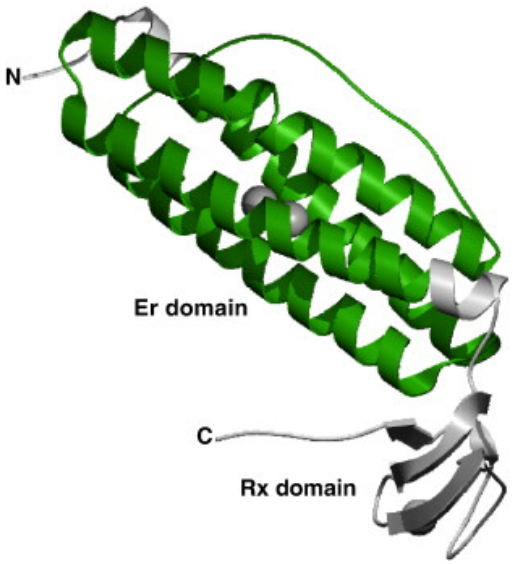

Figure 3: Crystal structure of the rubrerythrin of *Desulfovibrio vulgaris*. The ferritin-like di-iron motif (the one indicated as "Er domain") is shown here with the helices coloured green and the iron atoms represented as grey spheres. Image from [44].

Proteins known to contain a ferritin-like diiron domain are: Ferritin (Ftn), an eukaryotic intracellular protein that stores iron in a soluble, nontoxic, readily available form; Bacterioferritin (Bfr), a prokaryotic protein which may perform functions in iron detoxification and storage (as pointed out in section 1); Rubrerythrin (Rr), a non-heme protein isolated from anaerobic sulphate- reducing bacteria; Nigerythrin (Nr), a prokaryotic protein of unknown function. In the paper by Jocelyne et al. [4], genes encoding proteins belonging to Ferritin family and which are linked by the presence of a common ferritin-like di-iron motif displayed increase in mRNA level following irradiation, suggesting their critical role in removing free iron from solution and thereby limiting the production of hydroxyl radicals by Fenton chemistry, as explained in section 1. This motif is also found in DPS-like (DPSL) proteins that, as mentioned previously, might be an intermediate of the ferritin family evolution.

1.7 Sequence logo for motif representation

In bioinformatics, a sequence logo is a graphical representation, in a visually informative manner, of the sequence conservation of nucleotides (in a strand

of DNA/RNA) or amino acids (in protein sequences). A sequence logo is created from a collection of aligned sequences and depicts the consensus sequence and diversity of the sequences. They are frequently used to represent motifs such as protein-binding sites in DNA or functional units in proteins. [25]
Let us consider the case of TATA box, that is the DNA sequence motif where transcription factor TATA binding protein (TBP) binds as explained in section 1.1. The TATA binding protein can potentially bind to wide variety of sequences provided there is some degree of sequence conservation. Figure 4 shows an example of alignment of 15 TATA sequence variants, all of which accomodates TBP binding, and its sequence logo.

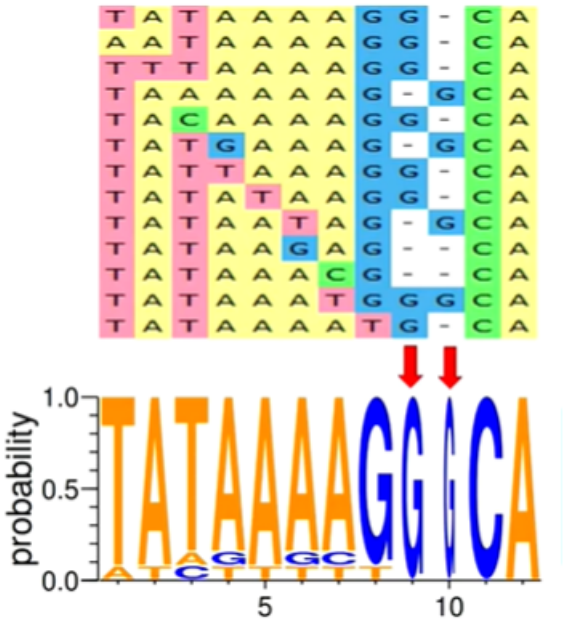

Figure 4: alignment of 15 TATA sequence variants and its corresponding logo. Image from [45].

In a logo plot, the vertical axis represents the probability (that ranges from 0 to 1) and on the horizontal axis there is the position of the character in the sequence. Therefore, the *height* of a character reflects the probability of finding the character in that given column, while the *width* of the character represents the presence of gaps in the alignment column (narrower the character is, higher the number of gaps in the relative column is).

This is the general philosophy. However it is often preferable to consider

Information Content (IC) measured in *bits* as the unit on the y axis of a sequence logo plot. The information content of position i for amino acids is given by:

$$IC_i = \log_2(20) - (H_i + e_n) \tag{2}$$

and for nucleic acids:

$$IC_i = \log_2(4) - (H_i + e_n) \tag{3}$$

where, for example in equation (3), $\log_2(4) = 2$ bits means that there are 4 possibilities per position and it takes 2 bits to encode all 4 possibilities, whereas H_i is the Shannon entropy that represents the *uncertainty* of position i:

$$H_i = -\sum_{b=1}^{t} f_{b,i} \times \log_2 f_{b,i} \tag{4}$$

Where $f_{b,i}$ is the relative frequency of base or amino acid b at position i, b ranges across all the various states (e.g. each nucleotide), and e_n is the small-sample correction for an alignment of n letters. Thus, we can "read" the equation (3) for calculating the IC_i for a specific position i, in case of nucleic acids, in this way:

$$IC_i = 2 - \text{the uncertainty of position } i \tag{5}$$

The height of each nucleotide b is the resulting IC multiplied by the relative frequency of that nucleotide b at the position i:

$$height = f_{b,i} \times IC_i \tag{6}$$

The more frequently the nucleotide b is found at that position, the lower the uncertainty and the higher the height. For example, if you have a position with 70% G and 30% C, ignoring the small sample correction e_n:

$$\text{IC} = 2 - [(-0.7 \cdot \log_2 0.7) + (-0.3 \cdot \log_2 0.3)] = 1.14 \text{ bits} \tag{7}$$

where 1.14 are bits of total height. Then, G gets a height of $1.14 \cdot 0.7 = 0.79$ bits, whereas C gets $1.14 \cdot 0.3 = 0.34$ bits.

Plotting IC rather than nucleotide frequency makes it a lot easier to interpret a sequence logo. In this way, we are not looking at frequencies for each base directly (in which case the stacks of letters would always reach the same height as it happens in figure 4). When looking at a logo, we are not really interested in the bases where there is equal probability of each residue. But in a probability rendering, these bases are presented to the

viewer with the same visual emphasis as bases where there is an (interesting, potentially biologically relevant) overabundance of one or two residues. In the entropy rendering, the bases positioned where there is equal probability of each residue are dismissed, so that the viewer's eye is directed to the residues that contribute most to the motif's information content. A site with no conservation will have IC of 0 bits (equal chance of getting A, C, G, or T, so no information content), whereas a completely conserved site has an IC of 2 bits. Figure 5 shows the sequence logo of the same 15 TATA sequence variants reported in figure 4 but with y-axis labeled as bits:

Figure 5: Sequence logo of 15 TATA sequence variants reported in figure 4 but with bits as y-axis height.

The sequence logo in figure 5 is created using the online software *WebLogo* available here: `https://weblogo.berkeley.edu/logo.cgi`. The option *Small Sample Correction* has been applied because when there are only a few sample sequences a straightforward calculation will tend to overestimate the entropy. To compensate, this option will subtract an approximation of this bias from the total entropy (it is the e_n in equations (2) and (3)). This small sample correction depends only on the number of symbol types (4 for RNA/DNA, 20 for protein) and the total amount of data in each column, which may differ from one column to another, since some columns will contain more gaps, and less data.

Thus in both cases (figure 4 and 5), the figure shows the same underlying data, but the difference is about presentation and what is trying to be communicated.

2 Software and methods

In order to detect gene amplification, a simple Python code (without extra libraries) that counts how many times a gene sequence is found in a genome

sequence has been written and reported in **Appendix C**.

As regards the motif analysis, the procedure is reported below.

2.1 De novo motif discovery

We take our set of related genes, that means they have similar expression patterns or similar functional annotation. We would like to find a short common pattern of nucleotides upstream of the transcription start sites of these genes, indicating a common transcription factor binding site responsible for their coordinate regulation or a common pattern of nucleotides downstream of the transcription start sites that could indicate enhancers or silencers, so transcriptional activators or transcriptional repressors binding sites. Given this set of sequences, which we have good reason to believe share a common binding motif, how do we search for new instances of a motif in this sea of DNA? [14] How do we extract these patterns from a set of sequences? De novo sequence motif finding is if you do not know anything about existing transcription factor binding motifs and there are three distinct approaches: enumeration, deterministic optimization and probabilistic optimization.

