UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA
IN FISICA E ASTRONOMIA

Ciclo XXVIII

Coordinatore Prof. Roberto Livi

Methods for a Genome-wide Analysis of DNA Promoters

Settore Scientifico disciplinare: FIS/02

Dottoranda: Lucia Pettinato

Tutore: Roberto Livi

Coordinatore: Roberto Livi

Anni 2012/2013 - 2013/2014 - 2014/2015

Nel farsi di ogni avvenimento
che poi grandemente si configura
c'è un concorso di minuti
avvenimenti, tanto minuti da
essere a volte impercettibili, che
in un moto di attrazione e di
aggregazione corrono verso un
centro oscuro, verso un vuoto
campo magnetico in cui
prendono forma: e sono,
insieme, il grande avvenimento
appunto. In questa forma, nella
forma che insieme assumono,
nessun minuto avvenimento è
accidentale, incidentale, fortuito:
le parti, sia pure molecolari,
trovano necessità - e quindi
spiegazione - nel tutto; e il tutto
nelle parti.

L. Sciascia, *L'affaire Moro*

Contents

Chapter 1

Introduction

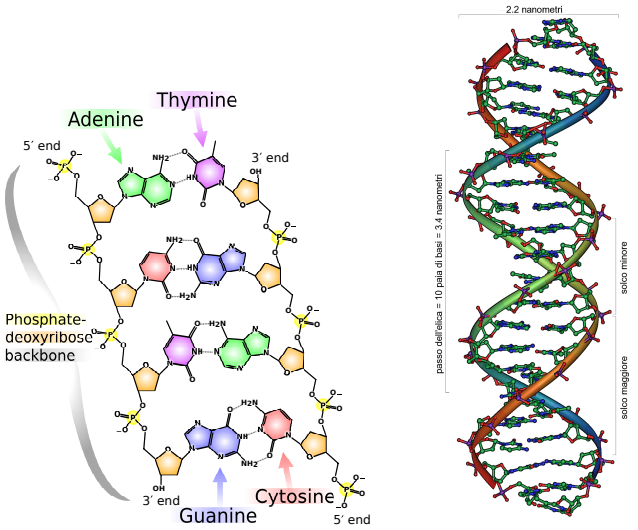
The work described in this Ph. D. Thesis consists in developing algorithms and new methods of analysis to study specific DNA sequences, the promoters. The key challenge of this analysis is to extract information from a great deal of data (the promoter sequences) to identify the relevant information from a biological point of view.

I will give here some essential biological background of my work; then I will give a general overview of the analyses performed, with a stress on the reasons why they have been undertaken.

1.1 DNA and promoters

DNA is a molecule that encodes the information used in the development and functioning of living organisms. It stores the information assembled during million of years of evolution, allowing this information to be inherited from one generation to the next. A DNA molecule consists of two complementary strands forming a double helix. Each strand is a polymer made of four different kinds of nucleotides, joined to one another by a phosphate-deoxiribose backbone (Figure 1.1(a)). The four different kinds of nucleotide differ in the nitrogen nucleobase, namely guanine (G), adenine (A), thymine (T), or cytosine (C). Between the two strands of DNA there are hydrogen bonds formed by the nucleobases, with a specific pairing rule: A pairs with T forming two hydrogen bonds, and C pairs with G forming three bonds. The two DNA strands coil around each other forming the DNA double helix [1,2] (Figure 1.1(b)).

The information in DNA is stored in the sequence formed by the nucleotides along a strand: such sequence can be thought as a string of symbols in a four letter alphabet. Since the very beginning of DNA studies, the main role of DNA molecule has been considered encoding information about protein structure in DNA sequences called genes. The gene sequence is first read and transcribed by a protein (RNA polymerase) in an RNA strand.



(a) (b)

Figure 1.1 DNA: nucleotides 1.1(a) and double helix 1.1(b)

This strand is then translated in the corresponding protein by ribosomes: they interpret the genetic code of the gene and use it as a template for determining the correct sequence of amino acids in the corresponding protein. The genetic code consists of three-nucleotide 'words' called codons; each codon corresponds to one of 21 possible aminoacids that form the protein polymer. This code is "universal", i.e. it is the same for all species [3].

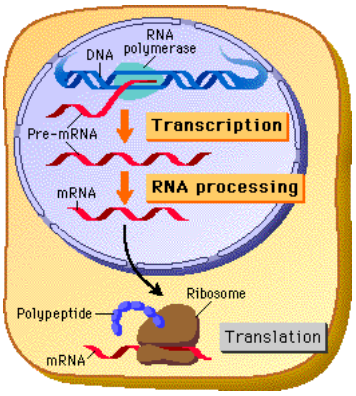


Figure 1.2 Transcription and translation processes.

Despite their capital importance, reducing DNA to a mere sequence of genes would be an oversimplification. 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. *For instance, only about 1.5% of the human genome is made by genes* [4]. The remaining 98.5% is non coding DNA: a

kind of a "dark matter" whose role is still on debate [5]. Anyway, some of these non-coding regions are known to carry important information as well: *they encode the instructions for a correct use of the information stored in genes*. Note that each cell of our organism possesses the complete genome (i.e. the complete set of genes); nevertheless, tissue specialization implies that only a part of those genes are used in a given cell, i.e. the genes necessary to the function of that tissue. Regulation mechanisms suitably inhibit or enhance the expression of each gene, directing when and in which tissue each gene must (or must not) be used. This means that they are responsible of the correct functioning and developing of an organism. Moreover, they also allow an adequate response to environmental stress activating specific genes when needed.

