Universidade de Caxias do Sul
Área de conhecimento de ciências da vida
Instituto de Biotecnologia
Programa de Pós Graduação em Biotecnologia

Characterization and annotation of archaeal promoter sequences

Gustavo Sganzerla Martinez

Caxias do Sul
2021

Gustavo Sganzerla Martinez

Characterization and annotation of archaeal promoter sequences

Tese apresentada ao Programa de Pós-graduação em Biotecnologia da Universidade de Caxias do Sul, visando a obtenção do grau de Doutor em Biotecnologia.

Orientadora: Prof. Dra. Scheila de Avila e Silva
Co-orientador: Prof. Dr. Aditya Kumar

Caxias do Sul
2021

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade de Caxias do Sul
Sistema de Bibliotecas UCS - Processamento Técnico

M385c Martinez, Gustavo Sganzerla
Characterization and annotation of archaeal promoter sequences [recurso eletrônico] / Gustavo Sganzerla Martinez. – 2021.
Dados eletrônicos.

Tese (Doutorado) - Universidade de Caxias do Sul, Programa de Pós-Graduação em Biotecnologia, 2021.
Orientação: Scheila de Avila e Silva.
Coorientação: Aditya Kumar.
Modo de acesso: World Wide Web
Disponível em: https://repositorio.ucs.br

1. Archaea. 2. DNA. 3. Biotecnologia. I. Silva, Scheila de Avila e, orient. II. Kumar, Aditya, coorient. III. Título.

CDU 2. ed.: 582.233

Catalogação na fonte elaborada pela(o) bibliotecária(o)
Carolina Machado Quadros - CRB 10/2236

Gustavo Sganzerla Martinez

Characterization and annotation of archaeal promoter sequences

Tese apresentada ao Programa de Pós-graduação em Biotecnologia da Universidade de Caxias do Sul, visando a obtenção do grau de Doutor em Biotecnologia.

Orientadora: Prof. Dra. Scheila de Avila e Silva
Co-orientador: Prof. Dr. Aditya Kumar

TESE APROVADA EM: ____/____/____

_____
Orientadora: Prof. Dra. Scheila de Avila e Silva

_____
Co-orientador: Prof. Dr. Aditya Kumar

_____
Prof. Dr. Sergio Echeverrigaray Laguna

_____
Prof. Dr. Venkata Rejesh Yella

_____
Prof. Dr. Edgardo Gálan Vásquez

Agradecimentos

A obtenção de um título de doutorado é algo grandioso, mais ainda é a pessoa que sabe como utilizar esse título em prol do bem. Espero poder fazer o melhor uso do meu. Devido a isso, preciso agradecer a todos que pude cruzar caminhos nesses anos de doutorando. Aos meus familiares, mãe – Teresinha, que possibilitou isso; avós, tios e primos. À Renata, estimada companheira que esteve comigo do início ao fim, em momentos de glória e frustração.

Agradecimentos não podem faltar a prof. Dra. Scheila de Avila e Silva, com quem tenho tido o prazer de trabalhar desde a graduação e conseguiu alocar as minhas aspirações científicas. Todo o grupo do Laboratório de Biologia Computacional da UCS, caros colegas com quem tive o prazer de crescer junto.

A Universidade de Caxias do Sul foi o local onde escolhi para fazer ciência. Tive sempre apoio incondicional da instituição. Um exemplo foi o trabalho que tive a oportunidade de fazer junto aos professores Sergio Echeverrigaray e Ana Paula Delamare em seu laboratório, o que ajudou a despertar a minha curiosidade pela biologia.

Desde que a professora Scheila e eu começamos a internacionalizar o PPG, e tivemos o prazer de convidar o Dr. Aditya Kumar para me co-orientar, exímio pesquisador que me acompanhou durante toda essa jornada. Também quero agradecer ao Dr. Ernesto Perez-Rueda, que foi um colaborador direto para que esse doutorado tenha sido concluído.

Acima de tudo, eu agradeço à verdade. Afinal, o cientista busca se aproximar da verdade.

"It is the small things, everyday deeds of ordinary folk that keeps the darkness at bay, simple acts of love and kindness." (J. R. R. Tolkien, *The Fellowship of the Ring*)

Summary

LIST OF FIGURES

LIST OF TABLES

List of abbreaviations

A – adenine
AI – artificial intelligence
ANN – artificial neural network
ATP – adenosine triphosphate
AUC – area under the curve
Bp – base pair
BMHT -
BRE – B recognition element
C – cytosine
CS – crystal structure
DDS – DNA duplex stability
DiProDB – Dinucleotide property database
DNA – deoxyribonucleic acid
E. coli – Escherichia coli
FN – false negative
FP – false positive
FPR – false positive rate
G - guanine
H. volcanii – Haloferax volcanii
LUCA – last unique common ancestor
ML – machine learning
mRNA – messenger ribonucleic acid
ORF – open reading frame
PIC – pre-initiation complex
pN – picoNewton
RNA – ribonucleic acid
RNAP – ribonucleic acid polymerase
rRNA – ribosomal ribonucleic acid
rprop – resilient backpropagation
ROC – receiver operator characteristic
T. kodakarensis – Thermococcus kodakarensis
T – thymine
Tanh – hyperbolic tangent
TPB – TATA-box binding protein
TFB – Transcription factor B binding protein
TP – true positive
TN – true negative
TSS – transcription start site
S. solfataricus – Sulfolobus solfataricus
SSID – stress induced DNA duplex destabilization

ABSTRACT

Archaea were not long ago included as a particular domain in the tree of life. This vast and unexplored domain requires a plethora of techniques aimed to produce reliable inferences on it since not much information has been annotated. The objective of this thesis is to create a bioinformatics pipeline that encounters promoters in unannotated genomes of archaea. To perform such task, the conservation reported in the binding site of specific transcription factor proteins is considered in order to draw a profile of promoter-like sequences derived from experimentally validated promoters of the archaea *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. Additionally, the conserved aspect of these organisms is employed is a classification task based on statistics and artificial neural networks. The results obtained in this thesis are displayed in the form of two scientific articles. Where in the first, the binding site of transcription factor proteins that assist RNAP binding was located even in promoter sequences that lack the canonical binding sites as well as unannotated organisms. Secondly, a combination of statistics and artificial neural network classification has succeeded in classifying archaeal promoters in a rate > 90%, which has shown a satisfactory way to deliver annotation upon sequences whose genome is not fully explored.

1 INTRODUCTION

The inclusion of the archaeal domain as a separate branch in the tree of life is recent (Woese, 1977). Since then, archaeal organisms have been sequenced and their information has become available. For a long time, archaea held the title of being extremophile organisms due to their abundance in places such as hot thermal vents, environments with a high salt concentration, freezing polar temperatures, acidic and alkaline waters, and other unfavorable conditions for life. However, new discoveries have suggested archaea are very abundant and found in basically every ecosystem of this planet (DeLong *et* al, 1994). There may also be inter-planetary evidence of these organisms (Maus *et* al, 2020). Nonetheless, archaea are reported to live in a mutualistic relation with other living beings, since they are found in the human digestive system, which helps to get rid of excessive hydrogen by converting it into methane (Nkamga, 2017).

Due to many archaea being found in extreme environments, their cultivation on a laboratory is hampered (Sun *et* al, 2020). Thus, *in silico* studies have gained attention and became to coexist with experimental approaches, facilitating hypotheses formulation and testing. Due to the data availability, a computational approach favors a preliminary analysis, decreasing the amount of experimental research to be done, which shrink costs and time on scientific research (Kuok *et* al, 2017).

Today, there is an estimated number of one nonillion ($1x10^{27}$) prokaryotic organisms, which all have complex cellular functions. Moreover, there have been reports that extensively explored the relationship between microorganisms' communities and their impact on shaping a suitable Earth (Hallam, 2019).

Therefore, through the unprecedented technological advances that are being experienced today, the genome of archaea might be better explored and its information extracted. This study aims to combine technological novelties in order to deliver a higher quality of genome annotation in the genome of archaea.

2 OBJECTIVES

2.1 GENERAL OBJECTIVE

The goal of this thesis is to create a pipeline that encompasses supervised learning and statistics to be used to annotate archaeal promoter elements.

2.2 SPECIFIC OBJECTIVES

- Capture signals that distinguish promoter and non-promoter sequences in archaeal genomes.
- Explore the genome of archaea apart from consensual sequences by codifying genetic information into numeric attributes.
- Classify archaeal promoters through an Artificial Neural Network approach.
- Classify archaeal promoters through a statistical model.
- Locate promoters in unannotated archaea.

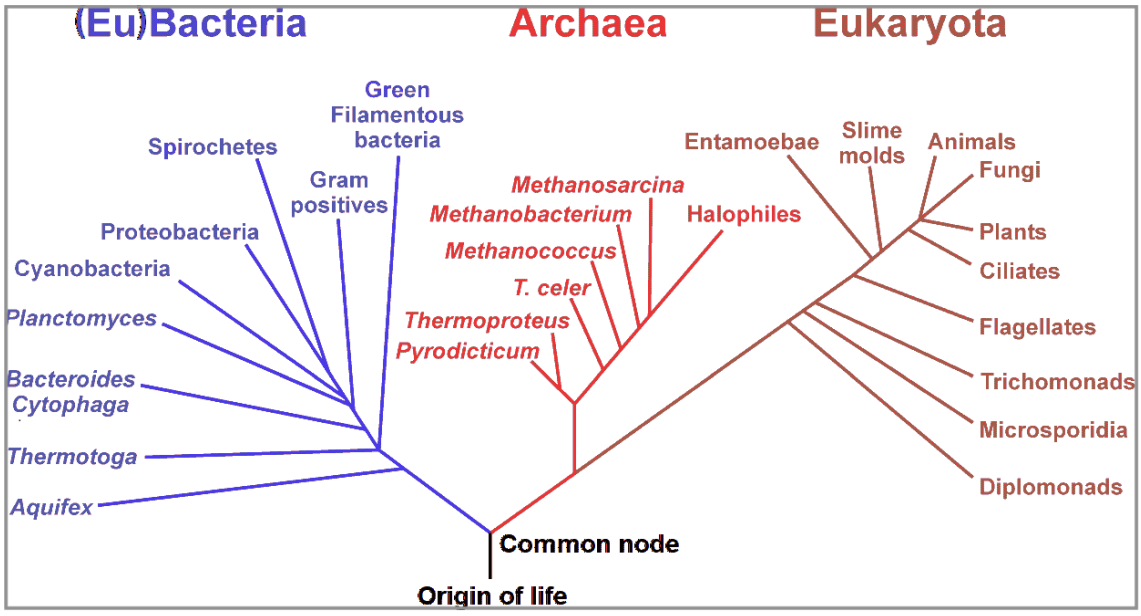
3 LITERATURE OVERVIEW

3.1 THE EVOLUTION OF ARCHAEA

Archaea are old. Carbon tracing techniques date these organisms to 2.7 million of years ago. This tracking measured the origin of methane whose origin is biological (i.e. methanogens archaea) (Gribaldo and Brochier-Armanet, 2006). Evolutionary traits place archaea in-between eukaryotes and bacteria (Gupta, 1998). A pre-archaea-tree of life would indicate that bacteria gave origin to eukaryotes, where they both shared a last universal common ancestor (LUCA). Although, the Woese and Fox (1977) inclusion of archaea in the tree of life changed this model. In this sense, the tree of life (depicted in Figure 3-1) started with a LUCA, which gave origin to both bacteria and archaea. Resemblances of these two domains have been studied (Lopez *et* al. 1999; Aravind, 1999; Zivanovic, 2002). Eukaryotes would arise only after evolutionary events in bacteria and archaea. Indeed, there are evidences that define many archaeal regulatory mechanisms as simplified version of the eukaryotic counterparts (Olsen and Woese, 1996).

Archaea are, functionally, placed in the middle of the other two domains, sharing aspects with both. Evolutionists still discuss about details, especially due to the difficulty of cultivating archaea (Forterre and Philippe, 2002). The correct tracing of the tree of life and, consequently, the discovery of the exact cell that gave origin to eukaryotes is one of the most enigmatic questions in biology; and archaea could play a major role on it (Imachi *et* al, 2020).

Figure 3-1: The three domains of life



3.2 ARCHAEA AS THE THRID DOMAIN

Until 1987, every single living organism in this planet would either fit in one of two domains: eukaryotes (organisms that possess a nuclear membrane, preventing the genetic material from wandering freely in the cytoplasm) and, prokaryotes (unicellular beings that do not own a nuclear membrane, i.e., the deoxyribonucleic acid (DNA) has no physical barrier within the cell) (Woese, 1987). At the time, the prokaryotic organisms were divided into Eubacteria (true bacteria) and Archaeabacteria (Gribaldo and Brochier-Armanet, 2006). This latter lineage of "bacteria" was known to inhabit extreme environments. This dichotomy consented by biologists at the time had never seemed to be static. It need to be remembered that once bacteria were considered plants (Woese and Fox, 1977).

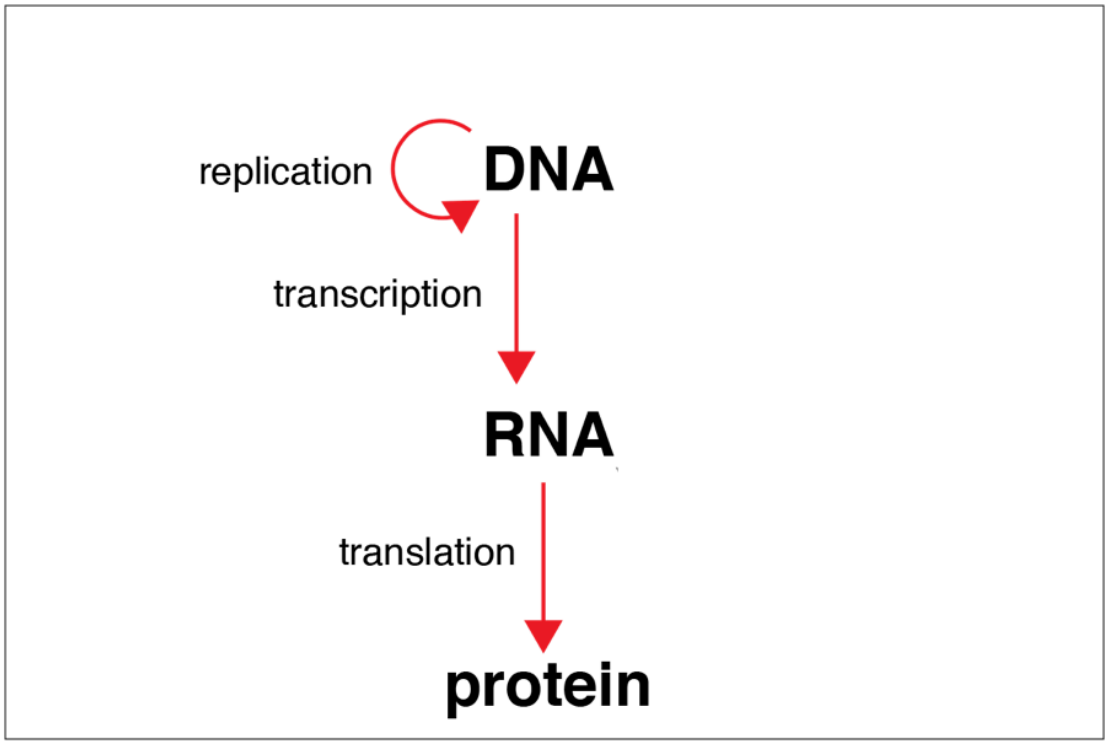
Following the idea that the ultimate record of the evolutionary history of any organism is its genome (Zuckerkandl and Pauling, 1965), the phylogenetic tree of all living beings drastically changed in 1987. Woese and Fox (1987) proposed the existence of a third domain of life that would differ from the other two. This change of paradigm in microbiology encompassed an analysis done by the researcher who found the ribosomal RNA (rRNA) of the Archaebacteria to be different from Eubacteria. Since rRNA is a well-conserved molecule found in every living being, it conveys to be a great phylogenetic marker.

The findings of Woese and Fox (1977) portrayed a unique paradigm. At the time, methanogens (prokaryotes that produce methane as a metabolic byproduct) were considered to be bacteria. The organisms mapped at the time to be archaebacterial were methanogens only. As time went through, the list of the representatives of this third domain substantially grew. After owning only methanogens, some extremophile organisms were included (i.e., organisms that inhabit acidic, alkaline, sulfurous, highly saline, and high-temperature environments). Then, microbiologists found that archaea were even more diverse. Being organisms that are widely distributed in nature and are prevalent in not so extreme habitats such as soil and oceans, turning archaea as significant contributors to the global carbon and nitrogen cycle (Gribaldo and Brochier-Armanet, 2006).

3.3 THE TRANSCRIPTION PROCESS IN ARCHAEA

The central dogma in molecular biology revolves around three distinct processes: replication, transcription and, translation, as Figure 3-2 shows. These processes are critical for the maintenance of the cellular functions in any cellular organism. In general terms, independently of the organism in question, the trace of genetic information starts when cells multiply, a copy the genetic material has to be transferred to the daughter cells. This process is known as replication. To convey gene expression, the information stored in the DNA is used to form an intermediate molecule, the ribonucleic acid (RNA). One of the reasons for this extra molecule to be present is to help to preserve DNA, preventing it from leaving the cell. Finally, after an RNA molecule is formed, it leaves the cell and is translated into a peptide chain, which will result in a functional protein (Krebs *et al*, 2017).

Figure 3-2: The central dogma of molecular biology (adapted from Watson and Crick, 1953).



The DNA is described as an information repository, needed to build RNA and, through it, a protein that has a function related to cell structure, regulation or catalysis. There are several mechanisms that ensure the correct genes are expressed in the right moment. This correct expression of certain genes grants the cell the ability of having its necessities supplied (Browning and Busby, 2016).

The transcription process is essential to life at any domain. This process is responsible for converting DNA into RNA. The element that orchestrates this is the enzyme RNA polymerase (RNAP). This multi subunit enzymatic complex is required to carry out transcription in the three domains of life. Since evolution has acted upon RNAP, its function is different among archaea, bacteria and, eukaryotes but, the final product of the enzyme is the same - RNA. However, the path that leads to the formation of mRNA is quite different in the tree of life. Bacterial RNAP alone is not capable of recognizing promoter elements and, consequently, genes. It employs one of its subunits to identify promoter sequences along the DNA strand. The σ subunit helps RNAP recruitment in bacteria (Krebs *et al*, 2017).

The eukaryotic/archaeal transcription resemble each other. In fact, archaea transcription is a simplified version of eukaryotic transcription, resembling the components of eukaryotic RNAP II (Gehring *et al*, 2016). The first drawn parallel among the likeness of transcription machinery of these two domains occurred when Zillig *et al*, (1979) analyzed the subunit composition of archaeal RNAP. It ended up being far more complex than the bacterial counterparts, while archaea have at least 10 subunits, bacteria have only five. Further analysis (Langer *et al*, 1995) compared eukaryotic genes that encode RNAP subunits and found them to be homologous to archaea.

The archaeal RNAP on its own is not capable of efficiently recognize a promoter element (Bell and Jackson, 2001) and then, starting RNA synthesis. There are two basal transcription factors that are minimally required to recruit RNAP and continue the transcription stage: *i*) a TATA-binding protein (TPB) and, *ii*) a Transcription Factor B protein (TFB).

There are three steps in which the transcription in archaea happens. The first, the initiation (Figure 3-3) complex, starts with the recruitment of RNAP to the DNA. This is mediated by the presence of a TBP and a TFB protein. The archaeal RNAP is generally found within 12 subunits (A', A'', B', B'', D, N, L, P, H, K, F, E). From which, the two A subunits recruit RNAP to the promoter element (Werner, 2007). Additional functions of the other subunits are included in Table 3-1. During the initiation, there is a pre-open complex (PIC), whose function is to open the DNA double helix allowing RNA synthesis.

Figure 3-3: Initiation, elongation and, termination of transcription.

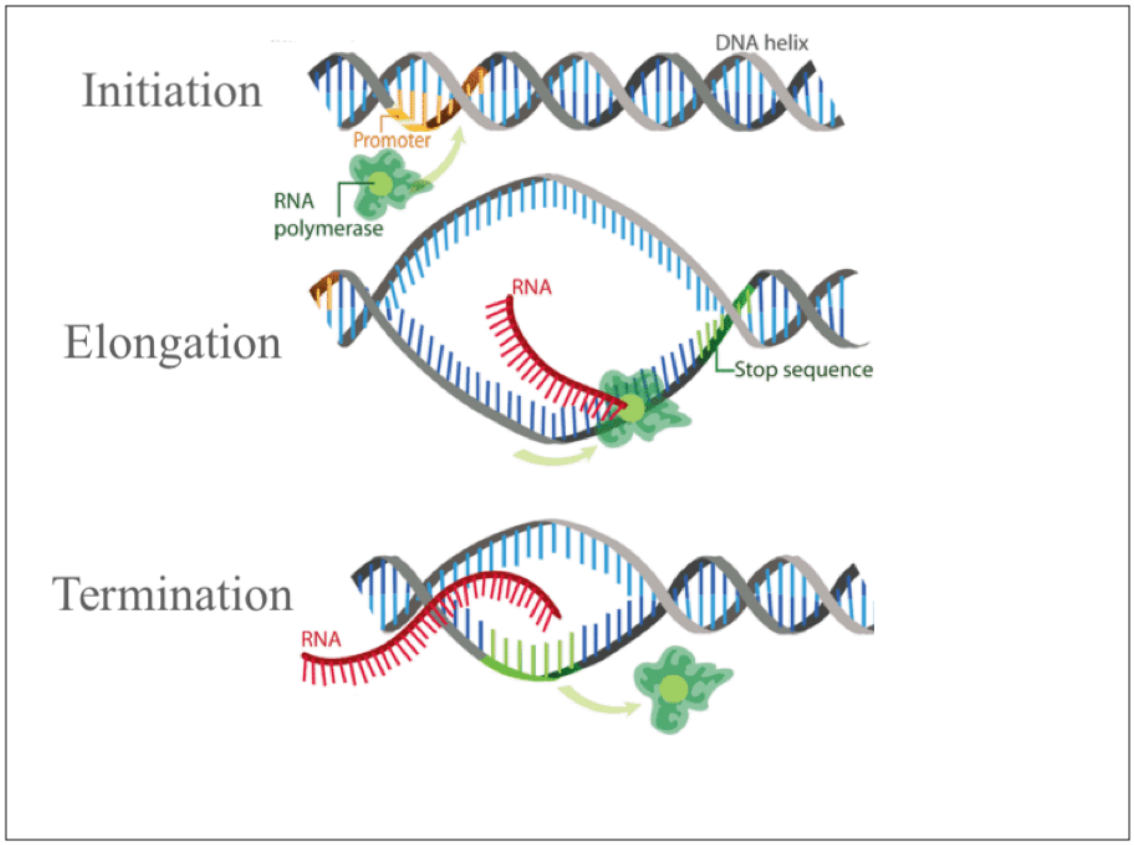


Table 3-1: Functions of the 12 RNAP subunits in archaea (Werner, 2007).

RNAP subunit	Function	Basal factor interactions
A B	DNA loading, stability of elongation complex.	TFE, Spt4/5
H A	Downstream duplex DNA binding	TFS domain II
F/E	Open complex formation, RNAP binding	TFE
A	RNAP recruitment to promoter	TFB
A	NTP entry, backtracked RNA exit	TFS domain III
A/B	Mg binding, polymerization and nucleolysis	TFS domain III
S	Substrate loading, catalysis, translocation	TFS, TFB
A	Nucleic acid strand handling	Spt4/5

Once RNAP is attached and DNA is unwound, the elongation process takes place. During it, RNAP changes its initiation subunits to elongation subunits. The two RNAP subunits that had been experimentally tested to function during archaea transcription are Spt4 and Spt5 (Gehring *et al*, 2016). Then, RNAP reads the DNA template strand and

adds nucleotides to the 3' end of the growing chain (Figure 3-3 – Elongation). It is interesting to point that the RNA formation varies among organisms. While *E. coli* has ~60 nucleotides (Vogel and Jansen, 1994) added per second, the model eukaryotic RNAPII varies between 1.7 – 33 nucleotides per second (Clancy, 2008). These two organisms thrive at 37°C. No measurements have been made upon archaea sequences so far.

The lack of cell nucleus, which capsules the genetic material, removes the presence of a physical barrier for archaeal cells. Hence, the genetic material has firstly to be coded into a messenger RNA (mRNA), so it can leave the nucleon. Only after this, ribosomes will start the translation process. On the other hand, nucleus-lacking cells are submitted through a process defined as coupled transcription and translation, where these two process happen simultaneously. There is still a lack of studies regarding the translation occurring when mRNA is not fully formed. However, evidences show that archaea make use of exploitation in regulatory mechanisms due to their limited genome size (French *et* al, 2007).

Finally, the termination process occurs when RNA is synthesized and RNAP is released. This process is characterized by RNAP failing to maintain contact with the DNA (Figure 3-3 – Termination). The loss of stability between DNA-RNA is explained many factors, including different sets of nucleotides that are more prevalent during termination (Gehring *et al*, 2016). The RNAP subunit involved in termination is Spt5 (Santangelo and Reeve, 2006).

One important remark regarding the transcription process in prokaryotic beings is the way that mRNA synthetization and its translation occurs. Indeed, the term prokaryotic has arisen many misconceptions. This term is not treated anymore as a synonym for bacteria (Pace, 2006).

3.4 ARCHAEAL PROMOTER ELEMENTS

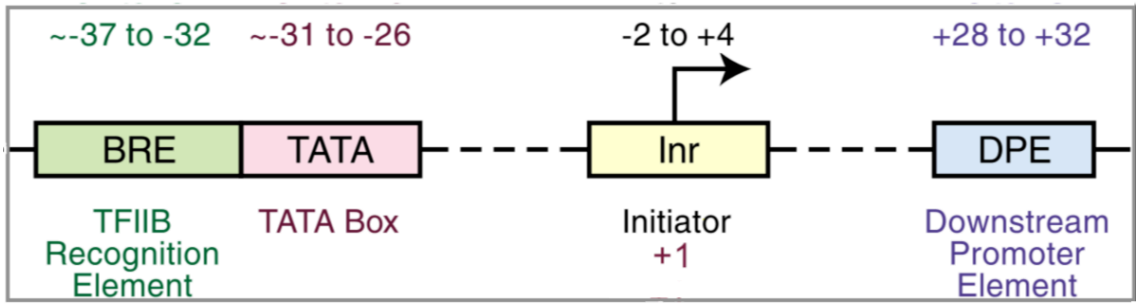
Key elements to convey the RNA transcription are the promoter sequences, they are evenly found in organisms sprung across the domains of life. These DNA segments are located upstream the Transcription Start Site (TSS), or +1. The promoter element is responsible for mediating the interaction between RNAP and DNA (Krebs *et* al. 2017). The three domains of life present uniqueness in their transcription machinery (Gehring *et* al. 2016). The smaller genome of archaea and bacteria requires a more complex orchestration of regulators. Indeed, the number of open-reading frames (ORFs) in some

archaeal genomes tends to be more significant than the number of promoters itself (Kooning and Wolf, 2008).

A promoter sequence will forward a conservation of its base-pairs (Yella and Bansal, 2017). In this sense, the start of transcription in archaea starts with a TATA-box being recognized by the basal transcription factor TBP protein at approximately 30 base pairs (bp) upstream the TSS. This machinery is no different when comparing archaea and eukaryotic organisms (Blombach and Grohman, 2017). When TBP has attached the archaeon promoter's TATA, there is the recruitment of a second basal transcription factor, a Transcription Factor B (TFB), which helps in the stabilization of TBP by saddling it in two opposing extremities. Alongside the TATA-box (i.e. a set of wTTATwww nucleotides), there is a region named B-recognition element (BRE) and a Proximal Promoter Element (PPE), which are shorter sets of nucleotides composed by ssnAA and TAC, respectively (Martinez *et* al, 2017). These two transcription factors acting together are sufficient to start archaeal transcription in a way that TBP recognizes the TATA-box and recruits TFB. A third transcription factor (Transcription Factor E, which has eukaryotic homology) has been reported to aid the PIC formation in archaea. However, it is not clear if its function is associated to transcription initiation or other transcription activities. (Gehring *et* al, 2016).

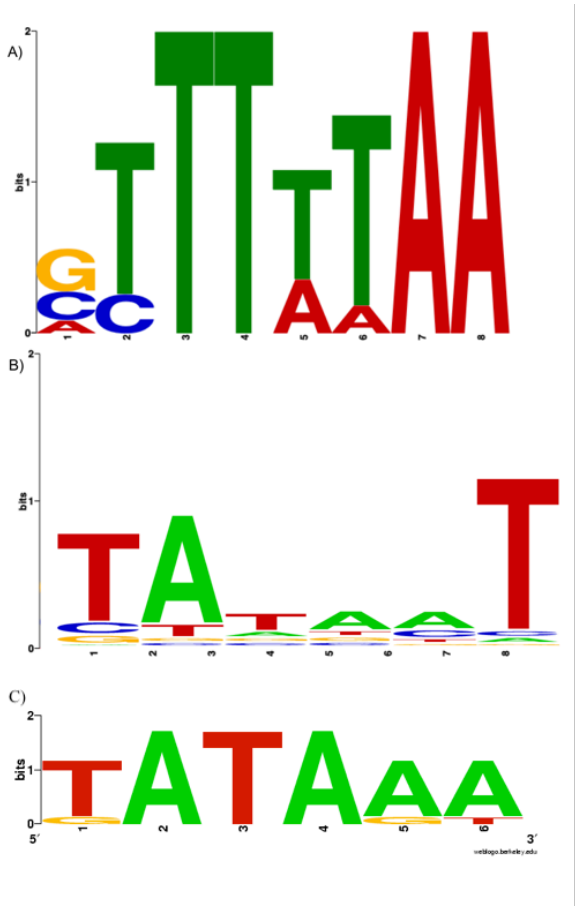
What is clear is the transcription machinery being shared between archaea and eukaryotes. Indeed, the archaea is known to have a simplified version of its eukaryotic counterpart (Gehring *et* al. 2016). What has been reported to encompass the core promoter region in eukaryotes is a sequence containing approximately 100 nucleotides (50 upstream to 50 downstream). This sequence transcribes all protein-coding genes (Haberle and Stark, 2018), defined as the core promoter. In Figure 3-4, a scheme of the conserved sites necessary to start transcription in archaea is depicted: TATA-box and BRE. These regions encompass the core promoter element.