Enumeration approach searches for consensus sequences. Motifs are predicted based on the enumeration of words and computing word similarities so this approach is sometimes called the word enumeration approach to solve Motif Problem with motif length (l) and a maximum number of mismatches (d). The algorithms based on the word enumeration approach exhaustively search the whole search space to determine which ones appear with possible substitutions and therefore it typically locates the global optimum. However, this also means that they are exponential-time algorithms that require a long time to detect the larger l and inefficient for handling dozens of sequences, so they are only suitable for short motifs . Moreover, these algorithms require many parameters determined by the users such as motif length, the number of mismatches allowed, and a minimum number of sequences that the motif has to appear [33].

Thus deterministic and probabilistic optimization approach are very popular to search motifs in the genome background. The first one is Expectation Maximization (EM) algorithm while the latter is known as the Gibbs Sampling. In this work, we will use the software MEME suite 5.4.1 where MEME stands for "Multiple Expectation Maximizations for Motif Elicitation" because the basic motif discovery algorithm used is the Expectation Maximization (EM) algorithm on many input sequences.

The MEME Suite allows to discover novel motifs in collections of unaligned nucleotide or protein sequences, and to perform a wide variety of other motif-based analyses. In addition to motif discovery, in fact, the MEME Suite provides a tool for comparing motifs to known motifs that is TomTom, that finds motifs that are similar to a given DNA or protein sequence motif by searching a database of known motifs. However, for the protein sequence motifs there are available only databases for Eukaryotes ("Eukaryotic Linear Motifs (ELM 2018) Motifs" and "Eukaryotic Linear Motifs (ELM 2017) Motifs"), so BLAST (Basic Local Alignment Search Tool) will be also used to compare protein sequences to sequence databases. In general, the program finds regions of similarity between biological sequences and calculates the statistical significance. In our case, the parameters are adjusted to search for a short input sequence, that is the protein sequence motif, in order to find it in other archaea or bacteria organisms and hypothesize its biological meaning.

2.2 What is the expectation maximization (EM) algorithm?

Probabilistic models, such as hidden Markov models or Bayesian networks, are commonly used to model biological data. Much of the popularity can be attributed to the existence of efficient and robust procedures for learning parameters from observations. Often, however, the only data available for training a probabilistic model are incomplete. The expectation maximization algorithm enables parameter estimation in probabilistic models with incomplete data.

As an example, consider a simple coin-flipping experiment in which we are given a pair of coins A and B of unknown biases, θ_A and θ_B, respectively (that is, on any given flip, coin A will land on heads with probability θ_A and tails with probability $1 - \theta_A$ and similarly for coin B). Our goal is to estimate $\theta = (\theta_A, \theta_B)$ by repeating the following procedure five times: randomly choose one of the two coins (with equal probability), and perform ten independent coin tosses with the selected coin. Thus, the entire procedure involves a total of 50 coin tosses (Figure 6 **a**). During our experiment, suppose that we keep track of two vectors $x = (x_1, x_2, ..., x_5)$ and $\pi = (\pi_1, \pi_2, ..., \pi_5)$, where $x_i \in 0, 1, ..., 10$ is the number of heads observed during the ith set of tosses, and $\pi_i \in A, B$ is the identity of the coin used during the ith set of tosses. Parameter estimation in this setting is known as *the complete data case* since the values of all relevant random variables in our model (that is, the result

of each coin flip and the type of coin used for each flip) are known. Here, a simple way to estimate θ_A and θ_B is to return the observed proportions of heads for each coin:

$$\hat{\theta_A} = \frac{\text{\# of heads using coin A}}{\text{total \# of flips using coin A}} \tag{8}$$

and

$$\hat{\theta_B} = \frac{\text{\# of heads using coin B}}{\text{total \# of flips using coin B}} \tag{9}$$

This intuitive guess is, in fact, known in the statistical literature as maximum likelihood estimation (roughly speaking, the maximum likelihood method assesses the quality of a statistical model based on the probability it assigns to the observed data). If $\log P(x, \pi; \theta)$ is the logarithm of the joint probability (or log-likelihood) of obtaining any particular vector of observed head counts x and coin types π, then the formulas in (8) and (9) solve for the parameters $\hat{\theta} = (\hat{\theta_A}, \hat{\theta_B})$ that maximize $\log P(x, \pi; \theta)$.

Now consider a more challenging variant of the parameter estimation problem in which we are given the recorded head counts x but not the identities π of the coins used for each set of tosses. We refer to π as hidden variables or latent factors. Parameter estimation in this new setting is known as *the incomplete data case*. This time, computing proportions of heads for each coin is no longer possible, because we don't know the coin used for each set of tosses. However, if we had some way of completing the data (in our case, guessing correctly which coin was used in each of the five sets), then we could reduce parameter estimation for this problem with incomplete data to maximum likelihood estimation with complete data.

One iterative scheme for obtaining completions could work as follows: starting from some initial parameters, $\hat{\theta}^{(t)} = (\hat{\theta_A^{(t)}}, \hat{\theta_B^{(t)}})$, determine for each of the five sets whether coin A or coin B was more likely to have generated the observed flips (using the current parameter estimates). Then, assume these completions (that is, guessed coin assignments) to be correct, and apply the regular maximum likelihood estimation procedure to get $\hat{\theta}^{(t+1)}$. Finally, repeat these two steps untill convergence (figure 6 **b**). As the estimated model improves, so too will the quality of the resulting completions.

In summary, the expectation maximization algorithm alternates between the steps of guessing a probability distribution over completions of missing data given the current model (known as the *E-step*) and then re-estimating the model parameters using these completions (known as the *M-step*). The name 'E-step' comes from the fact that one does not usually need to form the probability distribution over completions explicitly, but rather need only compute

'expected' sufficient statistics over these completions. Similarly, the name 'M-step' comes from the fact that model reestimation can be thought of as 'maximization' of the expected log-likelihood of the data.

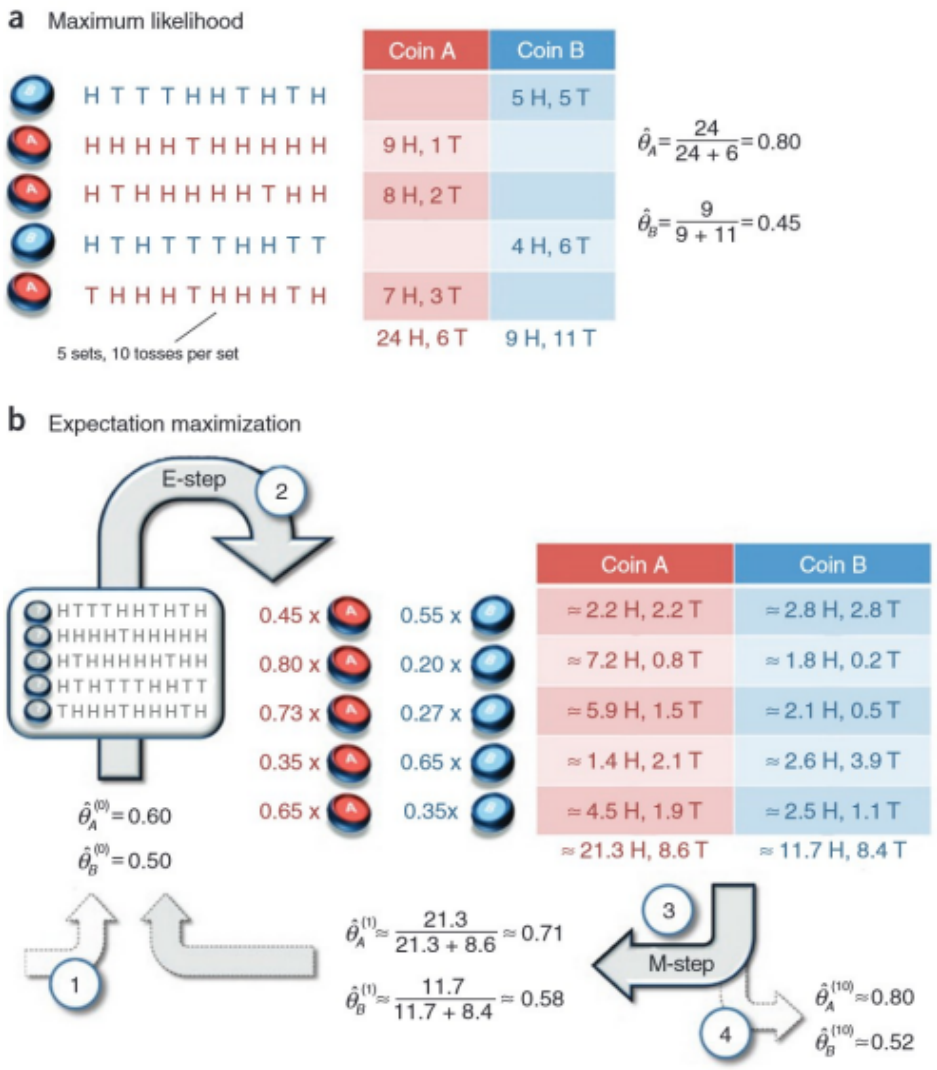

Figure 6: (**a**) Maximum likelihood estimation. (**b**) Expectation maximization procedure. Image from [34].

