UNIVERSIDADE DE CAXIAS DO SUL
ÁREA DE CONHECIMENTO DE CIÊNCIAS DA VIDA
INSTITUTO DE BIOTECNOLOGIA
PROGRAMA DE PÓS GRADUAÇÃO EM BIOTECNOLOGIA

Promoter sequence characterization through the analysis of enthalpy, entropy, stability and base-pair stacking values

Gustavo Sganzerla Martinez

Caxias do Sul

2018

GUSTAVO SGANZERLA MARTINEZ

Promoter sequences classification through the analysis of enthalpy, entropy, stability and base-pair stacking values

Dissertação apresentada ao Programa de Pós Graduação em Biotecnologia na Universidade de Caxias do Sul, visando a obtenção do título de Mestre em Biotecnologia

Orientadora: Prof. Dra. Scheila de Ávila e Silva

DISSERTAÇÃO APROVADA EM 19 DE OUTUBRO DE 2018.

Orientadora Prof. Dra. Scheila de Ávila e Silva

Prof. Dr. Sérgio Echeverrigaray

Prof. Dr. Julio Collado-Vides

Prof. Dr. Luis Fernando Saraiva Macedo Timmers

Caxias do Sul

2018

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

M385p Martinez, Gustavo Sganzerla
Promoter sequences classification through the analysis of enthalpy, entropy, stability and base-pair stacking values / Gustavo Sganzerla Martinez. – 2018.
xiii, 79 leaves : il. ; 30 cm

Dissertation (Masters) - University of Caxias do Sul, Graduate Biotechnology Program, 2018.
Advisor: Scheila de Ávila e Silva.

1. Enzymes. 2. RNA polymerases. 3. Enthalpy. 4. Entropy. I. Silva, Scheila de Ávila e, orient. II. Título.

CDU 2. ed.: 604.4:577.15

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

Acknowledgements

I am so thankful to:

all of my family members who have supported me through all the steps in this dissertation, specially my mother Teresinha, grandparents Darcy and Nair, my aunt and uncle Jane and Marcos, along with their children Lorenzo and Eduarda;

all the staff from Inclass school, my workplace, including my boss Carina and our secretary Marilis, they understood and supported me in all steps which required my absence in workdays;

my friends, who spent time and patience to understand my absences in key moments, specially João Pedro, Felipe, Eduardo, Lucas. The whole crew from our online gaming community who was patient enough to understand that I had other obligations rather than only playing games; the beloved classmates from undergraduate school who were present in several steps and helped me;

the examining board from UCS, honored professors Daniel and Sérgio who elucidated this dissertation with helpful advice;

my advisor, professor Scheila who was a key person helping me and presenting me the right guidance and decision making throughout the whole process;

all the members from PPGBIO in UCS, and the secretary always ready to assist Lucimara

The world is indeed full of peril, and in it there are many dark places; but still there is much that is fair, and though in all lands love is now mingled with grief, it grows perhaps the greater

J.R.R Tolkien

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

UCLUST	UCLUST algorithm
A	Adenine nucleotide
AI	Artificial intelligence
BACPP	Bacterial promoter prediction
BLAST	Basic local alignment search tool
BTSS	BTSS finder promoter tool
C	Cytosine nucleotide
CD-HIT	CD-HIT clustering and comparing tool
CNN Prom	CNN promoter prediction tool
Dace	DACE clustering algorithm
DNA	Deoxyribonucleic acid
FASTA	FASTA file format
G	Guanine nucleotide
H	Enthalpy
IT	Information technology
mRNA	Messenger ribonucleic acid
P	Pressure
pN	PicoNewton
RegulonDB	Regulon database
RNA	Ribonucleic acid
RNAP	Ribonucleic acid polymerase enzyme

SIDD	Stress induced deoxyribonucleic acid destabilization
T	Thymine nucleotide
TSS	Transcription starting site
U	Heat
UBLAST	UBLAST algorithm
V	Volume
Δs	Entropy variation

Abstract

Promoter sequence recognition by RNAp enzyme is a key step in gene transcription. Its location is found in a few base pairs before the coding region. An in-depth study of promoters sequences role might provide an enhanced foundation to understand how genes are expressed under different conditions and produce biological rules to be used in computer techniques such as data clustering. Somehow, a cell behaves similarly as man-made machine, thus, its processes involve the best possible use of energy sources without producing too much heat. By this means, some physical concepts applied to machines, might, as well, be applied in cells, such as entropy and enthalpy variation. The present dissertation looks to assess the role of physical properties of the DNA: entropy, enthalpy, base-pair stacking and stability, in the characterization of *Escherichia coli* (*E. coli)* promoter sequences. To do so, a clustering technique was used to group promoter sequences clusters including the beforementioned features. With the cluster results in hand, a profile of the physical aspects of the DNA in promoter sequences may be drawn and biological inferences made upon these. Currently, not a big number of promoter identification tools make the use of combined profiles of enthalpy, entropy, base-pair stacking and stability. This paper has reported a strong correlation between enthalpy, stability and base pair stacking, where each combination of these features behaves differently in promoter sequences recognized by different sigma factors. We understand, according to the literature, that promoter sequences are known to be different in comparison to other genomic sequences, the results displayed in this paper enable a wider comprehension of difference between promoters themselves. Where, according to the sigma factor that is associated to the RNA polymerase recognition, the physical profile tends to be different, and by this, this paper's results might bring a big acquisition to bioinformatics.

Keywords: Gene transcription; enthalpy; entropy; base-pair stacking; stability;

1 INTRODUCTION

Technological advances in several fields such as biology provide a huge amount of data to the scientific community. With this, emerges the need of computing techniques that are capable of predict, identify and classify this data, seeking to produce biological inferences.

The DNA molecule presents coding regions, which have their expression controlled by regulatory elements. According to Krebs, Goldstein and Kilpatrick (2014) the studying of these elements aids in understanding the gene function in different species and how they respond to environmental changes. In addition, comprehending the gene's functionalities provides the scientific ground to develop new drugs and to understand the biological mechanisms related to diseases, as instance.

One of the regulatory elements is the promoter sequence. This can be found before the coding region and the holoenzyme RNA polymerase DNA dependent (RNAP) recognizes the promoters, triggering the start of the transcription process. The promoter regions have in their nucleotide composition some segments with certain level of conservation, which helps its recognition by the RNAP. However, this biological pattern presents some degeneration, which hinders their computing analysis. Moreover, several works bring other features rather than the similarity in nucleotide composition. Some physical features are conserved in promoter regions and they can be used for their identification. These features: *i*) stability; *ii*) base-pair stacking; *iii)* entropy; and *iv*) enthalpy aid in RNAP recognition.

The stability of base pairs is a characteristic that refers to the amount of free-energy present in a base pair interaction. The base-pair stacking is a value found in the bond between the nucleotide duplexes. The entropy and enthalpy are physical and structural features of the DNA molecule. It is believed that certain molecular tasks performed only in the promoter region can cause changes in the entropy and enthalpy value, thus being able to differentiate promoters and non-promoters based on their entropy/enthalpy values. Several tools rely only in identifying promoters based on the similarities in their nucleotide composition, this computing task can be improved and become more precise with the use of a physical and structural feature of the DNA.

The first step to conclude this paper is to analyze how enthalpy, entropy, base-pair stacking and stability behave in different sites within the promoter sequence. Once these

results are achieved, with this present dissertation, it will be able to discuss in each different group of promoters the impact that the beforementioned physical features have in the transcription process. It is also proposed the use of a clustering analysis to identify and group together promoter sequences based on their physical features. It is important to mention that bright results may be inferred from a deep cluster analysis, where, depending on the outcomes, there can be stated whether or not promoter sequences are being captured by the clustering algorithm based on their physical and structural properties.

2 AIMS AND OBJECTIVES

The aim of this dissertation is to be able to present a profile on how enthalpy, entropy, base pair stacking and stability inside promoter regions. The interpretation of these profiles provides us to understand better what happens within promoter regions and to be able to differentiate bacterial promoters recognized by different RNAP σ factors.

2.1 SPECIFIC AIMS AND OBJECTIVES

The specific goals of this paper are:

- To analyze promoter sequences in terms of their DNA entropy, enthalpy, base-pair stacking and stability in different RNAP sigma groups;
- To clusterize promoter sequences according to their own physical aspects;
- To promote an overview of the physical aspects permeating promoter sequences;
- To use the knowledge gathered by the physical feature assessment to produce biological inferences;
- To enable the physical aspect comprehension of promoter sequences to be used in *in silico* analysis.

3 THEORETICAL REFERENCES

The upcoming section will explain the literature used to conduct this dissertation. Here, a multidisciplinary junction of biology, physics and computer science is presented. These three fields of study show synergy and vary from early biological and physical concepts to the state of art of computer techniques regarding artificial intelligence. This section is organized in a way that: *i*) biology concepts regarding molecular biology comes first, then; *ii*) a profile of physical characteristics concerning promoter sequences; *iii*) a physical review of enthalpy, entropy in thermodynamics point of view; and finally; *iv*) information technology's (IT) optics throughout clustering technique and how IT aids other areas such as biology.

3.1 GENE EXPRESSION AND ITS REGULATION

The deoxyribonucleic acid (DNA) sequence is described by the molecular biology as an information repository, which is necessary to build RNA, and through this – in most cases – a protein that has a function related to cell structure, regulation, or catalysis (DE ROBERTIS, 2003). There are several mechanisms that make sure the correct gene is expressed in the right moment. These mechanisms are defined as regulators of the genetic expression. Thus being, a cell, tissue or organism will be capable of improve, decrease, start or halt the production of RNA, proteins, and the gene's final products according to the metabolic demand. There are mechanisms responsible for controlling the genetic expression, when succeeded, grants to the organism – apart from its complexity (single or multi cell) – the ability of having its following necessities supplied (DE ROBERTIS, 2003; SANDERS and BOWMAN, 2014; BROWNING and BUSBY, 2016).

Figure 3.1 represents the central dogma of molecular biology, proposed by Watson and Crick in 1953 (WATSON and CRICK, 1953), in which the flow of genetic information is tracked. The first step represents the DNA replication, occurring when the organism needs to produce more DNA molecules starting from another DNA template. The step where the DNA carries information to form the ribonucleic acid (RNA) is the transcription, where DNA information is transferred to an intermediary molecule, the RNA in order to produce, in the last step portrayed by the Figure 3.1, protein molecules in the translation process, attending to an evolutionary response and granting survivability

to the organism itself (CASES *et* al. 2003; MCADAMS *et* al. 2004; KREBS, GOLDSTEIN and KILPATRICK, 2014).

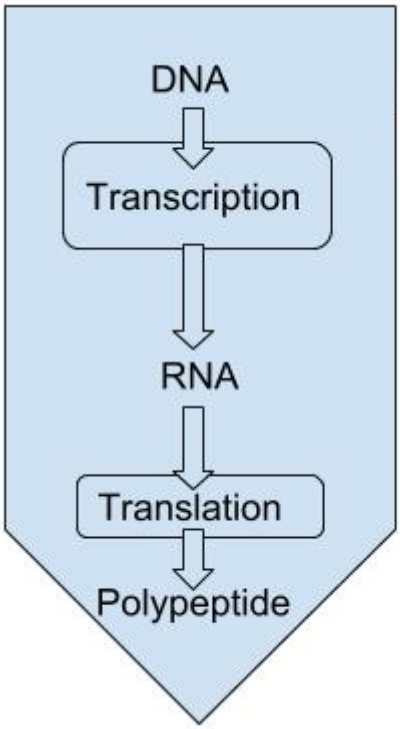


Figure 3.1: Molecular biology central dogma (adapted from KREBS, GOLDSTEIN and KILPATRICK, 2014)

Liu *et* al. (2018) proposed an orthogonal model to analyze the molecular biology central dogma. The model, shown in Figure 3.2, sought a comparison from the dogma itself with a computer software that must run in different platforms. Any specific big change on the software can affect its compatibility capacity in other systems. Thus, a software, in order to function in other platforms and to be compatible with other systems must diminish nuances that represent specificity of a single organism. This is the definition of an orthogonal system presented by the authors, where the components of this system (DNA, RNA, proteins) interact between themselves to achieve a specific goal, without interrupting or be interrupted by native cellular functions. On the other hand, universal rules are something that should be avoided in biologic sciences, tt is commonly said the main biology rule is that there are no rules. Exceptions can be found, basically, in every fundamental principle (KOONIN, 2012; LIU *et* al., 2018).

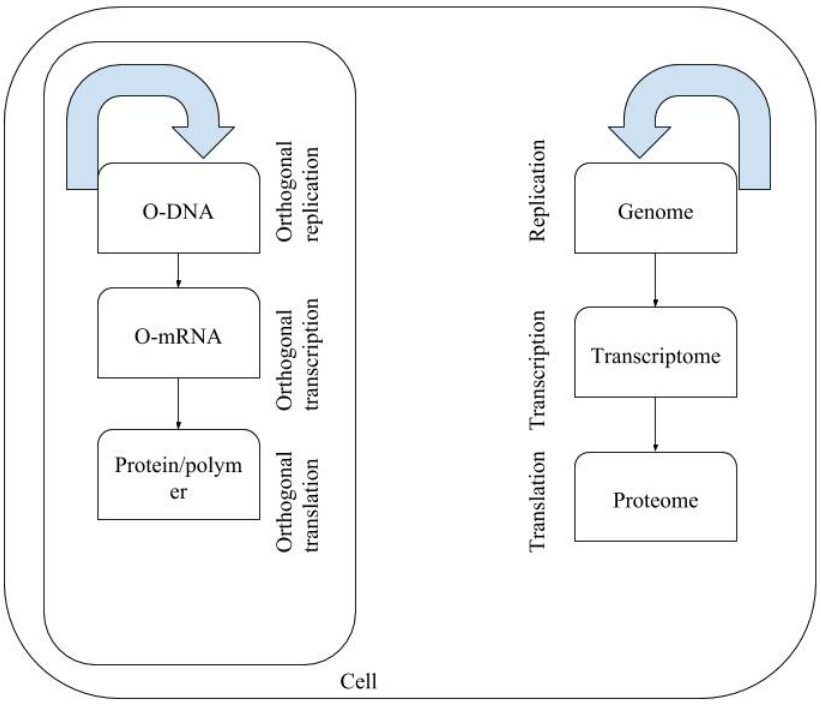


Figure 3.2: Orthogonal model from molecular biology central dogma (Adapted from LIU *et* al., 2018)

The DNA molecules is composed by: *i)* coding regions, which contain information regarding the final genetic products and; *ii)* regulating elements that act as controllers, assuring that the processes start and finish in the right place and time. The study of these regulating elements aids in understanding the genes' functions in different species and enabling for the organism to have its needs suited when facing environmental changes. Additionally, the genes' functionalities provide the scientific ground to develop new drugs and to understand the biological mechanisms related to diseases (KREBS, GOLDSTEIN and KILPATRICK, 2014).

3.1.1 PROMOTER SEQUENCES, TRANSCRIPTION AND THE RNAP ENZYME

The transcription process can be defined as one of the main steps in regulating the genetic expression in any organism. Bacteria are beings that live in soils, colonize plants and infect animal tissues and are likely to extreme environmental changes such as heat, humidity and acidity, which can affect the survivability of the cell if there is no suitable metabolic answer. In this meaning, bacteria rely on regulation systems that seek to optimize the metabolic answer, providing the cell the skill to make right decisions about which nutrient should have its production prioritized and which environmental changes

should be considered (CASES *et al*., 2004; MCADAMS *et al*. 2004; CASES and LORENZO, 2005).

The transcription process will produce RNA, a molecule that this is almost identical in terms of sequence to the DNA coding strand. The DNA coding strand follows the 5'-3' direction and is complementary to the sample strand, which follows a termination 3'-5' and works as a model for RNA synthesis. The synthesis of RNA is catalyzed by the enzyme DNA dependent RNA polymerase (RNAP), this will be detailed in the next section. The transcription process in Figure 3.3 initiates when RNAP identifies a promoter region in the upstream the gene. Starting from this position, the RNAP moves over the gene, performing RNA synthesis until it finds a terminator sequence that liberates the DNA molecule, finishing the transcription process. Previous sequences to the transcription starting site (TSS) are named upstream and the sequences that symbolize the coding region are named downstream. The sequences are generally written in a way that its transcription advances from left (upstream) to right (downstream) this corresponds to the writing in the messenger RNA (mRNA) in a direction 5'-3' (KREBS, GOLDSTEIN and KILPATRICK, 2014).

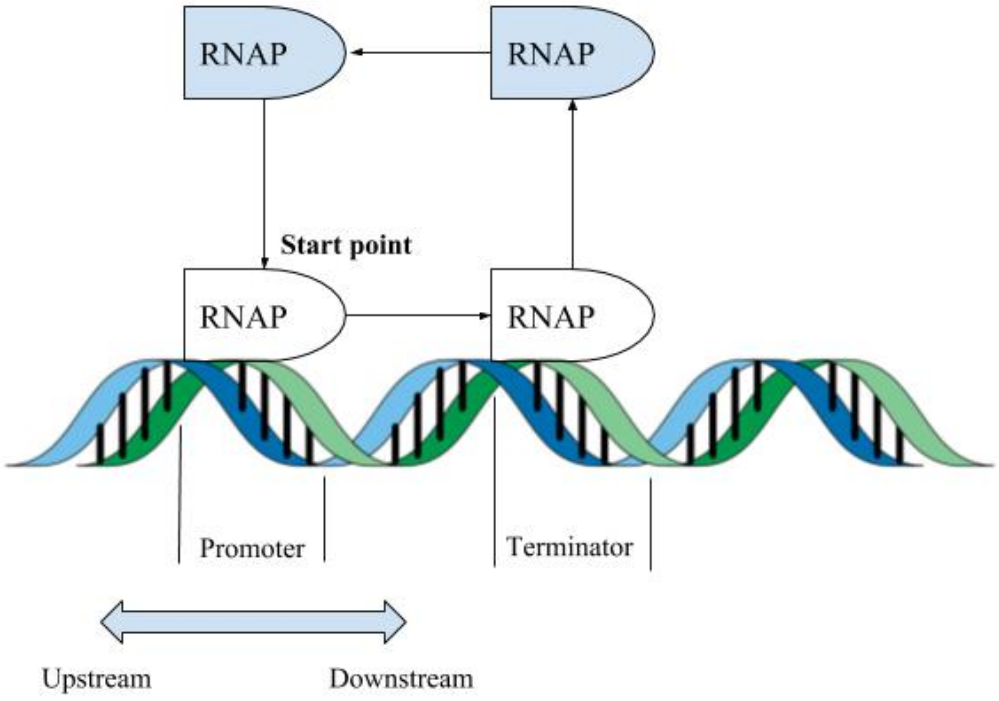


Figure 3.3: Transcription process (adapted from KREBS, GOLDSTEIN and KILPATRICK, 2014)

A key element in the transcription process is the promoter sequence. As shown in Figure 3.3, it tells the RNAP where exactly to start and end the mRNA writing. This element is characterized by being a segment of DNA that precedes the coding region. The first step involved in the transcription process is to the RNAP identify the promoter sequence (KIM *et al*., 2005; ABEEL *et al*., 2009).

Another key factor related to bacterial promoters is the action of the RNAP enzyme. The RNAP is a protein complex with its assigned function: to look specifically for the promoter region, and just after this acknowledgement the transcription process will continue. It is worth observing that RNAP plays a similar role in a variety of living organisms, which makes it an important element in transcription processes of all living organisms. In fact, this enzyme has suffered few alterations in the evolution process. The RNAP counts with six highly conserved units: two α subunits, one β subunit, one β' subunit; one subunit ω and σ. The σ subunit directs the bacterial RNAP to specific sites to connect to the DNA, matching environmental needs of the organism. The Table 3.1 shows each subunit present in the RNAP with their specific function.

Table 3.1: Description of the RNAP's subunits in *E. coli* (LEWIN, 2008)

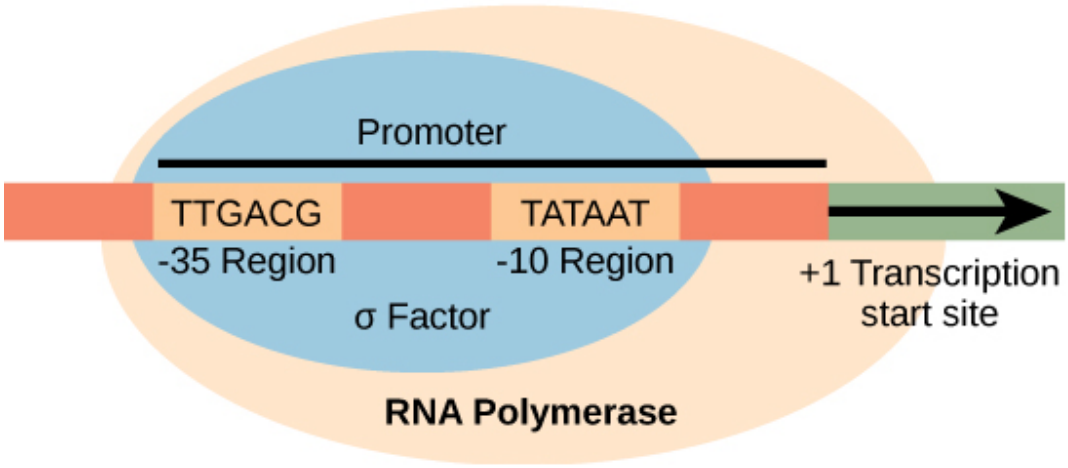
SUBUNIT	RNAP Function
α	Connecting regulatory elements
β	Phosphodiester bonds creation
β'	Connecting to the sample strand
σ	Promoter identification and initiation of transcription
Ω	Addition in the connection force between the units

The σ factor is responsible for mediating the interaction between RNAP and promoter. Each σ factor can start the transcription of different genes and groups of genes, being associated with promoters that regulate the expression of a group of genes during a specific cellular moment. The σ subunit is identified according to a molecular weight value of the cell (σ^{24}, σ^{28}, σ^{32}, σ^{38}, σ^{54} and σ^{70}). A feature involved in the interaction between promoter and RNAP are the consensus sequences. These are groups of nucleotides which face a level of base pair conservation, they are just one of the

identifying quotas of the RNAP. Any change in nucleotides in the consensus sequences can affect the efficiency and the speed of RNAP. As displayed in Figure 3.4, the sequences of nucleotides can be found in several promoters, creating a biological identification pattern. Thus, when recognizing a promoter, the RNAP looks for these regions, enhancing the idea of base conserving and the formation of consensuses (KREBS, GOLDSTEIN and KILPATRICK, 2014).