Figure 3-4: Basal transcription factors in of the archaeal core promoter (adapted from Gehring et al. 2016; Haberle and Stark, 2018).



Despite the organism being considered, there is some sort of conservation found in this cis-regulatory element (Putta and Mitra, 2010; De Avila e Silva *et* al, 2011; Gehring *et* al, 2016). As archaea and eukarya share similarities in their promoters (Olsen and Woese, 1996), in the following, they are discussed together. The most-conserved area found in the archaeal-eukaryotic promoter element is found at -27/-28 bps upstream the TSS (Haberle-Stark, 2018; whilst in bacteria, the conservation is located at -10 bps upstream the TSS (Llórens-Rico *et* al, 2015). In Figure 3-5, there is an eyeshot of the most-conserved motifs found in the promoters of archaea, eukarya, and bacteria. In addition, bacteria possess another conserved spot in the -35 region (Bi *et* al, 1997), while archaea/eukaryotes retain the BRE element found directly upstream the TATA-box (Lagrange *et* al, 1998).

Figure 3-5: The most-conserved cis-regulatory elements in: A) archaea, B) bacteria, and C) eukarya (Adapted from Crooks et al, 2004; Jager et al, 2014).



An overview of Figure 3-5 shows that both segments in the tree of life have an AT preference in their most-conserved areas found in the promoter. Moreover, the promoter has been identified as an AT rich locale (Akan and Deloukas, 2008; Meysman *et* al, 2014). The AT richness reported in the promoter region is a key feature to be captured by techniques that seek to inflict gene variance and identify promoter regions (Ning *et* al, 2014). Therefore, the conservation found in the promoter regions might yield in significant differences, especially when coded into numeric parameters (Ryasik *et* al, 2018) that resemble the primary structure of the DNA molecule. The features discussed in Section 3.5 are tuned with information that will benefit from the unique aspect promoter elements own (Fouqueau *et* al, 2018).

3.5 DNA CONVERSION IN STRUCTURAL PARAMETERS

It has been reported that there is a growing interest in well-annotated genes. Computer processing capacity has been growing substantially, hence, the workload has shifted to inputs that have been curated. Only then, a computer analysis can deliver better insights on reliable data (Koonin and Galparin, 2010; de Avila e Silva *et* al, 2020).

The number of experimentally identified promoters is constantly growing. Simultaneously, more robust analyses are required to deploy more information gain (Yella and Bansal, 2017). The nucleotide (textual) analysis of a given DNA/RNA sequence might yield information gathered in a way that the genetic alphabet is highly limited and the entropy observed in a four-symbols system is deficient (Shannon, 1948).

In this sense, there are options to approach a DNA segment and extract valuable information of it. A satisfactory way to codify information is by doing it into numeric attributes (Shannon, 1948), an "alphabet" with many possibilities. It is possible to find many intrinsic characteristics of the DNA into the literature; indeed, more than 100 numeric attributes have been reported (Friedel *et* al, 2009). In Table 3-2, a list of conformational and physicochemical features of the DNA strand is displayed. The numeric criterion for the selection of these parameters is the presence of studies that employed the given feature in order to identify, classify or characterize specific genomic regions (i.e. reported differentiation found between promoters and control sequences). The original authors that performed the calculations as long as the year of publication are also included.

Table 3-2: List of physicochemical and conformational DNA features.

DNA feature	**Type**	**Differentiation reported in**	**Original authors**
DNA free energy	Physicochemical	Kanhere and Bansal, 2005 Rangannan and Bansal, 2007	SantaLucia and Hicks, 2004
Enthalpy contribution	Physicochemical	Privalov and Crane-Robinson, 2018	SantaLucia and Hicks, 2004

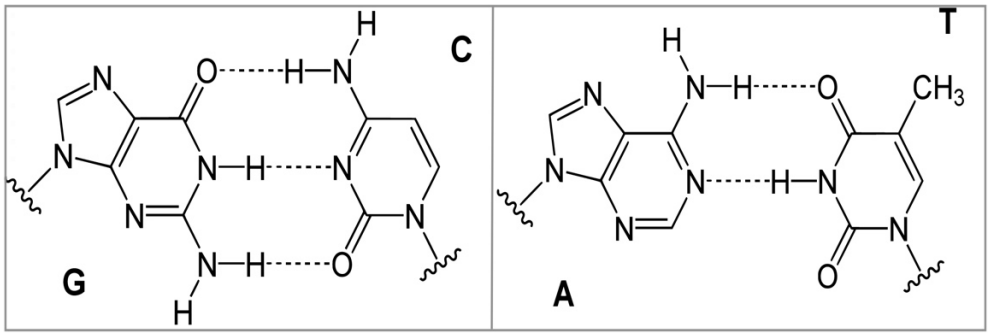
DNA bending	Conformational	Ozoline *et* al, 1999 Meysman *et* al, 2014	Karas *et* al, 1996
Intrinsic curvature	Conformational	Yella and Bansal, 2017	
Base-pair stacking	Physicochemical	Meysman *et* al, 2014 Zhang *et* al, 2015	Ornstein *et* al, 1978
SSID	Physicochemical	Wang *et* al, 2006	Wang *et* al, 2006

In the following subsections, each one of the features in Table 3-2 will be explored.

3.5.1 DNA DUPLEX STABILITY (DDS)

The stability values presented by the DNA molecule directly rely on the nucleotide sequence. When there is a bond between adenine and thymine (A, T), the quantity of hydrogen bonds is found to be two. On the other hand, when the nucleic acids turn to be cytosine and guanine (C, G), the number of bonds raises to three, as Figure 3-6 depicts. During the transcription process, there is a formation of an open complex, in other words, the DNA strand needs to be broken so its nucleotides are read by RNAP and synthesized into a mRNA molecule. This opening process involves energy being spent. For this process to be energetically viable, it is reasonable that the less stable DNA segment be broken, in that case, bonds between AT, which present a weaker chemical bond (Kanhere and Bansal, 2005; De Avila e Silva *et* al, 2014).

Figure 3-6: Hydrogen bonds between AT and CG.



There has been reports of DDS being employed as a distinguishing feature between promoters and other genomic sequences. Kanhere and Bansal (2005), de Avila e Silva *et* al (2014), and Yella and Bansal (2017) have, somehow, employed a

parametrization of eukaryotic and/or bacterial promoter sequences and were able to perceive differences in the resulting signal. The three branches in the three of life have differences found in their transcription apparatus and, consequently, their promoters (Basehoar, 2004; Werner, 2007; Blombach and Grohman, 2017). For instance, archaeal and eukaryal promoters both present an AT rich segment located directly upstream the TSS. This phenomenon is directly associated with the need to constantly open the DNA that was previously discussed. There has been reports stating the AT richness of promoters in the two domains (French *et* al, 2007; Yella and Bansal, 2017).

In terms of evolutionary biology, bacteria are known for, probably, being closer to the LUCA than archaea/eukarya (Gupta, 1998). Indeed, the transcriptional instruments of the three domains seem to follow a direct order: bacteria, archaea and, eukarya; where the first are the least exquisite and the last the most (Gribaldo and Brochier-Armanet, 2006). In this concern, there is homology of the upstream AT rich are found in eukaryotes/archaea into bacteria, a region known as the Pribnow-box (Pribnow, 1975).

Therefore, the configuration of a promoter element tends to have a leaning towards AT nucleotides, turning the promoter as a less-stable site.

3.5.2 ENTHALPY CONTRIBUTION

Over the past years, there has been efforts to visualize a living organism as a machine so our surroundings can be better understood (Davies *et* al, 2013). However, it is a simple reduction to bring cells and man-made machines to the same level (Rosen, 1991). The same author (Rosen, 1991) mainly states that biology gave origin to physics, not conversely. Thus being, the physics are a special case of biology, and following this premise, it can be stated that living-organisms and machines essentially differ. Notwithstanding, a cell performs a series of processes in order to maintain its functions and it is powered by a very important macromolecule: adenosine triphosphate (ATP) (Krebs *et* al, 2017). ATP is the main energy source for most complex biological processes and it is produced in the cell's powerhouse: the mitochondrion. This organelle, which may have once been another organism (Sagan, 1967), provides up to the weight of a human body for it to fulfil its energy requirements. When internal processes are happening (e.g. cellular division, transport, motility), they consume ATP and, consequently, exchange heat with their surroundings. Herzel *et* al., (1994) stated four moments where the heat variation can be measured inside the cell: *(i)* chemical bonds

leading to cell aggregation; *(ii)* mass transport in and outside the cell; *(iii)* heat generation caused by the metabolism and; *(iv)* information stored in the genetic code.

The recognition and binding of TPB and TFB to conserved sites in the promoters cause a change in enthalpy levels within the cell (Privalov and Crane-Robinson, 2017). The fact that promoters tend to be AT rich also contributes to the enthalpy levels found in these regulatory regions in a way that the necessary enthalpy contribution for a system to thermodynamically stabilize is rather different between ATs and CGs (Privalov and Crane-Robinson, 2017). While the first takes 7.4 ± 0.7 kcal/mol-bp^{-1}, the second requires a higher value, up to 10.75 ± 1.43 kcal/mol-bp^{-1}. In addition, Marmur and Doty (1962) stated the thermostability of the cell is impacted by the extra hydrogen bond GCs possess. Therefore, this calorimetric measurement of DNA molecule is presented as a possible identifier of specific intergenic regions.

3.5.3 DNA INTRINSIC BENDING

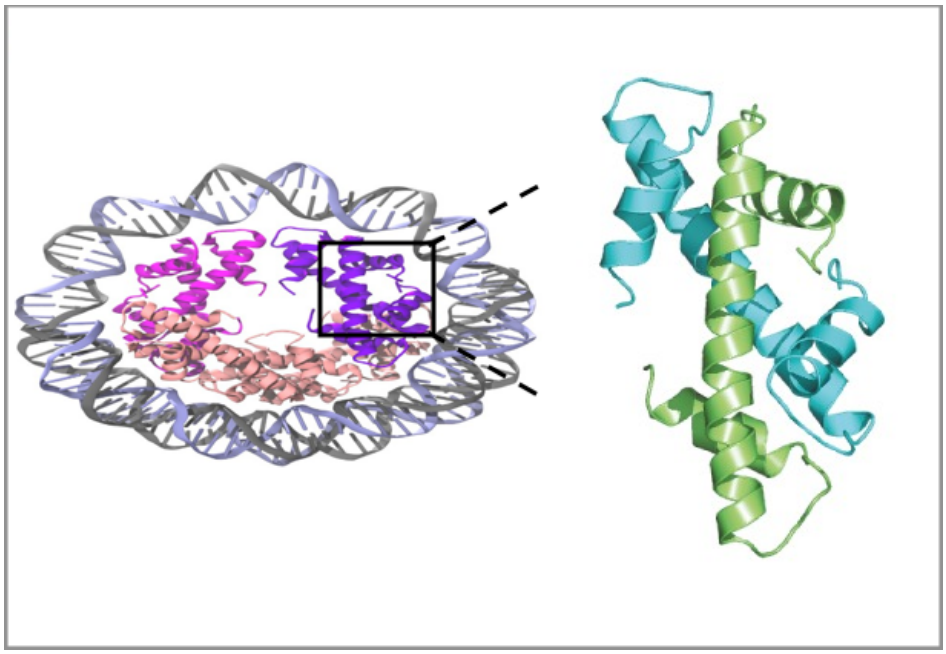
The genetic information is found in groups of genes stored in the long strands of DNA of any living organism. Complex organisms, such as eukaryotes have an extension of information that they need to use proteins called histones to help manage and organize the cell. In eukaryotes, a group of eight histones will form a structure namely nucleosome. When genetic information needs to be accessed, the eukaryotic organism will employ certain proteins that liberates DNA from the histone, turning the genetic material accessible to other proteins to act. However, the histone configuration in archaea differs; it does not form nucleosomes but coils the DNA into a structure defined as "slinky DNA" (Bowerman *et* al, 2021, Henneman *et* al, 2018). Figure 3-7 shows the way histone proteins bind the archaeal DNA.

The slinky DNA bent by the archaeal histone is a more flexible structure, i.e. the information can be easily accessed (in comparison to eukaryotes). Additionally, the archaeal cell does not need additional proteins to release DNA from its histones. This agrees with the need of the right nutrients being expressed at the right time, granting the archaeal cell survivability as well as the reduced genome of single-cell organisms (Bowerman *et* al, 2021; Browning and Busby, 2016).

There are studies that support the hypothesis that archaea are early descendants from eukaryotes (Spang *et* al, 2018; Brunk and Martin, 2019). Hence, the comprehension on the evolution of the 'minimalist' DNA organization on archaea might shed light on how the eukaryotic cell sprung.

In an attempt to optimize the DNA access and, consequently, the transcription, the RNAP enzyme needs to reach the segment to be coded, in an attempt to quantify the way DNA is bent, a pattern has been reported (Karas *et* al, 1996). Moreover, Dlakic *et* al, (2005) found bending peaks every 10 base-pairs (bps) with stronger bending every 20 bps intervals. Study models were formulated around the three-nucleotide model and their positioning (Satchwell *et* al, 1986). The position of certain trinucleotides might affect the DNA bending away or toward the nucleosome (Pedersen *et* al, 1998). Moreover, the conservation of certain nucleotides (e.g. archaeal TATA-boxes) might affect the DNA bending itself in a different angle (Leonard *et* al, 1997), especially those that are affected by certain proteins such as TPB and TFB. Archaea holds a transcriptional architecture similar to the organisms previously cited (Mattiroli *et* al, 2017), its TATA-box affects the DNA bending in a way that TA dinucleotides will angle the DNA in 6.74°, the most impactful angle bending from the 16 possible dinucleotide combinations (Karas *et* al, 1996).

Figure 3-7: The structure of an archaeal histone (Henneman et al, 2018).



3.5.4 INTRINSIC CURVATURE

The DNA structure is widely known as a double helix (Krebs *et* al, 2017). However, linear trajectories are unlikely to happen in the chaotic nature of things (Buchanan, 2011). The DNA molecule often suffers from a curved and twisted aspect since it is tightly packed to form its final structure (Jauregui *et* al, 2003). Notwithstanding,

it makes sense for these twists to follow any sort of pattern (Swanepoel, 2007). These rules might assist bioinformaticians to understand better the misty patterns nature presents itself.

It is essential to have robust experiments to determine a structural feature of the DNA. Therefore, the prediction of structures by bioinformatics models is more reliant (Kanhere and Bansal, 2003). A crystal-clear representative of the previous statement is the curvature as a parameter to code genetic information. There are many forms to extract information of curvature values (Gohlke, 2020). Di and trinucleotide models have succeeded in capturing specific curvature signals (Kozobay-Avraham *et* al, 2004). Thus being, it is essential for the researcher to know their data in order to apply the curvature model that is able to provide the highest information gain (Kanhere and Bansal, 2003).

The intrinsic curvature calculation results in the affect certain nucleotides (*di*, *tri*, and even *penta*) have in the linearity of the double-helical DNA (Jauregui *et* al, 2003). In addition, three parameters are employed in the global curvature calculation: tilt, roll, and twist. A visual representation of these parameters is available in Figure 3-8. Even though there are different models for curvature calculation, each one presenting strengths and weaknesses (Gohlke, 2020). In Table 3-3, five models for curvature calculation are displayed. A series of A nucleotides (every 10-11 bp) configures in alterations in the global curvature of the DNA (Koo *et* al, 1986). However, the effect other nucleotides have in the curved profile has also been widely studied (Bolshoy *et* al, 1991). Curvature parameters have been linked with transcription initiation in a way that the region upstream the TSS is more curved (Kanhere and Bansal, 2005). Studies that aimed to perpetrate specific genomic regions based on a curvature analysis have found the promoter region as a contributor to distinctive curvature profile (Olivares-Zavaleta *et* al, 2006; Kumar *et* al, 2016). Therefore, due to the particularities of this feature, it might be employed to capture genetic variance.

Figure 3-8: Parameters for calculating DNA curvature (Ghorbani and Mohammad-Rafiee, 2011).

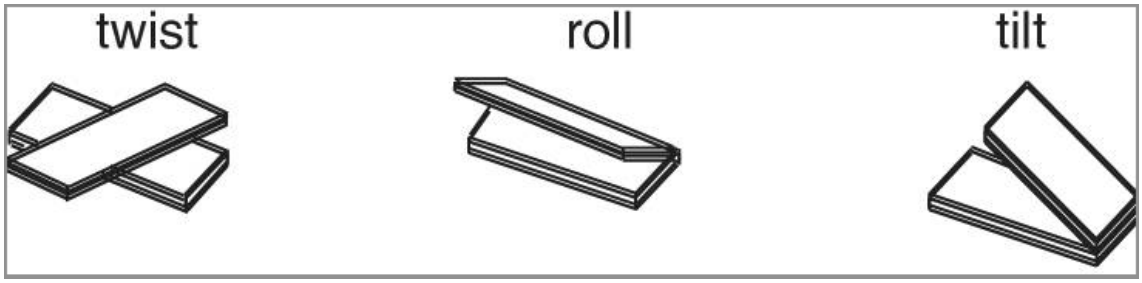


Table 3-3: Different models for DNA curvature calculation.

Model	Feature	Sensitive towards	Nucleotides	Reference
AA wedge	Certain nucleotides are responsible to protrude curved aspect based on hydrophobic interactions. Favorable hydrophobic interactions to avoid atom collision distorts DNA geometry.	AA	*Di*	Jernigan et al, (1987)
BMHT	Measurements made upon 16 dinucleotides (instead of AAs from AA wedge).	AT	*Di*, *penta*	Bolshoy *et* al, (1991)
CS	Based on mean values of dinucleotides obtained from B-DNA crystal structure. AA and TT duplexes don't own large roll/tilt values.	16 dinucleotides	*Di*	Nagaich *et* al, (1994)
Calladine	Repeated sequences have different gel migration speed when compared to random DNA sequences of the same length. This means that each dinucleotide has a different angular configuration specified by its values of twist and roll, this concept has two forms: a DNA rod or a superhelical structure, which the diameter is larger than a DNA rod.	16 dinucleotides	*Di*	Calladine *et* al, (1988)
Nucleotide Positioning	Certain triplets prefer a higher preference for being positioned in the DNA molecule.	64 triplets	*Tri*	Satchwell *et* al, (1986)

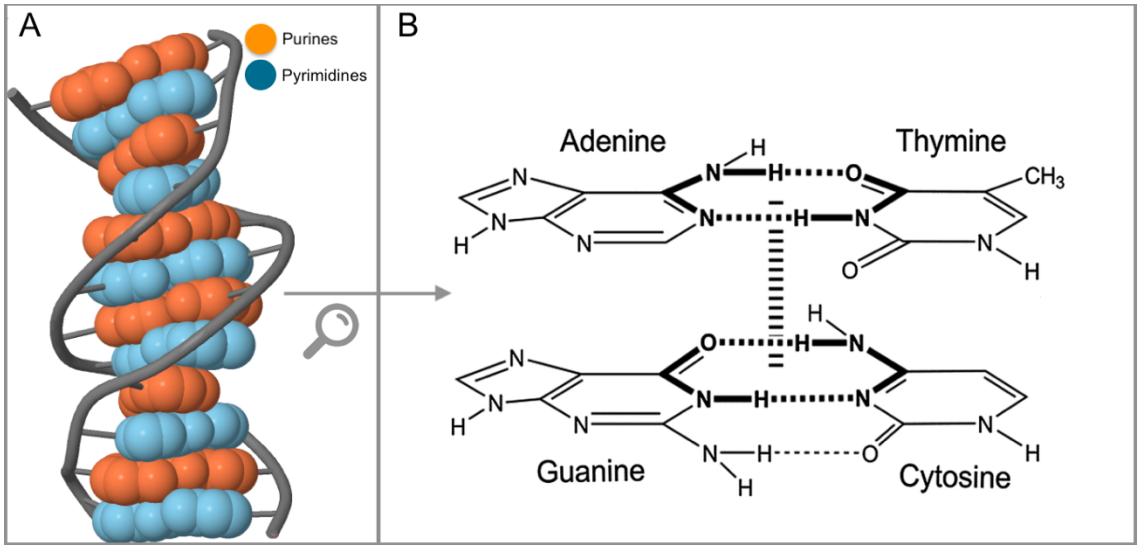
3.5.5 BASE-PAIR STACKING

Nucleic bases are arranged in a three-dimensional structure. Bases of the two strands are paired to form bps that stack along the helical axis (Krebs *et* al, 2017). The major components of this forces are hydrophobic and van der Waals. The stacking interaction between two adjacent base-pairs is an essential force component responsible for DNA stabilization and gene regulation (Zhang *et* al, 2015). One major component in order to understand the nuclear details of the gene expression is the thermodynamic stability of the double-stranded DNA. This stability is determined through interactions between bps that, in this particular case, are stacked among themselves (Häse and Zacharias, 2016).

The extent of the stacking interaction is determined by the sequence itself, where the hierarchy always follows: purine-purine; purine-pyrimidine; pyrimidine-pyrimidine (Friedman and Honig, 1995). Hence, stacking forces found in GCs contribute more to the stability of the molecule than ATs, this indicates that de degree of stabilization depends on the DNA sequence. There is also a cost involved in unstacking different bps. Zhang *et* al (2015) reported that ATs cost 14 picoNewton (pN) to unstack whilst GCs would cost 20 pN. Figure 3-9's panel A depicts the nature of stacked base-pairs, which when zoomed to panel B (this Figure ought to be visualized vertically not horizontally) shows stacking interactions between purines in the left side and pyrimidines in the right side. The stacking force is depicted in the middle of the panel and in this case, there is an optimal stabilization due to the components of it. The forces involved in this interaction would increase if pyrimidines were involved (Friedman and Honig, 1995).

Once again, the distinctive status of the promoter region when compared to other genomic areas is able to be coded into structural attributes such as base-pair stacking and be used to distinguish promoters (Ornstein *et* al, 1987; Meysman *et* al, 2014).

Figure 3-9: Stacking forces that drive DNA stabilization (Adapted from Friedman and Honing, 1995; Yakovchuk et al, 2006).



3.5.6 STRESS INDUCED DNA DUPLEX DESTABILIZATION (SSID)

The transcription initiation process involves DNA being untwisted (Krebs *et* al, 2017). This denaturation process (also referred as unwinding) must be highly controlled to maintain the energy at balance. This feature of the DNA is configured by relaxation in the hydrogen bonds between its base-pairs, which in constant separation decreases the use of energy needed to constantly open the DNA strand (Wang *et* al, 2004). This unwinding is then perceived in the super helical structure of the DNA, which is not related to the primary genetic structure. In this process, there is a difference between the energy spent in separating the strands to form an open complex, with the specific base pairs and the benefitted energy from the fractional relaxation in the super helical stress. It provides energy to control the SIDD process and for the DNA strand to remain open during the process when the mRNA is being written. It is known that in E. coli genome, the promoter sequences present a higher SIDD level. Some of the non-coding regions containing promoters are unstable, while coding regions are more stable under the stress imposed by negative super helical value. The variations in the super helical level in a promoter can show several effects in final product coded by the gene, one of them is the SIDD variation (Wang and Benham, 2006).

This feature has been employed as a reliable distinguisher between intergenic regions in *E. coli* (Wang and Benham, 2006) in a way that the SSID levels differ from promoters, intergenic regions, and coding regions. This enables the capacity of employing

SSID as a mean to code genetic information seeking to obtain specific signals along the DNA strand (De Avila e Silva *et* al, 2011).

3.6 PROMOTER CLASSIFICATION TOOLS

Biological data is characterized by huge amounts of data (Han and Liu, 2019); data that would need a lifetime to be analyzed and processed by humans (Huang *et* al, 2006). For instance, when new genes are identified, hypotheses need to be extensively analyzed to be either proven or rejected. As a solution, the employment of Artificial Intelligence (AI) techniques has been a reliable way to speed up the process (Conrad and Gerlich, 2010) of validating the findings made upon biological data (Mishra *et* al, 2020). It is important to state that experimental approaches are still essential for conducting reliable molecular experiments, among others.

A crucial task in enhancing gene regulation and optimizing biological processes is the identification of promoter sequences. For example, the broader comprehension of promoter activity could, in theory, grant the researchers the full control over starting and halting the expression of certain genes (Kernan *et* al, 2017). However, the prediction/identification of promoter sequences is a harsher task than it might look (Mishra *et* al, 2020). In Section 3.4, several conserved aspects of the promoter site have been presented, but the lone presence of these motifs is not enough to protrude a reliable promoter scanning. To corroborate this, reports have been gathered about: *(i)* transcription initiating in non-fixed locations over the genome (Wade and Grainger, 2014); *(ii)* an apparent overlapping of the promoter element into the TSS (Kanhere and Bansal, 2005); *(iii)* high gene density (Mishra *et* al, 2020), specially in organisms with a short genome such as archaea.

Thus, the promoter prediction task is not a simple thing. Indeed, the performance of different approaches under the same training dataset has shown to be quite varied (Klauck *et* al, 2020). Indeed, many of these methods present a way to code genetic information into numeric parameters (Ryasik *et* al, 2018) and use it as input of machine learning approaches. Moreover, promoter-specific tools are sprung among the three domains with the exception of the shortage on archaea. Table 3-4 illustrates some of the tools that have been created to predict promoters, as well as the classification technique implemented, including machine learning approaches. From this, there is a clear preference by artificial neural networks due to its classificatory nature. In the next section, this machine learning routine will be further explored.

Table 3-4: Comparison of promoter identification tools.

Tool	Organism	How does it find promoters?	Does it use machine learning?	Overall precision	Reference
DeePromoter	Human Mouse	A combination of ANN and long short-term memory recognizes TATA-less and TATA-containing promoters.	Yes, ANNs	Human TATA = 93% Human non-TATA = 97% Mouse TATA = 92% Mouse non-TATA = 91%	Oubounyt *et* al (2019)
Prompredict	*E. coli*	It calculates the stability difference between coding and promoter regions through a division of a sequence in overlapping windows of 15 nucleotides.	No	90% sensibility	Yella *et* al (2018)
bTSS Finder	*E. coli* Cyanobacteria	Presence and distance of promoter elements through the presence of physico-chemical features.	Yes, ANNs	*E. coli* = 89.22% Cyanobacteria = 79.26%	Shamuradov *et* al (2017)
BacPP	*E. coli*	Neural networks are trained with promoter and non-promoter examples; then biological rules are extracted from the learning technique.	Yes, ANNs	σ24 = 86.9%, σ28 = 92.8%, σ32 = 91.5%, σ38 = 89.3%, σ54 = 97.0%, σ70 = 83.6%	De Avila e Silva *et* al (2011)
CNN promoter	Human Rat *B. subtilis* *A. thaliana* *E. coli*	Analyzes promoter sequences from both prokaryotes and eukaryotes. The classification is made via coevolutionary ANN method and a deep learning approach.	Yes, ANNs	90% sensibility 96% specificity 0.84 correlation coefficient	Bedoya and Bustamante (2011)
PromH	Human	Statistically identifies conserved regions in human promoters.	No	TATA-boxes = 70%	Solovyev and Shamuradov (2003)
Promoter 2.0	Pol II promoters	A combination of elements similar to neural networks and genetic algorithms to recognize a set of	Yes, ANNs	Correlation coefficient = 0.63	Knudsen (1999)

discrete sub patterns with variable separation

3.6.1 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

The life sciences have a lot to gain when backed upon reliable techniques that allow its hypothesis to be fully stressed (Conrad and Gerlich, 2010). Thus being, the following section contextualizes two major – and confusing – misconceptions upon this chunk of computer science: Artificial Intelligence (AI) and Machine Learning (ML). It is essential to first define, in an epistemological approach, what learning is (Russel and Norvig, 2012).

The raw definition of learning has been a challenge to many psychologists over the past centuries (De Houwer *et* al, 2014). One widely accepted definition states the process of learning is defined as the acquiring of new understanding, knowledge, behavior, skills, values, attitudes, and preferences. In other words, learning might be defined as a change in behavior due to experience (e.g. being burned by a hot stove) (Lachman, 1997; Gross, 2015). This definition might be easily transmitted to a machine environment (Russel and Norvig, 2012).

It is not easy to find an area of learning with such a deep multidisciplinary focus than AI and ML. These disciplines contain mathematics, electrical engineering, statistics, signal processing, computer science, among others (Joshi, 2020). That is one of the reasons AI and ML are two different parts of computer science. A fine way to think of AI was brought up by Alan Turing, who defined the area as: "if there is a machine behind a curtain and a human is interacting with it and the human feels like he/she is interacting with another human, then the machine is artificially intelligent". Therefore, the objective of AI is not the building of such an incredible machine that is able to solve any problem in this universe but rather imitating human behavior, even considering its flaws. (Christian, 2013; Joshi, 2020). On the other hand, ML refers to a computer program that is able to produce a behavior that was not intended by its creators. This behavior performed by the machine is defined as learning. Next, this section will discuss the nature of AI/ML techniques around the concept of supervised learning.

