



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE

**DOTTORATO DI RICERCA IN  
FISICA E ASTRONOMIA**

CICLO XXXII

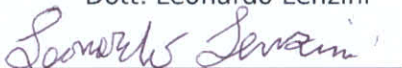
COORDINATORE Prof. Raffaello D'Alessandro

**The structure of bacterial intergenic sequences:  
relation with regulation and denaturation**

Settore Scientifico Disciplinare FIS/02

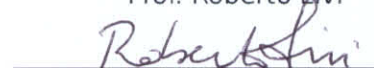
**Dottorando**

Dott. Leonardo Lenzi

  
(firma)

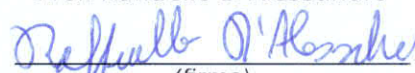
**Tutore**

Prof. Roberto Livi

  
(firma)

**Coordinatore**

Prof. Raffaello D'Alessandro

  
(firma)

Anni 2016/2019

*A mio nonno*

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Motivations and structure of the thesis . . . . .	4
1.2	Technical remarks about coding and noncoding DNA . . . . .	5
<b>2</b>	<b>Identification and clustering of intergenic sequences</b>	<b>8</b>
2.1	Identification of intergenic sequences . . . . .	8
2.1.1	Shine-Dalgarno sequences . . . . .	11
2.2	Spectral Clustering . . . . .	12
2.2.1	Sequences alignment . . . . .	13
2.2.2	The normalized Laplacian matrix . . . . .	15
2.2.3	Clustering algorithm . . . . .	17
2.2.4	Silhouette . . . . .	18
2.3	Base composition analysis . . . . .	20
2.4	Other bacterial species . . . . .	22
2.5	Resume of the method . . . . .	32
<b>3</b>	<b>Shared compositional and functional features of clustered intergenic sequences</b>	<b>35</b>
3.1	Results for <i>E. coli</i> . . . . .	37
3.1.1	Structural features of clusters . . . . .	37
3.1.2	Correlations between clustering and biological features . . . . .	39
3.2	Other bacterial species . . . . .	41
3.3	Materials and Methods . . . . .	45
3.3.1	About STRING . . . . .	45
3.3.2	COG categories enrichment . . . . .	47
3.4	Comparison with other methods identifying biological functions of IGSs . .	48

<b>4</b>	<b>Thermodynamics of DNA denaturation in a model of bacterial inter- genic sequences</b>	<b>50</b>
4.1	The Model of bacterial IGSs . . . . .	52
4.2	The denaturation dynamics . . . . .	55
4.2.1	Denaturation for IGSs of <i>E. coli</i> . . . . .	59
4.2.2	Other bacterial species . . . . .	65
4.3	Final remarks . . . . .	68
<b>5</b>	<b>Features of BCA in eukaryotes for coding and noncoding regions</b>	<b>69</b>
5.1	IGSs in eukaryotes . . . . .	70
5.1.1	Constrains in eukaryotic BCA . . . . .	72
5.1.2	Reproducing the spatial distribution of weak and strong bases for the BCA of eukaryotic IGSs . . . . .	76
5.2	BCA of genes in <i>H. sapiens</i> and <i>E. coli</i> . . . . .	83
<b>6</b>	<b>Conclusions</b>	<b>89</b>
<b>7</b>	<b>Ringraziamenti</b>	<b>91</b>



# Chapter 1

## Introduction

The work described in this physics Ph.D. thesis consists in applying methods and techniques of statistical mechanics to genetic problems, related to the study of DNA sequences. In this introductory chapter we present the philosophy that guided the development of this research, and afterwards we include a short technical introduction.

### 1.1 Motivations and structure of the thesis

The research problems tackled in this thesis concern four different items to which we devote separate chapters. For what concerns the first two chapters, our work has taken inspiration from a previous contribution [1] that focused on the study of noncoding (promoter) sequences in eukaryotes. Through a collaboration with the genetics group of Professor Renato Fani we have moved the object of our research towards noncoding sequences in some bacterial species. The first goal has been to define a method able to uniquely identify noncoding DNA sequences in proximity of genes in bacteria, similar to promoters in eukaryotes; the method is described in chapter 2. More precisely, following the approach proposed in [1], we have clustered these DNA sequences on the basis of structural similarities. Making use of updated databases, we have found a correspondence between structural features and biological properties. This analysis is illustrated in chapter 3 and has produced the following publication [2]: Lenzini L, Di Patti F, Livi R, Fondi M, Fani R, Mengoni A. A Method for the Structure-Based, Genome-Wide Analysis of Bacterial Intergenic Sequences Identifies Shared Compositional and Functional Features. *Genes*. 2019; 10(10):834.

The next goal has been to study the sequences from a point of view of the thermodynamics of denaturation taking into account most of the above identified noncoding

bacterial DNA sequences. In fact, taking advantage of the availability of real noncoding sequences, we have investigated the correspondence between structural and thermodynamic properties. This part of the research is described in chapter 4. The results of this study were reported in the following paper [3]: Leonardo Lenzini, Francesca Di Patti, Stefano Lepri, Roberto Livi, and Stefano Luccioli. Thermodynamics of dna denaturation in a model of bacterial intergenic sequences. *Chaos, Solitons & Fractals*, 130:109446, 2020.

The last part of this work, contained in chapter 5, concerns the reexamination of the intergenic eukaryotic sequences previously studied in [1]. In particular we have analyzed compositional features of different eukaryotic species along the phylogenetic tree. This systematic investigation has yielded the empirical observations about the presence of structural constraints characterizing such sequences. By this analysis it has emerged a correlation between evolutionary trends and compositional structures of noncoding sequences in analogy with what has been observed in coding components. Despite its interest and novelty, a clear interpretation about the presence of these constraints is still lacking. Certainly further investigations are necessary to possibly reach the goal of a convincing biological interpretation. In this perspective these results can be viewed as a preliminary step still insufficient for producing a publication.

## **1.2 Technical remarks about coding and noncoding DNA**

The four basic "bricks" of DNA are the nitrogenous bases called adenine (A), thymine (T), guanine (G) and cytosine (C). They are divided into purines (A and G) and pyrimidines (T and C) on the base of the number of heterocyclic rings that compose them (two for the first one, one for the second one). Each nitrogenous base can bind to a molecule of pentose sugar (deoxyribose) and to a phosphate group thus obtaining a nucleotide. A DNA molecule consists of two complementary strands, each of which is a polymer composed of the four nucleotides, connected to each other by forming covalent bonds between the phosphate group and the sugar molecule. The two filaments wrap around each other forming the well-known double helix structure thanks to the creation of bonds (i.e. the well-known hydrogen bonds) between the nitrogenous bases, with a very specific coupling rule: A mates with T forming two hydrogen bonds, while G with C forming three (see Figure 1.1). For this reason the former are called weak nucleotides, while the

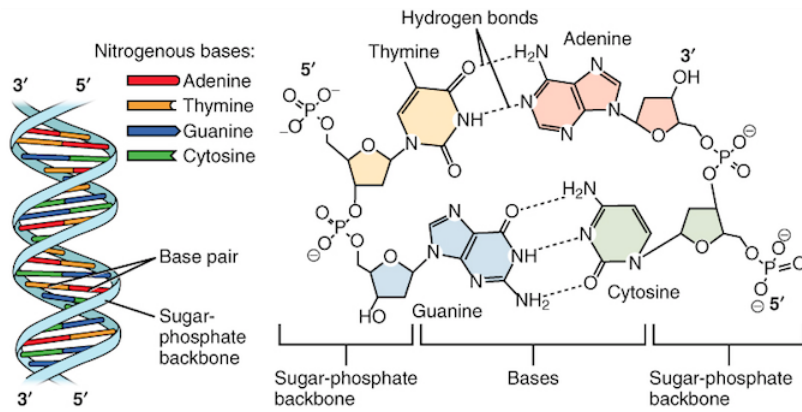


Figure 1.1: Structure of DNA.

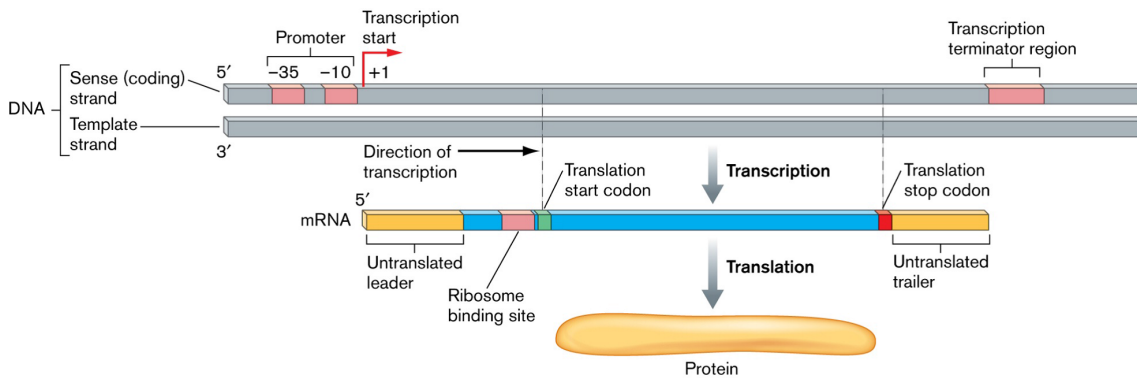


Figure 1.2: Scheme of protein synthesis in bacteria.

latter strong [4, 5].

DNA stores the biological information of an organism, its main role is to encode the information concerning the structure of proteins in sequences called genes. The gene sequence is first read and transcribed by an enzyme (RNA polymerase) in an RNA strand, which is translated into proteins by ribosomes according to the genetic code: each sequence of three nucleotides, called codon, is converted into an amino acid which will form the protein polymer (see Figure 1.2). This code is “universal”, i.e. it is the same for all species [6]. However there is a lot of information encoded in DNA besides genes. Consider that in many species genes account for only a small fraction of the genome: DNA is not a mere sequence of genes, there is a large part of noncoding DNA where instructions for the correct use of information expressed by genes are contained, a kind of “dark matter” whose role is still on debate [7].

Many regulatory mechanisms act at various stages of protein synthesis, we will be interested in regulation in the transcription phase. The element of DNA that plays a key role in this phase is called promoter, a DNA sequence located upstream the transcription start site (TSS) of a gene. In eukaryotes inside promoters we find short sequences,

called binding sites, whose task is binding specific proteins called Transcription Factors. Through complex biochemical mechanisms, transcription factors control the activity of the RNA polymerase, that transcribes the gene. In bacterial transcription RNA polymerase is bound to another protein, the sigma factor, responsible for anchoring the RNA polymerase to DNA by recognizing specific binding sites within the promoter, called consensus sequences. The promoters, by means of interaction with the sigma factor, are able to inhibit or intensify the transcription and expression of the corresponding gene.

Our aim is to obtain information about the general properties of the noncoding region located upstream a gene and their structure, instead of investigating the few base pairs that make up the binding sites. We don't limit our analysis to the study of the promoters itself but we will focus on a larger noncoding region. We therefore propose to find common features among these regions to find a relationship between the structure of their sequences and their biological function, a correlation between the composition of their groups and the genes they regulate. Note that gene regulation mechanisms may also have a key role in evolution. One could conjecture that the main mechanism for evolution of later, more complex organisms is the expansion and modification of gene families. Nevertheless, striking similarities in gene content have been observed since the beginning of genomic studies in the comparison of different species: see, for instance, [8], where chimpanzee and human biological differences are hypothesized to be due to regulatory mutations, since their genes and proteins are almost identical. This has been observed also for different but related phyla, that can share nearly identical sets of Hox genes (i.e. developmental genes controlling morphogenesis), despite their great morphological diversity and the long span of time since their divergence from a common ancestor [9]. There is evidence that morphological changes in animals have been shaped by evolutionary changes in developmental gene regulation, and not in genes themselves [9, 10, 11]. In other words, the differences at protein level between different species can be minimal; the different features we observe in different species are mostly due to how and when such proteins are produced and used in an organism.

# Chapter 2

## Identification and clustering of intergenic sequences

In this chapter we used *Escherichia coli* as a case study for illustrating the identification procedure of intergenic sequences (IGSs), that are expected to be correlated with regulation and expression of genes. Hereinafter we term IGS a noncoding portion of DNA with the peculiar features that we are going to describe in the following sections. The database of *E. coli* genome used in this thesis has been downloaded from NCBI (National Center for Biotechnology Information) [12], a part of the United States National Library of Medicine (NLM), the branch of the National Institutes of Health (NIH), that houses GenBank sequence database [13], an open access containing all publicly available nucleotide sequences with annotated TSC (Translation Start Codon) of all genes.

### 2.1 Identification of intergenic sequences

In analogy with what studied in the previous papers concerning the study of structural features of promoters in eukaryotes [1], we expect that noncoding regions of bacterial DNA close to the TSS are correlated with regulation/expression of genes. In what follows we call these regions intergenic sequences (IGSs). In order to identify IGSs we have to introduce first the concept of intergenic region (IGR): this region extends between the TSC of one gene and the end of the previous coding region on the same strand (see Figure 2.1). For instance, in *E. coli* the average length of the IGRs is close to 2000 base pairs (bps) on both strands (see Figure 2.2). Data employed to produce this figure have

been downloaded from NCBI (National Center for Biotechnology Information)<sup>1</sup> [12]. This database provides also the position of TSC, identified by an ATG triplet: in Figure 2.1 the TSC is represented by a black square.

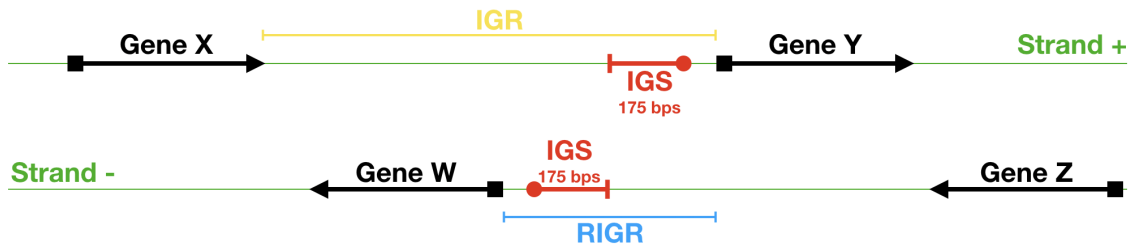


Figure 2.1: Scheme of identification of IGRs, RIGRs and IGSs in DNA bacterial strands: in green we denote the two DNA strands; the arrows denote the transcription direction; the black squares locate the TSCs, while the red circles locate the TSSs.

It is well known that the genome of bacterial species exhibits quite peculiar structures. For instance, they contain operons, i.e. groups of genes separated by short noncoding regions, that we assumed to be poorly relevant to our analysis. Accordingly, we have associated to the entire group of genes inside the operons a single IGR upstream the first TSC. In order to identify the genes contained in the operons we used DOOR (*Database of prokaryotic Operons*) [14, 15, 16], an operon database developed by Computational Systems Biology Lab (CSBL) at University of Georgia, covering 2072 bacteria genomes and with overall accuracy of 90%.

Moreover, most of these IGRs contain reverse complements of other genes on the opposite strand. A first step in the direction of the identification of IGSs amounts to restrict the extension of IGRs to the regions between genes, irrespectively of the strand they belong to. We term these regions restricted IGR (RIGR) (see Figure 2.1). The outcome of this procedure is a collection of RIGRs with different lengths, whose distribution is reported in Figure 2.3: these regions extend between the end of a gene and the beginning of the following gene if the latter belongs either to the same or to the opposite strand. Since the statistical methods that we are going to use in the following section necessitate sequences with the same length, we have to establish a criterion to define the reference length. We can observe that the average length of RIGRs is approximately 250 bps. In order to increase the statistical significance we have subtracted from this number the variance of this distribution, which amounts to about 50 bps, thus yielding sequences of length of 200 bps upstream the TSC.

<sup>1</sup>To avoid potential background noise, in this work we considered only chromosomal DNA and discarded plasmid DNA.

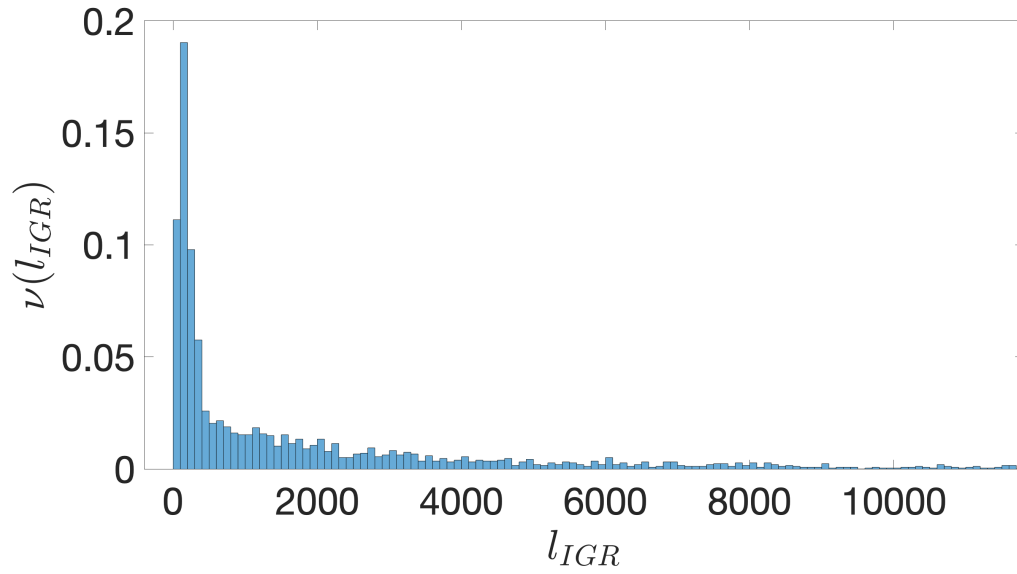


Figure 2.2: *E. coli*. The frequency  $\nu(l_{IGR})$  of IGRs versus their length,  $l_{IGR}$ , expressed in bps. The binning is over 100 bps. The distribution is truncated at 12000 bps.

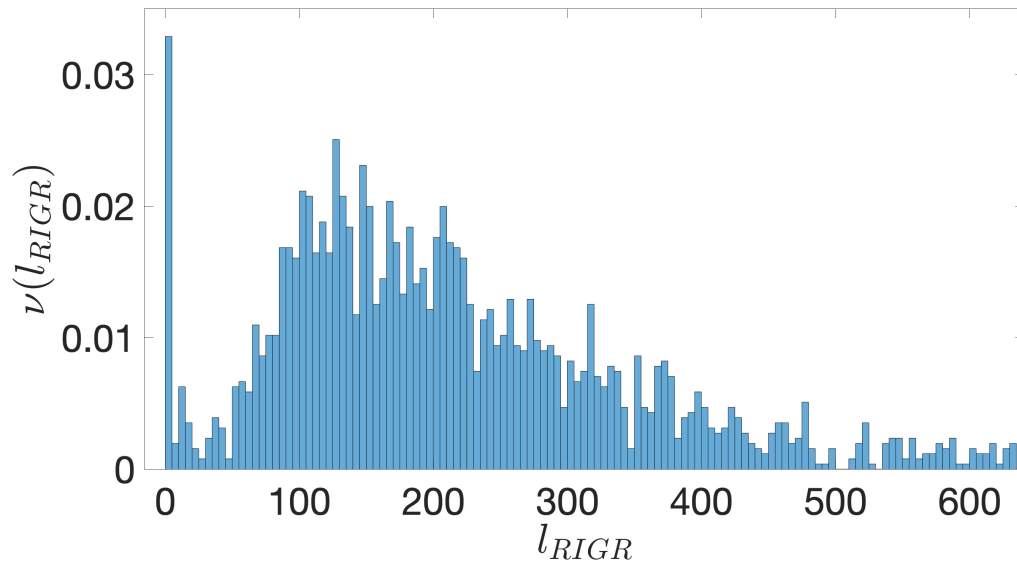


Figure 2.3: *E. coli*. The frequency  $\nu(l_{RIGR})$  of RIGRs versus their length,  $l_{RIGR}$ , expressed in bps. The binning is over 5 bps and the distribution is truncated at 600 bps. The peak close to 0 is due to the simplifying assumption of setting to 0 the contribution from overlapping coding regions.

### 2.1.1 Shine-Dalgarno sequences

As a final step for the identification of the IGSs, we have considered that regulatory features should be better ascribed to structural patterns belonging to sequences upstream the TSS (this is denoted by the red circle in Figure 2.1). We assume that the position of TSS in bacterial genomes corresponds to the first nucleotide of a Shine-Dalgarno sequence (SDS), everywhere this sequence is found. The SDS is a purine-rich ribosomal binding site, usually located a dozen bps upstream the TSC. The typical six-base consensus sequence is AGGAGG. The Base Composition Analysis (BCA) (see section 2.3) of the 200-bps-long noncoding sequences of *E. coli* (as well as those of the other bacteria analyzed in this thesis) exhibits a peak of the density of G nucleotides in the vicinity of 10 bps upstream the TSC, thus signaling the typical occurrence of SDS in this part of the noncoding sequences (data not reported). It is well known that the presence of a SDS is associated to the position of the TSS. This indicates that the 200-bps-long noncoding sequences of *E. coli* contain a noncoding region that is transcribed and not translated. Since our clustering analysis (see section 2.2) aims at characterizing structural similarities between strictly noncoding regions, we want to eliminate from any 200-bps-long noncoding sequence its portion upstream the TSC, that is transcribed and not translated. Moreover, following [17] we have considered as indicators of the TSS also all of its subsequences: GGA, GAG, AGG, GGAG, GAGG, AGGA, AGGAG, GGAGG. In more detail, the procedure for identifying SDSs is implemented as follows: for each of the 200-bps-long noncoding sequence we start looking for the longest SDS (AGGAGG) and if we do not find it we pass to shorter SDSs, proceeding in dissent order of length, up to the three bps long SDSs. The different kinds of SDSs and their frequency in *E. coli* genome are reported in Figure 2.4. For instance, in *E. coli* the SDSs are found in approximately 88% of the previously identified sequences, extending over 200 bps upstream the TSC. By performing a complete identification of SDSs in *E. coli* genome, we have checked that they are typically found in a range extending over the first 25 bps upstream the TSC (see Figure 2.5), so we look for the presence of a SDS in the first 25 bps upstream the TSC <sup>2</sup>. When we find a sequence of nucleotides matching with a SDS we annotate the position of its first nucleotide and we associate it to the TSS of the corresponding gene or operon. Accordingly, we adopt the criterion of considering as IGSs those sequences extending upstream 175 bps from

---

<sup>2</sup>The seemingly magic number 25 stems from the direct inspection of Shine-Dalgarno motifs in the noncoding sequences of *E. coli*: the probability of finding a Shine-Dalgarbo motif upstream the TSC practically vanishes beyond 25 bps.



the TSS <sup>3</sup>(see Figure 2.1). For those genes that are not preceded by a SDS, the TSS coincides with the TSC, and consistently we identify the IGS with the 175 bps upstream the TSC.

The overall selection procedure applied to the *E. coli* NCBI database provides us 2553 equal-length IGSs, each one made of 175 bps. The same criterion has been adopted for the other bacterial species (see section 2.4) and we have found that a length of 175 bps for equal-length IGSs applies also to the other species. It is evident that the criterion

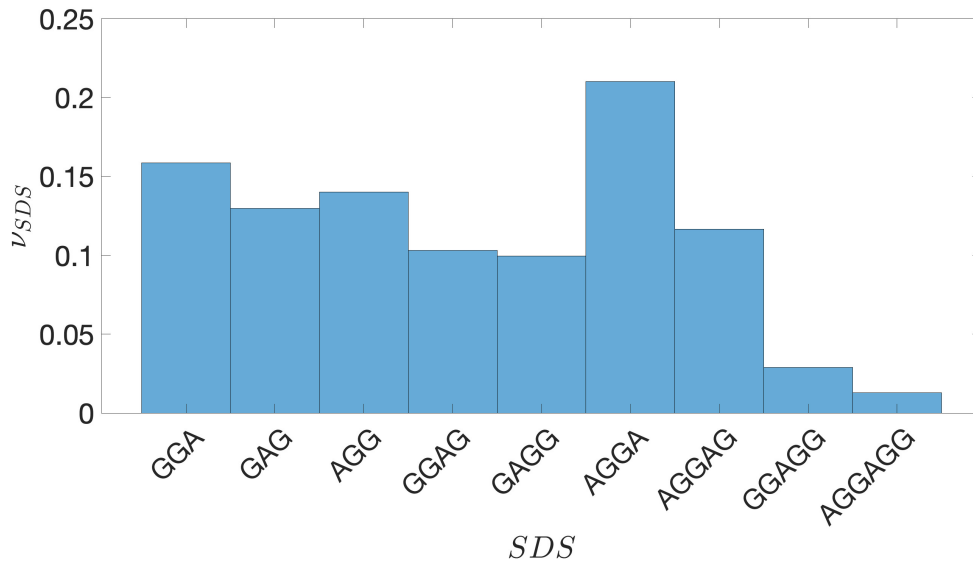


Figure 2.4: *E. coli*. The frequency  $\nu_{SDS}$  of the different SDSs located upstream the TSC, listed along the horizontal axis.

adopted for identifying equal-length IGSs unavoidably introduces portion of coding or reverse complement of coding sequences into the statistical sample of IGSs. On the other hand we have directly checked that in *E. coli* only 434 IGSs contain more than 50% of coding portions and this poorly affects the statistical significance of the chosen sample. Similar figures are found for the other bacterial species analyzed in this thesis.

## 2.2 Spectral Clustering

The aim of the procedures described in this section is to collect the annotated IGSs into clusters depending on the similarity between the sequences. This procedure consists of three main steps:

---

<sup>3</sup> In order to remove the ambiguity due to the possible presence of various SDSs in the noncoding sequence upstream the TSC we have adopted the criterion of taking the longest one as a reference for identifying the corresponding IGS.

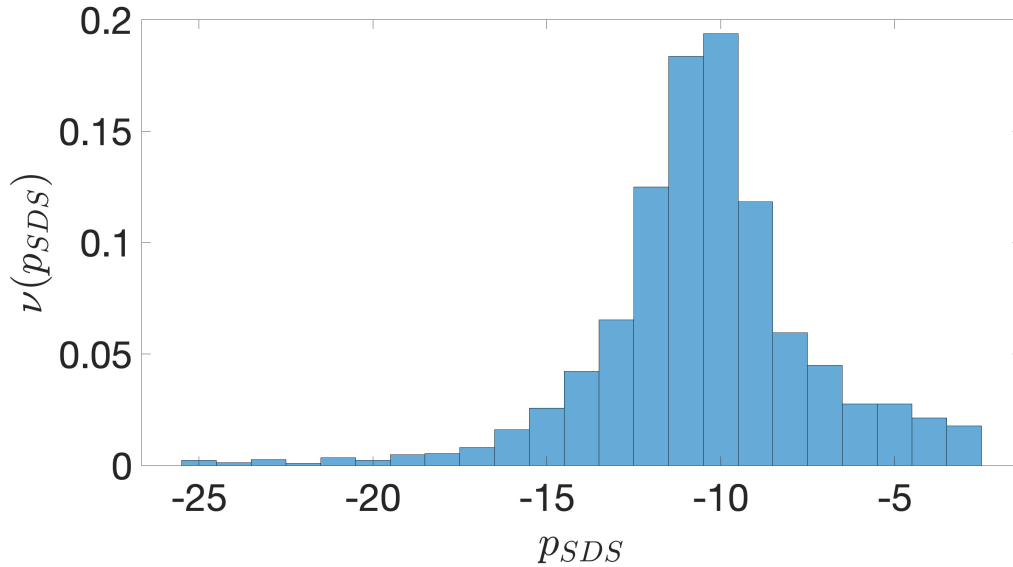


Figure 2.5: *E. coli*. The frequency  $\nu(p_{SDS})$  of the position  $p_{SDS}$  of the SDSs upstream the TSC.

- aligning each sequence with all the others (pairwise alignment), thus obtaining a matrix whose entries are similarity scores;
- analyzing the eigenvalues of the Laplacian matrix, computed by the similarity matrix, for determining the appropriate number of clusters;
- making use of the eigenvectors of the Laplacian matrix to work out the k-means algorithm, which allows us to associate each IGS to the selected clusters.

## 2.2.1 Sequences alignment

The basic idea of a sequence alignment is to identify regions of similarity that may be related with functional or structural properties as well as evolutionary relationships. Clearly, any alignment procedure cannot be based on a perfect match between sequences, but it has to take into account important biological features such as mutations and insertions or deletions occurred during the evolution. For this reason, the standard approach to this problem is to implement computational methods that make use of a substitution matrix to assign positive and negative scores to nucleotide matches or mismatches, and a gap penalty for matching a nucleotide in one sequence to a gap in the other one. These algorithms, in general, fall into two categories: global and local techniques. An example of alignment is reported in Figure 2.6.

A global algorithm spans the entire length of the sequence, while a local alignment focuses on identifying regions of similarity within long sequences that are often widely

```

# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 231
# Identity:      94/231 (40.7%)
# Similarity:   94/231 (40.7%)
# Gaps:        112/231 (48.5%)
# Score: 132.5
#
#
#=====
EMBOSS_001      1 ACCGCCAAAAGCGAGAAGTACGGGCAGGTGCTATGA--CCA----GGACT      44
                    |.||          |||.|  .||  |.|||
EMBOSS_001      1 -----ATTA-----TATAATGGCATTGTGTACT      23

EMBOSS_001     45 TTT--TGACCTGAAGTCCGGATAAAAAACAGCAACAATGTGAGCTTTGTTG     92
                    .|| | |||||.|.|  .|.|||.|||.|  |||.|
EMBOSS_001     24 CTTACT-ACCTGTATT-----CACCAAGTAAGA-----GTAG      53

EMBOSS_001     93 TAATTATATTGTAACA-TAT----TGCTAAA---TGTT----TTTAC     128
                    |||||.|||  || |||  |.|||||  |||.|  |.||.
EMBOSS_001     54 TAATCAGAT-----CAGTATGTCCGTACTAAAATCTGATACCAGTATAT     97

EMBOSS_001    129 ATCCACTACA---ACCA-----TAT--CATCACAAGT--     155
                    |.|.|||.|||  ||||  |||  |||||.|||
EMBOSS_001     98 ACCGACAACAGTGACCAGGAATAGCCCTGCCTGTATGGCATC-CCTGTAC     146

EMBOSS_001    156 GGTC-----AGACCTCCTA---CAAGTA-      175
                    ||||  ||||  |.||  |||.||
EMBOSS_001    147 GGTCGTCTATAGAC--CGTAGGACCAAATAC      175

```

Figure 2.6: We report an example of alignment of two sequences. The gap insertions (horizontal dashes) in both sequences are shown, as well as matches (vertical bars) and mismatches (black points) between sequences.

different overall. In this paper we have made use of the two most popular alignment methods, the Needleman-Wunsch global algorithm [18] and the Smith-Waterman local algorithm [19] implemented in the EMBOSS package version 6.6.0 [20].

A key aspect of the procedure, which may give rise to a marked difference in the best match score calculated by the two algorithms, is the choice of the penalty value to be assigned to the introduction of a new gap in the alignment (GAOPEN) and the value for each consecutive gap (GAPEXTEND); the scoring matrix for the nucleotide substitution has been taken equal to the standard EDNAFULL matrix for both methods:

$$\begin{matrix} & A & T & G & C \\ \begin{matrix} A \\ T \\ G \\ C \end{matrix} & \begin{pmatrix} +5 & -4 & -4 & -4 \\ -4 & +5 & -4 & -4 \\ -4 & -4 & +5 & -4 \\ -4 & -4 & -4 & +5 \end{pmatrix} & & & \end{matrix} \quad (2.1)$$

It assigns different scores to matched/mismatched nucleotides depending on the particular symbols aligned; these scores are computed based on the relative frequency of one nucleotide to be substituted to another in a collection of known alignments. The scoring

matrix is symmetric, meaning that the score assigned to substituting symbol  $x$  for symbol  $y$  is the same as that of substituting  $y$  for  $x$ .

Unfortunately there's no way to set a priori the optimal choice of parameters and thus the best option is to tune the values depending on the results obtained. Regarding our work, the trials we performed suggest to use a high GAPOPEN value (typically set equal to 10) and a low GAPEXTEND penalty (0.5) in order not to penalize long gap sequences. This setting favors the scores of very similar sequences yielding an easier detection of the suitable number of clusters (see subsection 2.2.2). Moreover, in the EMBOSS code, gaps inserted at the beginning or at the end of the sequence have no penalty. In this way, we do not observe a significant difference between the two algorithms, and the outcome of aligning  $N$  IGSs gives essentially the same similarity matrix  $S$  in both cases.

## 2.2.2 The normalized Laplacian matrix

The following step is the application of the same clustering strategy adopted for *H. sapiens* in [1], that takes into account the global properties of the identified IGSs, instead of specific short regulatory motifs. The clustering procedure is based on the spectral analysis of a similarity matrix: the entries of such matrix are obtained by the alignment algorithm, that quantifies the similarity between IGSs. Since the number of identified IGSs in *E. coli* is relatively small, the alignment protocol and the diagonalization of the similarity matrix can be performed with avoiding the computational limitations encountered for much larger sets of promoters, as those typically found in eukaryotes (see [1]).

A convenient way to represent the  $N \times N$  entries  $s_{ij}$  of the symmetric similarity matrix  $S$ , is to introduce a network, whose nodes coincide with the IGSs, while the entry  $s_{ij}$  represents the weighted link between sequence  $i$  and  $j$ . For the purpose of our work, however, dealing with a fully connected network is not the best approach. The risk is that the noise induced by the fact that even the alignment of two random sequences gives a positive score, may hide the real common features among IGSs, making the clustering procedure unfruitful. For this reason, it is of paramount importance to substitute  $S$  with a weighted adjacency matrix  $W$ , for which two nodes are connected only if their alignment score is larger than a certain threshold  $s^*$ , namely  $w_{ij} = s_{ij}$  if  $s_{ij} \geq s^*$  and  $w_{ij} = 0$  otherwise. To estimate  $s^*$ , we have used two methods that we describe in the following.

Finally, in order to manage a set of more homogeneous data, we have operated the normalization  $w_{ij} \rightarrow w_{ij}/\max\{w_{ij}\}$ . Following [21], once an appropriate similarity matrix

is obtained, the first step of the clustering procedure is the determination of the number of clusters. For this purpose, we introduce the normalized Laplacian  $L_{sym} = D^{-1/2}(D - W)D^{-1/2}$  where the degree matrix  $D$  is defined as the diagonal matrix with entries  $d_i = \sum_{j=1}^N w_{ij}$ . In some particularly successful cases,  $L_{sym}$  has a block structure, and the multiplicity of its null eigenvalue determines the number of connected components. In real cases, however, data is well mixed, and  $L_{sym}$  has a unique null eigenvalue corresponding to one connected component, which includes the whole data set. The solution of the problem comes from the matrix perturbation theory [22]. Indeed, given the spectrum  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$  of  $L_{sym}$ , the information about the number of clusters is carried by those eigenvalues which are located close to the null one. The idea is that the actual  $L_{sym}$  can be read as a perturbation of an *ideal* block matrix, and thus the first  $k$  values of the spectrum act as fluctuations of the corresponding null eigenvector of the *ideal* case, with multiplicity  $k$ . In practice, the more the first  $k$  eigenvalues are distant from the others, the more effective will be the separation of data into the  $k$  groups.