It is already known that bacterial promoters have a level of nucleotide conservation. This helps the RNAP when looking for promoters, the nucleotides tend to follow a certain biological pattern. In *E. coli* promoters regulated by the σ^{70} it is possible to find two regions presenting a higher nucleotide retention level. As indicated in Figure 3.4, this consensus regions are located in the nucleotides -10 and -35, the sequences are: TATAAT and TTGACA, respectively. The Figure 3.4 shows promoters recognized by σ^{70}, which as the σ factor that is responsible for starting the transcription processes in a handful of genes, this σ factor is identified as a housekeeping one, when the σ^{24}, σ^{28}, σ^{32}, σ^{38} and σ^{54} are known to be alternative σ factors (HAUGUEN, ROSS and GOUSE, 2008; LEWIN, 2008; DE ÁVILA E SILVA *et al.*, 2011, BABU, 2013).

During the RNAP and promoter interaction, two main moments can be highlighted. The transcription starts with the association between RNAP and the promoter sequence, forming a closed complex, in this process, the DNA remains untouched and protected by catalytic sites, which the function is to protect any single kind of alteration in the DNA that can cause unwanted mutations. After this first step, the closed complex is then converted in an open complex in which the DNA is partially untangled, initiating the RNA synthesis. Lastly, the σ subunit present in the RNAP detaches from the DNA and the process is ended when a terminator region is found (LEHNINGER, 2000; LEWIN, 2008).



σ factor	Moment of cellular activation	Consensus 35 spacer 10
24	Heat shock response	GGAACTT 15 bp GTCTAA
28	Flagellar genes	CTAAA 15bp GCCATAA
32	Heat shock response	CCCTTGAA 13-15bp CCCGATNT
38	Starvation response	TTGACA 16-18bp TATACT
54	Nitrogen metabolism	CTGGNA 6bp TTGCA
70	Housekeeping	TTGACA 16-18bp TATAAT

Figure 3.4: Different functions of the σ factor in E. coli. σ70 promoters (Adapted from PAGET and HELMANN, 2003; KREBS, GOLDSTEIN and KILPATRICK, 2014)

As depicted in Figure 3.5, the transcription process has it first step: the initiation, when the promoter is recognized by the RNAP and the RNA synthesis begins. Then, there is the elongation stage, in which the so-called DNA bubble is created and moves along the DNA strand, synthesizing RNA. At the end, there is the termination, where a terminator region is found, the RNAP detaches from the DNA, the transcript RNA is released and the bubble closed.

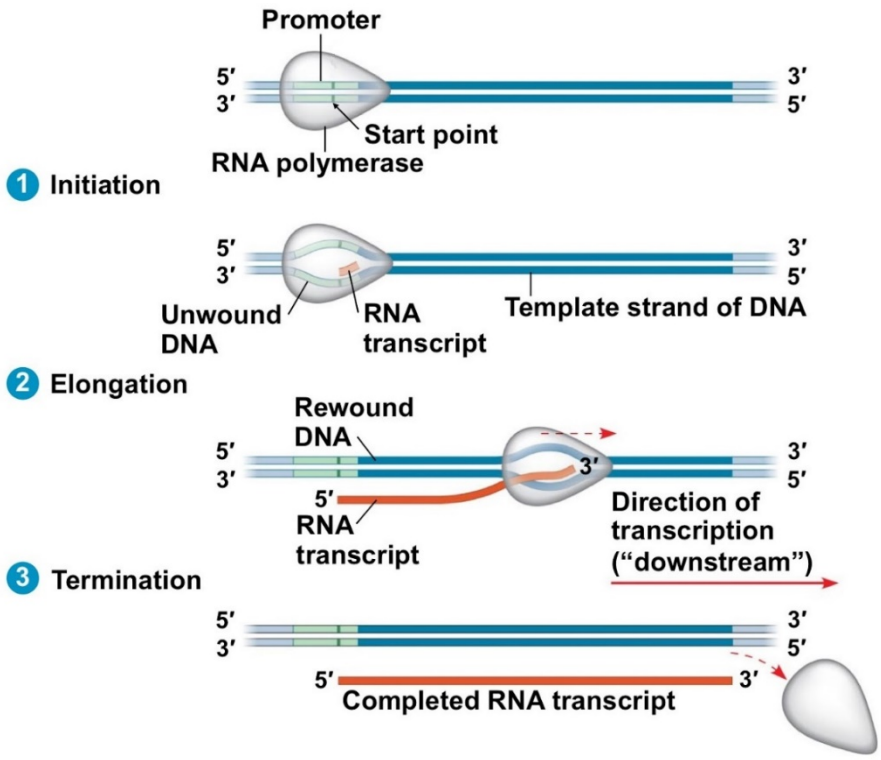


Figure 3.5: Bacterial Transcription process (Adapted from REECE *et* al., 2014).

Thus being, it is possible to contrast three main stages in bacterial transcription: *i)* initiation, where RNAP connects to the promoter; *ii)* elongation to form RNA, and; *iii)* termination. To start this process, the RNAP must associate the promoter with a σ factor, linking the RNAP with the promoter and the RNAP becoming a holoenzyme. After, approximately fourteen DNA nucleotides are merged in the upstream region towards the TSS forming an open complex. When about fourteen RNA nucleotides are synthesized the σ factor is released and an elongation complex is formed and begin to synthesize a RNA molecule a time, lastly when a terminator element is found, the RNAP is dissociated, allowing the starting of another transcription round (YARNELL and ROBERTS, 1999; BURGUESS and ANTHONY, 2001; SKORDALAKES and BERGER, 2003; MURAKAMI and DARST, 2003; KAPANIDIS *et al*., 2006; COOK and DEHASET, 2007; MA *et al., 2016*).

So far, the literature has presented that promoter sequences can be fairly distinguished than other genomic sequences. One of the parameters used by RNAP to identify promoters: the consensus regions, are not sites that presents absolute conservation in terms of sequence. The next section will explore how physical features

can be assessed to classify and identify promoters, enhancing the classification performed by *in silico* tools.

3.2 PHYSICAL FEATURES IN PROMOTER SEQUENCES

Now that it is well known what a promoter sequence is, as it was said in the previous section, there are a lot of critical factors that can tell promoters apart from other coding regions, these factors aid the RNAP targeting. A set of relevant, physical features of the promoter region will be explored in details through this section, providing a theoretical foundation for this paper. Computationally, the identification and classification of promoters can be a harsh task, one of the reasons for this is due to the fact that for some bacterial promoters there is an overlapping of the promoter into the coding region (RANGANNAN and BANSAL, 2007). This demands a detailed analysis up and downstream of the TSS, in order to overcome this challenge, a computer analysis is a way to look through the DNA molecule and extract more than just a simple physical feature and/or a determined sequence of nucleotides (RYASIK *et* al., 2018).

There are physical features inside the promoter region that can distinguish when compared to non-promoter regions. The features that will be explored are: stability, curvature, bendability, entropy, enthalpy and nucleotide composition. The level that these features appear inside promoter regions turn them unique when compared to other regions. It is worth mentioning that through literature review, sometimes the sheer sequence analysis does not show any conservation in the promoter region, but some functionalities remain conserved. It is believed that the presence of these traces make available the *in silico* promoter identification, in a way that there is a biological function – the RNAP identification, linked to the existence of these features (KANHERE and BANSAL, 2005).

3.2.1 BASE PAIR CONSERVATION

Taking the TSS as a reference point, where after it, the transcription process will produce mRNA, there can be found some patterns regarding the presence of certain base pairs as an addition to the energetic viability, approached in section 3.2.1 and the consensus, explored in section 3.1.1. KOZOBAY-AVRAHAM *et* al, (2008) have indicated that in intergenic regions – where promoters can be found, AT nucleotides are more common to happen. The authors have also concluded that the beginning of a sequenced gene is very rich in its GC content, this amount of GC will start to decrease at

the very end section of the gene, where the coding region gives places to intergenic sequences such as promoters and terminators.

The identification of certain patterns in relation to base pair presence in specific DNA segments does not stop here. As shown in Figure 3.6, is possible to perceive a profile where the nitrogenous bases found directly after the TSS. The GC content found in *E. coli* Examples is 42.32% upstream the TSS and 50.79% downstream the TSS. This difference in GC levels happened due to the promoter region being constantly opened by the RNAP. In this way, to open a DNA strand in the promoter region there is an energetic cost, owing to the number of hydrogen bonds beforehand mentioned, it makes sense – physical and biologically, that the region to be opened have base pairs that are easier to open.

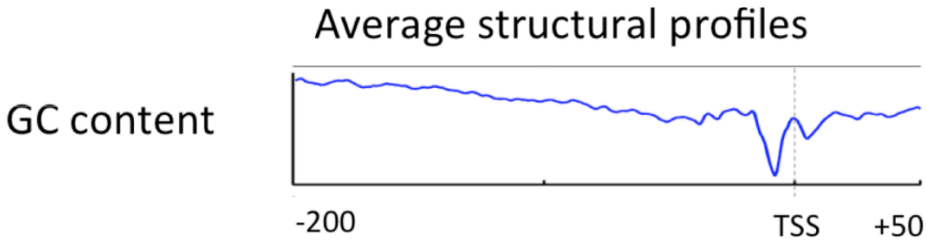


Figure 3.6: Presence of GC content in *E. coli* promoter sequences (Adapted from MEYSMAN et al., 2014)

Other studies have focused on the presence of certain bases along the promoter region. In one of these, Barrios, Valderrama and Morett (1999) have shown that for bacterial promoters recognized by RNAP σ^{54}, the TSS starts the transcription in the positions -12 and -24. Differently from sequences acknowledged by σ^{70}, with the consensus happening in positions -10 and -35 upstream the TSS. The authors proposed and experiment that sought to analyze 84 promoters σ54 dependent and came to the conclusions that a significant number of promoters had its TSS beginning in a purine, with mRNA transcription starting precisely 12 nucleotides upstream the retained purine. The Figure 3.7 shows this data distributed, construing the first nucleotide upstream the TSS (BARRIOS, VALDERRAMA and MORETT, 1999).

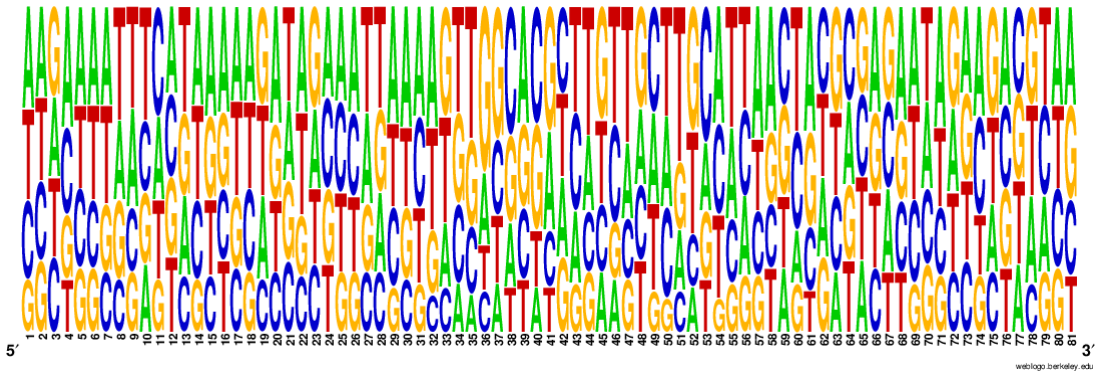


Figure 3.7: Nucleotides after -12 consensus in σ54 promoters (Adapted from BARRIOS, VALDERRAMA and MORETT, 1999).

Another study was conducted by Burr *et* al., (2000) where *E. coli* Promoters containing σ70 transcribed genes. It was found evidence that there is promoter activity found in the -14/-15 upstream positions. It is worth it mentioning that, as stated by the literature, the two motifs that help RNAP recognition in this σ are -10 and -35 upstream the TSS. The authors found a hierarchy in promoter activity that tends to be lower when TG base pairs are present. The reports have shown the TG presence aids in the open-closed complex transition. When the test is directed to the next 2 nucleotides, in positions -16/-17, again TG dinucleotides were found. The Figure 3.8 shows the dinucleotides through a histogram for the two mentioned positions: -14/-15 and -16/-17. The authors concluded in this study, using 300 *E. coli* Promoters, the right after the consensus motif, there is another identifier, enhancing the RNAP recognition (BURR *et* al., 2000).

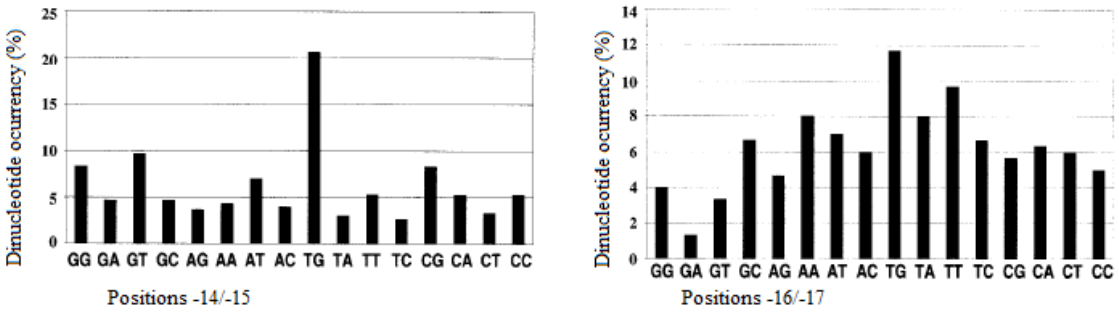


Figure 3.8: Dinucleotide presence in positions -14/-15 and -16/-17 the TSS (BURR et al., 2000)

One more work making analysis in the dinucleotide presence in *E. coli*. was done by Ezer, Zabet and Adrvan (2014), one of the goals was to check the evolutionary aspect of *E. coli*. promoters. The authors have identified a high conservation in locations where the RNAP binds to the DNA – called binding sites, through the analysis of base pair insertion/deletion rate in different regions located in the promoter sequence. The evolutionary traits were checked in the following groups: *i*) between the TSS and the first binding site; *ii*) inside the binding sites; *iii*) inside the binding sites with small base pairs spacing; *iv*) inside the binding sites, further than 100 bp starting and between the last binding site and the terminator sequence (EZER, ZABER and ADRYAN, 2014).

The results of this study are disposed in the Figure 3.9, in which is possible to perceive the evolutionary conservation in spots where the RNAP binds to the DNA. This conservation regarding the evolutionary aspect is due to the fundamental role RNA has in all living organisms. This is not different with the enzyme responsible for RNA synthetization, this suggests the frameworks of RNAP are not so different when comparing Archaea, Bacteria and Eukarya (WERNER and GROHMANN, 2011; EZER, ZABER and ADRYAN, 2014).

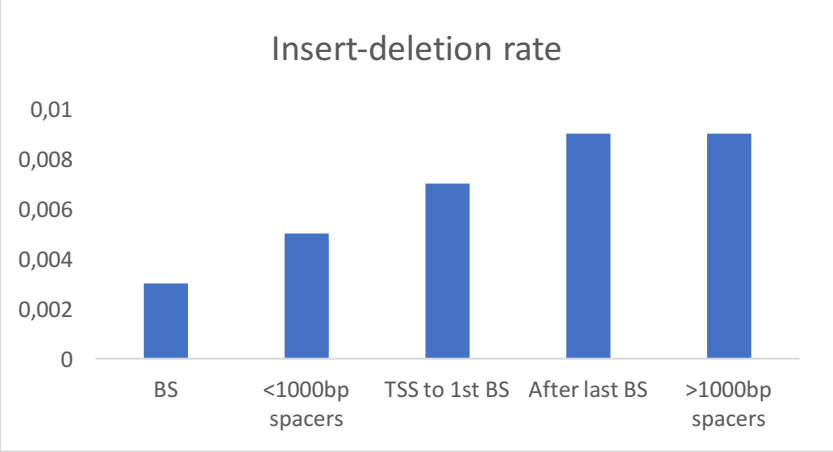


Figure 3.9: Analysis of the insertion/deletion rate of dinucleotides grouped in different segments through the promoter sequence (Adapted from EZER, ZABER and ADRYAN, 2014)

3.2.2 STABILITY

The stability values presented by the DNA molecule directly rely on the nucleotide sequence. In bonds between purines (adenine and guanine, A and G, respectively) chemical bonds of two hydrogen bounds are found, when there is union between

pyrimidines (thymine and cytosine, T and C, respectively) the number of hydrogen bonds is three, as shown on Figure 3.10. The transcription process involves the opening of a DNA strand, turning it into an open complex, hence, it is necessary to break the DNA strand, in other words, to spend power to get the DNA opened and separate the base pairs to write them to mRNA. For this process to be energetically viable, it is reasonable that the DNA segment with a lower stability between its base pairs, with a weaker chemical bond, is the one to be broken (KANHERE and BANSAL, 2005; RAMPRAKASH and SCHWARZ, 2007; DE ÁVILA E SILVA and ECHEVERRIGARAY, 2011).

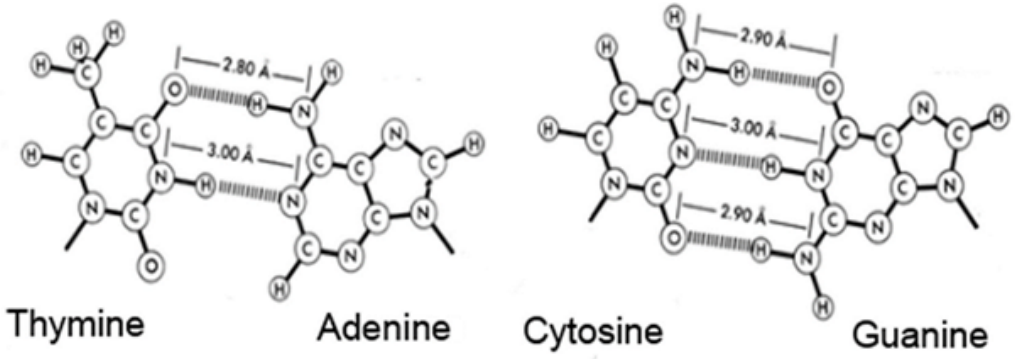


Figure 3.10: Hydrogen bonds on nucleic base pairs (WATSON and CRICK, 1953)

Figure 3.11 shows an energetic profile of a DNA slice containing the upstream and downstream region. The segment corresponds to 611 *E. coli* Promoters, the TSS is located in the position 0. When analyzing the image, it is possible to identify three stability peaks, matching the consensus regions -10, -35 and -50, spots that RNAP identifies promoters (RANGANNAN and BANSAL, 2007).

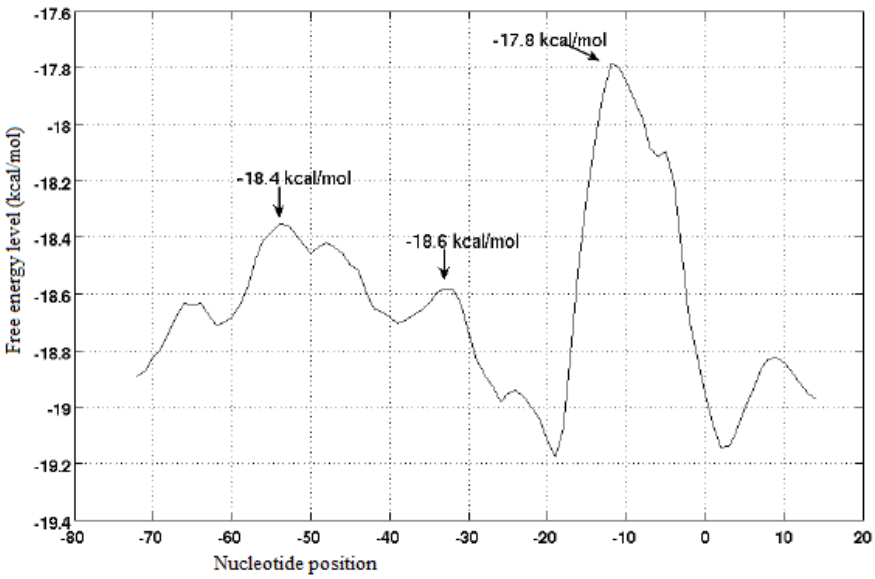


Figure 3.11: Stability profile in *E. coli*'s promoter region (RANGANNAN and BANSAL, 2007)

The following Table 3.2 shows the stability values for each base pair present in DNA sequences. This being, it is possible to highlight the genetic stability when the analysis object is the promoter region, whereas a whole genome can be checked through stability values and promoters be found (SANTALUCIA and HICKS, 2004).