Many probabilistic models in computational biology include latent variables. In motif finding, we are given a set of unaligned DNA sequences and asked to identify a pattern of length W that is present (though possibly with minor variations) in every sequence from the set. To apply the expectation maximization algorithm, we model the instance of the motif in each sequence

as having each letter sampled independently from a position-specific distribution over letters, and the remaining letters in each sequence as coming from some fixed background distribution.

The observed data x consist of the letters of sequences, the unobserved latent factors π include the starting position of the motif in each sequence and the parameters θ describe the position-specific letter frequencies for the motif. Here, the expectation maximization algorithm involves computing the probability distribution over motif start positions for each sequence (E-step) and updating the motif letter frequencies based on the expected letter counts for each position in the motif (M-step). Let us deepen the EM algorithm for motif finding.

2.3 EM algorithm for de novo motif discovery

In motif finding, the probabilistic model has the following objects:

- **seq**: a number of input sequences that you want to search for a motif;
- $\boldsymbol{\theta_0}$: non-motif (genome background) probability;
- $\boldsymbol{\theta}$: motif position probability matrix (PPM) or motif position weight matrix (PWM). It is what we really want to find from input data (see **Appendix D** for insights into the definitions of PPM and PWM);
- $\boldsymbol{\pi}$: location of where the motif occurs.

The problem we deal with is: given the input sequences **seq** (that is the letters of the sequences x in the notation of the previous subsection), given the non-motif (genome background) $\boldsymbol{\theta_0}$, we are trying to find both the motif matrix $\boldsymbol{\theta}$ and also where it appears in the input sequence $\boldsymbol{\pi}$: $\boldsymbol{P(\theta, \pi | seq, \theta_0)}$.

Unfortunately, at the beginning we do not know $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, therefore, EM and Gibbs sampler approach just initialize the algorithm with some random motif matrix and then they iteratively update. The EM procedure is the following: we have the input sequences **seq** and so we also have the genome background $\boldsymbol{\theta_0}$. If we know the motif matrix $\boldsymbol{\theta}$ (so the motif sequence), we can figure out where the motif appears in the input sequence because we are able to look at the sequence and see where there is a hit. By considering the motif matrix, we just scan all the k-mers and you consider the high scoring ones. Thus in this case we obtain: $\boldsymbol{\pi}$ by $\boldsymbol{P(\pi | \theta, seq, \theta_0)}$.

On the other hand, if we know $\boldsymbol{\pi}$, that is where the motif hits are in the data,

we can construct a probability matrix . This is the Gibbs Sampler approach that finds: $\boldsymbol{\theta}$ by $\boldsymbol{P(\theta|\pi, seq, \theta_0)}$.

As mentioned in the previous subsection, in EM algorithm there is an *expectation step* and a *maximization step*:

- ◇ *Expectation step* (E-step): if we already have the motif matrix $\boldsymbol{\theta}$, we try to estimate where the starting position $\boldsymbol{\pi}$ is in the sequence where this motif occurs. Similarly to the example of coin tosses: in that case, we guessed the values of the probability that coin A will land on heads, θ_A, and the probability that coin B will land on heads, θ_B, and we estimated the identities of the coins for each of the five set of flips on the basis of the knowledge of these completions (θ_A and θ_B) and the recorded head counts x that we already knew. Now it is the same: on the basis of $\boldsymbol{\theta}$ that we guess and on the basis of the letters of sequences **seq** that we already know, we estimate $\boldsymbol{\pi}$, that is the location of where the motif occurs. For example, we guess that the motif is only 5 nucleotides long and that its position probability matrix $\boldsymbol{\theta}$ of this 5-mer motif is:

Pos	**A**	**C**	**G**	**T**
1	0.7	0.1	0.01	0.2
2	0.01	0.01	0.8	0.1
3	0.32	0.02	0.3	0.18
4	0.03	0.42	0.1	0.47
5	0.2	0.5	0.1	0.2

Given the input sequence TTGACGACTGCACGT, we use the motif matrix to scan all the 5-mers:

TTGACGACTGCACGT
TTGAC LR_1
TGACG LR_2
GACGA LR_3
ACGAC LR_4
...

At each 5-mer in the input sequence, we calculate the Likelihood Ratio:

$$LR_1 = Likelihood\ Ratio = \frac{P(TTGAC|\theta)}{P(TTGAC|\theta_0)} \tag{10}$$

where $P(TTGAC|\theta)$ is the probability of the sequence TTGAC given the position probability matrix $\boldsymbol{\theta}$ and that, in this case, is:

$$P(TTGAC|\theta) = 0.2 \times 0.1 \times 0.3 \times 0.03 \times 0.5 \tag{11}$$

while $P(TTGAC|\theta_0)$ is the probability of seeing TTGAC in the genome background. For example, 30% of the basis in the human genome is A, 30% is T, 20% is C and 20% is G, the probability to have TTGAC in the genome background is:

$$\begin{aligned} P(TTGAC|\theta_0) = P_0T \times P_0T \times P_0G \times P_0A \times P_0C = \\ = 0.3 \times 0.3 \times 0.2 \times 0.3 \times 0.2 \end{aligned} \tag{12}$$

Therefore, we use the probability matrix $\boldsymbol{\theta}$ as well as the genome background $\boldsymbol{\theta_0}$ to calculate all the Likelihood ratios for every 5-mer in the input sequence.

◇ *Maximization step* (M-step): the next step is to find the motif probability matrix $\boldsymbol{\theta}$ parameters that maximize the likelihood of seeing the k-mers represented by $\boldsymbol{\theta}$ in the input sequences. Thus now for each k-mer we have a *score* (the likelihood ratio) that is used to weight each k-mer.

TTGACGACTGCACGT

$0.8 \times$ TTGAC

$0.2 \times$ TGACG

$0.6 \times$ GACGA

$0.5 \times$ ACGAC

...

For each position, the ratio between the sum of the nucleotides weighted by their respective score and the sum of all the scores is computed.

$$\begin{aligned} T_1\% &= \frac{0.8 + 0.2 + ...}{0.8 + 0.2 + 0.6 + 0.5 + ...} \\ G_2\% &= \frac{0.2 + ...}{0.8 + 0.2 + 0.6 + 0.5 + ...} \\ C_5\% &= \frac{0.8 + 0.5 + ...}{0.8 + 0.2 + 0.6 + 0.5 + ...} \end{aligned} \tag{13}$$

At the end, we have a new probability matrix that we use to score all the k-mers of the input sequence again. Then after having calculated the likelihood ratios, the Maximization step is computed again. This is done iteratively and "magically", at the end, it will converge at a motif represented by $\boldsymbol{\theta}$. As the estimated model improves, so too will the quality of the resulting completions $\boldsymbol{\pi}$.

Why does the algorithm converge? It can be understood by considering the starting PPM of the example already explained but, if we start, for example for simplicity, from 5-mers which all have score=25% , at the beginning there will be no preferences: all 5-mers will have the same score. If then we align them together, there will be 5-mers that occurs more times than others and so the PPM is updated towards the correct enrichment that we want to find. In the end, the over-represented 5-mer will arise. Nothing more, nothing less, certainly no magic involved.

This algorithm for de novo motif discovery is considered a deterministic approach because if we already initialize the motif matrix and we run the EM steps multiple times, the result will be always the same. MEME is one popular implementation of the expectation maximization algorithm and performs a single iteration for each k-mer in the target sequences, selects the best motif from this set and then iterates only that one to convergence, avoiding local maxima (so it runs for 10-30 seconds). This partially enumerative nature of MEME provides some assurance that the algorithm is unlikely to get stuck in a poor local maximum. Additional motifs present in the set of target sequences can be found by masking the sequences matched by the first motif and rerunning the algorithm [23].

In MEME software, in order to assess the quality of a motif (that is to decide if it can have a relevant biological meaning or not), we have to consider the following values:

- **E-value**: it is the statistical significance of the motif. MEME usually finds the most statistically significant (low E-value) motifs first. It is unusual to consider a motif with an E-value larger than 0.05 significant so, as an additional indicator, MEME displays these partially transparent.

  The E-value of a motif is based on its log likelihood ratio, width, sites, the background letter frequencies and the size of the training set.

  The E-value is an estimate of the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and site

count, that one would find in a similarly sized set of random sequences (sequences where each position is independent and letters are chosen according to the background letter frequencies).

- **p-value**: This is the combined match p-value. The combined match p-value is defined as the probability that a random sequence (with the same length and conforming to the background) would have position p-values such that the product is smaller or equal to the value calculated for the sequence under test.