3.6.2 ARTIFICIAL NEURAL NETWORKS

Supervised learning refers to prior knowledge about what the output should be. Furthermore, an ML system might be considered as a black box that is able to produce outputs based on the inputs. If there is a set of historical data that contains outputs for some given inputs, this type of learning is called supervised. An application of supervised ML is classification. This method consists in assigning a label (target variable) to a class. This initial step of using previous knowledge to learn more about the dataset is called training. Next, the testing of the model created with the training is performed (Russel and Norvig, 2012; Joshi, 2020).

There are many algorithms that follow a classificatory nature, such as Artificial Neural Networks (ANNs), decision trees, random forest, support vector machines, among others. As previously explored in Table 3.4, ANNs are a very common approach in classifying promoter sequences. The implementation of supervised ML employed in this thesis is ANN.

3.6.2.1 THE NATURE OF ARTIFICIAL NEURAL NETWORKS

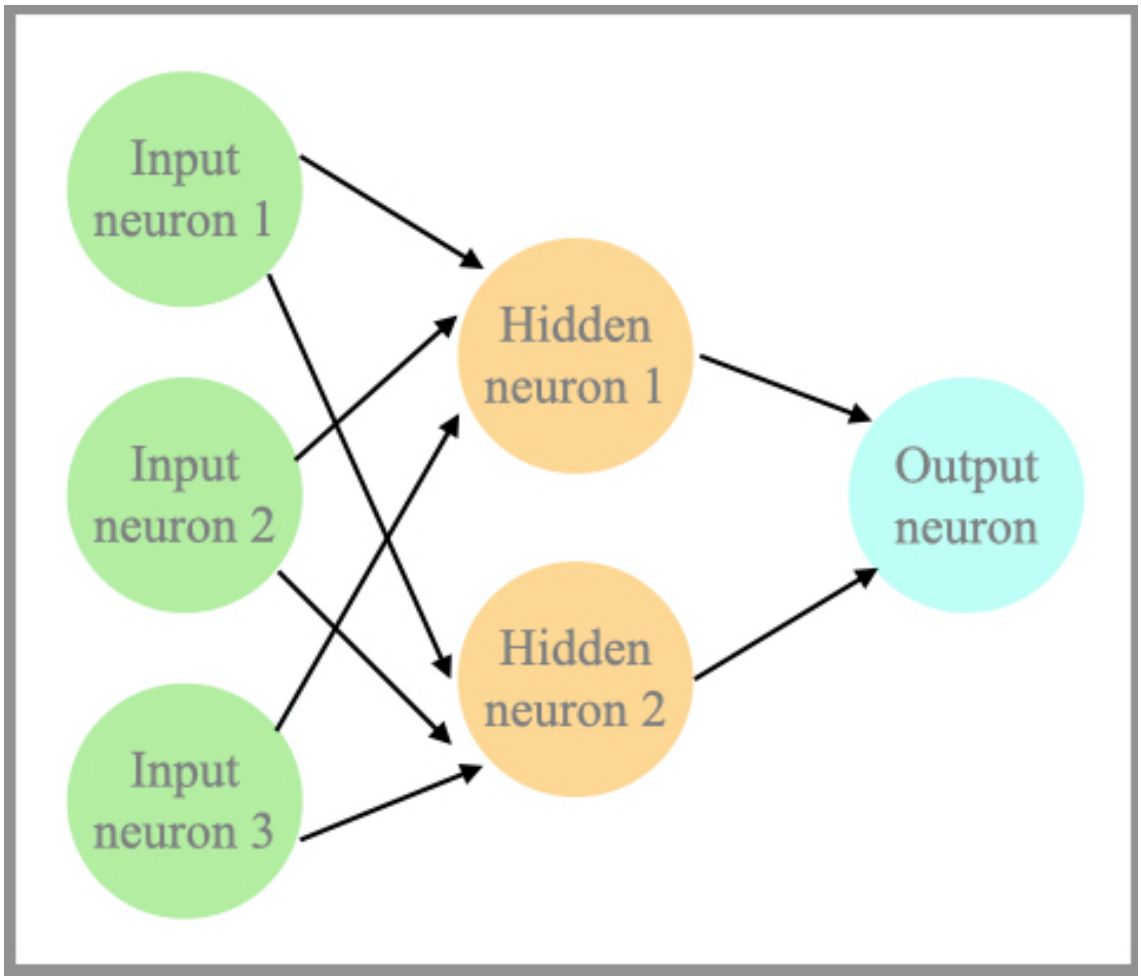
ANNs simulate the way a human brain processes information. This computational model is employed in several areas, such as engineering, biology, and economics, among others. When used in the biological sciences, ANNs might perform tasks to analyze genomic data and recognize patterns (Baldi and Brunak, 2001; De Avila e Silva and Coelho, 2020).

ANNs consist of a parallel distributed processor formed by simple units. These units aim to form knowledge upon data in a classificatory way. Moreover, an ANN consists of an *i* unit connected to a *j* unit. These units have a synaptic weight associated to them, namely W_{ij} (Mount, 2000). The connection of these units, i.e. neurons, which are divided into different layers. The neurons present in the layers, namely: input, hidden, and output have weights associated to them and these units all connected through a time span is one of the main premises of the functioning of an ANN (Russel and Norvig, 2012; De Avila e Silva and Coelho, 2020).

The ANN model depicted in Figure 3-10 is a multilayer perceptron, where all input neurons are indirectly connected to the output, which is represented by a variation of a discrete time signal. Then, the application of weights in a given neuron is only going to affect the output of this neuron, granting independence to the model. More than one

level of hidden neurons might be employed. In fact, the reason of the hidden neurons is to mediate the conversion of an input into an output (De Avila e Silva and Coelho, 2020).

Figure 3-10: Visual representation of an artificial neural network.



3.6.2.2 THE ARCHITECTURE OF ANNS

A given ANN is featured by the connection of its neurons, which consists on weights in neural connections and an activation function. Figure 3-11 traces the path an input signal is transformed into the output of the ANN model. The neurons are connected and they behave as the information processing unit which is responsible for the operation of an ANN. From Figure 3-11, three basic models of a neuronal model are identified:

I. A set if synapses or connection links, which are characterized by a weight or strength. A signal x_1 at the input of synapse j is connected to the neuron k is multiplied by the synaptic weight w_{jk}.

II. An adder for summing the input signals, which are weighted by the respective synapses of the neuron.

III. An activation function for establishing a limit to the amplitude of the output of a neuron. Typically, the normalized amplitude of a neuron is written as a closed unit interval (0, 1).

Additionally, Figure 3-11 presents an externally applied bias b_i, which function is to lower or increase the net input of the activation function. Mathematically, a neuron k is described by the following Equations 1 and 2:

$$U_k = \sum_{j=1}^{m} W_{kj} x_j \qquad (1)$$

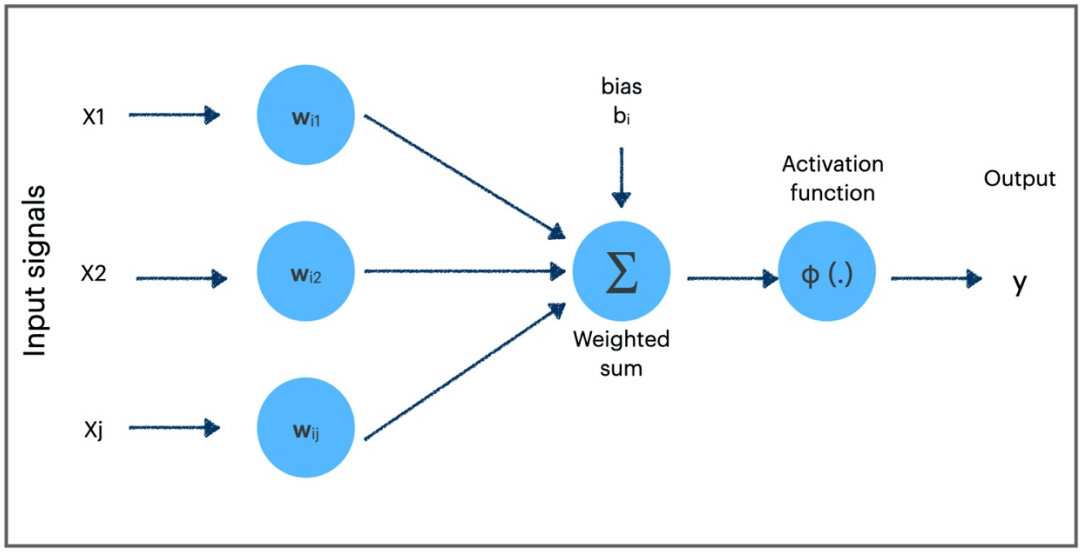
and

$$y_k = \varphi(u_k + b_k) \quad (2)$$

where $x_1, x_2, \ldots, x_m$ are the input signals; $w_{k1}, w_{k2}, \ldots, w_{km}$ are the synaptic weights of neuron k; u_k is the adder, which varies according to the input signal; b_i is the bias; φ is the activation function; and y_k is the output signal of a neuron. The use of the bias b_k transforms the output u_k of the adder as shown by Equation 3, which:

$$v_k = b_k + u_k \qquad (3)$$

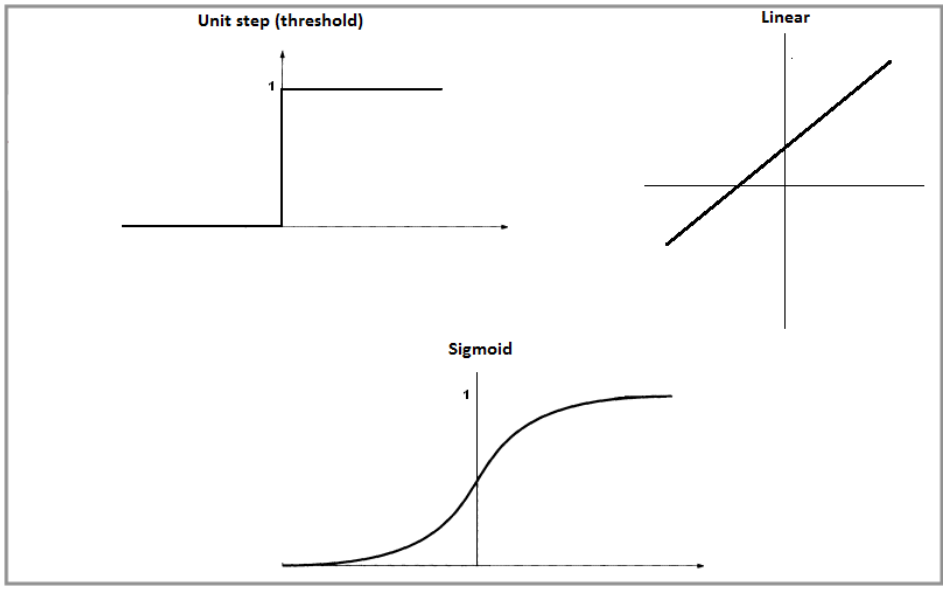
Figure 3-11: Model of an artificial neuron.



The rationale of the activation function φ is designed to convert the weighted sum of input and transform it into an output. Thus being, the behavior of the activation

function controls how well the neural network will learn the training dataset and its output will define the type of predictions the model can make (Wu, 1997; Joshi, 2020). There are different examples of activation functions, and they have to be carefully chosen considering the prediction task to be performed. Examples of activation functions are: threshold, linear, and sigmoid. The main difference between these functions are the values they can comprise when turning an input into an output. To begin with, the simpler activation function is the threshold function, these are limited due to comprise either 1 or 0 as a value. Moreover, linear functions allow multiple outputs (and not only 1s or 0s). In this, the input is multiplied by the weight of the neuron and creates an output proportional to the input. Finally, sigmoid functions are employed when there is the need to predict any normalized probability as an output, since probabilities range from 1 to 0 (Haykin, 1999; Russell and Norvig, 2012). A visual representation of a graph of each of the three functions described is found in Figure 3-12.

Figure 3-12: Activation functions of a neuron



3.6.2.3 TRAINING ANNS

Prior to their actual use, ANNs need to be trained. In other words, this process involves defining an architecture so the ANN functions well with a given set of inputs for any desired output. This happens through correct weights being applied to every unit of the network. Then, an input signal is tested according to its reaction to a specific weight. If the performance is not satisfactory – the root mean squared has not been reduced - the

training is repeated until it fits a given condition and the observed *vs* expected error is minimized (Wu, 1997), i.e. the ANN has converged.

The training of the neural network with a given set of weights through the entire training dataset is defined as an epoch. This hyper parameter grants that every single training data point had an opportunity to update its internal model parameter.

The walkthrough of all the weights in an ANN (randomly initiated) is defined as one epoch through the full training dataset. In general, after every epoch, the weights are updated. However, if an ANN is trained too much, overtraining might occur. This means that the model is able to memorize its input examples and performs well for a given dataset, being unable to generalize, i.e. the ability to handle unseen data. In order to solve this and deliver an acceptable ability to work with data different than the seen data, validation methods are employed. Examples of validation methods are *k*-fold-cross-validation, jackknife, holdout, among others. The chosen method of this study is *k*-fold-cross-validation, which consists in partitioning the data in *k* equal parts; then, *k*-1 partitions are employed to train the model and the unused k^{th} part is used in the evaluation of the ANN (Russell and Norvig, 2012; De Avila e Silva and Coelho, 2020).

3.6.2.4 LEARNING ALGORITHMS AND THE RPROP EXAMPLE

Neural networks are employed for data classification in a way that entry data will produce outcomes. For it to succeed, the observed output should be as close as possible of the desired one. This requires an external curator to interact with the model in order to minimize the observed output error. One example of error minimization function is represented by Equation 4:

$$E = \sum_{i=1}^{n} (y_i - h_w(x_i))^2 \qquad (4)$$

where, $h_w(x_i)$ is the expected output, y_i is the observed output, which is associated to a given set of parameters *w*. *n* is the number of input patterns. Additionally, an ANN might have more than one outcome unit, therefore, Equation 4 is replaced by Equation 5 in order to accommodate n_k outcomes.

$$E = \sum_{i=1}^{n_1} \sum_{j=1}^{n_k} (y_i - h_w(x_i))^2 \qquad (5)$$

Examples of learning algorithms are: perceptron (Rosenblat, 1958), backpropagation (Rumelhart *et* al, 1986) and resilient backpropagation (Rprop) (Riedmiller and Braun, 1993). The chosen algorithm in this study is the Rprop without weight backtracking because of its relative robustness and low computational cost (Riedmiller, 1994; Igel and Hüksen, 2003).

Such applications, which are based on gradient optimization, are commonly used in supervised learning models. The advantage is to control the weight update in every connection in an ANN in a way that at each iteration, the weights are updated in the opposite direction of the gradient, where the gradient itself gives the direction of the steepest (and the costliest) ascent. The cost in moving to reach the local minimum is determined the learning rate α. The gradient optimization is the process to reach the minimum of a function:

$$\theta^1 = \theta^0 - \alpha \nabla j(\theta) \quad (6)$$

where θ^1 is the next position to be reached, θ^0 is the current position, and $\alpha \nabla j$ a small step in the gradient times the direction of the fastest increase, where α is a constant of ∇, pre-determined by the learning rate, which usually varies are 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, and j is the gradient. This minimizes the oscillation and maximizes the update step-size. In each iteration, the updated weights are described by:

$$W_{ij}^{(t+1)} = W_{ij}^{(t)} + W_{ij}^{(t)} \quad (7)$$

In the Rprop algorithm, the direction of each weight update is based on the sign of the partial derivative $\partial E / \partial W_{ij}$, where each weight, i.e., Δ_{ij}, is adapted individually. One of the applications of the Rprop algorithm does not have the backtracking of its weights, being computed as:

$$\Delta W_{ij}^{(t)} = -sign\left(\frac{\partial E^{(t)}}{\partial W_{ij}}\right) \Delta_{ij}^{(t)} \quad (8)$$

the sign function returns +1 if the argument is positive or -1 if the argument is negative, else, it returns 0. The Δ_{ij} are initiated to a constant Δ_0. Each iteration of the Rprop algorithm is described by the adjustment of Δ_{ij} in Equation (9).

$$\Delta_{ij}^{(t)} = \begin{cases} \min\left(\eta^{+}\Delta_{ij}^{(t-1)}, \Delta_{max}\right) if \ \dfrac{\partial E^{(t-1)}}{\partial W_{ij}} \dfrac{\partial E^{(t)}}{\partial W_{ij}} > 0 \\ \min\left(\eta^{-}\Delta_{ij}^{(t-1)}, \Delta_{min}\right) if \ \dfrac{\partial E^{(t-1)}}{\partial W_{ij}} \dfrac{\partial E^{(t)}}{\partial W_{ij}} < 0 \\ \Delta_{ij}^{(t-1)} \ else \end{cases} \quad (9)$$

where $0 < \eta^{-} < 1 < \eta^{+}$. The updating of the weights is shown by the parameters Δ_{min} and Δ_{max}. The Equations 6 and 7 generate:

$$\frac{\partial E^{(t)}}{\partial \Delta_{ij}^{(t-1)}} = \frac{\partial E^{(t)}}{\partial w_{ij}^{(t)}} \frac{\partial w_{ij}^{(t)}}{\partial \Delta_{ij}^{(t-1)}} = -\frac{\partial E^{(t)}}{\partial W_{ij}} \, sign\left(\frac{\partial E^{(t-1)}}{\partial W_{ij}}\right) \quad (10)$$

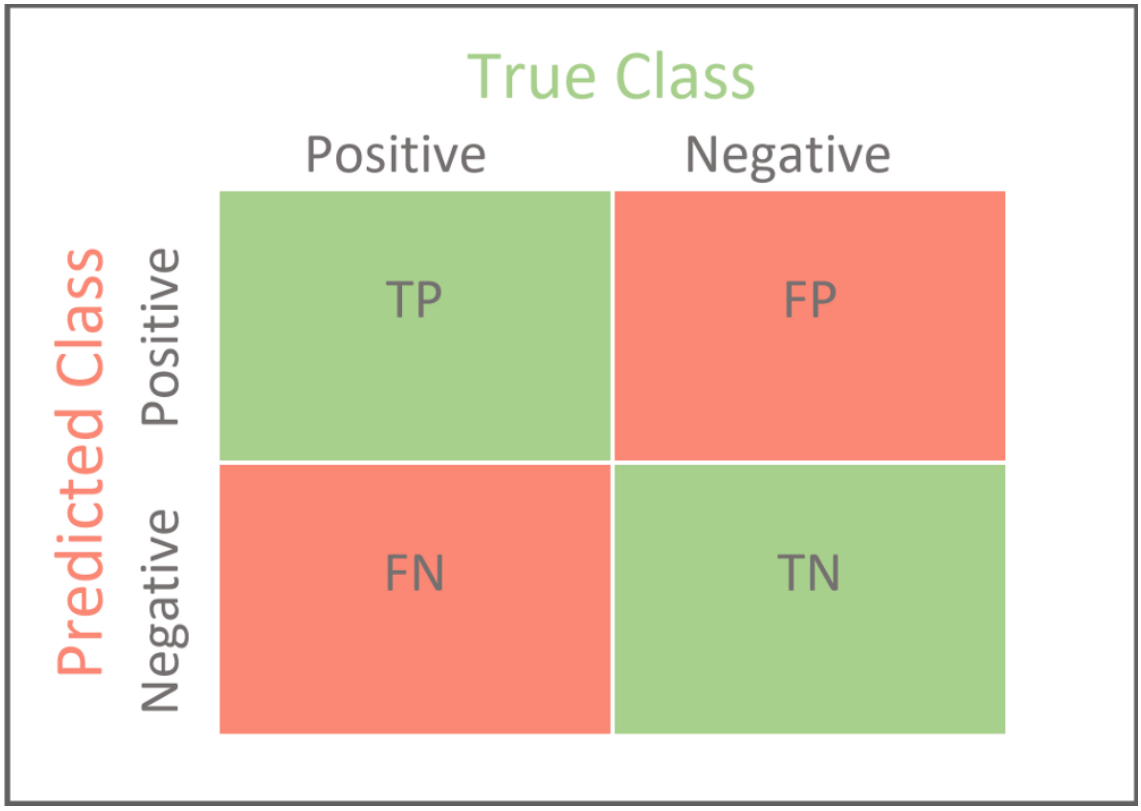
The direction for Δ_{ij} to change follows Equation 7 and fits a gradient-based optimization of the error of the network, mediated by the limits defined by Δ_{ij}.

3.6.3 EVALUATION OF MACHINE LEARNING TECHNIQUES

Machine learning techniques to have their results assessed, i.e. validated. To do so, there are performance metrics that can unveil the delicate and misty innards of machine learning algorithms. A good way to start is through a confusion matrix (Stehman, 1997). As known as error matrix, this table layout enables that supervised learning algorithms have their performance visualized (Powers, 2007).

A confusion matrix might be used to assess problems which the output contains two or more classes. The table that results in the matrix has predicted and actual values, which are employed to assess the performance of the technique. The elements that comprise a confusion matrix belong to two classes – positives and negatives, these elements are: *i*) True Positives (TPs), which are the number elements from one *c*-class correctly classified as *c*-class, if the element of *c*-class is classified as belonging to *d*-class then; *ii*) a False Negative (FN) is labeled. Now, if an element of *d*-class is classified as a member of *c*-class, thus, *iii*) a False positive (FP) is considered. Finally, if a non-*c*-class is correctly classified as not belonging to *c*-class, *iv*) a True Negative (TN) is assigned. Figure 3-13 depicts the organization of a two-class confusion matrix, where the main diagonal (i.e. the one formed by TP and TN) represents the correctly classified labels while the opposite diagonal (i.e. is represented by FP and FN) indicates the misclassification.

Figure 3-13: Confusion matrix



Once the confusion matrix is set, the performance of the ML algorithm might be assessed through:

I. Accuracy, which measures how well a 2-class classification correctly identify or exclude a class, i.e. the proportion of correct predictions (both true positives and true negatives) among the total number of observations. Its calculation is achieved through Equation 11.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (11)$$

II. Recall, which classifies how many of the actual positives are captured and labeled as TPs by the model. This metric is positively affected by high rates of true positives and low rates of false positives. Its calculation is done by Equation 12.

$$Recall = \frac{TP}{(TP + FN)} \quad (12)$$

III. Precision, which calculates how many of the classes classified as positive are actually positive. The cost of false positives is high when calculating precision through its formula, explained by Equation 13.

$$Precision = \frac{TP}{(TP + FP)} \qquad (13)$$

IV. F1 score, which makes a balance between precision and recall. This metric is used for any classifications which there is a considerable difference between precision and recall. Instead of using arithmetic mean, F1 score uses harmonic mean, being a function of two variables (precision and recall) at the same time and penalizing extreme values. Its calculation follows Equation 14.

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \qquad (14)$$

V. Specificity, which measures the rate of detecting TNs among the entire dataset. It is penalized by high numbers of false positives. Specificity is calculated through Equation 15.

$$Specificity = \frac{TN}{(TN + FP)} \qquad (15)$$

Another derivation from a confusion matrix is the Receiver Operating Characteristic (ROC) curve, which is a performance measurement for classification problems under different thresholds. Many machine learning applications such as ANN is able to predict the membership of a certain class. The decision for classifying an observation is given by a parameter namely decision threshold. By default, in ANNs whose activation function is the sigmoid, the decision threshold is set to 0.5, meaning that any outcome of the ANN equal or greater than 0.5 is predicted as a positive class. Values lesser than 0.5 are predicted as negative. However, the default decision threshold may not be the best way to represent an optimal interpretation of the data, its adjust is given by calculating ROC curves. Moreover, The ROC may also be used to calculate the Area Under the Curve (AUC), which measures the degree of separation between thresholds of the same classification method as well as a comparison of performance of different

classification techniques (e.g. linear regression vs. ANN). Higher an AUC is, the better the model predicted certain class (Joshi, 2020).

ROC curves may be calculated from any metric derived of a confusion matrix (Peres *et* al, 2015). A common calculation derives from two competing metrics: recall (as known as sensitivity), which measures the true positives rate (TPR); and false positives rate (FPR), which is estimated by 1- specificity. This ratio is calculated in a way that the greater the TPR is, the lesser the FPR will be, Equation 16 describes it.

$$FPR = 1 - \left(\frac{TN}{TN+FP}\right) \quad (16),$$

where, *FPR* measures the false positives rate in a 0-1 range.

Hence, the ROC calculation is able to provide visual identification of the relation between true and false positions in a classification model (Bewick *et* al. 2004).

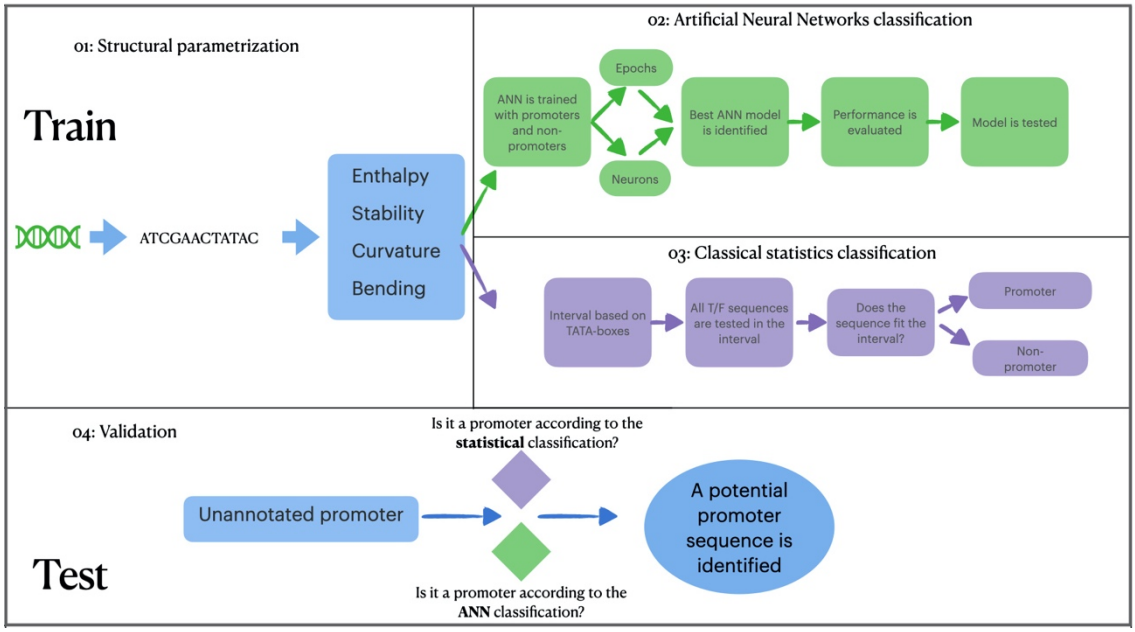
3.7 FINAL CONSIDERATIONS

Promoters are key elements in terms of controlling the gene expression within organisms in the three domains of life. The literature review conducted so far has shown how biology might benefit from state of the art computer techniques. However, specific tools that target archaeal promoters were not found until this literature review was finalized. The third domain is still a novelty in biological sciences and the availability of data from this organisms' genomes is still rising. Thus being, this work aims to employ the machine learning techniques hereby described to build a novel classificatory algorithm of archaeal promoter sequences.

4 DATASETS AND METHODS

In this chapter, an overview of the methods and datasets is described. To accomplish the objectives, genetic information is codified into the structural parameters namely enthalpy, curvature, stability, and bendability. Next, the numeric attributes brought by the previous step are used as an input for a classification through Artificial Neural Networks and an in-house statistical method for classification. The two models of classification are then validated with three different forms of control sequences, each one carrying particularities in order to stress the models. Finally, the outcome of each classificatory model is employed in identifying potential unannotated promoters in the archaeal genome. In order to describe a visual representation of the procedures, Figure 4-1 is provided.

Figure 4-1: Description of the methods.



4.1 DATASETS

The data used to conduct this research is comprised by ATCG sequences and is divided into: *i*) archaeal promoters and *ii*) control data.

4.1.1 ARCHAEAL PROMOTERS

A total of 3630 experimentally verified promoter sequences are analyzed in this study. These sequences belong to three organisms: *Haloferax volcanii* with 1340

promoters, *Sulfolobus solfataricus* with 1042 promoters and, *Thermococcus kodakarensis* with 1248 promoters. These particular genomes were selected for being model organisms in the family of Euryarchaeota (*H. volcanii* and *T. kodakarensis*) and the family of Crenarchaeota (*S. solfataricus*). Moreover, the genome of these three organisms have been completely sequenced (Wurtzel *et* al, 2009; Jäger *et* al, 2014; Babski *et* al, 2016) and have transcriptome data (RNAseq) available, which grants the possibility of retrieving experimentally validated promoter sequences. Only primary TSS data was considered due to its abundance (internal, antisense, and secondary TSSs were not considered).

Each experimentally validated promoter sequence contains 1000 nucleotides, ranging from -500 to +500, which was further sliced into a smaller region comprising the range -80 to +20. This sub-sequence is reported to contain the core promoter element (Haberle and Stark, 2018). The core promoter in archaea was observed as enough to convey archaeal transcription (Gehring *et* al, 2016) since it contains binding sites for the most important transcription factor proteins in archaea (Aptekman and Nadra, 2018). Therefore, the classification methods later explored encompassed these shorter sequences due to the high computational cost involved in working with a wider window of nucleotides that are not relevant for the promoter identification purpose. The full span (1000 nucleotides) was also considered in experiments that sought to boost the genetic variability of promoters when compared to intergenic regions.