Such regulation mechanisms can act at various stages of protein synthesis, controlling transcription, translation or even the proteins resulting from translation. We will focus here on transcriptional regulation, i.e. on the mechanisms that start and control the frequency at which a protein coding gene is transcribed into an RNA strand. The aim of this work is to investigate DNA elements that have a key role in transcriptional regulation mechanisms: promoters.

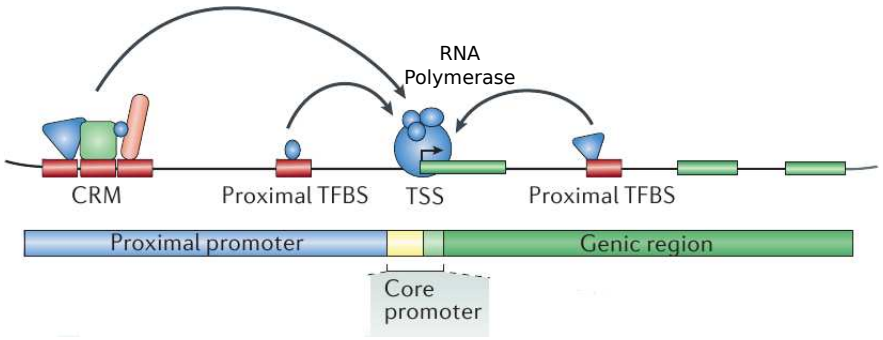


Figure 1.3 Scheme of promoter with Transcription Factors and Transcription Factors Binding Sites.

Promoters are non coding DNA sequences located immediately upstream of the Transcription Start Site (TSS)[1] of each gene. Their length is about 1000 nucleotides. Inside promoters we find short sequences, called binding sites, whose task is binding specific proteins called Transcription Factors (TF). Through complex biochemical mechanisms, transcription factors control the activity of an enzyme, the RNA polymerase, that transcribes the gene (Figure 1.3). Thus, promoters, by means of the interaction with Transcription Factors, are able to inhibit or enhance the transcription and the expression of the corresponding gene.

[1]TSS is the nucleotide at which the enzyme RNA polymerase starts to transcribe the gene sequence into an RNA strand. It is, in fact, the beginning of the gene sequence.

Note that gene regulation mechanisms may also have a key role in evolution (see also sec. 1.2.3). One would think that the responsible of the 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, [6], 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 [7]. There is evidence that morphological changes in animals have been shaped by *evolutionary changes in developmental gene regulation*, and not in genes themselves [7–9]. 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.

1.2 Methodological approach and scope of this work

The usual way to characterize a promoter is searching for the binding sites in its sequence, in order to identify the transcription factors that regulate the expression of the corresponding gene. Moreover, in literature, studies often focus on a specific promoter of a given gene. Although this method is certainly useful, the main approach of our work was completely different. We aim to gain information about *general* properties of promoters, i.e. properties that 1) characterize many promoters and/or 2) the entire promoter sequence instead of the few bases that constitute the binding sites. These are the main tracts that differentiates the study presented here from the mainstream approach applied to the study of promoters. We believe that this "blind" approach could be useful since it includes the whole promoter sequence "as it is" that, apart from binding sites, is quite unexplored, even if several hypotheses have been formulated about its role in promoter function. In fact, it has been observed that promoter sequences have a peculiar structure, often showing periodicities or other kind of regularities in nucleotide distribution.

We developed our approach in order to investigate especially the following topics:

1. Finding common features between promoters in order to search for a relation between the structure of their sequence and their function (for instance, we aim to find that promoters with similar features in the structure of their sequence regulate similar kind of genes). The state

of the art and the studies inspiring this point are described in section 1.2.1.

2. Characterizing promoter sequences especially identifying the main features of the regular structures observed in promoters (the reason why regular structures are interesting is reported in par. 1.2.2).

3. Searching for clues of how promoters can have a role in evolution, comparing the results of promoter analysis for different species in order to highlight how evolutionary selection acts on promoter sequences too. Current studies about promoter evolution are reported in par. 1.2.3

In the following we present some works that inspired the previous points on which we focused our investigation, and justify the approach we adopted.

1.2.1 Relation between promoter structure and function

With regard to point 1 we were mainly inspired by a previous works [10], in which evidence was found of a correlation between the compositional properties of groups of promoters and the kind of genes they regulate. These work relied upon the heuristic criterion of subdividing the database into two classes determined by the presence of a specific binding site, the TATA-box [11].

A TATA box is a DNA sequence that indicates to other Transcription Factors the starting point of the transcription of a genetic sequence. It is located near the TSS (about 30 bases upstream). The TATA box is one of the most conserved elements in promoter evolution, having an homologous sequence also in prokaryotes (bacteria). While in ancient species almost all promoters have a TATA box, we observe that during evolution this element has specialized its role, and only about 28% of human promoters have a TATA box.

The promoter classification criterion applied in [10] was inspired by the conjecture that the two classes (TATA and TATA-less) are usually related to different promoter regulatory activity: namely, promoters containing a TATA-box are usually associated with tissue–specific genes, while TATA-less promoters are related to housekeeping genes[2] [12]. The analysis of the average base composition showed that promoters containing the TATA-box (28% of the whole *H. sapiens* database) exhibit quite a different nucleotide composition (AT rich) with respect to the group of TATA–less promoters (CG rich) (see Figure 1.4). This was a very interesting result, if one considers that the difference between these classes of promoters is not limited to the region close to the TSS, but it extends over the entire promoter.

[2]Housekeeping genes are required for the maintenance of basic cellular function, and are expressed in all cells of an organism, while tissue specific genes are expressed only in some tissues or at precise stages of development.