Hence, the eigenvalues of the Laplacian matrix, associated to the similarity matrix, are expected to highlight the presence of possible clusters of IGSs for *E. coli*. The result of our analysis is shown in Figure 2.7, where we report these eigenvalues in ascending order. Symbols with different colors correspond to the eigenvalues obtained for two different values of the similarity threshold  $s^*$ . In particular, the red eigenvalues have been obtained by the unbiased averaging procedure adopted for estimating the similarity threshold in eukaryotes [1]. In this first method  $s^*$  has been estimated by reordering randomly the nucleotides of each one of the  $N$  annotated IGSs and then applying to this new set of  $N$  randomized sequences the alignment algorithm. The arithmetic mean of the  $s_{ij}$  obtained for the randomized sequences provides a preliminary estimate of  $s^*$ . By iterating this procedure to perform a further averaging over different estimates, we have checked that the preliminary estimate is pretty stable. The second method for estimating  $s^*$  is based on the computation of the alignment score of each one of the annotated IGSs with ten realizations of the random reordering of its nucleotides. Then  $s^*$  is computed as the arithmetic mean of the alignment scores obtained for all IGSs. Since in *E. coli* the length of IGSs is definitely smaller than the one of eukaryotic promoters, we have adopted this second, more effective statistical procedure for the determination of the similarity threshold, which actually yields a better discrimination of the eigenvalues. In fact, this method yields an estimate of the threshold that is typically higher than the one obtained with the former procedure. For instance, in Figure 2.7 the eigenvalues of the Laplacian

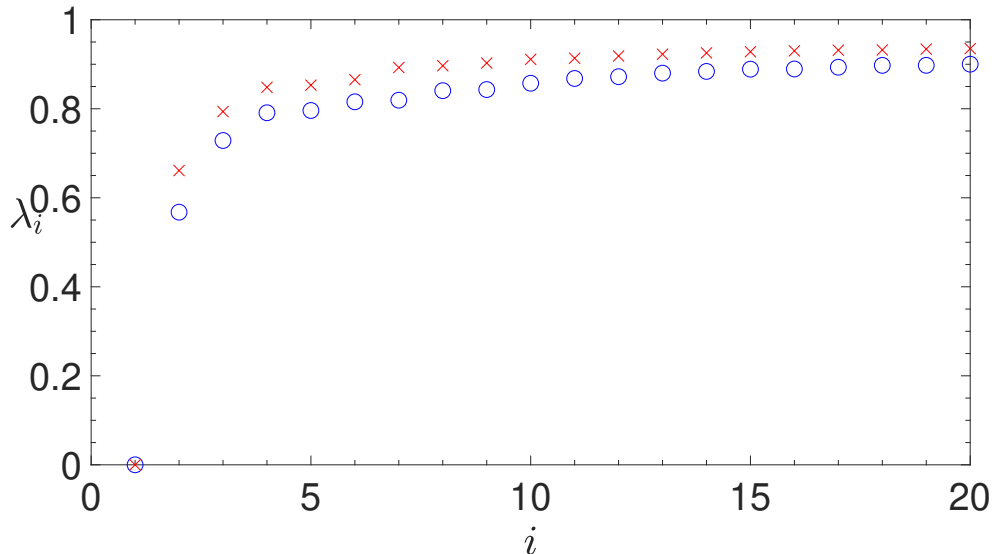


Figure 2.7: *E. coli*. The first twenty eigenvalues in ascending order of the normalized Laplacian matrix obtained by the alignment of the IGSs. Red crosses and blue circles correspond to different values of the similarity threshold, determined by the two statistical approaches described in the present section. A better discrimination of the three main eigenvalues is obtained for a higher similarity threshold (blue circles), which corresponds to the second statistical approach, better suited for short sequences, as the annotated IGSs.

matrix obtained by the second method (blue symbols) are lower than those obtained by the first method (red symbols), while their relative separation is more pronounced. Since we are interested in highlighting structural similarities between the relatively short IGSs of bacterial species, we have constructed the Laplacian matrix making use of the second method.

By this procedure we have obtained the blue eigenvalues shown in Figure 2.7. They allow us to identify three different clusters, corresponding to the three lowest nonzero eigenvalues, that can be distinguished from the total set, because of their sensibly different values between each other. Hereinafter they will be referred to as C0, C1 and C2.

### 2.2.3 Clustering algorithm

We are now able to apply the spectral clustering algorithm in order to assign each IGS to one of the clusters. The starting point is the computation of the first  $k$  eigenvectors  $u_1, \dots, u_k$  of  $L_{sym}$ , so as to form a new matrix  $U \in \mathbb{R}^{N \times k}$  containing the vectors  $u_1, \dots, u_k$  as columns. Let  $T \in \mathbb{R}^{N \times k}$  be the matrix obtained from  $U$  by normalizing the rows to norm 1, namely,  $t_{i,j} = u_{i,j} / (\sum_k u_{i,k}^2)^{1/2}$ . For  $i = 1, \dots, N$  we denote by  $y_i \in \mathbb{R}^k$  the vector corresponding to the  $i$ -th row of  $T$ . The last point consists in applying the k-

means algorithm to the points  $y_i$  so as to find  $C_1, \dots, C_k$  clusters. The iterative procedure of the algorithm works as follows: first, select  $k$  random points as initial centroids. Then, form  $k$  clusters assigning each point  $y_i$  to its closest centroid, according to Euclidean distance. Recompute the centroids as the mean of the points of each cluster. Repeat until the difference between the centroids coordinates of two consecutive steps reaches a fixed tolerance. For instance, in Figure 2.8 this tolerance was fixed to  $10^{-8}$ .

The reliability of this procedure is illustrated by representing the distribution of IGS in the so-called clustering space, shown in Figure 2.8. Each point in this space corresponds to an IGS, while IGSs with a high similarity score are represented as nearby points. Each of the 2553 IGS has been unambiguously associated to one of the three clusters.

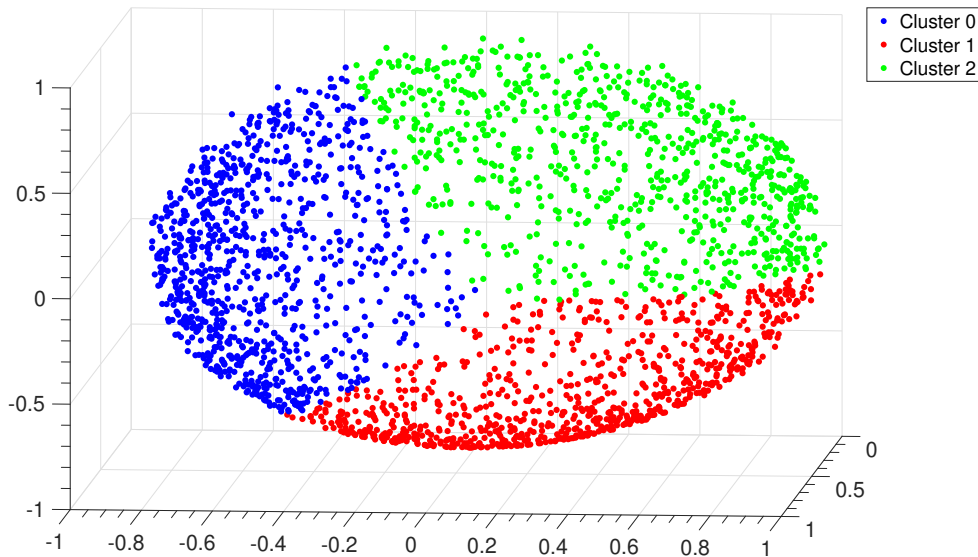


Figure 2.8: *E. coli*. Distribution of points in the clustering space relative to the alignment of the IGSs. Each point represents an IGS and the color code corresponds to the three clusters identified by the *Clustering Algorithm*.

## 2.2.4 Silhouette

Silhouette index allows to evaluate the consistency of a clustering procedure, measuring how similar an object is to its own cluster compared to other clusters. The silhouette value ranges from  $-1$  to  $+1$ , where positive values indicate that the assignment of the object to a cluster is good, while negative values stand for a bad assignment. The clustering configuration is more appropriate the more objects are characterized by a silhouette value close to  $+1$ . Now we illustrate how to calculate the silhouette.

For a data point  $i$  in the cluster  $C_k$  with  $N_k$  elements, we define

$$a(i) = \frac{1}{N_k - 1} \sum_{j \in C_k} d(i, j), \quad (2.2)$$

where the sum is over all the data points and  $d(i, j)$  is the Euclidean distance between the two points. The value  $a(i)$ , the average distance of  $i$  with all other data points in the same cluster, is as a measure of how well  $i$  is assigned to its cluster. Let

$$b(i) = \min_{h \neq k} \frac{1}{N_h} \sum_{j \in C_h} d(i, j) \quad (2.3)$$

be the smallest mean distance of  $i$  to all points in any other cluster  $C_h$  with  $N_h$  elements, of which  $i$  is not a member.

The silhouette value for a data point  $i$  is

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \quad (2.4)$$

So if  $a(i) < b(i)$  then  $s(i)$  will be positive, negative otherwise.

To ascertain the consistency of the clustering procedure we have calculated the silhouette values of each point. The distributions of these values for each cluster are reported in Figure 2.9. We can observe that the vast majority of values are positive, with a shift of the distribution towards the value +1, thus confirming that the clustering configuration is appropriate<sup>4</sup>.

The clustering method has been applied also to bacterial species different from *E. coli*, a Gram positive bacterium (*Bacillus subtilis*) and an extremophilic bacterium (*Pseudoalteromonas haloplanktis*). The results are reported in section 2.4. Altogether, the analysis based on clustering by alignment yields similar results for IGSs in different bacterial species. This indicates that, irrespectively of the considered bacterium, each identified cluster of IGSs is associated to the presence of global structural properties. Now, the main question concerns the identification of the structural features characterizing the different clusters.

---

<sup>4</sup>We have observed that the silhouette criterion improves for the division into two clusters (corresponding to the first two eigenvalues of the Laplacian matrix). On the other hand the heuristic rule to establish the number of appropriate clusters (see subsection 2.2.2) amounts to choose it for first eigenvalues which maintain a significant difference between each other. This is why we have chosen to consider three clusters. A posteriori this heuristic choice is justified by the significantly different structural features characterizing the BCA of the three clusters, as shown in Figure 2.12.



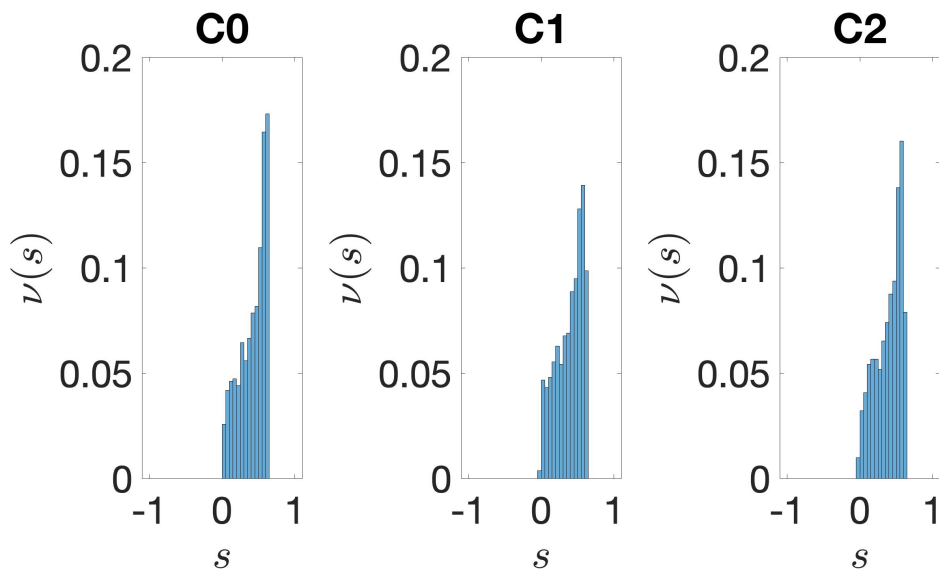


Figure 2.9: *E. coli*. Distribution of silhouette values relative to the clustering of the IGSs. On the vertical axis we report the frequency  $\nu(s)$  of IGSs versus the silhouette value  $s$ ; this value is between -1 and +1. The average values are 0.42 for cluster C0, 0.39 for C1 and 0.39 for C2.

## 2.3 Base composition analysis

In order to characterize a set of  $N$  equal-length sequences, it is useful to represent the spatial distribution of each nucleotide along the IGSs by the so-called Base Composition Analysis. In practice, we compute the density  $\rho_x(\ell)$  of each nucleotide  $x = A, T, G, C$  at position  $\ell$  along the IGS defined as

$$\rho_x(\ell) = \frac{1}{N} \sum_{i=1}^N s_i^x(\ell), \quad (2.5)$$

with  $s_i^x(\ell) = 1$  if in the  $i$ -th IGS the nucleotide  $x$  is present at position  $\ell$ ,  $s_i^x(\ell) = 0$  otherwise. For what concerns the annotated IGSs of the bacteria considered in this paper  $\ell = -175, \dots, -1$ , while the position 0 corresponds to the first nucleotide of the SDS, or to the TSC for those IGSs where the SDS is lacking.

In this section, we take the opportunity to report the BCA of the three clusters obtained after the clustering procedure, as we can see in Figure 2.11.

Since, for not too large values of  $N$ , BCA typically exhibits sensible fluctuations, the density profile of nucleotides can be better represented by a smoothing procedure, where

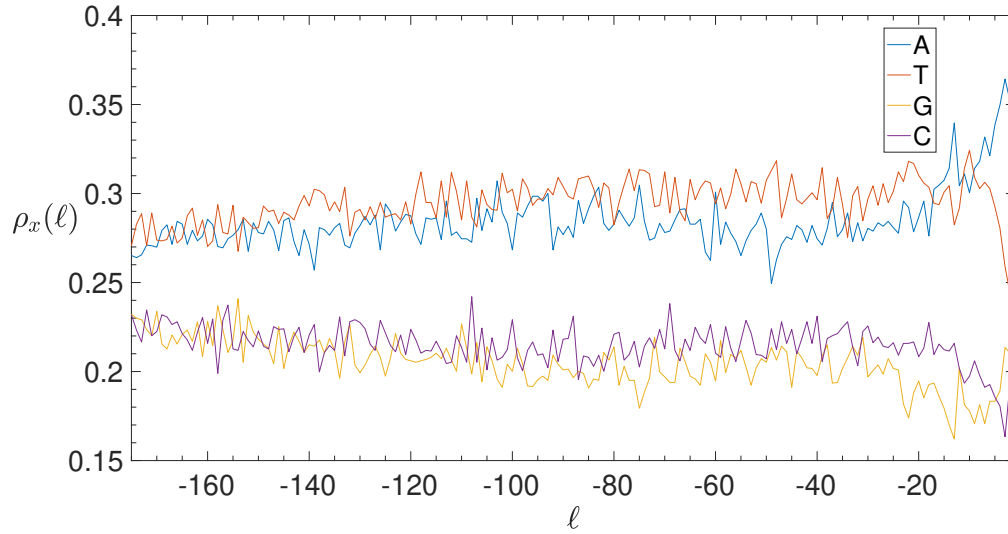


Figure 2.10: *E. coli*. BCA of the IGSs: on the vertical axis we report the density  $\rho_x(\ell)$  of each of the four nucleotides  $x = A$  (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the annotated 2553 IGSs.

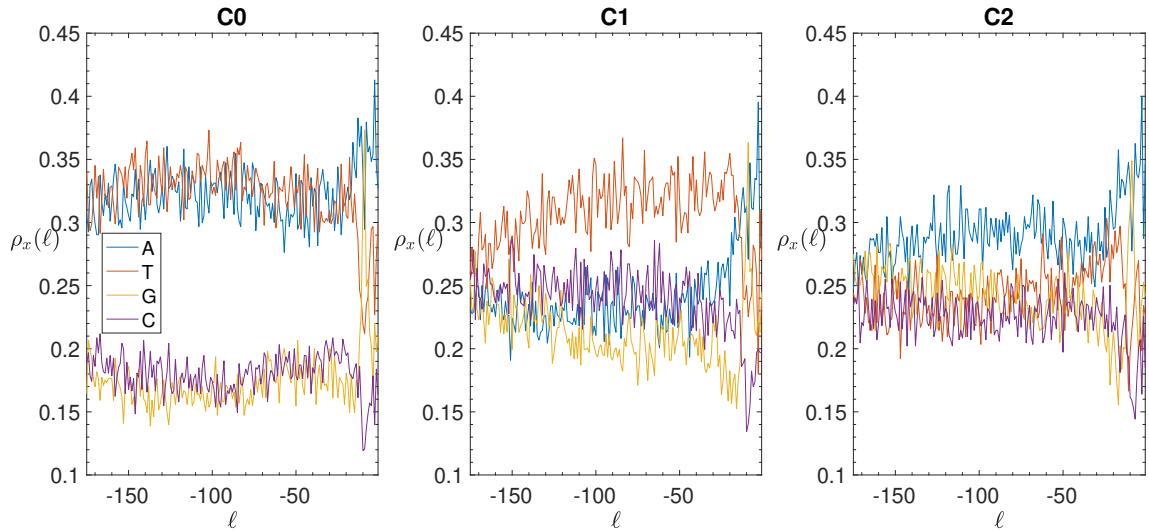


Figure 2.11: *E. coli*. Base composition analysis in the clusters of the IGSs: on the vertical axis we report the density  $\rho_x(\ell)$  of each of the four nucleotides  $x = A$  (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the IGS belonging to the clusters C0 (left panel), C1 (central panel) and C2 (right panel).

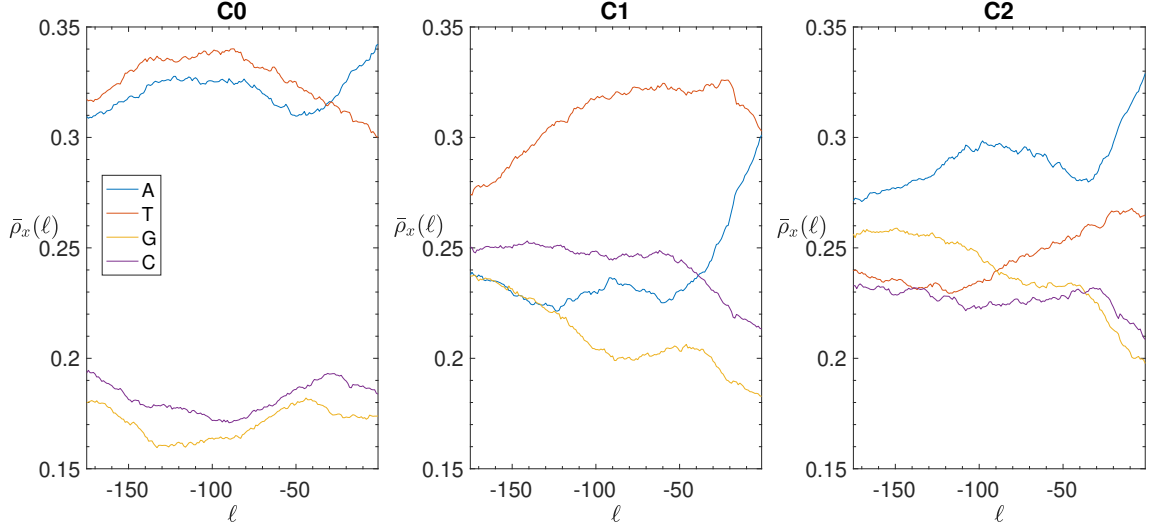


Figure 2.12: *E. coli*. Smoothed Base composition analysis in the clusters of the IGSs: on the vertical axis we report the averaged density  $\bar{\rho}(\ell)$  for  $a = 15$  bps of each of the four nucleotides A (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the IGSs belonging to the clusters C0 (left panel), C1 (central panel) and C2 (right panel).

we proceed to a further averaging of the density inside a "window" of  $2a$  bps; in formulae

$$\bar{\rho}_x(\ell) = \frac{1}{N} \sum_{i=1}^N \frac{1}{1+2a} \sum_{\ell'=\ell-a}^{\ell'+a} s_i^x(\ell'). \quad (2.6)$$

For *E. coli* we report in Figure 2.12 the result of this procedure.

## 2.4 Other bacterial species

Here we insert all information concerning other bacterial species: *B. subtilis* and *P. haloplanktis*. For each bacterium we highlight the major differences.

All the figures are qualitatively very similar to those of *E. coli*. We can notice a difference for what concerns the BCA of the cluster C0. In fact in *E. coli* and in *P. haloplanktis* the weak nucleotides show the same profile in their spatial distributions (first panel of Figure 2.11), while in *B. subtilis* have a specular behavior, when A grows T decreases and vice versa (see first panel of Figure 2.21).

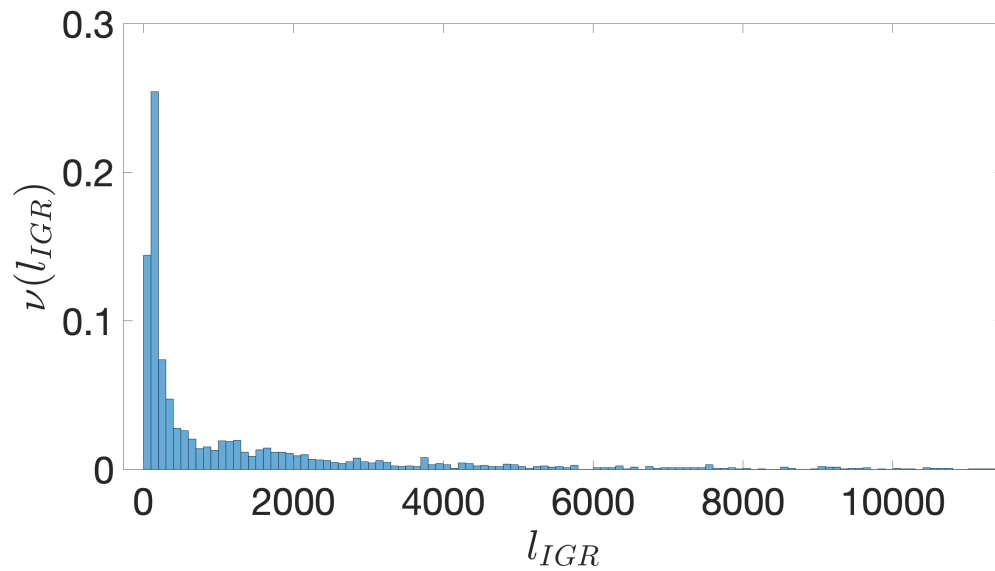


Figure 2.13: *B. subtilis*. The frequency  $\nu(l_{IGR})$  of IGRs versus their length,  $l_{IGR}$ , expressed in bps. The binning is over 100 bps. The distribution is truncated at 12000 bps.

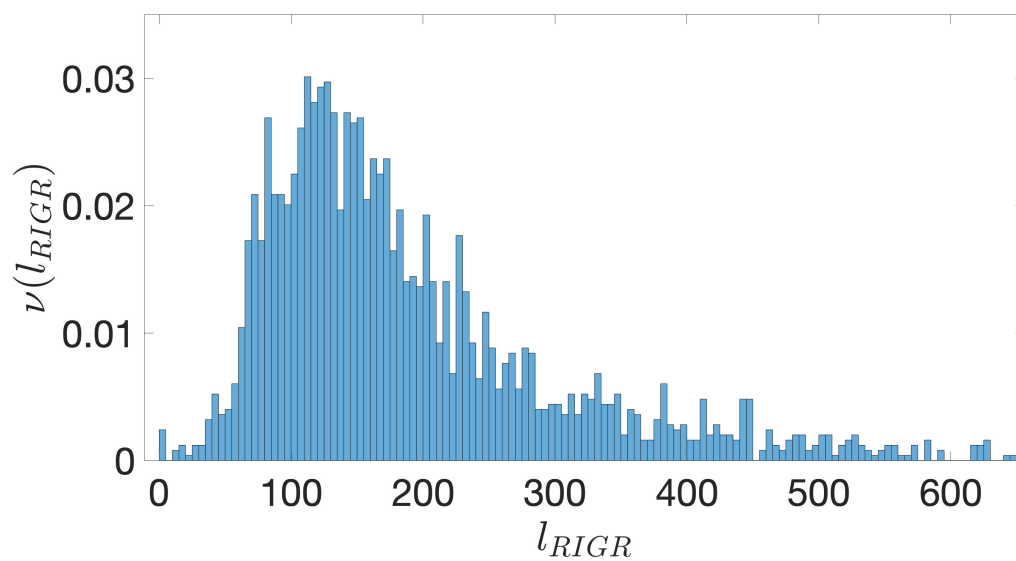


Figure 2.14: *B. subtilis*. The frequency  $\nu(l_{RIGR})$  of RIGRs versus their length,  $l_{RIGR}$ , expressed in bps. The binning is over 5 bps and the distribution is truncated at 600 bps. The peak close to 0 is due to the simplifying assumption of setting to 0 the contribution from overlapping coding regions.

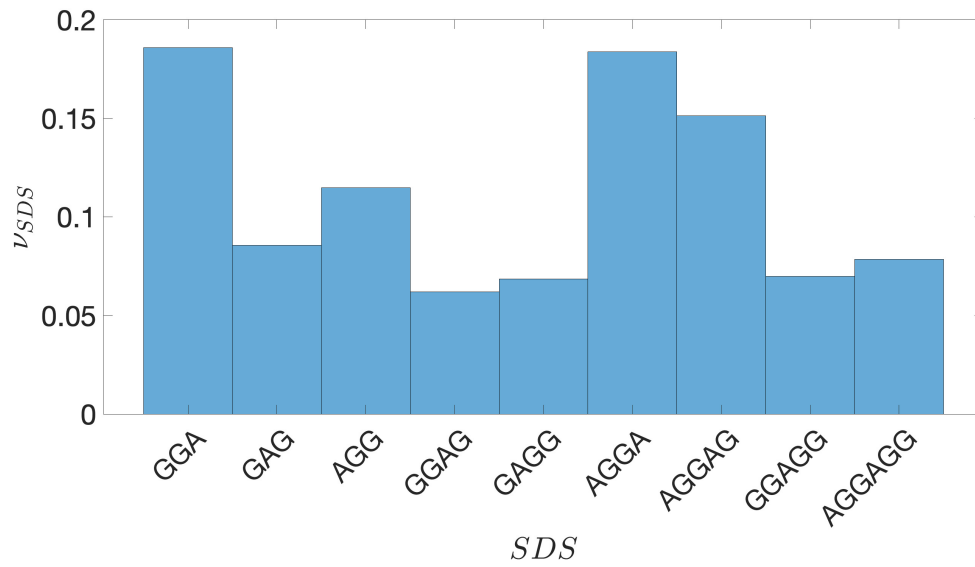


Figure 2.15: *B. subtilis*. The frequency  $\nu_{SDS}$  of the different SDSs located upstream the TSC, listed along the horizontal axis.

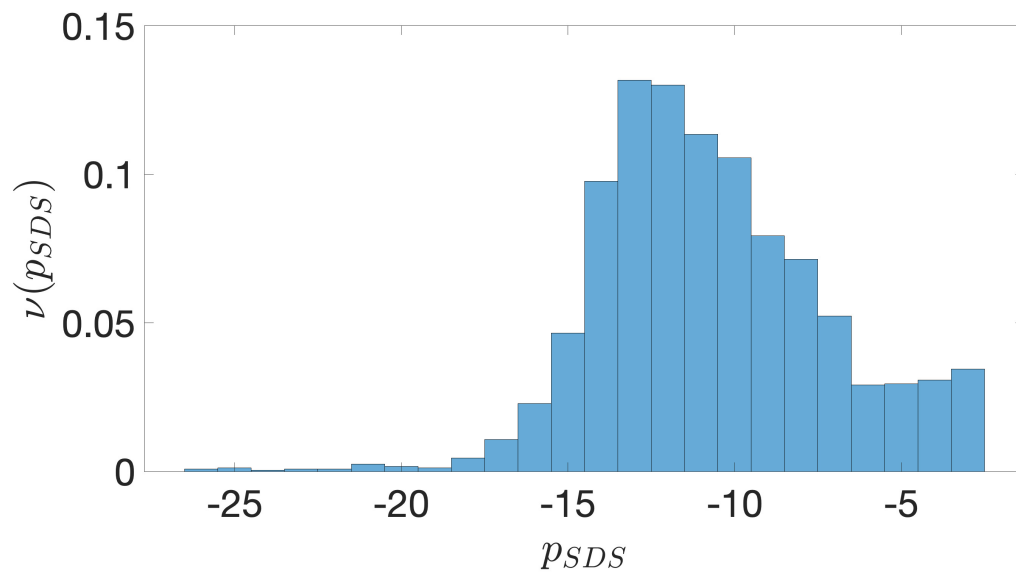


Figure 2.16: *B. subtilis*. The frequency  $\nu(p_{SDS})$  of the position  $p_{SDS}$  of the SDSs upstream the TSC.

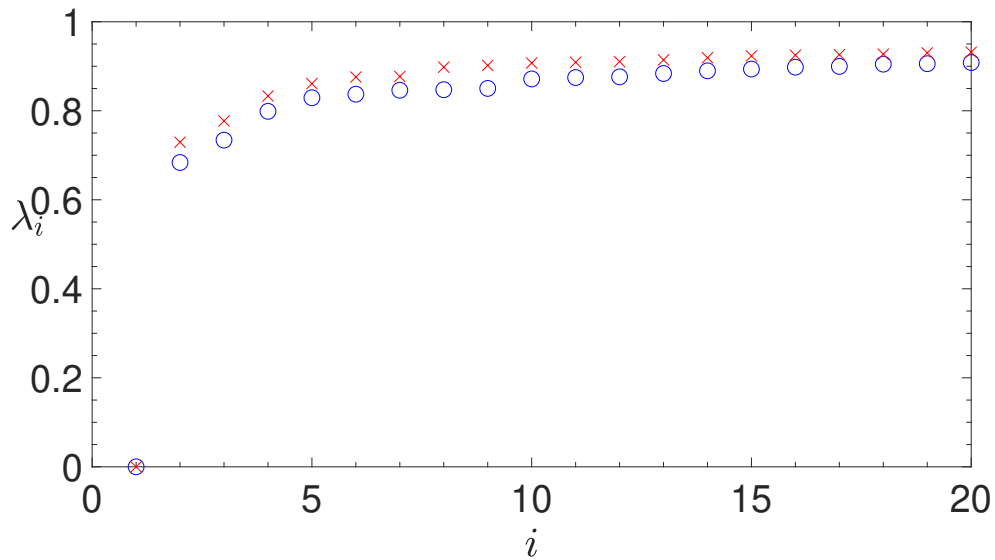


Figure 2.17: *B. subtilis*. The first twenty eigenvalues in ascending order of the normalized Laplacian matrix obtained by the alignment of the IGSs. Red crosses and blue circles correspond to different values of the similarity threshold, determined by the two statistical approaches described in subsection 2.2.2. A better discrimination of the three main eigenvalues is obtained for a higher similarity threshold (blue crosses), which corresponds to the second statistical approach, better suited for short sequences, as the annotated IGSs.

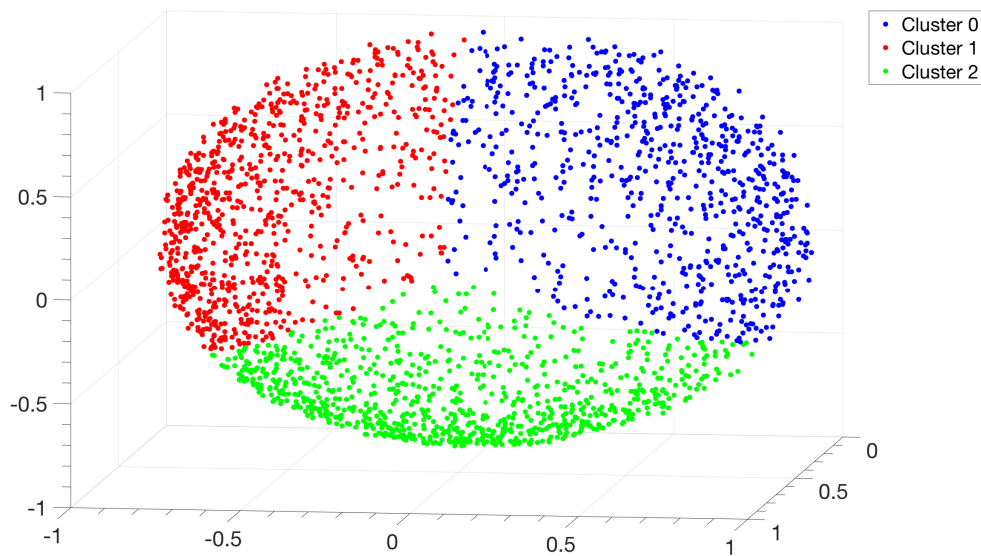


Figure 2.18: *B. subtilis*. Distribution of points in the clustering space relative to the alignment of the IGSs. Each point represents an IGS and the color code corresponds to the three clusters identified by the *Clustering Algorithm* described in subsection 2.2.3.

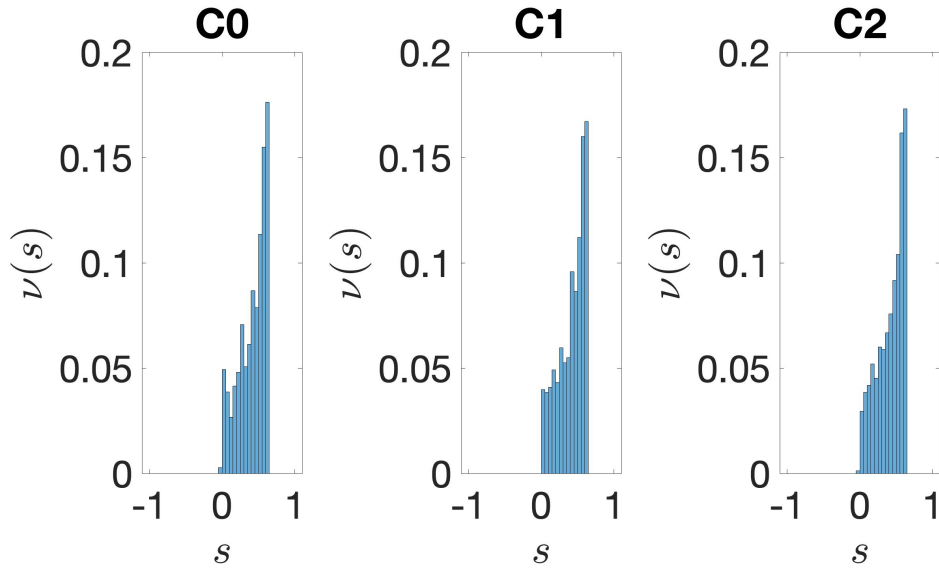


Figure 2.19: *B. subtilis*. Distribution of silhouette values relative to the clustering of the IGSs. On the vertical axis we report the frequency  $\nu(s)$  of IGSs versus the silhouette value  $s$ ; this value is between -1 and +1. The average values are 0.42 for all the clusters.