4.1.2 CONTROL SEQUENCES

In order to provide a reliable set of control sequences, three control datasets with varied nature were constructed. The first contains the original sequences with their order rearranged through a shuffling process. The second is a more specialized shuffling process, as proposed by Oubounyt *et* al, (2019), which first divides the 100 nucleotides sequences into five blocks and then mixes the blocks. According to the authors, this method of obtaining control sequences is a reliable way to preserve consensual sites that might contribute positively or negatively in a classification method. Finally, a third level of control is added by picking the next 100 nucleotides after the window that contains the core promoter, ranging from +21 to +121. The use of these intergenic regions is able to magnify the peculiarities found specifically in promoters (Kanhere and Bansal, 2005).

In each experiment, the number of archaeal promoters equals the number of control sequences, totaling three datasets of 3630 control sequences each. These sequences were achieved through in-house Python scripts.

4.2 STRUCTURAL PARAMETRIZATION

Four DNA physical/structural assets were chosen to code genetic information into numerical attributes. The features DNA duplex stability (DDS) (SantaLucia and Hicks, 2004), enthalpy contribution (SantaLucia and Hicks, 2004), and DNA bending (Karas *et* al, 1996) rely on dinucleotides sliding windows. Following a different form to be codified, DNA intrinsic curvature (Bolshoy *et* al, 1991) presents sliding windows containing five nucleotides instead of two.

The genetic alphabet is limited. In fact, all the information living beings need to maintain their biological functions is represented in a four-letter script. According to Ryasik *et* al, (2018), when genetic content is translated into numerical attributes, much more information is gathered. The database Dinucleotide Property Database (DiProDB) (Friedel *et* al, 2009) stores information on how to numerically represent genetic information. The database has 120+ properties with unique values that might be considered. Among these entries, the properties considered in this work are found. Table 4-1 indicates DDS, enthalpy contribution, and DNA bending values for the 16 dinucleotide possible combinations of the four nucleic acids.

Table 4-1: Enthalpy, DDS, and bendability parameters for every possible dinucleotide combination.

Dinucleotide	**Enthalpy (kcal/mol-bp-1)**	**DDS (kcal/mol-bp-1)**	**DNA bendability (degrees)**
AA	-7.6	-1.00	3.07
AT	-7.2	-0.88	2.6
AC	-8.5	-1.45	2.97
AG	-8.2	-1.3	2.31
TT	-7.6	-1	3.07
TA	-7.2	-0.58	6.74
TC	-7.8	-1.28	2.51
TG	-8.4	-1.44	3.58
CC	-8	-1.28	2.16
CA	-8.5	-1.45	3.58
CT	-7.8	-1.28	2.31
CG	-10.6	-2.24	2.81

GG	-8	-1.84	2.16
GA	-8.2	-1.3	2.51
GT	-8.4	-1.44	2.97
GC	-10.6	-2.24	3.06

Moreover, the parameters considered for calculating intrinsic curvature are available in Supplementary Table S1. Due to the greater amount of possible combinations (1024) of five nucleotides window as opposing to the 16 presented in Table 4-1, the parameters of curvature were moved to Supplementary Materials.

4.3 A CLASSIFICATION BASED ON CLASSICAL STATISTICS

A statistical analysis was elaborated by following the common rule archaeal promoter sequences have: there is a strong signal in converting the binding sites of the transcription factors TBP and TFB into structural properties. The conserved aspect found only in the promoter is used to create a window containing the thresholds for a particular sequence to be classified as a promoter. First, the arithmetic mean of each promoter's TBP + TFB is calculated following Equation 16,

$$\bar{x} = \sum_{n+1}^{n} p \qquad (16)$$

where $\bar{x}$ is the mean of the summation of p, which is a vector containing the n positions where the transcription factor proteins TFB (6 nucleotides which are represented by 5 values) and TBP (3 nucleotides which are converted into 2 values) bind the DNA. Once the means are obtained, the standard deviation is calculated through Equation 17.

$$Stdev = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \qquad (17)$$

where x_i is the individual value of each p (please see Equation 24), $\bar{x}$ is the mean, also obtained from Equation 24, and n is the number of occurrences. The calculation of Equation 24 and Equation 25 was employed in the creation of Table 4-2, which is formed by the thresholds achieved by the analysis of DNA Duplex Stability of promoter sequences in three archaea against the three levels of control sequences depicted in 4.1.2.

Table 4-2: Values used to create the intervals that perform the classical statistics analysis.

	Promoters	**Standard deviation**	**Control 01**	**Control 02**	**Control 03**
H. volcanii	-8.324	± 1.183	-11.82	-11.63	-11.47
S. solfataricus	-6.705	± 0.930	-9.022	-8.36	-8.86
T. kodakarensis	-7.135	± 0.836	-10.38	-10.07	-9.27

From Table 4-2, an interval was created in each organism following the limit of the Promoter means plus/minus the standard deviation. The obtaining of this interval is seen in Equation 18.

$$Interval = \bar{x}_{promoters} \pm stdev \quad (18)$$

4.4 A CLASSIFICATION BASED ON NEURAL NETWORKS

The ANN simulations took place in the R environment through the *neuralnet* package (Beck, 2018). The algorithm chosen to fit the ANN was the resilient backpropagation. This application of ANN has succeeded in classifying biological data before (Liu *et* al, 2020). The number of neurons in the hidden layer was kept as 2. The low number of hidden neurons has been described as enough in problems where no higher complexity is required. Additionally, when data has to fit a highly complex curve, the performance of such might get affected (Geman *et* al, 1992). The ANNs that presented the best performance through the assessment of precision, accuracy, recall, and specificity in the datasets (please see Section 4.1) were chosen. Among these, ANNs that presented a balanced computational cost were picked (no simulations that took more than 200,000 steps to converge were considered). Additionally, the number of epochs employed in each test was increased until the validation and training errors keep dropping (Afaq, 2020).

Moreover, in order to provide a form to assess the capacity of generalization of the ANN model, the k-fold-cross-validation was chosen, where $k = 10$. This method involves in reserving 1/10 of the dataset to be further tested. This grants that any biased data point gets covered. The validation process was done following the *sample* method in R (Stone, 1974).

5 RESULTS AND DISCUSSION

The results of this thesis are organized through two scientific articles, one is accepted for publication in the journal Microbiology Open and the other is having its draft finalized.

The article namely “Characterization of promoters in archaeal genomes based on DNA structural parameters seeks to employ structural and energetic profiles of the DNA molecule in order to located transcription factor binding sites throughout the genome of archaea. Moreover, this article aimed to divide the promoter sequences of archaea into two groups: conserved-TATA and degenerated-TATA. The observation of the two groups showed that promoter signals are still perceptible when a given promoter lacks a conserved TATA motif, indicating that the innards of archaeal transcription are more complex than the lone presence of consensual motifs. This article has been accepted for publication in the journal Microbiology Open.

The second article is titled: Machine learning and statistics shape a novel path in archaeal genome annotation. This article is in its final stage and will be submitted to BMC Bioinformatics. Through this article, we aimed to capture signals that turn promoters unmatched. Then, this rationale is captured by an in-house statistical model and Artificial Neural Networks in order to classify promoter sequences among varied forms of control. Lastly, the pattern observed in the promoters is carried out to annotate the genome of archaea whose promoters have not been – *in silico* and experimentally – validated.

5.1 CHARACTERIZATION OF PROMOTERS IN ARCHAEAL GENOMES BASED ON DNA STRUCTURAL PARAMETERS

Characterization of promoters in archaeal genomes based on DNA structural parameters

Running title: DNA structural properties of archaeal promoters

AUTHORS

Gustavo Sganzerla Martinez (gsmartinez@ucs.br)[1]

Scheila de Avila e Silva (sasilva6@ucs.br)[1]

Sharmilee Sarkar (milee93@tezu.ernet.in)[2]

Aditya Kumar (aditya@tezu.ernet.in)[2]

Ernesto Pérez-Rueda (ernesto.perez@iimas.unam.mx)[3]

AFFILIATIONS

[1] Universidade de Caxias do Sul, Programa de Pós-Graduação em Biotecnologia, Av. Francisco Getúlio Vargas, 1130-CEP 95070-560, Caxias do Sul-RS, Brasil.

[2] Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur 784028, Assam, India

[3] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Unidad Académica de Yucatán. Mérida, Yucatán, México.

Correspondence

Scheila de Avila e Silva (sasilva6@ucs.br)

Rua Francisco Getúlio Vargas, 1130, Petrópolis. Caxias do Sul RS Brazil. 95070-560.

Phone +55 54 3218 2100 Extension 2890

ABSTRACT

The transcription machinery of archaea can be roughly classified as a simplified version of eukaryotic organisms. The basal transcription factor machinery binds to the TATA-box found around 28 nucleotides upstream of the transcription start site; however, some transcription units lack a clear TATA-box and still have TBP/TFB binding over them. This apparent absence of conserved sequences could be a consequence of sequence divergence associated with the upstream region, operonic, and gene organization. Furthermore, earlier studies have found that a structural analysis gains more information compared to a simple sequence inspection. In this work, we evaluated and coded 3630 archaeal promoter sequences of three organisms, *Haloferax volcanii*, *Thermococcus kodakarensis*, and *Sulfolobus solfataricus* into DNA duplex stability, enthalpy, curvature, and bendability parameters. We also split our dataset into conserved TATA and degenerated TATA promoters in order to identify differences among these two classes of promoters. The structural analysis reveals variations in archaeal promoters' architecture, i.e., a distinctive signal is observed in the TFB, TBP, and TFE binding sites independently of these being TATA-conserved or TATA-degenerated. In addition, the promoter encountering method was validated with upstream regions of 13 other archaea, suggesting that there might be promoter sequences among them. Therefore, we suggest a novel method for locating promoters within the genome of archaea based on energetic/structural features.

Keywords: Archaea; Transcription; Structural features; Energetic features; TFBS.

1 INTRODUCTION

Archaea represent the third domain of life [1] and include an essential and vast variety of organisms with a large diversity of habitats and lifestyles. This cellular domain has many family divisions belonging to four superphylas: TACK, ASGARD, DPANN, and Euryarchaeota. However, well-known information is only available for two divisions, Euryarchaeota and Crenarchaeota, the latter being a member of the TACK superphylum. In recent years, with the advent of next-generation sequencing, the availability of archaeal genomes has increased, and more than 300 archaeal genomes have become available to the scientific community, allowing the exploration of diverse functional and evolutionary mechanisms, such as membrane origin, operon organization and the proteins devoted to regulating gene expression. Nevertheless, there is a lack of well-annotated archaeal genomic data [2], which enables a lush path towards genomic annotation such as regulatory sequences validation.

The transcription of DNA into RNA and its regulation are central processes in the genetic information flux. Research accumulated in the last few years has evidenced that transcription in archaeal organisms can be roughly described as a simplified version of its eukaryotic relatives [3]. The initiation process begins with the binding of a TATA-binding protein (TBP) and a transcription factor B (TFB) to a specific DNA segment, defined as a promoter, allowing the recruitment of the RNA polymerase (RNAP) enzyme. Additionally, the initiation might be optimized with the presence of a transcription factor E (TFE) protein. Subsequently, an open complex is assembled, followed by the elongation process whereby the RNAP carries out the synthesis of a messenger RNA molecule (mRNA)[4, 5]. In general, three main conserved DNA elements devoted to the

transcription process have been identified as common to all archaeal groups: (*i*) an initiator element (INR) around the transcription start site (TSS); (*ii*) the TATA-box element, centered around −26/27 relative to the TSS; and (*iii*) an element upstream the TATA box comprising two adenines at -34 and -33, which is designated as 'transcription factor B recognition element' (BRE). These elements, INR, TATA-box, and BRE, are crucial to initiating transcription in archaeal genes. They also present a close homology to eukaryotic transcriptional machinery[3, 4].

An in-depth analysis of archaeal promoter elements will provide comprehension of the gene functionality. As an example, there are advances in biotechnology that have employed promoter identification tools in order to enhance gene regulation and optimize biological processes. The broader comprehension of promoter activity could, in theory, enable full control over the start and halt of the expression of specific genes [6]. The production rise in biosynthetic processes is related to the control of regulatory pathways[7]. For example, clinical biology has benefited from promoter identification due to the increased mutation rate found in regulatory regions that may lead to antibiotic resistance. Evolutionary biology has also applied promoter identification as part of the process to understand better horizontal gene transfer between species of the three domains of life [8]. Bioinformatics tools employ physical assets of the genetic material and relate these with gene expression variance, enabling the distinction of specific regions such as promoters. The study of DNA structural features may give rise to more information about promoter activity than a primary sequence analysis [9-12]. Indeed, comparative analysis of bacterial and eukaryotic promoters have shown that Pribnow and TATA-boxes, respectively, differ at structure and sequence level from other random locations within and around the promoter [10, 13].

When converted into numeric attributes, genetic information will promote enough sensibility for capturing even the smallest alterations among the nucleic acids [14]. Hence, we consider the nucleotide conservation found in archaeal promoters [15, 16] will convey a sustained structural parametrization, enabling the characterization of archaeal promoters. In this work, we selected four structural parameters, namely, DNA duplex stability, enthalpy, curvature, and bendability, which are fundamental in understanding the molecular recognition that happens at a structural level [17].

2 DATASETS AND METHODS

2.1 Archaea promoter sequences

In order to determine the nucleotide composition, a total of 3630 promoter sequences of three archaeal organisms were evaluated, which are divided into 1340 sequences of *Haloferax volcanii* [18], 1248 of *Thermococcus kodakarensis* [19], and 1042 of *Sulfolobus solfataricus* [20]. These particular archaea were selected because they are model organisms and well-studied members of *Halobacteriales*, *Thermococcales*, and *Sulfolobales*, respectively. They also have available transcriptome data (RNAseq), enabling the possibility of retrieving promoter sequences from their published information. Internal and antisense promoters from the transcriptome dataset were not included due to limitation of data. The sequence IDs of the three organisms as well as the gene annotation of all *H. volcanii*, *T. kodakarensis* and, *S. solfataricus* promoters used in this analysis are available at http://doi.org/10.5281/zenodo.5137551

The original data covers 1000 nucleotide length sequences, which contains experimentally identified promoters with their transcription start site (TSS), spanning from -500 to +500. Only primary TSS (pTSS) was considered, a category that accounts for abundant transcripts from this original dataset. A shorter sequence was selected,

located at 80 nucleotides upstream and 20 nucleotides downstream of the TSS, i.e., the core promoter. This briefer region was chosen because it contains the core promoter element [21-23]. Accordingly, the core promoter has been detailed as sufficient to convey archaeal and eukaryotic transcription [2, 23, 24]. Indeed, promoters from halophilic archaea were reported to be located in the range proposed here; their TATA-boxes were found in a median distance of 31 base-pairs (bps) upstream the TSS [18]. Additionally, 96% of the pTSS TATA-boxes from *T. kodakarensis* are located in a median distance of 30 base-pairs upstream of the TSS [19]. The TATA-boxes identified in *S. solfataricus* were found in a median length of 35 base-pairs upstream of the TSS [25]. Each archaeal promoter sequence had a shuffled version assigned in order to have a control sequence. The shuffling process was performed by the Supplementary Script S4 (please see at http://doi.org/10.5281/zenodo.5137598).

Moreover, upstream regions from 13 other archaea found in the RSAT Prokaryote Database [42] were selected to validate the method formulated upon the experimentally verified promoters. *Aciduliprofundum boonei* (741 sequences), *Archaeoglobus fulgidus* (866 sequences), *Ferroplasma acidarmanus* (430 sequences), *Haloarcula marismortui* (1998 sequences), *Methanocaldococcus jannaschii* (1866 sequences), *Methanosarcina mazei* (822 sequences), *Methanospirillum hungatei* (1467 sequences), *Methanothermobacter thermautotrophicus* (1870 sequences), and *Pyrococcus furiosus* (1286 sequences) were selected as members of Euryarchaea. The following members of TACK archaea were selected: *Caldivirga maquilingensis* (1669 sequences), *Hyphertermus butylicus* (764 sequences), *Ignicoccus. hospitalis* (1005 sequences), and *Thermofilum pendens* (1926 sequences). DPANN and ASGARD archaea were not included due to their data unavailability. These particular organisms were selected

because of their key role in the evolution of archaea, posing as unique organisms in the archaeal tree of life [43].

2.2 Conversion in structural parameters

In order to convert the DNA sequences into structural parameters, four DNA structural features were selected, namely DNA duplex stability (DDS), enthalpy contribution, bendability, and intrinsic curvature. These features are biologically relevant to characterize promoter regions since they convert DNA information into numeric attributes [14]. These four parameters have previously been used and reflect in capturing specific signals not evident at sequence level [9-13, 26]. Moreover, the appointed features can be described as:

I. The DDS of double-stranded DNA is calculated as the sum of its base-pairs' free energy. It considers the free-energy values associated with the 16 possible combinations of dinucleotides [9].

II. Enthalpy parameters refer to thermodynamic processes that occur at a cellular level (e.g., chemical bonds, mass transport inside and outside the cell, and heat spawning) that affect the thermostability of the cell [27]. These numeric parameters have been taken from DNA melting studies [26].

III. DNA bendability is a sequence-dependent measurement, reflecting in the DNA bending itself because of the effect specific proteins have in the molecule's helical structure. By this means, DNA bending facilitates the assembly of transcription complexes [28]. TATA's bend angle is wider than GC-rich sequences; for instance, TA dinucleotides angle the DNA at 6.74°, the most impactful of the 16 dinucleotide combinations [29].

IV. Finally, intrinsic curvature reflects the capacity of DNA to form small circles around its helical axis [30]. To this end, we used a model based on DNA gel retardation values (BMHT) for its sensibility towards AT-rich sequences [30, 31]. BMHT calculation estimated 16 roll and tilt wedge angles based on independent gel mobility experiments performed on a training set of 54 different sequences [30].

All the four features selected are sequence-dependent and their combination yields more information gathered on a sequence [17]. The complete set of promoter sequences was converted into structural parameters through a self-developed Python script available in Supplementary Script S1 that adopts the numeric parameters available in Table I, except for intrinsic curvature. The curvature calculation hinges on five nucleotides (instead of di and resulted in 4^5 possible combinations). The 1024 numeric parameters are the result of BMHT calculations [30], and they are available in Supplementary Script S2 (please, see scripts at http://doi.org/10.5281/zenodo.5137598).

[Table 1]

The structural properties were computed in a one-nucleotide sliding window. All promoters were aligned relative to their TSS, and numerical values were averaged to get information in each position.

2.3 Classification of conserved TATA and degenerated TATA sequences

To classify the core promoters in conserved and degenerated TATA, the MEME Suite - a motif-based sequence analysis tool [32] was employed. All the sequences were scanned with MEME, and the motifs identified by it were extracted. A key motif for this research would be located in -27/-28, so the search was directed for this specific region to capture the TATAs. The following parameters on MEME were used in the organisms *H. volcanii*

and *T. kodakarensis*: *i*) 100 nucleotides sequence length, considering the -80 to +20 region, where the core promoter is located [21, 23]; *ii*) a 0-order background model generated from the supplied sequences; *iii*) zero or one occurrence (of a contributing motif site) per sequence; *iv*) 8 motifs were located; *v*) the width of the motifs varied between six and eight nucleotides [33]. The motif discovery had to follow different parameters in *S. solfataricus*, in which the width of the motifs was increased from six to fifteen nucleotides to capture the TATA-boxes adequately. Hence, TATA boxes and BRE elements were considered. The combination of these two consensuses was described as a critical feature in *Sulfolobaceae* family transcription [25].

Afterward, the dataset was classified through a self-developed Python script (available in Supplementary Script S3, see at http://doi.org/10.5281/zenodo.5137598), dividing it into two groups: conserved TATA, those motifs identified by MEME, and degenerated TATA, containing sequences which the previously identified motif was not present.

2.4 Statistical tests

Statistical tests were conducted in order to differentiate the two groups this study hinged on. First, the dataset was found not to be normally distributed through the rejection of the null hypothesis achieved by the Shapiro-Wilk test. Then, to determine if the difference between the groups is significant, the Wilcoxon test was applied. Additionally, the non-parametric Kruskal-Wallis test was conducted in order to determine the difference between variances in specific organisms. These tests were done in the R programming language in the *stats* package

3 RESULTS

3.1 Sequence composition

The nucleotide composition of the three archaeal organisms were evaluated to denote the genome configuration particular to each archaeon. Firstly, the 1000-nucleotide sequences are composed by 33.8% of AT in *H. volcanii* DS2, 49.3% in *T. kodakarensis* KOD1 and, 65.5% in *S. solfataricus* P2. Second, the core promoter elements (-80 to +20) in these organisms presented an AT value of 40% in *H. volcanii*, 56.8% in *T. kodakarensis* and 71.1% in *S. solfataricus*.

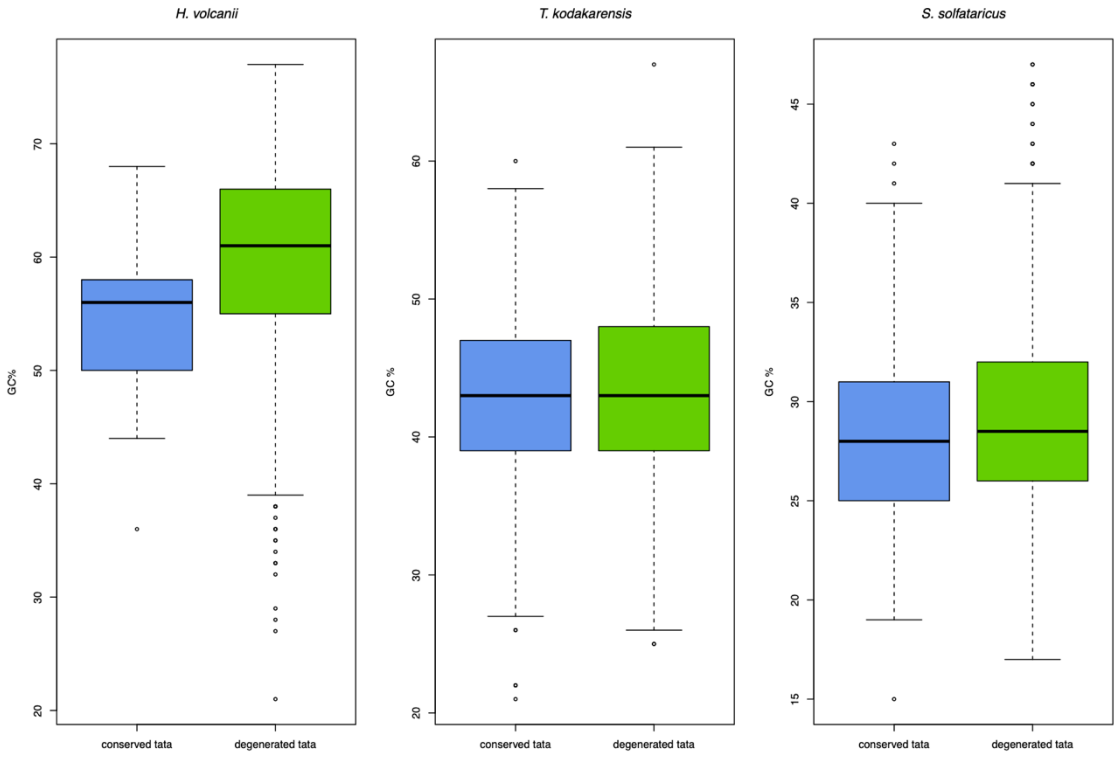
3.2 Conserved TATA and Degenerated TATA boxes.

The datasets were split in two groups to capture particularities and verify the hypothesis of the archaeal transcription being beyond TATA box conservation. The two groups are Conserved TATA and Degenerated TATA. Motifs of eight nucleotides were found in *H. volcanii* and *T. kodakarensis*. Simultaneously, the outcome of *S. solfataricus* encompassed 14 nucleotides. In an attempt to preserve the particularities each archaeon has; the analysis was individually done. The TATA-box motif of each organism is found in Figure A1, from where *H. volcanii* presented SYTTWWAA, *T. kodakarensis* TATA was identified as VYTTWWAA and *S. solfataricus* accounted for VYTTWWWW motifs.

When each one of the motifs were employed to split the dataset, the results of Table 2 were produced. To begin with, 1.56% of 1340 *H. volcanii* sequences presented the TATA motif previously identified. The number of sequences containing motifs in *T. kodakarensis* was 42.72%, and 80.6% in *S. solfataricus*. Then, the GC% of each group was evaluated to verify if they yield statistical significance. U-tests were performed due to the data not following a normal distribution. Figure 1 shows boxplots from which the means of the conserved and degenerated TATA in *H. volcanii*, *T. kodakarensis*, and *S. solfataricus* are $p = 0.0006556$, $p = 0.131$, and $p = 0.005365$, respectively.

[Table 2]

Figure 1: Boxplots of TATA-containing and TATA-less promoter sequences in three archaea



3.3 Structural profiles of archaeal promoter sequences vary when transcription factors binding sites

The entire promoter dataset was converted into enthalpy, DNA Duplex Stability (DDS), bendability and intrinsic curvature in order to capture specific signals in a wider genome analysis, ranging from -500 to +500. In addition, control sequences were added to elicit the strong signals promoter sequences have (Figure 2). A zoomed version, encompassing the promoter region only, was included in Figure 3, where there is a conserved region around the binding site of the transcription factor proteins: TBP (TATA-box, around -28), TFB (BRE, around 2 nucleotides upstream TBP), TFE, whose binding site is located

in position -10 (PPE – proximal promoter element) and +1, matching the INR (initiator element).

Figure 2: The structural/energetic profiles of 1000 nucleotides found in promoter and shuffled sequences.

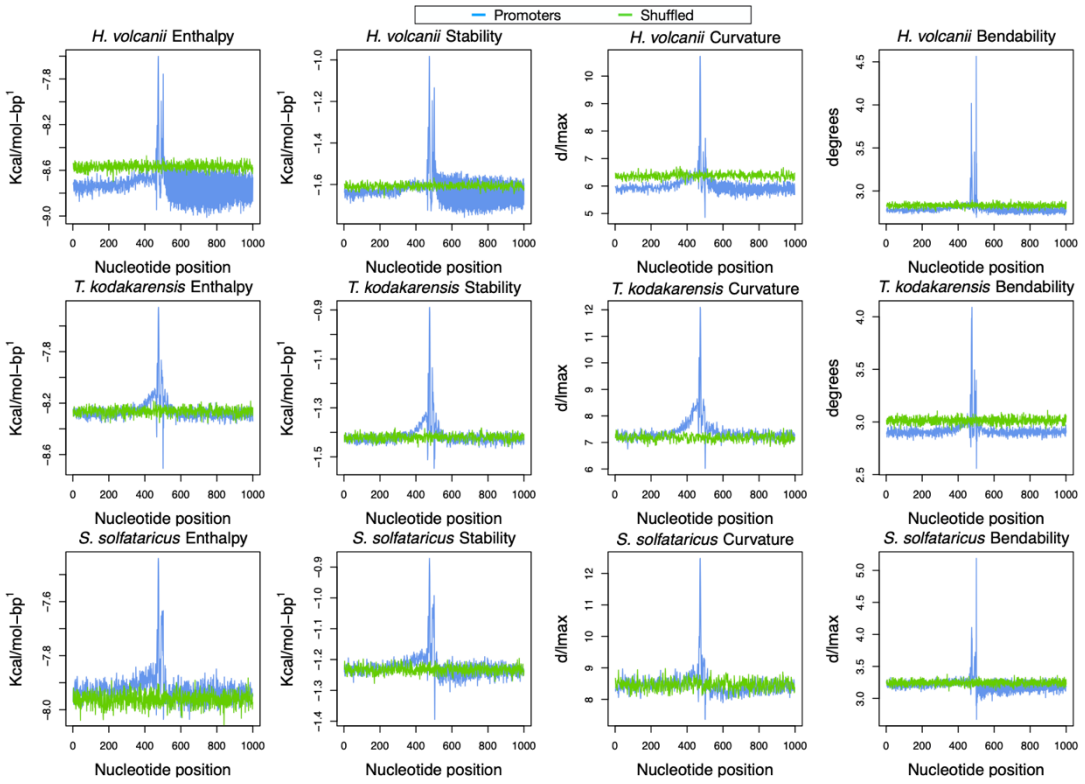
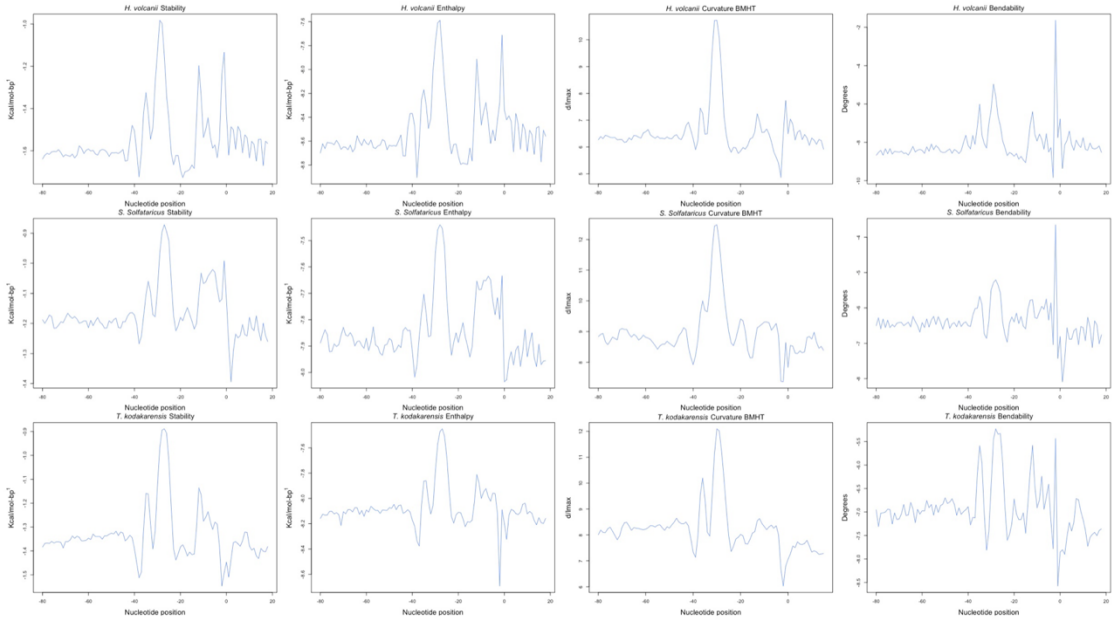


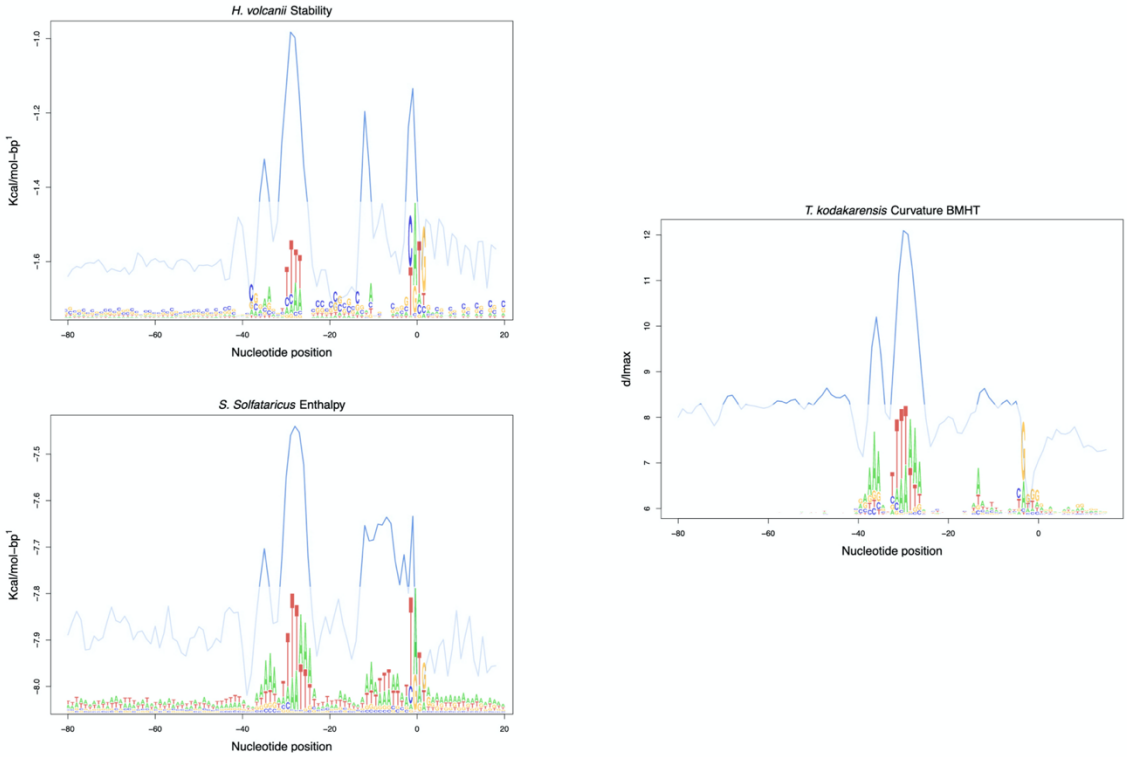
Figure 3: The structural/energetic core promoter profiles.