Table 3.2: Stability values for DNA base pairs (SANTALUCIA and HICKS, 2004)

Nucleotide Duplex	Stability Value (kcal/mol-bp^{-1})	Base pair	Stability Value (kcal/mol-bp^{-1})
AA	-1	TG	-1.44
AT	-0.88	GT	-1.44
TA	-0.58	TC	-1.28
AG	-1.3	CT	-1.28
GA	-1.3	CC	-1.84
TT	-1	CG	-2.17
AC	-1.45	GC	-2.24
CA	-1.45	GG	-1.84

3.2.3 STRESS INDUCED DNA DUPLEX DESTABILIZATION (SIDD)

The initiation and transcription process during the regulation both involve the untwisting of the DNA duplex, this process of separation must be controlled. In simple terms, according to Benham, Wang and Noordewier (2006), this feature includes a relaxation in the bonds between the base pairs, which in case of constant separation, decreases the use of energy needed to be constantly opening the DNA strand. This process occurs through superhelical stresses imposed on the duplex. This feature does not depend on the primary structure of the DNA strand nor the stability values. 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 superhelical 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 superhelical value. The variations in the superhelical level in a promoter can show several effects in final product coded by the gene, one of them is the SIDD variation (WANG, BENHAM and NOORDEWIER, 2004; WANG and BENHAM, 2006; DE AVILA E SILVA and ECVHERERRIGARAY, 2012).

The Figure 3.12 shows the destabilization level G(x) needed to the DNA strand with its base pairs remain open. Spots where the destabilization level is high have low G(x) values. As shown in the Figure 3.8, there are four sets of sequences: *i*) the promoter sequences that were identified immediately present in the upstream direction to the TSS; *ii*) coding sequences starting from the in the located TSS and extend up to 1001 base pairs towards the mRNA transcription; *ii*) intergenic regions, but no promoters; *iv*) and a random set of sequences. The authors (WANG, BENHAM and NOORDEWIER) have found that conserved SIDD sites show a higher tendency to avoid coding regions, where in intergenic regions, well documented promoter sequences indicate a higher SIDD value.

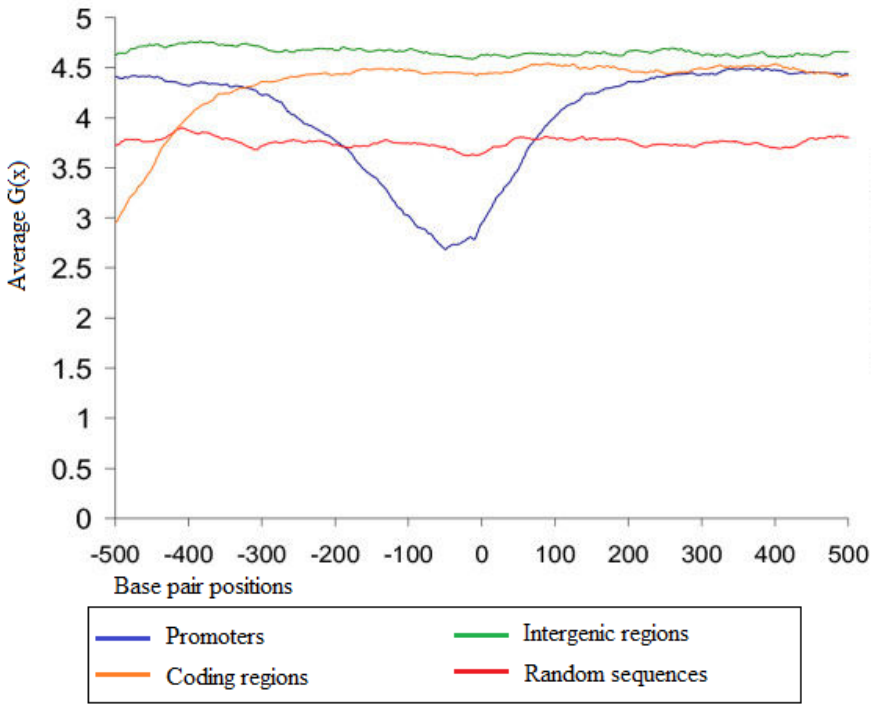


Figure 3.12: SSID profiles (Adapted from WANG and BENHAM, 2006)

3.2.4 DNA CURVATURE AND BENDABILITY

The term DNA curvature refers to the ability to the DNA twist itself without the aid of external forces. There are variations found in the DNA's linear trajectory that grant the DNA strand its curved shape. When analyzing curvature levels, this can be a feature to distinguish positions in the whole genome. Studies from Perez-Martin, Rojo and Lorenzo (1994) defined the DNA curved view as a feature that helps that transcription initiation, from the moment when RNAP binds to the DNA. In this way, is possible to find a difference. When comparing the curvature values between promoter regions and coding sequences (OLIVARES-ZAVALETA, JÁUREGUI and MERINO, 2006; KOZHOBAY-AVRAHAM *et* al., 2008).

The Figure 3.13 shows how the curvature causes influence in the DNA percentage. The gray region represents the curvature values in regulatory regions and the activation momentum of specific genes in *E. coli*, when compared to other DNA regions, the regulatory elements have a distinction, supporting the idea that RNAP can use the curved aspect to initiate transcription.

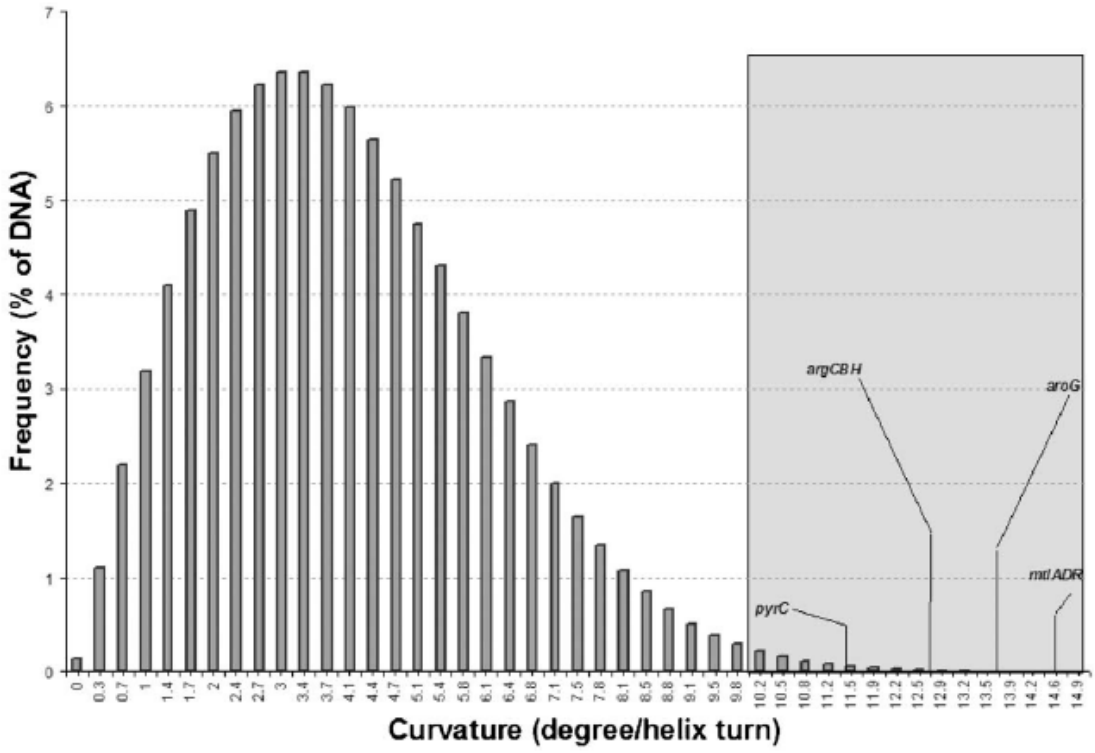


Figure 3.13: DNA curvature profiles from E. coli, comparing the DNA segments with the moment some specific genes (gray region) (OLIVARES-ZAVALETA, JÁUREGUI e MERINO, 2006).

The bendability of the DNA strand is another physical feature of the DNA molecule. This is differently presented in promoter regions due to the twist performed by the DNA when binding to RNAP. As shown in Figure 3.14, the hardness level contained in promoter regions have different levels when compared to other genomic sequences (MEYSMAN *et* al., 2014).

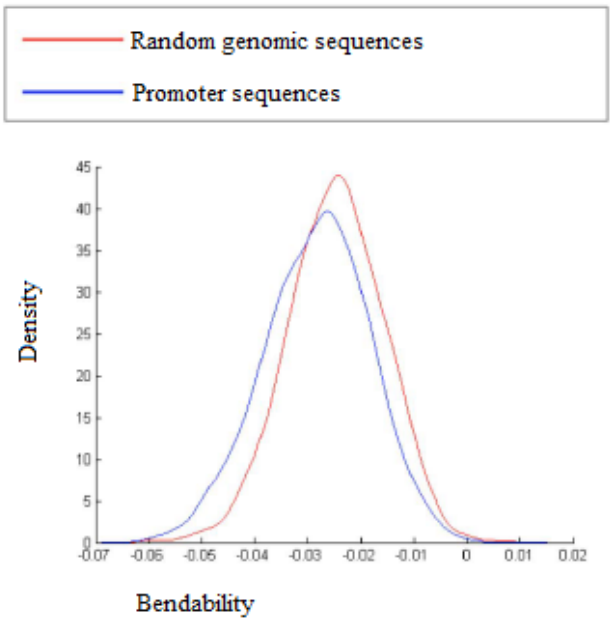


Figure 3.14 : Bendability profile of the region promoter compared to other genomic regions (Adapted from MEYSMAN et al., 2014)

3.2.5 BASE PAIR STACKING

Stacking interaction between two adjacent base pairs is an essential force component responsible for DNA stabilization and gene regulation (ZHANG *et* al., 2015). SATTIN *et* al., (2004) have indicated how this feature works, and due to the twisted structure of DNA strand, the relative force needed to unstack GC base pairs would cost more to unstack a base pair consisting of AT. This is due to the number of hydrogen bonds found in each one of the bindings (KANHERE and BANSAL, 2005). Previously, ZHANG *et* al., (2015) have yet another indicated the amount of binding strength between GC and AT base pairs, showing that the first base pair would cost 20.0 piconewton (pN) and the second set would cost 14.0 pN. This indicates a lower value in the strength binding between the base pairs most found in promoter sequences – which is AT (MEYSMAN *et* al., 2014).

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 value is determined by the interactions between the nucleic acid base pairs. The DNA is known for its helical structure, this structure is stabilized by the hydrogen bonds that

contain in between the beforementioned interactions (ZHANG *et* al., 2015; HASE and ZACHARIAS, 2016). The Figure 3.15 displays the mean value of the base stacking energy in *E. coli*. sequences. The Table 3.3 shows the stacking value for each pair of nucleoside (ORNSTEIN *et* al., 1978) and the AT base pair stacking value will count as -3.82, being the highest in Table 3.4, while GC connection present the lowest value of -14.59, encountering the data presented by ZHANG *et* al., 2015.

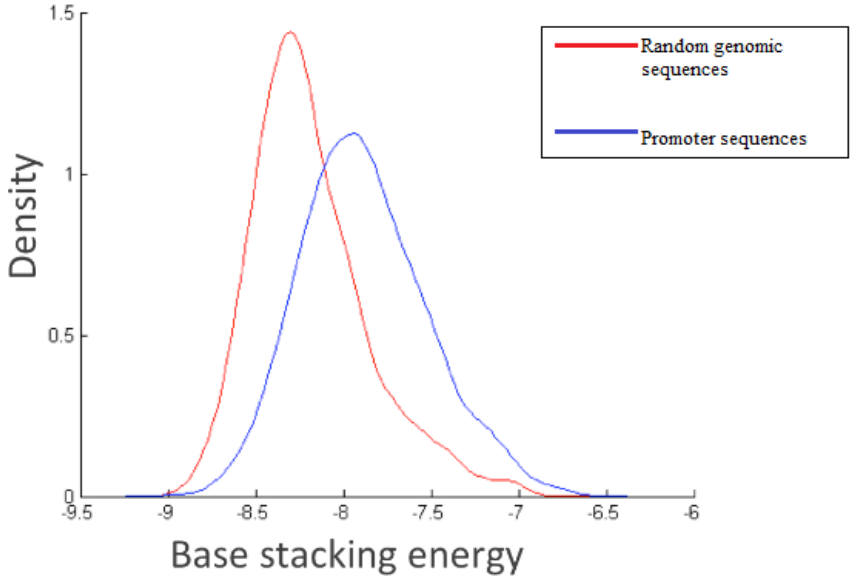


Figure 3.15: Base stacking energy profile (MEYSMAN et al., 2014)

Table 3.3: DNA nucleoside stacking values (ORNSTEIN et al., 1978)

Nucleotide Duplex	Stacking Value (kcal/mol-bp^{-1})	Base pair	Stacking Value (kcal/mol-bp^{-1})
AA	-5.37	TG	-6.57
AT	-6.57	GT	-10.51
TA	-3.82	TC	-9.81
AG	-6.78	CT	-6.78
GA	-9.81	CC	-8.26
TT	-5.37	CG	-9.69
AC	-10.51	GC	-14.59

CA	-6.57	GG	-8.26

3.2.6 ENTROPY, IRREVERSIBLE PROCESS AND ENTROPY VARIATION WITHIN THE DNA MOLECULE

According to the second law of thermodynamics, whenever energy is submitted to any sort of process resulting in its transformation, part of it becomes unused. This way, all natural-occurring phenomenon is classified as being an irreversible process, happening in a one hand way, as shown in Figure 3.16 where the ice cube will eventually melt and the ice will never lose heat for the water, this makes it impossible for the water to become hotter and the ice become colder (YOUNG and FREEDMAN, 2016).

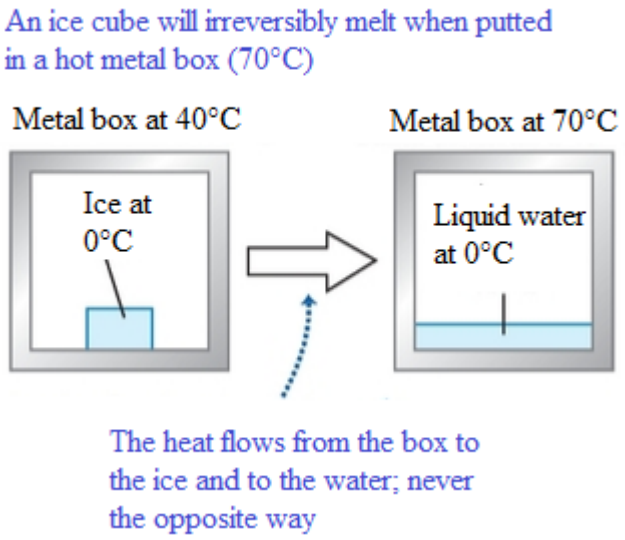


Figure 3.6: Irreversible process (adapted from YOUNG and FREEDMAN, 2016)

Now, when assaying nature laws, every event can only happen in a single sequence of events, this is determined through what physicians determine as: the arrow of time. What turns systems such as the one exemplified in Image 3.16 a single direction thermodynamic system where, according to the thermodynamics, the nature grants the ice will not give in heat to water (GASPAR, 2000; YOUNG and FREEDMAN, 2016).

The role played by the entropy in this scenario will be further explored. The concept presented here needs to be considered, where in every physical system, there is a single direction for the exchange of heat between the components inside these systems,

this same description of environment is found in a cell (DAVIES, RIEPER and TUSZYNSKI, 2013).

The main principle of the second law of thermodynamics says that no thermal system is able to fully convert its energy in work. Every thermal process product an amount of unused and dissipated energy. Adding irreversible process into this, and the energetic variation can be measured through what physicians define as entropy variation (Δs). The entropy is then, treated as a thermodynamic magnitude that measures the level of irreversibility in a system, and summoning thermodynamics' second law, where heat will never fully become work, the entropy of a system will always increase. In other words, the entropy can be used to describe the parcel of unused energy (HALLIDAY *et* al., 2006).

Any natural phenomenon leans to have its entropy raised. Inside the innumerous settings a system can have, the least organized is always the most probable and natural to occur. Under this point of view, it raises a notion that is a bad concept among physicians: the one that says entropy is responsible for measuring the disorder level in a universe. This idea needs to be more explained, and with the presented single direction that processes have, they always tend to move from a more organized to a less organized state. Thus, it is a simple-minded explanation to classify the entropy only as a disorder measurement without considering the energetic background explored (HALLIDAY, *et* al., 2006, DAVIES *et* al., 2013).

The Figure 3.18 depicts the raising disorder level in a system, when advancing in this system setting, a less organized state is always more possible, this is characterized by entropy raising. This same image shows the possible distributions a gas can achieve in its environments. The gas is not limited into a particular place in the environment, spreading itself equally through the whole system, this way, the entropy in situation D is higher than entropy in situations A, B and C; the state shown in situation D is always more probable as time moves. (GASPAR, 2000; HALLIDAY, *et* al., 2006; DE LIMA, 2007; YOUNG and FREEDMAN, 2016).

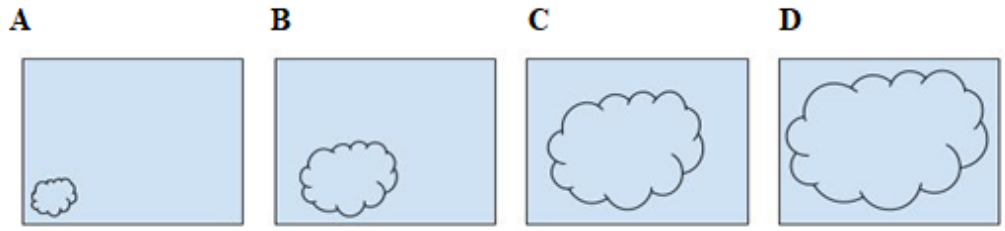


Figure 3.17: Portrayal of the possible distribution of a gas in a closed environment (Adapted from MAIA and BIANCHI, 2007)

Living beings or, living systems perform a high number of complex processes, all synchronized seeking to keep and maintain the system's biological functions. The information repository needed to store the steps required to perform the beforementioned processes is the DNA. The DNA has its own alphabet, consisting in four letters, or nitrogenous bases and the setting of these will indicate the actions to keep the organism living. Working similarly as other systems, the DNA required entry information (e.g. a chemical gradient), which will be processed and generate outputs in the system (e.g. a protein, DNA). To keep functioning, living organisms need a stable energy supply, which will be converted in useful work and keep the system in a constant physiological temperature. Therefore, the energetic production of a system is essential for its survival (DAVIES *et* al., 2013; MULLIGAN *et* al., 2015).

A cell works exchanging material and heat with an external surrounding, this defines the cell as an open system. In thermodynamic issues, a cell is similar as a machine, this was explored in a study from Davies *et* al., (2013). A cell needs to obey the laws of physics, and can have Δs measured in four distinct moments: *i*) chemical bonds leading to cell aggregation; *ii*) mass transport in and out the cell; *iii*) heat generation due to cellular metabolism; and *iv*) information stored in genetic code. Thus, the entropy level of DNA sequences will always be the highest possible (HERZEL *et* al., 1994; DAVIES *et* al., 2013). The following Table 3.4 shows the entropy values for Watson-Crick base pairs in termination 5'-3' (WATSON and CRICK, 1953). The authors Santalucia and Hicks (2004) presented the thermodynamic values for base pairs. The parameters used to get to these numbers were derived from linear regressions with 108 sequences taken as examples.

Table 3.4: Entropy values for DNA base pairs (Adapted from SANTALUCIA and HICKS, 2004)

Nucleotide duplex	Entropy Value (kcal/mol-bp^{-1})	Base pair	Entropy Value (kcal/mol-bp^{-1})
AA	-21.3	TG	-22.4
AT	-20.4	GT	-22.4
TA	-21.3	TC	-22.4
AG	-21	CT	-22.2
GA	-21	CG	-27.2
TT	-21.3	GC	-24.4
AC	-22.7	CC	-19.9
CA	-22.7	GG	-19

Then, it is evident that the entropy is a physic magnitude present in living organisms, which makes it possible to calculate the entropy values present in the DNA of any living organism.

3.2.7 ENTHALPY

Enthalpy is defined as the amount of heat that is present in a system containing few to none pressure variation. Mathematically, enthalpy can be defined as: $H = U + Pv$, where: P is the pressure of a system and v is the volume. As U, P and v are state functions, the result of this equation H, the enthalpy is also a state variable. In a way that the variation seen in the enthalpy, moving from a start to an endpoint will take the whole system into another state. Thus, in a system where volume and pressure are constant, in can be asserted that the enthalpy corresponds to the amount of heat added or removed from the system. Adding or removing heat relies on the comparison of the products and reagents in a system, where if a system between its timeline is characterized as: *i*) an endothermal process where there is heat absorption, and the enthalpy variation is higher in the products than the reagents; and *ii*) exothermal, where there is heat dispersion, turning the enthalpy variation negative (ATKINS and DE PAULA, 2012).