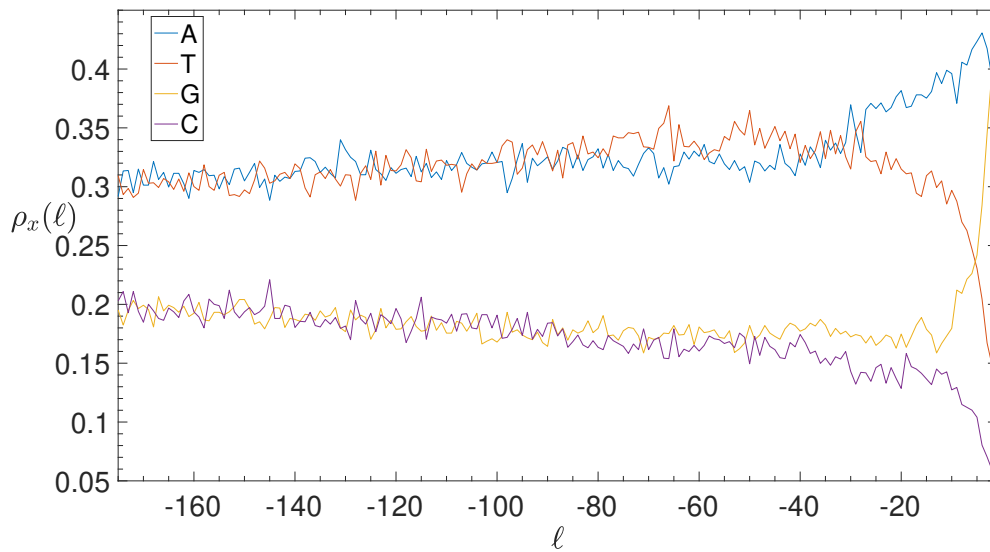


Figure 2.20: *B. subtilis*. BCA of the IGSs: on the vertical axis we report the density  $\rho_x(\ell)$  (see section 2.3) of each of the four nucleotides  $x = A$  (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the annotated 2338 IGSs.

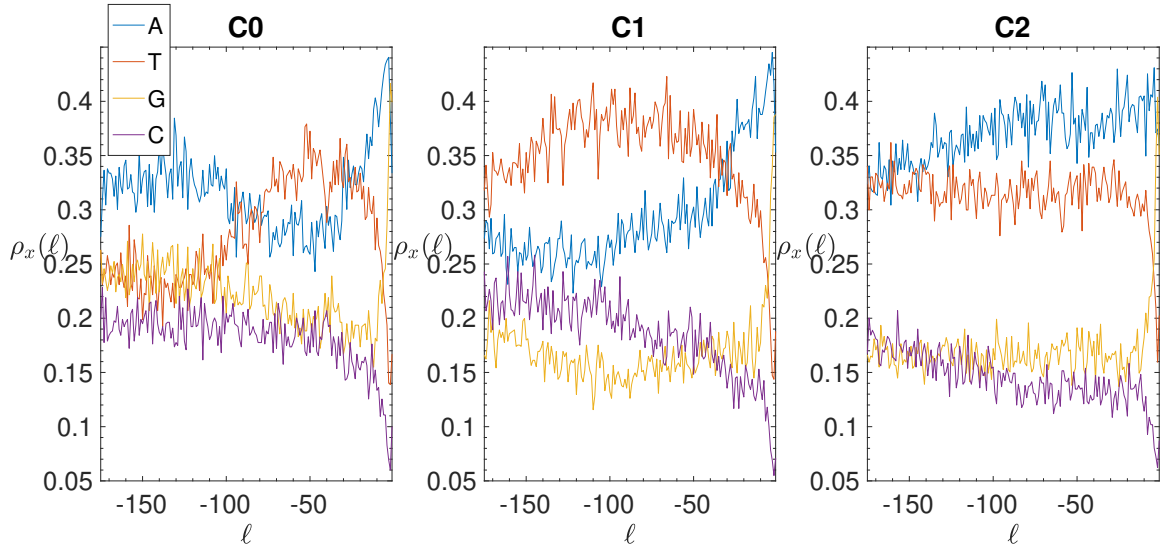


Figure 2.21: *B. subtilis*. Base composition analysis in the clusters of the IGSs: on the vertical axis we report the density  $\rho_x(\ell)$  (see section 2.3) of each of the four nucleotides  $x = A$  (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the IGS belonging to the clusters C0 (left panel), C1 (central panel) and C2 (right panel).

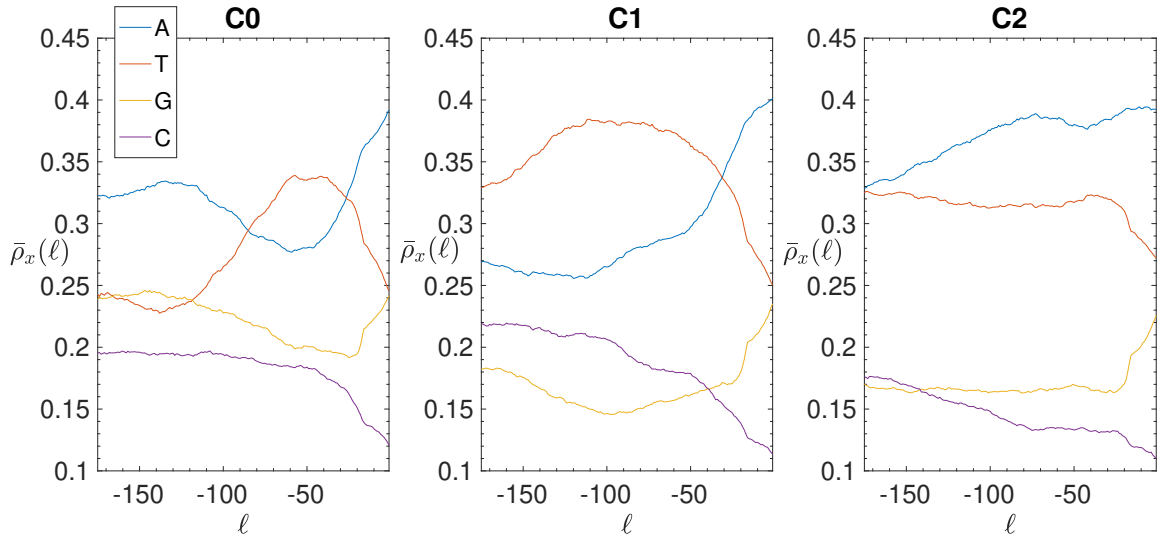


Figure 2.22: *B. subtilis*. Smoothed Base composition analysis in the clusters of the IGSs: on the vertical axis we report the averaged density  $\bar{\rho}(\ell)$  for  $a = 15$  bps (see section 2.3) of each of the four nucleotides A (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the IGSs belonging to the clusters C0 (left panel), C1 (central panel) and C2 (right panel).

For *P. haloplanktis* the more remarkable difference with *E. coli*, but also with *B. subtilis*, is the shape of the distribution of points in the clustering space (see Figure 2.28). This seemingly unusual distribution is due to the presence of scattered points which correspond to IGSs very far from the centroid of the different clusters. This is peculiar of this bacterium; on the other hand if these few atypical IGSs would be eliminated from the



sample one should recover a point distribution very similar to those reported in Figure 2.8 and Figure 2.18.

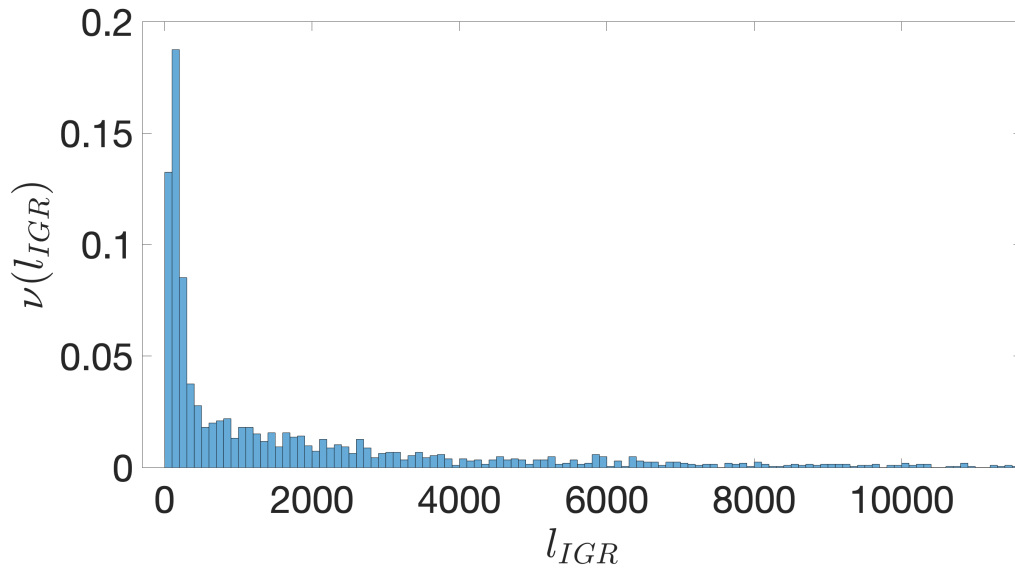


Figure 2.23: *P. haloplanktis*. The frequency  $\nu(l_{IGR})$  of IGRs versus their length,  $l_{IGR}$ , expressed in bps. The binning is over 100 bps. The distribution is truncated at 12000 bps.

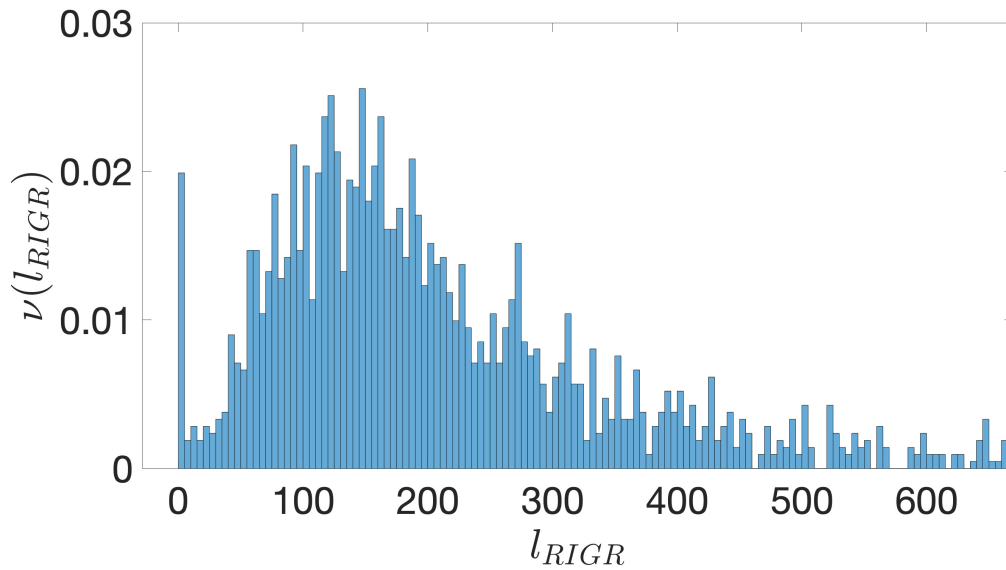


Figure 2.24: *P. haloplanktis*. The frequency  $\nu(l_{RIGR})$  of RIGRs versus their length,  $l_{RIGR}$ , expressed in bps. The binning is over 5 bps and the distribution is truncated at 600 bps. The peak close to 0 is due to the simplifying assumption of setting to 0 the contribution from overlapping coding regions.

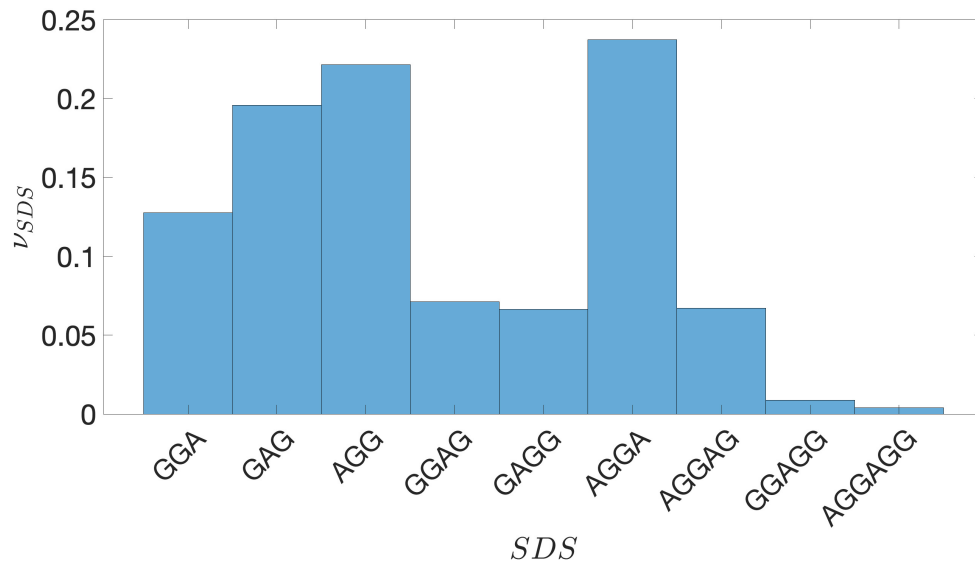


Figure 2.25: *P. haloplanktis*. The frequency  $\nu_{SDS}$  of the different SDSs located upstream the TSC, listed along the horizontal axis.

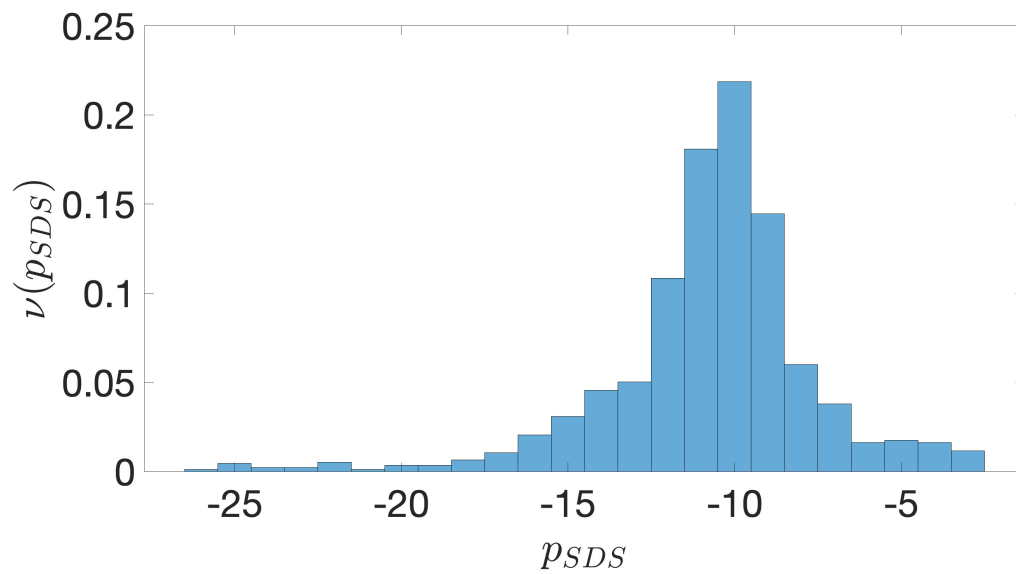


Figure 2.26: *P. haloplanktis*. The frequency  $\nu(p_{SDS})$  of the position  $p_{SDS}$  of the SDSs upstream the TSC.

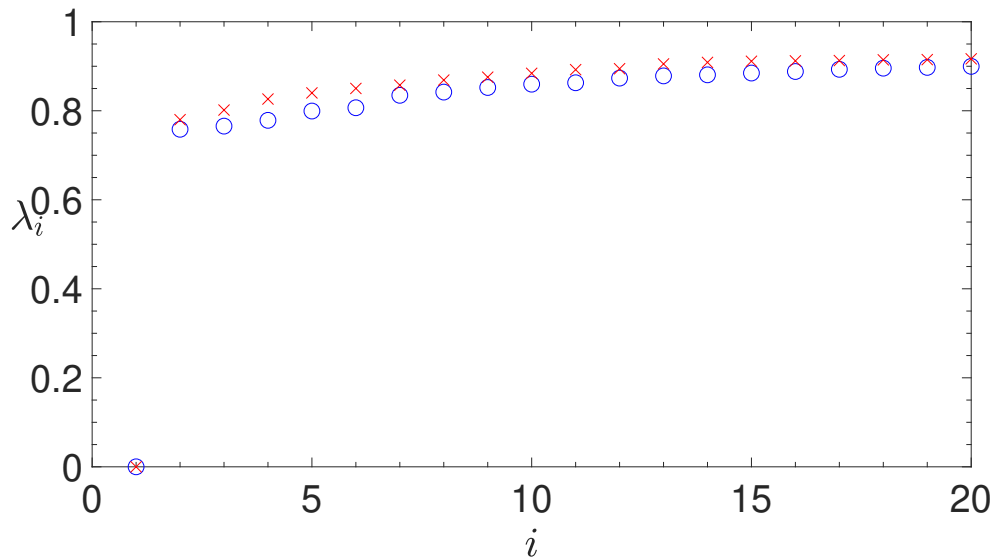


Figure 2.27: *P. haloplanktis*. The first twenty eigenvalues in ascending order of the normalized Laplacian matrix obtained by the alignment of the IGSs. Red crosses and blue circles correspond to different values of the similarity threshold, determined by the two statistical approaches described in subsection 2.2.2. A better discrimination of the three main eigenvalues is obtained for a higher similarity threshold (blue crosses), which corresponds to the second statistical approach, better suited for short sequences, as the annotated IGSs.

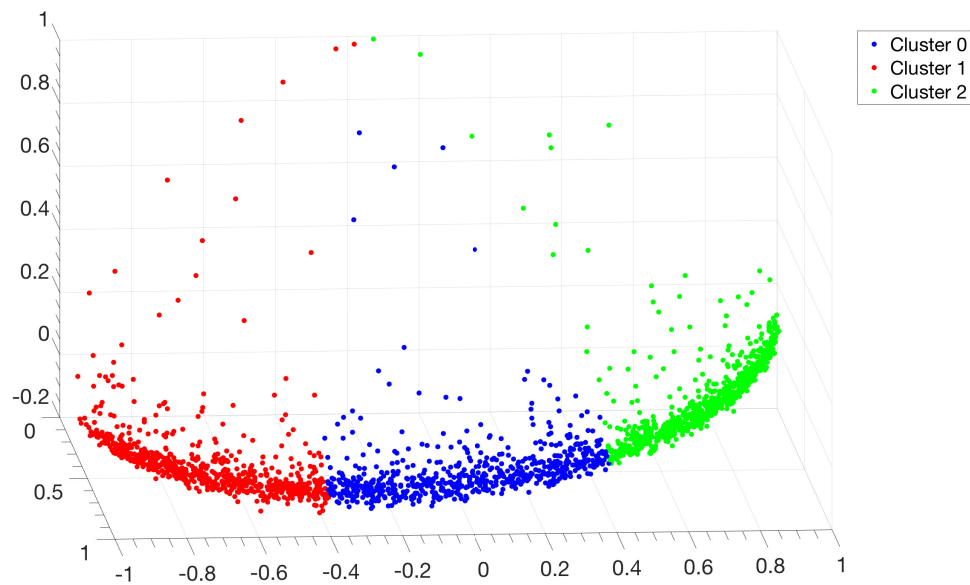


Figure 2.28: *P. haloplanktis*. Distribution of points in the clustering space relative to the alignment of the IGSs. Each point represents an IGS and the color code corresponds to the three clusters identified by the *Clustering Algorithm* described in subsection 2.2.3.

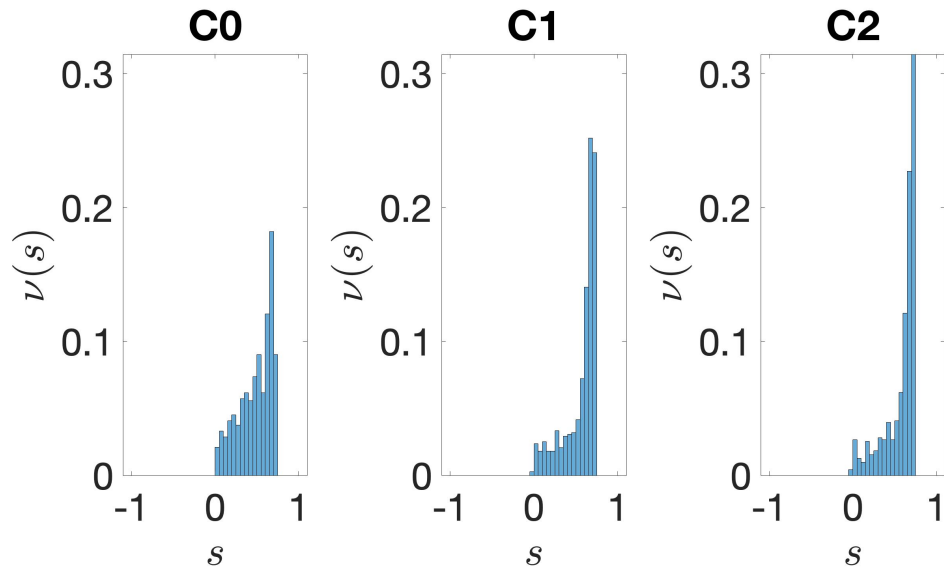


Figure 2.29: *P. haloplanktis*. Distribution of silhouette values relative to the clustering of the IGSs. On the vertical axis we report the frequency  $\nu(s)$  of IGSs versus the silhouette value  $s$ ; this value is between -1 and +1. The average values are 0.48 for cluster C0, 0.56 for C1 and 0.58 for C2.

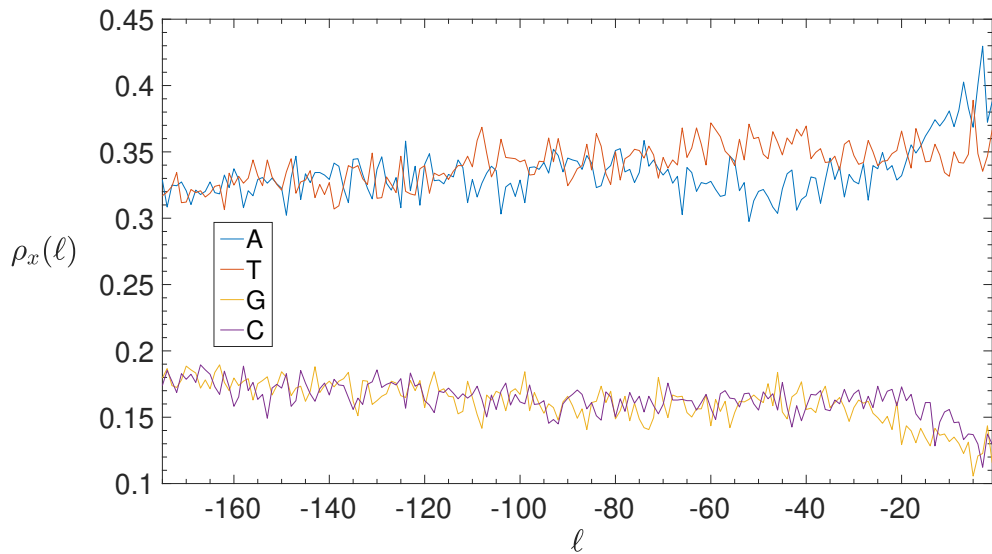


Figure 2.30: *P. haloplanktis*. BCA of the IGSs: on the vertical axis we report the density  $\rho_x(\ell)$  (see section 2.3) of each of the four nucleotides  $x = A$  (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the annotated 2091 IGSs.

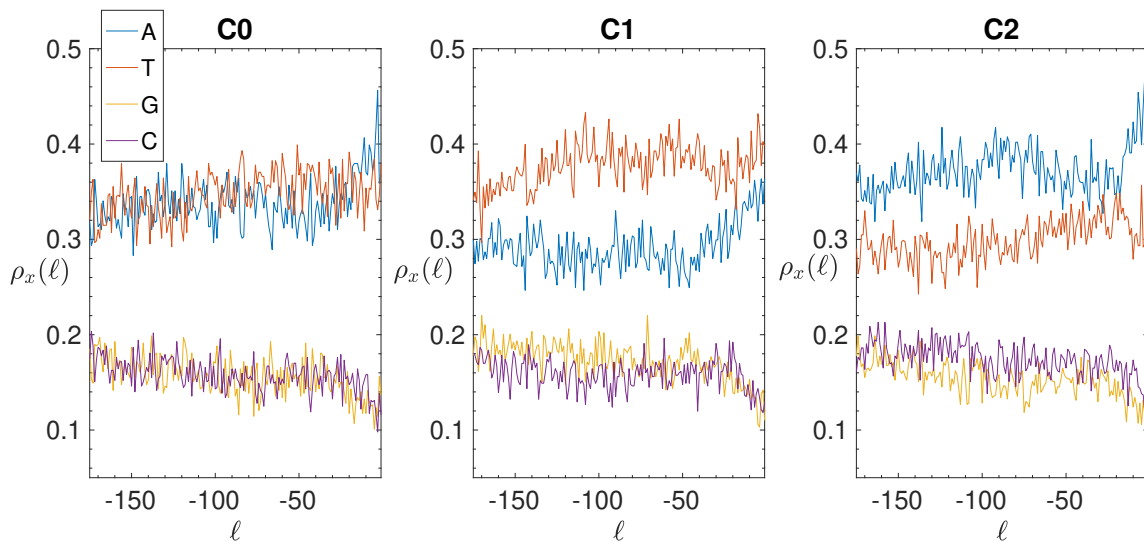


Figure 2.31: *P. haloplanktis*. Base composition analysis in the clusters of the IGSs: on the vertical axis we report the density  $\rho_x(\ell)$  (see section 2.3) of each of the four nucleotides  $x = A$  (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the IGS belonging to the clusters C0 (left panel), C1 (central panel) and C2 (right panel).

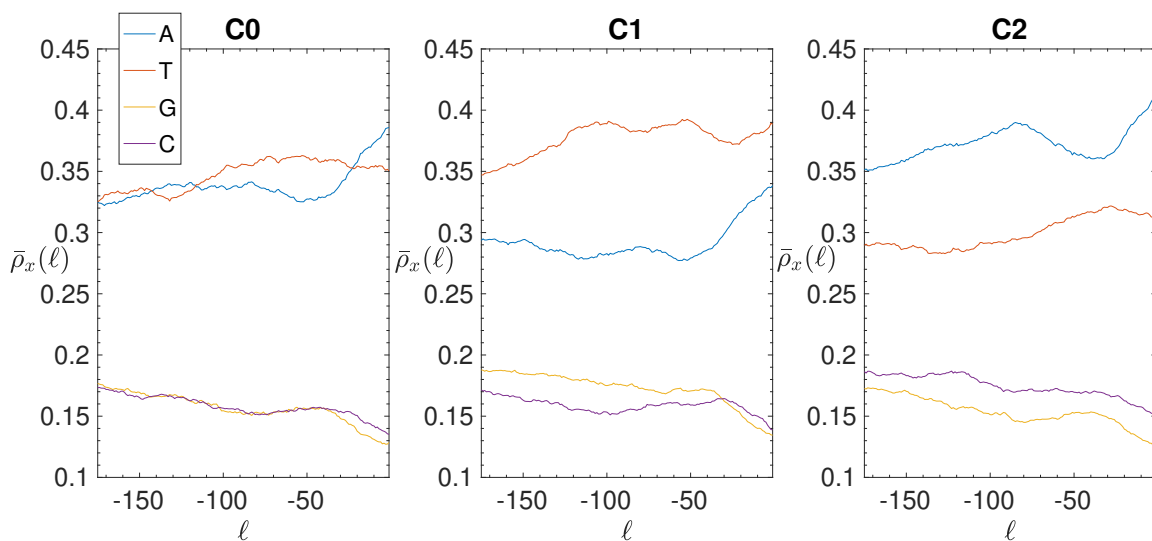


Figure 2.32: *P. haloplanktis*. Smoothed Base composition analysis in the clusters of the IGSs: on the vertical axis we report the averaged density  $\bar{\rho}_x(\ell)$  for  $a = 15$  bps (see section 2.3) of each of the four nucleotides A (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the IGSs belonging to the clusters C0 (left panel), C1 (central panel) and C2 (right panel).

## 2.5 Resume of the method

In summary, the proposed clustering method, summarized in Figure 2.33, allows us to detect specific similarities among IGSs associated also to relatively short regular subsequences. This method can be sketched in five steps. As shown in section 2.4 these features

are conserved in other bacterial species.

- 1- First we identify the IGRs, all the noncoding portions that are upstream the TSCs annotated in the genome considered, including also the reverse complement of the genes on the opposite strand.
- 2- Starting from the IGRs we can build the set of the RIGRs by selecting only the non-coding part between the TSC of a gene and the end of the previous one, regardless of the strand where is located. With the help of an operon database only those that precede the TSC of a transcriptional unit (single gene or operon) are selected. We annotate the length in term of bps of each RIGR. Calculating the distribution of the lengths, mean and standard deviation, we can choose a common length for all the RIGRS.
- 3- The presence of the SDSs is useful to detect approximately the position of the TSSs so as to eliminate the transcribed and not translated part for each sequence.
- 4- These “cleaned up” noncoding sequences, the IGSs, can be compared using alignment algorithms that provide a similarity score between them.
- 5- Similarity matrix containing these scores is processed with a clustering algorithm and the IGSs are divided into clusters based on compositional similarities.

Finally, it’s possible “to interweave” the information contained in each cluster with the ones associated with biological-type databases in order to check if they are expression of functional characteristics, as we can see in the next chapter.

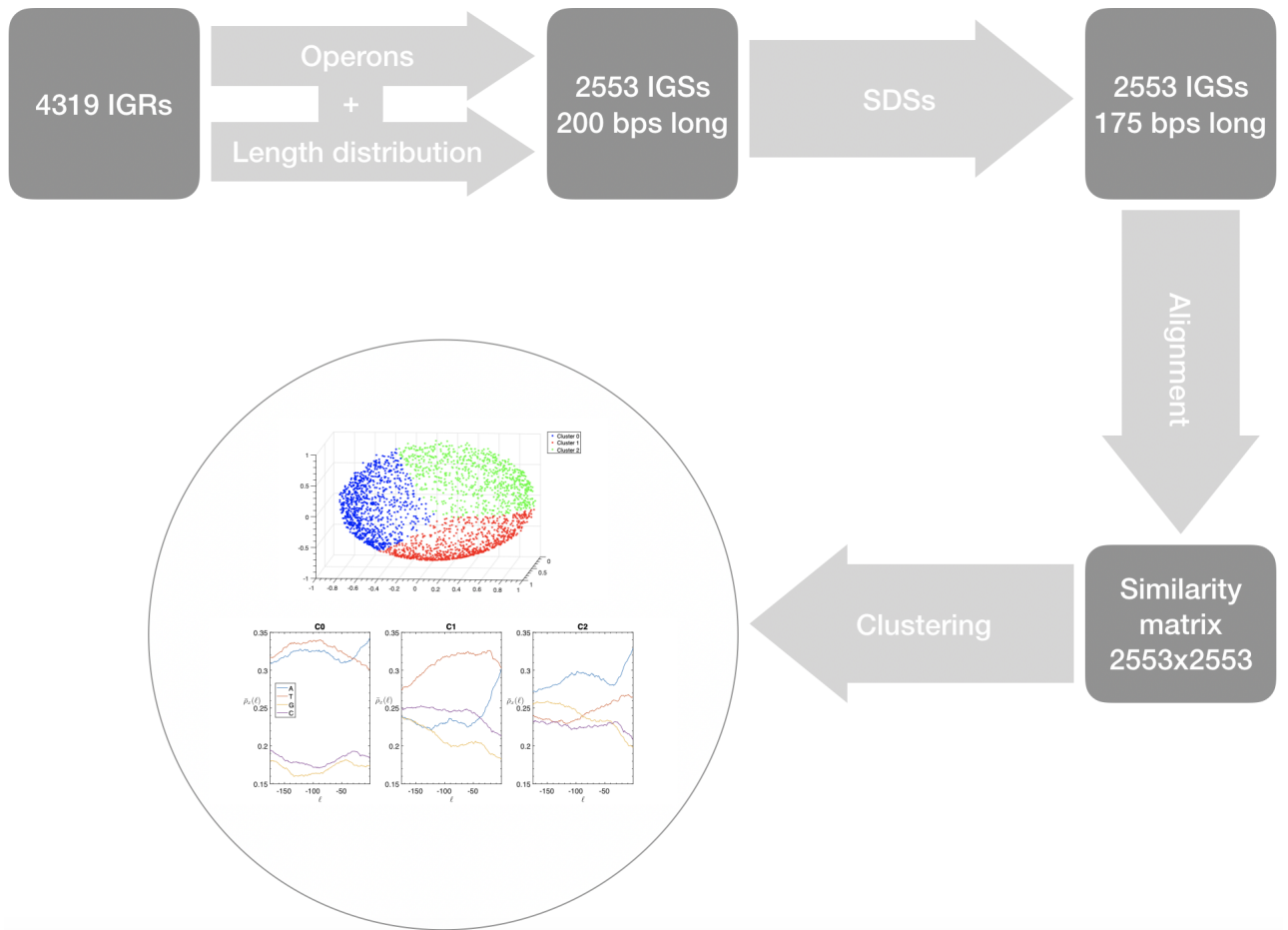


Figure 2.33: Workflow of the clustering procedure described in this chapter.

## Chapter 3

# Shared compositional and functional features of clustered intergenic sequences

It is well known that noncoding regions of DNA contain important functional elements, that mainly concern regulatory activities and changes in gene expression. Such functional elements have been identified as the participation in reproducible biochemical events, for instance Transcription Factor (TF) association, chromatin structure- or histone-modification [7]. Moreover, noncoding DNA is expected to play the role of a major substrate for critical changes, driving phenotypic modifications and differences between species or individuals, thus representing the basis for evolution as well as for disease-associated regulatory variants [8, 23, 24, 25]. In particular, the genetic programming of complex eukaryotes appears to be significantly correlated to the variability of noncoding DNA [26, 27]. Accordingly, considerable efforts have been devoted by several research groups to the study of noncoding DNA regions, mainly in eukaryotes. Traditional *in silico* approaches are based on comparative genomics, that relies upon evolutionary conservation as a property for identifying functional regions. For instance, pairwise or multiple sequence alignments have been used for predicting noncoding RNA transcripts or TF binding sites [28, 29, 30, 31, 32, 33]. By comparing genomic DNA from closely and distantly related species, functional elements may be recognized on the basis of their conservation. Comparative analyses can be applied also within a species to find paralogous regions deriving from duplication events within a genome [34] or even function-related patterns based on sequence similarities [35]. These sequence-based analyses, together with experimental techniques [36, 37, 38], have



proved quite effective for predicting functional noncoding sequences and their biological implications [39]. On the other hand, as a consequence of the variability of regulatory regions, it is quite difficult to establish the accuracy of such methods in estimating the TF binding or the transcriptional output [40, 41]. In fact, it is well known that, at variance with coding sequences that are well conserved even across distantly related species, regulatory regions are relatively flexible, since most TFs tolerate considerable variations in target sequences [42]. The high turnover rate both in adjacent putatively non-functional DNA and in duplicated TF binding sites often disrupts sequence conservation and makes alignments impossible (e.g., see [43, 44, 45]). Moreover, transcriptional rewiring [46] may explain events of sequence similarity loss, but retention of similar function. Accordingly, in noncoding DNA, sequence similarity may not necessarily correspond to functional homology.