3.4 Definition of a promoter-like profile

By following the profiles brought by Figure 3, a promoter-like profile was formed upon the average per position (100 nucleotides) of each feature in the validated promoter dataset. By combining nucleotide information (sequence logo profiles) with the structural parametrization brought by this work, Figure 4 was created. In this, the strong DDS, enthalpy, bendability, and BMHT curvature signals are overlaid with transcription factor binding sites.

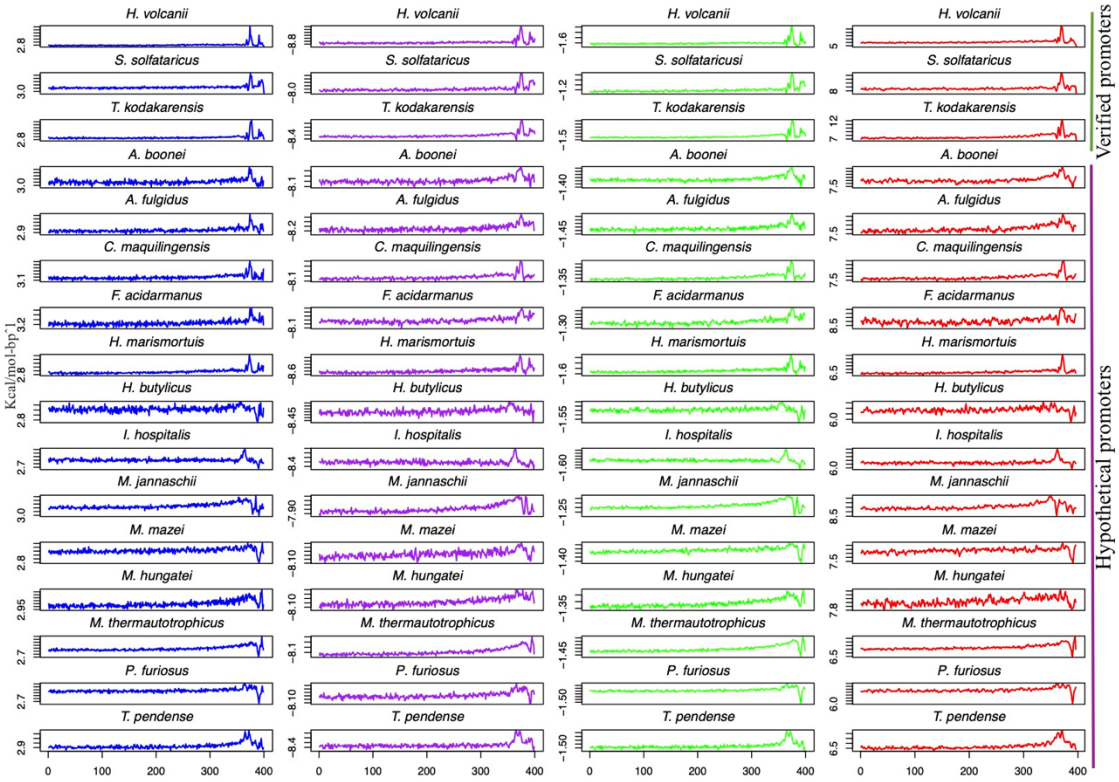
Figure 4: Transcription factor binding sites represented by signals regarding structural/energetic profiles of the core promoter



3.5 Validation of the results with 13 other archaea

Upstream regions of thirteen other archaea divided into: four TACKs and nine *Euryarchaea* were included in order to test the validity of the findings. Figure 5 holds the genomic information of each archaeon plotted against DNA bendability, BMHT curvature, enthalpy, and DDS. In all cases, a strong signal around the ending of the upstream regions was located.

Figure 5: The structural/energetic upstream profiles in thirteen archaea.



Moreover, we included a comparison of the upstream regions found in 13 archaea against the promoter-like profile established in 3.6. In order to perform a comparative analysis, the promoter-like profile was compared to upstream regions of 13 other archaea split into their phylogenetic family (Figure 6). Since the profiles observed in Figure 6 are the same when another physical feature is tested, comparisons following DDS, enthalpy, and bendability are included in Figures A2, A3, and A4, respectively. Analysis of variance tests indicated each organism is significantly different that the other by presenting $p < 2e\text{-}16$ in TACK archaea and $p < 2e\text{-}16$ in *Euryarchaea*. The statistical analysis of the two archaeal families is visualized in boxplots available in Figure 7.

Figure 6: Bendability signal comparison of promoters and upstream regions of thirteen other archaea.

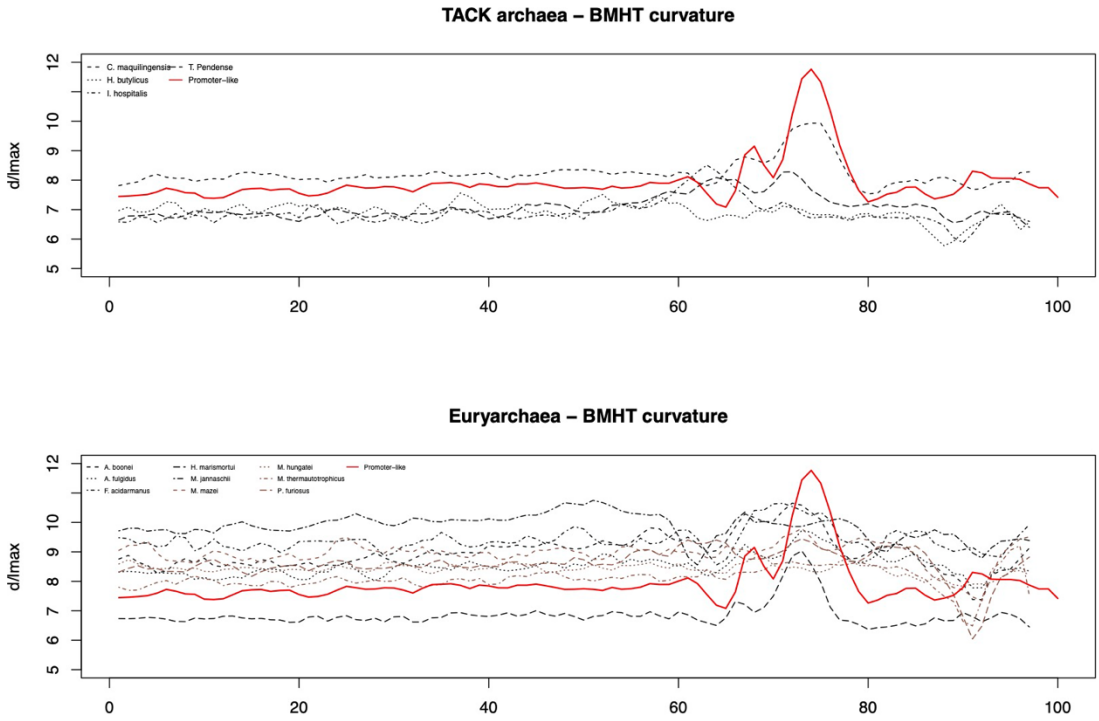
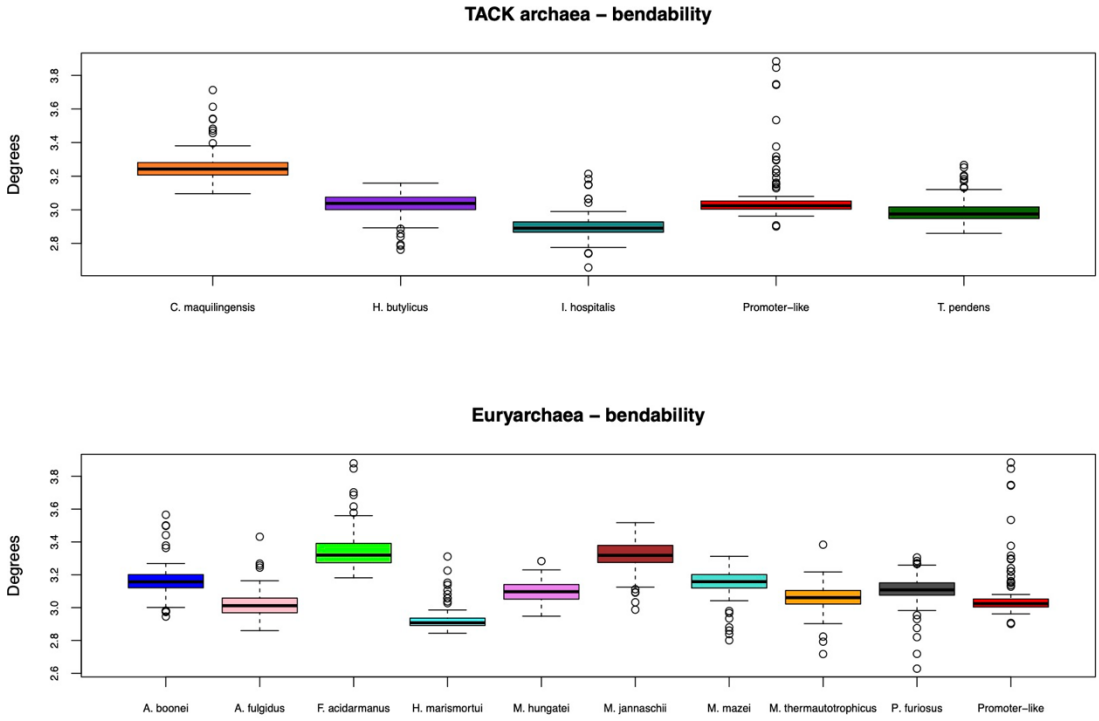


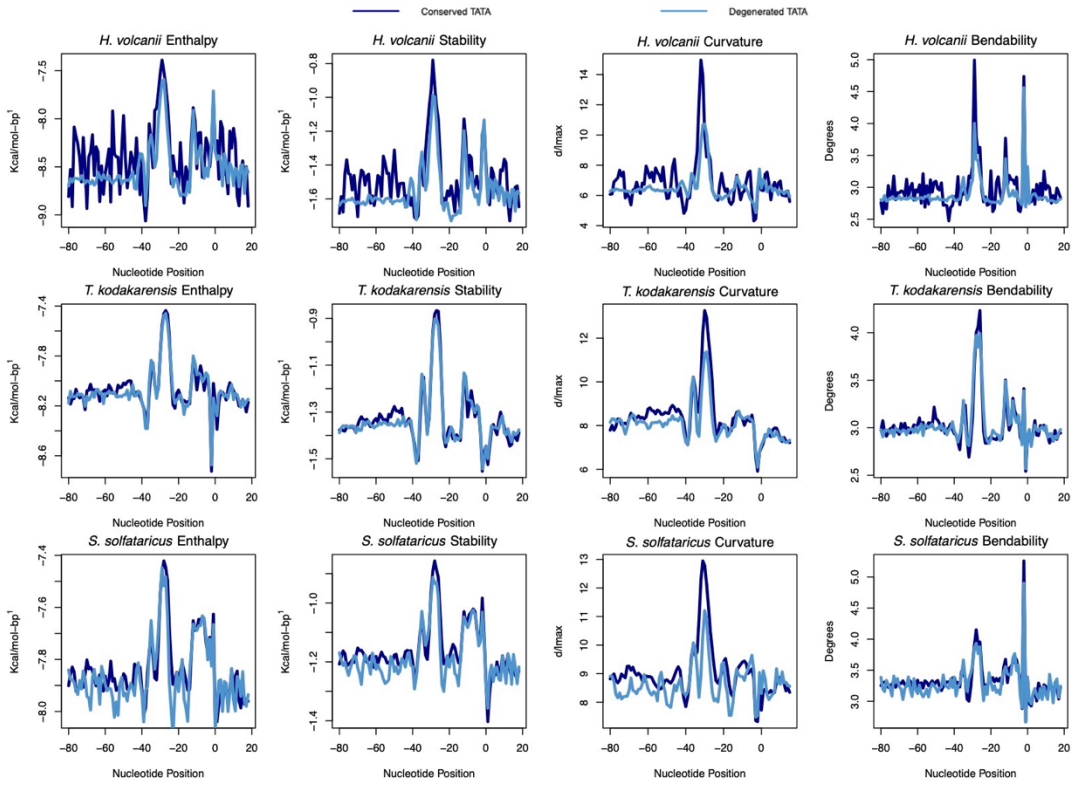
Figure 7: Boxplots of promoters and upstream regions of thirteen other archaea converted to bendability.



3.6 Conserved and degenerated TATA groups

The core promoters belonging to conserved and degenerated TATA groups were converted into energetic and structural properties in order to indicate RNAP action in both groups (Figure 8). The two groups presented overlapping lines with strong signals being located around -28.

Figure 8: The structural/energetic profiles of conserved and degenerated TATA promoters.



4 DISCUSSION

4.1 Nucleotide content

The results of this study suggest that TATA-boxes slightly vary between organisms, supporting the archaeal diversification reported by [34]. Additionally, the AT content was found differently in each archaeon.

When the archaeal promoters were evaluated as owning either a conserved or a degenerated TATA consensus, the GC% each organism has explained the conservation found upon TATA-boxes, so the organism with higher genome GC% was the one that presented the least amount of TATAs, this is no news. However, the binding of TBP, TFB, and TFE to a TATA+BRE motif and TFE binding to PPE/INR were found through this *in-silico* approach to be off from a primary sequence inspection, just as that conservation found around these motifs is not mandatory. Moreover, promoter activity is still observed when promoters lack a clear TATA-motif [22]. Therefore, the uneven number of conserved TATA sequences sprung around archaea is explained by the dynamics of biology. The two groups of promoters (conserved and degenerated TATA) have also presented statistical significance in *H. volcanii* and *S. solfataricus* when the GC content was employed as a possible explanation for each group. This reassures the hypothesis that the probability of TATA-boxes to be found depends directly on the genome composition of a given archaeon.

4.2 Energetic and structural parameters define promoter-like profiles

Promoter sequences might be defined by a set strong signals around their Transcription Factor Binding Sites (TFBS), i.e. TFB, TBP, and TFE. In this study, the conversion of genetic information into physical attributes has protruded distinctive signals around TFBS of the proteins, whilst shuffled sequences did not. These strong signals are in favor of the location of these basal Transcription Factor (TF) location, which is explained by the laws ruling the promoter area. Both enthalpy and stability are energetic-related features, the base-pairs that are more commonly found in promoters are AT and their chemical conformation reflects in more energy available [13, 26, 27, 37]. The distinct signals

represented by curvature and bendability are explained by the TFBS being more rigid and more curved, which acts against the formation of nucleosomes [38].

The profiles obtained in this study indicate conserved aspect around the binding site of proteins that are key elements in the Pre-Initiation Complex (PIC) formation. In vitro studies advocated for TBP+TFB being enough to begin transcription. Indeed, our results show conserved signals around this site (-27 2nt spacer -31). However, the inclusion of signal in the vicinity of -10 and +1, which matches TFE binding site, also contributes to promoter definition. This TF protein was reported to optimize the formation of PIC in TACK and other families as well [44].

The signal located in the -10 region of three archaea is also an important factor in bacterial transcription[40]. Both bacteria and archaea share the same last unique common ancestor, and consequently, share similarities despite their evolution taking place in different branches of the tree of life [16].

The lack of annotation in the genome of many archaea creates the possibility for such methods. When the validation of the promoter identification method was tested in upstream regions of thirteen archaea, the same rationale was inferred. Mining published information upon transcripts has enabled the definition of a promoter-like profile through a combination of strong signals in the binding sites of TBP, TFB, and TFE (-27, -31, -10, and +1, respectively). When data that does not encompass experimentally validated promoter sequences only was assessed, strong signals were observed in the ending of the sequences, suggesting that there might be promoter elements found in these intergenic areas, as identified by [45].

The observation of Figure 6 (and Figures A2, A3, and A4) assures the possibility of locating promoters in upstream regions due to their physical profile. Two archaea have shown TFBS signals similar to the promoter-like profile: *A. boonei* and *T. pendens*. Even

though there are differences in the signals protruded by promoters and potential promoters, resulting in significant differences between the groups' averages, the second group poses for the rise of methods for promoter identification as the one brought by this study.

4.3 Promoter signal beyond TATA boxes.

TATA boxes are likely the most conserved sites that distinguish both archaeal/eukaryotic promoters. The initiation of the transcription in archaea has been reported to start with TPB and TFB proteins attaching to the promoter [3] and enhanced by the presence of TFE[44], this binding is assisted by the conservation found around the binding site of these proteins. Promoters have been grouped in terms of their TATA analysis in [12, 38], both authors performed structural conversions such as this study did. Divergent results could be observed in which TATA-conserved sequences did not show significant differences when compared to TATA-degenerated ones.

In this study, both TATA-conserved and TATA-degenerated groups have shown the same strong signals around the binding sites of TFB, TBP, and TFE. There are differences that might protrude mathematical variance, e.g. the TFB and TBP binding sites analyzed in the curvature profile of three archaea and *H. volcanii*'s bendability and DDS. This feature defines the promoter (either TATA-conserved or not) as a promoter-like sequence, which is a novel approach in identifying and finding new promoter sequences in archaea.

5 CONCLUSIONS

The results we demonstrated in this study encourage the DNA codification into energetic/structural attributes that reveal transcription factor proteins binding sites where

a primary sequence inspection failed. Hence, this study poses a novel method to be used in genome annotation regarding archaeal promoters.

Author Contributions

Gustavo Martinez

Conceptualization-Equal, Data curation – Equal; Formal analysis – Equal; Funding acquisition – Supporting; Investigation – Equal; Methodology – Equal; Project administration – Equal; Resources – Equal; Software – Equal; Supervision – Equal; Validation – Equal; Visualization – Equal; Writing original draft – Equal; Writing-review & editing – Equal.

Sharmilee Sarkar

Conceptualization-Equal, Data curation – Equal; Formal analysis – Equal; Funding acquisition – Supporting; Investigation – Equal; Methodology – Equal; Project administration – Equal; Resources – Equal; Software – Equal; Supervision – Equal; Validation – Equal; Visualization – Equal; Writing original draft – Equal; Writing-review & editing – Equal.

Aditya Kumar

Conceptualization-Equal, Data curation – Equal; Formal analysis – Equal; Funding acquisition – Supporting; Investigation – Equal; Methodology – Equal; Project administration – Equal; Resources – Equal; Software – Equal; Supervision – Equal; Validation – Equal; Visualization – Equal; Writing original draft – Equal; Writing-review & editing – Equal.

Ernesto Perez-Rueda

Conceptualization-Equal, Data curation – Equal; Formal analysis – Equal; Funding acquisition – Leading; Investigation – Equal; Methodology – Equal; Project administration – Equal; Resources – Equal; Software – Equal; Supervision – Equal; Validation – Equal; Visualization – Equal; Writing original draft – Equal; Writing-review & editing – Equal.

Scheila Silva

Conceptualization-Equal, Data curation – Equal; Formal analysis – Equal; Funding acquisition – Supporting; Investigation – Equal; Methodology – Equal; Project administration – Equal; Resources – Equal; Software – Equal; Supervision – Equal; Validation – Equal; Visualization – Equal; Writing original draft – Equal; Writing-review & editing – Equal.

Acknowledgements: We are grateful to the support received from Universidade de Caxias do Sul (UCS), Coordernação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Department of Biotechnology (DBT), Govt. of India for project Junior Research Fellowship, and the Department of Biotechnology, Govt. of India for the DBT Twinning project grant (BT/PR24927/NER/95/911/2017). This research was supported by grants from Dirección General de Asuntos del Personal Académico-Universidad Nacional Autónoma de México (UNAM) (IN-209620).

Ethics statement: None required.

Conflict of Interests: None declared.

Data availability statement: The experimentally verified promoter sequences of *H. volcanii* (doi: 10.1186/s12864-016-2920-y), *S. solfataricus* (doi: 10.1101/gr.100396.109), and *T. kodakarensis* (doi: 10.1016/0022-2836(91)90492-O) were retrieved from their original publications. The upstream regions employed in the method validation step were extracted from RSAT Database. (Please see DOIs). Gene annotation for the three archaea tested in this study is available at http://doi.org/10.5281/zenodo.5137551.

REFERENCES

1. Woese, C. R. (1987). Bacterial evolution. Microbiol Mol Biol Rev. 51:2, 221-271.
2. Zuo, G., Xu, Z., & Hao, B. (2015). Phylogeny and taxonomy of archaea: A comparison of the Whole-Genome-Based CVtree approach with 16s rRNA sequence analysis. Life. 5:1, 949-968.
3. Gehring, A. M., Walker, J. E., Santangelo, T. J. (2016). Transcription regulation in archaea. J Bacteriol.. 198:14, 1906-1917. doi: 10.1128/JB.00255-16
4. Soppa, J. (1999). Transcription initiation in Archaea: Facts, factors and future aspects. Mol Microbiol. 31:5, doi: 10.1046/j.1365-2958.1999.01273.x
5. Smollet, K., Blombach, F., Fouqueau, T., Wernerm F. (2017) A Global Characterisation of the Archaeal Transcription Machinery. In: Clouet-dO'rval, B. (eds) RNA metabolism and Gene Expression in Archaea. Nucleic Acids Mol Biol, 32.
6. Kernan, T., West, A. C., Banta, S. (2017). Characterization of endogenous promoters for control of recombinant gene expression in Acidithiobacillus ferrooxidans. Biotechnol Appl Biochem. 64:6, doi: 10.1002/bab.1546.
7. Ren, H., Shi, C., Zhao, H. (2020). Computational Tools for Discovering and Engineering Natural Product Biosynthetic Pathways. IScience. 23:1, doi: 10.1016/j.isci.2019.100795
8. Khademi, S. M., Sazinas, P., Jelsbak, L. (2019). Within-Host adaptation mediated by intergenic evolution in Pseudomonas aeruginosa. Genome Biol Evol. 11:5, 1385-1397.
9. Kanhere, A., Bansal, M. (2005). A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinformatics. 6:1. doi: 10.1186/1471-2105-6-1
10. De Avila e Silva, S., Echeverrigaray, S., Gerhardt, G. J. L. (2011). BacPP: Bacterial promoter prediction-A tool for accurate sigma-factor specific assignment in enterobacteria. J Theor Biol. 287, 92-99.
11. Bansal, M., Kumar, A., Yella, V. R. (2014). Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. Curr Opin Struct Biol. 25, 77-85.

12. Yella, V. R., & Bansal, M. (2017). DNA structural features of eukaryotic TATA-containing and TATA-less promoters. FEBS Open Bio. 7(3), 324-334. https://doi.org/10.1002/2211-5463.12166
13. Yella, V. R., Kumar, A., & Bansal, M. (2018). Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. Sci Rep. 8, 4250. doi: 10.1038/s41598-018-22129-8
14. Benham, C. J. (1996). Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. J Mol Biol. 225:3, 425-434.
15. Londei, P. (2005). Evolution of translational initiation: New insights from the archaea. FEMS Microbiol Rev. 29:2, 185-200.
16. Gribaldo, S., Brochier-Armanet, C. (2006). The origin and evolution of Archaea: A state of the art. Philos Trans R Soc Lond B Biol Sci. 361: 1470, https://doi.org/10.1098/rstb.2006.1841
17. Ryasik, A., Orlov, M., Zykova, E., Ermak, T., Sorokin, A. (2018). Bacterial promoter prediction: Selection of dynamic and static physical properties of DNA for reliable sequence classification. J Bioinform Comput Biol. 16:1. doi: 10.1142/S0219720018400036
18. Babski, J., Haas, K. A., Näther-Schindler, D., Pfeiffer, F., Förstner, K. U., Hammelmann, M., Hilker, R., Becker, A., Sharma, C. M., Marchfelder A., Soppa, J. (2016). Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). BMC Genomics. 17: 629. doi: 10.1186/s12864-016-2920-y
19. Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J., Reeve, J. N. (2014). Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. BMC Genomics. 222:5, 495-508. doi: 10.1016/0022-2836(91)90492-O
20. Wurtzel, O., Sapra, R., Chen. F., Zhu, Y., Simmons, B. A., Sorek, R. (2009). A single-base resolution map of an archaeal transcriptome. Genome Res. 20: 133-141.
21. Kadonaga, J. T. (2012). Perspectives on the RNA polymerase II core promoter. Wiley Interdiscip Rev Dev Biol. 1:1, doi: doi.org/10.1002/wdev.21
22. Aptekman, A. A., Nadra, A. D., (2018). Core promoter information content correlates with optimal growth temperature. Scientific Reports. 8, 1313. https://doi.org/10.1038/s41598-018-19495-8

23. Haberle, V., Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 19, 621-637. doi: 10.1038/s41580-018-0028-8
24. Bartlett, M. S., Thomm, M., Geiduschek, E. P. (2000). The orientation of DNA in an archaeal transcription initiation complex. Nat Struct Biol. 7, 782-785.
25. Le, T. N., Wagner, A., Albers, S., (2017). A conserved hexanucleotide motif is important in UV-inducible promoters in Sulfolobus acidocaldarius. Microbiology (N Y) 163:5, 778-788.
26. SantaLucia, J., Hicks, D. (2004). The Thermodynamics of DNA Structural Motifs. Annu Rev Biophys Biomol Struct. 33, 415-440.
27. Privalov, P. L., Crane-Robinson, C. (2018). Forces maintaining the DNA double helix and its complexes with transcription factors. Prog Biophys Mol Biol. 135, 30-48.
28. Leonard, D. A., Rajaram, N., Kerppola, T. K. (1997). Structural basis of DNA bending and oriented heterodimer binding by the basic leucine zipper domains of Fos and Jun. Proc Natl Acad Sci U S A. 94:10, 4913-4918.
29. Karas, H., Knuppel, R., Schuiz, W., Sklenar, H., Wingender, E. (1996). Combining structural analysis of dna with search routines for the detection of transcription regulatory elements. Bioinformatics. 12:5, 441-446.
30. Bolshoy, A., McNamara, P., Harrington, R. E., Trifonov, E. N. (1991). Curved DNA without A-A: Experimental estimation of all 16 DNA wedge angles. Proc Natl Acad Sci U S A 88:6, 2312-2316.
31. Kanhere, A., Bansal, M. (2003). An assessment of three dinucleotide parameters to predict DNA curvature by quantitative comparison with experimental data. Nucleic Acids Res. 31:10, 2647, 2658.
32. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Jingyuan, R., Wilfred, W. L., Noble, W. S. (2009). MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. 37, Issue suppl_2, W202-W208. doi: 10.1093/nar/gkp335
33. Hausner, W., Frey, G., Thomm, M. (1991). Control regions of an archaeal gene. A TATA box and an initiator element promote cell-free transcription of the tRNAVal gene of Methanococcus vannielii. J Mol Biol. https://doi.org/10.1016/0022-2836(91)90492-O

34. DeLong, E. F., Wu, K. Y., Prézelin, B. B., Jovine, R. V. M. (1994). High abundance of Archaea in Antarctic marine picoplankton. Nature. 371, 695-697.
35. Ao, X., Li, Y., Wang, F., Feng, M., Lin, Y., Zhao, S., Liang, Y., Peng, N., (2013). The sulfolobus initiation element is an important contributor to promoter strength. J Bacteriol. 195:22, 5216-5222.
36. De Avila e Silva, S., Forte, F., T.S. Sartor, I., Andrighetti, T., J.L. Gerhardt, G., Delamare, A. P. L.; Echeverrigaray, S. (2014). DNA duplex stability as discriminative characteristic for Escherichia coli σ54- and σ28- dependent promoter sequences. Biologicals. 42(1), 22-28. https://doi.org/10.1016/j.biologicals.2013.10.001
37. Allawi, H. T., & SantaLucia, J. (1997) Thermodynamics and NMR of internal G-T Mismatches in DNA. Biochemistry. 97(36), 10581-10594. https://doi.org/10.1021/bi962590c
38. Tirosh, I., Berman, J., & Barkai, N. (2007). The pattern and evolution of yeast promoter bendability. Trends in Genetics. 23:7, 318-321.
39. Blombach, F., & Grohmann, D. (2017). Same same but different: The evolution of TBP in archaea and their eukaryotic offspring. Transcription. 8:3, 162-168.
40. Lloréns-Rico, V., Lluch-Senar, M., Serrano, L. (2015). Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. Nucleic Acids Res. 43:7, 3442-3453.
41. Anish, R., Hossain, M. B., Jacobson, R. H., & Takada, S. (2009). Characterization of transcription from TATA-less promoters: Identification of a new core promoter element XCPE2 and analysis of factor requirements. PLoS ONE. https://doi.org/10.1371/journal.pone.0005103
42. Nguyen, N. T. T., Contreras-Moreira, B., Castro-Mondragon, J. A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C. D., … Thomas-Chollier, M. (2018). RSAT 2018: Regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky317
43. Williams, T. A., Szöllosi, G. J., Spang, A., Foster, P. G., Heaps, S. E., Boussau, B., … Martin Embley, T. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1618463114

44. Hanzelka, B. L., Darcy, T. J., & Reeve, J. N. (2001). TFE, an archaeal transcription factor in Methanobacterium thermoautotrophicum related to eucaryal transcription factor TFIIEα. *Journal of Bacteriology*. https://doi.org/10.1128/JB.183.5.1813-1818.2001
45. Yella, V. R., Kumar, A., & Bansal, M. (2018). Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. *Scientific Reports*. https://doi.org/10.1038/s41598-018-22129-8

Tables

Table 1 - Enthalpy, stability, and bendability parameters for every possible dinucleotide combination.