Studies have found that the enthalpy level in promoter sequences tend to be different than other genomic sequences. According to the literature, the GC content, which is lower in promoters, needs an enthalpy value of 10.75 ± 1.43 kcal/mol-bp for its stabilization, while AT nucleotides need 7.4 ± 0.7 kcal/mol-bp to stabilize. At first sight, it would make sense for promoter sequences to have a different enthalpy value due to its low GC content. However, other studies have shown that the forces involved in the bond between protein (RNAP) and DNA strand are not simple. They need to be weak enough to allow the protein to easily scan the DNA and, simultaneously, must be strong enough for longer-living connections (PRIVALOV and CRANE-ROBINSON, 2018).

In the same way as it was previously done with the entropy values, in study from Santalucia and Hicks (2004) the values for enthalpy in DNA base pairs were calculated. The values are shown in Table 3.5.

Table 3.5: Enthalpy values of DNA base pairs (SANTALUCIA and HICKS, 2004)

Nucleotide duplex	Enthalpy Value (kcal/mol-bp^{-1})	Nucleotide duplex	Enthalpy Value (kcal/mol-bp^{-1})
AA	-7.6	TG	-8.4
AT	-7.2	GT	-8.4
TA	-7.2	TC	-7.8
AG	-8.2	CT	-7.8
GA	-8.2	CC	-8
TT	-7.6	CG	-10.6
AC	-8.5	GC	-10.6
CA	-8.5	GG	-8

3.2.8 ENTROPIC AND ENTHALPIC CONTRIBUITIONS TO DNA STABILIZATION

So far, this dissertation has already explained in section 3.2.4 the difference in AT and GC contents regarding regulatory sequences, the AT prevalence also has a link with thermodynamics measurements that are found in the DNA double strand. The enthalpic contribution on AT base pairs is somehow larger than GC base pairs. Previous studies

(MARMUR and DOTY, 1962) believed that the thermostability present in the DNA double helix would be increasing when the GC content is higher due to its extra hydrogen bond as shown in section's 3.2.1 Figure 3.7. According to the literature, early studies have shown that due to the GC content, promoters would have a different enthalpy and entropy value in comparison to other genomic sequences. These studies have shown that the entropy value for an AT base pair stabilization would be 7.4 ± 0.7 kJ/mol-bp and the average enthalpic contribution for a GC base pair stabilization is 10.75 ± 1.43 kJ/mol-bp. Regarding the entropy needed for base pair stabilization, AT is 20.55 ± 2.39 kJ/mol-bp and GC is 27.24 ± 3.58 kJ/mol-bp. Promoters are known to be poor in their GC content. The GC affects the DNA stabilization both in the enthalpic and entropic contribution. The high AT levels presented by promoters comes from water that links to AT, this more complex system increases the entropy (PRIVALOV and CRANE-ROBINSON, 2018).

Entropy and enthalpy are both thermodynamic features of the DNA and these two are closely related. Nevertheless, the way that these two measurements behave is quite different. Enthalpy refers to the system as a whole, as stated in section 3.3.2, in addition, this system being composed of water-enzyme-DNA has high entropy on its interaction. Water is a major ingredient that can be found permeating the hydrogen bonds and affecting final entropy measurements, in other words, this means the system is disorganized due to the amount of its components. While RNAP moves around the strand and water connects to the DNA we have a final product of a more organized system. On the other hand, enthalpy does not refer to all the system's components involved in the beforementioned system, enthalpy only gets affected by the hydrogen bonds found in the connecting nucleotides. This indicates that entropy is a system measurement, while enthalpy is used to check particular components (ZU, ZHI and LENG, 2012; MORGUNOVA *et* al., 2018; PRIVALOV and CRANE-ROBINSON, 2018)

3.3 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Some recent advances in technology have provided the science an exponential growth in the availability of genomic data. With the data being ready, it comes up the need for ways to work with this large amount of data. One form of science that can be handy in this moment is the computer science, which is able to provide reliable tools in terms of identifying, registering and charting all the sequenced genomic data. With more and more trustworthy tools, the bioinformatics plays a key role in unscrambling and deciphering genomic, transcriptomic and proteomic datasets. Then, the bioinformatics

can be classified as a needful field of study integrated to biological sciences (RANGANNAN and BANSAL, 2007).

Since the end of 19th century with the industrial revolution, science has developed and machines have been used to replace manual work. Furthermore, in 20th century, with the arise of the computer and information technology machines that were previously used to simply replace the physical effort performed by humans have started to do the work of a human brain. This led to the emergence of artificial intelligence (AI). In the 21st century, the advent of cloud computing, mobile computing has changed the lifestyle of people, which indicated the coming of a new era to computer science. AI is defined as a variety of human behaviors such as: perception, memory, emotion, judging, rationalizing, acknowledgment, comprehension, design, thinking and creation in a way that all activities can be artificially done by a machine, system or network (LI and DU, 2017).

AI, thus, can be defined as machines designed to perform automated activities that require intelligence, such as decision making, learning and problem solving. More AI applications try to simulate human intelligence by performing its neural associations in an algorithm. The spectrum of AI's acting is wide, one of its functions is to identify patters among big datasets. A common problem found in biological sciences (RUSSEL and NORVIG, 2003; CHRISTIAN, 2013).

Inside the AI field, the agents perform an important set of roles. One of the AI's purposes is to design an agent software, this implements functions and maps the feelings around its environment in actions. The agent is executed in a computer device called platform. There are agents with different intentions, they can vary from agents based in models, agents based in goals, agents based in utility to agents designed to learn. Any computer agent that works with learning allows the agent to operate in unknown surroundings, and over time, become proficient in how to deal with its environment (RUSSELL and NORVIG, 2003).

The learning, as shown in this section can be used as an AI technique to formulate patterns in a large amount of fuzzy data. A task that would take a tremendous amount of time when executed by a human brain compared to a computer. The next section will focus on a specific AI technique used in this paper: clustering.

3.3.1 CLUSTERING AND THE *K* MEANS ALGORITHM

Clustering is an AI technique that can be employed to solve several problems with the most diverse nature, such as image recognition, biological application, document grouping in similar topics, climate data junctions and geographic analysis. The boom in the availability of information that has been recently happening gives the scientific community whopping amount of data. In this sense, when divided in its respective groups, can provide to the scientist some new interpretations and points of view, until then, hidden.

The goal of data clustering is well defined: find patterns that facilitate the natural grouping of a dataset. According to Webster dictionary, the cluster analysis uses as technique to classify data seeking that individuals of a population belong to different groups. Nevertheless, clustering is an AI technique that is capable of recognize features in the data that at first sight are not perceived. Clustering techniques deal with different ways to group data in a space on *n* dimensions, so that every element in a group have some link between its other group conjuncts. The use of clustering techniques is named discovery tools, which grants the user a bigger understanding regarding the data structure and the set of data. As demonstrated in Figure 3.18 (a) the entry data does not show any similarity. The next step, Figure 3.18 (b) displays the data grouped in distinct groups, sharing common resemblances (DUBES and JAIN, 1976; MERRIAN-WEBSTER DICTIONARY, 2016; JAIN, 2010).

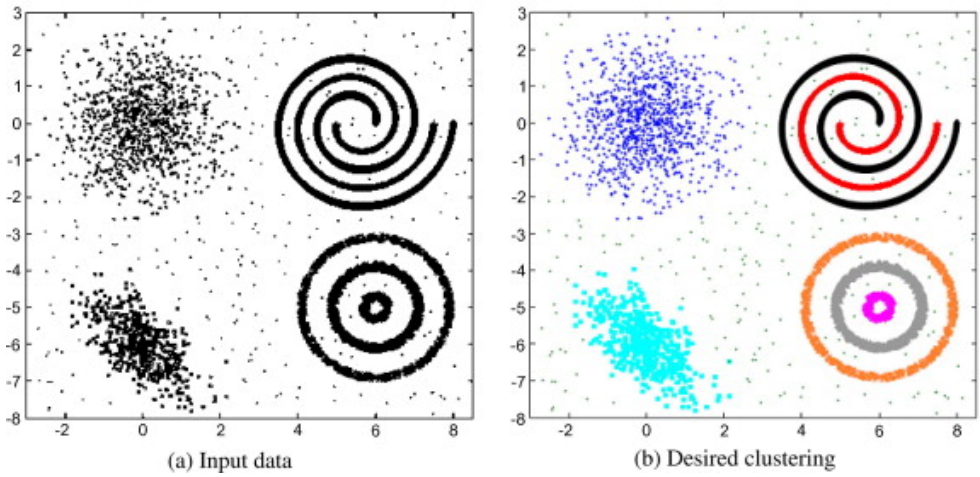


Figure 3.18: Data pre(a) and post(b) clustering (JAIN, 2010)

More studies tried to define some of the main purposes of the clustering technique, Jain (2010) propounded three main goals of clustering: *i*) subjacent structures: used to raise the judging of data, formulate hypothesis, detect anomalies and identify ledges; *ii*) natural classification: seeks to identify the resemblance level between individuals; *iii*) comprehension: employed as a method or group data and summarize them according their specific clusters.

Connel and Jain (2002) used clustering to identify subclasses in handwriting through an online tool. In this study, different users wrote the same character in different ways. According to the literature, when the variance of an element is high, the efficiency in the clustering raises. This is a common example in the real world, where not every data is always the same, and can be displayed in different ways, still being the same data though. As shown in Figure 3.19, the parameter of classification for the same cluster can be presented in distinct ways. It is up to the clustering algorithm to find out the same data is being displayed in a different way. The better the algorithm is, the higher its recognizing precision will be (CONNEL and JAIN, 2002; JAIN, 2010).

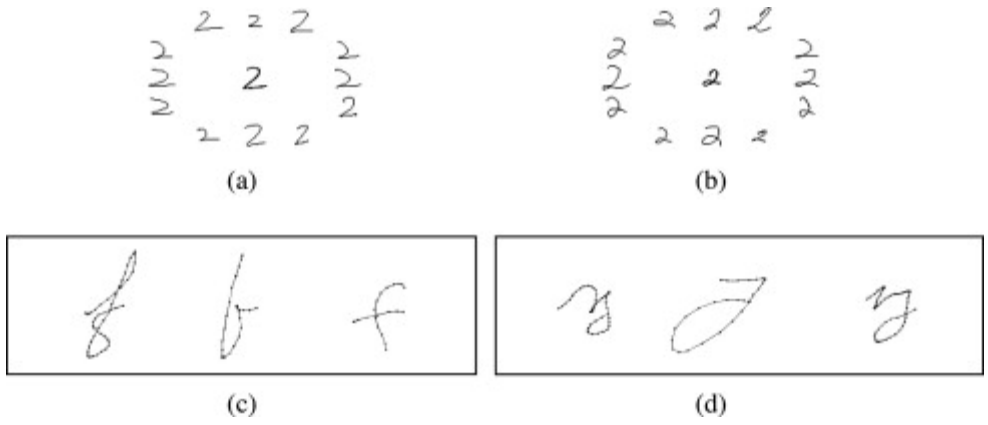


Figure 3.19: Different recognition patterns in the same cluster (CONNEL and JAIN, 2002)

It is possible to divide the clustering algorithms in two groups: hierarchical and partitioned. The hierarchical algorithms recursively find the nested clusters in two means: *i*) in a agglomerative way, where the algorithm starts with a data point in its own cluster and blends to the pair of more similar clusters, successively till it forms a cluster hierarchy; *ii*) in a divisive way, where the algorithm starts with all data points in one cluster and recursively divides each cluster into smaller clusters.

On the other hand, the partitioned algorithms find all clusters simultaneously, while hierarchical algorithms are represented by a matrix *n* x *n*, where *n* is the number of objects to be clustered. The partitioned algorithm makes use of a *n* x *d* matrix, where *n* objects are embedded in a *n* dimension space (JAIN, 2010).

The most common partition cluster algorithm is the *K*-means which was implemented more than 50 years ago and is still used in large scale due to its simplicity, efficiency and empirical success. A dataset consisting in *n* dimensional points can be divided in a set of *k* clusters. The *K*-means algorithm has as its goal to partition this data so that the squared error between the cluster arithmetic average and the points inside the cluster is minimized. This way, the *K*-means tries to diminish the squared error in a set of *k* clusters. The *K*-means starts with a division of *k* clusters and assigns patterns to the clusters, these patterns are based in the called cluster center, where in an optimal universe, all elements inside a cluster should be as close as possible as the center of its cluster. As the squared error always decreases when the number of clusters raise, *K*-means will only minimize the squared error in a fixed number of clusters, the steps performed by *K*-means are disposed in Table 3.7 (JAIN and DUBES, 1998; JAIN, 2010).

Table 3.6: Description of the steps performed by the K-means algorithm (Adapted from JAIN and DUBES, 1998)

Step	Description
1	Select a partition with *k* clusters Repeat steps 2 and 3 until the error is minimum
2	Generate a new partition, assigning the closest value from the cluster center
3	Calculate new cluster centers

The Figure 3.21 illustrated the acting of k-means in a two dimension set with 3 clusters. The first step (a) the input data are presented; (b) three cluster centers are initially selected; (c) and (d) demonstrates middle iterations updating the cluster center and; (e) the final grouping done by k-means (JAIN, 2010).

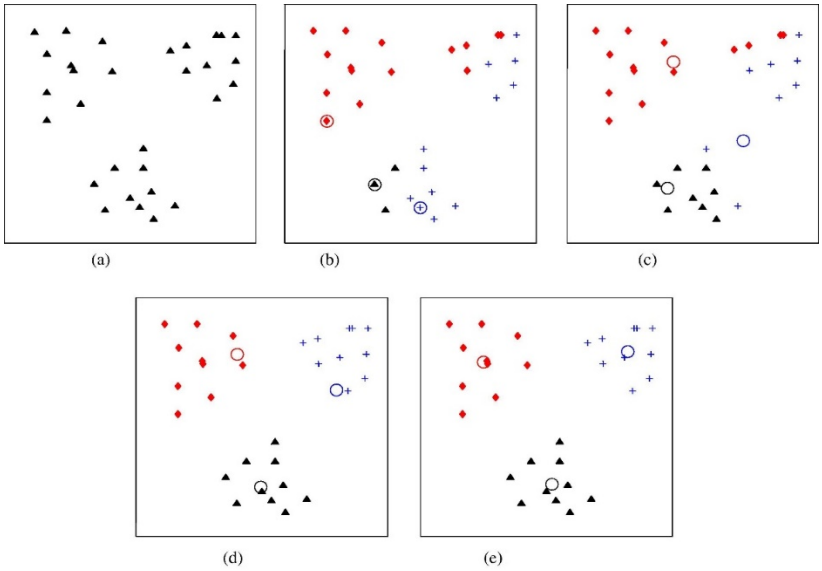


Figure 3.20: K-means illustrated (JAIN, 2010)

In a study from Edgar (2010), the author compared grouping tools to find out a good way to search genetic sequences. The study compared a new version of BLAST, the UBLAST, used alongside with a clustering algorithm, UCLUST, where for certain query, sequence databases are organized in a way the number of matching words is minimized. Taking advantage that small sequences have small sets of words in common, considering this, for an equivalence found in a database, is highly probable that this equivalence is found among the first candidates, and this probability quickly drops when the number of failed attempts of a match increase. This leads to a faster search, with less matchings being analyzed (EDGAR, 2004; EDGAR, 2010).

The comparing study between the tool resulted in the UCLUST algorithm classifying better quality clusters, where the similarity between the results from CD-HIT were inferior in every tested case. The following Table 3.10 shows the comparison between the two clustering tools, UCLUST and CD-HIT. The UBLAST and UCLUST were used and introduce a new paradigm to sturdily group biological data, its use decreases the resource consumption to classify large scale-sequences (EDGAR, 2010).

Table 3.7: Comparative between clustering tools UCLUST and CD-HIT (Adapted from EDGAR, 2010).

Algorithm	Identity (%)	Size	Similarity (%)	Time (mins)	Memory (Mb)
UCLUST	85	536	91.5	1.44 min	36
CD-HIT	95	343	88.9	570 min	349
UCLUST	90	230	96	1.8 min	40
CD-HIT	90	175	92.1	62 min	349
UCLUST	95	73	97.7	134 min	55
CD-HIT	95	68	95.9	61.15 min	349
UCLUST	99	11	99.5	789 min	165
CD-HIT	99	15	99.1	123 min	411

Identity is the clustering limit; Size, the average size of each cluster (higher is better); Similarity is the average identity between a cluster member and its representative sequence (higher is better); Time is the CPU time; Memory refers to the RAM amount used by the software.

4 MATERIALS AND METHODS

This section will explore on how this dissertation was conducted in terms of every tool and step that had to be used to produce results.

4.1 PROMOTER DATA AQUISITION

The data used to conduct this study consists of promoter regions from six different groups, regarding the σ recognition by RNAP, all of these were retrieved from the biological database RegulonDB (GAMMA-CASTRO *et* al., 2015). Table 4.1 represents how many examples of promoter profiles divided into the 6 different sigma factors present in the bacterial genetic expression. These examples are DNA sequences in direction 5'-3' with 81 nucleotides in the same way RegulonDB displays their promoter sequences.

Table 4.1: Distribution of *E. coli*. promoter sequences through this research.

σ factor	Number of sequences
24	508
28	133
32	299
38	157
54	83
70	1869

4.2 DATA TRANSFORMATION

After the examples on Table 4.1 where loaded and converted into numerical values corresponding to the different DNA physical features: entropy, enthalpy, base-pair stacking and stability. The example of each σ group was converted through a Python script, designed in order to automatically convert the examples of each σ group. This algorithm consisted in analyzing a file with all the examples from all σ groups shown in Table 4.1 and had the 81-nucleotide sequence transformed in its correspondent value. The values for entropy, enthalpy, stacking and stability are shown in section 3, under each physical feature sub-section.

Section's 3 Tables 3.2, 3.3, 3.4 and 3.5 indicate a different scale in terms of the physical features. To soothe this difference and bring the results all gathered under the same scale, a normalization algorithm was used to transform the data in the interval 0-1. The algorithm that performed this data normalization was developed by the authors. The data normalization used in this dissertation sought to transform the present data in the 0-1 range, by using the mathematical formula: $ndata = \frac{(x-\min)}{max-min}$ (JUSZCZAK *et* al., 2002), where, *ndata* is the normalized value between 0-1 range; *x* is the information being normalized at moment; *min* is the lowest value found in the whole dataset; *max* represents the highest value found in the dataset.

4.3 K-MEANS CLUSTERING

As soon as the data was ready to perform the promoter sequence characterization, some actions needed to happen to produce the discussed results, these steps are divided into two section, where both explore different usages of the K-means algorithm.

The first approach, presented by section 5.2 indicates a K=2. Two is the minimum value K can assume in terms of data clustering (BHOLOWALIA and KUMAR, 2014), since in this section, the goal was to be able to distinguish - in terms of physical aspects, promoters associated to housekeeping genes, recognized by RNAP σ^{70} and alternative genes, recognized by RNAP σ^{24}, σ^{28}, σ^{32}, σ^{38} and σ^{54} factors.

The second usage of K-means, depicted in Results section 5.3 sought to cluster the data. To cluster the data, the K value had to be found, since there are not only two groups of promoters to be clustered, and previous work from Dal'Alba *et* al., (2018, unpublished data) have clustered stability values and have identified optimal K values (Table 4.2).

Table 4.2 – K in stability (DAL'ALBA *et* al., 2018)

σ factor	K value
24	7
28	3
32	5

38	3
54	2
70	3

However, when the presented K values were used to cluster other physical properties of promoter sequences, some clusters have presented a low amount of sequences in it, thus, the K value is not optimal (JAIN, 2010; BHOLOWALIA and KUMAR, 2014). In order to produce clusters with a higher purity, the K value had to reset following Elbow Method, which indicates a validation to determine the appropriate number of clusters in a given dataset. This method is based on the idea that the percentage of variance is explained by the chosen amounts of clusters, the first cluster will add a lot of information about the profile that data has inside this cluster, however, the amount of information will significantly decrease, this method's idea is to start K=2 and increase by 1 until the cost of this K addition is optimal. Once it starts dropping, the true K for a given dataset is found (BHOLOWALIA and KUMAR, 2014). Table 4.3 presents the clusters with their K value recalculated.

Table 4.3 – New K values for entropy, enthalpy, base-pair stacking and stability values

σ factor	Entropy	Enthalpy	Stacking	Stability
24	4	4	7	7
28	3	3	3	3
32	4	5	5	5
38	2	3	3	3
54	2	2	2	2
70	3	3	3	3

Once the clusters the new clusters were ready, the next step was to perform a series of intersections between this clusters with a straightforward goal: determine in promoters share the same physical properties. To perform the intersections, a Python script had been developed and it checked all the promoters in the most populated cluster from each physical feature. Then, if the same promoter was present in the enthalpy, entropy, base-pair stacking and stability most populated cluster, it could be said that the promoter sequence indicates a higher level of conservation in terms of its physical aspects and its physical profile is worth to be checked.

The workflow represented in Figure 4.1 represents the key moments to reach the discussion section, where, firstly, data is acquired from RegulonDB database, then it has to be converted into its entropy, enthalpy, base-pair stacking and stability values. Then our results are split into a global biological analysis of the physical profile in promoter sequences identified by different RNAP σ factors. The second wave of results include a data clustering using the K-means algorithm, where in the first, the K value is set = 2 and this paper tries to distinguish between housekeeping and alternative RNA σ factors. The second, uses the Elbow approach for setting the K and enabling a physical profile for each σ factor.