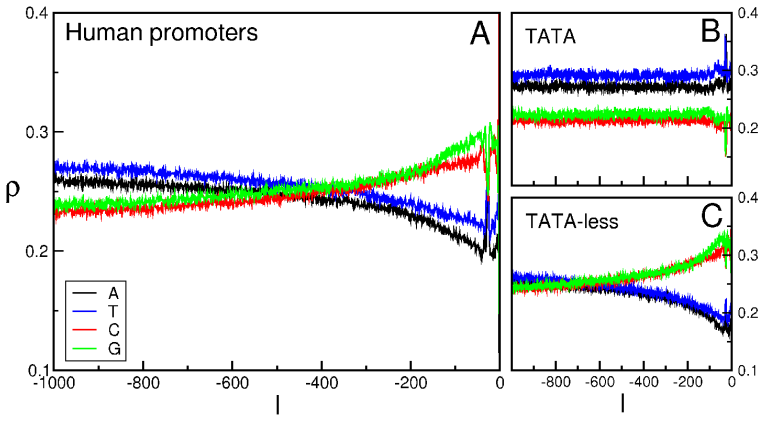


Figure 1.4 BCA of human promoters. BCA of the entire repertoire of human promoters (panel **A**) and of the two sets of TATA and TATA-less promoters (panels **B** and **C**). We report the frequency ρ of each of the four nucleotides A (black), T (blue), C (red) and G (green) as a function of the position l along the promoter (0 corresponds to the TSS). Figure from [10].

Such results suggested that the idea of further investigation of compositional properties of groups of promoters instead of considering one promoter at a time could produce biologically significant results. Moreover, regarding the choice of taking into account the entire promoter sequence and not just the binding sites, these results are encouraging since they highlight significant compositional properties of the whole sequence, and these properties have a biological significance. We will show in this thesis and in the conclusion how the results we obtain confirm our working hypothesis.

1.2.2 Regular sequences in promoters

Nucleotide sequences in promoters are characterized by the alternation of regular and disordered regions of different length. In particular, the regular ones exhibit various structures in nucleotide distribution, ranging from homogeneous to periodic and palindromic [13]. Such regular sequences have been shown to posses peculiar structural properties involved in regulatory functions.

For instance, in the works of Sela and Lukatsky [14–18] it is discussed the role of a specific kind of regular sequences (such as poly(dA:dT) and poly(dC:dG) tracts, i.e. tracts with a homogeneous composition made only of weak or only of strong bases) in TF recruitment. In fact, it is an open problem at the moment the mechanism by which a Transcription Factor can find its corresponding Binding Site in DNA. Sela and Lukatsky hypothesize that poly(dA:dT) and poly(dC:dG) form a non specific binding with the TF. This binding keeps the TF anchored to DNA but, due to its weakness,

it allows the TF to slide along DNA until it finds its corresponding Binding Site. Thus, poly(dA:dT) and poly(dC:dG) convert the 3D diffusion motion of the TF in cell nucleus in 1D diffusion motion along the DNA, considerably speeding up the process.

Other possible roles for regular sequences have been hypothesized: due to the different chemo-physical properties of the weak and strong nucleotides, homogeneous tracts of weak and strong bases are supposed to influence the dynamical properties of the promoter double helix in terms of stiffness, bending, formation of DNA bubbles (necessary for transcription initiation) and coiling [19–23]. Such physical properties of the promoter double helix determine its regulatory function, influencing its interaction with transcription factors and especially with RNA polymerase.

Therefore, due to the evidence that the regular structures are an intrinsic property of promoter regions, and play an important role in its functioning, in this thesis we will also focus on identification and characterization of such sequences, as reported in point 2 on page 5.

1.2.3 Evolutionary role of promoters

Gene mutation is not the only possible evolutionary mechanism: evidences have been found that also promoter sequences undergo evolutionary selection. This kind of evolution does not act on genes and corresponding proteins, but it acts on the gene expression mechanisms [24, 25]. A comparison between promoter structure in many species [26] covering the whole phylogenetic tree highlighted phylogenetic trends throughout evolution of promoter sequence base distributions. Particularly, in all cases either GC-rich or AT-rich monotone gradients in the direction of the TSS were observed: the former being present in eukaryotes, the latter in bacteria along with strand biases. Moreover, within eukaryotes, GC-rich gradients increased in length from unicellular organisms to plants, to vertebrates and, within them, from ancestral to more recent species (Figure 1.5).

Results show a possible correlation between nucleotide distribution patterns, evolution, and the putative existence of differential selection pressures, deriving from structural and/or functional constraints, between and within prokaryotes and eukaryotes.

In order to further investigate this topic, one of the purposes of this thesis work, i.e. point 3 on page 5, was to compare the results of promoter analysis for different species, in order to highlight evolutionary trends.