Dinucleotide	Enthalpy (kcal/mol-bp-1)	Stability (kcal/mol-bp-1)	DNA bendability (degrees)
AA	-7.6	-1.00	3.07
AT	-7.2	-0.88	2.6
AC	-8,5	-1.45	2.97
AG	-8.2	-1.3	2.31
TT	-7.6	-1	3.07
TA	-7.2	-0.58	6.74
TC	-7.8	-1.28	2.51
TG	-8.4	-1.44	3.58
CC	-8	-1.28	2.16
CA	-8.5	-1.45	3.58
CT	-7.8	-1.28	2.31
CG	-10.6	-2.24	2.81
GG	-8	-1.84	2.16
GA	-8.2	-1.3	2.51
GT	-8.4	-1.44	2.97
GC	-10.6	-2.24	3.06

Table 2 – Conserved TATA and degenerated TATA upon core promoter sequences in three archaeal organisms.

		Conserved TATA		**Degenerated TATA**	
Organism	**Genome GC%**	**Number of promoters (%)**	**GC%**	**Number of promoters (%)**	**GC%**
H. volcanii	66.13	21 (1.56%)	54.09	1319 (98.44%)	60.03
T. kodakarensis	50.67	506 (42.72%)	42.55	742 (57.28%)	43.39
S. solfataricus	34.48	840 (80.6%)	28.42	202 (19.4%)	30.68

Figure legends

Figure 1: Boxplots of TATA-containing and TATA-less promoter sequences in three archaea. We divided 1340 *H. volcanii*, 1248 *T. kodakarensis* and 1042 *S. solfataricus* sequences into two groups: TATA-containing and TATA-less by following Materials and Methods 2.3. Then, we calculated the GC% of each sequence in the groups and created boxplots alongside U-tests to discover significance between the groups. The p values in the non-parametric U-tests were: 0.0006556, 0.131, and 3.241e-09 in *H. volcanii*, *T. kodakarensis* and, *S. solfataricus*, respectively.

Figure 2: The structural/energetic profiles of 1000 nucleotides found in promoter and shuffled sequences. Energetic/structural features of three archaea. We plotted the average value in each one of the 1000 positions. The highest peak is seen at position -28 in three archaea, four measurements. The blue line represents the promoter sequences and the green line indicates a shuffled version of the promoters. The shuffling process was carried out by a Python script available in Supplementary Script S4.

Figure 3: The structural/energetic core promoter profiles. Energetic and sequence-dependent features of three archaea. We plotted the average of the core promoter positions reported by Kadonaga, 2012; Haberle and Stark, 2018. Our plots indicated a strong signal in i) the TATA-box and BRE positions; ii) the PPE area.

Figure 4: Transcription factor binding sites represented by signals regarding structural/energetic profiles of the core promoter. Nucleotide information (sequence logo profiles) are overlaid with signals that represent the core promoter content.

Figure 5: The structural/energetic upstream profiles in thirteen archaea. Thirteen other archaea were selected from 42 in order to validate the promoter-like behavior observed. These organisms have 400 nucleotide-long sequences corresponding to upstream sequences where no annotation towards promoter finding was done. The blue lines represent bendability profiles, the purple enthalpy, the green refers to DDS, and the red is BMHT curvature.

Figure 6: Bendability signal comparison of promoters and upstream regions of thirteen other archaea. The red line (promoter-like) represents the average formed upon experimentally validated promoter of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis* to be compared with upstream sequences of thirteen other archaea divided into two phylogenetic families. The remaining DDS, enthalpy, and BHMT curvature are found in Figures A2, A3, and A4, respectively.

Figure 7: Boxplots of promoters and upstream regions of thirteen other archaea converted to bendability. The boxplots represent statistical comparisons between the promoter-like profile, (red), formed upon experimental data of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. The $p < 2e\text{-}16$ values obtained by the nonparametric Kruskal-Wallis test conveyed statistical significance in the averages of both groups. Additional analyses encompassing BMHT curvature, enthalpy, and DDS are found in Figures A5, A6, and A7, respectively.

Figure 8: The structural/energetic profiles of conserved and degenerated TATA promoters. The conserved and degenerated TATA core promoter profiles are plotted. The lines represent the average value each group and organism showed. The navy-blue lines represent sequences that had a MEME-identified TATA motif, the light blue depicts sequences in which the specific TATA motif was not found.

Appendix Figures

Figure A1: *H. volcanii*, *T. kodakarensis* and *S. solfataricus* TATA-motifs identified by the MEME suite. S1 (a) indicates *H. volcanii*, from which the resulting motif was found in a median position downstream of the TSS of 31 bps. The e-value of this motif is 1.8 e-092; it has been found in 382 sites; its relative entropy is 12.1 and; information content = 13.2. S1 (b) indicates *T. kodakarensis*, motifs located in a median distance of 30 bps downstream of the TSS. Its e-value is 1.3 e-017; this motif has been located in 257 sites; relative entropy and information content 12.3 and 12.4, respectively. S1 (c) represents the TATA-motif found in *S. solfataricus* found in a median distance of 30 bps downstream the TSS. The e-value = 6.1e-022, site count 192, relative entropy = 12.5 and, information content = 16.2.

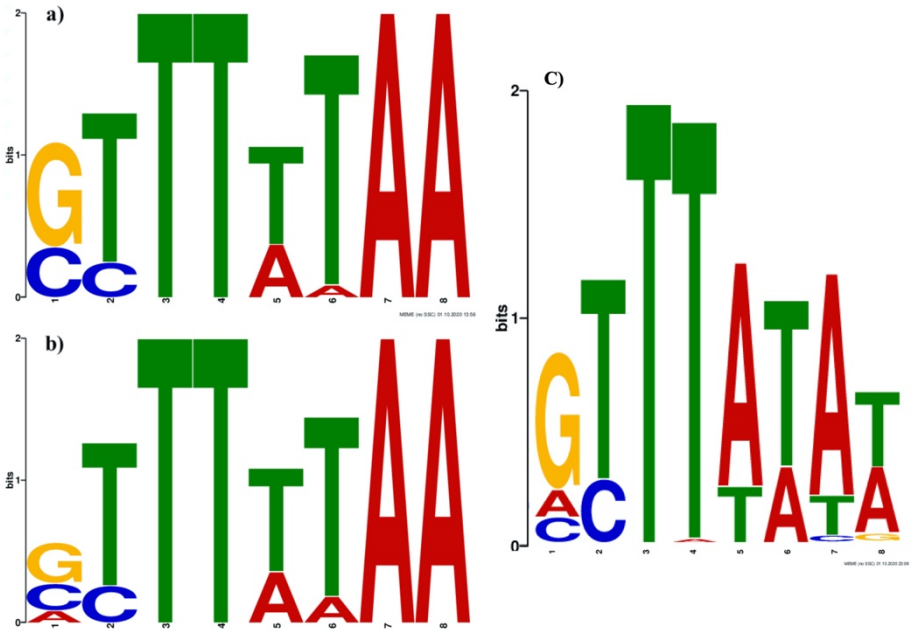


Figure A2: DNA Duplex Stability signal comparison of promoters and upstream regions of thirteen other archaea. The green line (promoter-like) represents the average formed upon experimentally validated promoter of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis* to be compared with upstream sequences of thirteen other archaea divided into two phylogenetic families.

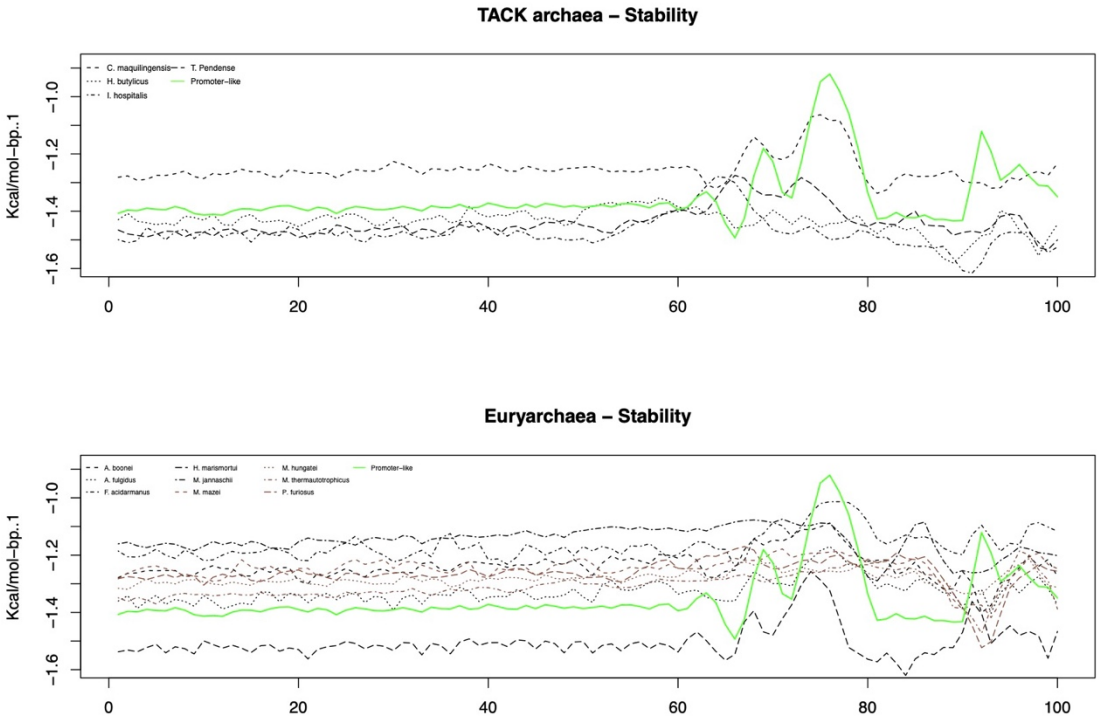


Figure A3: Enthalpy signal comparison of promoters and upstream regions of thirteen other archaea. The purple line (promoter-like) represents the average formed upon experimentally validated promoter of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis* to be compared with upstream sequences of thirteen other archaea divided into two phylogenetic families.

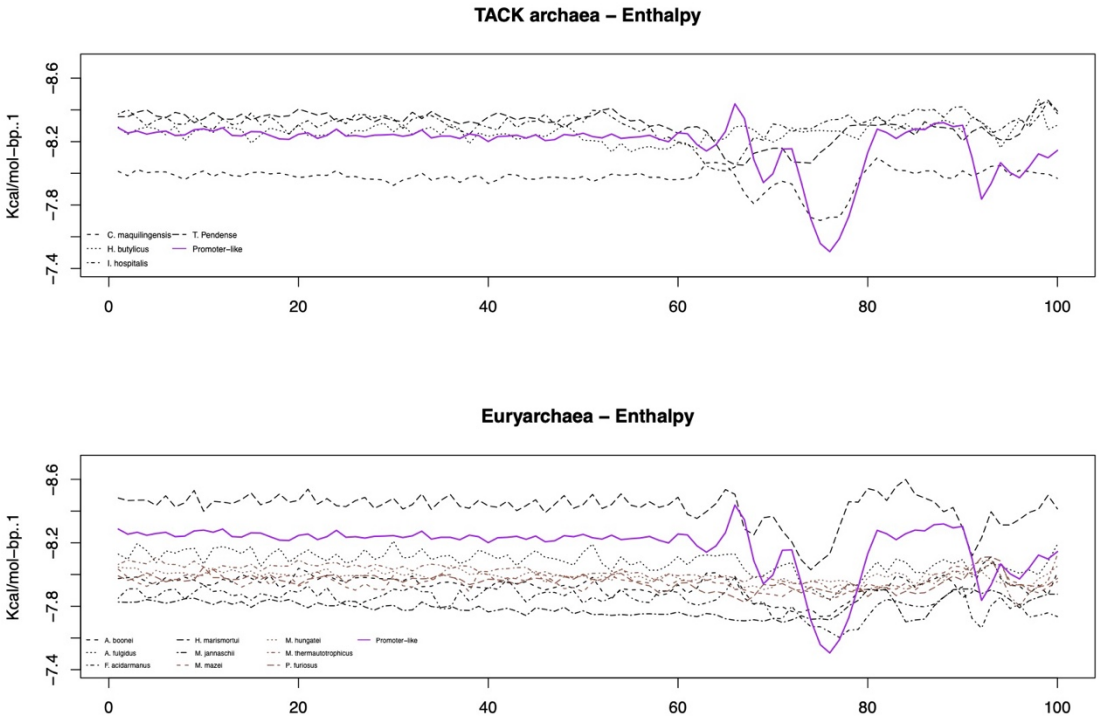


Figure A4: Bendability signal comparison of promoters and upstream regions of thirteen other archaea. The blue line (promoter-like) represents the average formed upon experimentally validated promoter of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis* to be compared with upstream sequences of thirteen other archaea divided into two phylogenetic families.

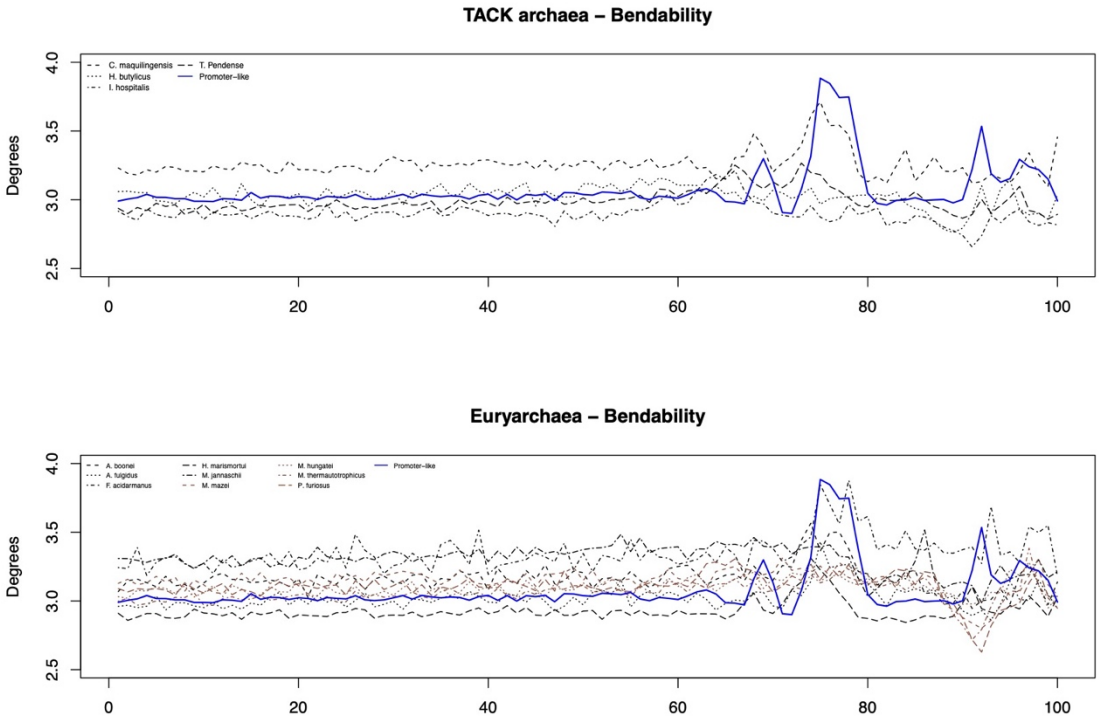


Figure A5: Boxplots of promoters and upstream regions of thirteen other archaea converted to BMHT curvature. The boxplots represent statistical comparisons between the promoter-like profile, (red), formed upon experimental data of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. The p < 2e-16 values obtained by the nonparametric Kruskal-Wallis test conveyed statistical significance in the averages of both groups.

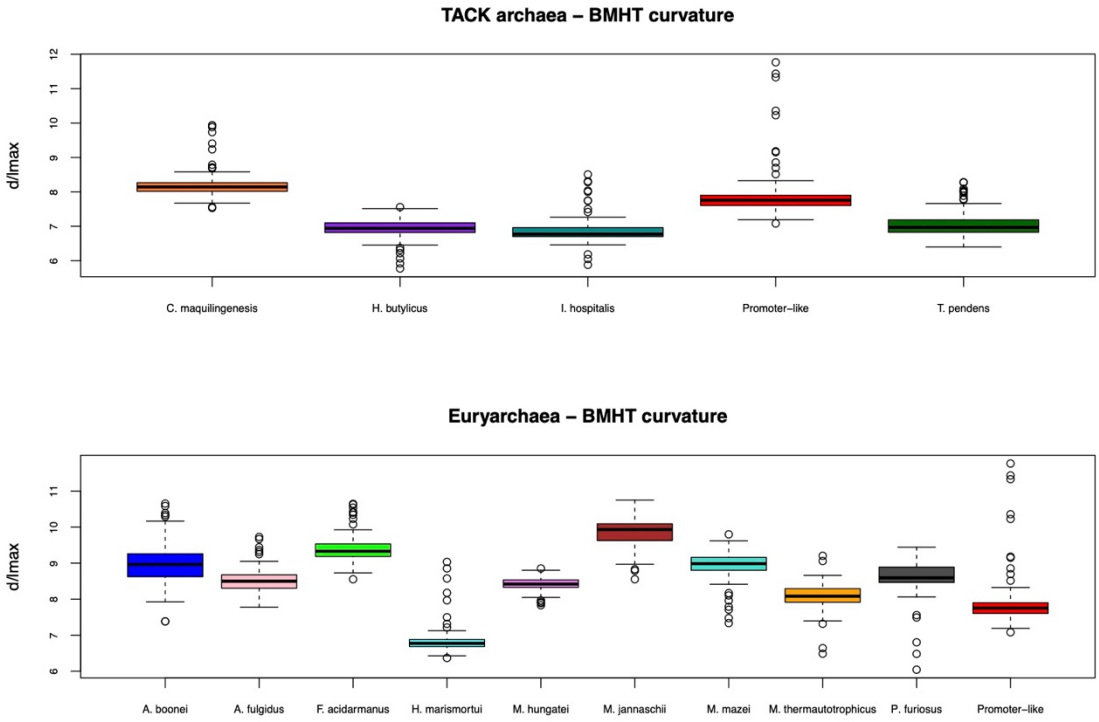


Figure A6: Boxplots of promoters and upstream regions of thirteen other archaea converted to enthalpy. The boxplots represent statistical comparisons between the promoter-like profile, (red), formed upon experimental data of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. The p < 2e-16 values obtained by the nonparametric Kruskal-Wallis test conveyed statistical significance in the averages of both groups.

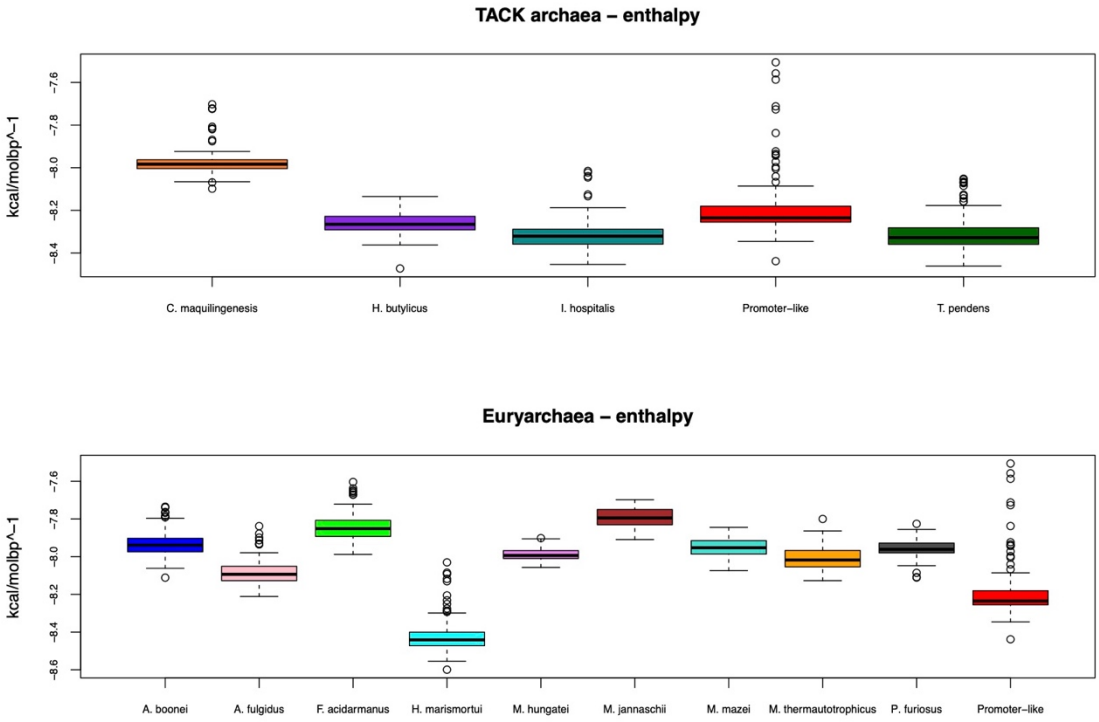
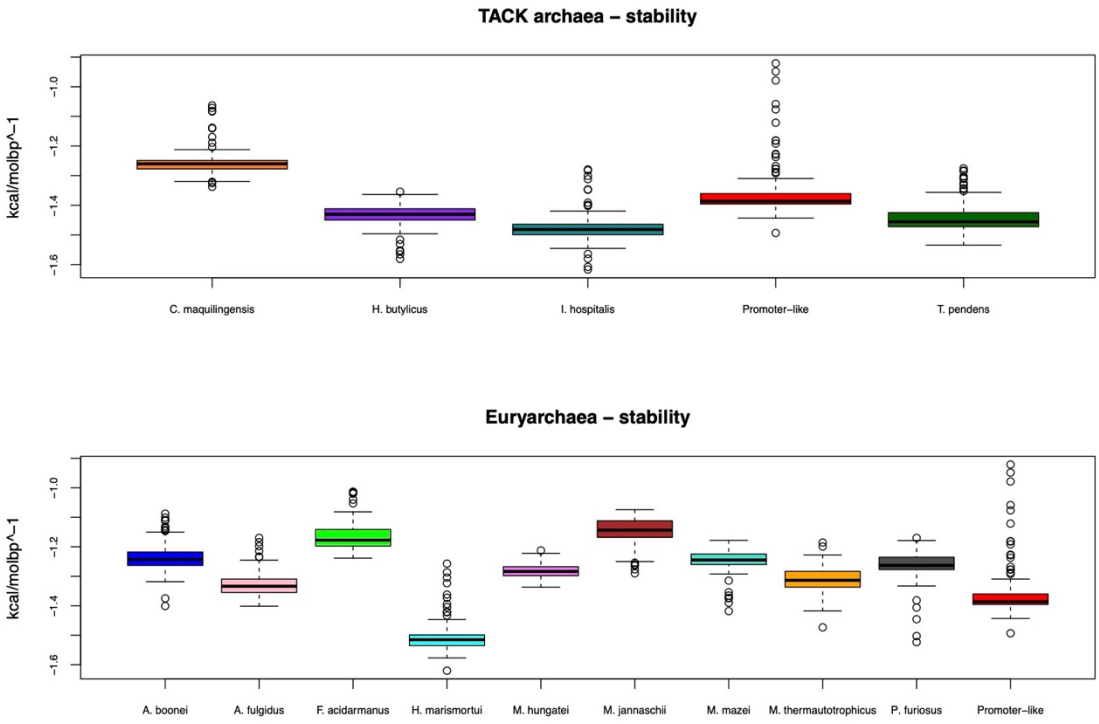


Figure A7: Boxplots of promoters and upstream regions of thirteen other archaea converted to stability. The boxplots represent statistical comparisons between the promoter-like profile, (red), formed upon experimental data of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. The p < 2e-16 values obtained by the nonparametric Kruskal-Wallis test conveyed statistical significance in the averages of both groups.



5.2 MACHINE LEARNING AND STATISTICS SHAPE A NOVEL PATH IN ARCHAEAL GENOME ANNOTATION

Machine learning and statistics shape a novel path in archaeal genome annotation

AUTHORS

Gustavo Sganzerla Martinez (gsmartinez@ucs.br)[1]

Ernesto Pérez-Rueda (ernesto.perez@iimas.unam.mx)[3]

Sharmilee Sarkar (milee93@tezu.ernet.in)[2]

Aditya Kumar (aditya@tezu.ernet.in)[2]

Scheila de Avila e Silva (sasilva6@ucs.br)[1]

AFFILIATIONS

[1] Universidade de Caxias do Sul, Programa de Pós-Graduação em Biotecnologia, Av. Francisco Getúlio Vargas, 1130-CEP 95070-560, Caxias do Sul-RS, Brasil.

[2] Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur 784028, Assam, India

[3] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Unidad Académica de Yucatán. Mérida, Yucatán, México.

ABSTRACT

Background

Archaea are a vast and unexplored domain. Bioinformatics techniques might enlighten the path to a higher quality genome annotation in varied organisms. Promoter sequences of archaea have the action of a plethora of proteins upon it. The conservation found in a structural level of the binding site of proteins such as TBP, TBP, and TFE aids RNAP DNA binding and makes the archaeal promoter prone to be explored by statistical and machine learning techniques.

Results

In this study, experimentally verified promoter sequences of the organisms *Haloferax volcanii*, *Sulfolobus solfataricus*, and *Thermococcus kodakarensis* were coded into DNA duplex stability attributes (i.e. numerical variables) and were classified through Artificial Neural Networks and an in-house statistical method of classification, being tested with three forms of controls. The recognition of these promoters enabled its use to validate unnanoted promoter sequences in other organisms. As a result, the binding site of basal transcription factors was located through a DNA duplex stability codification. Additionally, the classification presented satisfactory results (above 90%) among varied levels of control.

Conclusions

The classification models were employed to perform genetic annotation into the archaea *Aciduliprofundum boonei* and *Thermofilum pendens*, from which potential promoters have been identified and uploaded into public repositories.

BACKGROUND

The righteous introduction of the archaeal domain to the tree of life dates no longer than half a century. Since then, a lush path towards discovering new insights in order to benefit archaeal genome annotation arose. The archaeal domain is diverse [4, 5], they inhabit the Earth's most extreme environments to our guts. Hence, finding a model organism that represents the whole expanse of this domain is rather a simple-minded and reductionist task. At least 13 families in the archaea phylogenetic tree might be spotted, which have huge dissimilarities both in their genetic and phenotype setting [6, 7], as well as elements that orchestrate the cell necessities.

Single cell organisms rely on finely regulated cellular processes. The production of the right nutrient at the right moment grants the cell survivability. Instances of these processes

include the transcription of an RNA molecule. This mid-step operation is carried out by the RNAP enzyme and configures a central process in the genetic information flux across all domains. The way the transcription occurs in archaea roughly resembles the eukaryotes. In fact, these two domains are evolutive siblings and archaea might have given origin to eukarya [32]. The overall structure of this process in these two domains presents a certain level of conservation. Indeed, the eukaryotic model poses as a more specialized version of its archaeal counterpart. For instance, while archaea employs a single RNAP to transcribe all genes, animals and plants make use of three and five different enzymes, respectively [28, 33].