  The position p-value of a match of a given position within a sequence to a motif is defined as the probability that a randomly selected position in a randomly generated sequence (with the same length and conforming to the background) would have a match to the motif under test with a score greater or equal to that of the given position [46]. See **Appendix D** for insights into the definition of the match score.

2.4 Known motif search

FIMO, that is part of MEME suite, scans a set of sequences for individual matches to each of the motifs you provide. The name FIMO stands for 'Find Individual Motif Occurrences'. The program searches a set of sequences for occurrences of known motifs, treating each motif independently.

FIMO converts each input motif into a log-odds PSSM (Position-Specific Scoring Matrix) and uses each PSSM to independently scan each input sequence. It reports all positions in each sequence that match a motif with a statistically significant log-odds score (the procedure is the same reported in **Appendix D** for the match score calculation). You can decide the match p-value that is considered significant, and whether or not FIMO reports matches on both strands when the sequence alphabet is complementable (e.g., DNA or RNA).

In this case we have to pay attention to the **p-value** of a motif occurrence that is defined as the probability of a random sequence of the same length as the motif matching that position of the sequence with as good or better score.

3 Results and discussion

The first step of the analysis is to check if there are overrepresented Ferritin gene sequences in COM1 genome sequence. By using the Python code reported in **Appendix**, the 8 gene sequences which encode the Ferritin-like proteins (whose gene symbols are PF1193, PF0742, PF0138, PF1042, PF1190, PF1196, PF1199, PF1325) that were upregulated following gamma irradiation in the experiment described in the paper by Jocelyne et al [4], were searched in the COM1 genome and in the Reference genome and each one is found only once in each genome. The ratio between the number of Ferritin-like gene sequences found in COM1 and in Reference is 1, therefore we can not say that there is gene amplification in COM1 relative to the genes involved in protection against oxidative damage following the inactivation and disruption of DNA repair genes occured in COM1.

As a consequence, in the second part of the analysis we search known motifs associated to radioresistance and do de novo motif discovery in particular we try to understand if there are motifs typical of Ferritin-like genes and to see if they are overrepresented in COM1 genome with respect to the Reference genome, as explained in section 1. The results obtained using the software MEME are reported below.

3.1 Known motif search

As explained in section 2.4, we use FIMO to scan a set of sequences for individual matches to the motifs provided and as said in section 1.6 the TATA-box motif **TATAWAWAN** is particularly interesting because it is one of the most common motifs in the three domains of life and also because genes containing the TATA-box tend to be involved in stress-responses. For our purposes, we will use FIMO to scan the P. Furiosus reference genome sequence and its genetically tractable strain COM1 genome sequence to the TATAWAWAN motif in order to check if this motif is overrepresented in one of these genomes. The results are the following:

- There were 239 motif occurences with a p-value less than 0.0001 in the Reference genome that has 1908256 residues. The best possible match is TATATATAT.
- There were 239 motif occurences with a p-value less than 0.0001 in the COM1 genome that has 1909827 residues. The best possible match is

TATATATAT.

There is no overrepresentation of TATA-box motif in COM1 with respect to the reference genome, therefore the radio-resistance of COM1 following destruction and inactivation of DNA repair genes can not be attributed to amplification of TATA-containing genes.

Scanning an entire eukaryotic genome with transcription factor motifs using FIMO is usually a bad idea. The first problem with scanning a genome is that genomes are very large (it has a lot of DNA that isn't part of any gene) and transcription factor affinities are not very specific. Motifs typically only have 8 bases or less of specificity, and, for example, all 8-mers occur many millions of times in a eukaryotic genome. However, in this case, we are dealing with archaeal genomes that resemble bacterial genomes with respect to the number, length, and density of genes they encode: they are smaller than eukaryotic genomes and archaeal genes tend to be either adjacent to the neighboring genes or separated by less than about 200 bp resulting in high gene densities with minimal non-coding regions [51]. Moreover, it has been known for a long time that intergenic regions do contain functionally important elements such as promoters and enhancers [52]. So, if we want to see how common the motif is, in general, in this case we can look at the whole genome without too many problems and having obtained the same number of motif occurences, we can conclude that there is no TATA-box enriched genome.

3.2 De novo motif discovery

As explained in section 1.4, a protein sequence motif can be defined as a subsequence of a protein sequence that has a specific structure and that can be used for prediction of protein function and as a base of protein classification. Figure 7 shows a protein sequence motif discovered by MEME considering 11 proteins which belong to the Ferritin family and to the Ferritin-like superfamily [36] which are linked by the presence of a common four-helical bundle domain (as described in section 1.6) and whose relative encoding genes were upregulated following gamma irradiation in the experiment by Joceline et al. [4].

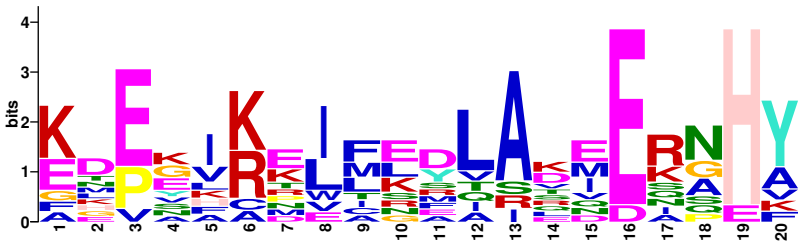

Figure 7: Discovered protein sequence motif by MEME software with $E-value = 6.9 \times 10^{-5}$.

The consensus sequence, that is the theoretical representative amino acid sequence in which each amino acid is the one which occurs most frequently at that site in the different sequences, is: **KDEKIKEIFEDLAKEERNHY**. If we search this motif using BLAST with parameters adjusted to search for a short input sequence, we find that the above consensus sequence is present in many Ferritin-like proteins of other archaeon species or bacteria (*Ignisphaera aggregans*, *Thermoprotei archaeon*, *Firmicutes bacterium* with $E-value$[1] ≤ 0.05) which are hyperthermophiles or also radioresistant, which may suggest that it is a distinctive protein pattern of these organisms having particular survival mechanisms that they have developed in extreme environments typical of many archaea species.

[1]The BLAST E-value is the number of expected hits of similar quality (score) that could be found just by chance.

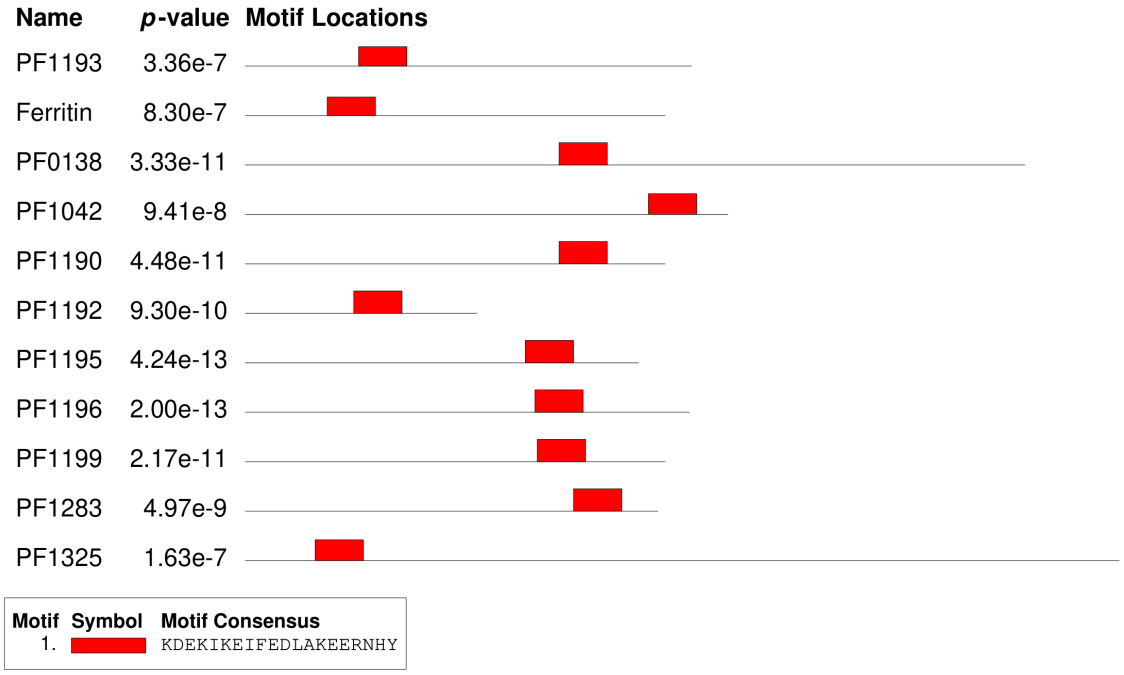

Figure 8: Protein sequence motif sites representation by MEME on the 11 Ferritin-like proteins.