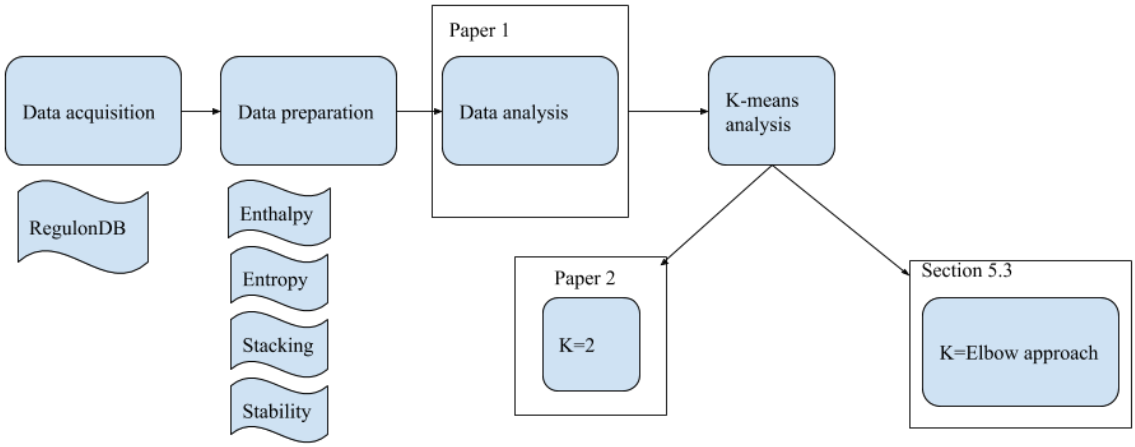


Figure 4.1: Process workflow in this dissertation.

5 RESULTS AND DISCUSSION

This dissertation's results are organized through two articles. The first one seeks to propose a biological analysis in promoter sequences physical features. The physical aspects' comparison of promoter sequences recognized by different σ groups enables a better comprehension on how promoter sequences behave in comparison to other metrics used to classify promoters: the presence of consensual motifs.

The second article, yet to be submitted to publishing, uses the same physical aspects presented by the Article 1 in terms of the use of the clustering technique. This second paper aims to be able to look deeper in promoter sequences that share similar values in terms of their physical features. The sequences were submitted to an intersection assessment to determine which cluster has indicated a larger number of promoter sequences, then, each cluster is separately analyzed, and the examples that are present in all features are extracted.

The third section in the results, deals with an intersection analysis of our most populated clusters. The main step here was to select the clusters from enthalpy, stability and base-pair stacking profiles in different sigma groups and intersect these promoters, enabling to assess the profile of all the promoters that shared the same features.

5.1 ARTICLE 1 – COMPARISON OF ENTROPY, ENTHALPY, STABILITY AND BASE-PAIR STACKING PROFILES OF *E. COLI* IN PROMOTER SEQUENCES RECOGNIZED BY DIFFERENT σ FACTORS

COMPARISON OF ENTROPY, ENTHALPY, STABILITY AND BASE-PAIR STACKING PROFILES OF *E. COLI* IN PROMOTER SEQUENCES RECOGNIZED BY DIFFERENT σ FACTORS

Gustavo Sganzerla Martinez[1]*, Scheila de Ávila e Silva[1]

University of Caxias do Sul - Biotechnology Institute

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

ABSTRACT

The transcription of gene expression in bacteria counts with the presence of a promoter sequences. These are conserved DNA sequences that precede the coding region and tells the RNAP enzyme where to start producing RNA, different sigma factors that are recognized by a subunit in the RNAP can start the transcription of genes with different functionalities. Although, the somewhat conserved promoter regions show variations that difficult their identification by simple nucleotide sequence analysis. The conservation of the nucleotide content this DNA site is not absolute for all promoter sequences. In face of this, there are physical aspects of DNA, such as enthalpy, entropy, stability and base-pair stacking, that aids promoter prediction. In this paper, we propose to analyze the beforementioned measurements to help in promoter sequence understanding. There are not many promoter sequence recognition and identification tools that make use of combined physical aspects of the DNA. The results that were concluded in this paper tell us that the tested physical aspects, are, somehow, entwined and each different group of promoter sequences behaves differently in terms of their physical point of view. The results produced here may aid in bacterial promoter recognition, by delivering this area a stronger set of biological inferences.

Key-words: bacteria promoter; DNA structural properties; gene transcription

1 INTRODUCTION

The transcription process is a key factor during gene expression. This is no different in bacteria, which are sensitive to environmental changes, and requires a regulating system that prioritized which stresses conditions should be considered, granting survivability to the cell [1, 2, 3, 4].

A key element in the transcription process is the promoter sequence. In a simple definition, a promoter sequence is a DNA segment placed upstream of the coding region [5]. When analyzing a genome, an important footmark for researchers is to search for genes firstly, and then seek for the promoter sequence that precedes the gene in question. This being, we can comprehend more way about the gene transcription [6].

During transcription, an important actor is the enzyme RNA polymerase (RNAP). This protein is formed by different polypeptides and units, among which there is the sigma (σ) subunit that is responsible for identification and attachment to specific DNA sequences called promoters. Different σs can start the transcription of different gene groups by associating to different promoters regulating gene expression of a group of genes [3].

There are several critical factors that can tell promoters apart from other DNA regions and help RNAP targeting. An example of this are the consensual motifs, around the -10 and -35 upstream the gene. The study of these consensual regions has shown that promoters are somehow similar but not identical and have sites of a conservation level, however, these regions are not all the same [3]. Additionally, some bacterial promoters there is an overlapping of the promoter into the coding region, demanding a detailed analysis up and downstream of the site where RNA begins to have its nucleotides inserted, the transcription start site (TSS). To overcome this challenge and improve the *in-silico* analysis, approaches can be stated considering a way to look through the DNA molecule and extract more than just a simple nucleotide composition analysis linked to the consensual regions [7].

Despite the effort to computationally identify bacterial promoters, they still represent a challenge due to the wide variety of profiles that can be found.

There are lots of efforts to computationally identify promoters. But this is still a challenge due to the wide variety of profiles that can be found in these regulatory

sequences. When identifying a promoter, features need to be considered to enhance the sensitivity of the tool. It is biologically known that promoter regions tend to be conserved when compared to coding regions. Some tools have been developed over time to aid in this field, some of them are: BTSS Finder [9], CNN promoter [11] and BacPP [8].

There are some features that RNAP uses to identify promoters, these are DNA physical aspects such as entropy, enthalpy, stability, base-pair stacking values, stress induced DNA duplex destabilization and DNA curvature and bendability. Those used in this paper are: entropy, enthalpy, stability and base-pair stacking. The entropy and enthalpy are thermodynamic values that measure the amount of energy that is not converted into labor. Living beings perform a number of processes trying to keep their biological functions, requiring energy to occur. Thus, the organism needs a stable energy supply to keep functioning. The chemical bonds, mass transport in and outside the cell, heat spawning and information stored in the genetic code are all examples of processes that cause entropy/enthalpy variation in a cell. The stability of base pairs is a characteristic that refers to the amount of free-energy present in a base pair interaction. The base-pair stacking is a value found in the bond between the RNAP and the DNA molecule [11, 12, 13].

The study of physical properties of a sequence enables a wide comprehension on how enthalpy, entropy, stability and base pair stacking correlate with each other enlighten in promoter sequences comprehension [14]. The understanding how the DNA physical aspects behave on promoter sequences may bring the scientific community a better understanding on the molecular structure itself. In this context, this paper aims to rely on the physical features in order to distinguish different promoter sequences, recognized by different sigma groups. It is believed, according to the literature that a deeper comprehension of these features may aid in promoter recognition tools. According to the literature, entropy, enthalpy, base pair stacking and stability show different behavior inside the promoter region when compared to other DNA sites. These differences can be used as parameters in promoter identification and prediction tools.

2 MATERIALS AND METHODS

The data used to conduct this paper was promoter regions from 6 different groups retrieved from the biological database RegulonDB [15]. Table 1 shows the number of examples divided into the 6 different sigma factors of gram-negative bacteria.

Table 1: *E. coli* A-T-G-C promoter sequences displayed in 5'-3' termination.

σ factor	Number of sequences
24	508
28	133
32	299
38	157
54	83
70	1869

All the examples shown in Table 1 were converted into values for the four physical features that were tested in this paper: entropy, enthalpy, base-pair stacking and stability (Table 2) considering the 16 duplexes. Once the values were converted in each distinct DNA physical feature, this data was normalized due to their different in scale in order to analyze how these features correlate with each other. The data normalization technique chosen was 0-1 normalization.

Table 2 – Entropy, enthalpy, stacking and stability values in the DNA [11, 16].

Nucleotide Duplex	Entropy	Enthalpy	Base-pair stacking	Stability
AA	-21.3	-7.6	-5.37	-1
AT	-20.4	-7.2	-6.57	-0.88
TA	-21.3	-7.2	-3.82	-0.58
AG	-21	-8.2	-6.78	-1.3
GA	-21	-8.2	-9.81	-1.3
TT	-21.3	-7.6	-5.37	-1
CC	-19.9	-8	-8.26	-1.84
GC	-24.4	-10.6	-9.69	-2.27
AC	-22.7	-8.5	-10.51	-1.45
CA	-22.7	-8.5	-6.57	-1.45
TG	-22.4	-8.4	-6.57	-1.44
GT	-22.4	-8.4	-10.51	-1.44
TC	-22.4	-7.8	-9.81	-1.28
CT	-22.2	-7.8	-6.78	-1.28
CG	-27.2	-10.6	-14.59	-2.24
GG	-19.9	-8	-8.28	-1.84

The final step was to use the tool Weblogo [17] to analyze the conserved nucleotides in positions through the 81-length promoter sequence. The goal of the use of this tool is to check among a set of *n* examples the most common nucleotide.

3 RESULTS AND DISCUSSION

The results involved the analysis promoter sequences from six different σ groups of *E. coli* and the main goal was to assess how entropy, enthalpy, base-pair stacking and stability behave in each one of the groups. It its already know that there are different consensual regions in the promoter sequences recognized by each σ factor. These consensual regions will aid whenever the RNAP enzyme σ subunit needs to look for promoter sequences and bind itself in the DNA. These consensuses are, somehow, conserved around their -10 and -35 positions. Apart from sequence similarity [3] there are other factors that RNAP look for when its σ subunit searches for promoter sequences, these factors are the local and distinctive physical features of the promoter regions.

It is worth mentioning that a cell performs a handful of processes to survive, these processes can cause a thermal variation in the cell and its environment. Some components such as mitochondria. Thus being, in thermodynamic means, a cell is not so different from a machine, where the energetic balance is always sought [14].

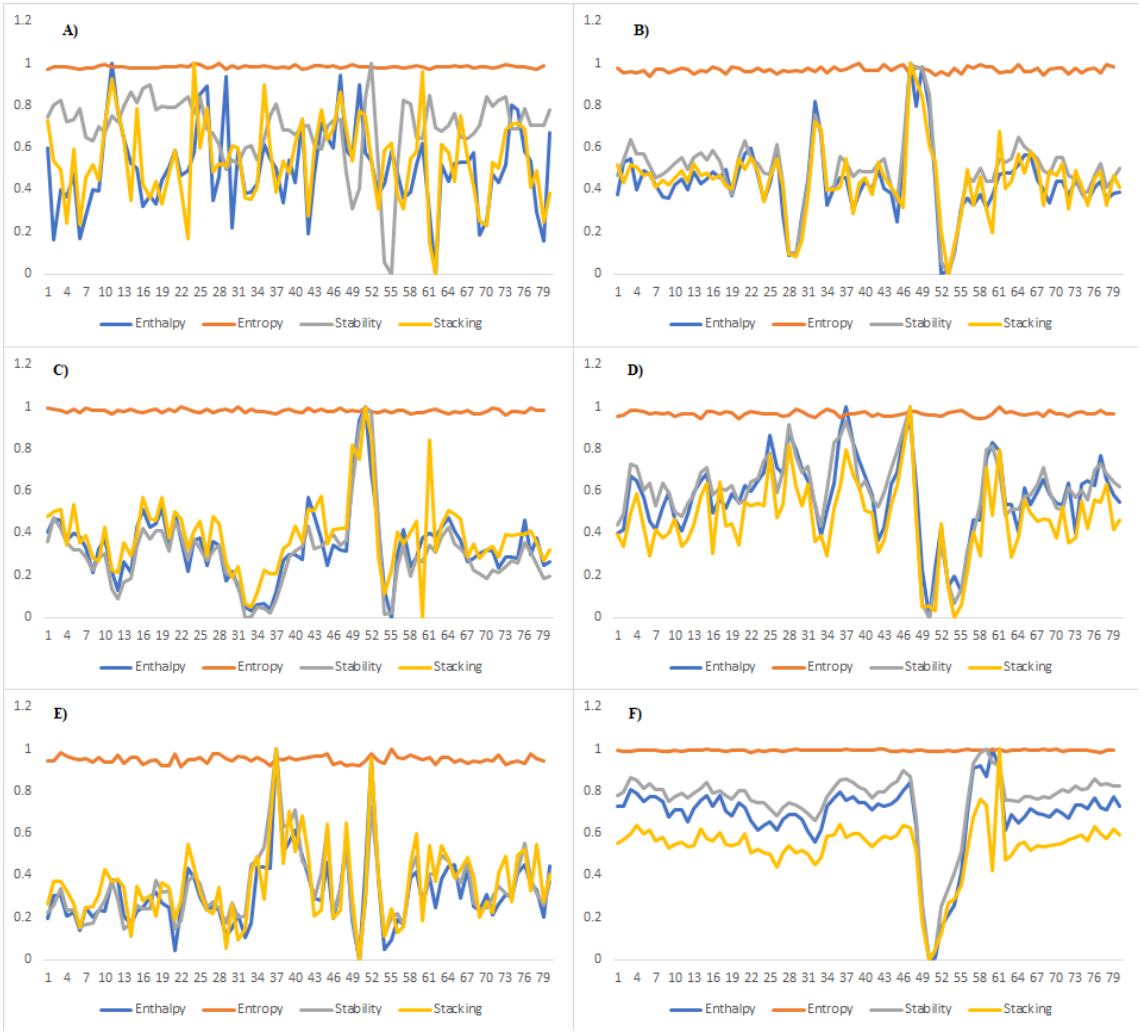
In the results displayed in Figure 1, it is possible to perceive that around position -10, represented by the 49 on x axis, there is a variation on the lines for three features. This means that enthalpy, stacking and stability are meaningful features to distinguish promoters in between different sigma groups. As the image also shows, the values are different in every single group, this indicates that the groups themselves are not alike, and not just because they are all promoters, their physical features behave in a similar way.

In terms of the σ^{24} promoters, they showed noisy results Figure 1 (A) and the lines did not exhibit overlapping. It is important to print out that almost 80% of σ^{24} promoters listed in RegulonDB were predicted *in silico* and not confirmed by biological experiments [18].

The other σ promoters, σ^{28} (B), σ^{32} (C), σ^{38}(D), σ^{54}(E) σ^{70}(F) have shown that on their series there is a clear and similar protuberance around -10 and -35, all of them have presented a similar variation regarding enthalpy, stacking and stability. These peaks match the consensual motifs reported in the literature presents for *E. coli.* promoters [3]. These variations around the motifs indicate that there are structural/physical differences in binding sites where the RNAP connects to the DNA molecule. This suggests that the σ subunit also looks for the presence of this distinguished set of physical profiles that

promoter sequences have around the -10 and -35 regions, not only relying on the consensual regions that each σ factor presents in *E. coli.* promoter recognition.

Figure 1 – Each physical feature and its averages per position in the sigma groups



This figure represents the average value per position in entropy, enthalpy, base-pair stacking and stability values in all promoters tested in this paper, displayed in Table 4.1 under Materials and Methods. (A) indicates promoters recognized by σ^{24}, (B) σ^{28}, (C) σ^{32}, (D) σ^{38}, (E) σ^{54} and (F) σ^{70}.

The online tool Weblogo [17] was used to check the average nucleotide composition in each one of the 81 length promoter sequences presented in Table 1 and the results are: *i*) σ^{24} have a prevalence of A nucleotides 25^{th} – 28^{th} nucleotide A, and a predominance of A/T from the range 49^{th} to 56^{th}; *ii*) from the 133 σ^{28} promoters, the -10 position has a TGCAAT sequence surrounding the -10 position, while the -35 region is composed by A/T nucleotides; *iii*) σ^{32} promoter sequences recognized by sigma 32, around the degenerated -35 and -10 [19] the first have indicated a slight advantage towards T nucleotides a higher C presence on the second; *iv*) sigma 38 examples show a

leaning to A/T nucleotides around its -10 and a subtle A/T predominance around its -35; *v*) sigma 54 examples has just presented a low G content on its -10 and no predominance at all on its -35; *vi*) finally, sigma 70 promoters have presented a high T composition around its -35 and the classic TATAAT motif on its -10. The results from Weblogo are displayed on Figure 2.

The stability and stacking series shown on Figure indicated that in most cases, this feature behaves differently when approaching the -10 and -35 consensual regions. The literature asserts that the whole stability level in promoter sequences tends to be lesser than that of coding genomic sequences due to is constant opening. The promoter is a region that during the transcription process is opened by RNAP and have its nucleotides written to the mRNA. For this process to be energetically viable, it is reasonable that the promoter has a lower stability that other genomic regions. A/T bindings present two hydrogen bonds whereas G/C base pairs have 3 hydrogen bonds, which makes promoters regions pro present a higher A/T presence due to its constant opening [12, 20, 8].

Analyzing the entropy and enthalpy values, it is clear that in a promoter sequence context, there is no alteration regarding the consensual regions when the feature is entropy. Early studies [13] have shown that due to the GC content, promoter sequences would have a different enthalpy and entropy value in comparison to other genomic sequences. Authors [13] have shown that the entropy value for an AT base pair stabilization would be 31±3 KJ/mol-bp and the average enthalpic contribution for a GC base pair stabilization is 45±6 KJ/mol-bp. Regarding the entropy necessary for base pair stabilization, AT is 86±10 KJ/mol-bp and GC is 114±15 KJ/mol-bp. Promoters are known to be poor in their GC content. The GC amount affects DNA stabilization both in the enthalpic and entropic contribution. During the RNAP binding to the promoter sequences there are other elements present in this RNAP – DNA binding. According to the literature, the more complex a system is, the bigger the entropy of the same system is. If we analyze, during this connection between RNAP and DNA, the other elements involved will raise the complexity of the system itself, there are water molecules, minerals that increase the entropy level. This higher complexity brought by AT nucleotides binding to water molecules explains the higher entropy levels [13, 21, 22] – basically all near the maximum 1 value in our graph - that Figure 1 shows.

Figure 2 – Weblogos for the six σ groups tested in this paper

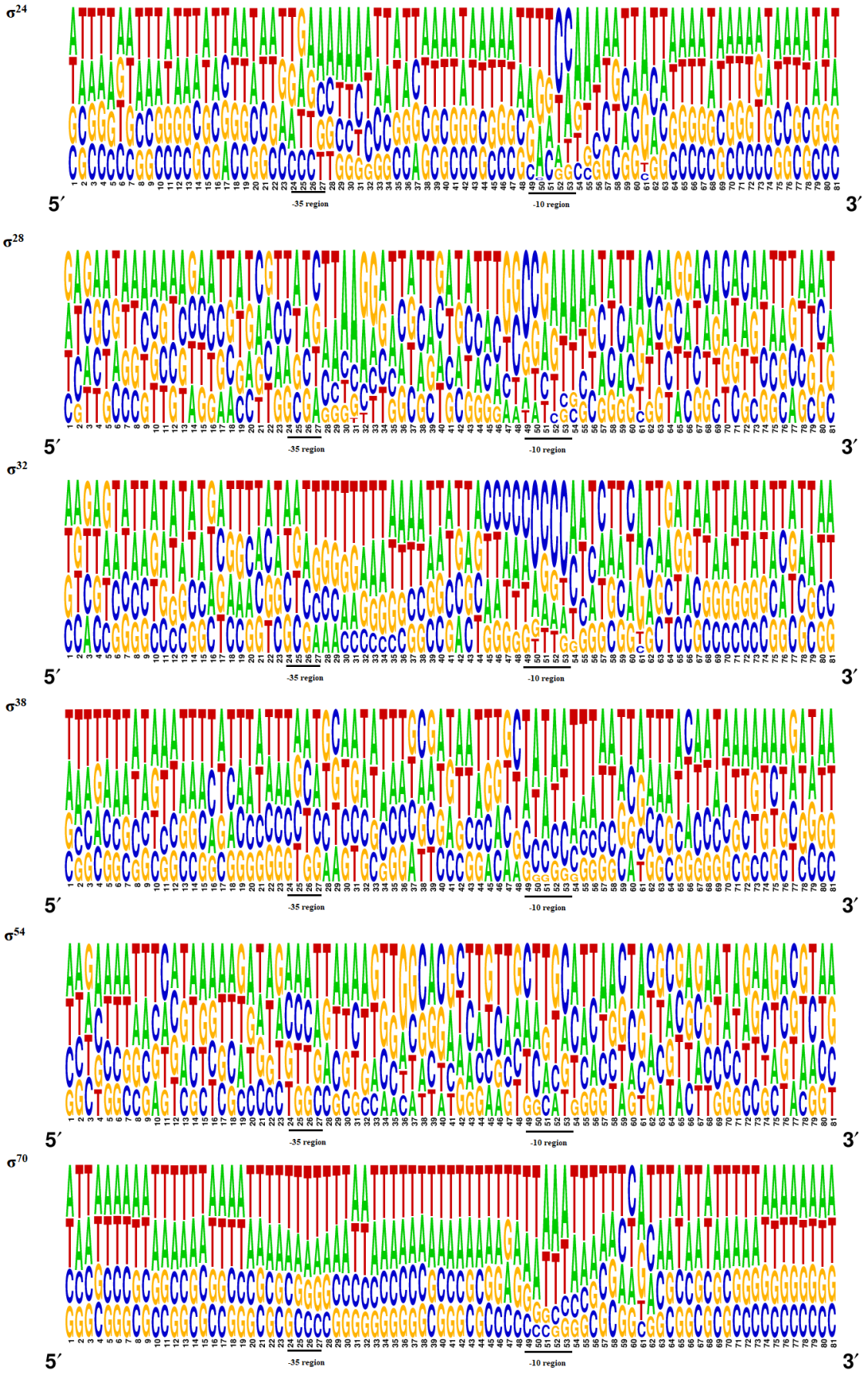


Figure also 1 shows that the entropy levels carried close to no alteration regarding consensual regions, so a question is raised: how different entropy is and behaves from the other tested DNA features? We analyzed that entropy and enthalpy are similar features, both related to thermodynamics of the DNA, but the way that these two measurements behave is different. The literature asserts that enthalpy refers to the system as a whole, in addition, we can perceive a system presence in this DNA-enzyme-water interaction – water is a major component that is present in the hydrogen bonds and directly affects entropy levels – this whole system is not organized, the entropy tends to decrease as the RNAP moves along the DNA strand, with the water connecting to the DNA. On the other hand, enthalpy does not refer to all the components involved in this systems, featuring as a standalone characteristic, the enthalpy is only affected by the hydrogen bonds found in the nucleotide connection. This leads us to the remark that entropy is a system measurement and feature, but at the same time, it corresponds to DNA thermodynamics, it cannot be measured alone, it is highly affected by the other components in the system [13, 21].