For all these reasons the comparative approach among specific sequence elements in the noncoding regions of DNA is certainly useful, but insufficient to obtain an exhaustive description of DNA double helix functional properties. Many other approaches have been proposed to fill the gap. Among them we just mention the various techniques that run motif-finding algorithms on sets of sequences and incorporate the information of experimentally known TF binding sites in position-specific weight matrices [47, 48, 49], or rely on the study of the three-dimensional structure of DNA [50, 51] and on neural network optimization procedures [52, 53]. For instance, more recently other methods or databases aiming at identifying prokaryotic promoters have been proposed [54, 55].

Along this direction, a structure-based genome-wide analysis of the eukaryotic promoters was proposed as a new approach to a comprehensive identification of the correlations between the structural properties of promoter sequences and the kind of genes they regulate [56, 35]. In particular, BCA and specific entropic indicators were employed for identifying structural similarities among different classes of promoters [57, 58]. Moreover, the region around the TSS was shown to exhibit a very distinctive structural profile, which seems to be actively maintained by non-neutral selective constraints. Such structural profile is primarily related to a non-random distribution of nucleotides along the promoter close to the TSS [56, 35]. This kind of approach has been further refined in [1], where it was found that promoter sequences in *Homo sapiens*, can be classified into three main groups: two of them are distinguished by the prevalence of weak or strong nucleotides and are characterized by short compositionally biased sequences, while the most frequent regular sequences in the third group are strongly correlated with transposons. Moreover,

the comparison of the promoter database of *H. sapiens* with those of other species indicates that structural complexity characterizes also the evolutionary content appearing in mammalian promoters, at variance with ancestral species in the phylogenetic tree, that exhibit a definitely lower level of differentiation among promoters. This notwithstanding, evolutionary selection of regulatory DNA sequences is at work in all organisms [59, 60, 61] and it is reasonable to expect that in prokaryotes also a genome-wide approach can be effective in identifying possible correlations between structure and regulation/expression of genes.

In this chapter we develop a computational method based on the compositional analysis of bacterial IGSs to analyze the structure of noncoding sequences close to the TSS in various bacterial species, while searching for possible correlations with the expression, regulation and biological functions of the genes they correspond to. The overall strategy of this approach is illustrated making use of *E. coli*, as a primary case study (although the method has been applied to other prokaryotes, mostly leading to overlapping results).

## 3.1 Results for *E. coli*

### 3.1.1 Structural features of clusters

The complex structure of nucleotide sequences in the IGSs considered in this thesis is essentially due to the presence of some regular patterns, that allows for a structural clustering. For instance, although obviously regulatory motifs exist also in bacteria, they are much less complex (typically, homogeneous sequences) and much shorter than in eukaryotes [40, 62, 63]. In Figure 2.10 we report the BCA of all the 2553 IGSs of *E. coli*: it has been obtained by measuring the positional density of nucleotides along both strands. The first feature that emerges is the well-known dominance of weak nucleotides (A and T) with respect to strong ones (G and C). Only close to the null position one observes peculiar peaks, corresponding to the typical enrichment of purines close to the TSS. This is an indication that our selection procedure of IGSs consistently identifies such known enrichment [64].

If one subdivides the IGSs into the three clusters represented in the clustering space shown in Figure 2.8, one obtains the smoothed BCA (see (2.6) in section 2.3, where is reported also the original BCA) reported in the panels of Figure 2.12. Cluster C0 is quite similar to the total BCA, although the separation between weak and strong nucleotides is

amplified. Cluster C1 is characterized by the dominance of T nucleotides, the depression of G nucleotides, while A and C nucleotides exhibit a similar intermediate dependence on the position. Finally, Cluster C2 shows similar trends with respect to C1, with weak and strong nucleotides exchanging their role between themselves. *A posteriori* we can conclude that the clustering procedure is effective in identifying differences and similarities among the annotated IGSs. Anyway, at variance with eukaryotes, the noncoding regions of bacterial species exhibit a definitely lower level of complexity.

A more careful inspection of the BCA analysis of *E. coli* indicates that the structural differences among the three clusters is associated to the presence of regular motifs of weak nucleotides, like homogeneous patches of A and T, or period-2 sequences made of AT pairs. In fact, we have found that C1 and C2 contain IGSs that are typically enriched by homogeneous segments of T and A nucleotides, respectively. These segments extend over a few to some tens of nucleotides, while their most frequent length (as observed also in eukaryotes [1]) is close to six nucleotides. For instance, the number of homogeneous T-segments of length equal or larger than 6 nucleotides in the IGSs of C1 is approximately four times larger than homogeneous A-segments and AT-segments. Similarly, in C2 homogeneous A-segments occur twice with respect to homogeneous T-segments and five times more than AT-segments. Conversely, in C0 there are more regular A-, T- and AT-segments than those found in the other clusters and their absolute numbers are comparable (238, 292 and 185, respectively). Actually, we have also found that there is a sort of symmetry between the IGSs in C1 and C2, where homogeneous segments of weak nucleotides of an IGS in one strand appear as reverse complements in another IGS on the opposite strand. Such homogenous motifs have been recognized as typical sequences, favoring the diffusion of transcription factors along the DNA chain in search of the TSS [65].

Anyway, we are aware that the annotated IGSs of *E. coli* contain coding portions. More precisely, only 1356 IGSs do not contain any coding portion and half of them are found to belong to C0, while the remaining IGSs are approximately equally shared between C1 and C2. The average length of noncoding portions in the remaining 1197 IGSs is 102 bps and again they are almost equally shared in C1 and C2, while only 262 are contained in C0. This figures indicate that our clustering analysis is certainly influenced by the presence of coding portions, despite they play a minor role with respect to noncoding ones. On the other hand, there is not a sharp correspondence between the content of coding portions and the IGSs contained in the three clusters and we can conclude that

the statistical significance of our clustering analysis is sufficient for identifying structural differences and similarities among the annotated IGSs.

### 3.1.2 Correlations between clustering and biological features

Once grouped IGSs into clusters the further step was to investigate possible correlations inside each cluster with biological properties. This task has been accomplished making use of the STRING database [66], which provides us information about various features related to the interactions in genetic networks. In particular, we have focused our analysis on genetic co-expression and co-occurrence in *E. coli*: details about the content of biological information associated to such features and the way they are quantified by a score is shortly discussed in subsection 3.3.1.

We have considered all genes and operons associated with the IGSs belonging to a cluster and we have constructed the corresponding genetic network, whose nodes represent single genes as well as genes belonging to an operon (we indicate with  $N_{\text{genes}}$  the total number of nodes of the network and with  $N_{\text{IGS}}$  the number of IGSs in a cluster). A network link is established between two genes if the corresponding element in the matrix determined by the score of the STRING algorithm overtakes a threshold value, that we have fixed to 0.7, in order to obtain a sparse matrix with a high level of "affinity" between pairs of connected nodes (see subsection 3.3.1). Then, we have computed the size,  $N_{\text{LCC}}$ , and the total numbers of links per node,  $N_{\text{link}}$ , of the largest connected component (LCC) of the network. The results are reported in Table 3.1 and Table 3.2, together with the average values ( $\overline{N}_{\text{LCC}}$  and  $\overline{N}_{\text{link}}$ ) and the variances ( $\sigma_{\text{LCC}}$  and  $\sigma_{\text{link}}$ ) of the same quantities, obtained by averaging over a 1000 random samplings of the IGSs (and of the corresponding genes) in the networks, built up by grouping the same total number of IGSs in each cluster. The values obtained by our clustering method correspond to values of the co-expression and co-occurrence indicators, that are typically close to, or just beyond, the border of the variance range.

Table 3.1: Co-expression networks for *E. coli*. We compare the features of the co-expression networks for each cluster between the three clusters obtained by the clustering method and other three clusters obtained by averaging over a 1000 random samplings of the IGSs.

	$N_{\text{IGS}}$	$N_{\text{genes}}$	$N_{\text{LCC}}$	$\bar{N}_{\text{LCC}}$	$\sigma_{\text{LCC}}$	$\frac{N_{\text{LCC}} - \bar{N}_{\text{LCC}}}{\sigma_{\text{LCC}}}$	$N_{\text{link}}$	$\bar{N}_{\text{link}}$	$\sigma_{\text{link}}$	$\frac{N_{\text{link}} - \bar{N}_{\text{link}}}{\sigma_{\text{link}}}$
C0	930	1543	32	53.6	21.4	-1.01	82	341.9	238.8	-1.09
C1	812	1451	62	43.3	17.6	1.06	707	269.4	194.0	2.26
C2	811	1325	59	42.7	17.8	0.92	261	263.8	197.2	-0.01

Table 3.2: Co-occurrence networks for *E. coli*. We compare the features of the co-occurrence networks for each cluster between the three clusters obtained by the clustering method and other three clusters obtained by averaging over a 1000 random samplings of the IGSs.

	$N_{\text{IGS}}$	$N_{\text{genes}}$	$N_{\text{LCC}}$	$\bar{N}_{\text{LCC}}$	$\sigma_{\text{LCC}}$	$\frac{N_{\text{LCC}} - \bar{N}_{\text{LCC}}}{\sigma_{\text{LCC}}}$	$N_{\text{link}}$	$\bar{N}_{\text{link}}$	$\sigma_{\text{link}}$	$\frac{N_{\text{link}} - \bar{N}_{\text{link}}}{\sigma_{\text{link}}}$
C0	930	1543	338	267.7	35.4	1.99	1022	665.5	123.2	2.89
C1	812	1451	146	214.4	35.4	-1.93	326	503.8	110.5	-1.61
C2	811	1325	179	213.8	35.2	-0.99	447	497.1	111.1	-0.45

Hence, clustering IGSs by structural similarity suggests the existence of a correlation with co-expression and co-occurrence. Establishing more precise relations, if any, with specific motifs appearing in the IGSs belonging to each cluster demands a deeper inspection about the mechanisms associated to gene expression and regulation. However, this issue is beyond the aims of this thesis.

In order to understand whether genes belonging to a specific biological function were over-represented in any of the identified clusters, we performed a functional enrichment analysis using COG categories and evaluating statistical significance (if any) using a negative binomial test. Data obtained for *E. coli* are shown in Figure 3.1, whereas results for *B. subtilis* and *P. haloplanktis* are reported in section 3.2 (see Figure 3.2 and Figure 3.3).

Overall we observed a few enriched functional categories for each of the *E. coli* clusters. In particular, 3 COG functional categories were found to be over- and down-represented in C0, respectively. The first set included genes involved in the transport and metabolism of inorganic ion, in the production and conversion of energy as well as genes lacking a

functional annotation. The second set included genes involved in information processing (translation, ribosomal structure and biogenesis), coenzyme and nucleotide metabolism. The other two clusters included over-represented categories (cell membrane biogenesis and metabolism and information processing of C1 and coenzyme metabolism of C2) as well as down-represented ones (inorganic ion transport and energy production and metabolism of C2). Despite only a few cellular processes displayed a significant trend in the clustering of IGSs, it is worth noticing that some functional categories fall in more than one cluster, but always with an opposite trend, according to the functional enrichment analysis. For instance, this is the case of COG J over-represented in C1 and down-represented in C0, COG P and C, over-represented in C0 and down-represented in C2 and COG H over-represented in C2 and down-represented in C0. It might be tempting to speculate that this peculiar distribution of genes belonging to the same process category among the identified clusters and due to the structure of their IGSs could reflect differences in the regulatory features of the corresponding genes. Additional analyses/experiments will be needed to evaluate the robustness of this association.

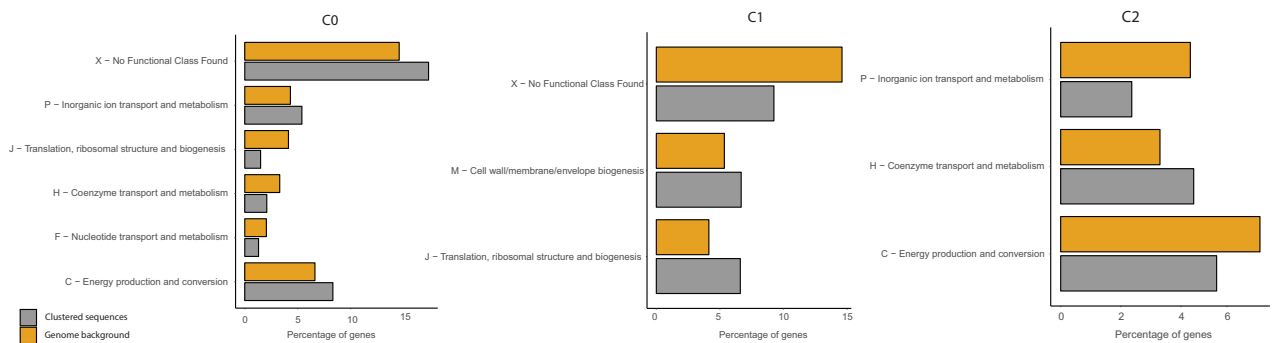


Figure 3.1: COG functional enrichment analysis of clustered genes in *E. coli*. We report the significantly enriched or depleted COG functional categories belonging to each of the identified clusters (C0, C1, C2) in respect to the genome background. *Clustered Sequences* refers to the functional annotation of the sequences that were clustered according to our method (i.e. after the analysis of IGSs) in each of the three clusters. *Genome background* refers to the functional annotation of the entire genome (i.e. of each gene of the considered organism).

## 3.2 Other bacterial species

Here we insert all information concerning other bacterial species: *B. subtilis* and *P. haloplanktis*. Comparing Table 3.1 with Table 3.3, Table 3.5 and Table 3.2 with Table 3.4, Table 3.6 what stands out immediately is the smaller number of clusters for which the co-expression and co-occurrence networks are statistically significant. This can be at-

tributed to the lower amount of data available for these two bacteria. In particular for *P. haloplanktis* the cluster C1 seems to be the only statistically relevant.

Table 3.3: Co-expression networks in *B. subtilis*. We compare the features of the co-expression networks for each cluster between the three clusters obtained by the clustering method and other three clusters obtained by averaging over a 1000 random samplings of the IGSs.

	$N_{\text{IGS}}$	$N_{\text{genes}}$	$N_{\text{LCC}}$	$\bar{N}_{\text{LCC}}$	$\sigma_{\text{LCC}}$	$\frac{N_{\text{LCC}} - \bar{N}_{\text{LCC}}}{\sigma_{\text{LCC}}}$	$N_{\text{link}}$	$\bar{N}_{\text{link}}$	$\sigma_{\text{link}}$	$\frac{N_{\text{link}} - \bar{N}_{\text{link}}}{\sigma_{\text{link}}}$
C0	749	1341	25	25.7	14.4	-0.05	63	169.6	189.7	-0.56
C1	856	1346	41	30.6	15.6	0.67	417	199.1	201.2	1.08
C2	884	1487	16	32.3	16.5	-0.99	28	217.5	212.2	-0.89

Table 3.4: Co-occurrence networks in *B. subtilis*. We compare the features of the co-occurrence networks for each cluster between the three clusters obtained by the clustering method and other three clusters obtained by averaging over a 1000 random samplings of the IGSs.

	$N_{\text{IGS}}$	$N_{\text{genes}}$	$N_{\text{LCC}}$	$\bar{N}_{\text{LCC}}$	$\sigma_{\text{LCC}}$	$\frac{N_{\text{LCC}} - \bar{N}_{\text{LCC}}}{\sigma_{\text{LCC}}}$	$N_{\text{link}}$	$\bar{N}_{\text{link}}$	$\sigma_{\text{link}}$	$\frac{N_{\text{link}} - \bar{N}_{\text{link}}}{\sigma_{\text{link}}}$
C0	749	1341	73	50.2	16.6	1.38	162	119.8	60.9	-0.69
C1	856	1346	71	63.1	17.4	0.45	198	158.3	77.0	0.52
C2	884	1487	38	67.4	18.4	-1.60	85	170.1	84.2	-1.01

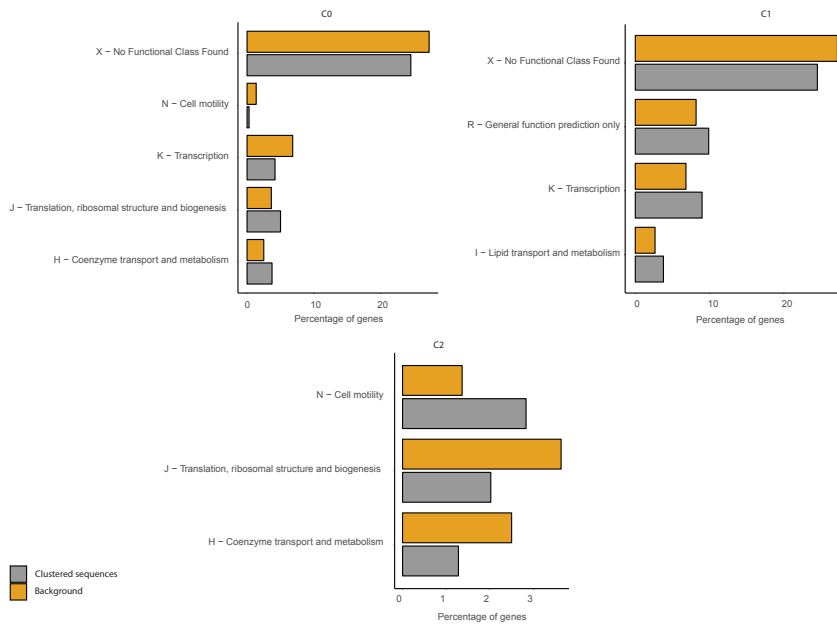


Figure 3.2: COG functional enrichment analysis of clustered genes in *B. subtilis*. We report the significantly enriched or depleted COG functional categories belonging to each of the identified clusters (C0, C1, C2) in respect to the genome background. *Clustered Sequences* refers to the functional annotation of the sequences that were clustered according to our method (i.e. after the analysis of IGSs) in each of the three clusters. *Genome background* refers to the functional annotation of the entire genome (i.e. of each gene of the organism considered).

Table 3.5: Co-expression networks in *P. haloplanktis*. We compare the features of the co-expression networks for each cluster between the three clusters obtained by the clustering method and other three clusters obtained by averaging over a 1000 random samplings of the IGSs.

	$N_{\text{IGS}}$	$N_{\text{genes}}$	$N_{\text{LCC}}$	$\bar{N}_{\text{LCC}}$	$\sigma_{\text{LCC}}$	$\frac{N_{\text{LCC}} - \bar{N}_{\text{LCC}}}{\sigma_{\text{LCC}}}$	$N_{\text{link}}$	$\bar{N}_{\text{link}}$	$\sigma_{\text{link}}$	$\frac{N_{\text{link}} - \bar{N}_{\text{link}}}{\sigma_{\text{link}}}$
C0	664	1074	22	23.0	12.0	-0.08	36	102.7	108.8	-0.61
C1	718	1182	47	26.8	13.8	1.46	305	124.6	128.4	1.40
C2	709	1079	24	26.1	12.0	-0.16	42	122.1	121.6	-0.66



Table 3.6: Co-occurrence networks in *P. haloplanktis*. We compare the features of the co-occurrence networks for each cluster between the three clusters obtained by the clustering method and other three clusters obtained by averaging over a 1000 random samplings of the IGSs.

	$N_{\text{IGS}}$	$N_{\text{genes}}$	$N_{\text{LCC}}$	$\bar{N}_{\text{LCC}}$	$\sigma_{\text{LCC}}$	$\frac{N_{\text{LCC}} - \bar{N}_{\text{LCC}}}{\sigma_{\text{LCC}}}$	$N_{\text{link}}$	$\bar{N}_{\text{link}}$	$\sigma_{\text{link}}$	$\frac{N_{\text{link}} - \bar{N}_{\text{link}}}{\sigma_{\text{link}}}$
C0	664	1074	58	45.3	21.2	0.60	129	104.8	71.7	0.34
C1	718	1182	89	51.4	21.5	1.75	441	120.3	77.1	4.16
C2	709	1079	36	51.8	22.5	-0.70	76	122.6	86.2	-0.54

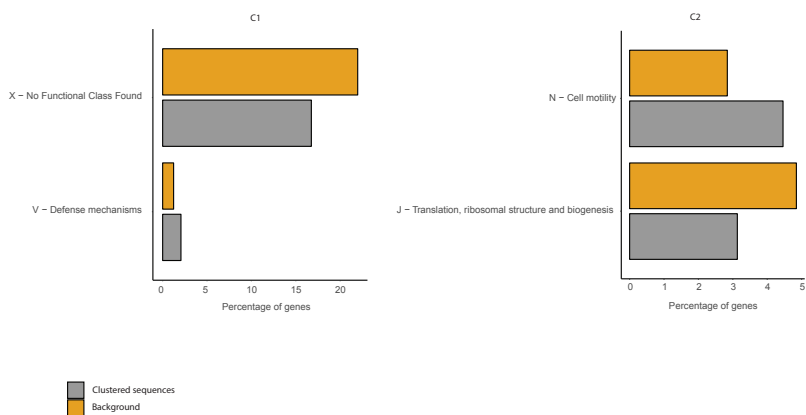


Figure 3.3: COG functional enrichment analysis of clustered genes in *P. haloplanktis*. We report the significantly enriched or depleted COG functional categories belonging to each of the identified clusters (C0, C1, C2) in respect to the genome background. *Clustered Sequences* refers to the functional annotation of the sequences that were clustered according to our method (i.e. after the analysis of IGSs) in each of the three clusters. *Genome background* refers to the functional annotation of the entire genome (i.e. of each gene of the organism considered).

The scarcity of data also concerns the COG database of *B. subtilis* and *P. haloplanktis* (the cluster C0 in the last bacterium doesn't present any enriched functional category). Nevertheless, we have an evident similarity concerning the COG functional enrichment analysis of clustered genes in *B. subtilis* and *P. haloplanktis*, and in particular the ones in cluster C2 for both. In fact, we can see from the last panel on the right in Figure 3.2 and Figure 3.3 that COG N (cell motility) is over-represented, while COG J (translation, ribosomal structure and biogenesis) is down-represented.

## 3.3 Materials and Methods

### 3.3.1 About STRING

STRING (*Search Tool for the Retrieval of Interacting Genes*) is a database of known and predicted protein-protein interactions (PPI). The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases. Thanks to this database we can build the network of predicted associations for a particular group of genes (or proteins). The network nodes are the genes. The weighted edges represent the predicted functional associations. In fact, STRING provides a score for each protein-protein association. The scores take values from zero to one and indicate the estimated likelihood that a given interaction is biologically meaningful, specific and reproducible, given the supporting evidence. There are seven *evidence channels* that together contribute to providing the total evidence, that is the protein-protein association score.

We have analyzed the channels separately. In particular we focused on the co-expression and co-occurrence channels. (i) The co-expression channel: for this channel gene expression data originating from a variety of expression experiments are normalized, pruned and then correlated. Pairs of proteins that are constantly similar in their expression patterns, under different conditions, will receive a high interaction score. (ii) The co-occurrence channel: in this channel, STRING calculates the phylogenetic distribution of the orthologs of all proteins in a given organism. If two proteins show a high similarity in this distribution, i.e. if their orthologs tend to be present or absent in the same subsets of organisms, then an association score is assigned.

#### Co-expression

During the course of its life, a bacterium must regulate the type, quantity and activity of proteins and other macromolecules they produce. After the transcription of DNA in RNA, information is translated for the production of specific proteins. These processes are known by the collective name of gene expression. From gene expression profiles under different experimental conditions, we go to check which pairs of genes exhibit the same behavior, since the transcription levels of two co-expressed genes rise and lower simultaneously. Apart from these direct evidences, STRING also takes into account properties

transferred from one organism to another. This so-called “interolog” transfer is based on the observation that co-expressed orthologs in an organism often are co-expressed also in another (this inference is the more confident the better orthology relationships that can be established). However, mostly poorly studied organisms benefit of the transfers of interaction between genes, where the fraction of interactions supported by transfers from other organisms can even reach 99%. On the other hand, in well-studied model organisms as *E. coli*, the corresponding fraction is below 20%. In Figure 3.4 it is reported a co-expression pattern for a set of genes present in *E. coli*.

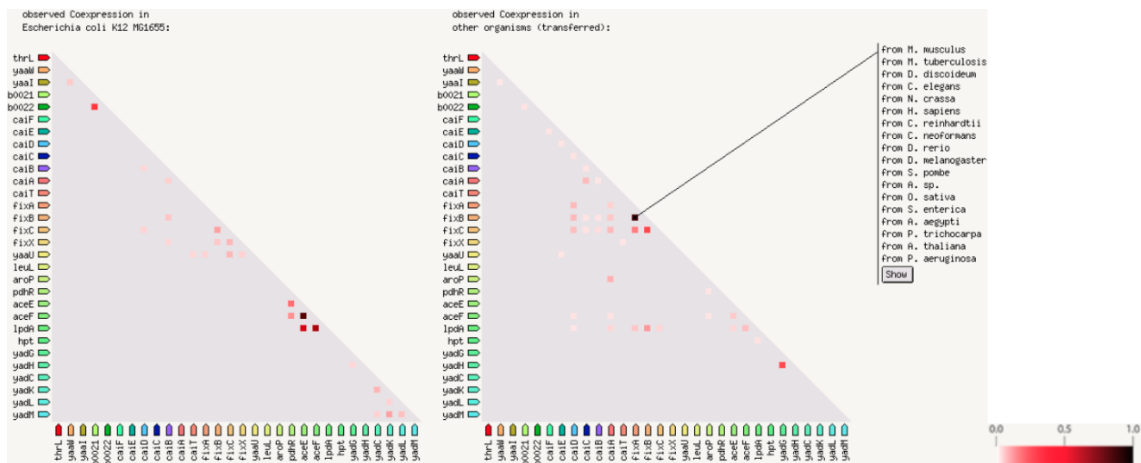


Figure 3.4: In the triangular matrices above, the intensity of the color indicates the confidence level with which two proteins are functionally associated, considering the total expression data in the organism. On the left are reported the data found directly for *E. coli*, on the right those relating to the same orthologs present in other organisms.

Thanks to the one-to-one correspondence between one IGS and the corresponding transcriptional unit (formed by a gene or an operon), from the clustering we can build a co-expression network, where the nodes are the genes and the weighted edges are given by the PPI score of the co-expression channel. We choose a threshold for the score, below which, the link is deleted, above it is arbitrarily set to one. We analyzed the characteristic properties of this undirected and unweighted graph based on gene co-expression.

### Co-occurrence

We adopted the same procedure for the co-occurrence channel. Because of the so-called divergent evolution, it may be that proteins or genes that share a common origin may have reached very different structures and functions. The genes, in fact, can undergo an independent evolution, to give rise to the so-called paralogous genes, that is, genes present within the same genome that code for different products but originate from a single

ancestral gene. The co-occurrence concerns the so-called orthologous genes, which are instead homologous genes, present in different but related species, which code for proteins with similar structures and functions. Figure 3.5 shows how STRING has organized this data.

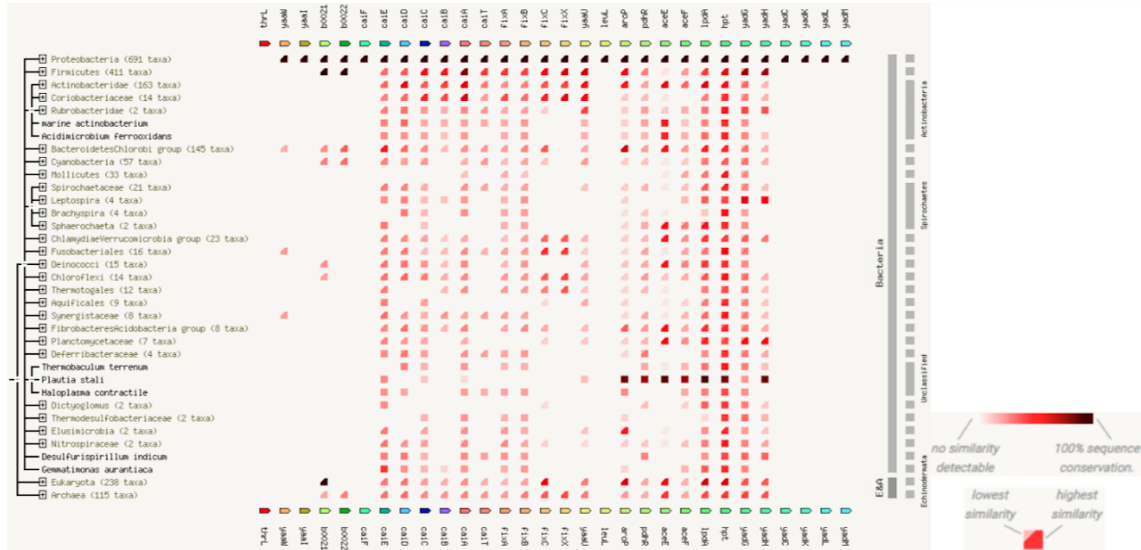


Figure 3.5: Below and above this diagram, the names of some genes are reported, on the left the phylogenetic tree. The colored squares indicate the presence of the gene in the correspondent clade. The color denotes, for the gene of interest, the similarity of the nucleotide sequence with that one of a given genome contained in STRING. For groups of genomes that have collapsed into the phylogenetic tree, two distinct colors indicate the lowest and the highest similarity observed in that clade.

At the base there is a phylogenetic tree built through cladistic methods: the living are classified on the basis of the degree of kinship, or on the distance in time of the last common ancestor; organisms are divided in clades, i.e. groups having a common ancestor and all its descendants. For each gene of a given organism we check in what other organisms it is present and with what similarity. In this channel, STRING evaluates the phylogenetic distribution of orthologs of all genes in a given organism. If two genes show a high similarity in this distribution, that is their orthologs are present or absent roughly in the same subsets of organism, then a high score of the co-occurrence channel is assigned.

### 3.3.2 COG categories enrichment

To conduct functional enrichment, each gene whose upstream intergenic region was clustered in one of the three clusters was assigned to a specific functional category using a BLAST [67] search against the COG database [68], with default parameters and considering a hit as significant if E-value  $< 1e^{-20}$ . The exact binomial test implemented in the R

package [69] was used to assess over- and under-represented functional categories against the corresponding genomic background. The Blast2Go package [70] was used to assess over- and under-represented GO terms in each cluster.

### **3.4 Comparison with other methods identifying biological functions of IGSs**

Now that we have illustrated our method for classifying IGSs a priori on the basis of structural properties and a posteriori on the one of biological functionalities, it makes sense to compare it with other similar methods, in particular the ones already mentioned in [28, 29, 30, 31, 32, 33], to highlight the differences. In [28, 29, 30] noncoding conserved sequences are taken into consideration. In [28] it is shown that conserved noncoding segments contain an enrichment of transcription factor binding sites, when compared to the sequence background in which the conserved segments are located and that this enrichment of binding sites was not observed in coding sequences. Also the comparative sequence analysis executed in [29] for identifying sequences that are conserved across multiple species revealed substantial fraction of the bases within this sequences (approximately 70%) resides within noncoding regions. Initial characterization of these “Multi-species Conserved Sequences” has revealed sequences that correspond to clusters of transcription factor-binding sites, noncoding RNA transcripts, and other candidate functional elements. In [30] it is found that conserved noncoding sequences are significantly more conserved than protein-coding genes and noncoding RNAs within the mammalian class, from primates to monotremes to marsupials. The pattern of substitutions in conserved noncoding sequences differed from that seen in protein-coding and noncoding RNA genes and resembled that of protein-binding regions. A three-way multiple alignment between the genomes (human, mouse and rat) carried on in [31] to detect noncoding sequences is at the base of a graph theoretic clustering algorithm, akin to the highly successful methods used in elucidating protein sequence family relationships. The algorithm is applied to a highly filtered set of about 700000 human-rodent evolutionarily conserved regions, not resembling any known coding sequence. From these, roughly 12000 non-singleton clusters have been obtained, dense in significant sequence similarities. Reference [32] contains a method that can accurately identify pairs of functional noncoding orthologs at evolutionarily diverged loci by searching for conserved transcription factor binding

sites arrangements, detecting approximately 300 pairs of diverged elements that are likely to share common ancestry and have similar regulatory activity. It can be argued that transcription factor binding sites composition is often necessary to retain and sufficient to predict regulatory function in the absence of overt sequence conservation, revealing an entire class of functionally conserved, evolutionarily diverged regulatory elements. In [33] a comparative method for genome-wide identification of families of regulatory RNA structures had been proposed: it has been applied to a 41-way genomic vertebrate alignment in order to find regulatory RNA structures that are often members of families with multiple paralogous instances across the genome. Family members share functional and structural properties, which allow them to be studied as a whole, facilitating both bioinformatic and experimental characterization. Known families identified include both noncoding RNAs and cis-regulatory structures. They also identify tens of new families supported by strong evolutionary evidence and other statistical evidence, such as GO term enrichments. These findings exemplify the diversity of post-transcriptional regulation and provide a resource for further characterization of new regulatory mechanisms and families of noncoding RNAs.

In the light of the methods described above, ours differs significantly from all of them, because it aims at the identification of structural elements or properties inherent the whole set of IGSs inside a species and, then, at a comparison among different species. In particular, the three main identifying features of our method are listed hereafter.

- 1- The object of our research, the IGSs, are sequences of DNA upstream the TSS, charged with regulation at its very first step, since it is noncoding non-transcribed DNA (unlike noncoding RNA).
- 2- IGSs belonging to the same organism are considered and the structural similarities are identified among sequences upstream the TSS unambiguously determined by the identification procedure, regardless of whether they are conserved or not.
- 3- The IGSs are 175-bps long and the alignment procedure takes into consideration the whole sequence globally in its length without focusing specifically on the transcription factor binding sites allowing a correspondence between functional properties and large-scale structural features.