Figure 8 shows the position of the protein sequence motif discovered by MEME in each protein sequence. We can notice that for 8 proteins, this motif is located at the bottom of the amino acid sequence, whereas for 3 proteins it is located upstream the protein sequence. The 3 proteins are Ferritin protein and those annotated as PF1193, whose protein name in UniProt is "DNA protection during starvation protein", and PF1325, whose protein name in UniProt is "Rubrerythrin domain-containing protein", which belong to the Ferritin family; the 8 proteins are rubrerythrin proteins (PF1199, PF1196, PF1190, PF0138 and PF1283), hypothetical protein (PF1042 and PF1195) and PF1192, that is an other one whose protein name in UniProt is "Rubrerythrin domain-containing protein".

Figure 9 shows DNA sequence motif discovered by MEME software by considering 8 genes (whose DNA sequence is available) of the 11 genes encoding the proteins which belong to the Ferritin-like superfamily and which have the common protein sequence motif represented in figure 7.

Figure 9: Discovered DNA sequence motif by MEME software with $E - value = 1.7 \times 10^{-3}$.

We can notice that this DNA sequence motif is 21 nucleotides long that is a quite common prokaryotes TF binding site length as described in the paper by Alexander et al. [37] where they analyze conservation of TFs binding sites characteristics across diverse taxa as mentioned in the introductory section 1. At first sight, we notice that in the motif there is a a predominance of Guanine base and that Adenine is completely completely conserved in second site and the seventeenth one (in this one, however, the height is a little bit lower than 2 bits so corresponding to this position there are more gaps). Moreover, figure 10 shows the DNA sequence motifs locations on the 8 gene sequences.

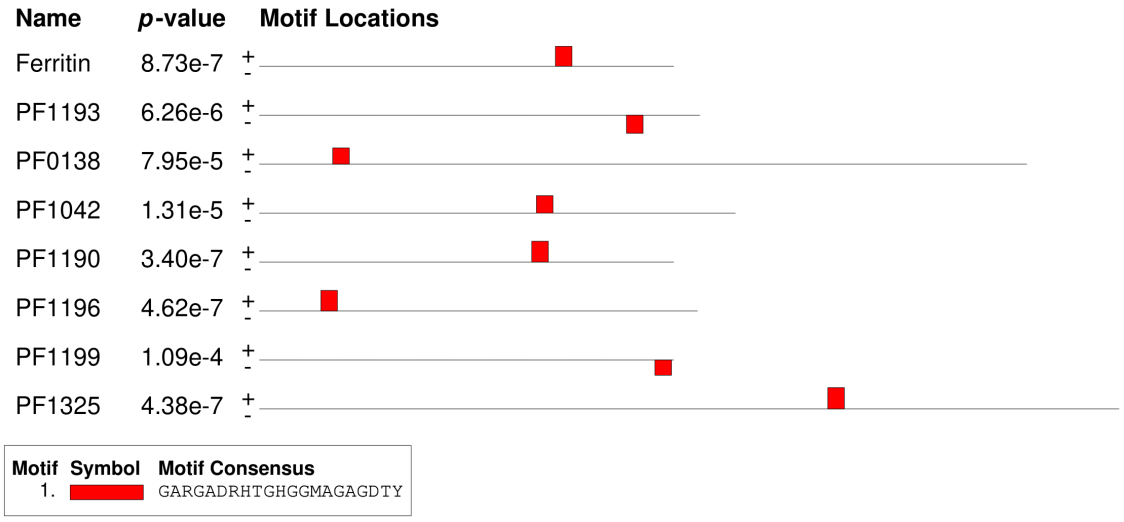

Figure 10: DNA sequence motif sites representation by MEME on the 8 genes encoding Ferritin-like proteins.

We can notice that for Ferritin, DNA protection during starvation protein PF1193, the hypothetical protein encoding gene PF1042, rubrerythrin

encoding genes PF1190, PF1199 and the Rubrerythrin domain-containing protein encoding gene PF1325 the detected motif is located at the bottom of the gene sequences (in particular for PF1193 and for PF1199 it is positioned on the negative strand) while for the rubrerythrin genes PF0138 and PF1196 it is located upstream of the genes. In the theoretical section 1.2 we have discussed that typically the DNA sequence motifs represent transcription factors binding sites and, since the analysed DNA sequences represent functionally related genes that are all upregulated after gamma irradiation [4], this motif could be a regulatory element such as an activator binding site.

In order to check if this motif has a relevant meaning, one way is to search this motif against a known motif functional database. As explained in section 2.1, MEME has an option to do this that is TomTom that allows motif comparison. The problem in this case is that the only regulatory motif databases available are for Eukaryotes and Bacteria, while for Archaea there is nothing available neither in MEME nor in other websites. However, as said in section 1.2, many of the transcription factors binding sites in Archaea are not yet identified and it is not yet well known how regulation of archaeal transcription is achieved, but we know that this domain has many aspects in common with the bacteria and eukaryotic domains, thus a comparison with the other two better known domains of life can reveal new elements of transcriptional regulation and new similarities of transcriptional regulatory mechanisms to eukaryotic or bacteria organisms.

We can firstly compare our motif **GARGADRHTGHGGMAGAGDTY** with the *CollecTF* database, that is a database of transcription factor binding sites (TFBS) in the Bacteria domain. Both Bacteria and Archaea are prokaryotes, single-celled microorganisms with no nuclei and the fossil record indicates that the first living organisms were prokaryotes (Bacteria and Archaea), and eukaryotes arose a billion years later [38]. Moreover, a number of transcription factors in Archaea govern the transcription process with homologs in bacteria.

Thus, by selecting the database category *prokaryotic DNA* and then the database *CollecTF*, we search the DNA sequence motif **GARGADRHTGHGGMAGAGDTY**. Figure 11 shows the first interesting motif that has similar profile to our motif:

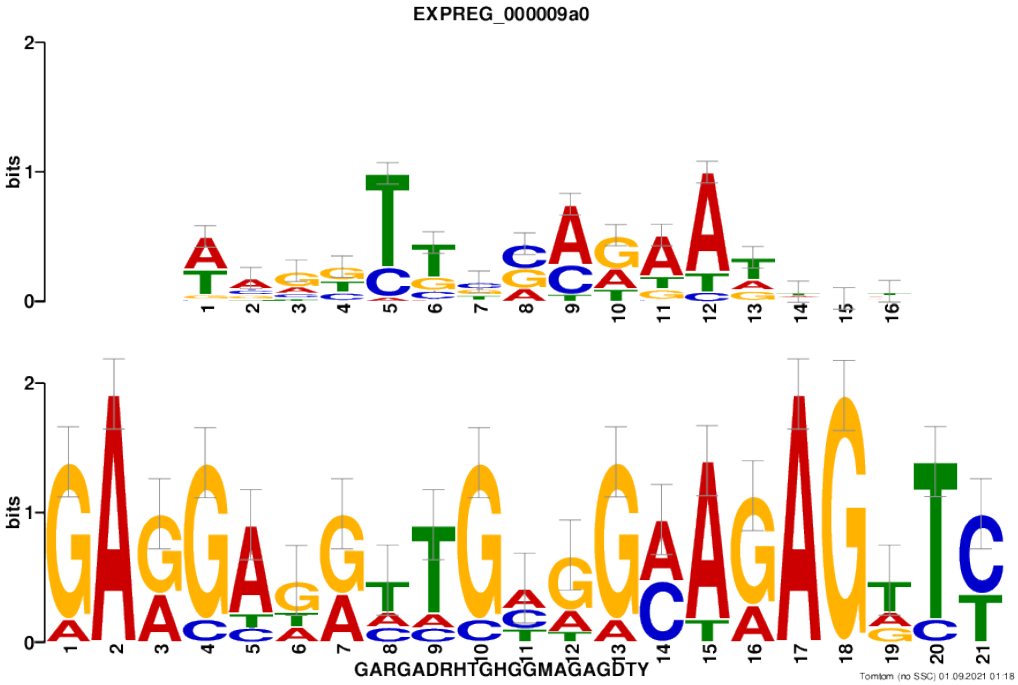

Figure 11: Optimal alignment between the binding site motif of *Pseudomonas aeruginosa* PAO1 (at the top) in CollecTF database and the query **GARGADRHTGHGGMAGAGDTY** motif (at the bottom). Error bars indicate the confidence of a motif based on the number of sites used in its creation.

The query motif matches with the *AmpR P.aeruginosa* motif (whose ID is `EXPREG_000009a0`) that represents the binding site for *HTH-type transcriptional activator AmpR*. In UniProt we can find that this protein is a positive regulator of gene expression of beta-lactamase (AmpC) that is an enzyme which convey resistance to antibiotics such as penicillins, second and third generation cephalosporins and cephamycins. In this case, $p - value$[2] $= 1.30 \times 10^{-2}$.