The recruitment of RNAP to the DNA is mediated by a DNA segment defined as a promoter sequence whose presence is necessary for the initiation of the transcription. The typical archaeal promoter element possesses three basal transcription factor binding sites. These additional proteins are TATA-box Binding Protein (TBP), Transcription Factor B (TFB), and Transcription Factor E (TFE) and they are needed in correctly directing RNAP to its precise site of action [34]. On a nucleic acid level, these proteins bind to: *i*) a wTTATwww set of nucleotides, located at -25, matching the TBP binding site; *ii*) an ssnAA sequence located around two nucleotides upstream TATA and a TAC sequence located in the range of -1/-10, due to its two-extremity binding, TFB stabilizes TBP and the two combined create the Pre Initiation Complex (PIC); *iii*) a TFE protein has the function of assisting PIC formation, hence, its binding preference varies according to the promoter and organism [27, 34].

The conservation found around the binding site of transcription factor proteins in the archaeal genome might be used as input in a way that the recognition of these regulators is able to provide a more reliable annotation. The promoter prediction task is well developed in other branches of life than archaea. Such tools have succeeded in classifying these regulators in eukarya and bacteria. However, due to the particularities archaea have, a universal promoter classifier gets jeopardized.

Thus, the main objective of this study is to systematically locate the potential promoters of unannotated archaea. To do so, the use of the well-reported conserved elements belonging to an archaeal promoter are explored.

MATERIALS AND METHODS

2.1 Promoter sequences

In order to maintain balance in different archaeal families, Halobacteriales, Sulfolobales and Thermococcales were picked. These have transcriptome data available, i.e. promoter sequences can be extracted, from which we selected 3630 archaeal promoter sequences of the organisms *Haloferax volcanii* [8], *Sulfolobus solfataricus*[9] and, *Thermococcus kodakarensis* [10]. These organisms are members of two different archaea phylums. *H. volcanii* and *T. kodakarensis* belong to Euryarchaeota. This particular family has been reported as presenting more similarities to Eukaryotes[1]. Additionally, *S. solfataricus* is included in the Crenarchaeota, being known for sharing more similarities towards bacteria.

The original data contains 1001 nucleotides per sequence which contain experimentally verified promoters with their Transcription Start Site (TSS) mapped. Only primary TSS (pTSS) was considered because of data abundance. Next, a sub-sequence containing 100 nucleotides, i.e. -80 to +20 was extracted. This region comprises the reported core promoter [11, 12]. The core promoter has been reported as sufficient to initiate transcription in archaea[12]. Furthermore, the precise location of these organism's promoters was reported to be located in the proposed range [12, 13]. In addition, promoters from the three organisms tested in this study had their promoters located in the aforementioned areas [8, 9, 10].

2.2 Control datasets

The classification methods of this study were stressed with three forms of control. First, we, through a self-developed Python script, shuffled the 100-nucleotides original sequences. A second control dataset was by selecting the downstream sequences from +21 to +121. By this, we wanted to test the validity of our method by assessing sequences that do not indicate promoter activity nor have a TATA-box. Third, and finally, we followed the approach proposed by [14] to perform a second method for shuffling sequences. We divided our 100 nucleotide sequences into 5 blocks of 20 nucleotides each, then, we shuffled each one of the blocks. By doing this, we would preserve consensual motifs such as TATA-boxes in a way that our identification method is tensioned.

2.3 Structural parametrization

The totality of the sequences of this study (promoters and controls) have been submitted through a structural coding in order to represent genetic information into numeric attributes. This process has shown promising and is reflected in the capture of specific

regulatory regions such as promoters[15]. The parameter chosen for this study is DNA Duplex Stability (DDS). This particular feature has been employed as a way to represent the richness of GC base-pairs due to their extra hydrogen bond [16-20].

Equation 1 was employed in the calculation of the DNA duplexes reported in[18]. It hinges on the assignment of a numeric attribute in sliding dinucleotide windows.

$$G = \Delta^{0}_{i,i+1} \quad (1)$$

2.4 Classification through a statistical approach

Firstly, position-specific slices of 8 nucleotides (6, 2 spacer, 2) were extracted and averaged in each dataset (promoters, three controls). Then, an interval was set ranging from the plus and minus values of the standard deviation formed upon the promoter dataset (Equation 2).

$$Interval = Mean_{promoter} \pm stdev_{promoter} \quad (2)$$

Finally, a sequence was labeled as a promoter if its TATA+BRE nucleotides belongs to the rangeInterval. Otherwise, it was classified as a non-promoter. A visual representation of the statistical method for classifying archaeal promoter sequences is available at: https://doi.org/10.5281/zenodo.5154110.

2.5 Classification through an artificial neural network approach

In order to grant validation to the simulation process, a *k*-fold-cross validation method was employed, where *k*=10. This method involves in reserving 1/10 of the dataset to be used in the testing. The training is done with the remaining 9/10 shares. This grants that any biased data point gets covered. The validation process was done following the *sample* method in R [21].

Artificial Neural Network (ANN) simulations took place in the R environment through the *neuralnet* package [22]. The algorithm chosen to fit the ANN was the resilient backpropagation, since it has already succeeded in classifying genomic data [23]. The number of neurons in the hidden layer was set to 2 since a too complex curve to fit data points might be seen as a non-productive decision in machine learning [24]. The number of

iterations over the training dataset, i.e. epochs, was increased until the validation and training errors kept dropping [25]. Finally, the maximum number of steps the ANN was allowed to reach until convergence was 200000 as an attempt to balance computational costs. The R script that performed the ANN simulation is available at.

2.6 Validation of the methods

In order to provide validation for the methods proposed in this study, upstream sequences whose promoter activity has not been encountered yet have been selected in the RSAT prokaryotic database. Two archaea have presented a promoter-like profile[26]: *Aciduliprofundum boonei* and *Thermofilum pendens*. Hence, the method proposed in this study aimed to hand in a regulatory annotation upon these two organisms. The nonparametric Kruskal Wallis test was employed in order to prove if the groups of experimental and potential promoters hold statistical differences. Finally, lists of annotated potential promoters of these two organisms have been provided.

3 RESULTS

3.1 DNA duplex stability parametrized archaeal promoters differ from control sequences

In order for getting the binding sites of transcription factor proteins represented by DNA Duplex Stability (DDS), genetic information was coded into this attribute. Promoter sequences have already been well represented by DNA duplex stability (DDS) [16]. Concerning the coding of genetic information into DDS as well as locating areas of interest for turning promoters unmatched, Figure 1 has been provided. The plotting of promoter sequences and their control reveal peculiarities in the true sequences (Figure 1) in a way that the line representing promoters differs from the three levels of control. Secondly, two distinctive signals are observed in the promoter line, i.e. around position -28 and in the range of -10 to +1.

[FIGURE 1]

3.2 Statistical classification succeeds in distincting promoter sequences

In order to promote a classification method, Table 1 was created containing the mean values of TATA+BRE sites of promoter sequences as well as three levels of control converted into DDS. In every observation, the average of the promoter sequence differs

from the three levels of control, where the closer from the promoter score is the shuffled sequence, followed by the downstream, and finally, the block shuffling process.

[TABLE 1]

Table 1 enables the statistical form of classification included in this study. The promoter interval was ranged and all the sequences got their data classified into promoter or non-promoter. The results were then computed on a confusion matrix (Table 2), from which the precision value remains the same in every organism, since it's calculation relies on positive values. The assessment of Table 2 indicates a higher recall value. The most satisfactory scores were achieved by the block form of control whilst the last was found in shuffled sequences.

[TABLE 2]

In an attempt to stress the statistical method more, a cross-organism classification was conducted. In this form, the rationale of one archaeon is tested to classify data from another archaeon. In this sense, Table 3 was created, whose results have indicated a similar logistic towards *S. solfataricus* and *T. kodakarensis* and also, a high recall value when *H. volcanii* data was classified with the other two archaea principles.

[TABLE 3]

3.3 Artificial Neural Network conveys a sturdier classification

In order to achieve a more robust classification score, Artificial Neural Networks (ANNs) were used. In the ANN simulation, the architecture that protruded satisfactory scores follows: *i*) seven neurons in the input layer; *ii*) two neurons in the hidden layer and; *iii*) one neuron in the output layer. Table 4 indicates the results achieved by the ANN simulation. Table 4 was achieved with a default tradeoff value of .5 in computing the output of the model. In order to observe the behavior of the outcomes with different values, a ROC (Receiver Operator Characteristic) curve was presented (Figure 2).

[FIGURE 2]

A second application of ANNs was conducted in order to test the hypothesis that the pattern of one archaeon might be employed to classify another. Following this rationale,

a new simulation was achieved in which the ANN is trained with one organism and tested with another. The results of this new simulation are available in Table 5, from which there is a leaning towards *S. solfataricus* and *T. kodakarensis*. The *H. volcanii* logistics could have proven satisfactory in identifying promoters from other organisms.

3.4 ANN and statistics employed in finding potential archaeal promoters

Upstream regions of *Aciduliprofundum boonei* and *Thermofilum pendens* were selected in order to extract potential promoters from. The statistical and ANN models found in *S. solfataricus* and *T. kodakarensis* were employed in the validation dataset. *H. volcanii* was left out due to its unparalleled AT content; such inclusion would have jeopardized the validation. An upstream region was considered as a promoter if the statistics of *S. solfataricus* and *T. kodakarensis* and the ANN of *S. solfataricus* and *T. kodakarensis* flagged the given sequence as a promoter. From the 742 and 1927 sequences from *A. boonei* and *T. pendens*, respectively, the method encountered 145 promoters of the Euryarchaea and 243 promoters of the Crenarchaea. The lists containing sequence ID and the nucleotides are available at.

A final attempt to validate the newly identified promoters aimed to compare them with experimentally verified promoters. Hence, Figure 2 holds information of the DDS profile of *A. boonei* and *T. pendens* as well as the other three archaea. In Figure 3, there is a conserved region in the binding site of TBP, TFB and TFE proteins for all observations. A statistical analysis of the slice -40 to -1 of Figure 2 was provided in Figure 4, from which unannotated promoters of *A. boonei* resemble the averages of *S. solfataricus*, while *T. pendens* match *T. kodakarensis*. The whole analysis of the datasets present a p=3.241^-14.

[Figure 3]

[Figure 4]

4 DISCUSSION

4.1 Distinctive signals matches archaeal TFBS

The analysis of Figure 1 has portrayed the site of binding of basal transcription factor (TF) proteins. In fact, the archaeal open complex formation is initiated by a TATA-box Binding Protein (TBP) and a Transcription Factor IIB protein (TFB) [27, 28]. Moreover, a second strong signal was observed around positions -10 and +1, matching the Proximal

Promoter Element, an area where the Transcription Factor Protein E (TFE) acts and helps in order to optimize a Pre-Initiation Complex (PIC) formation [3]. Thus, the DDS conversion of archaeal promoters reveals TFBS as it has previously been reported [26].

4.2 Statistical classification

The statistical method of classification reported in this study has proven satisfactory in a way that it did not employ techniques encompassing machine learning. Firstly, the low recall value the method presented suggests the model underperformed in classifying False Negatives, this agrees with the diversification found in archaea [5]. Secondly, the most fine counts were achieved in the block form of control, matching the identification of Table 1, in which the means of the blocks are the furthest from the promoters. Additionally, the method has presented fine scores regarding recall, a metric that is sensible towards False Negatives. This phenomenon is explained due to the presence of TBP+TFB + TFBS being reported as key elements to convey archaeal transcription [26]. The stress of the statistical model, brought by an inter-archaea classification, similarities have been found in *S. solfataricus* and *T. kodakarensis*, confirming what was proposed by Takemasa *et* al (2011) [29] in order to turn *T. kodakarensis* and *S. solfataricus* as regulatory chassis for hyperthermophilic archaea. Hence, the statistical method proposed in this work poses a promising way to encounter archaeal promoters.

4.3 ANN classification

The results brought by the ANN classification suggest the model succeeded in classifying archaeal promoters, distincting them from three variations of control. In fact, this machine learning approach has succeeded in encountering promoters [14, 19, 20]. By outshining the statistical classification, the mathematical robustness of the ANN method [30] has proven uneven. Also, the rationale found in such method has matched the statistical classification, but overcame it. A good indicator to observe prediction validity is brought by ROC curves, which plots the specificity cost in gaining more recall [31]. The most evident characteristics are observed in the block control, which is found in the upper left corner of the plotting areas, confirming what the statistical analysis has found.

The verification of inter-organism rationale of classification has evidenced that *S. solfataricus* and *T. kodakarensis* share similarities, evidenced by the level of classification between these two archaea. Additionally, the ANN trained with *H. volcanii*

has outperformed the statistics classification and protruded a satisfactory way to encounter promoters of other organisms despite the higher GC content this organism promoters' have [26].

4.4 Validation

Potential promoters of *A. boonei* and *T. pendens* are a byproduct of this study. Due to many factors such as the diversity of archaea, their relatively recent discovery creates the need for high quality genome annotation. This is the moment when *in-silico* approaches provide help to experimental biology by curating data [17]. The boxplots portrayed in Figure 3 showed two groups of organisms. No taxonomic inferences could be made upon these since *T. kodakarensis* and *A. boonei* are Euryarchaea while *S. solfataricus* and *T. pendens* belong to the Crenarchaeota division, the statistical resemblance of these organisms require further analysis. We also suggest using the model of *H. volcanii* in order to locate promoters in archaea that have high AT content. The statistical similarity found between verified and potential promoters advocate the robustness of the method proposed.

5 CONCLUSIONS

The results gathered on this study reflect a novel way of encountering promoter sequences in unannotated archaeal genomes through a combination of artificial neural networks and statistics. The structural parametrization of genetic information has been able to locate key areas within upstream regions, areas that were successfully classified and its knowledge provided a new dataset of potential promoters of *A. boonei* and *T. pendens*.

ACKNOWLEDGEMENTS

FUNDING

CONFLICT OF INTEREST

REFERENCES

1. Werner, F. Structure and function of archaeal RNA polymerases. Molecular Microbiology (2007), 65(6), 1395–1404.
2. Zuo, G., Xu, Z., & Hao, B. (2015). Phylogeny and taxonomy of archaea: A comparison of the Whole-Genome-Based CVtree approach with 16s rRNA sequence analysis. Life. 5:1, 949-968.
3. Hanzelka, B. L., Darcy, T. J., & Reeve, J. N. (2001). TFE, an archaeal transcription factor in Methanobacterium thermoautotrophicum related to eucaryal transcription factor TFIIEα. *Journal of Bacteriology*. https://doi.org/10.1128/JB.183.5.1813-1818.2001.
4. DeLong, E. F., Wu, K. Y., Prézelin, B. B., & Jovine, R. V. M. (1994). High abundance of Archaea in Antarctic marine picoplankton. *Nature*. https://doi.org/10.1038/371695a0
5. Baker, B. J., De Anda, V., Seitz, K. W., Dombrowski, N., Santoro, A. E., & Lloyd, K. G. (2020). Diversity, ecology and evolution of Archaea. In *Nature Microbiology*. https://doi.org/10.1038/s41564-020-0715-z
6. Coulson, R. M. R., Touboul, N., & Ouzounis, C. A. (2007). Lineage-specific partitions in archaeal transcription. *Archaea*. https://doi.org/10.1155/2006/629868
7. Leigh, J. A., Albers, S. V., Atomi, H., & Allers, T. (2011). Model organisms for genetics in the domain Archaea: Methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS Microbiology Reviews*. https://doi.org/10.1111/j.1574-6976.2011.00265.x
8. Babski, J., Haas, K. A., Näther-Schindler, D., Pfeiffer, F., Förstner, K. U., Hammelmann, M., … Soppa, J. (2016). Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics*. https://doi.org/10.1186/s12864-016-2920-y
9. She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M. J., … Van Der Oostg, J. (2001). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.141222098
10. Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J., & Reeve, J. N. (2014). Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics*. https://doi.org/10.1186/1471-2164-15-684
11. Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/s41580-018-0028-8
12. Kadonaga, J. T. (2012). Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology*.
13. Bartlett, M. S., Thomm, M., & Geiduschek, E. P. (2000). The orientation of DNA in an archaeal transcription initiation complex. *Nature Structural Biology*. https://doi.org/10.1038/79020
14. Oubounyt, M., Louadi, Z., Tayara, H., & To Chong, K. (2019). Deepromoter: Robust promoter predictor using deep learning. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2019.00286
15. Ryasik, A., Orlov, M., Zykova, E., Ermak, T., & Sorokin, A. (2018). Bacterial promoter prediction: Selection of dynamic and static physical properties of DNA for reliable sequence classification. *Journal of Bioinformatics and Computational Biology*. https://doi.org/10.1142/S0219720018400036

16. Yella, V. R., Kumar, A., & Bansal, M. (2018). Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. *Scientific Reports*. https://doi.org/10.1038/s41598-018-22129-8
17. Martinez, G. S., de Ávila e Silva, S., Kumar, A., & Pérez-Rueda, E. (2021). DNA structural and physical properties reveal peculiarities in promoter sequences of the bacterium Escherichia coli K-12. *SN Applied Sciences*. https://doi.org/10.1007/s42452-021-04713-2
18. SantaLucia, J., & Hicks, D. (2004). The Thermodynamics of DNA Structural Motifs. *Annual Review of Biophysics and Biomolecular Structure*. https://doi.org/10.1146/annurev.biophys.32.110601.141800
19. Kanhere, A., & Bansal, M. (2005). Structural properties of promoters: Similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gki627
20. de Avila e Silva, S., Echeverrigaray, S., & Gerhardt, G. J. L. (2011). BacPP: Bacterial promoter prediction-A tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of Theoretical Biology*. https://doi.org/10.1016/j.jtbi.2011.07.017
21. Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society: Series B (Methodological). https://doi.org/10.1111/j.2517-6161.1974.tb00994.x
22. Beck, M. W. (2018). NeuralNetTools: Visualization and analysis tools for neural networks. Journal of Statistical Software. https://doi.org/10.18637/jss.v085.i11
23. Liu, X., Guo, Z., He, T., & Ren, M. (2020). Prediction and analysis of prokaryotic promoters based on sequence features. BioSystems. https://doi.org/10.1016/j.biosystems.2020.104218
24. Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*. https://doi.org/10.1162/neco.1992.4.1.1
25. Afaq, S., & Rao, S. (2020). Significance Of Epochs On Training A Neural Network. International Journal of Scientific and Technology Research.
26. Martinez, Sarkar, Kumar, et al. Characterization of promoters in archaeal genomes based on DNA structural parameters. *Authorea.* July 22, 2021. DOI: 10.22541/au.162698044.49256534/v1
27. Soppa, J. (1999). Transcription initiation in Archaea: Facts, factors and future aspects. Mol Microbiol. 31:5, doi: 10.1046/j.1365-2958.1999.01273.x
28. Smollet, K., Blombach, F., Fouqueau, T., Wernerm F. (2017) A Global Characterisation of the Archaeal Transcription Machinery. In: Clouet-dO'rval, B. (eds) RNA metabolism and Gene Expression in Archaea. Nucleic Acids Mol Biol, 32.
29. Takemasa, R., Yokooji, Y., Yamatsu, A., Atomi, H., & Imanaka, T. (2011). Thermococcus kodakarensis as a host for gene expression and protein secretion. *Applied and Environmental Microbiology*. https://doi.org/10.1128/AEM.01005-10
30. Mangal, R., Nori, A. V., & Orso, A. (2019). Robustness of neural networks: A probabilistic and practical approach. *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results, ICSE-NIER 2019*. https://doi.org/10.1109/ICSE-NIER.2019.00032
31. Xu, Y., Wang, X. B., Ding, J., Wu, L. Y., & Deng, N. Y. (2010). Lysine acetylation sites prediction using an ensemble of support vector machine

classifiers. *Journal of Theoretical Biology*. https://doi.org/10.1016/j.jtbi.2010.01.013

32. Eme, L., Spang, A., Lombard, J., Stairs, C. W., & Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*. https://doi.org/10.1038/nrmicro.2017.133
33. Fouqueau, T., Blombach, F., Cackett, G., Carty, A. E., Matelska, D. M., Ofer, S., … Werner, F. (2018). The cutting edge of archaeal transcription. *Emerging Topics in Life Sciences*. https://doi.org/10.1042/ETLS20180014
34. Martinez-Pastor, M., Tonner, P. D., Darnell, C. L., & Schmid, A. K. (2017). Transcriptional Regulation in Archaea: From Individual Genes to Global Regulatory Networks. *Annual Review of Genetics*. https://doi.org/10.1146/annurev-genet-120116-023413

TABLES

Table 1 - Flagships of archaeal classification based on statistics

	Promoters	Standard deviation	Blocks	Downstream	Shuffled
H. volcanii	-8.36	±1.15	-11.66	-11.44	-10.66
S. solfataricus	-6.72	±0.85	-8.87	-8.58	-8.73
T. kodakarensis	-7.13	±0.93	-10.16	-10.07	-9.28

Table 2 - Results of the statistical method of classification

		Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
H. volcanii	Blocks	79.13	67.81	87.64	73.76
	Downstream	78.76	67.81	86.8	73.6
	Shuffled	72.04	67.81	74.06	70.33
S. solfataricus	Blocks	78.55	77.44	79.19	77.93
	Downstream	78.22	77.44	78.65	77.8
	Shuffled	78.55	77.44	79.19	77.93
T. kodakarensis	Blocks	81.37	70.11	90.48	75.6
	Downstream	81.16	70.11	90.02	75.52
	Shuffled	74.63	70.11	77.09	72.59

Table 3 - Results of the inter-organism statistical method of classification

		H. volcanii	*S. solfataricus*	*T. kodakarensis*
H. volcanii	Accuracy (%)	-	61.68	68.92
	Precision (%)	-	24.7	40.89
	Recall (%)	-	95.05	93.34
	Specificity (%)	-	56.71	62.11
S. solfataricus	Accuracy (%)	32.17	-	78.44
	Precision (%)	21.78	-	77.44
	Recall (%)	27.51	-	79.01
	Specificity (%)	35.23	-	77.88
T. kodakarensis	Accuracy (%)	50.47	82.9	-
	Precision (%)	40.62	72.43	-
	Recall (%)	50.99	91.86	-
	Specificity (%)	50.21	77.17	-

Table 4 - Results of the ANN-based classification.

		Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
H. volcanii	Blocks	92.48	93.05	92.03	92.96
	Downstream	91.08	90.67	91.45	90.77
	Shuffled	84.55	84.86	84.27	84.94
S. solfataricus	Blocks	89.03	91.43	87.01	91.18
	Downstream	87.36	86.93	88.23	86.48
	Shuffled	86.63	84.56	88.27	85.17
T. kodakarensis	Blocks	94.96	93.39	96.21	93.83
	Downstream	91.35	91.69	91.31	91.46
	Shuffled	86.46	84.1	89.12	84.36

		H. volcanii	*S. solfataricus*	*T. kodakarensis*
H. volcanii	Accuracy (%)	-	70.63	81.61
	Precision (%)	-	41.75	66.56
	Recall (%)	-	95.95	94.68
	Specificity (%)	-	63.89	74.72
S. solfataricus	Accuracy (%)	66.51	-	86.46
	Precision (%)	98.89	-	95.45
	Recall (%)	60.87	-	81.04
	Specificity (%)	96.23	-	94.42
T. kodakarensis	Accuracy (%)	75.42	82.48	-
	Precision (%)	97.52	71.56	-
	Recall (%)	68.9	91.48	-
	Specificity (%)	94.34	77.01	-