# Chapter 4

## Thermodynamics of DNA

### denaturation in a model of bacterial intergenic sequences

In the past decades DNA denaturation attracted the interest of various researchers, which introduced and studied statistical and dynamical models of this fundamental biological process. In a historical perspective one basic step in this direction was accomplished by the celebrated Lifson-Poland-Scheraga (LPS) model [71], where the mechanism of denaturation was described by a simplified stochastic dynamics of H-bonds breaking, driven by thermal fluctuations. The statistical mechanics of this model has been widely investigated, since the crucial feature of the LPS model is that the effective interaction is long-ranged, so that this 1D model may exhibit a true phase transition. The nature of this phase transition has been a longly debated problem. In fact, it depends on the long-range parameter, usually denoted by  $\alpha$ , and also on the presence of disorder as a manifestation of the random-like structure in the sequence of weak and strong bases forming the DNA double-strand structure. The present scenario can be summarized as follows. The homogeneous LPS-model exhibits a continuous phase transition for  $1 < \alpha < 2$ , while the transition turns to a first-order one for  $\alpha > 2$  [72]. When structural weak disorder is considered it has been rigorously proven that the phase transition is continuous for any value of  $\alpha$  [73]. This notwithstanding, even an approximate estimate of the values of the critical exponents in the disordered case for  $\alpha > 2$  is still an open problem (e.g., see [74]). Anyway, it has been shown that the LPS model identifies an universality class of wetting processes, that are equivalent to specific random walk processes in the upper half-plane

(e.g., see [73, 74]). The dynamics and statistics of various models of DNA denaturation, taking into account explicitly its 3D structure, have been also widely investigated by numerical simulations, e.g. see [75, 76, 77, 78, 79, 80, 81]. All of these studies have pointed out again the difficulty of a clear-cut identification of the order of the phase transition (e.g., see [82]). Experimental studies on DNA denaturation also seem to confirm the ambiguities inherent the identification of the denaturation phase transition (e.g., see [83, 84, 85]). One can argue that this is due to various factors affecting experimental studies, including the way experimental protocols are defined, finite size effects, etc. An alternative approach to DNA denaturation relies on a dynamical description, where this process can be represented as a crossover between different dynamical regimes of relatively short strands, rather than a true thermodynamic transition, which, by definition, stems from the singularities inherent the thermodynamic limit. The basic features of the LPS model has provided inspiration also for the definition of dynamical models. For instance, a stochastic dynamics of the LPS model has been analyzed in [86], pointing out that above the critical temperature the system undergoes a gelation phenomenon, i.e. relaxation to equilibrium in a finite time, due to the long-range nature of the interaction. Hamiltonian models of DNA denaturation, where the nature of both the stacking and the basis interactions forces acting in a double strand together with isotopic disorder effects are explicitly taken into account, have also been proposed in the past. Among these models the most popular is certainly the Peyrard-Bishop one [87] (see also [88]). The LPS approach has inspired further refinements of this model aiming at introducing a cooperative effect in bond denaturation by a suitable nonlinear contribution to the base-pair stacking interaction potential [89]. The importance of this ingredient is that, while previous models, like the Peyrard-Bishop one [87], required unrealistically weak stacking interactions to avoid a too high denaturation temperature, it increases significantly the entropy associated with melting and, accordingly, it lowers the denaturation temperature in such a way that more realistic stacking potential parameters can be used. The adequacy of this phenomenological model for describing the melting transition in short oligonucleotides made of a few tens of base-pairs (bps) has been checked in [90]. It is also important to mention that the presence of the melting transition in DNA is reproduced also by this class of Hamiltonian models on the basis of statistical mechanics arguments. In fact, despite the short-range nature of interactions in such 1D models, a true phase transition can be characterized by an entropic stabilization of a nonlinear field configuration, i.e. a soliton-like domain wall, with macroscopic energy content [91, 92]. This peculiar condensation phenomenon



is analogous to the gelation process outlined in [86] for the stochastic dynamics of the LPS model. A recent and complete review about the physical aspects concerning DNA dynamics is available in [93].

In this thesis we adopt the same model introduced in [89], together with the physical values of its phenomenological parameters. In particular, in this manuscript we want to study the phenomenon of denaturation in DNA IGSs of different bacterial species, extracted from the NCBI database [12]. In section 4.1 we shortly review the main features of the model applied to the IGS identified in chapter 2. In fact, these IGSs correspond to the DNA regions containing relevant motifs of gene expression and regulation. Studying the features of their denaturation process is particularly important for understanding the suitable thermodynamic conditions under which the transcription process of the genes they regulate can be efficiently performed. The protocol employed for studying the denaturation process in these IGSs is described in section 4.2. The results obtained for *E. coli* are illustrated in subsection 4.2.1, while the ones obtained for different bacterial species are shown in subsection 4.2.2, where we also comment on the similarities and differences among IGSs of bacteria living in different environments, that influence significantly their metabolism.

## 4.1 The Model of bacterial IGSs

The model adopted in this paper is the one introduced in [89], describing a simplified double-strand dynamics of DNA. It represents interacting nucleotides A,T, G and C as mass points displaced along two chains with longitudinal and transverse interactions. The displacement of the  $i$ -th nucleotide from its equilibrium position is denoted by  $w_i$  on the upper strand and by  $v_i$  on the lower one. The interaction between neighboring nucleotides on the same strand is characterized by a stacking potential, containing a harmonic component and an effective nonlinear interaction, yielding a cooperative effect in the denaturation process. As discussed in [89], this is a crucial ingredient for observing the denaturation phase transition in this one-dimensional model. The transverse interaction between bps is governed by a Morse potential, that is designed for including the effect due to hydrogen bonds and also the repulsive interaction of the phosphate groups, partly screened by the surrounding solvent action.

Since in this manuscript we are interested to analyze the thermodynamics of the denaturation process in real IGSs of bacterial species, we have explicitly taken into account

that the four different nucleotides have different masses: the symbols  $m_i^+$  and  $m_i^-$  denote the mass of the  $i$ -th nucleotide in the upper and lower strands, respectively,  $i = 1, \dots, N$  labelling the  $N$  bps along the double strand.

Now we illustrate the passages yielding the Hamiltonian used for this thesis. From [87], the system is described by the Hamiltonian

$$H = \sum_i \left[ \frac{1}{2} m \dot{x}_i^2 + \frac{k}{2} (x_{i+1} - x_i)^2 \right] + \left[ \frac{1}{2} m \dot{y}_i^2 + \frac{k}{2} (y_{i+1} - y_i)^2 + D_i (e^{-a_i y_i} - 1)^2 \right], \quad (4.1)$$

where  $x_i = (w_i + v_i)/\sqrt{2}$ ,  $y_i = (w_i - v_i)/\sqrt{2}$  and  $m$  is the average mass of the four nucleotides. Let's add the stacking potential as in [89], which takes into account the interaction between the electronic distributions of adjacent bases:

$$H = \sum_i \left[ \frac{1}{2} m \dot{x}_i^2 + \frac{k}{2} (x_{i+1} - x_i)^2 \right] + \left[ \frac{1}{2} m \dot{y}_i^2 + \frac{k}{2} (1 + \rho e^{-\alpha(y_{i+1} + y_i)}) (y_{i+1} - y_i)^2 + D_i (e^{-a_i y_i} - 1)^2 \right] \quad (4.2)$$

Let's recover the degeneracy in the masses of the four nucleotides and, writing the Hamiltonian with the real coordinates and momenta, it becomes

$$H = \sum_i \frac{1}{2} m_i^+ \dot{w}_i^2 + \frac{k}{2} (w_{i+1} - w_i)^2 + \frac{1}{2} m_i^- \dot{v}_i^2 + \frac{k}{2} (v_{i+1} - v_i)^2 + \frac{k}{2} \rho \exp \left( -\alpha \frac{w_{i+1} + w_i - v_{i+1} - v_i}{\sqrt{2}} \right) \times \left( \frac{w_{i+1} - w_i + v_{i+1} - v_i}{\sqrt{2}} \right)^2 + D_i \left( \exp \left( -a_i \frac{w_i - v_i}{\sqrt{2}} \right) - 1 \right)^2, \quad (4.3)$$

where  $m_i^+$  and  $m_i^-$  are the masses of the nucleotide in the position  $i$  on the upper and lower strand respectively. The complete Hamiltonian reads

$$H = \sum_{i=1}^N (H_h^+(i) + H_h^-(i) + V(i)), \quad (4.4)$$

where

$$H_h^+(i) = \frac{1}{2}m_i^+\dot{w}_i^2 + \frac{k}{2}(w_{i+1} - w_i)^2 \quad (4.5)$$

is the harmonic energy of the upper strand,

$$H_h^-(i) = \frac{1}{2}m_i^-\dot{v}_i^2 + \frac{k}{2}(v_{i+1} - v_i)^2 \quad (4.6)$$

is the harmonic energy of the lower strand and

$$\begin{aligned} V(i) = & \frac{k}{2}\rho \exp\left(-\frac{\alpha}{\sqrt{2}}(w_{i+1} + w_i - v_{i+1} - v_i)\right) \times \\ & \frac{1}{2}(w_{i+1} - w_i - v_{i+1} + v_i)^2 + \\ & D_i \left( \exp\left(-a_i \frac{w_i - v_i}{\sqrt{2}}\right) - 1 \right)^2. \end{aligned} \quad (4.7)$$

The first term in the above equation stems from the stacking interaction, while the second one is the Morse potential. Notice that we have to rewrite the overall Hamiltonian introduced in [89] as a function of the real particle coordinates, because considering different nucleotide masses does not allow us to use the reduced form of the Hamiltonian therein reported. Another consequence of this choice is that the spectrum of linear waves, obtained via the harmonic approximation, contains an (almost) acoustic component, corresponding to slow collective motion of the backbone. As shown later, this will affect the low-frequency dynamics in the non-denaturated phase.

The values of the physical parameters have been borrowed from [90]:  $k = 0.025 \text{ eV}/\text{\AA}^2$ ,  $\rho = 2$ ,  $\alpha = 0.35 \text{ \AA}^{-1}$ ,  $D_{AT} = 0.05 \text{ eV}$ ,  $D_{GC} = 0.075 \text{ eV}$ ,  $a_{AT} = 4.2 \text{ \AA}^{-1}$  and  $a_{GC} = 6.9 \text{ \AA}^{-1}$ . The masses of the four nucleotides are known to take the values  $m_A = 0.0343 \text{ eV ps}^2/\text{\AA}^2$ ,  $m_T = 0.0333 \text{ eV ps}^2/\text{\AA}^2$ ,  $m_G = 0.0360 \text{ eV ps}^2/\text{\AA}^2$  and  $m_C = 0.0318 \text{ eV ps}^2/\text{\AA}^2$ .

The IGSs of bacteria species have been selected according to the procedure described in chapter 2. In what follows we investigate the thermodynamics of denaturation of these IGSs. Due to their structural differences, we expect that also the denaturation temperature should be different for IGSs belonging to the three structural clusters. Moreover, our study will concern various bacterial species, in order to point out that the denaturation transition in their IGSs is significantly correlated to their structural features.

## 4.2 The denaturation dynamics

All numerical simulations of model (4.4) reported in this paper have been performed by imposing periodic boundary conditions. This choice is appropriate, because the IGSs are portions of longer chains, so that neither free nor fixed boundary conditions can be considered close to the real conditions in DNA. For instance, we have checked that free boundary conditions fasten the denaturation process, while fixed boundary conditions anomalously favour the process of bps recombination, thus yielding unrealistic underestimates and overestimates, respectively, of the denaturation temperature (data not reported).

Molecular dynamics has been performed making use of a symplectic integration algorithm, based on the fourth order Runge-Kutta method with coefficients taken from [94]. Initial conditions have been sampled by choosing typical thermal configurations from a canonical Gibbs distribution. Most of the simulations reported in this paper correspond to a canonical setup, i.e. the temperature of the double-chain is fixed by putting in contact its dynamical variables with a thermal reservoir. For this purpose we have adopted a standard scheme of a Maxwellian thermostat [95]. More precisely, the velocities of all mass particles (nucleotides) at integer multiples of a given time interval  $t_{th}$  are updated by assigning them new values drawn from a Maxwell-Boltzmann distribution at given temperature  $T$ . Between two actions of the thermostat, energy is conserved thanks to the symplectic integration algorithm. As we shall discuss in detail in the following section, we have also performed some molecular dynamics simulations in a microcanonical setup, i.e. making the double-chain evolve deterministically, after having switched-off the thermostat.

Suitable time scales for the molecular dynamics simulations can be tuned by considering the typical frequencies present in model (4.4). Given the adopted physical parameters of the model, one can easily realize that there is a minimum and a maximum frequency, namely

$$\nu_{\min} = \frac{1}{N} \sqrt{\frac{k}{\bar{m}}} \approx 5 \times 10^6 \text{ Hz} \quad (4.8)$$

$$\nu_{\max} = \frac{1}{2\pi} \sqrt{\frac{2D_{GC}a_{GC}^2}{m_C}} \approx 2.4 \times 10^9 \text{ Hz}, \quad (4.9)$$

where  $\bar{m} = 0.0339 \text{ eV ps}^2/\text{\AA}^2$  is the average mass of nucleotides. The former corresponds to a period  $\mathcal{T} = \frac{1}{\nu_{\min}} = 203 \text{ ps}$ , which indicates the minimum duration for any meaningful

simulation run, while the latter, according to the Nyquist-Shannon theorem, allows to estimate a lower bound for the sampling frequency  $\nu_s > 2\nu_{\max}$ , implying that the sampling time interval  $t_s = \nu_s^{-1}$  has to be smaller than 0.21 ps:  $t_s = 0.2$  ps has been chosen henceforth. Anyway, we have checked that in all the explored range of parameters this choice guarantees a sufficient statistical sampling, while avoiding the storage of exceedingly large data sets.

The integration time step  $\Delta t$  of the symplectic integration algorithm must be sufficiently smaller than  $t_s$ : we have taken  $\Delta t = 10^{-2}$  ps, because it allows to keep energy conservation up to the sixth significant figure.

One further important time scale to be properly tuned in the canonical setup is  $t_{th}$ , i.e. the time interval between successive actions of the thermal reservoir. In order to cope with this task we have performed preliminary molecular dynamics simulations in both setups by measuring the (unnormalized) dynamical structure factor of the coordinate of the upper strand for an IGS of *E. coli*,

$$S_w(k, \nu) = \left\langle \left| \int_0^\tau \sum_{j=1}^N w_j(t) e^{i2\pi(\nu t - \frac{k}{N}j)} dt \right|^2 \right\rangle, \quad (4.10)$$

where  $\tau$  is the total integration time and the brackets represent an average over an ensemble of independent molecular dynamics trajectories of the same IGS in the dynamical phase below the denaturation transition<sup>1</sup>. The choice of measuring  $S_w(k, \nu)$  is not a prejudice of generality, since very similar results are obtained by measuring the structure factor of the coordinate of the lower strand.

The upper panel of Figure 4.1 shows  $S_w(k, \nu)$  for the isolated double-chain: as expected the peaks in the low frequency region correspond to  $\nu_{\min}$  and its harmonics, while in the high frequency region one can single out fast decay tails immediately after  $\nu_{\max}$ . The three lower panels, from top to bottom, show the same quantity for canonical simulations with  $t_{th} = 10^2, 10$  and 1 ps. For  $t_{th} = 10^2$  ps some spectral components of the isolated chain are still distinguishable, but the thermalization process results to be extremely slow (i.e.  $\mathcal{O}(10^8 \Delta t)$ ). The high-frequency spectrum of the isolated double-chain is still reasonably reproduced for  $t_{th} = 10$  ps and the thermalization time results to be two orders of magnitude smaller (i.e.  $\mathcal{O}(10^6 \Delta t)$ ). Upon this outcome, this appears as a suitable compromise for a thermalized dynamics keeping relevant dynamical features of

---

<sup>1</sup>In the calculation of the structure factor the  $w_i$  has been subtracted by their average value to avoid a peak at zero frequency.

the double-chain, while allowing a reasonable thermalization time. In fact, in the third case,  $t_{th} = 1$  ps, an even faster thermalization is attained, but the price to pay is the wiping of any signature of the deterministic dynamics of the double-chain. In summary, this spectral analysis allows one to properly tune, with respect to the typical time scales of the model, the response time of the thermostat in canonical molecular dynamics.

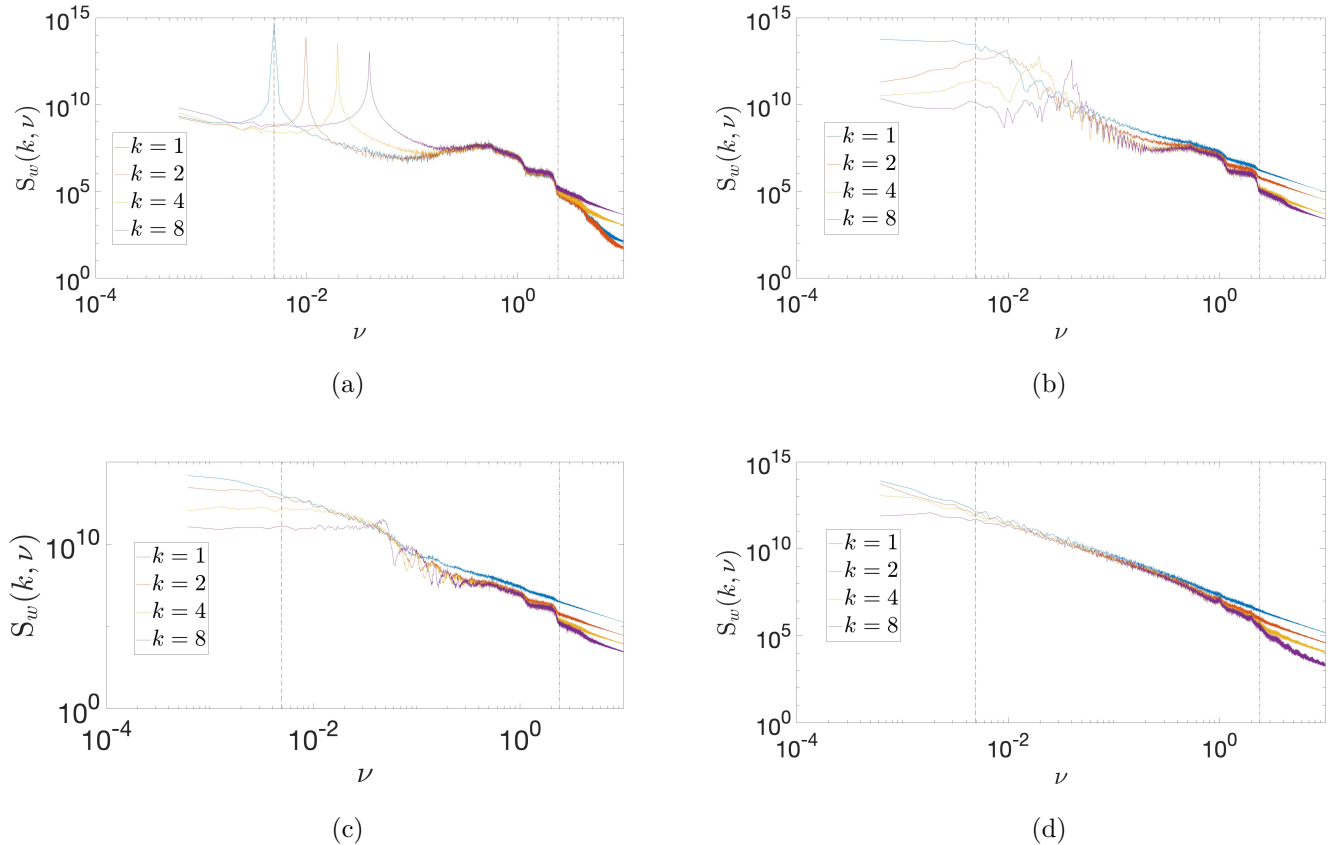


Figure 4.1: *E. coli*. The structure factor  $S_w(k, \nu)$  of the upper-strand coordinates of an IGS for different values of the time scale  $t_{th}$  of the thermal reservoir: isolated chain (a),  $t_{th} = 100$  ps (b),  $t_{th} = 10$  ps (c) and  $t_{th} = 1$  ps (d). The dashed line on the left and the dot-dashed line on the right single out the values of  $\nu_{\min}$  and  $\nu_{\max}$ , respectively. In each panel we report data for some of lowest wavenumber values  $k = 1, 2, 4, 8$  (long-wavelength limit). Averages have been performed over 25 independent molecular dynamics trajectories. Simulation parameters are  $\Delta t = 10^{-3}$  ps, run time  $t_I = 1628$  ps,  $t_s = 10^{-2}$  ps.

Since we aim at characterizing the denaturation process in bacteria IGSs, modeled by Hamiltonian (4.4), we have to establish a microscopic condition for bps denaturation. Making reference to the adopted physical parameters, we use the common criterion [87, 89, 90] that the chemical bond at site  $i$  is open (or denaturated) if the relative displacement of the nucleotides in the asymmetric Morse potential (see (4.7)) overtakes a given typical length, specifically

$$(w_i - v_i)/\sqrt{2} > 2\text{\AA}. \quad (4.11)$$

In fact, this choice is equivalent to a dynamical configuration, where the interaction force originated by the Morse potential in (4.7) becomes negligible.

In order to characterize the denaturation dynamics of an IGS of length  $N$  we consider three different indicators:

- the one that indicates if the IGS is denaturated or not

$$D(t) = \begin{cases} 1 & \text{if the sequence} \\ & \text{is fully denaturated at time } t, \\ 0 & \text{otherwise;} \end{cases} \quad (4.12)$$

- the relative length of the largest open bubble at time  $t$

$$b(t) = \frac{1}{N} \sum_{i \in S_b} \delta_i(t), \quad (4.13)$$

where  $S_b$  represents the support of the largest open bubble at time  $t$  and

$$\delta_i(t) = \begin{cases} 1 & \text{if the bond at site } i \text{ is open,} \\ 0 & \text{otherwise;} \end{cases} \quad (4.14)$$

- the one that indicates if the transcription bubble (the first 14 bps upstream the TSS) is open or not

$$c(t) = \begin{cases} 1 & \text{if the trascription bubble} \\ & \text{is fully denaturated at time } t, \\ 0 & \text{otherwise.} \end{cases} \quad (4.15)$$

The first quantity,  $D(t)$ , is a natural order parameter of the denaturation transition and provides a global statistical measure of the denaturated configuration. The “bubble” order parameter  $b(t)$  aims at catching the role of entropic, i.e. configurational, fluctuations in the denaturation process, while  $c(t)$  points directly to a specific feature associated with the denaturation of the IGS region, where the RNA-polymerase can intrude the double-chain to start the gene transcription process [96].

### 4.2.1 Denaturation for IGSs of *E. coli*

In this section we first focus our study on the canonical setup for the dynamics of the IGSs of *E. coli*. In particular, we study the denaturation transition by averaging over a subset of typical sequences belonging to the three clusters C0, C1 and C2 introduced in chapter 2. In particular, in order to obtain an effective statistical sampling, we have chosen  $N_s$  different IGSs from each cluster and we have studied their denaturation dynamics for different values of the thermostat temperature  $T$ . The numerical protocol adopted for measuring  $\langle D \rangle$  has been worked out by considering for any value of  $T$  the fraction of IGSs that are fully denaturated in the set of the chosen IGSs, whose dynamics has been sampled at integer multiples of the sampling time  $t_s = 0.2$  ps, over an integration time  $t_I \sim \mathcal{O}(10^6 \Delta t) = \mathcal{O}(10^4)$  ps. In fact, this choice of  $t_I$  guarantees that any IGS has reached a thermal equilibrium state for every value of  $T$  in the explored range around the denaturation transition temperature  $T_d$ . In order to improve the statistics we have also averaged over  $N_r$  different realizations of the dynamics of each IGS, in formulae

$$\langle D \rangle = \frac{t_s}{t_I} \frac{1}{N_s} \frac{1}{N_r} \sum_{j=1}^{N_r} \sum_{m=1}^{N_s} \sum_{n=1}^{t_I/t_s} D_m^{(j)}(n), \quad (4.16)$$

where  $D_m^{(j)}(n)$  is the quantity defined in (4.12) for the  $j$ -th realization of the  $m$ -th sequence at discrete time  $n$ . The results are reported in Figure 4.2, where the error bars have been computed as the *standard error of the mean* defined by the formula

$$\sigma_D = \sqrt{\frac{\sum_{j=1}^{N_r} \sum_{m=1}^{N_s} (\langle D_m^{(j)} \rangle - \langle D \rangle)^2}{(N_s N_r - 1)(N_s N_r)}}, \quad (4.17)$$

where

$$\langle D_m^{(j)} \rangle = \frac{t_s}{t_I} \sum_{n=1}^{t_I/t_s} D_m^{(j)}(n). \quad (4.18)$$

By adopting the criterion of identifying  $T_d$  when  $\langle D \rangle = 0.5$ , we observe that the structural features of the three clusters identify three different values of  $T_d$ . More precisely, IGSs from clusters C1 and C2 have very close values of  $T_d$ , because of their structural similarity, while  $T_d$  for the IGSs from cluster C0 is significantly smaller. Anyway, for all clusters  $T_d$  is found to be close to 350 K, quite a realistic temperature for the thermal denaturation of a DNA strand. It is important to point out that the minimum energy involved in the process of a single bond breaking, 0.050 eV, is significantly larger than the



energetic equivalent of the denaturation temperature  $k_B T_d \approx 0.030$  eV, thus confirming that even in relatively short IGSs, like those of *E. coli* (and also of other bacterial species, as we show hereafter), the thermal denaturation process is a cooperative dynamical effect, which is driven by statistical fluctuations.

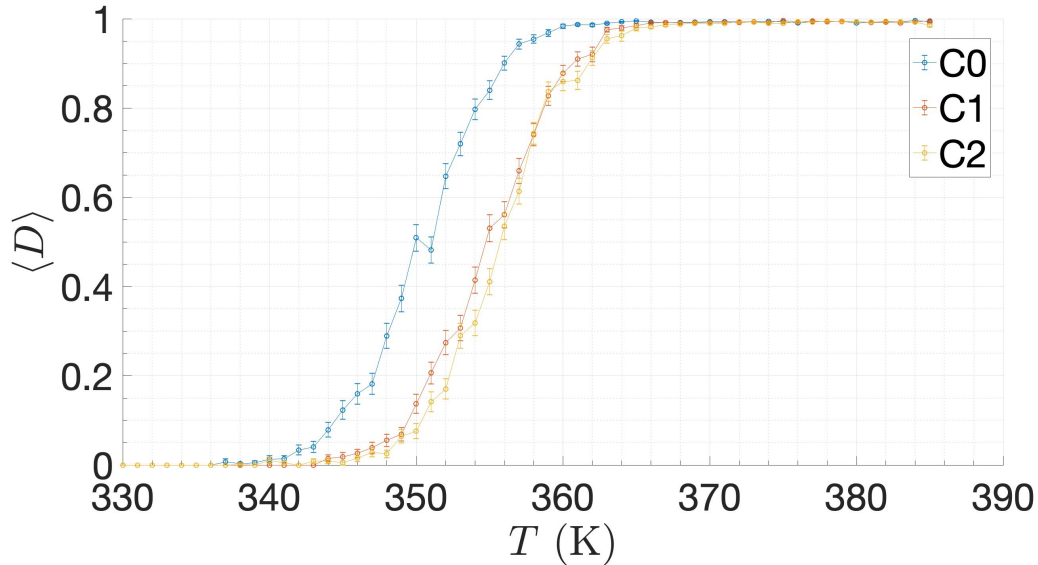


Figure 4.2: *E. coli*. The average fraction of denaturated IGSs  $\langle D \rangle$  of the three different clusters, C0, C1 and C2, as a function of temperature  $T$ . The parameters adopted in these numerical simulations are  $t_I = 10^6 \Delta t$ ,  $N_s = 18$  and  $N_r = 20$ .

In order to obtain a complete characterization of the thermodynamics of denaturation in bacteria IGSs we have performed also molecular dynamics simulation of isolated double-chains, i.e. in a microcanonical setup. As we have pointed out in the introduction, a comparison between the results obtained by the canonical setup and by the microcanonical one is worthwhile, because the presence of a phase transition in the thermodynamic limit for the 1D model (4.4) does not guarantee for the statistical ensemble equivalence and one could expect to recover some trace of such inequivalence also in the denaturation dynamics. In order to save computational time we limit this analysis to the set of sequences, belonging to the cluster C0 of *E. coli*. In Figure 4.3 we compare the denaturation curves in the canonical and in the microcanonical ensembles: the difference in the denaturation temperatures is sizeable.

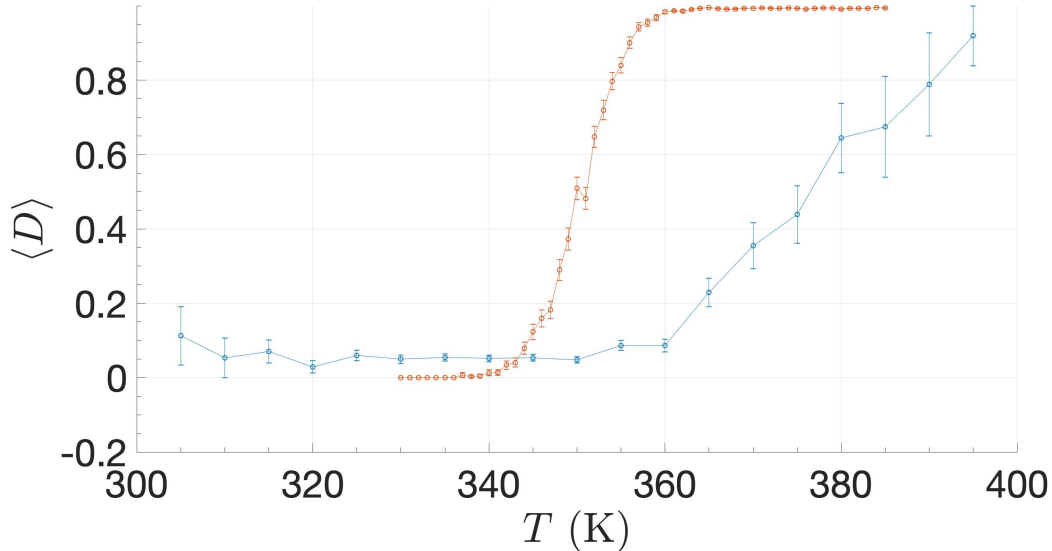


Figure 4.3: *E. coli*. The average fraction of denaturated intergenic sequences  $\langle D \rangle$  from cluster C0 for different values of the temperature  $T$ . In this figure we compare the results obtained by the canonical (red) and the microcanonical (blue) setups. We observe a remarkable difference, which is due to the crucial effect played by thermal fluctuations in determining the denaturation transition, driven by cooperative interactions between nucleotides. The parameters adopted in these numerical simulations are  $t_I = 10^6 \Delta t$ ,  $N_s = 6$  for cluster,  $N_r = 15$ .

This result can be interpreted by considering that the denaturation transition in the canonical ensemble of model (4.4) is due to the entropic stabilization of a soliton-like domain wall, which contains a finite amount of the macroscopic energy [91, 92]. Such a condensation phenomenon may not necessarily occur in a microcanonical description, where energy conservation does not allow for the large amplitude fluctuations induced by the contact of the double-chain with a thermal reservoir. Accordingly, one can conjecture that the cooperative effect of fluctuations in determining the denaturation transition could be significantly weakened in a microcanonical scenario. For finite double-chains, like those considered in this paper, the finding of different denaturation temperatures in the two setups provides a partial confirmation of this conjecture.

We have also studied the denaturation dynamics in the canonical setup by performing the same kind of statistical measurements for the average extension of the *largest bubble*,  $\langle b \rangle$ , and of the denaturation of the *transcription bubble*,  $\langle c \rangle$ , as defined in (4.16) by replacing  $D(t)$  with  $b(t)$  and  $c(t)$ , respectively. The results obtained for  $\langle b \rangle$  are shown in Figure 4.4, together with the corresponding *standard error of the mean*  $\sigma_b$ . We can assume that  $T_d$  in the three clusters can be identified looking by the position of the peaks of  $\sigma_b$ : actually, the higher the fluctuations of the *largest bubble* the closer we are to  $T_d$ . We can observe that the estimates of  $T_d$  are in very good agreement with those obtained

by the heuristic criterion adopted for  $\langle D \rangle$ .

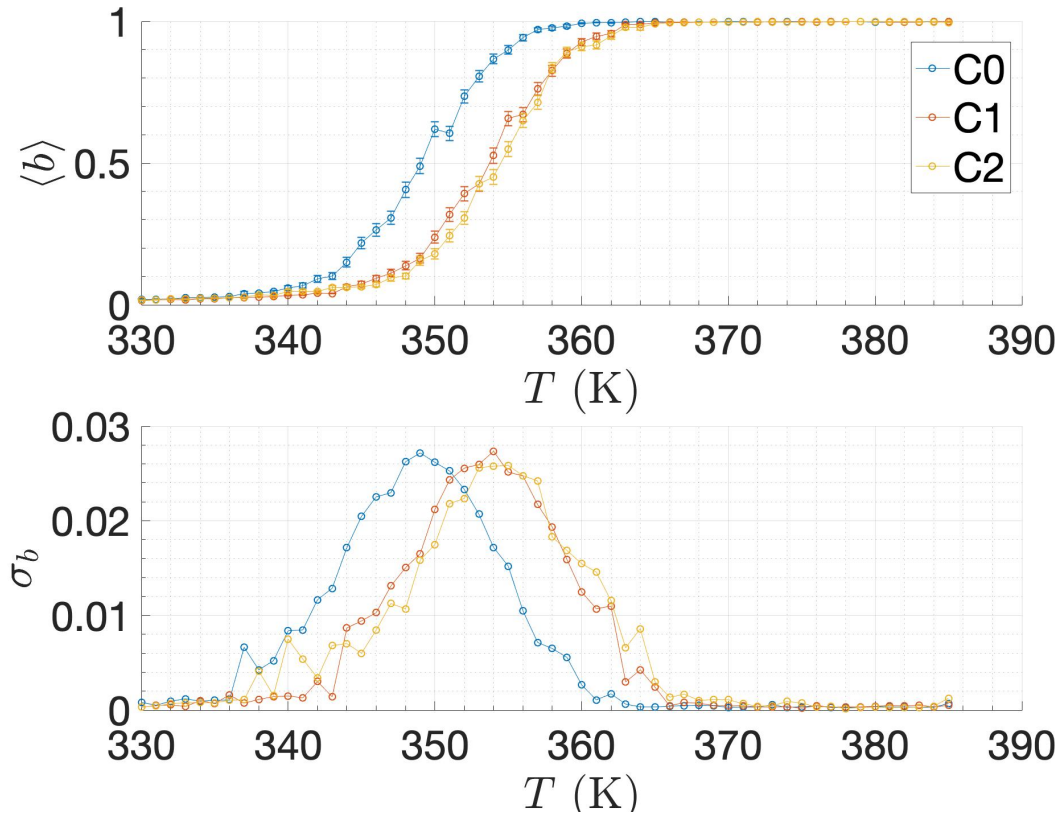


Figure 4.4: *E. coli*. The average value (upper panel) and the variance (lower panel) of the largest bubble length from the three different clusters, C0, C1 and C2, for different values of the temperature  $T$ , expressed in K degrees. The parameters adopted in these numerical simulations are the same of Figure 4.2.

For what concerns  $\langle c \rangle$  (see Figure 4.5) we find that  $T_d$ , identified as the value at which half of the configurations have the transcription bubble denaturated (i.e.  $\langle c \rangle = 0.5$ ), is slightly smaller than the estimate obtained by the previous statistical analyses. This is due to the increasing gradient of weak basis in this region (see section 2.3), although the actual values are very close to the previous ones. We can conclude that in the canonical setup the identification of  $T_d$  doesn't depend significantly on the choice of the indicator.

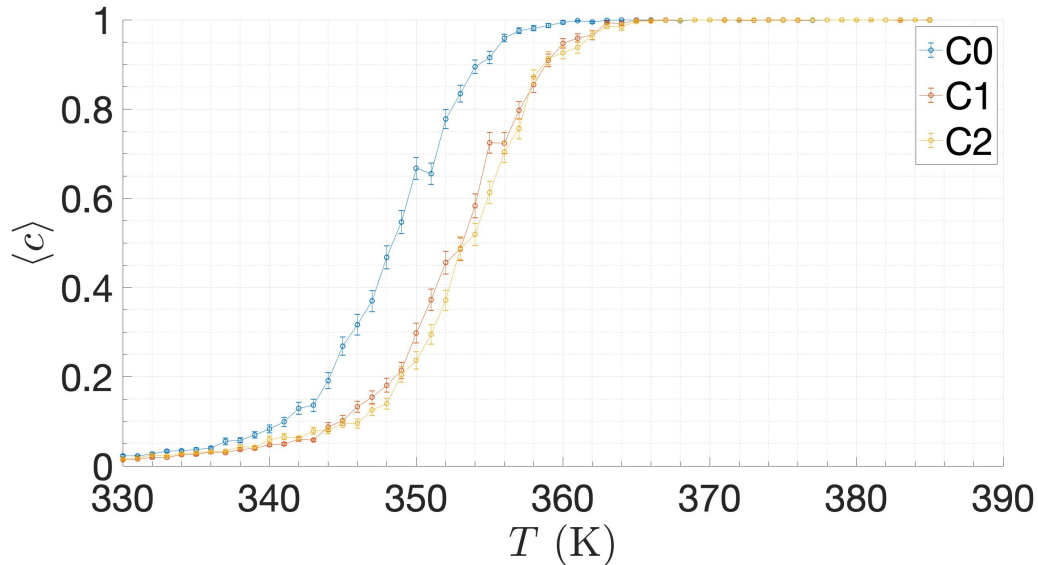


Figure 4.5: *E. coli*. The average fraction of the denaturated transcription bubble  $\langle c \rangle$  of intergenic sequences from the three different clusters, C0, C1 and C2, for different values of the temperature  $T$ , expressed in K degrees. The parameters adopted in these numerical simulations are the same of Figure 4.2.