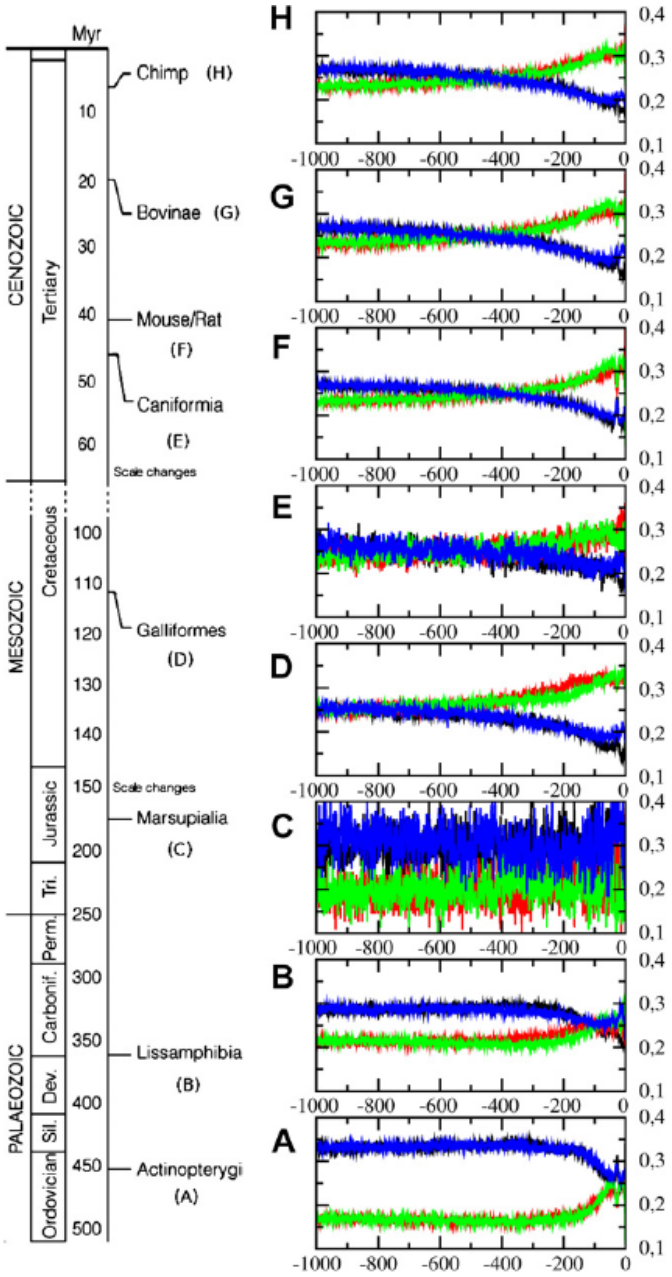


Figure 1.5 BCA in vertebrates. Promoter sequences from -1000 bp to -1 bp relative to the TSS in (A) Danio rerio, (B) Xenopus tropicalis, (C) Monodelphis domestica, (D) Gallus gallus, (E) Canis familiaris, (F) Mus musculus, (G) Bos bovis, (H) Pan troglodytes. Species are arranged according to divergence-time estimates for mammalian orders and major lineages of vertebrates, as given in Kumar and Hedges (1998), based on molecular time estimations. On the left: a molecular timescale for vertebrate evolution, adapted from Kumar and Hedges (1998). All times indicate Mya separating humans and the Order, Family or Genus shown. Letters (A–H) relate every represented species with the phylogenetic group they belong to. Figure from [26].

1.3 Structure of the thesis

In this manuscript we report the analyses performed on the promoters of *H. sapiens* and their results. As already stated, our aim was to investigate the points reported in section 1.2. Such points have been addressed with different works, each reported in a chapter of this thesis. Each of the works reported here addresses some or all of the principal inspiring points reported in the previous section.

In Chapter 2 we describe the procedure we developed to compare promoter sequences of a given species in order to classify them in groups, each group characterized by specific compositional features of the promoters it contains. It is also reported a first characterization of the groups obtained, showing biologically significant results.

Then, in order to deepen the results of this first analysis, we also decided to further characterize these groups, broadening our field of analysis beyond the mere nucleotide sequence of the promoter. In Chapter 3 we treat promoter sequences as strings of text, and we employ a text analysis tool that, through suitable entropic indicators, is able to 1) quantify the variability across different promoter sequences and 2) identify keywords in the text.

In Chapter 4 we explore the relationship between our promoter classification and a gene network. The main idea is to color the nodes of the network on the basis of the group (obtained as described in Chapter 2) the promoter belongs to: then we can study the topology of this node-colored network in order to identify whether nodes of a specific color have specific features, and to study also the topology of single color subnetworks on themselves and in relation to the whole network. This part of my work also involves notions of Random Matrix Theory.

Finally, part of the work of my PhD course has been devoted to simulations of the dynamical properties in nonlinear regime of the DNA chains that constitute promoters. We have tried to highlight how the interplay between inhomogeneities in nucleotide composition and nonlinearity influences the energy localization properties of these chains. We have also tried to assess whether the different compositional properties of these groups is reflected in different dynamical properties. Nevertheless, due to the huge effort that it wold have required, this part of my work has been left behind because it was not compatible, in terms of time, with the other research paths we had undertaken, since it would have required a full-time investigation. The results about this topic remained quite incomplete and will not be shown in this thesis.

Chapter 2

Genome-Wide Analysis of Promoters: Clustering and Analysis of Regular Patterns

2.1 Introduction

In this chapter we present the results we obtained from our genome-wide analysis of promoter sequences that has been published in [27]. In particular the analysis is focused on *H. sapiens* but a comparison with other species is also presented.

In this work we take into account the entire promoter sequence. We develop a method that takes into account all the promoters of a given species and organizes them into groups on the basis of the similarity between their sequences. We also search for a characterization of the groups obtained on the basis of biologically significant features. In particular, we developed and combined two mathematical methods that allow us to

1. Classify promoters into groups characterized by specific global structural features

2. Recover, in full generality, any regular sequence in the different classes of promoters. By regular sequence we mean any sequence with a periodicity or homogeneity in nucleotide distribution. We focus our attention on regular sequences because many of them have been shown to posses peculiar structural properties involved in regulatory functions (see par. 1.2.2) [14–23].

In this work we address all of the points described in par. 1.2 as the scope of this thesis.

As for point 1, the method we developed makes use of a clustering algorithm, that groups promoters via an alignment procedure [28–30] that takes into account the whole sequence. This alignment procedure compares promoters two by two, estimating their similarity, and is the starting point of the procedure that groups promoters in clusters on the basis of their similarities. The second method identifies regular sequences characterizing the different clusters. In this framework, the promoter is modeled as a chain of oscillators according to the Peyrard–Bishop model [31–33]: from the analysis of the vibrational properties of the promoter chain it is possible to identify all the regular sequences. The results and the methods of this work are described in detail in the following sections.