Now, since Archaea and Eukaryotes share different aspects (for instance the core transcription machinery is more similar to eukaryotic transcription [16]), the same job is worth doing for an eukaryotic database. Thus we select the category database *Eukaryote DNA* and the database JASPAR, that contains published and experimentally defined transcription factors binding sites for

[2]In TomTom the p-value is the probability that a random motif of the same width as the target would have an optimal alignment with a match score as good or better than the target's.

eukaryotes. The first hit is the ABR1 motif that is the transcription binding site relative to Ethylene-responsive transcription factor ABR1 of the organism *Arabidopsis thaliana (Mouse-ear cress)*. Figure 12 shows the optimal alignment. In this case, $p - value = 1.95 \times 10^{-5}$.

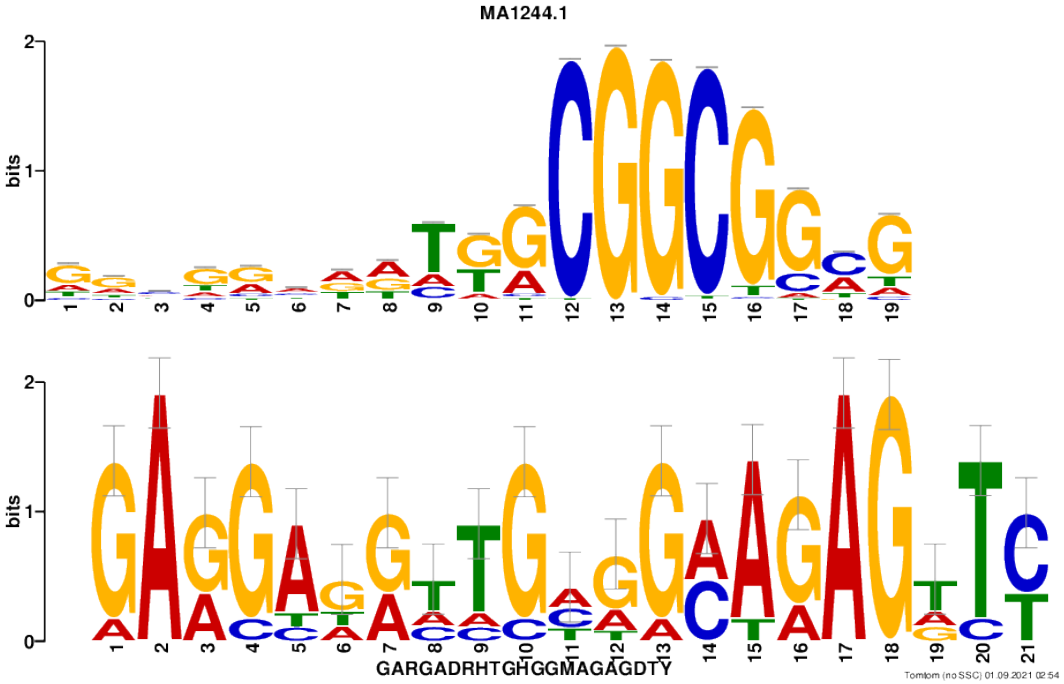

Figure 12: Optimal alignment between the binding site motif of *Arabidopsis thaliana (Mouse-ear cress)* in JASPAR database (at the top) and the query **GARGADRHTGHGGMAGAGDTY** motif (at the bottom). Error bars indicate the confidence of a motif based on the number of sites used in its creation.

We can notice that the p-value is lower than the case of the alignment with the bacterial motif indicating a better match. UniProt shows that the Ethylene-responsive transcription factor ABR1 is a negative regulator of the abscisic acid (ABA) signaling pathway involved in seed germination and in responses to stress conditions. Probably acts as a transcriptional activator. May be involved in the regulation of gene expression by stress factors and by components of stress signal transduction pathways (By similarity). Therefore this can suggest that the analysed 8 Ferritin-like genes of the P. Furiosus organisms, that were upregulated after gamma irradiation [4], can be similarly recognized and its transcription regulated in a similar manner to the *Arabidopsis thaliana (Mouse-ear cress)* genes regulated by Ethylene-responsive

transcription factor ABR1. A system similar to that in *Arabidopsis thaliana (Mouse-ear cress)* may exist in the archaeon. We can do a similar consideration for the alignment with *P. aeruginosa* motif that is involved in resistance to antibiotics and so, also in this case, it is involved in response to stress conditions. Moreover, as we have already observed, the DNA sequence motif is Guanine enriched and this reminds us the property of genes imparting resistant to radiation in radioresistant bacteria, such as *D. Radiodurans*, that have enriched promoter-PG4 motifs (discussed in subsection 1.3).

The last step of our analysis is to check if this consensus sequence motif **GARGADRHTGHGGMAGAGDTY** is overrepresented in COM1 with respect to the reference genome. The results by searching with FIMO are the following:

- There were 1819 motif occurences with a p-value less than 0.0001 in the Reference genome that has 1908256 residues. The best possible match is GAAGATATTGTGGAAGAGTTT.
- There were 1832 motif occurences with a p-value less than 0.0001 in the COM1 genome that has 1909827 residues. The best possible match is GAAGATATTGTGGAAGAGTTT.

There is an overrepresentation of GAAGATATTGTGGAAGAGTTT motif in COM1 with respect to the reference genome. Both this last motif and the TATA-box motif were searched on both positive and negative strand, since there are no experimental evidences that claim that motif orientation matters, neither what regards transcription factor binding site motifs in gene promoter regions [39] nor regulatory DNA sequences such as enhancers [47] because they can be positioned in both forward or reversed sequence orientations and still affect gene transcription.

4 Conclusions

This report has provided an overview of what motifs in molecular biology are both in DNA and protein sequences and their respective structure with an insight into one of the methods to search for motifs in sequences, that is the Expectation Maximization algorithm and its popular implementation MEME software, and its application to the hyperthermophilic archaeon Pyrococcus Furiosus organism.

The research has shown that Archaea still represent a sort of "blind spot" in our understanding of natural diversity and, in particular, here we have

recognized this issue in the archaeon transcriptional anatomy in the investigation of DNA sequence motifs. Many transcription factors and their relative binding sites await to be discovered and, in fact, one of the most important difficulties of this study has been the lack of transcription factor databases available for Archaea. However, the comparison with the transcription factor binding sites of the other two domains has allowed to enstablish similarities between the archaeal Pyrococcus Furiosus organism and Eukaryotic and Bacteria organisms from the transcriptional regulatory machinery point of view.

We have, in fact, discovered a DNA sequence motif in 8 DNA sequences of Ferritin-like genes **GARGADRHTGHGGMAGAGDTY** ($E - value = 1.7 \times 10^{-3}$) that has been compared with transcription factor binding sites of bacteria and eukaryotic domain databases (respectively CollecTF and JASPAR) thanks to TomTom and found that this motif is similar to the one that allows to be recognized by HTH-type transcriptional activator AmpR in *P. aeruginosa* (in CollecTF), that is a positive regulator of gene expression of beta-lactamase (AmpC) that is an anzyme which convey resistance to antibiotics, and to the one that allows to be recognized by Ethylene-responsive transcription factor ABR1 of the organism *Arabidopsis thaliana* (*Mouse-ear cress*) (in JASPAR) that is involved in the regulation of gene expression in response to stress conditions. The p-values are respectively 1.30×10^{-2} and 1.95×10^{-5}, suggesting a higher similarity to the eukaryotic DNA sequence motif as we could expect since the core transcription machinery is more similar to the eukaryotic one. However, in both cases TomTom has detected a similarity to motifs involved in response to stress conditions and this can suggest that our **GARGADRHTGHGGMAGAGDTY** motif might be similarly recognized and its transcription regulated in a similar manner to those occuring in *P. aeruginosa* and in *Arabidopsis thaliana*. A system of gene regulation similar to that in *P. aeruginosa* and in *Arabidopsis thaliana* may exist in the archaeon. To go any further, one would need to identify the gene for the putative transcription factor, show that it affected the transcription directly or perform genetic experiments that would allow the same conclusion. Anyway, our results already confirm that **GARGADRHTGHGGMAGAGDTY** might have a relevant biological meaning.

The first purpose of this research was to investigate the COM1 radioresistance following its DNA repair genes inactivation. Assuming a mechanism of compensation given by the "reinforcement" of protection system against oxidative damage, we searched for duplication of ferritin-like genes in the COM1 genome with respect to the reference, using the Python code reported in **Appendix C**, but they were found in both organisms 1 time. We searched

the well known TATAWAWAN motif using FIMO, but there were found 239 motif occurences with a p-value less than 0.0001 in both genomes. Then, since we can consider the discovered DNA sequence motif **GARGADRHTGHGGMAGAGDTY** as a "fingerprint" of genes encoding proteins which belong to Ferritin family, or of genes with similar functions, we searched it using FIMO and in this case there is an overrepresentation of this DNA sequence motif in COM1 with respect to the reference genome and this can confirm our hypothesis of a "reinforced" DNA protection system in the radioresistance machinery of COM1 highlighting the great role of these genes, which are devoted to DNA protection against oxidative damage, in the radioresistance of the P. Furiosus species.