In terms of base-pair stacking as a feature to distinguish promoter sequences in different σ groups our results indicated that base-pair stacking in Figure 1 demonstrates an overlapping with stability. ZHANG *et* al. (2015) performed an analysis using an Atom Force Microscopic to evaluate the base pair hydrogen bond strength and base pair stacking force in DNA strands. The authors have come up with the data showing that GC binding strength would consist as 20 piconewton (pN) and AT base pairs 14pN, the stacking force in adjacent base pairs is estimated by the authors by being 2pN. The binding strength in GC duplexes makes sense when in promoter regions where the GC amount does not exceed the AT content. Turning promoters in a so-called weaker sequence in terms of its bindings and adjacent base pairs [12, 23, 24]. This close connection between base-pair stacking strength and DNA stability explains the superposition displayed in Figure 1.

There may be a leaning towards AT nucleotides composing the promoter sequence, as some sigma groups show on Table 3, it would perfectly make sense for promoters having more AT nucleotides than GC due to the amount of force involved with these base pairs. However, [13] have portrayed that the forces involved in the bond between protein (RNAP) and DNA strand are not simple. They need to be weak enough to allow the protein to easily scan the DNA and, simultaneously, must be strong enough

for longer-living connections. Studies have presented that the enthalpy value of a sequence follows the increases of the stability value, as Image 2 depicts, in all cases, the enthalpy line follows the stability line. This is due to the enthalpy being associated with the bond between RNAp and the DNA strand [12, 13].

The results shown in Figure 1 have indicated a strong correlation in terms of enthalpy, base-pair stacking and stability profiles for five out of the six σ groups that got tested. This correlation is explained due to the physical connection that these three features have between themselves [8, 12, 13, 24]. Promoter sequences recognized by different σ groups have shown that in each σ groups, the way these features behave is different, the values on Figure 1 vary according to each group. Even though, they all presented some sort of leaning towards the consensual motifs, which aids the RNAP recognition.

The Table 3 indicates the correlation level between the four physical features tested. The correlation teste used was Spearman, due to the data not following normal distribution. All the features indicate a level of correlation.

Table 3 – Correlations

	Enthalpy	Entropy	Stability	Stacking
Enthalpy	1.0	.336**	.815**	.821**
Entropy	.336**	1.0	.391**	.310**
Stability	.815**	.391**	1.0	.707**
Stacking	.821**	.310**	.707**	1.0

*. The correlation is significant at the 0.05 level (2 extremities).

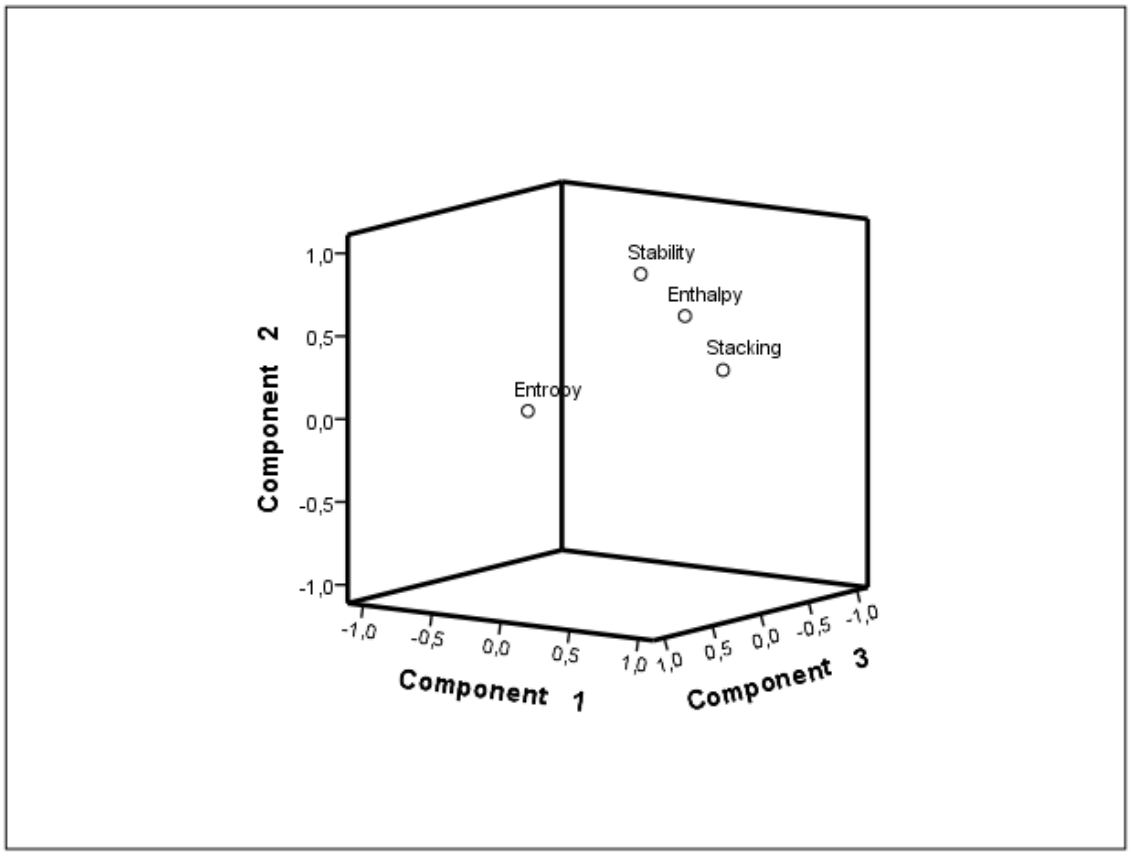
Table 4 shows a main component analysis, where 3 components explain 93.3% of our data variance. At the same time, Figure 3 shows a rotated space component diagram, where it can be perceived exactly what Figure 1 has indicated: enthalpy, base-pair stacking and stability are entwined, while entropy is not.

Table 4 – Main components analysis

	Component		
	1	2	3
Enthalpy	.250	.304	-.081

Entropy	-.045	-.202	1.055
Stability	-.648	1.204	-.114
Stacking	1.181	-.689	-.019

Figure 3 - a rotated space component diagram



4 CONCLUSIONS

A better comprehension between chemical bonds and DNA can aid to develop better tools to predict how strong two molecules will interact only by knowing their structure. The results presented on Figure 1 let us know that, except for σ^{24}, every other sigma group behaves differently in terms of their physical features. This gives us the opportunity to distinguish promoters with their associated sigma factor. There is only one feature that could not be used to support this idea, which is the entropy.

It is already known that promoter prediction is not a simple task due to several deformations, mutations and overlapping found in these regulatory sequences. Tools that only seek for sequence similarity may sound outdated and limited. The results gathered

here can enlighten *in silico* promoter prediction and characterization by providing the bioinformatics field a deeper analysis in the physical features present in promoters that got tested in this paper. In a future study, more segments of a genome may be compared with the promoter sequence to analyze the entropy, enthalpy, base-pair stacking and stability level outside the promoter sequence and combine these biological inferences in artificial intelligence techniques.

5 REFERENCES

[1] Cases, I., de Lorenzo, V. & Ouzounis, C. A. (2003). Transcription regulation and environmental adaptation in bacteria. **Trends Microbiol.** 11, 248–253.

[2] McAdams, H. H., Srinivasan, B. & Arkin, A. P. (2004). The evolution of genetic regulatory systems in bacteria. **Nature Rev. Genet.** 5, 169–178.

[3] Krebs, J. E.; Goldstein, E. S., Kilpatrick, S. T. Lewin's GENES XI, Jones & Bartlett Publ., 2014.

[4] Barnard, A., Wolfe, A. & Busby, S. (2004). Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. **Curr. Opin. Microbiol.** 7, 102–108.

[5] Kim, T. H.; Barrera, L. O.; Zheng, M.; Qu, C.; Singer, M. A.; Richmond, T. A.; WU, Y.; Green, R. D.; Ren, B. (2005). A high-resolution map of active promoters in the human genome. **Nature**, v. 436, n. 7052, p. 876–880, ago.

[6] Abeel, T.; Peer, Y. Van de; Saeys, Y. (2009). Toward a gold standard for promoter prediction evaluation. **Bioinformatics**, v. 25, n. 12, p.313–320.

[7] 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**. Vol. 16 (1). 16 p.

[8] De Ávila 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**, v.287, p.92-99.

[9] 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**, *33*(3), 334–340. http://doi.org/10.1093/bioinformatics/btw629

[10] Bedoya, Ó., & Bustamante, S. (2011). Cnn-promoter , New Consensus Promoter Prediction Program Based on Neural Networks.

[11] Santalucia, J., Jr.; Hicks, D. (2004). **Annu. Rev. Biophys. Biomol. Struct:** The thermodynamics of DNA structural motifs. 33. p. 415-440.

[12] Kanhere,A. and Bansal,M. (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. **Nucleic Acids Res**., 33, 3165–3175.

[13] Privalov, P. and Crane-Robinson, C. (2018). **Progress in biophysics and molecular biology:** forces maintaining the DNA double helix and its complexes with transcription factors. 135, 30-48. https://doi.org/10.1016/j.pbiomolbio.2018.01.007

[14] Davies, P. C. W., Rieper, E., Tuszynski, J. A. (2013). Self-organization and entropy reduction in a living cell. **National Institute of Health**. January ; 111(1): 1–10**.**

[15] Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A, Hernández-Koutoucheva A, Moral-Chávez VD, Rinaldi F, Collado-Vides J.

[16] Ornstein, R. L., Rein, R. , Breen, D. L. and Macelroy, R. D. (1978), An optimized potential function for the calculation of nucleic acid interaction energies I. Base stacking. **Biopolymers**, 17: 2341-2360. doi:10.1002/bip.1978.360171005

[17] Crooks, G. E., Hon, G., Chandonia, J. M., Brenner, S., E. (2004). WebLogo: A sequence logo generator, ***Genome Research***, 14:1188-1190.

[18] Shimada, T., Tanaka, K., & Ishihama, A. (2017). The whole set of the constitutive promoters recognized by four minor sigma subunits of *Escherichia coli* RNA polymerase. **PLoS ONE**, *12*(6), e0179181. http://doi.org/10.1371/journal.pone.0179181

[19] Burr, T., Mitchell, J., Kolb, A., Minchin, S., Busby, S. (2000). DNA sequence elements located immediately upstream of the -10 hexamer in *Escherichia coli* promoters: a systematic study. **Nucleic Acids Research**, vol. 28. n. 9. p. 1864-1870.

[20] Ramprakash, J.; Scwarz, F. P. (2007). Identification and annotation of promoters regions in microbial genome sequences on the basis of DNA stability. **Journal of Biosciences** 32: 851-862.

[21] Xu, X., Zhi, X., & Leng, F. (2012). Determining DNA Supercoiling Enthalpy by Isothermal Titration Calorimetry. **Biochimie**, *94*(12), 2665–2672. http://doi.org/10.1016/j.biochi.2012.08.002

[22] Morgunova, E., Yin, Y., Das, P. K., Jolma, A., Zhu, F., Popov, A., … Taipale, J. (2018). Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. **eLife**, *7*, e32963. http://doi.org/10.7554/eLife.32963

[23] Meysman, P., Collado-Vides, J., Morett, E., Viola, R., Engelen, K., Laukens, K. (2014). Structural properties of prokaryotic promoter regions correlate with functional structures. **PLOS ONE**. v. 9 n. 2.

[24] Zhang, T., Zhang, C., Dong, Z., & Guan, Y. (2015). Determination of Base Binding Strength and Base Stacking Interaction of DNA Duplex Using Atomic Force Microscope. **Scientific Reports**, *5*, 9143. http://doi.org/10.1038/srep09143

5.2 ARTICLE 2 – A LOOKTHROUGH TO CLUSTERS CONTAINING ENTROPY, ENTHALPY, BASE-PAIR STACKING AND STABILITY PROFILES OF *E. COLI* PROMOTER SEQUENCES IDENTIFIED BY DIFFERENT σ FACTORS.

A LOOKTHROUGH TO CLUSTERS CONTAINING ENTROPY, ENTHALPY, BASE-PAIR STACKING AND STABILITY PROFILES OF *E. COLI* PROMOTER SEQUENCES IDENTIFIED BY DIFFERENT σ FACTORS

ABSTRACT

The presence of a promoter sequence is a key element to start gene transcription, since the enzyme RNA polymerase targets this specific DNA segment. An element that the RNAP employs to find promoter sequences is the recognition of specific consensual regions with some degree of nucleotide retention. However, not all the sequences present the same consensual region, which turns the *in-silico* promoter prediction a harsh task. Other than the simple presence of a set of nucleotides, there are aspects of the DNA that are found in a different level in promoter sequences: entropy, base-pair stacking and stability. We aim to use these features as a way to characterize promoter sequences recognized by different σ groups. In order to produce biological inferences on the behavior of promoter sequence in terms of their physical profile the clustering technique was used to assess specific promoter sequences in terms of their physical properties. The results this paper presented enabled a link between RNAP to find consensual regions in promoters, when analyzed, these sites have presented some degree of conservation.

1 INTRODUCTION

The DNA molecule is known for storing genetic information on how to produce critical products to the functioning of a cell, which results from an important molecular process: the gene transcription (KREBS, GOLDSTEIN and KILPATRICK, 2014). One important actor on how effectively the transcription works is the presence of regulator elements. These elements ensure that the correct gene has its production prioritized when the cell needs nutrients in order to answer environmental changes and thus, guarantees the cell survivability. The understanding on how these elements function provides the

scientific field a deeper comprehension in biological mechanisms, leading researches to better understand organisms (MACADAMS, 2004; SANDERS and BOWMAN, 2014).

One common regulator element found in all living beings is promoter sequences. A promoter sequence is a DNA segment found in intergenic regions, before the DNA segment to be written into RNA. Acting as a flag to the RNA Polymerase (RNAP) enzyme, the promoter sequences aid organisms to optimize their metabolic answers, providing the cell a full control on its transcription process. Promoter sequences are known to present some levels of nucleotide retention, the GC content present in these sequences are lower when the rest of the genome is taken as an example. This specific nucleotide retention will help RNAP to identify these sequences (KREBS, GOLDSTEIN and KILPATRICK, 2014).

Both *eukaryote* and *prokaryote* organisms count with this enzyme in their gene transcription. In bacteria, RNAP presents six major σ subunits (σ^{24}, σ^{28}, σ^{32}, σ^{38}, σ^{54} and σ^{70}). The σ factor checks the specificity of the transcription process, since each σ can start the transcription of different genes. One of the metrics RNAP σ subunit uses to find promoters is the presence of a conserved sequence in positions -10 and -35 upstream the transcription starting site (TSS). As instance, the organism tested in this paper, *E. coli*, presents in its promoter sequences recognized by σ70 the motifs TTGACA and TATAAT, respectively (ABEEL *et* al., 2009; KREBS, GOLDSTEIN and KILPATRICK, 2014).

With the presence of conserved sequences, the task to identify promoters *in silico* seems simple. A software must only look the genome and find these conserved regions in order to find promoter sequences, task that can be easily done with the current level of computer processing power. However, the nucleotide composition of the promoter sequences is not the same for sequences recognized by a given σ factor. Besides the conservation of the consensual motifs are not exactly present in all sequences. In face of this, this paper proposes to include the assessment of promoter sequences the presence of DNA physical features. The features that are tested in this paper are: stability, enthalpy, entropy and base-pair stacking. Stability refers to the amount of free energy found in base pair interaction. The literature asserts that in GC bonds, the number of hydrogen bonds is 3 and within AT links the hydrogen bond number is 2. When RNAP binds to DNA, there is a process that opens the DNA strand, this opening, mediated by RNAP will read the DNA nucleotides and insert the correspondent information into the RNA strand. This opening process involves the breaking of DNA strand, which tends to be energetically

viable to open segments that are less stable. Base-pair stacking refers to twisted look that a DNA molecule has. Stacking values refers to the interaction of adjacent nucleosides that connects to the DNA phosphate strand, it has been found that the relative force needed to unstack AT and GC base pairs is different. So, it is expected that promoters present different stacking values in comparison to other genomic regions (KANHERE and BANSAL, 2005; DAVIES *et* al., 2013; PRIVALOV and CRANE-ROBINSON 2018).

The processes that involve gene regulation will cause the cell to have a thermodynamic variation. Entropy and enthalpy are measurements of DNA sequences that contribute to the DNA stabilization. MARMUR and DOTY, 1962 have found that the GC content in promoters present a different entropy and enthalpy values when compared to other genomic sequences. During the RNAP and DNA connection, there are other elements found, such as water and minerals. The presence of these elements directly affects the entropy levels due to the raise of the complexity of the system (ZU, ZHI and LENG, 2012; MORGUNOVA *et* al., 2018; PRIVALOV and CRANE-ROBINSON, 2018)

As mentioned before, *in silico* promoter recognition must not rely only in the presence of consensual regions. The addition of physical aspects of the DNA will enhance the analysis performed by promoter recognition tools. The use of artificial intelligence techniques is another step in terms of a reliable promoter sequence characterization. Clustering technique is an artificial intelligence usage commonly used in several areas, one of them is biology. The goal of clustering technique is to discover patterns that enable a wider comprehension of any given dataset. This is done through the recognition of certain features of a cluster that could not be perceived in advance, this enables a deeper understanding of a population based on any specific feature they may present (CONNEL and JAIN, 2002).

By these means, this paper seeks to analyze clusters containing promoter sequences from different σ groups. These clusters are organized in terms of four physical aspects of DNA: entropy, enthalpy, base-pair stacking and stability. The deep analysis provided by the clusters indicates how these promoter sequences behave in terms of their physical magnitudes, granting the biotechnology field a better comprehension on these specific regulatory sequences with a solid alternative to characterize promoters.

2 MATERIALS AND METHODS

The promoter sequences examples used in this paper were retrieved from the database RegulonDB (GAMA-CASTO *et* al., 2015) and comprehend sequences organized into six different groups, associated with the σ factor that recognizes this promoter by RNAP enzyme. In total, it was used ATCG sequences in termination 5'-3', divided in the σ factor that recognizes the sequence (Table 1).

Table 1: Distribution of examples through the research.

σ factor	**Number of sequences**
24	508
28	133
32	299
38	157
54	83
70	1869

Once the examples were available, the next step was to convert the nucleotides of promoter sequences in terms of four of their physical features: enthalpy, base-pair stacking and stability according to (ORNSTEIN *et* al., 1978; SANTALUCIA and HICKS, 2004). The values are disposed on Table 2.

Table 2 – Entropy, enthalpy, stacking and stability values in the DNA (ORNSTEIN *et* al., 1978; SANTALUCIA and HICKS 2004)

Nucleotide Duplex	**Entropy**	**Enthalpy**	**Base-pair stacking**	**Stability**
AA	-21.3	-7.6	-5.37	-1
AT	-20.4	-7.2	-6.57	-0.88
TA	-21.3	-7.2	-3.82	-0.58
AG	-21	-8.2	-6.78	-1.3
GA	-21	-8.2	-9.81	-1.3
TT	-21.3	-7.6	-5.37	-1
CC	-19.9	-8	-8.26	-1.84
GC	-24.4	-10.6	-9.69	-2.27
AC	-22.7	-8.5	-10.51	-1.45
CA	-22.7	-8.5	-6.57	-1.45
TG	-22.4	-8.4	-6.57	-1.44
GT	-22.4	-8.4	-10.51	-1.44
TC	-22.4	-7.8	-9.81	-1.28
CT	-22.2	-7.8	-6.78	-1.28
CG	-27.2	-10.6	-14.59	-2.24
GG	-19.9	-8	-8.28	-1.84

With the values converted, a data normalization had to be done to eliminate the difference in scale presented in Table 2. The data normalization used in this paper sought to transform the present data in the 0-1 range, by using the mathematical formula: $ndata = \frac{(x-\min)}{max-min}$, where, *ndata* is the normalized value between 0-1 range; *x* is the information being normalized at moment; *min* is the lowest value found in the whole dataset; *max* represents the highest value found in the dataset (JUSZCZAK *et* al., 2002). The next step was to use the clustering technique.