### Specific heats and caloric curves

For completeness here we discuss another method for determining the denaturation temperature. Through the fluctuation-dissipation relation, let us define the specific heat for a single IGS:

$$C = \frac{\langle E^2 \rangle - \langle E \rangle^2}{Nk_B T^2}, \quad (4.19)$$

where  $E$  is the total energy of the system and the average is made on temporal sampling and on  $N_r$  different realizations of the dynamics. Due to finite size effects, this quantity signals the occurrence of a phase transition exhibiting a bump at the critical value of the temperature. We execute the same calculation for one IGS for each cluster of *E. coli*. From Figure 4.6 we can see that the three curves obtained from a polynomial fit of the real values of  $C$  have the desired shape and provide comparable denaturation temperatures.

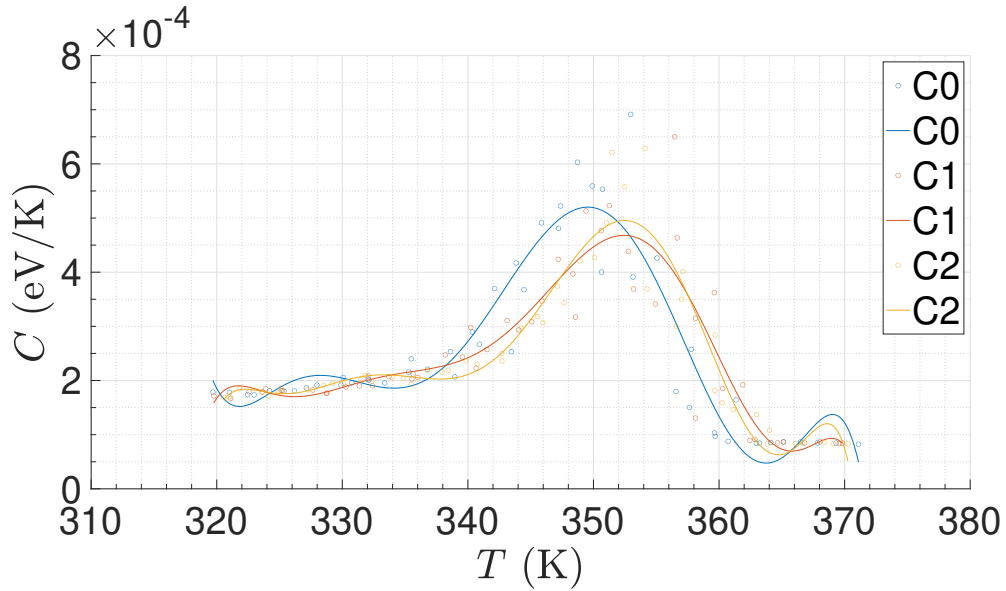


Figure 4.6: *E. coli*. Specific heat of the three different clusters, C0, C1 and C2, as a function of temperature  $T$ . The circles correspond to the values obtained by the numerical simulations, the continuous lines to the best polynomial fit. The parameters adopted in these numerical simulations are  $t_I = 10^6 \Delta t$  and  $N_r = 20$ .

We also report the caloric curves (see Figure 4.7), one for each cluster. All three have two branches with a defined slope corresponding to the two distinct phases, the non-denatured one at low energies and the denatured one at high energies, connected by flatter segment that corresponds to the position of the bump of the specific heat  $C$ .

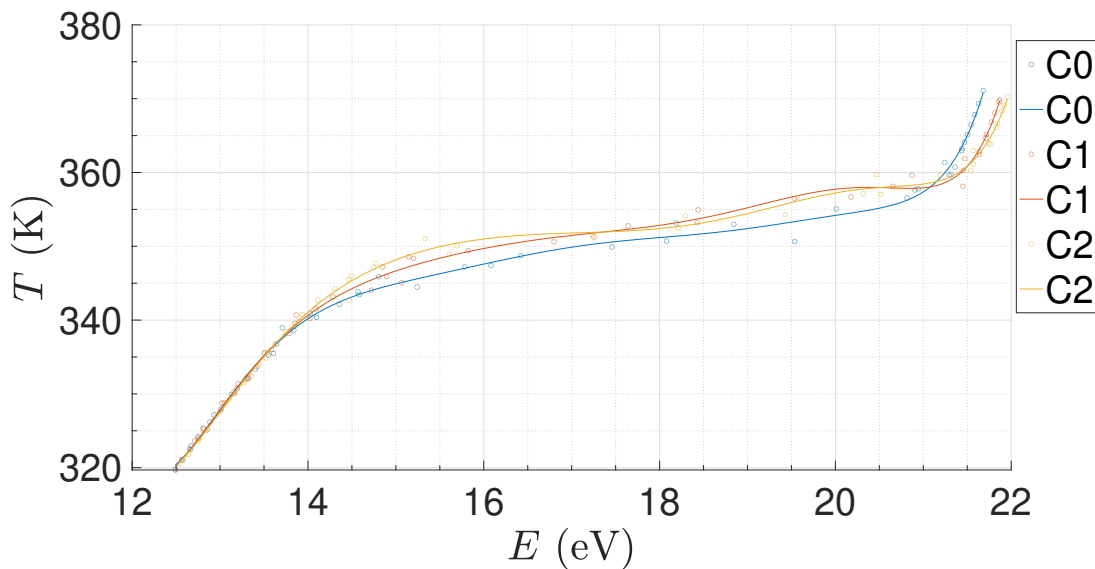


Figure 4.7: *E. coli*. Caloric curves of the three different clusters, C0, C1 and C2. The circles correspond to the values obtained by the numerical simulations, the continuous lines to the best polynomial fit. The parameters adopted in these numerical simulations are  $t_I = 10^6 \Delta t$  and  $N_r = 20$ .

## 4.2.2 Other bacterial species

We have compared the results obtained for *E. coli* IGSs with those of three other bacterial species, namely *B. subtilis*, *P. haloplanktis* and *S. coelicolor*. The BCA of the first three species are also reported in the figures of section 2.3 and section 2.4. Here we show the BCA of the fourth bacterium, *S. coelicolor* (see Figure 4.8).

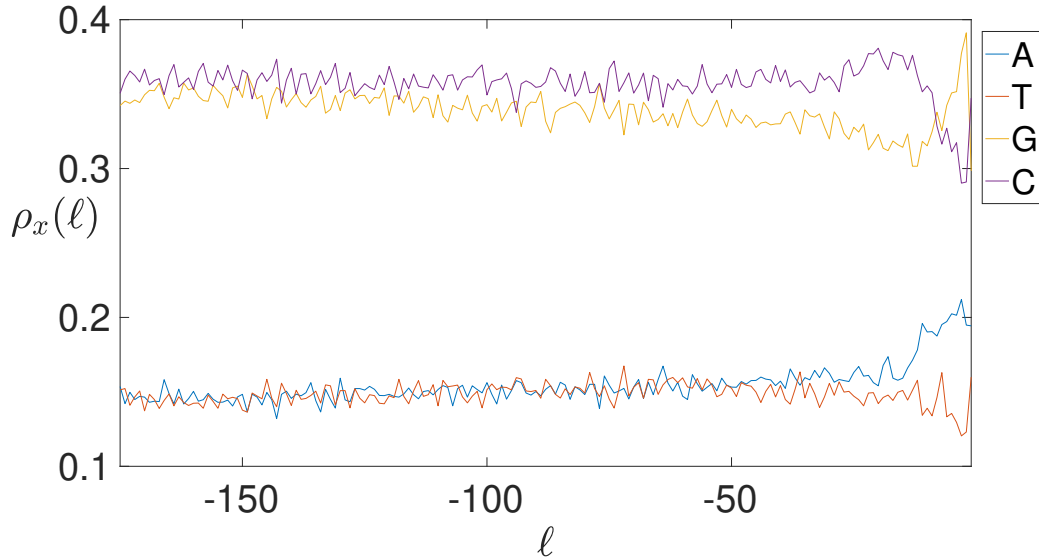


Figure 4.8: *S. coelicolor*. BCA of the IGSs: on the vertical axis we report the density  $\rho_x(\ell)$  of each of the four nucleotides  $x = A$  (blue), T (red), G (yellow), C (purple) as a function of the position  $\ell$  along the annotated 8022 IGSs.

As a first step of this comparison, in Figure 4.9 we report  $\langle D \rangle$ , averaged over the whole samples of the IGSs for each species. The  $T_d$  of *B. subtilis* and *P. haloplanktis* are quite close to each other, but a bit smaller than the  $T_d$  of *E. coli*. This is consistent with the strong similarity of the BCA of the former two species (see again section 2.4), while the moderately higher contribution of strong bases in the BCA of *E. coli* yields a higher value of  $T_d$ . The IGSs of the fourth species, *S. coelicolor*, exhibits very different structural properties, not only because of the dominance of the strong bases with respect to the weak ones (see Figure 4.8), but also because the clustering procedure identifies a single gigantic cluster containing most of the IGSs. In fact, for this species we obtain  $T_d \approx 363$  K, which is definitely higher than the  $T_d$  of the other species. Altogether, this statistical studies confirm that the estimate of  $T_d$  is consistent also in these bacterial species, as we have observed for *E. coli*.

For what concerns the division in clusters we can see almost the same scenario for *E. coli* and *P. haloplanktis* (compare Figure 4.2 and Figure 4.11), while for *B. subtilis* (see

Figure 4.10) the cluster C1 exhibits a lower denaturation temperature respect to the other two (for *E. coli* and *P. haloplanktis* this role is played by cluster C0); the cause could be attributed to a higher concentration of the nucleotide T respect to the one of A, and of C respect to the one of G. In fact, between the two weak bases T has the smallest mass, the same is true for C for the strong bases. This fact makes understand the importance of the difference of the masses among the four nucleotides in the denaturation process which should not only be attributed to the number of covalent bonds between the two strands, and therefore to the mere division between weak and strong bases.

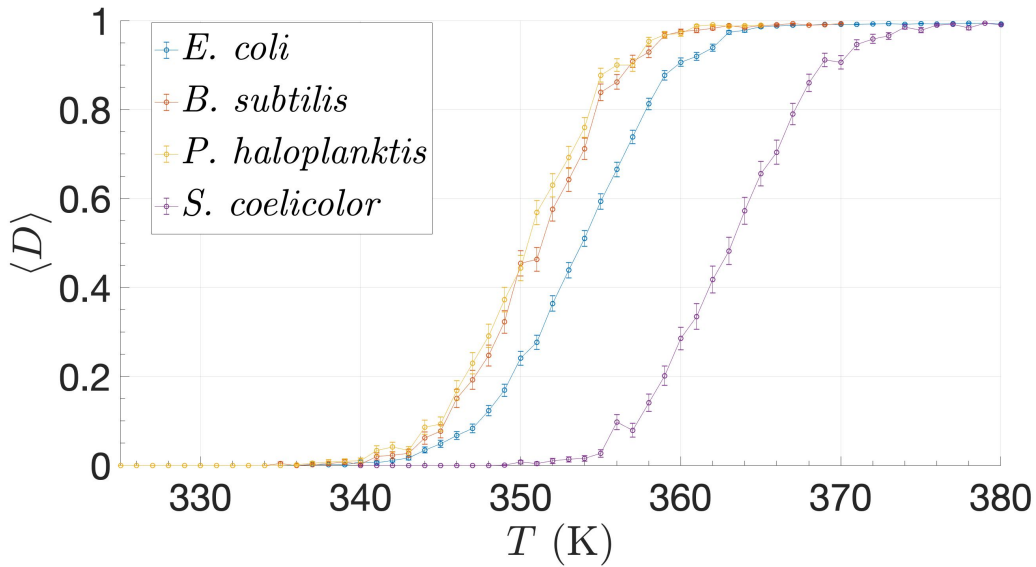


Figure 4.9: The average fraction of denaturated intergenic sequences  $\langle D \rangle$  of *E. coli*, *B. subtilis*, *P. haloplanktis* and *S. coelicolor* for different values of the temperature  $T$ , expressed in K degrees. The parameters adopted in these numerical simulations are  $t_I = 10^6 \Delta t$ ,  $N_s = 6$  and  $N_r = 20$ .

In order to provide further details about the denaturation indicator  $\langle D \rangle$  for different clusters, hereafter we report the data for *B. subtilis* (see Figure 4.10) and *P. haloplanktis* (see Figure 4.11).

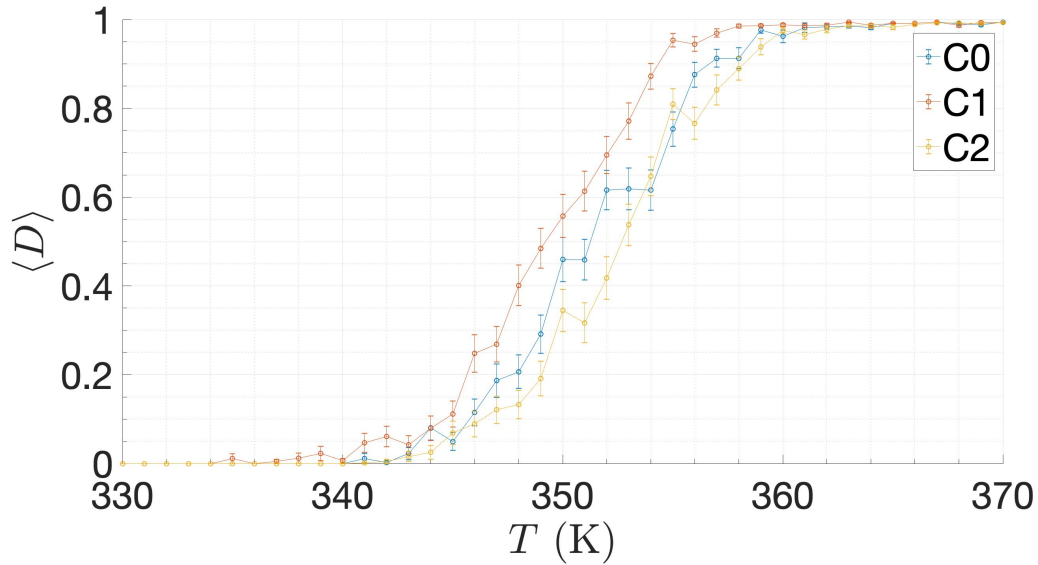


Figure 4.10: *B. subtilis*. The average fraction of denaturated intergenic sequences  $\langle D \rangle$  of the three different clusters, C0, C1 and C2, as a function of temperature  $T$ . The parameters adopted in these numerical simulations are  $t_I = 10^6 \Delta t$ ,  $N_s = 12$  and  $N_r = 15$ .

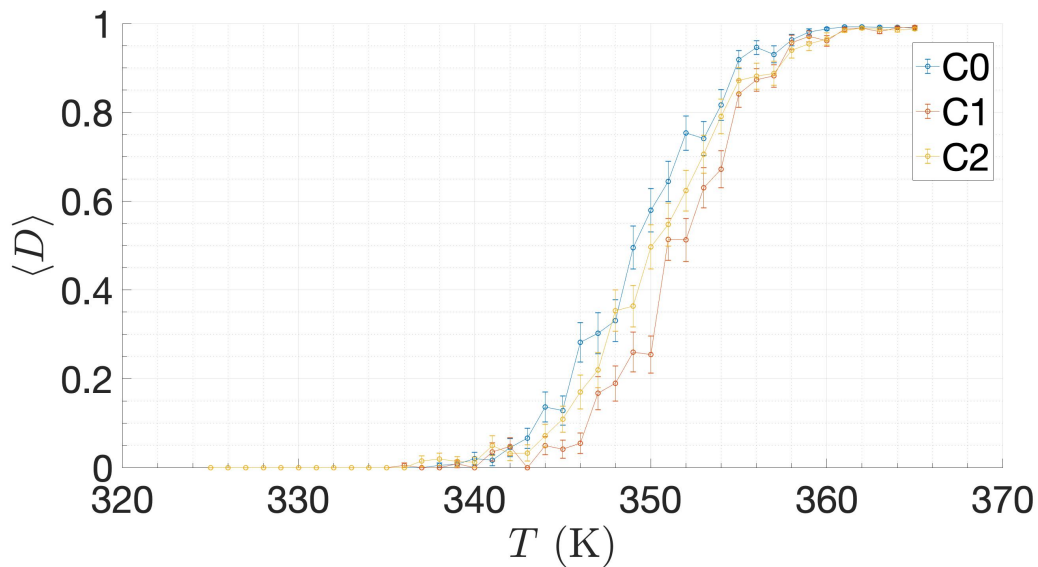


Figure 4.11: *P. haloplanktis*. The average fraction of denaturated intergenic sequences  $\langle D \rangle$  of the three different clusters, C0, C1 and C2, as a function of temperature  $T$ . The parameters adopted in these numerical simulations are  $t_I = 10^6 \Delta t$ ,  $N_s = 12$  and  $N_r = 15$ .

Finally, for the indicators  $\langle b \rangle$  and  $\langle c \rangle$  of the other species qualitatively similar results to *E. coli* are found (data not reported).



### 4.3 Final remarks

In the physical literature devoted to the study of DNA denaturation much attention has been attracted by the definition of suitable model ingredients able to capture the cooperativeness intrinsic to this phenomenon. A systematic analysis of this phenomenon for IGSs preceding the TSS of coding sequences (genes) to our knowledge was still lacking.

In this thesis, we have focused our attention on IGSs of several bacterial species. They extend over a not too large number of nucleotides (typically 175 bps), thus making possible a careful and extended computational study of the denaturation process. Our numerical simulations confirm that thermal fluctuations play a crucial role in amplifying the collective effects ruling denaturation, while indicating that the actual denaturation temperature is associated to the structural features of the IGSs. For instance, specific algorithms allow to group the IGSs of *E. coli* and other bacterial species into three similar populations, each one characterized by different structural features and, accordingly, different values of the typical denaturation temperature. Moreover, we have reported the study of a peculiar bacterial species (*S. coelicolor*), whose IGSs are characterized by a dominance of strong nucleotides (C- and G-type). In this case the adopted model consistently predicts a sensibly higher denaturation temperature, with respect to the other bacterial species, where the structure of IGSs is dominated by weak nucleotides (A- and T-type).

Similar studies could be extended to any other species along the phylogenetic tree, although in eukaryotes the relatively large extension of IGSs, even limited to promoters, would demand a considerable computational effort to be properly worked out.

# Chapter 5

## Features of BCA in eukaryotes for coding and noncoding regions

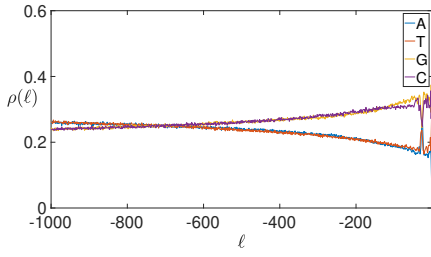
In the previous chapters we dealt with IGSs of prokaryotic organisms. Among the various features we have found that, due for the peculiarities of the genome consisting largely of coding regions, we have considered few hundred bps upstream the TSS to find sequences that are significantly distinct from the coding ones. In eukaryotes the situation is completely upside down (i.e. in *H. sapiens* noncoding DNA is approximately 98% of the genome [97]), so it makes sense to examine several hundred bps long IGSs.

In the first part of the present chapter we try to find common features between non-coding regions of different eukaryotic organisms by analyzing their BCA. We show that some conservation laws exist in the abundance of the four nucleotides. Besides, we notice that the density profile of the weak nucleotides and the strong ones is not randomly arranged in the non coding sequences upstream the TSS, but it has a very specific functional dependence on the distance from the TSS. Surprisingly, we highlight that this function is the same for all the eukaryotes, unless the parameters contained in the function itself.

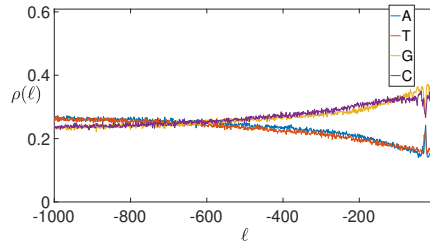
In the second part of the chapter we deal with the properties of the coding regions of *Homo sapiens*, in particular focusing on the well-known three-base periodicity property of the genes. We give an explanation a posteriori, based on the different abundance of amino acids in nature. To reinforce this conclusion we repeat the same analysis also for *E. coli*.

## 5.1 IGSs in eukaryotes

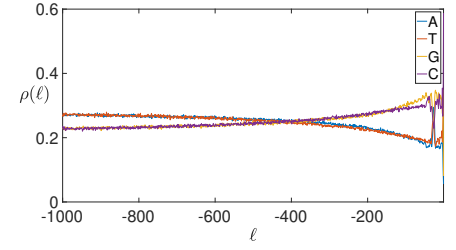
A promoter is a region of DNA where transcription of a gene is initiated. Eukaryotic promoters are much more complex and diverse than prokaryotic promoters. Eukaryotic promoters span a wide range of DNA sequences. There are three main portions that make up a promoter: core promoter, proximal promoter, and distal promoter. The final portion of the promoter region is called the distal promoter which is upstream of the proximal promoter. The distal promoter also contains transcription factor binding sites, but mostly contains regulatory elements. It is not unusual to have several regulatory elements such as enhancers several thousands bps away from the TSS. So for eukaryotes we can consider 1000 bps as the length of IGSs. In Figure 5.1 we report the BCA of several eukaryotic organisms (data have been download from the Eukaryotic Promoter Database [98, 99]), the ones we have examined in this thesis.



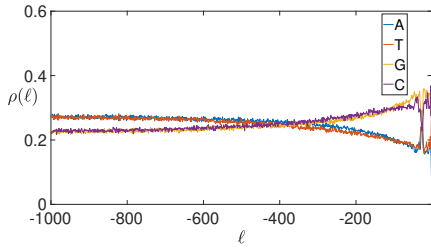
(a) *H. sapiens*



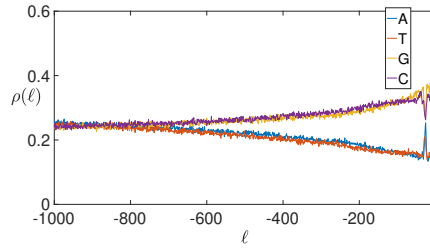
(b) *M. mulatta*



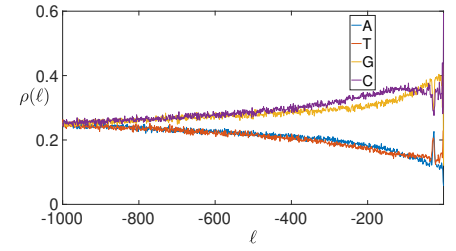
(c) *M. musculus*



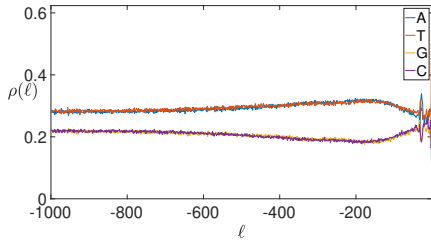
(d) *R. norvegicus*



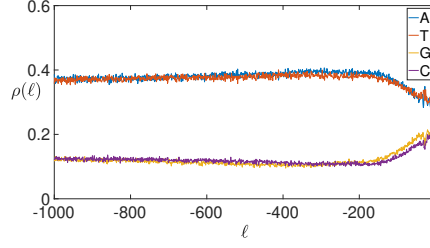
(e) *C. familiaris*



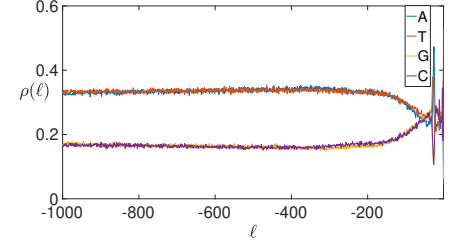
(f) *G. gallus*



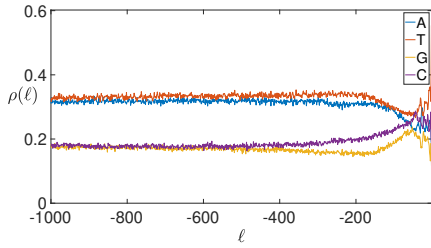
(g) *D. melanogaster*



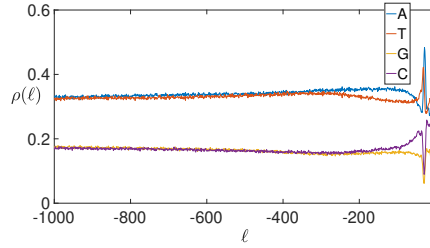
(h) *A. mellifera*



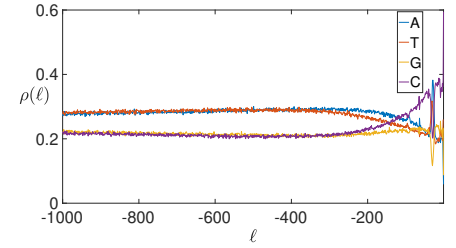
(i) *D. rerio*



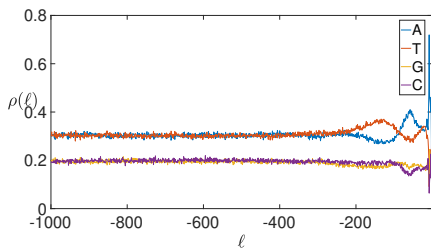
(j) *C. elegans*



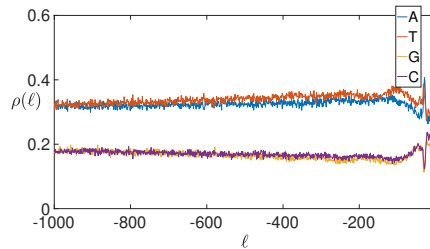
(k) *A. thaliana*



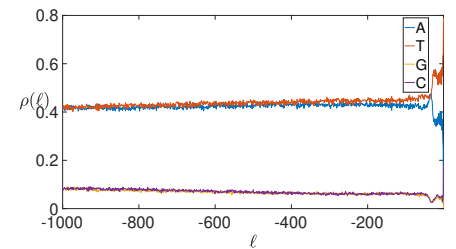
(l) *Z. mays*



(m) *S. cerevisiae*



(n) *S. pombe*



(o) *P. falciparum*

Figure 5.1: BCA of some eukaryotic organisms.

We can see that, for what concerns Eukaryotes, the general feature, distinctive of their position along the phylogenetic tree, is the presence of a positive gradient of strong bases (CG) and a complementary negative one in the weak bases (AT). It is worth pointing out that the BCA in eukaryotic promoters results from the merging of two different populations (obeying different statical rules, if any), i.e. the TATA-less promoters that exhibit the above mentioned gradient and the TATA-with promoters, where a split between weak and strong basis (in favour of the former) is observed, apart a very short region close to the TSS. Going down the phylogenetic tree one observes that the gradient region becomes shorter until it disappears. On the other hand, in *E. coli* and other similar bacteria, the BCA of IGSs exhibits a positive gradient of weak bases (and a negative one of strong bases), although there are exceptions, like the bacterium *S. coelicolor*, that lives in hot environments.

### 5.1.1 Constrains in eukaryotic BCA

First of all let us introduce the discrete variable  $\ell = -1000, \dots, -1$  to indicate the position of a nucleotide along a sequence and the functions  $A_r(\ell)$ ,  $T_r(\ell)$ ,  $G_r(\ell)$  and  $C_r(\ell)$  that are the real profiles of the nucleotide frequencies appearing in the BCA of a chosen organism. By observing in Figure 5.1 the trends of the four nucleotides in function of the position in different organisms we can speculate that some constraints among their abundance exist. To better analyze this aspect we introduce also the ideal continuous functions  $A(\ell)$ ,  $T(\ell)$ ,  $G(\ell)$  and  $C(\ell)$  (we evaluate them only in the discrete positions located by  $\ell$ ) that respect exactly these constraints. This is an heuristic hypothesis that we'll verify a posteriori. There is the obvious relation  $A(\ell) + T(\ell) + G(\ell) + C(\ell) = 1$  due to the definition of BCA. The other constrains we hypothesize for the continuous functions, valid for the range  $\ell = -1000, \dots, -45$  (in order to avoid the TATA-box region where we can find particular motifs), are:

- 1)  $A(\ell) = T(\ell)$  and  $G(\ell) = C(\ell)$ ,
- 2)  $A(\ell) + T(\ell) \equiv W(\ell) = ae^{b\ell} + ce^{d\ell} + f$ ,
- 3)  $A(\ell) + G(\ell) = 0.5$  and  $T(\ell) + C(\ell) = 0.5$ ,
- 4)  $A(\ell) + C(\ell) = 0.5$  and  $T(\ell) + G(\ell) = 0.5$ ,

where  $a, b, c, d, f$  are constants that depend on the considered organisms and have been found by a fit procedure (this aspect will be analyzed diffusely in the next section). The

constrain 3) can be read as the abundance of the purines and pyrimidines for every position along the chain is constant, and in particular it is 50% for both.

To quantify the correctness of the previous relations we use the following procedure that illustrate, for simplicity, for the first one. We calculate the best fit function  $P_{20}(\ell)$  (we used a degree twenty polynomial) for the difference  $A_r(\ell) - T_r(\ell)$  and we evaluate the corresponding chi-squared defined as

$$\chi_r = \sqrt{\sum_{\ell=-1000}^{-45} (A_r(\ell) - T_r(\ell) - P_{20}(\ell))^2}. \quad (5.1)$$

We do the same for the function that corresponds to the constraint, in this case the null constant function, obtaining the chi-squared

$$\chi = \sqrt{\sum_{\ell=-1000}^{-45} (A_r(\ell) - T_r(\ell) - 0)^2}. \quad (5.2)$$

In general, the goodness of the  $j$ -th bond relation with  $j = 1, 2, 3, 4$  can be estimated thanks to the parameter obtained by the ratio between the two chi-squared:

$$c_j = \frac{\chi}{\chi_r}. \quad (5.3)$$

We will say that the lower the value of  $c_j$  is the better the hypothesized constrain for that organism is imposed. In the following figures we report the  $c_j$  values for every species (similar colors indicate closer species in the phylogenetic tree).

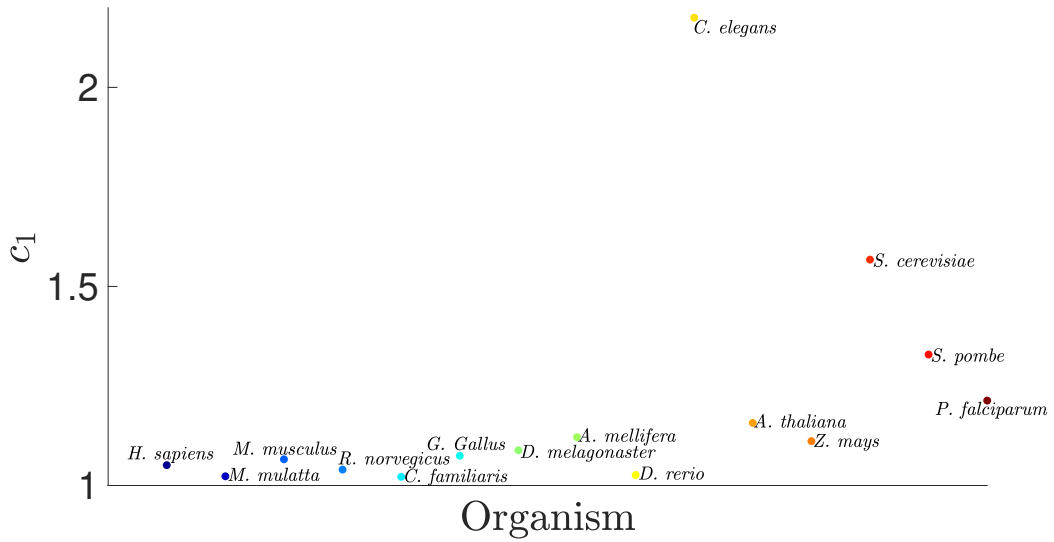


Figure 5.2: On the vertical axes the values of  $c_1$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

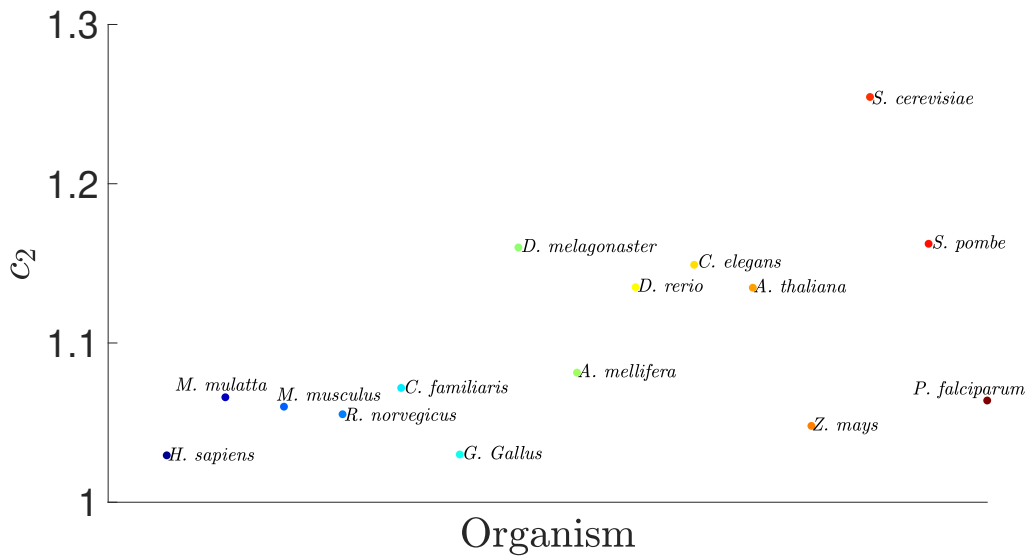


Figure 5.3: On the vertical axes the values of  $c_2$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

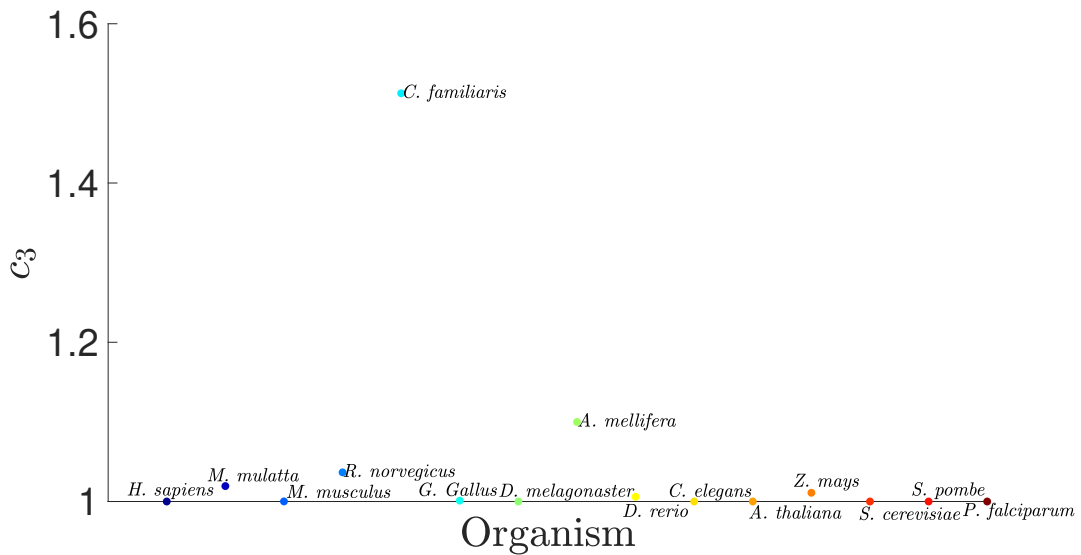


Figure 5.4: On the vertical axes the values of  $c_3$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

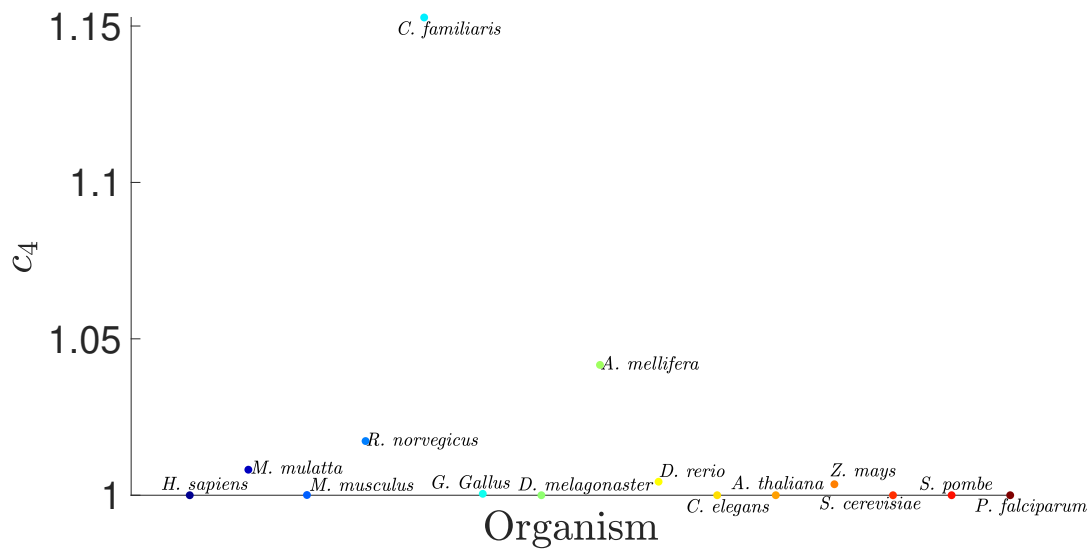


Figure 5.5: On the vertical axes the values of  $c_4$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

We can see the all the four constraints are valid in first approximation for every studied organism. We'll analyze in more detail the consequences represented by the second constraint.