A great attention of this work has been dedicated to the genes encoding proteins which belong to Ferritin family because in the experiment of Joceline et al. [4] these were upregulated following gamma irradiation. In this project, we considered the 11 amino-acid sequences of these proteins and found a common protein sequence motif using MEME that is **KDEKIKEIFEDLAKEERNHY** with $E - value = 6.9 \times 10^{-5}$. Using BLAST, we have found that this amino-acid sequence pattern is present in several hyperthermophiles or also radioresistant organisms (*Ignisphaera aggregans*, *Thermoprotei archaeon*, *Firmicutes bacterium* with $E - value \leq 0.05$) which may suggest that it is a distinctive protein pattern of these particular organisms having particular survival mechanisms that they have developed dealing with extreme environments where many archaea species are found. One further step of this work could be that to identify this protein sequence in the protein structure or also to look at the protein structure of a DNA sequence motif if we are looking for a DNA sequence motif which, when translated, produces a protein domain and so that is only relevant within coding sequences. By considering that archaea genomes typically have minimal non-coding regions, if there is a DNA sequence motif, it is very likely that it will fall within the coding regions.

Thus, bioinformatic tools help us to clarify these "blind spots" in our understanding of natural diversity. With a limited set of tricks, the expectation maximization algorithm provides a simple and robust tool for parameter estimation in models with incomplete data. In theory other numerical optimization techniques, such as gradient descent or Newton-Raphson, could be used instead of expectation maximization: in practice, however, expectation maximization has the advantage of being simple, robust and easy to implement. Here we have exploited one popular implementation MEME that is very fast in detecting common patterns in a given set of sequences, in fact for our 8 DNA sequences, the time was ~ 10 seconds and for our 11 protein

sequences, the time was ~ 15 seconds. Moreover, its deterministic approach avoids to stuck in poor local maxima giving us the optimal result.

Appendix A
DNA

DNA (DeoxyriboNucleic Acid) consists of long chains of units called nucleotides. As in Figure 13, each nucleotide is composed by a nitrogenous base, a sugar and a phosphate group bond together.

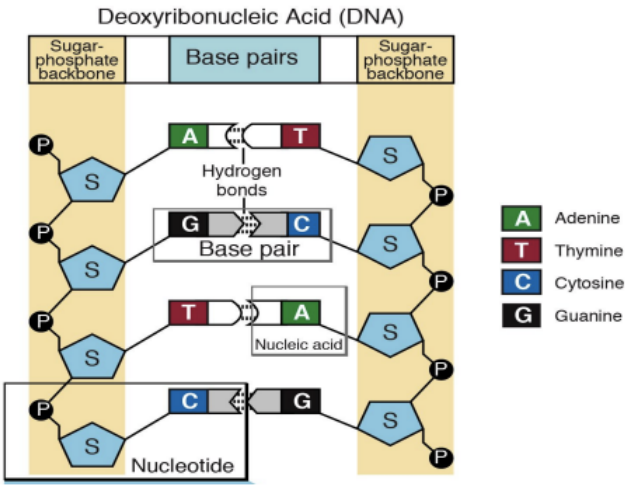

Figure 13: DNA portion, with two strands of 4 nucleotides held together by hydrogen bonds. Image from the National Human Genome Research Institution.

Adjacent nucleotides are joined by a phosphodiester linkage, which consists of a phosphate group that links the sugars of two nucleotides. This bonding results in a backbone with a repeating pattern of sugar-phosphate units. Only certain bases in the double helix are compatible with each other. Adenine (A) always pairs with thymine (T), and guanine (G) always pairs with cytosine (C). Thus, the two strands of the double helix are complementary.

We can say that a polynucleotide has a built-in directionality along its sugar-phosphate backbone. The two sugar-phosphate backbones runs in opposite direction: one is called the forward strand or positive strand (from 5' to 3') and the other the reverse strand or negative strand (from 3' to 5') as shown in figure 14.

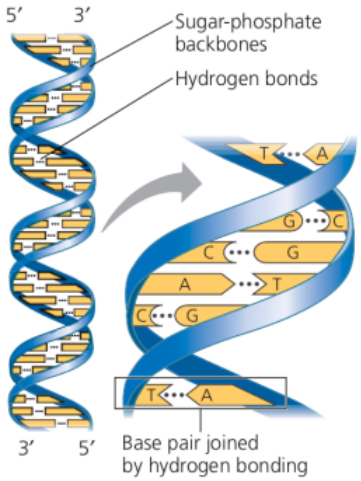

Figure 14: DNA double helix.

Appendix B

Nucleotide symbol	Full name
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Guanine / Adenine (purine)
Y	Cytosine / Thymine (pyrimidine)
K	Guanine / Thymine
M	Adenine / Cytosine
S	Guanine / Cytosine
W	Adenine / Thymine
B	Guanine / Thymine / Cytosine
D	Guanine / Thymine / Cytosine
H	Adenine / Cytosine / Thymine
V	Guanine / Cytosine / Adenine
N	Adenine / Guanine / Cytosine / Thymine

Table 1: Nucleotide symbols

Appendix C

Python code for gene amplification detection

```
#Input

Input_Sequence=input("Enter the gene sequence to be searched=")


#Convert into one line gene sequence

x=''.join(Input_Sequence)
One_Line_Input=x.replace('\n', '')

#Open files in read mode

Genome_Sequence_COM1=open("COM1_Complete_Sequence.FASTA",'r')
Genome_Sequence_Ref=open("Reference_Complete_Sequence.FASTA","r")

#Tuples
Genomes=(Genome_Sequence_COM1,Genome_Sequence_Ref)
names=('COM1','Reference')


#Zip joins two tuples together
for Genome,name in zip(Genomes,names):

    #Convert into one line genome sequence
    y=''.join(Genome)
    One_Line_File=y.replace('\n', '')

    #Count how many times the entered sequence appears in the
    ↪  genome sequence
    N_Times=One_Line_File.count(One_Line_Input)

    #Output
    if One_Line_Input in One_Line_File:
        print("\nThe sequence {} is found {} times in {}
        ↪  genome".format(Input_Sequence,N_Times,name))
    else:
```

```
print("\nThe sequence {} is not found in {}
↪  genome".format(Input_Sequence,name))
```

Appendix D

Position weight matrix [32]

A position weight matrix (PWM), also known as a position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM), is a commonly used representation of motifs (patterns) in biological sequences. PWMs are often represented graphically as sequence logos that is, in fact, an other common rapresentation of motifs (as explained in section 1.7).
PWMs are usually derived from a set of aligned sequences that are thought to be functionally related and have become an important part of many software tools for computational motif discovery.

Conversion of sequence to position probability matrix

A PWM has one row for each symbol of the alphabet (4 rows for nucleotides in DNA sequences or 20 rows for amino acids in protein sequences) and one column for each position in the pattern (or viceversa). In the first step in constructing a PWM, a basic position frequency matrix (PFM) is created by counting the occurrences of each nucleotide at each position. From the PFM, a position probability matrix (PPM) can now be created by dividing that nucleotide count at each position by the number of sequences, thereby normalising the values. Formally, given a set X of N aligned sequences of length l, the elements of the PPM **M** are calculated:

$$M_{k,j} = \frac{1}{N} \sum_{i=1}^{N} I(X_{i,j} = k) \tag{14}$$

where i ∈ (1,...,N), j ∈ (1,...,l), k is the set of symbols in the alphabet and $I(X_{i,j} = k)$ is an indicator function where $I(X_{i,j} = k)$ is 1 if $k = K$, where K is the nucleotide that we are considering, and 0 otherwise. For example, given the following DNA sequences:

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT

TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

The corresponding PFM is:

$$M = \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{matrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

Therefore, the resulting PPM is:

$$M = \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0 & 0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0 & 0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 0.1 & 0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0 & 0.1 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

Both PPMs and PWMs assume statistical independence between positions in the pattern, as the probabilities for each position are calculated independently of other positions. From the definition above, it follows that the sum of values for a particular position (that is, summing over all symbols) is 1. Each column can therefore be regarded as an independent multinomial distribution. This makes it easy to calculate the probability of a sequence given a PPM, by multiplying the relevant probabilities at each position. For example, the probability of the sequence S = GAGGTAAAC given the above PPM **M** can be calculated:

$$p(S|M) = 0.1 \times 0.6 \times 0.7 \times 1.0 \times 1.0 \times 0.6 \times 0.7 \times 0.2 \times 0.2 = 0.0007056 \quad (15)$$

This is usually referred as the *score* of the sequence and in particular we usually consider the *match score*. The match score of a motif to a position in a sequence is the product of the score from each row of the position-dependent scoring matrix **M** corresponding to the letter at that position in the sequence [46]. For example, if the sequence is:

TAATA<u>GAGGTAAAC</u>GTTTTTGTGGCATCGGGCGAGAATAGCGC

then the match score of the fifth position in the sequence (underlined) would be found by equation (15).