One of the main findings of this analysis is that *H. sapiens* promoters can be classified into four main groups. Two of the groups are distinguished by the prevalence of weak (A,T) or strong (C,G) nucleotides and are characterized by short compositionally biased sequences; instead, the most frequent regular sequences in the other two groups (that are to be considered together as a single group) are strongly correlated with transposons. Taking advantage of the generality of these mathematical procedures, we compared the promoter database of *H. sapiens* with those of other species. We have found that the above mentioned features characterize 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.

In section 2.2 the main results obtained for *H. sapiens* promoters are reported and discussed. A brief overview on the methods developed and an overall comparison with other species is reported. The main findings are summarized in the Conclusion section (par. 2.5). The results obtained for other species are reported in more details in Supplementary Materials (par. 2.3). A detailed outline of the methods developed for this study can be found in Methods (par. 2.4).

2.2 Results and discussion

2.2.1 Clustering of promoters

The database of *H. sapiens* promoters contains 32122 sequences associated to protein–coding genes (see par. 2.4.1 in Methods). Each promoter is represented by 1000 nucleotides upstream of the TSS of all annotated genes. We also analyzed databases of other species: two mammals, namely *P. troglodytes* and *M. musculus* (respectively chimpanzee and mouse), *D. rerio* (a fish) and *A. thaliana* (a plant). Nevertheless, we focus here on the results obtained for *H. sapiens*. More details on the results obtained for other species are reported in Supplementary Materials (sec. 2.3).

A first classification of the promoters of this database was proposed in [10] and is described in par. 1.2.1. In order to obtain a more refined promoter classification, we adopt here a general clustering strategy of *H. sapiens* promoters that takes into account the global properties of the whole promoter sequence instead of specific short regulatory motifs. In other words, we adopt a blind approach and we take into account the whole promoter sequence as it is, neglecting any kind of biological information like Binding Sites. This clustering procedure is described in details in section 2.4.3 of Methods. It is based on a three-step procedure:

1. Evaluate the similarity between promoters. This is done by applying an alignment algorithm [28–30] that compares the promoters two by two and gives a score that estimates their similarity. We obtain the matrix of the scores whose element (i,j) is the score of the alignment of the i-th promoter with the j-th promoter. The robustness of the method has been first verified by comparing two different alignment algorithms, namely Needleman–Wunsch [28] and Waterman-Smith [29]. We have found that, although the entries of the similarity matrix are quite different, both alignment algorithms yield essentially the same cluster organization. Accordingly, we have decided to report here only the result of the Needleman–Wunsch alignment algorithm, whose parameters have been fixed by a suitable optimization procedure (see section 2.4.3 in Methods). The main computational limitations of this clustering procedure stem from the alignment protocol and from the diagonalization of the similarity matrices. Therefore we have been able to consider similarity matrices of rank up to 2880, meaning that each run of the clustering algorithm can be applied to a sample of 2880 promoters.

2. From the matrix of the scores, with simple mathematical steps, we build the Laplacian matrix [34]. From its eigenvalues we get the right number of clusters in which to divide the sample. The eigenvalues of the Laplacian matrix, associated to the similarity matrix, highlight the presence of four clusters for *H. sapiens* (see Figure 2.16 in Methods). With the eigenvectors we build a distribution of points in an abstract space (the clustering space) such that each point represents a promoter. In this representation each point represents a promoter: promoters with a high similarity score correspond to near points (see Figure 2.1).

3. We unambiguously associate each of the promoters to one of the four clusters with the K-means algorithm. The whole procedure is extensively described in section 2.4.3 in Methods.

In panel A of Figure 2.1 we make use of a four–color representation, where each color corresponds to a cluster, while in panel B we show, by a

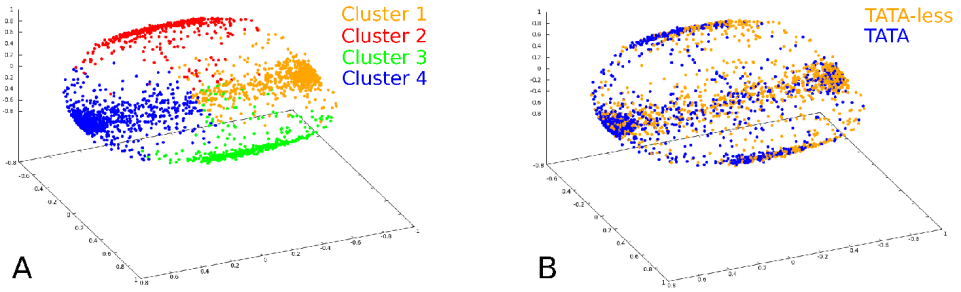


Figure 2.1 Distribution of points in the clustering space (see Methods) relative to the alignment of 2880 human promoters. Each point represents a promoter of the sample. **A**. The color code represents the four clusters. **B**. The color code represents the TATA (blue dots) and TATA–less (orange dots) classification.