### 5.1.2 Reproducing the spatial distribution of weak and strong bases for the BCA of eukaryotic IGSs

In this section we consider the IGSs of each organism as binary sequence of two nucleotides: A and T are substituted by a W, G and C by a S. The constraint 2) tells us that, for the IGSs of an eukaryote<sup>1</sup>, the average behavior of the weak nucleotides in function of the position along the sequence can be described in first approximation by the sum of two exponentials translated by a proper constant. We report in Figure 5.6 the BCA for every studied organism calculated for the binary IGSs with the fits obtained<sup>2</sup> estimating the parameters of the function

$$W(\ell) = ae^{b\ell} + ce^{d\ell} + f. \quad (5.4)$$

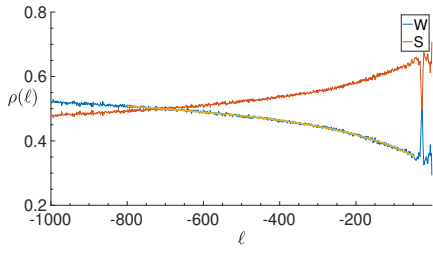
If we consider the corresponding continuous function  $W(x)$  of the variable  $x \in [-800, -45]$ , we have that  $W(x)$  satisfy the following second order differential equation:

$$\frac{d^2}{dx^2}W(x) - (b + d)\frac{d}{dx}W(x) + bdW(x) - bdf = 0. \quad (5.5)$$

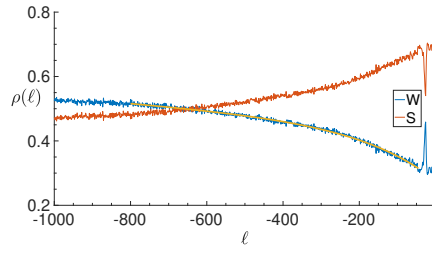
---

<sup>1</sup>We can speculate that the same constraint exists also for prokaryotes but we have to reduce the length in term of bps of the IGSs since the noncoding portion of the genome is very restricted.

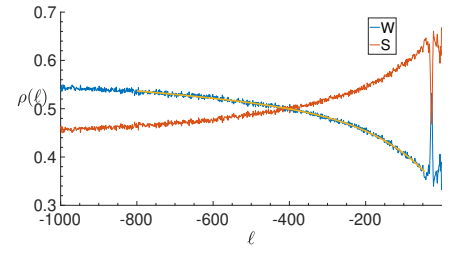
<sup>2</sup>We have fitted the function  $W_r(\ell)$  from  $\ell = -800$  to  $\ell = -45$  to minimize boundary effects.



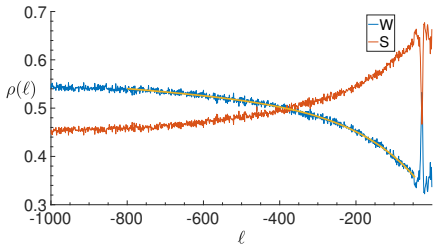
(a) *H. sapiens*



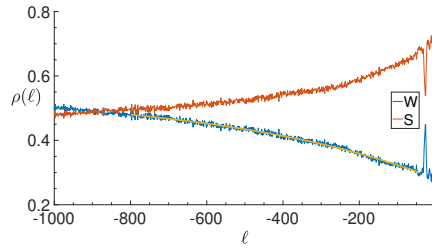
(b) *M. mulatta*



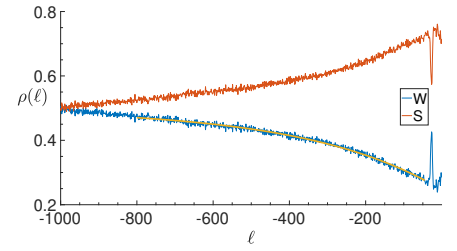
(c) *M. musculus*



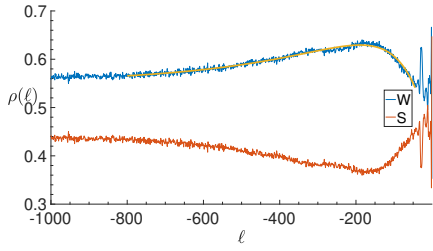
(d) *R. norvegicus*



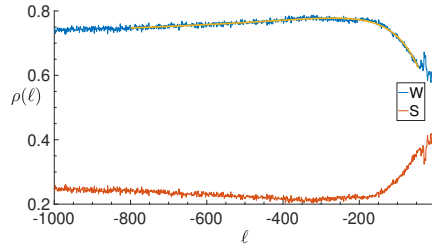
(e) *C. familiaris*



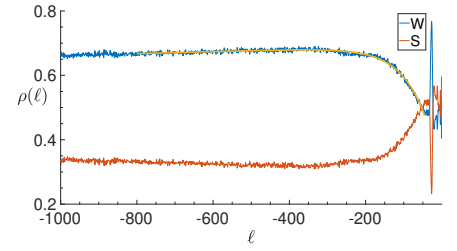
(f) *G. gallus*



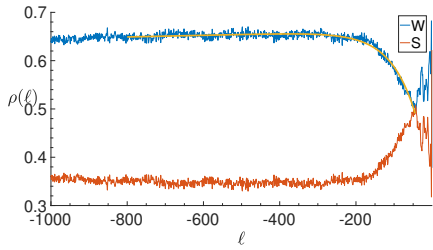
(g) *D. melanogaster*



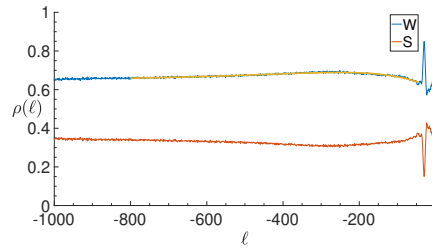
(h) *A. mellifera*



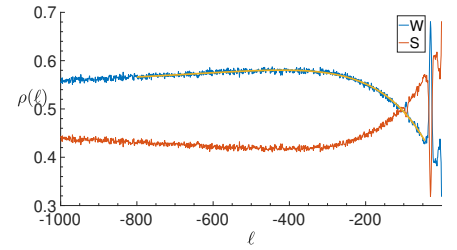
(i) *D. rerio*



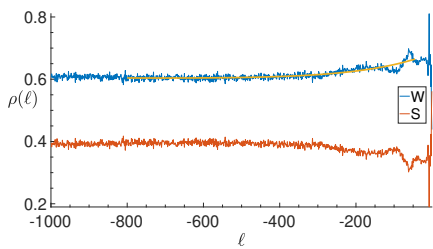
(j) *C. elegans*



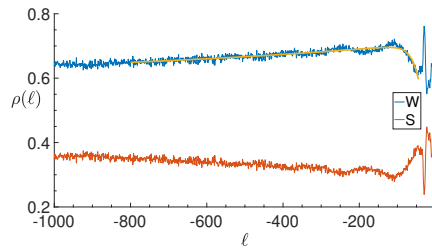
(k) *A. thaliana*



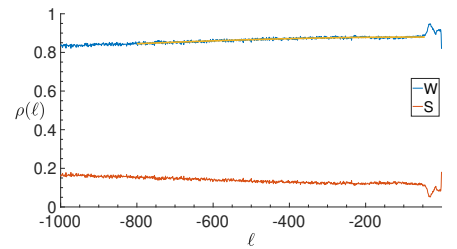
(l) *Z. mays*



(m) *S. cerevisiae*



(n) *S. pombe*



(o) *P. falciparum*

Figure 5.6: BCA of weak W and strong S bases in some eukaryotic organisms. The thick yellow lines correspond to the functions  $W(\ell) = ae^{b\ell} + ce^{d\ell} + f$ .

Each species is characterized by a different set of parameters that we can divide into two types based on their dimensions:  $a$ ,  $c$  and  $f$  with the dimension of a frequency (rate of a nucleotide for a given position  $\ell$ ) on one hand,  $b$  and  $d$  on the other hand that have the dimensions of an inverse length. We can construct a three-dimensional space where to arrange the species taken into consideration and where the coordinates are given by the three parameters  $a$ ,  $c$  and  $f$ , see Figure 5.7.

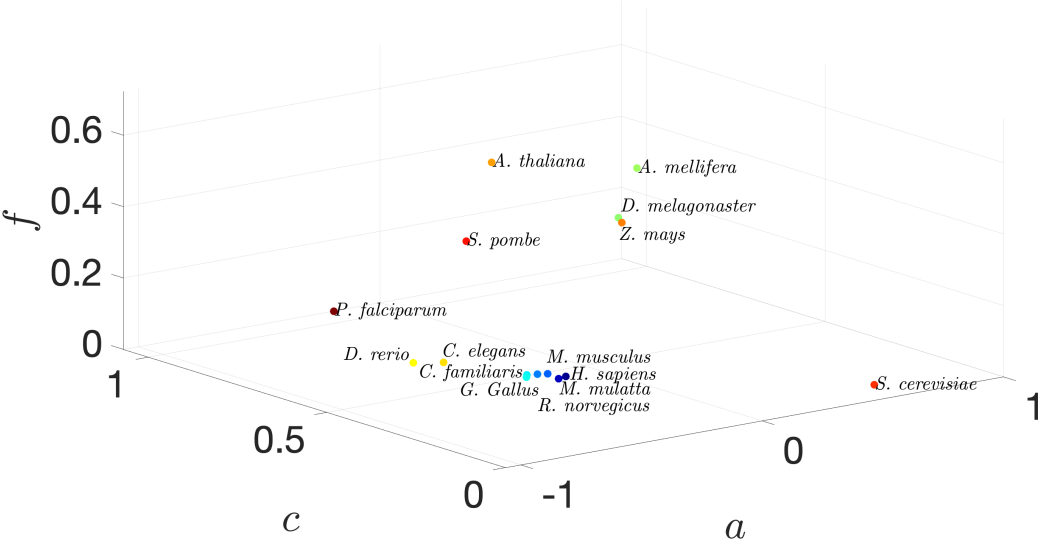


Figure 5.7: Parameter space with coordinates  $a$ ,  $c$  and  $f$ .

We can do the same for the couple of parameters  $b$  and  $d$  obtaining the two-dimensional space represented in Figure 5.8.

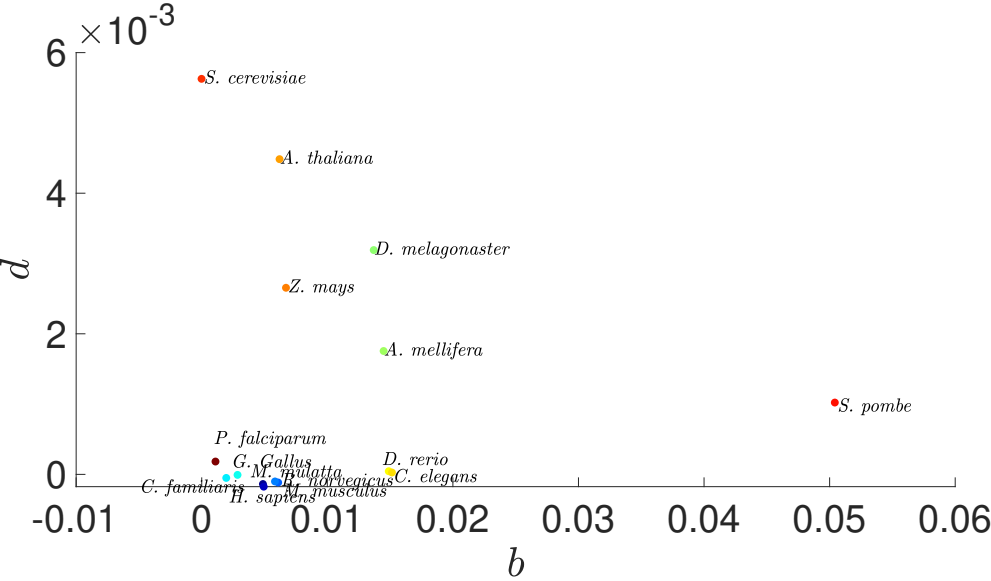


Figure 5.8: Parameter space with coordinates  $b$  and  $d$ .

Now we can evaluate the euclidian distance of a point representing a species from

the point of *H. sapiens*. This distance can quantify how much the species is far from *H. sapiens* from an evolutive point of view. Let  $D_C$  the distance in the three-dimensional space and  $D_L$  the one in the other space calculated for every organisms. In Figure 5.9 and Figure 5.10 we report the values of these distances for every species.

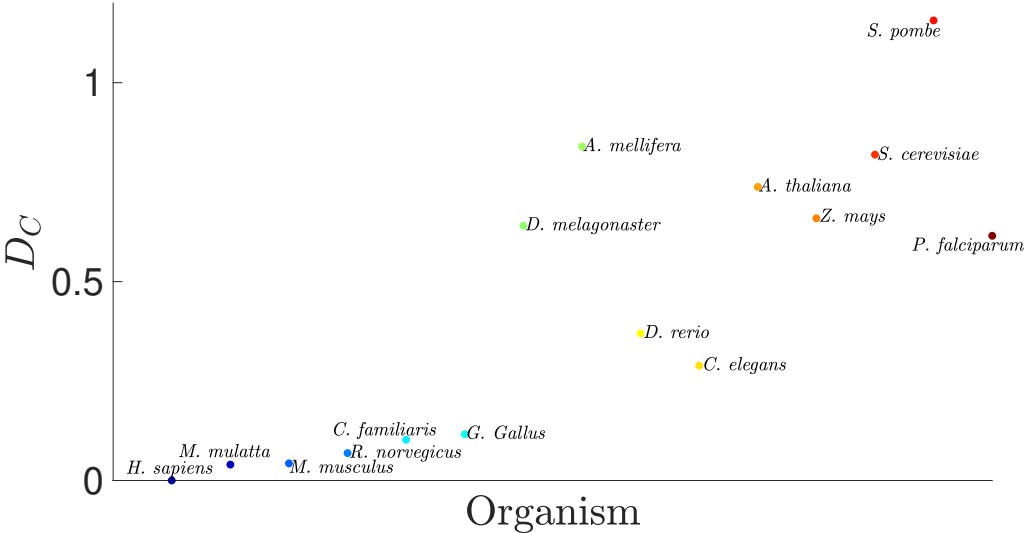


Figure 5.9: On the vertical axes the values of  $D_C$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

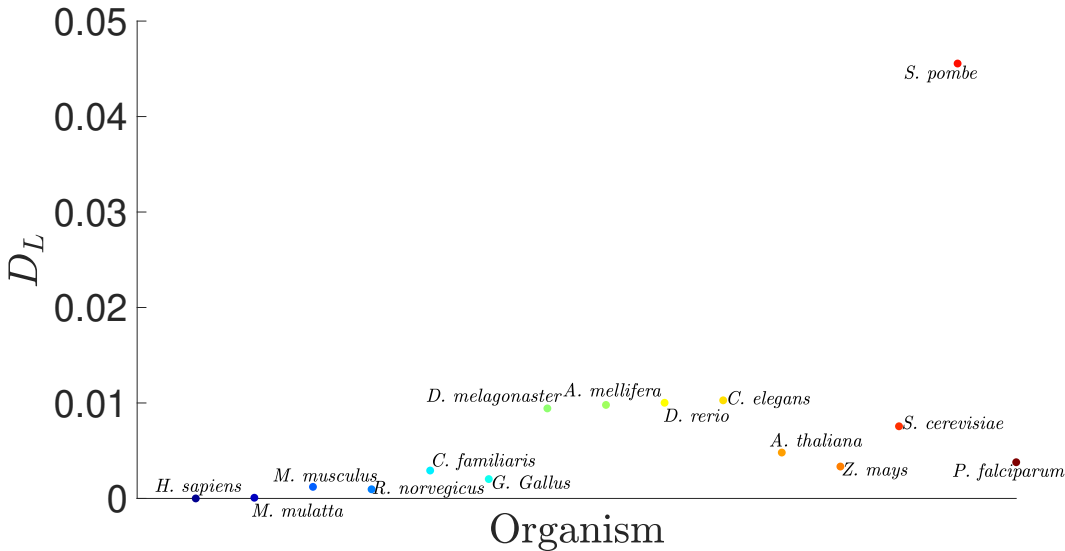


Figure 5.10: On the vertical axes the values of  $D_L$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

We can notice that the distribution of the distance  $D_C$  shows a growing trend moving away from *H. sapiens*, more pronounced than that of  $D_L$ .

Always observing the obtained fit, we can notice that, especially for the less evolved

organisms, the true profiles present a sort of undulation. We hypothesize that a better fit function could be the following one with two sinusoidal modulations, one for each exponential:

$$W_{\sin}(\ell) = ae^{b\ell}[1 + g_1 \sin(h_1\ell + \phi_1)] + ce^{d\ell}[1 + g_2 \sin(h_2\ell + \phi_2)] + f. \quad (5.6)$$

To evaluate the goodness of the previous choice for the fit function we calculate again the chi-squared

$$\chi_{\sin} = \sqrt{\sum_{\ell=-800}^{-45} (A_r(\ell) + T_r(\ell) - W_{\sin}(\ell))^2}, \quad (5.7)$$

and we compare it with the corresponding chi-squared calculated for the function  $W(\ell)$  (5.4):

$$\chi = \sqrt{\sum_{\ell=-800}^{-45} (A_r(\ell) + T_r(\ell) - W(\ell))^2}, \quad (5.8)$$

The results are reported in Figure 5.11 and show an improvement with the choice of this second fit function.

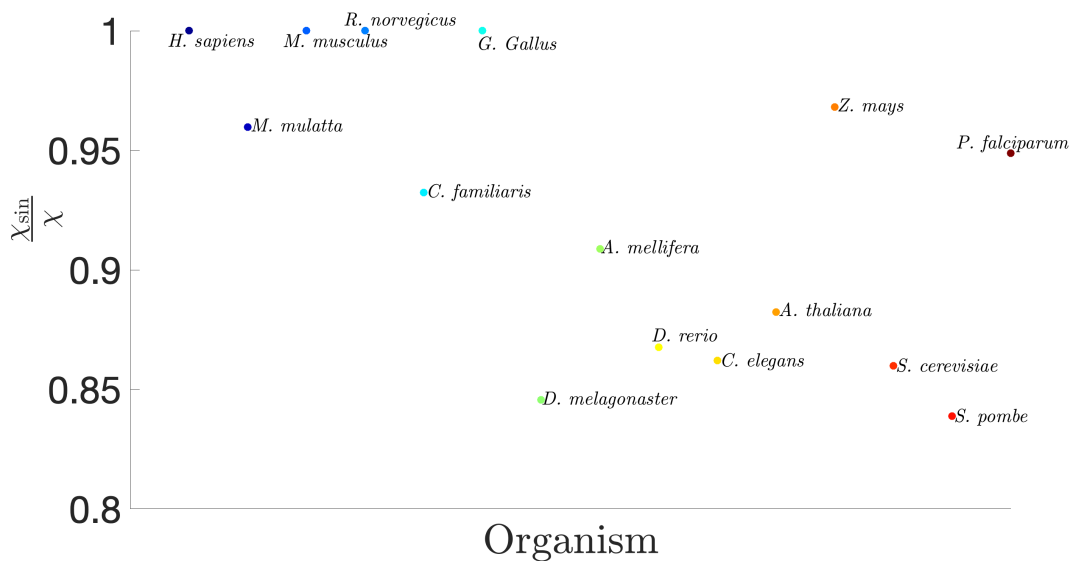


Figure 5.11: On the vertical axes the ratio of  $\chi_{\sin}/\chi$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

The “new” fits obtained with the function of (5.6) are shown in Figure 5.12: the quality of the fits is so good that it is hardly appreciable by the sides of figures.

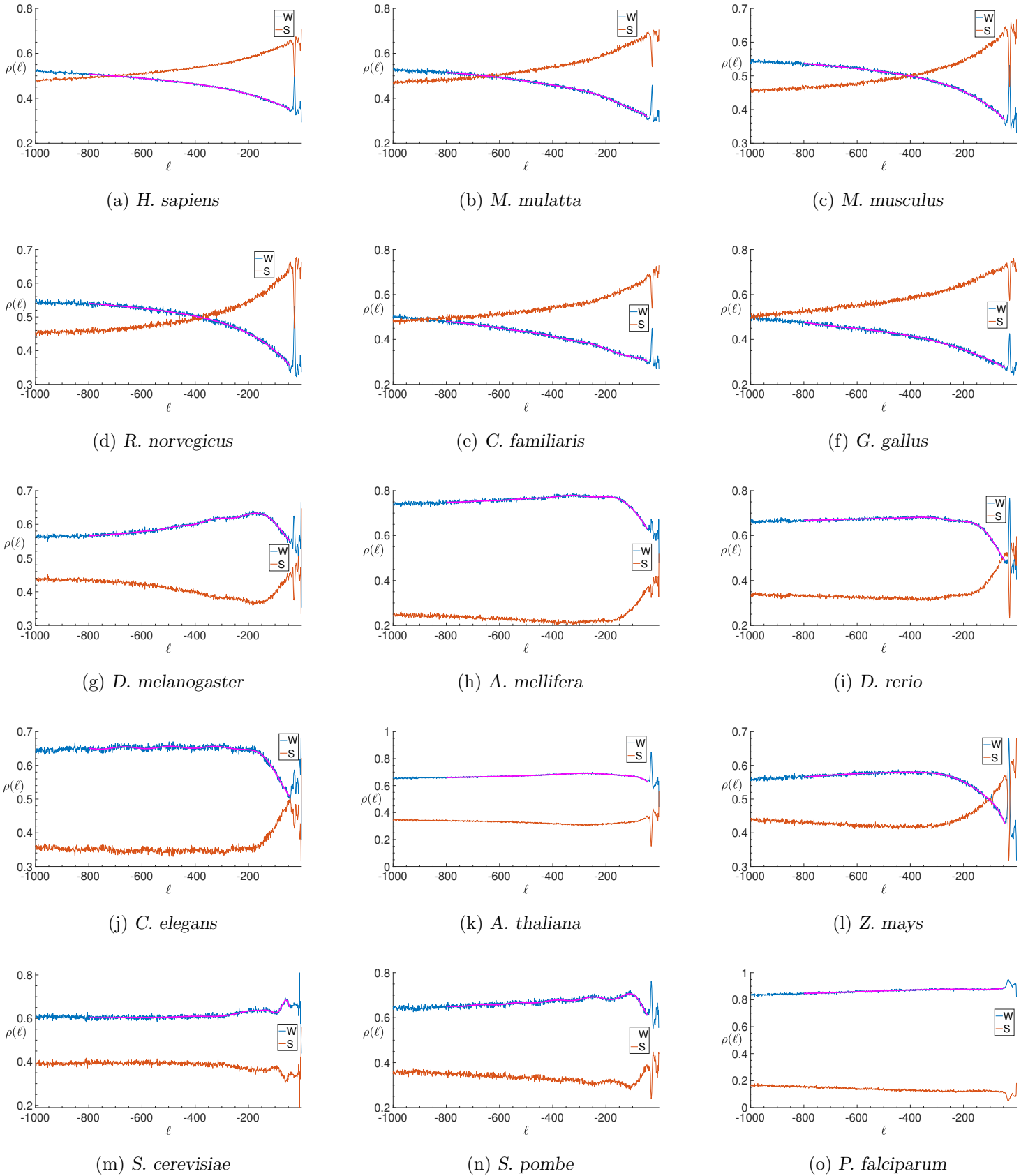


Figure 5.12: BCA of weak W and strong S bases in some eukaryotic organisms. The thick magenta lines correspond to the functions  $W_{\sin}(\ell) = ae^{b\ell}[1 + g_1 \sin(h_1\ell + \phi_1)] + ce^{d\ell}[1 + g_2 \sin(h_2\ell + \phi_2)] + f$ . The overlap between the fit functions and the profiles of the BCA makes the latter not clearly visible.

Following the previous logic we can build the two artificial spaces: the second one will become four-dimensional because we add the third and fourth coordinates to  $b$  and  $d$ , i.e.  $h_1$  and  $h_2$ , which have the same dimension. We proceed with the calculation of the distance of every species from  $H. sapiens$  point.

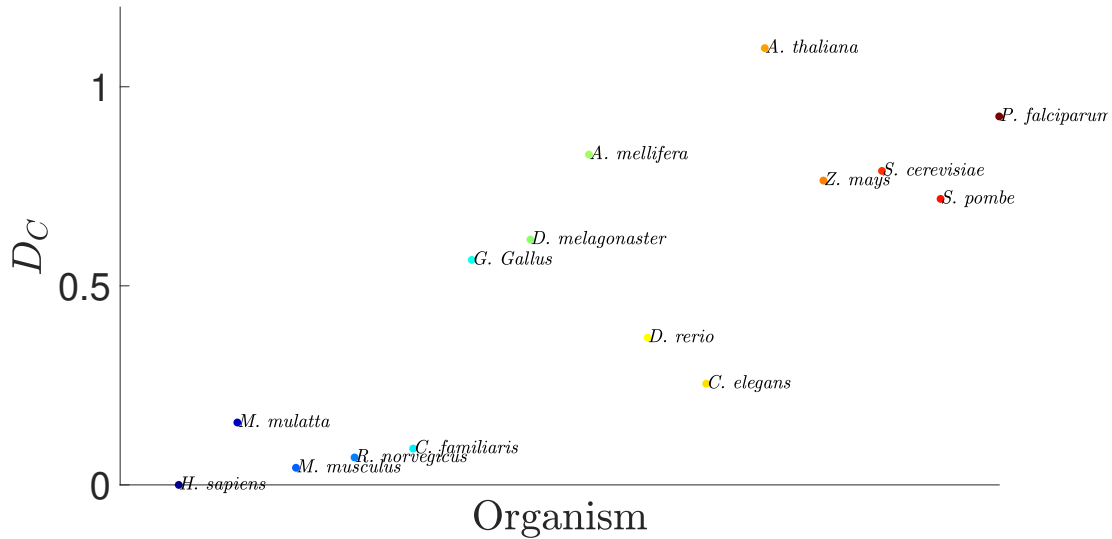


Figure 5.13: On the vertical axes the values of  $D_C$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

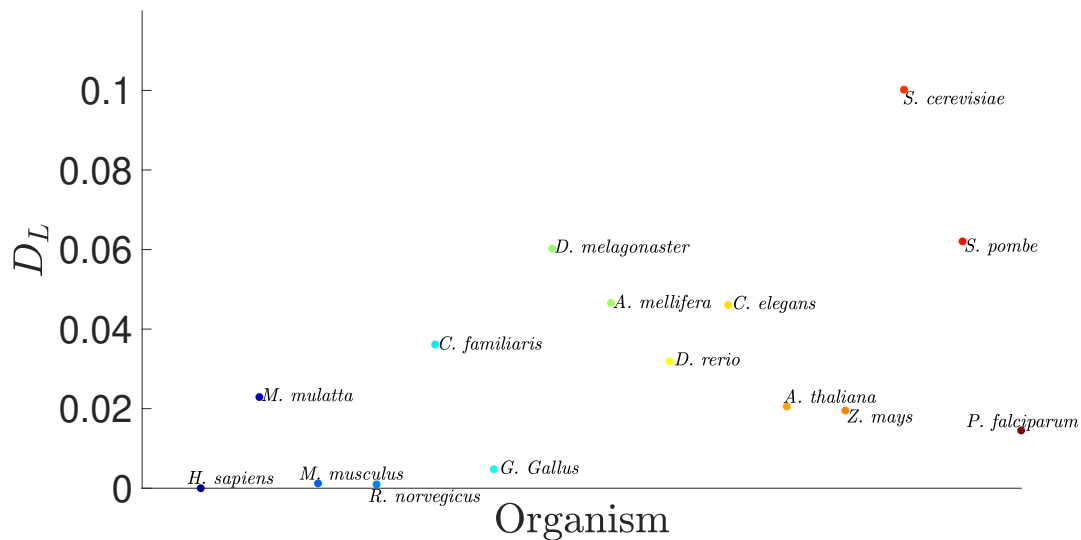


Figure 5.14: On the vertical axes the values of  $D_L$  are reported for different species ordered on the horizontal axis according to the phylogenetic tree.

Now the growing trend is more pronounced also for the distance  $D_L$ .

This classification based on the fits is consistent with the one of the phylogenetic tree, even if not completely equivalent. We can conclude that the structural properties of the

IGSs are correlated with evolution.

## 5.2 BCA of genes in *H. sapiens* and *E. coli*

Protein-coding DNA sequences are characterized by the long-known three-base periodicity property. The reason for this periodicity is probably due to the biased distribution towards codon triplets, which is a consequence of genetic code degeneracy. However, the way in which three-base periodicity is produced has not been elucidated [100]. The initial discovery was made in 1980 observing periodicity in DNA sequences by applying the autocorrelation function to chromatin DNA [101]. It was confirmed by a spectral analysis of DNA sequences defining the Fourier transform of a sequence of bases [102], and then was compared the Fourier analysis of coding, noncoding and random sequences and was proposed a reason for the 3-periodicity property found in coding sequences [103].

First we define the BCA of the coding sequences, aligned starting from the TSC. The genes have different lengths in terms of base pairs, so in the BCA formula the normalization constant also depends on the position along the sequence:

$$\rho_x(\ell) = \frac{1}{N(\ell)} \sum_{i=1}^{N(\ell)} s_i^x(\ell). \quad (5.9)$$

We consider sequences of  $L = 1000$  nucleotide maximum length, from  $\ell = 0$  (first site of the TSC) to  $\ell = 999$ . We visualize the gene as a sequence of codons in order to identify the first, second and third nucleotide according to the position that they occupy in the codon which they belong to. In the panels on the left of Figure 5.15 and Figure 5.16 the BCA of a coding region are shown, respectively for *H. sapiens* and *E. coli*. The four signals exhibit a periodic behavior. Let's introduce now the power spectrum of the signal of G (we could have chosen any other nucleotide):

$$P(k) = \left| \sum_{\ell=0}^{L-1} G_r(\ell) e^{-\frac{2\pi k \ell}{L} i} \right|^2, \quad (5.10)$$

where the index  $k$  of the frequency response corresponds to an angular frequency  $\omega(k) = 2\pi k/N$ . In the panels on the right of Figure 5.15 and Figure 5.16 we report the power spectrum of  $G_r(\ell)$ , respectively for *H. sapiens* and *E. coli*, from which we can see that there is a peak for  $k = N/3$ , that is  $\omega = 2\pi/3 \simeq 2.1$ , corresponding to a period of 3 bps.

To better understand this aspect we can think of a gene as a discrete quaternary signal



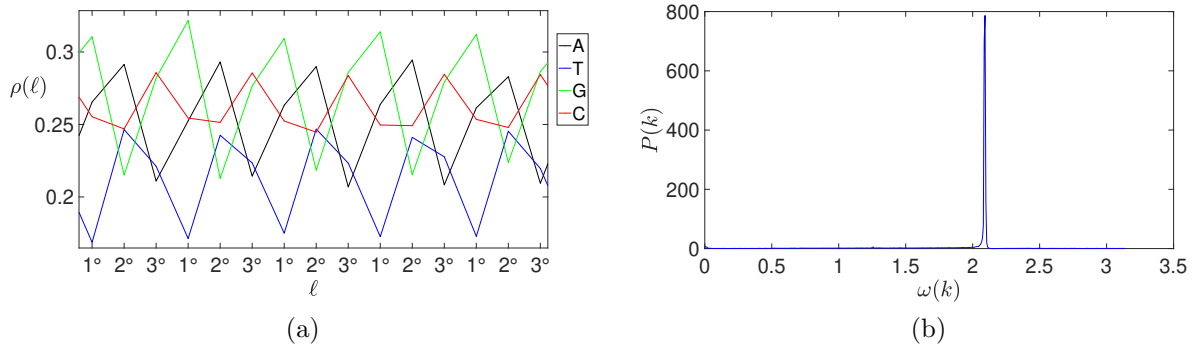


Figure 5.15: *H. sapiens*. BCA of a coding portion and relative power spectrum.