Figures

Figure 1

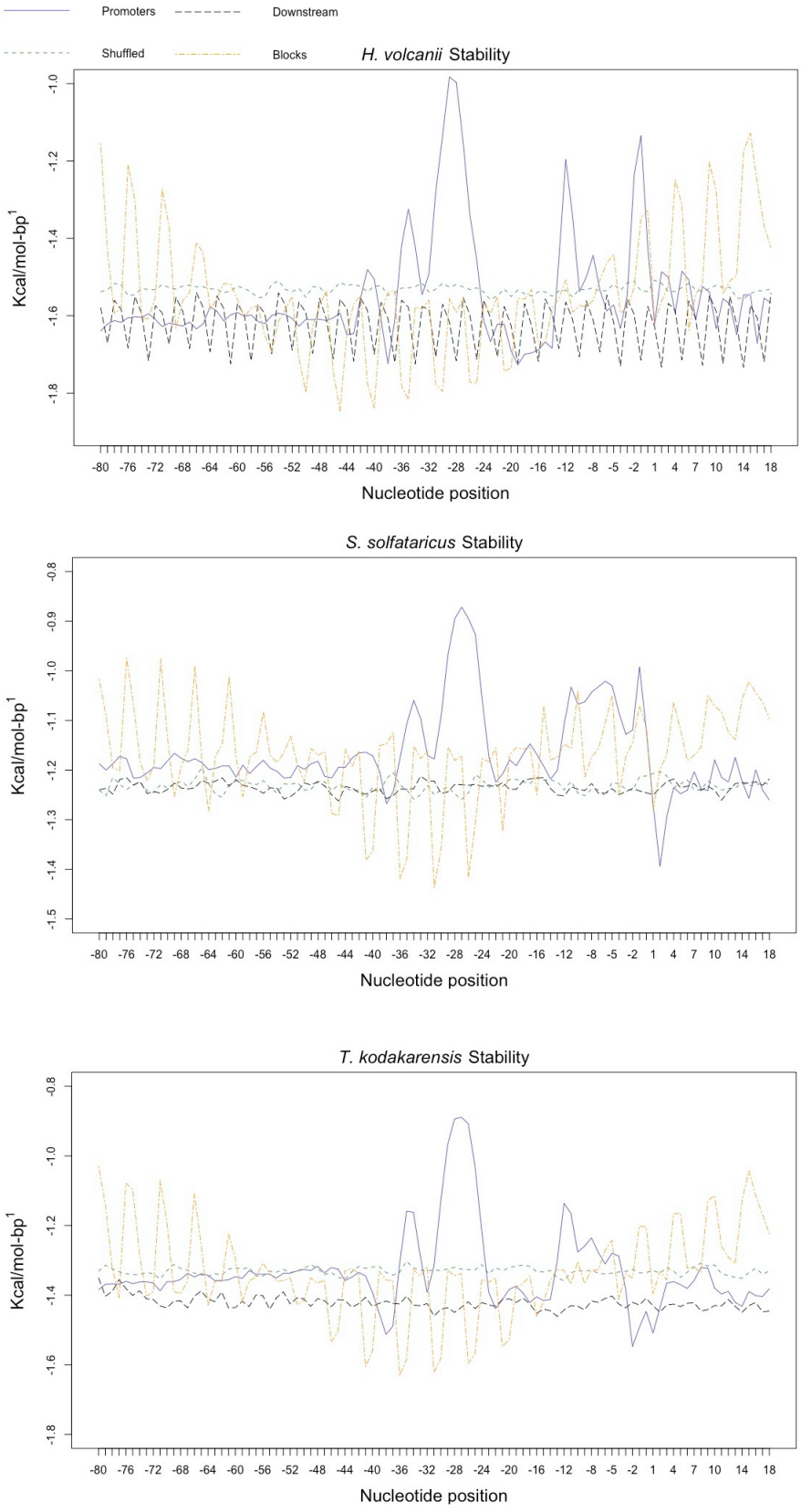


Figure 2

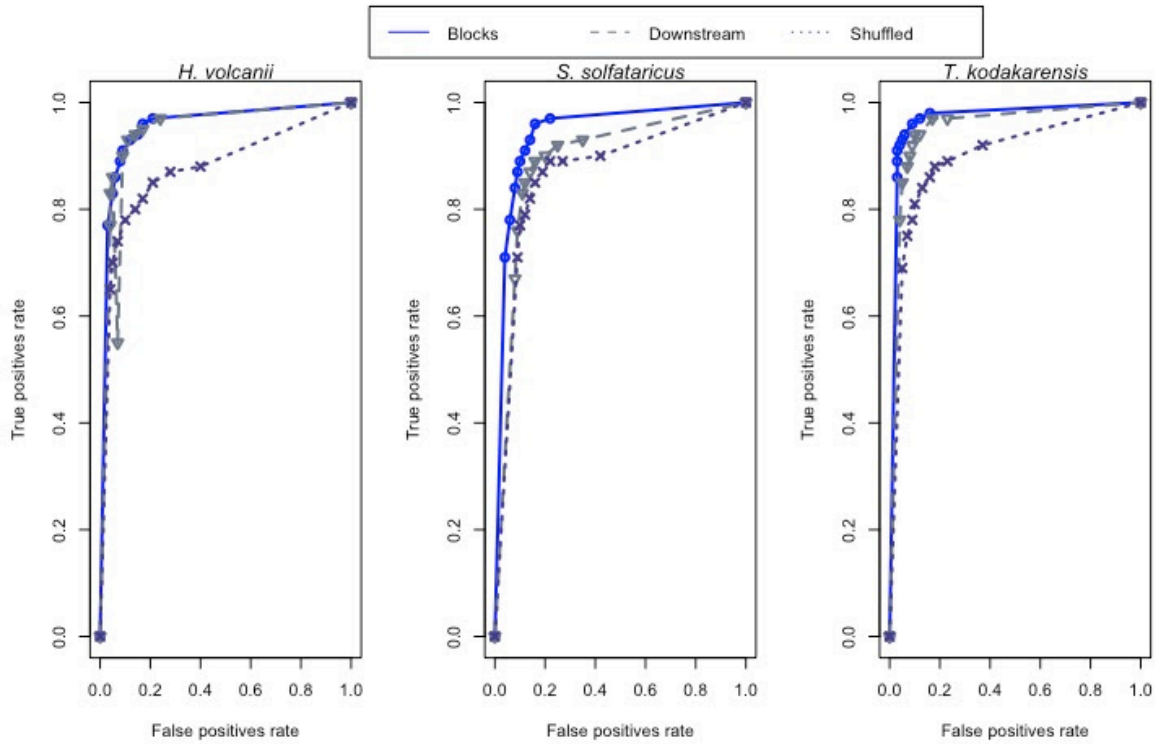


Figure 3

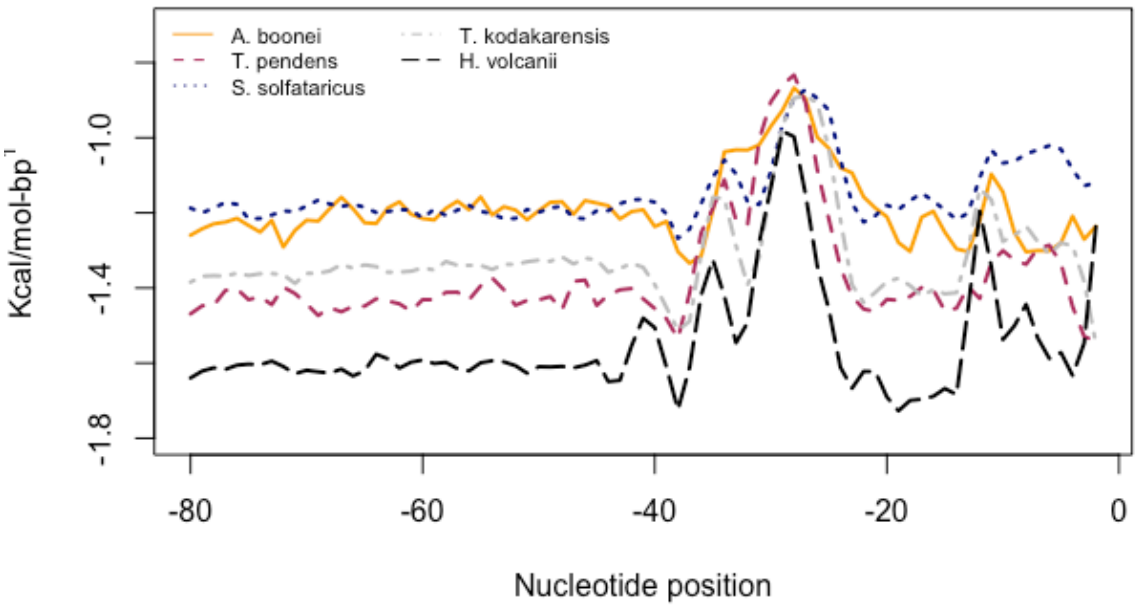
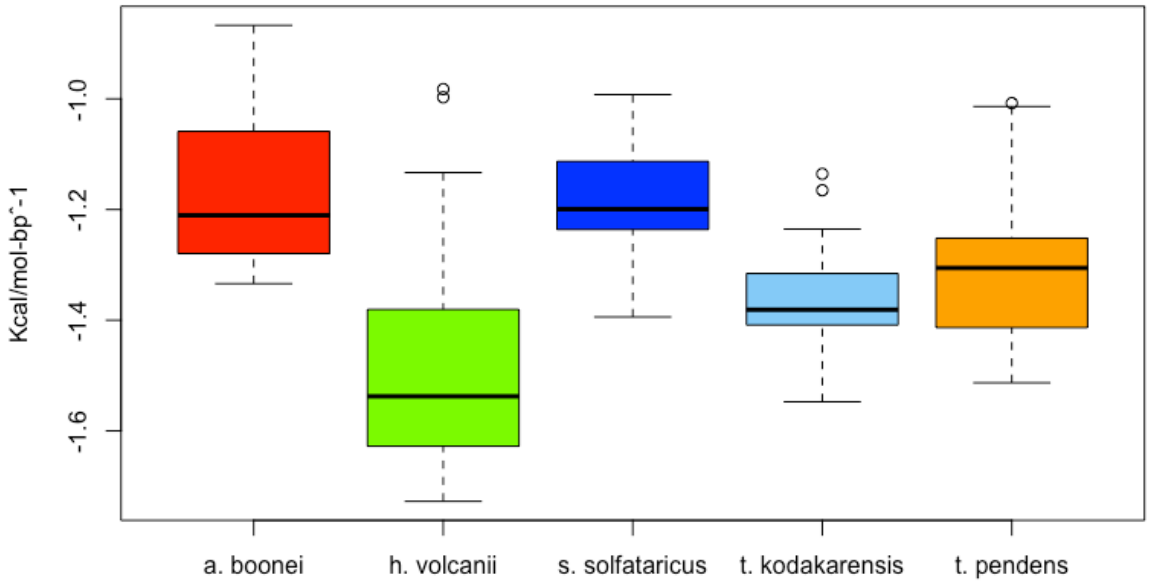


Figure 4


Kcal/mol-bp^-1
-1.0
-1.2
-1.4
-1.6
a. boonei
h. volcanii
s. solfataricus
t. kodakarensis
t. pendens

6 CONCLUSIONS

Archaea have been discovered to be a unique domain not long ago. There are many hypothesis being tested in order to understand how eukaryotes evolved and the comprehension of archaeal mechanisms might play an important role on this achievement.

In this thesis, through a characterization of archaeal promoter elements, it was possible to show the dissimilarities these organisms have in their transcription factor binding sites. Moreover, a conversion of genetic information into structural parameters was able to provide different analysis where a primary sequence inspection failed.

The results gathered in this thesis are being employed in order to create a predictor of archaeal promoter sequences, based both on statistics and machine learning.

The discovery of archaeal genetic mechanisms such as their broader comprehension is still an important mark in bioinformatics. The advances of computer science enable hypothesis formulation and testing.

Therefore, the quest of exploring the unknown, i.e. archaea, remains. Through the curated results achieved in this thesis it was possible to shed some light and continue the endeavor to explore the innards of microbiology.

7 REFERENCES

Adami, A, G., & Adami, A, M. Análise por agrupamentos. In: de Avila e Silva, S., Notari, D. L., & Dall'Alba, G. (2020). *Bioinformática contexto computacional e aplicações* (1st ed.). Caxias do Sul: EDUCS.

Afaq, S., & Rao, S. (2020). Significance Of Epochs On Training A Neural Network. *International Journal of Scientific and Technology Research*.

Aptekmann, A. A., & Nadra, A. D. (2018). Core promoter information content correlates with optimal growth temperature. *Scientific Reports*. https://doi.org/10.1038/s41598-018-19495-8

Akan, P., & Deloukas, P. (2008). DNA sequence and structural properties as predictors of human and mouse promoters. *Gene*. https://doi.org/10.1016/j.gene.2007.12.011

Aravind, L. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Research*. https://doi.org/10.1093/nar/27.23.4658

Babski, J., Haas, K. A., Näther-Schindler, D., Pfeiffer, F., Förstner, K. U., Hammelmann, M., … Soppa, J. (2016). Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics*. https://doi.org/10.1186/s12864-016-2920-y

Baldi, P.; Brunak, S. (2001). *Bioinformatics: the machine learning approach* (2nd ed). MIT.

Basehoar, A. D., Zanton, S. J., & Pugh, B. F. (2004). Identification and distinct regulation of yeast TATA box-containing genes. *Cell*. https://doi.org/10.1016/S0092-8674(04)00205-3

Beck, M. W. (2018). NeuralNetTools: Visualization and analysis tools for neural networks. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v085.i11

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1989). The New S Language. *Biometrics*. https://doi.org/10.2307/2531523

Bedoya, Ó., & Bustamante, S. (2011). CNN-PROMOTER, NUEVO PROGRAMA PARA LA PREDICCIÓN DE PROMOTORES BASADO EN REDES NEURONALES. *Revista EIA*.

Bell, S. D., & Jackson, S. P. (2001). Mechanism and regulation of transcription in archaea. *Current Opinion in Microbiology*. https://doi.org/10.1016/S1369-5274(00)00190-9

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: Receiver operating characteristics curves. *Critical Care*. https://doi.org/10.1186/cc3000

Bi, W., Wu, L., Coustry, F., De Crombrugghe, B., & Maity, S. N. (1997). DNA binding specificity of the CCAAT-binding factor CBF/NF-Y. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.272.42.26562

Bolshoy, A., McNamara, P., Harrington, R. E., & Trifonov, E. N. (1991). Curved DNA without A-A: Experimental estimation of all 16 DNA wedge angles. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.88.6.2312

Bowerman, S., Wereszczynski, J., & Luger, K. (2021). Archaeal chromatin 'slinkies' are inherently dynamic complexes with deflected dna wrapping pathways. *ELife*. https://doi.org/10.7554/eLife.65587

Blombach, F., & Grohmann, D. (2017). Same same but different: The evolution of TBP in archaea and their eukaryotic offspring. *Transcription*. https://doi.org/10.1080/21541264.2017.1289879

Browning, D. F., & Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*. https://doi.org/10.1038/nrmicro.2016.103

Brunk, C. F., & Martin, W. F. (2019). Archaeal Histone Contributions to the Origin of Eukaryotes. *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2019.04.002

Buchanan, M. (2011). The irregularity of reality. *Nature Physics*. https://doi.org/10.1038/nphys1943

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., & Chou, K. C. (2015). PseKNC-General: A cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu602

Christian, B. (2011). The most human human: what talking with computers teaches us about what it means to be alive (1st ed). Anchor.

Calladine, C. R., Drew, H. R., & McCall, M. J. (1988). The intrinsic curvature of DNA in solution. *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(88)90444-5

Clancy, S. (2008) DNA transcription. *Nature Education* 1(1):41

Conrad, C., & Gerlich, D. W. (2010). Automated microscopy for high-content RNAi screening. *Journal of Cell Biology*. https://doi.org/10.1083/jcb.200910105

Davies, P. C. W., Rieper, E., & Tuszynski, J. A. (2013). Self-organization and entropy reduction in a living cell. *BioSystems*. https://doi.org/10.1016/j.biosystems.2012.10.005

De Avila e Silva, S., Forte, F., T.S. Sartor, I., Andrighetti, T., J.L. Gerhardt, G., Longaray Delamare, A. P., & Echeverrigaray, S. (2014). DNA duplex stability as

discriminative characteristic for Escherichia coli σ54- and σ28- dependent promoter sequences. *Biologicals*. https://doi.org/10.1016/j.biologicals.2013.10.001

De Avila e Silva, S., & Coelho, R., Redes Neurais Artificiais: Introdução e Definições. In: de Avila e Silva, S., Notari, D. L., & Dall'Alba, G. (2020). *Bioinformática contexto computacional e aplicações* (1st ed.). Caxias do Sul: EDUCS.

De Avila e Silva, S., Notari, D. L., & Dall'Alba, G. (2020). *Bioinformática contexto computacional e aplicações* (1st ed.). Caxias do Sul: EDUCS.

de Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-013-0386-3

Dlakic, M., Ussery, D. W., Søren, B., In: Ohyama, T. (2005). DNA Conformation and Transcription. In *DNA Conformation and Transcription*. https://doi.org/10.1007/0-387-29148-2

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. *New York: John Wiley, Section*.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*. https://doi.org/10.1080/01969727408546059

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq461

Forterre, P., Brochier, C., & Philippe, H. (2002). Evolution of the Archaea. *Theoretical Population Biology*. https://doi.org/10.1006/tpbi.2002.1592

Fouqueau, T., Blombach, F., Cackett, G., Carty, A. E., Matelska, D. M., Ofer, S., … Werner, F. (2018). The cutting edge of archaeal transcription. *Emerging Topics in Life Sciences*. https://doi.org/10.1042/ETLS20180014

French, S. L., Santangelo, T. J., Beyer, A. L., & Reeve, J. N. (2007). Transcription and translation are coupled in Archaea. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msm007

Friedel, M., Nikolajewa, S., Sühnel, J., & Wilhelm, T. (2009). DiProDB: A database for dinucleotide properties. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkn597

Friedman, R. A., & Honig, B. (1995). A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophysical Journal*. https://doi.org/10.1016/S0006-3495(95)80023-8

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts565

Gehring, A. M., Walker, J. E., & Santangelo, T. J. (2016). Transcription regulation in archaea. *Journal of Bacteriology*. https://doi.org/10.1128/JB.00255-16

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*. https://doi.org/10.1162/neco.1992.4.1.1

Ghorbani, M., & Mohammad-Rafiee, F. (2011). Geometrical correlations in the nucleosomal DNA conformation and the role of the covalent bonds rigidity. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq952

Gribaldo, S., & Brochier-Armanet, C. (2006). The origin and evolution of Archaea: A state of the art. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2006.1841

Gross, R. (2015). Psychology: The science of mind and behaviour (7th ed.). *Psychology: The Science of Mind and Behaviour (7th Ed.)*.

Gohlke, C. (2020). DNA Curvature Analysis 2020.1.1. [Computer Software]. Retrieved from https://www.lfd.uci.edu/~gohlke/dnacurve/

Govek, K. W., Yamajala, V. S., & Camara, P. G. (2019). Clustering-independent analysis of genomic data using spectral simplicial theory. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1007509

Gupta, R. S. (1998). What are archaebacteria: Life's third domain or monoderm prokaryotes related to Gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. *Molecular Microbiology*. https://doi.org/10.1046/j.1365-2958.1998.00978.x

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/s41580-018-0028-8

Hallam, S. J. (2019). MICROCOSMOS. In *Memory*. https://doi.org/10.2307/j.ctvbtzpfm.25

Han, H., & Liu, W. (2019). The coming era of artificial intelligence in biological data science. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-019-3225-3

Häse, F., & Zacharias, M. (2016). Free energy analysis and mechanism of base pair stacking in nicked DNA. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw607

Haykin, S. (1999). Neural Networks: A Comprehensive Foundation (3rd Edition). In *The Knowledge Engineering Review*.

Henneman, B., van Emmerik, C., van Ingen, H., & Dame, R. T. (2018). Structure and function of archaeal histones. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1007582

Herzel, H., Ebeling, W., & Schmitt, A. O. (1994). Entropies of biosequences: The role of repeats. *Physical Review E*. https://doi.org/10.1103/PhysRevE.50.5061

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*. https://doi.org/10.1016/0893-6080(89)90020-8

Huang, G. Bin, Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2005.12.126

Igel, C., & Hüsken, M. (2003). Empirical evaluation of the improved Rprop learning algorithms. *Neurocomputing*. https://doi.org/10.1016/S0925-2312(01)00700-7

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., … Takai, K. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature*. https://doi.org/10.1038/s41586-019-1916-6

Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J., & Reeve, J. N. (2014). Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics*. https://doi.org/10.1186/1471-2164-15-684

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2009.09.011

Jáuregui, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., Collado-Vides, J., & Merino, E. (2003). Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkg882

Jernigan, R. L. J., Sarai, A., Shapiro, B., & Nussinov, R. (1987). Relationship between curved dna conformations and slow gel migration. *Journal of Biomolecular Structure and Dynamics*. https://doi.org/10.1080/07391102.1987.10507660

Joshi, A. V. (2020). Machine Learning and Artificial Intelligence. In *Machine Learning and Artificial Intelligence*. https://doi.org/10.1007/978-3-030-26622-6

Kanhere, A., & Bansal, M. (2003). An assessment of three dinucleotide parameters to predict DNA curvature by quantitative comparison with experimental data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkg362

Kanhere, A., & Bansal, M. (2005). Structural properties of promoters: Similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gki627

Karas, H., Knuppel, R., Schuiz, W., Sklenar, H., & Wingender, E. (1996). Combining structural analysis of dna with search routines for the detection of transcription regulatory elements. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/12.5.441

Kernan, T., West, A. C., & Banta, S. (2017). Characterization of endogenous promoters for control of recombinant gene expression in Acidithiobacillus ferrooxidans. *Biotechnology and Applied Biochemistry*. https://doi.org/10.1002/bab.1546

Klauck, H. A., Dall'Alba, G., Avila e Silva, S. de, & Longaray Delamare, A. P. (2020). Prediction and Recognition of Gram-Negative Bacterial Promoter Sequences: An Analysis of Available Web Tools. *Journal of Biotechnology Research*. https://doi.org/10.32861/jbr.67.90.97

Knudsen, S. (1999). Promoter2.0: For the recognition of PolII promoter sequences. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/15.5.356

Koo, H. S., Wu, H. M., & Crothers, D. M. (1986). DNA bending at adenine · thymine tracts. *Nature*. https://doi.org/10.1038/320501a0

Koonin, E. V., & Wolf, Y. I. (2008). Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkn668

Koonin, E. V., & Galperin, M. Y. (2010). Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. In *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics.*

Kozobay-Avraham, L., Hosid, S., & Bolshoy, A. (2004). Curvature distribution in prokaryotic genomes. *In Silico Biology*.

Krebs, Jocelyn E.; Goldstein, Elliot S.; Kilpatrick, S. T. (2017). *Lewin's Gene XII* (12th ed.). Jones & Bartlett Learning.

Kumar, A., Manivelan, V., & Bansal, M. (2016). Structural features of DNA are conserved in the promoter region of orthologous genes across different strains of Helicobacter pylori. *FEMS Microbiology Letters*. https://doi.org/10.1093/femsle/fnw207

Kuok, C. F., Hoi, S. O., Hoi, C. F., Chan, C. H., Fong, I. H., Ngok, C. K., … Fong, P. (2017). Synergistic antibacterial effects of herbal extracts and antibiotics on methicillin-resistant Staphylococcus aureus: A computational and experimental study. *Experimental Biology and Medicine*. https://doi.org/10.1177/1535370216689828

Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., & Ebright, R. H. (1998). New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB. *Genes and Development*. https://doi.org/10.1101/gad.12.1.34

Langer, D., Hain, J., Thuriaux, P., & Zillig, W. (1995). Transcription in archaea: Similarity to that in eucarya. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.92.13.5768

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl158

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., & Chou, K. C. (2015). Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv458

Liu, X., Guo, Z., He, T., & Ren, M. (2020). Prediction and analysis of prokaryotic promoters based on sequence features. *BioSystems*. https://doi.org/10.1016/j.biosystems.2020.104218

Lloréns-Rico, V., Lluch-Senar, M., & Serrano, L. (2015). Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv170

López-García, P., & Forterre, P. (1999). Control of DNA topology during thermal stress in hyperthermophilic archaea: DNA topoisomerase levels, activities and induced thermotolerance during heat and cold shock in Sulfolobus. *Molecular Microbiology*. https://doi.org/10.1046/j.1365-2958.1999.01524.x

Marmur, J., & Doty, P. (1962). Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *Journal of Molecular Biology*. https://doi.org/10.1016/S0022-2836(62)80066-7

Martinez-Pastor, M., Tonner, P. D., Darnell, C. L., & Schmid, A. K. (2017). Transcriptional Regulation in Archaea: From Individual Genes to Global Regulatory Networks. *Annual Review of Genetics*. https://doi.org/10.1146/annurev-genet-120116-023413

Mattiroli, F., Bhattacharyya, S., Dyer, P. N., White, A. E., Sandman, K., Burkhart, B. W., … Luger, K. (2017). Structure of histone-based chromatin in Archaea. *Science*. https://doi.org/10.1126/science.aaj1849

Maus, D., Heinz, J., Schirmack, J., Airo, A., Kounaves, S. P., Wagner, D., & Schulze-Makuch, D. (2020). Methanogenic Archaea Can Produce Methane in Deliquescence-Driven Mars Analog Environments. *Scientific Reports*. https://doi.org/10.1038/s41598-019-56267-4

Meysman, P., Collado-Vides, J., Morett, E., Viola, R., Engelen, K., & Laukens, K. (2014). Structural properties of prokaryotic promoter regions correlate with functional features. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0088717

Mishra, A., Mishra, A., Dhanda, S., Siwach, P., Siwach, P., Aggarwal, S., … Jayaram, B. (2020). A novel method SEProm for prokaryotic promoter prediction based on DNA structure and energetics. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz941

Mount, David W. Bioinformatics: sequence and genome analysis. New York: Cold SpringHarbor Laboratory, 2000.

Nagaich, A. K., Bhattacharyya, D., Brahmachari, S. K., & Bansal, M. (1994). CA/TG sequence at the 5' end of oligo(A)-tracts strongly modulates DNA curvature. *The Journal of biological chemistry*, *269*(10), 7824–7833

Ning, L. W., Lin, H., Ding, H., Huang, J., Rao, N., & Guo, F. B. (2014). Predicting bacterial essential genes using only sequence composition information. *Genetics and Molecular Research*. https://doi.org/10.4238/2014.June.17.8

Nkamga, V. D., Henrissat, B., & Drancourt, M. (2017). Archaea: Essential inhabitants of the human digestive microbiota. *Human Microbiome Journal*. https://doi.org/10.1016/j.humic.2016.11.005

Olivares-Zavaleta, N., Jáuregui, R., & Merino, E. (2006). Genome analysis of Escherichia coli promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. *Genomics*. https://doi.org/10.1016/j.ygeno.2005.11.023

Olsen, G. J., & Woese, C. R. (1996). Lessons from an Archaeal genome: What are we learning from Methanococcus jannaschii? *Trends in Genetics*. https://doi.org/10.1016/0168-9525(96)30092-9

Ornstein, R. L., Rein, R., Breen, D. L., & Macelroy, R. D. (1978). An optimized potential function for the calculation of nucleic acid interaction energies I. Base stacking. *Biopolymers*. https://doi.org/10.1002/bip.1978.360171005

Oubounyt, M., Louadi, Z., Tayara, H., & To Chong, K. (2019). Deepromoter: Robust promoter predictor using deep learning. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2019.00286

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Aameh, F., … Aadebiyi, E. (2016). Clustering algorithms: Their application to gene expression data. *Bioinformatics and Biology Insights*. https://doi.org/10.4137/BBI.S38316

Ozoline, O. N., Deev, A. A., & Trifonov, E. N. (1999). Dna bendability—;a novel feature in e. coli promoter recognition. *Journal of Biomolecular Structure and Dynamics*. https://doi.org/10.1080/07391102.1999.10508295

Pace, N. R. (2006). Time for a change. *Nature*. https://doi.org/10.1038/441289a

Patil, C., & Baidari, I. (2019). Estimating the Optimal Number of Clusters k in a Dataset Using Data Depth. *Data Science and Engineering*. https://doi.org/10.1007/s41019-019-0091-y

Pedersen, A. G., Baldi, P., Chauvin, Y., Søren, B. (1998). DNA structure in human RNA polymerase II promoters. *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1998.1972.

Peres, D. J., Iuppa, C., Cavallaro, L., Cancelliere, A., & Foti, E. (2015). Significant wave height record extension by neural networks and reanalysis wind data. *Ocean Modelling*. https://doi.org/10.1016/j.ocemod.2015.08.002

Powers, D. M. W. (2007). Evaluation: from precision, recall and f-factor. *Technical Report SEI-07-001*.

Privalov, P. L., & Crane-Robinson, C. (2018). Forces maintaining the DNA double helix and its complexes with transcription factors. *Progress in Biophysics and Molecular Biology*. https://doi.org/10.1016/j.pbiomolbio.2018.01.007

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.72.3.784

Putta, P., & Mitra, C. K. (2010). Conserved short sequences in promoter regions of human genome. *Journal of Biomolecular Structure and Dynamics*. https://doi.org/10.1080/07391102.2010.10508574

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Riedmiller, M., Braun, H. A direct adaptive method for faster backpropagation learning: the RPROPalgorithm, in: E.H. Ruspini (Ed.), Proceedings of the IEEE International Conference on NeuralNetworks, IEEE Press, New York, 1993, pp. 586–591.

Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons - From backpropagation to adaptive learning algorithms. *Computer Standards and Interfaces*. https://doi.org/10.1016/0920-5489(94)90017-5

Rangannan, V., & Bansal, M. (2007). Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. *Journal of Biosciences*. https://doi.org/10.1007/s12038-007-0085-1

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*. https://doi.org/10.7717/peerj.2584

Rosen, R.,1991. Life Itself: A Comprehensive Inquiry into the Nature, Origin and Fabrication of Life. Columbia University Press, New York.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*. https://doi.org/10.1037/h0042519

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. https://doi.org/10.1016/0377-0427(87)90125-7

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*. https://doi.org/10.1038/323533a0

Russel, S., & Norvig, P. (2012). Artificial intelligence—a modern approach 3rd Edition. In *The Knowledge Engineering Review*. https://doi.org/10.1017/S0269888900007724

Sagan, L. (1967). On the origin of mitosing cells. *Journal of Theoretical Biology*. https://doi.org/10.1016/0022-5193(67)90079-3

Samson, R. Y., & Bell, S. D. (2014). Archaeal chromosome biology. *Journal of Molecular Microbiology and Biotechnology*. https://doi.org/10.1159/000368854

SantaLucia, J., & Hicks, D. (2004). The Thermodynamics of DNA Structural Motifs. *Annual Review of Biophysics and Biomolecular Structure*. https://doi.org/10.1146/annurev.biophys.32.110601.141800

Santangelo, T. J., & Reeve, J. N. (2006). Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2005.10.06

Sarle, W. S., Jain, A. K., & Dubes, R. C. (1990). Algorithms for Clustering Data. *Technometrics*. https://doi.org/10.2307/1268876

Satchwell, S. C., Drew, H. R., & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(86)90452-3

Shahmuradov, I. A., Mohamad Razali, R., Bougouffa, S., Radovanovic, A., & Bajic, V. B. (2017). bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli. *Bioinformatics (Oxford, England)*. https://doi.org/10.1093/bioinformatics/btw629

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Solovyev, V. V., & Shahmuradov, I. A. (2003). PromH: Promoters identification using orthologous genomic sequences. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkg525

Spang, A., Eme, L., Saw, J. H., Caceres, E. F., Zaremba-Niedzwiedzka, K., Lombard, J., … Ettema, T. J. G. (2018). Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1007080

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*. https://doi.org/10.1016/S0034-4257(97)00083-7

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Sun, Y., Liu, Y., Pan, J., Wang, F., & Li, M. (2020). Perspectives on Cultivation Strategies of Archaea. *Microbial Ecology*. https://doi.org/10.1007/s00248-019-01422-7

Swanepoel, A. C. (2007). Irregular regularity: A chaos-theory reading of the ecology presented in coleridge's "frost at midnight." *Journal of Literary Studies*. https://doi.org/10.1080/02564710701786483

Theodoridis, S., & Koutroumbas, K. (2006). Pattern Recognition, Third Edition. *Publishing Research Quarterly*.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*. https://doi.org/10.1007/BF02289263

Trifonov, E. N., & Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.77.7.3816

Vlachos, A. (2011). Evaluating unsupervised learning for natural language processing tasks. *Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP*.

Vogel, U., & Jensen, K. F. (1994). The RNA chain elongation rate in Escherichia coli depends on the growth rate. *Journal of Bacteriology*. https://doi.org/10.1128/jb.176.10.2807-2813.1994

Wade, J. T., & Grainger, D. C. (2014). Pervasive transcription: Illuminating the dark matter of bacterial transcriptomes. *Nature Reviews Microbiology*. https://doi.org/10.1038/nrmicro3316

Wang, H., Noordewier, M., & Benham, C. J. (2004). Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters. *Genome Research*. https://doi.org/10.1101/gr.2080004

Wang, H., & Benham, C. J. (2006). Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-7-248

Watson, J. D., & Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Journal of the American College of Cardiology*. https://doi.org/10.1016/s0735-1097(03)00800-3

Werner, F. (2007). Structure and function of archaeal RNA polymerases. *Molecular Microbiology*. https://doi.org/10.1111/j.1365-2958.2007.05876.x

Wittig, S., & Wittig, B. (1982). Function of a tRNA gene promoter depends on nucleosome position. *Nature*. https://doi.org/10.1038/297031a0

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.74.11.5088

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*. https://doi.org/10.1128/mmbr.51.2.221-271.1987

Wu, C. H. (1997). Artificial neural networks for molecular sequence analysis. *Computers and Chemistry*. https://doi.org/10.1016/S0097-8485(96)00038-1

Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B. A., & Sorek, R. (2010). A single-base resolution map of an archaeal transcriptome. *Genome Research*. https://doi.org/10.1101/gr.100396.109

Yakovchuk, P., Protozanova, E., & Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkj454

Yella, V. R., & Bansal, M. (2017). DNA structural features of eukaryotic TATA-containing and TATA-less promoters. *FEBS Open Bio*, *7*(3). https://doi.org/10.1002/2211-5463.12166

Yella, V. R., Kumar, A., & Bansal, M. (2018). Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. *Scientific Reports*. https://doi.org/10.1038/s41598-018-22129-8

Zhang, T. B., Zhang, C. L., Dong, Z. L., & Guan, Y. F. (2015). Determination of Base Binding Strength and Base Stacking Interaction of DNA Duplex Using Atomic Force Microscope. *Scientific Reports*. https://doi.org/10.1038/srep09143

Zilling, W., Stetter, K. O., & Janecovic, D. (1979). DNA-Dependent RNA Polymerase from the Archaebacterium Sulfolobus acidocaldarius. *European Journal of Biochemistry*. https://doi.org/10.1111/j.1432-1033.1979.tb13074.x

Zivanovic, Y. (2002). Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Research*. https://doi.org/10.1093/nar/30.9.1902

Zhou, X., Li, Z., Dai, Z., & Zou, X. (2011). Predicting methylation status of human DNA sequences by pseudo-trinucleotide composition. *Talanta*. https://doi.org/10.1016/j.talanta.2011.05.043

Zou, Q., Lin, G., Jiang, X., Liu, X., & Zeng, X. (2018). Sequence clustering in bioinformatics: An empirical study. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bby090

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*. https://doi.org/10.1016/0022-5193(65)90083-4