The purpose of data clustering is to find hidden patterns in a given dataset, providing a better comprehension of the data (JAIN, 2010). The most popular use of non-hierarchical data clustering is the K-means algorithm. This algorithm seeks to find a way to part the data in a way that the squared error between the sample mean of the cluster and the points inside the cluster are minimal. First, the clusters are divided into *k* clusters, the K means algorithm assigns patterns to each cluster, these patterns are based in the central point of the cluster, where, in an optimal setting, every data point belonging to a cluster must be as close as possible from the central point of the cluster it belongs (JAIN, 2010).

This algorithm performs three steps during its execution:

1. Select an initial partition with *k* clusters. Repeat steps 2 and 3 until the error rate is minimum.
2. Generate a new partition, assigning the closest value from the cluster central point.
3. Calculate new cluster central points.

In order to cluster the data, we have set a goal: to be able to fully analyze clusters containing housekeeping σ factor (σ^{70}) and alternative σ factors ($^{\sigma 24,}$ σ^{28}, σ^{32}, σ^{38} and σ^{54}). Due to the existence of two profiles of promoters, we have set a K=2, in order to stablish two profiles of each σ dependent promoter sequence. As the literature says, K=2 is the minimum value assigned to any use of K-means (BHOLOWALIA and KUMAR, 2014). These tests were performed four times, due to four physical features being presents in the simulations.

3 RESULTS AND DISCUSSION

The cluster profile analysis provides two main profiles for each σ dependent promoter sequence. The quantity of promoter sequences belonging to each one of the clusters is present in Table 3.1.

Table 3.1 – Promoter sequences distribution in clusters.

Crossing	Feature	Cluster	Quantity	Crossing	Feature	Cluster	Quantity
$\sigma^{70 \times} \sigma^{54}$	Enthalpy	C1	1090	$\sigma^{70 \times} \sigma^{54}$	Entropy	C1	1098
		C2	875			C2	867
		C1(σ^{70})	1040			C1(σ^{70})	1074
		C2(σ^{70})	842			C2(σ^{70})	808
		C1(σ^{54})	50			C1(σ^{54})	24
		C2(σ^{54})	33			C2(σ^{54})	59
$\sigma^{70 \times} \sigma^{54}$	Stability	C1	873	$\sigma^{70 \times} \sigma^{54}$	Stacking	C1	739
		C2	1092			C2	1226
		C1(σ^{70})	847			C1(σ^{70})	726
		C2(σ^{70})	1035			C2(σ^{70})	1156
		C1(σ^{54})	26			C1(σ^{54})	13
		C2(σ^{54})	56			C2(σ^{54})	40
$\sigma^{70 \times} \sigma^{38}$	Enthalpy	C1	890	$\sigma^{70 \times} \sigma^{38}$	Entropy	C1	854
		C2	1151			C2	1187
		C1(σ^{70})	802			C1(σ^{70})	772
		C2(σ^{70})	1080			C2(σ^{70})	1110
		C1(σ^{54})	88			C1(σ^{54})	82
		C2(σ^{54})	71			C2(σ^{54})	77
$\sigma^{70 \times} \sigma^{38}$	Stability	C1	885	$\sigma^{70 \times} \sigma^{38}$	Stacking	C1	796
		C2	1156			C2	1245
		C1(σ^{70})	834			C1(σ^{70})	752
		C2(σ^{70})	1048			C2(σ^{70})	1130
		C1(σ^{54})	51			C1(σ^{54})	44
		C2(σ^{54})	108			C2(σ^{54})	115
$\sigma^{70 \times} \sigma^{32}$	Enthalpy	C1	958	$\sigma^{70 \times} \sigma^{32}$	Entropy	C1	994
		C2	1209			C2	1173
		C1(σ^{70})	842			C1(σ^{70})	799
		C2(σ^{70})	1040			C2(σ^{70})	1083
		C1(σ^{54})	116			C1(σ^{54})	195
		C2(σ^{54})	169			C2(σ^{54})	90
$\sigma^{70 \times} \sigma^{32}$	Stability	C1	1022	$\sigma^{70 \times} \sigma^{32}$	Stacking	C1	1345
		C2	1145			C2	822
		C1(σ^{70})	882			C1(σ^{70})	1123
		C2(σ^{70})	1000			C2(σ^{70})	759
		C1(σ^{54})	140			C1(σ^{54})	222
		C2(σ^{54})	145			C2(σ^{54})	63
$\sigma^{70 \times} \sigma^{28}$	Enthalpy	C1	1121	$\sigma^{70 \times} \sigma^{28}$	Entropy	C1	890

		C2	890			C2	1121
		C1(σ^{70})	1040			C1(σ^{70})	795
		C2(σ^{70})	842			C2(σ^{70})	1087
		C1(σ^{54})	81			C1(σ^{54})	95
		C2(σ^{54})	48			C2(σ^{54})	34
$\sigma^{70 \times} \sigma^{28}$	Stability	C1	1048	$\sigma^{70 \times} \sigma^{28}$	Stacking	C1	747
		C2	963			C2	1264
		C1(σ^{70})	985			C1(σ^{70})	725
		C2(σ^{70})	897			C2(σ^{70})	1157
		C1(σ^{54})	63			C1(σ^{54})	22
		C2(σ^{54})	66			C2(σ^{54})	107
$\sigma^{70 \times} \sigma^{24}$	Enthalpy	C1	1079	$\sigma^{70 \times} \sigma^{24}$	Entropy	C1	1312
		C2	1308			C2	1075
		C1(σ^{70})	932			C1(σ^{70})	1214
		C2(σ^{70})	950			C2(σ^{70})	668
		C1(σ^{54})	147			C1(σ^{54})	98
		C2(σ^{54})	358			C2(σ^{54})	407
$\sigma^{70 \times} \sigma^{24}$	Stability	C1	1635	$\sigma^{70 \times} \sigma^{24}$	Stacking	C1	818
		C2	752			C2	1569
		C1(σ^{70})	1219			C1(σ^{70})	698
		C2(σ^{70})	663			C2(σ^{70})	1184
		C1(σ^{54})	416			C1(σ^{54})	120
		C2(σ^{54})	89			C2(σ^{54})	385

Table 3.1 represents the crossing values between σ^{70} and all alternative σ factors. Each simulation contains a K=2, in other words, 2 clusters. The column representing the clusters has first on the C1 and C2 attributes the number of promoters in both clusters. C1(70) and C2(70) only shows the amount of σ^{70} promoters in the cluster. Same applies for the alternative σs.

The next step, was to analyze how each feature behaves in the two clusters. To do so, we have calculated the average value per position (81 length sequences) and plotted these lines. Each cluster has its own graphic as Image 3.1 indicates, and presents few differences in terms of cluster 1 and cluster 2, this turns the clustering technique unable to catch differences and be able to fully separate the promoter sequences.

The capacity of RNAP to recognize promoter regions is a critical factor in gene transcription, in bacteria, σ^{70} is responsible for initiating the transcription of majority of genes, it is called a housekeeping factor. While σ^{24}, σ^{28}, σ^{32}, σ^{38}, σ^{54} start the transcription of genes used in specific moments of the live of the cell (PUPOV *et* al., 2013; PAYNE *et* al., 2018). In this paper, we sought to determine the profiles of physical aspects inside the clusters and related them to the action of RNAP enzyme in order to understand how RNAP is aided in terms of recognizing promoters due to their physical features.

In transcription process, bacterial RNAP – this is conserved in three domains of life, recognizes promoter DNA through a σ factor, the enzyme recognizes two conserved

regions in the promoter sequence, the -10 and -35 (KREBS, GOLDSTEIN and KILPATRICK, 2014) and initiates the mRNA production (MURAKAMI, 2016).

Figure 3.1 is presenting a variation in the physical profiles in all its segments (A, B, C, D, E, F, G, H, I, J). Around positions 47 to 55 we can see some level of variation in the lines representing all enthalpy, entropy, stacking and stability averages inside the clusters. This range perfectly matches the -10 consensual region that promoters are known to present (KREBS, GOLDSTEIN and KILPATRICK). This variation around -10 and its extended -10 (BURR *et* al., 2000) is present in both clusters 1 and 2.

The line that has presented the most linear series in Figure 3.1 is the entropy. The literature shows that entropy is a feature that connects to other elements present in the interaction DNA-enzyme such as water and minerals. The binding between DNA and RNAP raises the entropy level due to the complexity added to the system with the presence of these other elements (PRIVALOV and CRANE-ROBINSON, 2018). During the attaching of RNAP and promoter sequence to start transcription, there is a conserved adenine nucleotide in -11 position, this triggers the opening of DNA double strand, this A nucleotide is a key element to separate the strand and form an open complex. RNAP recognizes this A nucleotide through a stacking interaction from DNA and the amino acid tyrosine present in RNAP. It was expected that -10 region presented in Figure 1 and perceived by the clustering had nuances, after all, -10 and -35 are the sites where RNAP binds the DNA strand. They do present some level of degeneration and mutation, but it is the physical site where RNAP binds, so it is expected that, in terms of physical aspects the -10 indicates some differences (SCHROEDER *et* al., 2016; MURAKAMI, 2018). In terms of thermodynamics of RNAP binding, there is a highly negative enthalpy and small negative entropy during the connection RNAP – DNA. This difference in these two thermodynamics factors implies on the same nuance that Image 1 presented around -10 (BHOWRNIK, BHARDWAJ, CHATTERJI, 2017).

To enhance the importance of comprehension of gene transcription, PUPOV *et* al. (2013) have conducted an experimental test using a drug that inhibits the action of diseases. Due to a detailed understanding on how different sigma factors recognize promoter elements provide important insights towards mechanisms in transcription and can purpose new models to be used by synthetic biology. HUSSEY and MCMILLAN (2018) have developed a programmable transcription factor, where a phage is used to transcribe an *E. coli*. Orthogonal and programmable regulating elements use the state of

art in biotechnology, and this can only be achieved with a full comprehension on how transcription works in different organisms.

Figure 3.1 – Enthalpy, entropy, stacking and stability profiles for crossings.

Figure 3.1 represents the cluster crossing between alternative and housekeeping σ factors, whereas A) and B) represents clusters 1 and 2 from σ^{70} and σ^{54}; C) and D) both clusters from σ^{70} and σ^{38}; E) and F) cluster 1 and 2 from crossing between σ^{70} and σ^{32}; G) and H) the clusters resulted from σ^{70} and σ^{28} crossing and lastly; I) and J) σ^{70} and σ^{24} clusters.

4 CONCLUSIONS

Promoter sequence identification in *prokaryote* is not a simple task. The conserved regions these beings present are, somehow, not a universal rule and there are some situations where the promoter sequences is extended to the coding region (RANGANNAN and BANSAL, 2005). The bioinformatics field of study can make a good use of techniques that employ deeper analysis and provide a sound comprehension of these regulatory elements. The results concluded in this paper bring the scientific community a set of rules about how promoters behave by not assessing simple sequence similarities.

Studies from PUPOV *et* al. (2013) and HUSSEY and MCMILLAN (2018) have presented some interesting developing from new drugs that can be used from the inhibition of a gene expressed by a given σ factor, presenting new developments to the application of works and knowledge produced by mechanisms that seek to better comprehend the gene regulation. LIU *et* al. (2018) indicated an orthogonal model for biology's central dogma where, similarly as a computer software, any mechanism from synthetic biology can be reused in different systems, due to the authors, the rise of new methods to diminish nuances between specific organisms can help in compatibility problems.

The results presented in this paper at first may sound a little underwhelming due to the inability to form pure clusters. Through a deeper analysis of what is going on inside of each cluster, we were able to identify what happened. It does not matter the physical feature taken as example, it behaves in a similar way in all promoter sequences, making no distinction between σ groups that are known to be both structural, functional and biologically different. Entropy, enthalpy, base-pair stacking and stability profiles have all indicated a variation around its -10 consensual region, with no distinction between housekeeping and alternative σs. However, no differences were spotted linking the -35

region to physical aspects of DNA molecules, further studies need to be developed in these terms.

Our literature review has indicated that in the whereabouts of -10, RNAP binding to this specific site, where conservation can be found in a level of conservation that is significant to be distinguished from other genomic regions. The cluster outcomes have shown this exact point around -10, where there is variation in levels for all of the tested physical features of the DNA. A deeper understanding on how promoters are recognized by specific σ factors bring the scientific community a better elucidation of transcription factors exists in all living systems, any mutation that occurs during the transcription may lead to consequences of deletion or overexpression of transcription factors that may jeopardize the survivability of a cell.

5 REFERENCES

ABEEL, T.; PEER, Y. Van de; SAEYS, Y. Toward a gold standard for promoter prediction evaluation. **Bioinformatics**, v. 25, n. 12, p. i313–i320, 2009.

BARRIOS, H., VALDERRAMA, B., MORETT, E., Compilation and analysis of σ54-dependent promoter sequences. **Nucleic Acids Research**, vol. 27, n. 22, p. 4305-4313. 1999.

CONNEL, S.D., JAIN, A.K., 2002. Writer adaptation for online handwriting recognition. **IEEE Trans. Pattern Anal. Machine Intell**. 24 (3), 329–346.

DAVIES, P. C. W., RIEPER, E., TUSZYNSKI, J. A.; Self-organization and entropy reduction in a living cell. **National Institute of Health**. 2013 January ; 111(1): 1–10**.**

Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A, Hernández-Koutoucheva A, Moral-Chávez VD, Rinaldi F, Collado-Vides. J. **RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond.**

Nucleic Acids Res. 2016 Jan 4;44(D1):D133-43. doi: 10.1093/nar/gkv1156. Epub 2015 Nov 2.

JAIN, A. K.; Data clustering: 50 years beyond K-means. **Pattern Recognition Letters.** 31. 2010 pgs. 651-666.

P. Juszczak, D. Tax, R. P. W. Duin, E. Deprettere, A. Belloum, J. Heijnsdijk, F. van der Stappen, "Feature scaling in support vector data description", *ASCI 2002 8th Annual Conf. of the Advanced School for Computing and Imaging*, pp. 95-102, 2002.

KANHERE, A. and BANSAL, M. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes**. Nucleic Acids Res**., **33**, 3165–3175, 2005.

KREBS, J. E.; GOLDSTEIN, E. S., KILPATRICK, S. T. Lewin's GENES XI, Jones & Bartlett Publ., 2014.

Meysman P, Collado-Vides J, Morett E, Viola R, Engelen K, et al. (2014) Structural Properties of Prokaryotic Promoter Regions Correlate with Functional Features. PLoS ONE 9(2): e88717. doi:10.1371/journal.pone.0088717

MCADAMS, H. H., SRINIVASAN, B., ARKIN, A. P. The evolution of genetic regulatory systems in bacteria. **Nature Rev. Genet.** 5, 169–178 (2004).

MORGUNOVA E., YIN Y., DAS P.K., Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. Ben-Tal N, ed. *eLife*. 2018;7:e32963. doi:10.7554/eLife.32963.

Murakami, K. S. (2015). Structural Biology of Bacterial RNA Polymerase. *Biomolecules*, *5*(2), 848–864. http://doi.org/10.3390/biom5020848

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, 17: 2341-2360. doi:10.1002/bip.1978.360171005

PAYNE, S. R., PAU, D. I., WHITING, A. L., KIM, Y. J., PHAROAH, B. M., MOI, C., BODDY, C. N., BERNAL, F., 2018. Inhibition of bacterial gene transcription with an RpoN-base stapled peptide. **Cell Chemical Biology** 25, 1–8

PRIVALOV, P. L., CRANE-ROBINSON, C., Forces maintaining the DNA double helix and its complexes with transcription factors. **Progress in Biophysics and molecular biology**. 135. p. 30-48. 2018.

PUPOV, D., ESYUNINA, D., FEKLISTOV, A., KULBACHINSKIY. A., (2013) Single-stranded promoter traps for bacterial RNA polymerase. **Biochem J.** 452, 241-248.

SANDERS, M. F., BOWMAN, J. L. Análise genética – uma abordagem integrada. Pearson. 2014.

SANTALUCIA, J., Jr.; HICKS, D.; The thermodynamics of DNA structural motifs. **Annu. Rev. Biophys. Biomol. Struct** 33. p. 415-440, 2004.

XU X, ZHI X, LENG F. Determining DNA Supercoiling Enthalpy by Isothermal Titration Calorimetry. *Biochimie*. 2012;94(12):2665-2672. doi:10.1016/j.biochi.2012.08.002.

5.3 – AN ASSESSMENT REGARDING INTERSECTED PROMOTERS SEQUENCES RECOGNIZED BY DIFFERENT σ FACTORS IN THE MOST POPULATED ENTHALPY, BASE-PAIR STACKING AND STABILITY CLUSTERS

The K value to perform the clustering technique in this section has been presented in section 4, Table 4.2, the next step was to use the clustering tool in the promoter sequences.

Table 5.1 – Clustering outcomes.

Sigma	Feature	Cluster	Quantity	Sigma	Feature	Cluster	Quantity
24	Enthalpy	1	32%	24	Entropy	1	18%
		2	24%			2	10%
		3	34%			3	13%
		4	10%			4	1%
		5	-			5	12%
		6	-			6	6%
		7	-			7	40%
	Stacking	1	28%		Stability	1	26%
		2	14%			2	17%
		3	3%			3	11%
		4	19%			4	10%
		5	12%			5	11%
		6	13%			6	13%
		7	11%			7	12%
28	Enthalpy	1	31%	28	Entropy	1	11%
		2	27%			2	39%
		3	42%			3	50%
	Stacking	1	50%		Stability	1	31%
		2	37%			2	39%
		3	13%			3	30%
32	Enthalpy	1	29%	32	Entropy	1	17%
		2	4%			2	41%
		3	17%			3	29%
		4	33%			4	12%
		5	17%			5	-
	Stacking	1	33%		Stability	1	22%
		2	17%			2	12%
		3	18%			3	28%
		4	7%			4	25%
		5	25%			5	13%
38	Enthalpy	1	69%	38	Entropy	1	46%
		2	27%			2	54%
		3	4%			3	-
	Stacking	1	30%		Stability	1	33%
		2	25%			2	32%
		3	45%			3	45%

54	Enthalpy	1	99%	54	Entropy	1	72%
		2	1%			2	28%
	Stacking	1	26%		Stability	1	47%
		2	76%			2	53%
70	Enthalpy	1	62%	70	Entropy	1	40%
		2	8%			2	31%
		3	30%			3	29%
	Stacking	1	30%		Stability	1	39%
		2	41%			2	30%
		3	29%			3	31%

5.3.1 INTERSECTION ANALYSIS OF THE MOST POPULATED CLUSTERS IN ORDER TO VERIFY SIMILARITIES

Checking the conclusions carried by Article 1, it can be perceived how intricately enthalpy, base-pair stacking and stability relate between themselves. These 3 are sequence-dependent features (KANHERE and BANSAL, 2005; XU, ZHI and LANG, 2012; PRIVALOV and ROBINSON, 2018). The only feature that is system-dependent is enthalpy (XU, ZHI and LANG, 2012; MORGUNOVA *et* al., 2018), this means that it would be a complicated task to evaluate the entropy as a standalone feature, in other words, it would jeopardize the cluster intersection analysis to include entropy values because it depends on other components that are involved in the interaction between the DNA strand, the RNAP enzyme such as water molecules and some minerals. With this conclusion, it is worth checking how these three features interact with each other in terms of clustering.

The intersection analysis sought to consider separated clusters of a physical aspect. The most populated cluster from each physical feature was selected and all its content was checked. If an *n* promoter is present in the most populated cluster from each feature, it can be said that this promoter sequence presents a profile that is worth to be checked, distinguishing this specific promoter from others. When compared to the raw number of promoters depicted in Materials and Methods Table 4.1, the number of intersected sequences is portrayed in Figure 5.1: σ^{24} = 2.7%, σ^{28} =4.5%, σ^{32} = 0%, σ^{38} = 12.7%, σ^{54} = 26.5% and σ^{70} = 26%. These numbers, especially for σ^{24}, σ^{28} and σ^{32} may sound low. The goal was not to have a large number of promoter sequences, rather than that, we sought to promote a better understanding in these particular sequences acting as promoter characterizer and analyze how they behave in terms of their physical aspects. Once the most populated clusters from each physical feature has been found, the next step

was to promote an assessment of how these DNA aspects interact with each other. Figure 5.2 depicts the lines representing the 81-length sequence normalized average in enthalpy, base-pair stacking and stability. These three features were selected due to the results presented by Article 5.1, where entropy was not taken as a standalone measurement of promoter sequences due to its high link with the whole system involved the gene transcription (PRIVALOV and CRANE-ROBINSON, 2018; XU, ZHI and LANG, 2012; MORGUNOVA *et* al., 2018).