two-color representation, the partition into TATA and TATA–less promoters. The former (latter) are preferentially located on the left (right) side. Accordingly, if we compare our results with those reported in [10], we can conclude that our clustering algorithm yields a different and more refined classification of promoter sequences with respect to the mere partition of the sample into TATA and TATA–less promoters. In fact, our clustering method takes into account global properties of promoters, while the one adopted in [10] relies upon a local criterion, i.e. the presence of the TATA–box in a specific promoter region. Such a difference also emerges from the comparison of the BCA for the two families of TATA and TATA-less promoters of the whole database (see Figure 1.4) with the one of promoters in the four clusters (see Figure 2.2). The latter exhibit two clusters dominated by CG and AT nucleotides, denoted as cluster 1 (C1) and cluster 4 (C4), respectively; the other clusters , 2 (C2) and 3 (C3), on the contrary, are characterized by a more uniform distribution of nucleotides. The clusters (that correspond to those of panel A of Fig. 1) contain 934 (C1), 408 (C2), 409 (C3) and 1129 (C4) promoters (see also Fig. 2.3). We observe a different content of TATA promoters in each clusters: in C1 the percentage is about 28%, in C4 is 67% while in C2 and C3 it is 51%. As already mentioned, promoters containing a TATA-box (as most of those in C4) are usually associated with tissue–specific genes, while TATA-less promoters (mostly in C1) are related to housekeeping genes [12]. From the BCA no interesting property of C2 and C3 emerges, but further analyses will reveal quite interesting features of these two clusters.

It is known from the literature that the region around the TSS of animal promoters is typically CG enriched [35, 36]. On the other hand, the result of our clustering procedure indicates that a strong CG bias is present in a specific subset of promoters, i.e. those contained in C1. Although a commonly accepted explanation of the CG enrichment in mammalian promoters

is the presence of the so-called CpG islands, in a previous work [37] it has been shown that all the strong dinucleotide combinations increase with the same rate towards the TSS in mammalian promoters. The same scenario is recovered here for the promoters in C1 (see Fig. 2.14). For more details, see section 2.3.4 in Supplementary Materials.

The same partition into four clusters has been obtained also for *P. troglodytes* and *M. musculus* (see Fig. 2.6 in Supplementary Materials). This suggests that, at least for mammals, there is a general organization of promoters into structurally similar clusters. This clustering method, that takes into account the entire promoter, has been applied also to species different from mammals. For instance, we have studied *D. rerio* and *A. thaliana*, but in this case we do not observe any indication of a clustering. As shown in section 2.3.1 in Supplementary Materials, a clustering for these species can be recovered by limiting the alignment algorithm to a shorter and more specialized region of the promoter, i.e. the first 100 nucleotides upstream the TSS. This seems to suggest that regions much further than 100 nucleotides from the TSS can be considered intergenic regions, that do not correspond to any specific function. This conjecture is also confirmed by other studies of the functional regions of the genome in different species [26, 37, 38]. Altogether, the clustering analysis indicates that promoters in mammals exhibit common features, that depend on global structural properties. Conversely, in other species the clustering strategy is effective only when limited to relatively small regions (typically 100 nucleotides) close to the TSS.

Now, the main question concerns the identification of the structural features characterizing the different clusters.

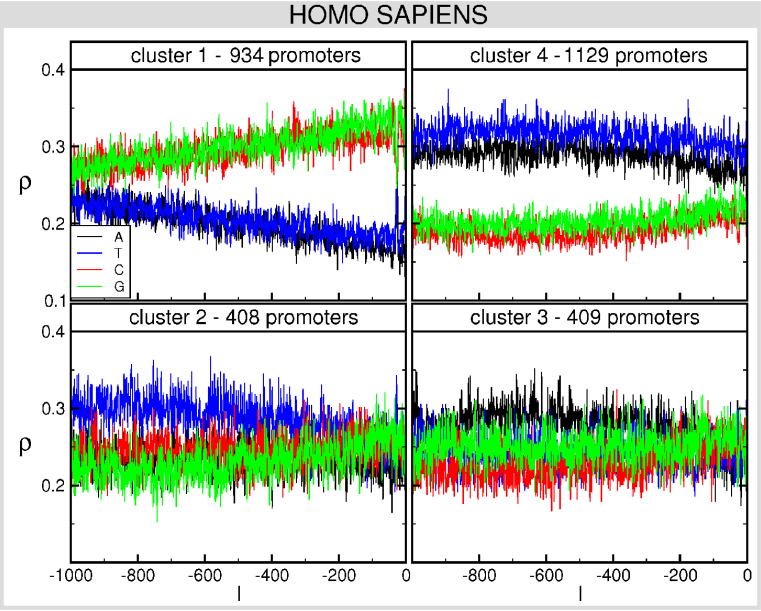


Figure 2.2 BCA of each of the clusters obtained with the clustering algorithm for *H. sapiens*. We report the frequency ρ of each of the four nucleotides A (black), T (blue), C (red) and G (green) as a function of the position l along the promoter (0 corresponds to the TSS).

2.2.2 Regular sequences in promoters

The complex structure of nucleotide sequences in promoters is due to the alternation of regular and disordered regions. These regions can be completely identified by computing the eigenvalues and the eigenvectors of the Hessian Matrix derived from the harmonic approximation of a simple double-strand DNA model, the Peyrard-Bishop model [31–33]. The main idea of this algorithm is the following. Translational invariance assures that, for a perfectly periodic chain, the vibrational modes extend on the whole length of the sequence. On the other hand, disorder causes the localization of the modes, according to the Anderson mechanism [39]. Thus, regular sequences are identified as those on which a "fair" number of delocalized modes overlap. Clearly, the algorithm required to conveniently set several parameters for regular sequence recognition, e.g. the minimum threshold for a mode to be considered delocalized and the minimum number of extended modes on a sequence in order to consider it regular. The choice of these parameters is arbitrary but was guided by common sense and heuristic methods. For instance, the minimum arbitrary threshold chosen for the extent of the modes reflects on the minimum length of regular sequences found, 7 nucleotides. Such threshold was chosen according to the typical length of functional modules in promoters, i.e. 7-8 nucleotides. The details of the algorithm and on the choices of the thresholds are reported in section 2.4.4 of Methods.