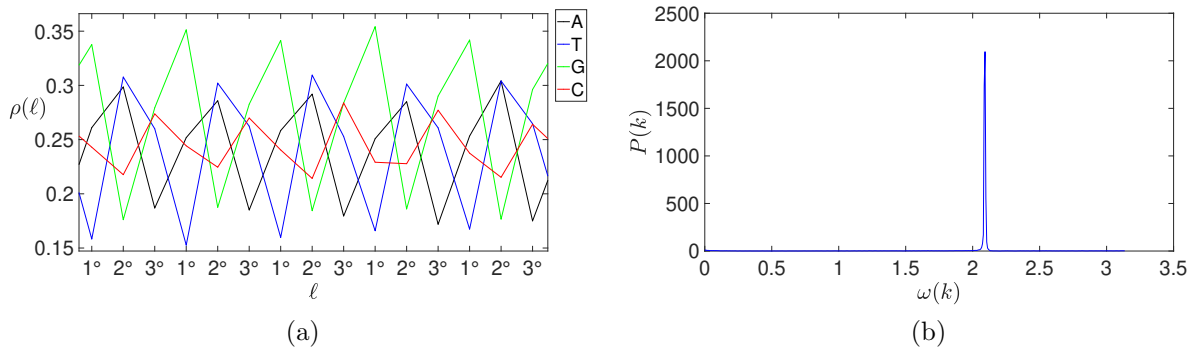


Figure 5.16: *E. coli*. BCA of a coding portion and relative power spectrum.

characterized by the 3-base periodicity property. In principle, by averaging signals not in phase but with the same periodicity, we should obtain a same periodicity signal in which the amplitude is more decreased the greater the number of added signals. Contrary to what we would have expected, we see from Figure 5.15 and Figure 5.16 that the BCA maintains a clear 3-base periodicity. Moreover the same positions within a triplet exhibit similar nucleotide frequencies. The differences of the abundance of the four nucleotides in the three positions can be attributed to the codon usage bias, that refers to differences in the frequency of occurrence of synonymous codons (different codons that encode the same amino acid) in coding DNA (see data of GenScript Codon Usage Frequency database in Figure 5.17 and Figure 5.18).

For example, we consider all the codons that have the nucleotide A in the second position; the sum of their frequency will give the probability of finding A in the second position of the triplette. We repeat the same calculation for all the nucleotides in every position (values are reported in Table 5.1 and Table 5.2 in the columns indicated by  $1^\circ_{GS}$ ,  $2^\circ_{GS}$  and  $3^\circ_{GS}$ ). Now we want to compare these values with the frequencies obtained in the BCA. We calculate the average frequencies of each nucleotide for all the positions,  $1^\circ$ ,  $2^\circ$  and  $3^\circ$ . For example, to obtain the average frequency of A in the second position in

Triplet	Amino acid	Fraction	Frequency/ Thousand	Number	Triplet	Amino acid	Fraction	Frequency/ Thousand	Number
TTT	F	0.45	16.9	336562	TCT	S	0.18	14.6	291040
TTC	F	0.55	20.4	406571	TCC	S	0.22	17.4	346943
TTA	L	0.07	7.2	143715	TCA	S	0.15	11.7	233110
TTG	L	0.13	12.6	249879	TCG	S	0.06	4.5	89429
TAT	Y	0.43	12.0	239268	TGT	C	0.45	9.9	197293
TAC	Y	0.57	15.6	310695	TGC	C	0.55	12.2	243685
TAA	*	0.28	0.7	14322	TGA	*	0.52	1.3	25383
TAG	*	0.20	0.5	10915	TGG	W	1.00	12.8	255512
CTT	L	0.13	12.8	253795	CCT	P	0.28	17.3	343793
CTC	L	0.20	19.4	386182	CCC	P	0.33	20.0	397790
CTA	L	0.07	6.9	138154	CCA	P	0.27	16.7	331944
CTG	L	0.41	40.3	800774	CCG	P	0.11	7.0	139414
CAT	H	0.41	10.4	207826	CGT	R	0.08	4.7	93458
CAC	H	0.59	14.9	297048	CGC	R	0.19	10.9	217130
CAA	Q	0.25	11.8	234785	CGA	R	0.11	6.3	126113
CAG	Q	0.75	34.6	688316	CGG	R	0.21	11.9	235938
ATT	I	0.36	15.7	313225	ACT	T	0.24	12.8	255582
ATC	I	0.48	21.4	426570	ACC	T	0.36	19.2	382050
ATA	I	0.16	7.1	140652	ACA	T	0.28	14.8	294223
ATG	M	1.00	22.3	443795	ACG	T	0.12	6.2	123533
AAT	N	0.46	16.7	331714	AGT	S	0.15	11.9	237404
AAC	N	0.54	19.5	387148	AGC	S	0.24	19.4	385113
AAA	K	0.42	24.0	476554	AGA	R	0.20	11.5	228151
AAG	K	0.58	32.9	654280	AGG	R	0.20	11.4	227281
GTT	V	0.18	10.9	216818	GCT	A	0.26	18.6	370873
GTC	V	0.24	14.6	290874	GCC	A	0.40	28.5	567930
GTA	V	0.11	7.0	139156	GCA	A	0.23	16.0	317338
GTG	V	0.47	28.9	575438	GCG	A	0.11	7.6	150708
GAT	D	0.46	22.3	443369	GGT	G	0.16	10.8	215544
GAC	D	0.54	26.0	517579	GGC	G	0.34	22.8	453917
GAA	E	0.42	29.0	577846	GGA	G	0.25	16.3	325243
GAG	E	0.58	40.8	810842	GGG	G	0.25	16.4	326879

Figure 5.17: *H. sapiens*. GenScript Codon Usage Frequency Table. For each codon, the table displays the frequency of usage of each codon (second column) and the relative frequency of each codon among synonymous codons (first column).

Triplet	Amino acid	Fraction	Frequency/ Thousand	Number	Triplet	Amino acid	Fraction	Frequency/ Thousand	Number
TTT	F	0.58	22.1	80995	TCT	S	0.17	10.4	38027
TTC	F	0.42	16.0	58774	TCC	S	0.15	9.1	33430
TTA	L	0.14	14.3	52382	TCA	S	0.14	8.9	32715
TTG	L	0.13	13.0	47500	TCG	S	0.14	8.5	31146
TAT	Y	0.59	17.5	63937	TGT	C	0.46	5.2	19138
TAC	Y	0.41	12.2	44631	TGC	C	0.54	6.1	22188
TAA	*	0.61	2.0	7356	TGA	*	0.30	1.0	3623
TAG	*	0.09	0.3	989	TGG	W	1.00	13.9	50991
CTT	L	0.12	11.9	43449	CCT	P	0.18	7.5	27340
CTC	L	0.10	10.2	37347	CCC	P	0.13	5.4	19666
CTA	L	0.04	4.2	15409	CCA	P	0.20	8.6	31534
CTG	L	0.47	48.4	177210	CCG	P	0.49	20.9	76644
CAT	H	0.57	12.5	45879	CGT	R	0.36	20.0	73197
CAC	H	0.43	9.3	34078	CGC	R	0.36	19.7	72212
CAA	Q	0.34	14.6	53394	CGA	R	0.07	3.8	13844
CAG	Q	0.66	28.4	104171	CGG	R	0.11	5.9	21552
ATT	I	0.49	29.8	109072	ACT	T	0.19	10.3	37842
ATC	I	0.39	23.7	86796	ACC	T	0.40	22.0	80547
ATA	I	0.11	6.8	24984	ACA	T	0.17	9.3	33910
ATG	M	1.00	26.4	96695	ACG	T	0.25	13.7	50269
AAT	N	0.49	20.6	75436	AGT	S	0.16	9.9	36097
AAC	N	0.51	21.4	78443	AGC	S	0.25	15.2	55551
AAA	K	0.74	35.3	129137	AGA	R	0.07	3.6	13152
AAG	K	0.26	12.4	45459	AGG	R	0.04	2.1	7607
GTT	V	0.28	19.8	72584	GCT	A	0.18	17.1	62479
GTC	V	0.20	14.3	52439	GCC	A	0.26	24.2	88721
GTA	V	0.17	11.6	42420	GCA	A	0.23	21.2	77547
GTG	V	0.35	24.4	89265	GCG	A	0.33	30.1	110308
GAT	D	0.63	32.7	119939	GGT	G	0.35	25.5	93325
GAC	D	0.37	19.2	70394	GGC	G	0.37	27.1	99390
GAA	E	0.68	39.1	143353	GGA	G	0.13	9.5	34799
GAG	E	0.32	18.7	68609	GGG	G	0.15	11.3	41277

Figure 5.18: *E. coli*. GenScript Codon Usage Frequency Table. For each codon, the table displays the frequency of usage of each codon (second column) and the relative frequency of each codon among synonymous codons (first column).

BCA data, we do the average of the frequency of A for all the 2° position in the first 1000 bps:

$$\bar{A}_{2^\circ} = \frac{1}{N_{2^\circ}} \sum_{\ell \in \{2^\circ\}} A_r(\ell), \quad (5.11)$$

where the sum runs over the second position of the codons and  $N_{2^\circ}$  is the number of these positions in the first 1000 bps. The corresponding error will be given by the standard deviation. The values are reported in Table 5.1 and Table 5.2 in the columns indicated by 1°, 2° and 3°. The values not compatible with those of the GenScript Codon Usage Frequency database are highlighted in red.

Table 5.1: *H. sapiens*. The columns relative to the position with  $_{GS}$  contain the value in percentage obtained from GenScript Codon Usage Frequency, in the others the ones calculated from BCA.

	1° <sub>GS</sub>	1°	2° <sub>GS</sub>	2°	3° <sub>GS</sub>	3°
A	26.68	26.4 ± 3.9	31.17	29.1 ± 1.7	18.88	20.8 ± 1.3
T	17.03	16.9 ± 1.1	26.44	24.7 ± 4.0	21.83	22.0 ± 1.2
G	31.65	31.6 ± 1.9	19.10	21.5 ± 1.3	29.07	28.7 ± 3.8
C	24.64	25.0 ± 1.5	23.29	24.8 ± 1.4	30.22	28.6 ± 1.8

Table 5.2: *E. coli*. The columns relative to the position with  $_{GS}$  contain the value in percentage obtained from GenScript Codon Usage Frequency, in the others the ones calculated from BCA.

	1° <sub>GS</sub>	1°	2° <sub>GS</sub>	2°	3° <sub>GS</sub>	3°
A	26.24	25.2 ± 4.1	29.62	28.7 ± 2.2	19.38	18.2 ± 2.0
T	16.05	15.9 ± 1.2	29.69	30.6 ± 4.0	27.28	26.1 ± 1.9
G	34.58	34.8 ± 2.5	17.98	18.0 ± 1.4	27.84	28.9 ± 4.2
C	23.13	24.1 ± 1.6	22.71	22.6 ± 1.6	25.50	26.7 ± 1.8

We want to verify if also for coding sequences the constraints 3) and 4) of subsection 5.1.1 are respected. Results are reported in Table 5.3 and Table 5.4 (the values highlighted in red are not compatible with the constraints). We can see that both con-

Table 5.3: *H. sapiens*. Constraints of the BCA (the values are reported in percentage).

	1°	2°	3°
A+G	58.1 ± 2.4	50.5 ± 2.7	49.5 ± 2.9
T+C	41.9 ± 2.4	49.5 ± 2.7	50.5 ± 2.9
A+C	51.5 ± 2.6	53.9 ± 2.9	49.3 ± 2.7
T+G	48.5 ± 2.6	46.1 ± 2.9	50.7 ± 2.7

Table 5.4: *E. coli*. Constraints of the BCA (the values are reported in percentage).

	1°	2°	3°
A+G	60.0 ± 2.4	46.7 ± 3.0	47.1 ± 3.1
T+C	40.0 ± 2.4	53.3 ± 3.0	52.9 ± 3.1
A+C	49.3 ± 2.9	51.3 ± 3.2	44.9 ± 2.7
T+G	50.7 ± 2.9	48.7 ± 3.2	55.1 ± 2.7

straints are not always respected, but that the validity of one often excludes the validity

of the other. Therefore we can conclude that the type of satisfied constraint for nucleotide abundances depends on the position considered in the codon triplet, but the sum of the abundance of a weak base with a strong one (the purines A+G and the pyrimidines T+C or A+C and T+G) is always compatible with 50% in all of the three positions.

# Chapter 6

## Conclusions

In this thesis we have developed a method to identify for a given prokaryotic organism same length noncoding sequences of DNA (the IGSs), that do not even undergo the transcription process by RNA polymerase. The sequences obtained so far have been compared pairwise and then, thanks to a clustering procedure, they have been divided into groups based on structural similarities highlighted by BCA.

As next step we have characterized from a biological point of view the obtained clusters for each organism. We have used two databases in order to do this. The first one, STRING, has allowed us to create biological networks among the genes regulated by the corresponding IGSs. We have built co-expression and co-occurrence networks and with a statistical procedure we have seen which cluster gave rise to networks with different features respect to networks created by choosing randomly IGSs. The second database used, COG, has allowed us to univocally associate a biological function to each IGS; then, through a functional enrichment analysis, a functional category in a cluster has been seen to be over- or down-represented respect to the genome background. Results have highlighted that, for the bacteria for which we have made the previous analysis, all the clusters (with the exception of C0 in *P. haloplanktis*) exhibit both over-represented and under-represented functions. Sometimes the same function is over-represented in one cluster and down-represented in another. Moreover for the cluster C2, the one with a greater predominance of A, we have obtained the same identical results both for *B. subtilis* and for *P. haloplanktis*. It would be interesting to further investigate biological correlations using other databases and to expand the study to other prokaryotic organisms by looking for possible similarities between IGSs belonging to different species.

Focusing on the fact that in correspondence with the IGSs the DNA is opened to begin

transcription, we have performed a thermodynamic analysis at the denaturing process. To model the nucleotide chain we have used the Dauxois-Peyrard-Bishop model in which, among the various contributes to the potential energy, there is the one given by the Morse potential which takes into account the transverse bonds between the two opposite strands, introducing a difference for what concerns weak and strong bases. This effect is evident in the denaturation curves obtained for *E. coli*, *B. subtilis* and *P. haloplanktis* compared with that of *S. coelicolor*. The IGSs of the last bacterium are rich of strong bases, unlike those of the other three that instead have a similar BCA with a predominance of weak bases; this has the consequence of having obtained a denaturation temperature higher for *S. coelicolor*. The study of the denaturing properties for individual clusters had the merit of highlighting the importance of the changes we have made to Dauxois-Peyrard-Bishop model. In fact, if in Dauxois-Peyrard-Bishop model the four nucleotides had the same mass, we have introduced a “degeneration” in the masses in the symplectic algorithm reproducing the dynamics of the chain coupled with a thermostat. The cluster C1 and C2 in *B. subtilis*, on the base of the division in weak and strong bases, should have provided the same result for  $T_d$ . On the other hand, we can observe that the  $T_d$  is sensibly lower for C1. In fact, among the found clusters the C1 of *B. subtilis* is the one with the largest separation between T, the nucleotide with the smallest mass, and G, the nucleotide with the biggest mass. A natural development of this line of research would be to apply the study of denaturation dynamics to eukaryotic IGSs, although it must be taken into account that this would involve a much higher computational cost.

For what concerns the understanding of the constraints observed in the BCA of non-coding sequences in eukaryotes, it would be important to improve and extend this analysis in order to reveal if and how the constraints are correlated with evolution or depending on combination with other causes. At the present stage of the research it is a fully open problem.

# Chapter 7

## Ringraziamenti

Prima di tutto vorrei ringraziare Roberto Livi, che mi ha guidato per tutti questi tre anni del dottorato, un grande esempio di scienziato e di profonda umanità. Francesca Di Patti, supporto fondamentale soprattutto nella fase iniziale, senza la quale probabilmente sarei ancora a premere il pulsante di accensione del computer. Marco Fondi, Renato Fani e Alessio Mengoni per aver collaborato e avermi aiutato nella stesura del mio primo lavoro. Stefano Lepri e Stefano Luccioli per aver condiviso con me la loro esperienza e avermi dato un contributo fondamentale per il mio secondo lavoro. Duccio Fanelli, per le conversazioni stimolanti e gli aiuti provvidenziali.

Un ringraziamento speciale va alla mia famiglia, mamma, babbo e Cosimo, per avermi supportato e sopportato, sì lo so, sono stato un tantino piaga, ma si sa, sono tanto sensibbbile. Inoltre vorrei ringraziare le nonne, Vera e Rosi, che mi hanno fornito le due metà del mio patrimonio genetico (poi c'è la terza nonna, Angela, I love you). La zia Anna, con cui prima o poi riuscirò a prendere il volo. Il Bera, io “manine tese”, lui “padrone di Jim”. Gemma e Ale, padre e figlia favolosamente cignalissimi. I miei amici da una vita, Laura, Elena, Alessandra e Filippo, Cutigliano ci aspetta da vecchi quando riusciremo a ricongiungerci, finalmente rimbambiti del tutto. La Cate, la Silvia e la Marzia, zie per scelta, che mi hanno insegnato a godermela (sì sì, ho scritto proprio “godermela” in una tesi di dottorato). La Sofia e la Taddei, non riusciamo a farne uno neanche in una fototessera. La Susie, amica di palcoscenici e di turbe da camerino. La Camilla, amica e dottoressa a distanza. Silvano, amico “particolare” con cui condivido la provenienza da Pentolina. Matteo e Filippo, che riescono ad alleviare i miei momenti da “nervosetto”. La gang del CSDC che ha contribuito ad alleggerire questi tre anni,



facendomi sentire leggiadro come un... volatile.

Infine un enorme riconoscimento va alle altre due figure paterne che hanno fatto parte della mia vita e che più di tutti mi hanno stimolato nello studio, nonno Franco e zio Sergio.

# Bibliography

- [1] Lucia Pettinato, Elisa Calistri, Francesca Di Patti, Roberto Livi, and Stefano Luccioli. Genome-wide analysis of promoters: Clustering by alignment and analysis of regular patterns. *PLOS ONE*, 9(1):1–14, 01 2014.
- [2] Leonardo Lenzini, Francesca Di Patti, Roberto Livi, Marco Fondi, Renato Fani, and Alessio Mengoni. A method for the structure-based, genome-wide analysis of bacterial intergenic sequences identifies shared compositional and functional features. *Genes*, 10(10), 2019.
- [3] Leonardo Lenzini, Francesca Di Patti, Stefano Lepri, Roberto Livi, and Stefano Luccioli. Thermodynamics of dna denaturation in a model of bacterial intergenic sequences. *Chaos, Solitons & Fractals*, 130:109446, 2020.
- [4] Rosalind E. Franklin and R. G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741, 1953.
- [5] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.
- [6] M Nirenberg, P Leder, M Bernfield, R Brimacombe, J Trupin, F Rottman, and C O’Neal. Rna codewords and protein synthesis, vii. on the general nature of the rna code. *Proceedings of the National Academy of Sciences*, 53(5):1161–1168, 1965.
- [7] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.
- [8] MC King and AC Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, 1975.
- [9] Sean B Carroll. Endless forms: The evolution of gene regulation and morphological diversity. *Cell*, 101(6):577 – 580, 2000.

- [10] Kevin Chen and Nikolaus Rajewsky. The evolution of gene regulation by transcription factors and micornas. *Nature Reviews Genetics*, 8(2):93–103, 2007.
- [11] Sean B Carroll. Evolution at two levels: On genes and form. *PLOS Biology*, 3(7), 07 2005.
- [12] David J. Lipman, David L. Wheeler, Dennis A. Benson, Ilene Karsch-Mizrachi, and James Ostell. GenBank. *Nucleic Acids Research*, 33(suppl\_1):D34–D38, 01 2005.
- [13] David J. Lipman, Dennis A. Benson, Ilene Karsch-Mizrachi, James Ostell, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 39(suppl\_1):D32–D37, 11 2010.
- [14] Chuan Zhou, Fenglou Mao, Hanyuan Zhang, Jincai Yang, Qin Ma, Wei Lai, Xin Chen, Xizeng Mao, and Ying Xu. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Research*, 42(D1):D654–D659, 11 2013.
- [15] Fenglou Mao, Jacky Chou, Phuongan Dam, Victor Olman, and Ying Xu. DOOR: a database for prokaryotic operons. *Nucleic Acids Research*, 37(suppl\_1):D459–D463, 11 2008.
- [16] Kyle Harris, Phuongan Dam, Victor Olman, Ying Xu, and Zhengchang Su. Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Research*, 35(1):288–298, 12 2006.
- [17] Damilola Omotajo, Travis Tate, Hyuk Cho, and Madhusudan Choudhary. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics*, 16(1):604, Aug 2015.
- [18] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.
- [19] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
- [20] P. Rice, I. Longden, and A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16:276–277, 2000.
- [21] U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput*, 17:395–416, 2007.

- [22] G. Stewart and J. Sun. *Matrix perturbation theory*. Academic Press, 1990.
- [23] S.B. Carroll. Evo-devo and the expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134:25–36, 2008.
- [24] Yoichiro Shibata, Nathan C. Sheffield, Olivier Fedrigo, Courtney C. Babbitt, Matthew Wortham, Alok K. Tewari, Darin London, Lingyun Song, Bum-Kyu Lee, Vishwanath R. Iyer, Stephen C. J. Parker, Elliott H. Margulies, Gregory A. Wray, Terrence S. Furey, and Gregory E. Crawford. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet*, 8:e1002789, 2012.
- [25] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutuyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337:1190–1195, 2012.
- [26] Ryan J. Taft, Michael Pheasant, and John S. Mattick. The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays*, 29:288–299, 2007.
- [27] Chris P. Ponting. The functional repertoires of metazoan genomes. *Nat Rev Genet*, 9:689–698, 2008.
- [28] Samuel Levy, Sridhar Hannenhalli, and Christopher Workman. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics*, 17:871–877, 2001.
- [29] Elliott H. Margulies, Mathieu Blanchette, NISC Comparative Sequencing Program, David Haussler, and Eric D. Green. Identification and characterization of multi-species conserved sequences. *Genome Research*, 13(12):2507–2518, 2003.
- [30] Emmanouil T. Dermitzakis, Alexandre Reymond, Nathalie Scamuffa, Catherine Ucla, Ewen Kirkness, Colette Rossier, and Stylianos E. Antonarakis. Evolution-

- ary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, 302(5647):1033–1035, 2003.
- [31] Gill Bejerano, David Haussler, and Mathieu Blanchette. Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics*, 20:i40–i48, 2004.
- [32] Leila Taher, David M. McGaughey, Samantha Maragh, Ivy Aneas, Seneca L. Bessling, Webb Miller, Marcelo A. Nobrega, Andrew S. McCallion, and Ivan Ovcharenko. Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Research*, 21(7):1139–1149, 2011.
- [33] Brian J. Parker, Ida Moltke, Adam Roth, Stefan Washietl, Jiayu Wen, Manolis Kellis, Ronald Breaker, and Jakob Skou Pedersen. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Research*, 21(11):1929–1943, 2011.
- [34] Masatoshi Matsunami, Kenta Sumiyama, and Naruya Saitou. Evolution of conserved non-coding sequences within the vertebrate Hox clusters through the two-round whole genome duplications revealed by phylogenetic footprinting analysis. *Journal of Molecular Evolution*, 71(5-6):427–436, 2010.
- [35] Elisa Calistri. *Variability and constraints in promoter evolution*. PhD thesis, Non-linear dynamics and complex systems, 2008.
- [36] Anirudh Natarajan, Galip Grkan Yardmc, Nathan C. Sheffield, Gregory E. Crawford, and Uwe Ohler. Predicting cell-typespecific gene expression from regions of open chromatin. *Genome Research*, 22(9):1711–1722, 2012.
- [37] Shane Neph, Jeff Vierstra, Andrew B. Stergachis, Alex P. Reynolds, Eric Haugen, Benjamin Vernot, Robert E. Thurman, Sam John, Richard Sandstrom, Audra K. Johnson, Matthew T. Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shiny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R. Scott Hansen, Tanya Kutyaavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang, Gayathri Balasundaram, Rachel Byron, Michael J. MacCoss, , Joshua M. Akey, M. A. Bender, Mark Groudine, Rajinder Kaul, and John A. Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489:83–90, 2012.

- [38] Jane M. Landolin, David S. Johnson, Nathan D. Trinklein, Shelly F. Aldred, Catherine Medina, Hennady Shulha, Zhiping Weng, and Richard M. Myers. Sequence features that drive human promoter function and tissue specificity. *Genome Research*, 20(7):890–898, 2010.
- [39] Martin Hemberg, Jesse M. Gray, Nicole Cloonan, Scott Kuersten, Sean Grimmond, Michael E. Greenberg, and Gabriel Kreiman. Integrated genome analysis suggests that most conserved non-coding sequences are regulatory factor binding sites. *Nucleic Acids Research*, 40(16):7858–7869, 2012.
- [40] Zeba Wunderlich and Leonid A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10):434 – 440, 2009.
- [41] Peggy J. Farnham. Insights from genomic profiling of transcription factors. *Nat Rev Genet*, 10:605–616, 2009.
- [42] Bart Deplancke, Daniel Alpern, and Vincent Gardeux. The genetics of transcription factor dna binding variation. *Cell*, 166(3):538 – 554, 2016.
- [43] Dominic Schmidt, Michael D. Wilson, Benoit Ballester, Petra C. Schwalie, Gordon D. Brown, Aileen Marshall, Claudia Kutter, Stephen Watt, Celia P. Martinez-Jimenez, Sarah Mackay, Iannis Talianidis, Paul Flicek, and Duncan T. Odom. Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040, 2010.
- [44] Matthew T. Weirauch and Timothy R. Hughes. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends in Genetics*, 26(2):66 – 74, 2010.
- [45] Guillaume Bourque, Bernard Leong, Vinsensius B. Vega, Xi Chen, Yen Ling Lee, Kandhadayar G. Srinivasan, Joon-Lin Chew, Yijun Ruan, Chia-Lin Wei, Huck Hui Ng, and Edison T. Liu. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research*, 18:1752–1762, 2008.
- [46] Sandeep Venkataram and Justin C. Fay. Is transcription factor binding site turnover a sufficient explanation for cis-regulatory sequence divergence? *Genome Biology and Evolution*, 2:851–858, 2010.

- [47] E. Blanco, R. Guigo, and X. Messeguer. Multiple non-collinear TF-map alignments of promoter regions. *BMC Bioinformatic*, 8:138, 2007.
- [48] A.S. Bais, S. Grossmann, and M. Vingron. Incorporating evolution of transcription factor binding sites into annotated alignments. *J. Biosci.*, 32:841–850, 2007.
- [49] Outi Hallikas, Kimmo Palin, Natalia Sinjushina, Reetta Rautiainen, Juha Partanen, Esko Ukkonen, and Jussi Taipale. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47 – 59, 2006.
- [50] Stephen C. J. Parker, Loren Hansen, Hatice Ozel Abaan, Thomas D. Tullius, and Elliott H. Margulies. Local DNA topography correlates with functional noncoding regions of the human genome. *Science*, 324(5925):389–392, 2009.
- [51] Thomas Abeel, Yvan Saeys, Eric Bonnet, Pierre Rouz, and Yves Van de Peer. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Research*, 18(2):310–323, 2008.
- [52] Martin G Reese. Application of a time-delay neural network to promoter annotation in the drosophila melanogaster genome. *Computers & Chemistry*, 26(1):51 – 56, 2001.
- [53] B Demeler and GW Zhou. Neural network optimization for E. coli promoter prediction. *Nucleic Acids Res*, 19:1593–1599, 1991.
- [54] Hilal Tayara, Muhammad Tahir, and Kil To Chong. Identification of prokaryotic promoters and their strength by integrating heterogeneous features. *Genomics*, 2019.
- [55] Scheila de Avila e Silva, Sergio Echeverrigaray, and Gnther J.L. Gerhardt. Bacpp: Bacterial promoter predictiona tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of Theoretical Biology*, 287:92 – 99, 2011.
- [56] Elisa Calistri, Roberto Livi, and Marcello Buiatti. Evolutionary trends of GC/AT distribution patterns in promoters. *Molecular Phylogenetics and Evolution*, 60(2):228 – 235, 2011.
- [57] Albin Sandelin, Piero Carninci, Boris Lenhard, Jasmina Ponjavic, Yoshihide Hayashizaki, and David A. Hume. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet*, 8:424–436, 2007.

- [58] Boris Lenhard, Albin Sandelin, and Piero Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*, 13:233–245, 2012.
- [59] Sacha A. F. T. van Hijum, Marnix H. Medema, and Oscar P. Kuipers. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiology and Molecular Biology Reviews*, 73(3):481–509, 2009.
- [60] Michael Lässig. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics*, 8(6):S7, Sep 2007.
- [61] Marco Galardini, Matteo Brilli, Giulia Spini, Matteo Rossi, Bianca Roncaglia, Alessia Bani, Manuela Chianciani, Marco Moretto, Kristof Engelen, Giovanni Bacci, Francesco Pini, Emanuele G. Biondi, Marco Bazzicalupo, and Alessio Mengoni. Evolution of intra-specific regulatory networks in a multipartite bacterial genome. *PLOS Computational Biology*, 11(9):1–24, 09 2015.
- [62] Nicolas E. Buchler, Ulrich Gerland, and Terence Hwa. On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences*, 100(9):5136–5141, 2003.
- [63] Mayetri Gupta and Jun S. Liu. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proceedings of the National Academy of Sciences*, 102(20):7079–7084, 2005.
- [64] J. Shine and L. Dalgarno. Determinant of cistron specificity in bacterial ribosomes. *Nature*, 254(5495):34–38, 1975.
- [65] Itamar Sela and David B. Lukatsky. DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophysical Journal*, 101(1):160 – 166, 2011.
- [66] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2015.



- [67] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.
- [68] Roman L Tatusov, Eugene V Koonin, and David J Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.
- [69] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.
- [70] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.
- [71] Douglas Poland and Harold A. Scheraga. Phase transitions in one dimension and the helixcoil transition in polyamino acids. *The Journal of Chemical Physics*, 45(5):1456–1463, 1966.
- [72] Yariv Kafri, David Mukamel, and Luca Peliti. Why is the dna denaturation transition first order? *Phys. Rev. Lett.*, 85:4988–4991, Dec 2000.
- [73] G. Giacomin and F. L. Toninelli. Smoothing of depinning transitions for directed polymers with quenched disorder. *Phys. Rev. Lett.*, 96:070602, Feb 2006.
- [74] H. Kunz and R. Livi. DNA denaturation and wetting in the presence of disorder. *EPL (Europhysics Letters)*, 99(3):30001, aug 2012.
- [75] Maria Serena Causo, Barbara Coluzzi, and Peter Grassberger. Simple model for the dna denaturation transition. *Phys. Rev. E*, 62:3958–3973, Sep 2000.
- [76] Enrico Carlon, Enzo Orlandini, and Attilio L. Stella. Roles of stiffness and excluded volume in dna denaturation. *Phys. Rev. Lett.*, 88:198101, Apr 2002.
- [77] Marco Baiesi, Enrico Carlon, and Attilio L. Stella. Scaling in dna unzipping models: Denaturated loops and end segments as branches of a block copolymer network. *Phys. Rev. E*, 66:021804, Aug 2002.
- [78] M. Baiesi, E. Carlon, Y. Kafri, D. Mukamel, E. Orlandini, and A. L. Stella. Interstrand distance distribution of dna near melting. *Phys. Rev. E*, 67:021911, Feb 2003.

- [79] Maria Barbi, Simona Cocco, Michel Peyrard, and Stefano Ruffo. A twist opening model for dna. *Journal of Biological Physics*, 24(2):97–114, Jun 1999.
- [80] Simona Cocco and Rémi Monasson. Statistical mechanics of torque induced denaturation of dna. *Phys. Rev. Lett.*, 83:5178–5181, Dec 1999.
- [81] Maria Barbi, Stefano Lepri, Michel Peyrard, and Nikos Theodorakopoulos. Thermal denaturation of a helicoidal dna model. *Phys. Rev. E*, 68:061909, Dec 2003.
- [82] Ralf Blossey and Enrico Carlon. Reparametrizing the loop entropy weights: Effect on dna melting curves. *Phys. Rev. E*, 68:061911, Dec 2003.
- [83] Roger M. Wartell and Albert S. Benight. Thermal denaturation of dna molecules: A comparison of theory with experiment. *Physics Reports*, 126(2):67 – 107, 1985.
- [84] R. D. Blake and Scott G. Delcourt. Thermal stability of DNA. *Nucleic Acids Research*, 26(14):3323–3332, 07 1998.
- [85] G. Zocchi, A. Omerzu, T. Kuriabova, J. Rudnick, and G. Grner. Duplex-single strand denaturing transition in dna oligomers, 2003.
- [86] Hervé Kunz, Roberto Livi, and András Sto. The structure factor and dynamics of the helix–coil transition. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06004–P06004, jun 2007.
- [87] M. Peyrard and A. R. Bishop. Statistical mechanics of a nonlinear model for dna denaturation. *Phys. Rev. Lett.*, 62:2755–2758, Jun 1989.
- [88] Marc Joyeux and Sahin Buyukdagli. Dynamical model based on finite stacking enthalpies for homogeneous and inhomogeneous dna thermal denaturation. *Phys. Rev. E*, 72:051902, Nov 2005.
- [89] Thierry Dauxois, Michel Peyrard, and A. R. Bishop. Entropy-driven dna denaturation. *Phys. Rev. E*, 47:R44–R47, Jan 1993.
- [90] Alessandro Campa and Andrea Giansanti. Experimental tests of the peyrard-bishop model applied to the melting of very short dna chains. *Phys. Rev. E*, 58:3585–3588, Sep 1998.

- [91] Thierry Dauxois, Nikos Theodorakopoulos, and Michel Peyrard. Thermodynamic instabilities in one dimension: Correlations, scaling and solitons. *Journal of Statistical Physics*, 107(3):869–891, May 2002.
- [92] Michel Peyrard. Nonlinear dynamics and statistical physics of DNA. *Nonlinearity*, 17(2):R1–R40, jan 2004.
- [93] Manoel Manghi and Nicolas Destainville. Physics of base-pairing dynamics in dna. *Physics Reports*, 631:141, May 2016.
- [94] R I McLachlan and P Atela. The accuracy of symplectic integrators. *Nonlinearity*, 5(2):541–562, mar 1992.
- [95] Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.
- [96] Donna N. Lee and Robert Landick. Structure of rna and dna chains in paused transcription complexes containing escherichia coli rna polymerase. *Journal of Molecular Biology*, 228(3):759 – 777, 1992.
- [97] Consortium IHGS. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.
- [98] Ren Dreos, Giovanna Ambrosini, Romain Groux, Rouada CavinPrier, and Philipp Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Research*, 45(D1):D51–D55, 11 2016.
- [99] Ren Dreos, Giovanna Ambrosini, Rouayda Cavin Prier, and Philipp Bucher. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Research*, 43(D1):D92–D96, 11 2014.
- [100] Joaquin Snchez and Imelda Lpez-Villaseor. A simple model to explain three-base periodicity in coding dna. *FEBS Letters*, 580(27):6413 – 6422, 2006.
- [101] E N Trifonov and J L Sussman. The pitch of chromatin dna is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences*, 77(7):3816–3820, 1980.
- [102] B.D. Silverman and R. Linsker. A measure of dna periodicity. *Journal of Theoretical Biology*, 118(3):295 – 300, 1986.

- [103] Anastasios A. Tsonis, James B. Elsner, and Panagiotis A. Tsonis. Periodicity in dna coding sequences: Implications in gene evolution. *Journal of Theoretical Biology*, 151(3):323 – 331, 1991.