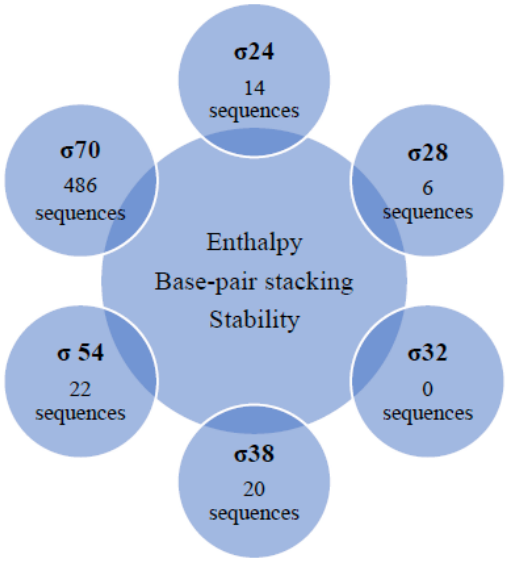


Figure 5.1 – Intersection representation

Figure 5.1 (A), refers to σ^{24} promoter sequences, and in a similar way as Article 5.1 had shown, this is the tested σ that presented noisier results and few inferences can be taken from that. The profile that was reproduced by this σ factor does not represent a solid profile due to the nature of these promoter sequences found in RegulonDB (GAMA-CASTRO *et* al., 2015) where more than 80% of the examples from σ^{24} presented as evidence that supports the existence of the promoter sequence as inferred computationally without human inference. This means that the consensus motifs that these promoter sequences have are not as accurate as other σ factors (SHIMADA, TANAKA and ISHIHAMA, 2017).

In what concerns the others σ groups that presented intersections in terms of promoter sequences in the most populated enthalpy, base-pair stacking and stability clusters we can perceive a high correlation of the data, in a similar way that Article 1 did, considering all the promoter sequences in RegulonDB (GAMA-CASTRO *et* al., 2015).

The Figure 5.2 did not include any graph in terms of σ^{32}, as Figure 5.1 has shown, no intersections were found in promoter sequences recognized by this σ factor. The other profiles presented in Figure 2, (B) σ^{28}, (C) σ^{38}, (D) σ^{54} and (E) σ^{70} have indicated a closer correlation between the physical features in a similar way that Article 5.1 did. In these four σ groups, the -10 consensual motif can be found due to a difference in terms of how physical features behave in these regions. σ^{28} indicates a high peak around -10 followed by a drop upstream; another peak in σ28 series was found in an extended -10 region, around -18 (BARRIOS, VALDERRAMA and MORRET, 1999). σ^{38} indicated a low average on the values in -10 and a sudden rise around -35, matching the most common conserved consensual regions. σ^{54} lines are entirely overlapping, the higher correlation between these values are in enthalpy and stacking series, stability matches these other two features in an extended -10 region (citation). Ultimately, σ^{70}, in a similar way Article 5.1 did, presented a high conservation and similarity between the three features around -10.

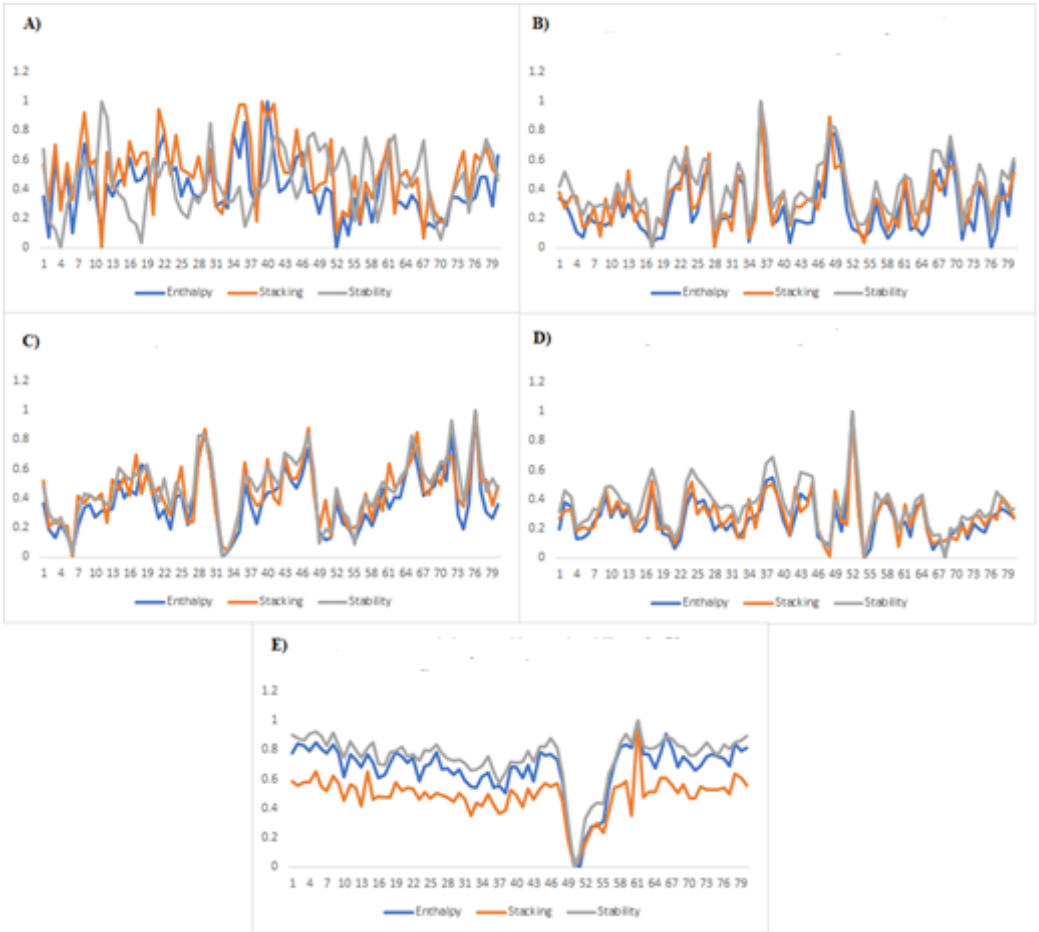


Figure 5.2 – Enthalpy, stacking and stability profiles in intersected promoter sequences recognized by different σ factors.

6 CONCLUSIONS

The results that were presented on this paper has led us to some important remarks regarding the physical features in promoter sequences. The literature asserts (RANGANNAN and BANSAL, 2009) that promoter sequences *in silico* identification and classification in eukaryotes might not be considered as a simple task, there are some genes which their promoter is located not only in the intergenic regions, but stretching to coding regions, which leads to artificial intelligence techniques usage. When properties regarding the physical set up from promoters is added to the *in silico* identification, the tools that make use of this knowledge may be improved through a substantial analysis.

Section 5 has indicated that in specific RNAP binding locations, mainly around the -10 and -35 motifs, the way that enthalpy, base-pair stacking and stability behave is different. These two sites should present higher conservation levels in terms of their sequence pattern (KREBS, GOLDSTEIN and KILPATRICK, 2014) due to RNAP action (ABEEL *et* al., 2009). But again, in biology sciences, no rule can be taken for granted, and exception may surge when less expected (KOONIN, 2012). This uncertainty permeating biology severely impacts *in silico* identification, so a tool that only seeks for sequence patterns to identify promoter sequences may become obsolete in no time at all (RANGANNAN and BANSAL, 2009).

It is believed that, the addition of physical aspects of the DNA to promoter recognition brings to the bioinformatics field a grand plunder. In the tests performed to conclude this work, it was clear that around the consensual motifs, promoter sequences associated with RNAp σ^{28}, σ^{32}, σ^{38}, σ^{54} and σ^{70} presented a significant difference in terms of their enthalpy, base-pair stacking and stability values. The only feature that was concluded to not being a standalone measurement was entropy, where, it should be analyzed as a system characteristic, and when it is assayed alone, it clarifies the entropy series presenting no significant variation shown in Article 5.1 Figure 1.

As the literature has brought, the comprehension of how cellular elements interact between themselves during the transcription process may aid scientific community to develop new drugs that seek to inhibit the activity of genes related to a specific σ factor, and when orthogonal models are added, there can be seen a way to utilize knowledge that had previously been used in a specific model in order to produce new inferences in other organisms.

A limitation to this study is the focus that was given to promoter sequences, in future endeavors, considering what was here concluded, it might be worth it checking how the four physical features behave inside and outside promoter sequences. This will extend the results gathered and enable an enhanced understanding in the role that the physics of the DNA play in the transcription processes. Another remark that could be analyzed through this genomic-wise analysis of the DNA physics would be the entropy variation, where, when disregarded with the RNAp action, the entropy levels could be different, being able to distinguish entropy from promoters and non-promoter sequences.

7 REFERENCES

ABEEL, T.; PEER, Y. Van de; SAEYS, Y. Toward a gold standard for promoter prediction evaluation. **Bioinformatics**, v. 25, n. 12, p. i313–i320, 2009.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., LIMPMAN, D. J.; Basic local alignment search tool. **Journal of Molecular Biology.** 1990. Vol. 215(3). p. 403 – 410.

ATKINS, P.W., DE PAULA, J., Físico-química. 9 ed.: LTC Rio de Janeiro, RJ. 411p. 2012.

BABU. M. M., Bacterial Gene Regulation and Transcriptional Networks 1 ed.: Caister Academic Press. Cambridge UK. 2013. 277p.

BARRIOS, H., VALDERRAMA, B., MORETT, E., Compilation and analysis of σ54-dependent promoter sequences. **Nucleic Acids Research**, vol. 27, n. 22, p. 4305-4313. 1999.

BHOLOWALIA, P., KUMAR, A., EBK-Means: a clustering technique based on Elbow Method and K-Means in WSN. **International Journal of Computer Applications**. 105. p. 17-24. 2014.

BROWNING D. F., BUSBY S. J., Local and global regulation of transcription initiation in bacteria. **Nat rev microbiol.** 2016. V. 16(10). P. 638-650.

BURGESS, R. R., ANTHONY, L. How sigma docks to RNA polymerase and what sigma does. **Curr Opin Microbiol** 4 (2001) 126 –131.

BURR, T., MITCHELL, J., KOLB, A., MINCHIN, S., BUSBY, S., DNA sequence elements located immediately upstream of the -10 hexamer in *Escherichia coli* promoters: a systematic study. **Nucleic Acids Research**, vol. 28. n. 9. p. 1864-1870. 2000.

CASES, I., DE LORENZO, V. & OUZOUNIS,C. A. Transcription regulation and environmental adaptation in bacteria. **Trends Microbiol.** 11, 248–253 (2003).

CHRISTIAN, B., O humano mais humano: o que a inteligência artificial nos ensina sobre a vida. 1 ed. Companhia das letras SP. 368 p. 2013.

COOK, V. M., DEHASET, P. L., Strand opening-deficient Escherichia coli RNA polymerase facilitates investigation of closed complexes with promoter DNA: effects of DNA sequence and temperature. **J Biol Chem** 282 (2007) 21319 –21326.

CONNEL, S.D., JAIN, A.K., 2002. Writer adaptation for online handwriting recognition. **IEEE Trans. Pattern Anal. Machine Intell**. 24 (3), 329–346.

CROOKS, G. E., HON, G., CHANDONIA, J.M., BRENNER, S.E., WebLogo: A sequence logo generator, **Genome Research,** 14:1188-1190, (2004)

DAVIES, P. C. W., RIEPER, E., TUSZYNSKI, J. A.; Self-organization and entropy reduction in a living cell. **National Institute of Health**. 2013 January; 111(1): 1–10**.**

DE AVILA e SILVA, S.; ECHEVERRIGARAY, S.; GERHARDT, G. J. L. BacPP: Bacterial promoter prediction - A tool for accurate sigma-factor specific assignment in enterobacteria. **Journal of Theoretical Biology**, v.287, p.92-99, 2011.

DE AVILA E SILVA, S., ECHEVERRIGARAY, S., Bacterial Promoter Features Description and Their Application on E. coli in silico Prediction and Recognition Approaches. **IntertechOpen**. 2012.

DE LIMA, A. A., Físico-química. 1ed. São Paulo: Pearson do Brasil, 2014.

DE ROBERTIS, E. M.; Biologia Molecular e Celular. 14 ed. 413 p. Guanabara Koogan. Rio de Janeiro. 2003.

DUBES, R., JAIN, A. K.; Clustering techniques: the user's dilemma. **Pattern recognition: Pergamon Press.** 1976. Vol. 8. Pgs. 247-260.

EDGAR, R. C.; Local homology recognition and distance measures in linear time using compressed amino acid alphabets. **Nucleic Acids Research.** 2004. V. 32(1). p. 380-385

EDGAR, R.C.; Search and clustering orders of magnitudes faster than BLAST. **Bioinformatics.** 2010. Vol. 26(19). p. 2460-2461.

EZER, D., ZABET, N., R., ADRYAN, B., Physical constraints determine the logic of bacterial promoter architecture. **Nucleic Acids Research**, vol. 42, n. 7. p. 4196-4207. 2014.

Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A, Hernández-Koutoucheva A, Moral-Chávez VD, Rinaldi F, Collado-Vides. J. RegulonDB version 9.0: high-level

integration of gene regulation, coexpression, motif clustering and beyond. **Nucleic Acids Res.** 2016 Jan 4;44(D1):D133-43. doi: 10.1093/nar/gkv1156. Epub 2015 Nov 2.

GASPAR A., Física: Ondas, Ótica, Termodinâmica. 1 ed. Ática: São Paulo – SP. 368 p. 2000

HALLIDAY D., RESNICK R., WALKER J. Fundamentos de Física. Gravitação, ondas e termodinâmica. 7 ed. Rio de Janeiro– RJ. 324 p. 2006.

HAUGUEN, S., P., ROSS, W., GOURSE, R. L., Advances in bacterial promoter recognition and its control by factors that do not bind DNA. **Nat rev microbiol.** 6(7) 507-519. 2008.

HASE F. E ZACHARIAS M. (2016) Free energy analysis and mechanism of base pair stacking in nicked DNA. **Nucleic Acid Research**. Vol. 45(15) p. 7100-7108.

HERZEL, H., EBELING, W., SCHIMITT, A. O. Entropies of biosequences-the role of the repeats. **Phys. Rev.** 1994. v. 50, pgs. 5061-5071.

JAIN, A. K.; Data clustering: 50 years beyond K-means. **Pattern Recognition Letters.** 31. 2010 pgs. 651-666.

JAIN, A. K., DUBES, R. C. Algorithms for clustering data. Englewood Cliffs: Prentice Hall. 304 p. 1988.

P. Juszczak, D. Tax, R. P. W. Duin, E. Deprettere, A. Belloum, J. Heijnsdijk, F. van der Stappen, "Feature scaling in support vector data description", *ASCI 2002 8th Annual Conf. of the Advanced School for Computing and Imaging*, pp. 95-102, 2002.

KANHERE, A. and BANSAL, M. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes**.** **Nucleic Acids Res**., **33**, 3165–3175, 2005.

KAPANIDIS, A. N., MARGEAT, E., HO, S. O., KORTKONJIA, E., WEISS, S., EBRIGHT, R. H., Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. **Science** 314 (2006) 1144 –1147.

KIM, T. H.; BARRERA, L. O.; ZHENG, M.; QU, C.; SINGER, M. A.; RICHMOND, T. A.; WU, Y.; GREEN, R. D.; REN, B. A high-resolution map of active promoters in the human genome. **Nature**, v. 436, n. 7052, p. 876–880, ago. 2005.

KOONIN, E., V.; Does the central dogma still stand? **Biology direct** 2012. p. 7-27.

KOZOBAY-AVRAHAM,L.; HOSID, S.; VOLKOVICH, Z.; BOLSHOY, A. Prokaryote Clustering based on DNA curvature distributions. **Discrete Applied Mathematics,** v.11, p.2378-2387**,** 2008.

KREBS, J. E.; GOLDSTEIN, E. S., KILPATRICK, S. T. Lewin's GENES XI, Jones & Bartlett Publ., 2014.

LEWIN, B. **Genes IX**. Porto Alegre: Artes Medicas, 2008. 955 p.

LI, W., GODZIK, A.; Cd-hit: a fast program for clustering and comparing large set of protein or nucleotide sequences. **Bioinformatics.** 2006. Vol. 22(13). p. 1658-1659.

LIU, C. C., JEWETT, M. C., CHIN, J. W., VOIGT, C, A., Toward an orthogonal central dogma. **Nature chemical biology**. Vol. 14 (2018) p. 103-106.

JAIN, A. K.; Data clustering: 50 years beyond K-means. **Pattern Recognition Letters.** 31. 2010 pgs. 651-666.

LAUKENS, K., Structural properties of prokaryotic promoter regions correlate with functional structures. **PLOS ONE**. v. 9 n. 2. 2014.

LI, D., DU, Y., Artificial Intelligence with Uncertainity. 2 ed: CRC Press (Boca Raton, FL, EUA) 289 pgs.

MA, C., YANG, X., LEWIS, P. J., Bacterial Transcription as a Target for Antibacterial Drug Development. **Microbiology and Molecular Biology Reviews** 80(1) (2016) 39-80.

MAIA, D. J., BIANCHI, J. C. de A. Química geral: fundamentos. São Paulo: Pearson Prentice Hall, 448 p. 2007.

MARMUR, J., DOTY, P., Determination of the base composition of deoxyribonucleic acid from its thermal melting temperature. **J. Mol. Biol**. 5 p.109-118. 1962.

MCADAMS, H. H., SRINIVASAN, B., ARKIN, A. P. The evolution of genetic regulatory systems in bacteria. **Nature Rev. Genet.** 5, 169–178 (2004).

Meysman P, Collado-Vides J, Morett E, Viola R, Engelen K, et al. (2014) Structural Properties of Prokaryotic Promoter Regions Correlate with Functional Features. PLoS ONE 9(2): e88717. doi:10.1371/journal.pone.0088717

Morgunova E, Yin Y, Das PK, et al. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. Ben-Tal N, ed. *eLife*. 2018;7:e32963. doi:10.7554/eLife.32963.

MULLIGAN, P. J., CHEN, Y., PHILIPS, R., SPAKOWITZ, A. J., Interlay of protein binding interactions, DNA mechanics and entropy in DNA looping kinetics. **Biophysics journal**. 109, p. 618-629. 2015.

MURAKAMI, K. S., DARST, S. A., Bacterial RNA polymerases: the whole story. **Curr Opin Struct Biol** 13 (2003) 31-39.

OLIVARES-ZAVALETA, N, JÁURENGUI, R, MERINO E; Genome analysis of Escherichia coli promoter sequence evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. **Genomics** vol. 87 (3). Pgs. 329-337. 2006.

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, 17: 2341-2360. doi:10.1002/bip.1978.360171005

PRIVALOV, P. L., CRANE-ROBINSON, C., Forces maintaining the DNA double helix and its complexes with transcription factors. **Progress in Biophysics and molecular biology**. 135. p. 30-48. 2018.

RAMPRAKASH, J., SCWARZ, F. P. Identification and annotation of promoters regions in microbial genome sequences on the basis of DNA stability. **Journal of Biosciences** 32: p. 851-862, 2007.

RANGANNAN, V., BANSAL, M., Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. **J. Biosc.** p. 851-862 v.32 n. 5. Agosto 2007.

REECE, J. B., URRY, L. A., CAIN, M. L., WASSERMAN, S. A., MINORSKY, P. V., JACKSON, R. B., Campbell Biology 10th ed. Pearson USA. 2014. 1488p.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. 2.ed. New Jersey: Prentice Hall International. 1080 p. 2003.

RYASIK, A., ORLOV, M., ZYKOVA, E., ERMAK, T., SOROKIN, A., Bacterial promoter prediction: Selection of dynamic and static physical properties of DNA for reliable sequence classification. **Journal of Bioinformatics and Computational Biology**. Vol. 16 (1). 16 p. 2018.

SANDERS, M. F., BOWMAN, J. L. Análise genética – uma abordagem integrada. Pearson. 2014.

SANTALUCIA, J., Jr.; HICKS, D.; The thermodynamics of DNA structural motifs. **Annu. Rev. Biophys. Biomol. Struct** 33. p. 415-440, 2004.

SATTIN, B. D., PELLING, A. E., GOH, M. C., DNA base pair resolution by single molecule force spectroscopy. **Nucleic acids research** 32, p. 4876-4883, 2004.

SZKLARCZYK et al. Nucleic Acids Res. 2015 43(Database issue):D447-52

SKORDALAKES, E., BERGER, J. M., Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. **Cell** 114 (2003)135–146.

The Merriam-Webster dictionary. (2016). Springfield, MA: Merriam-Webster, Incorporated.

WANG, H.; BENHAM, C. J. (2006). Promoter prediction and annotation of microbial genomes based on DNA sequences and structural responses to superhelical stress. *BMC Bioinformatics*, Vol. 7:248.

WANG, H., BENHAM, C. J., NOORDEWIER, M., Stress induced DNA duplex destabilization (SIDD) in the E. coli. genome: SIDD sites are closely associated with promoters. **Genome res.** 14(8), 1575-1584, 2014.

WATSON, J. D., & CRICK, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

Xu X, Zhi X, Leng F. Determining DNA Supercoiling Enthalpy by Isothermal Titration Calorimetry. *Biochimie*. 2012;94(12):2665-2672. doi:10.1016/j.biochi.2012.08.002.

YARNELL, W. S., ROBERTS, J. W., Mechanism of intrinsic transcription termination and antitermination. **Science** 284 (1999) 611–615.

YOUNG H. D., FREEDMAN R. A. Física II: Termodinâmica e ondas. 14 ed. Pearson do Brasil: São Paulo – SP. 352 p. 2016.

WERNER, F., GROHMANN, D., Evolution of multisubunit RNA polymerase in the three domains of life. **Nature reviews microbiology.** v. 9, p. 85-98. 2011.

ZHANG, T., ZHANG, C., DONG, Z., GUAN, Y. (2015). Determination of Base Binding Strength and Base Stacking Interaction of DNA Duplex Using Atomic Force Microscope. *Scientific Reports*, *5*, 9143. http://doi.org/10.1038/srep09143