One major limitation of this simple model is that it does not distinguish the four nucleotides (A, T, C, G) but only weak (A,T) from strong (C,G) nucleotides[1]. It could be argued that this binary representation introduces a strong bias, because a regular sequence in a weak (W) and strong (S) binary code is not necessarily regular in the natural (A, T, C, G) quaternary code. On purely heuristic grounds, we can say that in most of the promoters many "regular" sequences in the binary code are still "regular" in the quaternary code. Moreover, as testified by the results discussed hereafter, we have checked that a good deal of the regular sequences in the binary code (that may appear less regular in the quaternary code) still play a relevant role in characterizing structural features of the different clusters.

We have found that regular sequences are distributed all along the promoters and cover a relevant portion of them: on average, about 40% of the promoter length in *H. sapiens*, *P. troglodytes* and *M. musculus*, while they reach 50% in *D. rerio* and *A. thaliana* (see Fig. 2.3).

In this section, we focus on the investigation of the properties of the regular sequences in the four clusters of *H. sapiens*. Although they have been identified in the (W,S) binary code, it is worth representing them in the natural quaternary code. Given the huge number of regular sequences in each cluster, we decided to focus our analysis only on the 15 most frequent

[1] C and G form 3 hydrogen bonds and are classified as strong bases, while A and T form only 2 hydrogen bonds and are classified as weak bases

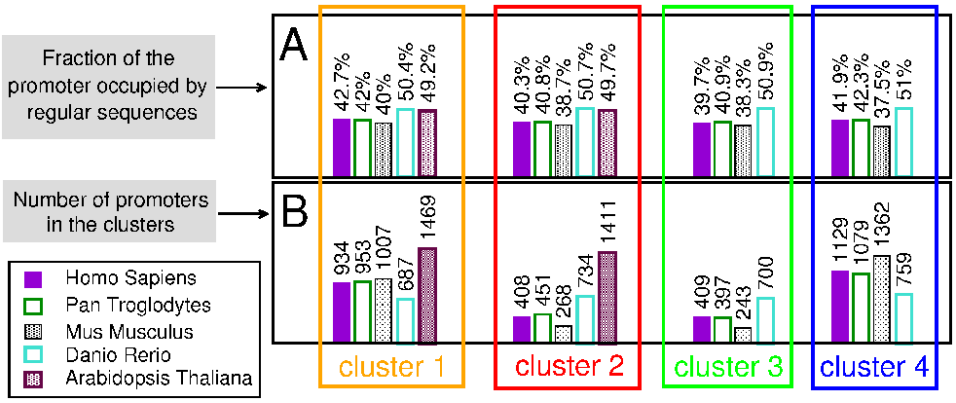


Figure 2.3 Occurrence of regular sequences in the clusters of promoters of different species. **A**. Average fraction of the promoter occupied by regular sequences. **B**. Number of promoters within the clusters.

regular sequences, conjecturing that their over-representation is related to their importance. Anyway, we do not claim that they are the only interesting ones.

The most frequent regular sequences found in C1 and C4 (see Fig. 2.4) extend over 7 nucleotides, i.e. the minimum length of a regular sequence detected by the algorithm (see par. 2.4.4). These short sequences exhibit a prevalence of S-nucleotides in C1 and of W-nucleotides in C4, consistently with the results obtained from the BCA (see Figure 2.2). Their structure as well as their frequency in C1 and C4 are essentially similar. In most cases they are composed of a homogeneous sequence of five nucleotides flanked by two identical nucleotides of different nature in the (W,S) binary code, namely TCCCCCT, ACCCCCA, TGGGGGT, AGGGGGA, CTTTTTC, GAAAAAG, GTTTTTG. A first interesting quantitative feature is that each of these sequences appears in approximately 10% of the promoters of the cluster (see Fig. 2.4). We have also counted how many times each sequence is contained in these host promoters. The large majority contain the regular sequence just once, while only a small fraction of them contains it at most twice. In fact, the average number of each of these regular sequences in host promoters amounts to approximately 1.1: this indicates that each sequence is mostly spread across different promoters.

Also sequence AGGAGGA (as well as its complementary TCCTCCT) appears among the most frequent ones in all clusters. This sequence is fundamental in Prokaryotes, since it corresponds to a consensus sequence for the ribosome-binding site [40]. Its structural properties have been investigated [41, 42] together with its presence in promoters, where it has been found to interact with a stage-specific factor during the late stages of erythropoiesis [43].

One could wonder if over-expressed regular sequences in C1 and C4 are correlated with any biologically relevant function. For instance, taking inspiration from the literature, they could be associated with structural properties of the double helix [14, 23, 44], with the binding of basal transcription factors and RNA polymerase to DNA [13, 45], or to the possibility that homogeneous tracts could play the role of hotspots for mutations [24, 46]. On the same ground, one cannot exclude that they could interact with specific TF [45,47]: we have checked this possibility with various tools and databases (i.e., [48–50]), but we have not found unambiguous outcomes corresponding to these motifs. Anyway, a verification of such conjectures is worthwhile, but goes beyond the aims of this work and will be considered elsewhere. On the other hand, we have selected these sequences on the basis of their regularity and frequency, so that they are not necessarily associated with the specificity of regulatory signal typical of a TF binding site. In their turn, TF binding sites are variously dislocated along the genome (in enhancers, introns, etc.) and they are neither necessarily over-expressed nor regular, as they need a high information content for the specificity of their signal [45, 51].