The sequence score gives an indication of how different the sequence is from a random sequence. The score is 0 if the sequence has the same probability of being a functional site and of being a random site. The score is greater than 0 if it is more likely to be a functional site than a random site, and less than 0 if it is more likely to be a random site than a functional site.

Conversion of position probability matrix to position weight matrix

Most often the elements in PWMs are calculated as log likelihoods. That is, the elements of a PPM are transformed using a background model **b** so that:

$$M_{k,j} = \log_2(M_{k,j}/b_k) \tag{16}$$

describes how an element $M_{k,j}$ in the PWM can be calculated. The simplest background model assumes that each letter appears equally frequently in the dataset, that is, the value of $b_k = 1/|k|$ for all symbols in the alphabet (0.25 for nucleotides and 0.05 for amino acids). When the PWM elements are calculated using log likelihoods, the score of a sequence (that is the probability of a sequence S given the PWM) can be calculated by adding (rather than multiplying) the relevant values at each position in the PWM.

References

[1] G. Castellani, teaching material of *Physical Methods of Biology* course, University of Bologna.

[2] Stephanie L. Bridger, W. Andrew Lancaster, Gerrit Jan Schut, Farris L. Poole, *Genome Sequencing of a Genetically Tractable Pyrococcus Furiosus Strain Reveals a Highly Dynamic Genome.*

[3] Gina L Lipscomb, Karen Stirrett, Gerrit J Schut, Fei Yang, Francis E Jenney Jr, Robert A Scott, Michael W W Adams, Janet Westpheling, *Natural competence in the hyperthermophilic archaeon Pyrococcus furiosus facilitates genetic manipulation: construction of markerless deletions of genes encoding the two cytoplasmic hydrogenases.*

[4] Ernest Williams, Todd M. Lowe, Jeffrey Savas, Jocelyne DiRuggiero, *Microarray analysis of the hyperthermophilic archaeon Pyrococcus Furiosus exposed to gamma irradiation.*

[5] Mary Ann Knovich, Jonathan A. Storey, Lan G. Coffman, and Suzy V. Torti, *Ferritin for the Clinician.*

[6] Lina Bai, Ting Xie, Qingqing Hu, Changyan Deng, Rong Zheng, Wanping Chen, *Genome-wide comparison of ferritin family from Archaea, Bacteria, Eukarya, and Viruses: its distribution, characteristic motif, and phylogenetic relationship.*

[7] Lecture 8:Motifs and Motifs finding (with a section on Chip-Seq), Principles of Computational Biology, Teresa Przytycka, PhD.

[8] Alexandra M. Gehring, Julie E. Walker, Thomas J. Santangelo, *Transcription Regulation in Archaea.*

[9] Transcription Factor/Transcription Factors, Scitable by nature education.

[10] Promoter (genetics)
url: https://en.wikipedia.org/wiki/Promoter_(genetics)

[11] Oleg V. Bylino, Airat N. Ibragimov and Yulii V. Shidlovskii, *Evolution of Regulated Transcription.*

[12] Re: Do archaea have INTRONS and do they undergo splicing ?
url:http://www.madsci.org/posts/archives/2003-07/1059068302.Ge.r.html

[13] H Xu and T R Hoover, *Transcriptional regulation at a distance in bacteria.*

[14] Patrik D'haeseleer, *What are DNA sequence motifs ?*
url: https://www.nature.com/articles/nbt0406-423

[15] CHAPTER 3: REGULATORY MOTIFS
url: http://web.mit.edu/manoli/www/thesis/Chapter3.html

[16] Archaeal transcription
url: https://en.wikipedia.org/wiki/Archaeal_transcription

[17] Transcription factors
url:https://en.wikipedia.org/wiki/Transcription_factor

[18] Michael Y.Galperin, Dmitrij Frishman, *Towards Automated Prediction of Protein Function from Microbial Genomic Sequences*, Methods in Microbiology, Volume 28, 1999, Pages 245-263.

[19] Peer Borkab, Eugene VKooninc, *Protein sequence motifs*, Current Opinion in Structural Biology, Volume 6, Issue 3, June 1996, Pages 366-376.

[20] Protein Domains, Motifs, and Folds in Protein Structure.
url: https://bio.libretexts.org

[21] Structural Motifs
url: https://en.wikipedia.org/wiki/Structural_motif

[22] L Aravind, Vivek Anantharaman, Santhanam Balaji, M Mohan Babu, Lakshminarayan M Iyer, *The many faces of the helix-turn-helix domain: transcription regulation and beyond.*

[23] Patrik D'haeseleer, *How does DNA sequence motif discovery work?*
url: https://www.nature.com/articles/nbt0806-959

[24] The MEME suite
url: https://meme-suite.org/meme/doc/overview.html?man_type=web

[25] Sequence logo
url: https://en.wikipedia.org/wiki/Sequence_logo

[26] TATA-box
url: https://en.wikipedia.org/wiki/TATA_box

[27] Jorg Soppa, *Transcription initiation in Archaea: facts, factors and future aspects.*

[28] Ferritin-like diiron domain
url: https://www.ebi.ac.uk/interpro/entry/InterPro/IPR009040/

[29] DPS-like protein, ferritin-like diiron-binding domain
url: https://www.ebi.ac.uk/interpro/entry/InterPro/IPR033921/

[30] Proteins and Enzymes: What is the difference between motifs and domains?
url:https://www.life.illinois.edu/mcb/150/private/faq/pdf/1390.pdf

[31] Nicolas Beaume, Rajiv Pathak, Vinod Kumar Yadav, Swathi Kota, Hari S. Misra, Hemant K. Gautam and Shantanu Chowdhury, *Genome-wide study predicts promoter-G4 DNA motifs regulate selective functions in bacteria: radioresistance of D. radiodurans involves G4 DNA-mediated regulation.*

[32] Position weight matrix
url: https://en.wikipedia.org/wiki/Position_weight_matrix

[33] Fatma A. Hashim, Mai S. Mabrouk and Walid Al-Atabany, *Review of Different Sequence Motif Finding Algorithms.*

[34] Chuong B Do and Serafim Batzoglou, *What is the expectation maximization algorithm ?*

[35] STAT115 Chapter 10.2 Expectation Maximization for Motif Finding url: https://www.youtube.com/watch?v=kq5NAd4pnkU

[36] Ferritin-like superfamily domain assignments.

[37] Alexander J Stewart, Sridhar Hannenhalli, Joshua B Plotkin, *Why Transcription Factor Binding Sites Are Ten Nucleotides Long.*

[38] Organismal Biology url:https://organismalbio.biosci.gatech.edu/biodiversity/prokaryotes-bacteria-archaea-2/

[39] Monika Lis and Dirk Walther, *The orientation of transcription factor binding site motifs in gene promoter regions: does it matter?*

[40] Archaea and the Tree of Life url:https://www.ibiology.org/microbiology/archaea/#

[41] How to Odentify Protein Motifs from Protein Sequences. Bioinformatics: Methods Express. Edited by Paul H. Dear, Scion, 2007.

[42] Walker motifs url:https://en.wikipedia.org/wiki/Walker_motifs

[43] C. Ramakrishnan, V.S. Dani, T. Ramasarma, *A conformational analysis of Walker motif A [GXXXXGKT (S)] in nucleotide-binding and other proteins.*

[44] Simon C. Andrews, *The Ferritin-like superfamily: Evolution of the biological iron storeman from a rubrerythrin-like ancestor.*

[45] Sequence Logo url:https://www.youtube.com/watch?v=UnU5M7rYYvE

[46] MAST - Motif Alignment and Search Tool url:http://www.cbil.upenn.edu/EpoDB/release/version_2.2/meme/mast-output.html

[47] Enhancer, Scitable by nature education.

[48] Eveline Peeters and Daniel Charlier, *The Lrp Family of Transcription Regulators in Archaea.*

[49] Stephen D Bell, Stephen S Cairns, Robert L Robson and Stephen P Jackson, *Transcriptional Regulation of an Archaeal Operon In Vivo and In Vitro.*

[50] Yoh-ichi Watanabe, Shin-ichi Yokobori, Toshiro Inaba, Akihiko Yamagishi, Tairo Oshima, Yutaka Kawarabayasi, Hisasi Kikuchi and Kiyoshi Kita, *Introns in protein-coding genes in Archaea.*

[51] D.W. Grogan, *Archaea* in Brenner's Encyclopedia of Genetics (Second Edition), 2013.

[52] Intergenic region
url:https://en.wikipedia.org/wiki/Intergenic_region

[53] Protein structure prediction in 1D, 2D, and 3D
url:https://www.rostlab.org/papers/1998_encyclopedia/paper.html

[54] Meaning of 'motif' in molecular biology https://biology.stackexchange.com/q/101835/61683