Anyway, more relevant features differentiate C1 and C4 from C2 and C3, whose regular sequences typically exhibit a different structure. First of all, in C2 and C3 there are long regular sequences, up to 19 nucleotides (i.e. CTAATTTTTGTATTTTTAG and CTAAAAATACAAAAATTAG), among the most frequent ones. Moreover, the most frequent regular sequences appear in about 48% of promoters, at variance with C1 and C4, where they cover at most 14% of the promoters of the cluster. Last but not least, almost all regular sequences found in C2 have a companion sequence in C3 that corresponds to its reverse complement [2]. As we are going to discuss in the following section, this observation indicates a relation of the most frequent regular sequences in C2 and C3 with transposons. This is by far the most interesting and distinctive feature of regular sequences in C2 and C3.

We want to conclude this section by adding two remarks.

Our analysis indicates that the clustering algorithm is able to detect specific similarities among promoters. In C1 and C4 similarities seem to stem just from the prevalence of S or W nucleotides, respectively, while in C2 and C3 they are mostly associated to the presence of specific regular sequences.

With regard to the comparison with other species, we want to point out

[2] Given a DNA sequence, the reverse complement is the sequence we find in the complementary DNA strand. The reverse complement sequence is obtained converting A with T, T with A, C with G and G with C (according to the pairing rule of nucleobases) and then reading the obtained sequence in the opposite sense. The sequence must be reversed because each DNA strand has a reading sense (3' $\rightarrow$ 5') determined by its chemical structure. This sense is opposite for the two strands. For instance, the reverse complement of ATCG is CGAT.

that *P. troglodytes* exhibits the same most frequent regular sequences (including the 19–nucleotide one) found for *H. sapiens*. However, *M. musculus* exhibits rather different features (see Fig. 2.8). In *D. rerio* and *A. thaliana* the search for regular sequences has been performed in all the 1000 nucleotides of each promoter, even if the clusters differentiate only in the 100 nucleotides upstream the TSS. We have found that, at variance with mammals, the most frequent regular sequences are essentially the same in all the clusters (see Fig. 2.9). This is not completely unexpected, because of the low level of differentiation between promoters outside of a small region near the TSS.

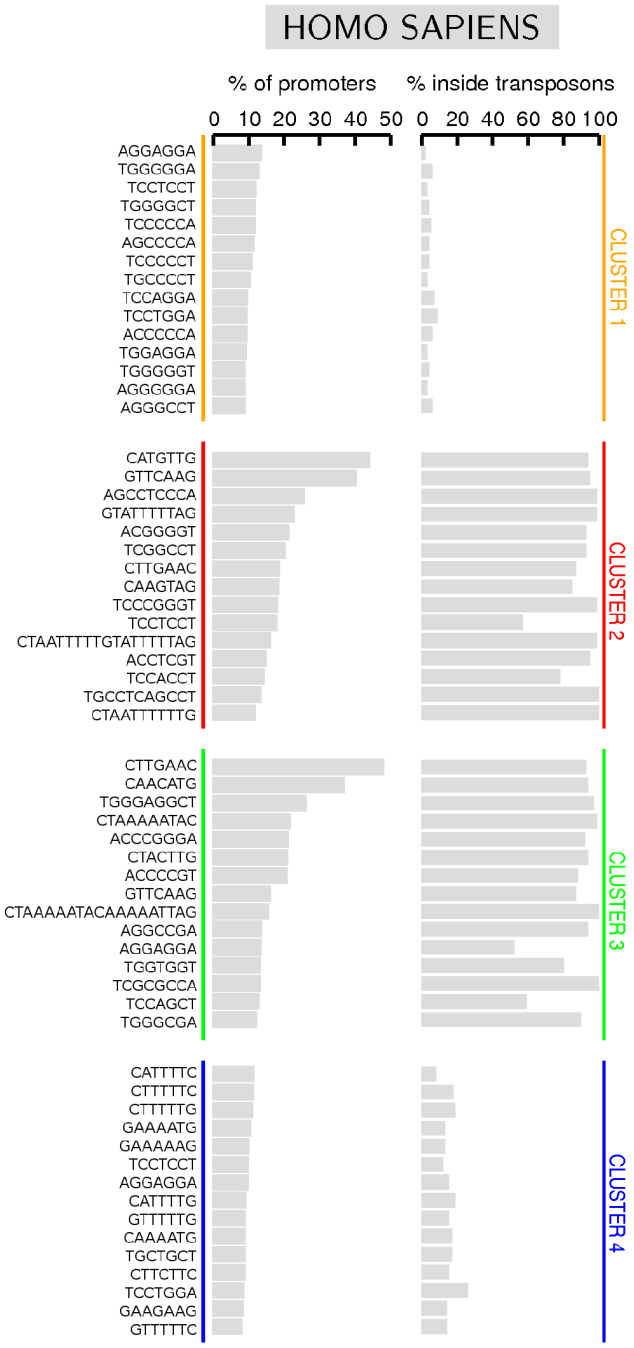


Figure 2.4 The most frequent regular sequences found in the clusters of *H. sapiens*. We report the percentage of promoters of the cluster in which the sequence appears at least once (left column), and the percentage of times the sequence is found inside a transposon (right column): it is calculated dividing the number of times it appears in a transposon by the total number of times it appears in the cluster.

2.2.3 Transposons and regular sequences

In order to identify correlations of regular sequences with specific elements in promoters, we have focused our attention on transposons, that are conjectured to be associated with promoter evolution, while playing a role in gene regulation and expression [25, 52, 53].

Transposons are DNA elements capable to move from one position in the genome to another, creating copies of themselves that insert in other positions (retrotransposons) or deleting themselves from one position to "jump" somewhere else in the genome (DNA transposons). While genes are only 1.5%, transposons account for about 45% of the human genome [4]. Despite the danger they represent (they can cause diseases if they insert inside genes), it is supposed that during evolution they have been domesticated by the host genome and they have been rendered relatively harmless. Actually, it is conjectured that they had an important role in promoter evolution: transposons are known to carry inside their sequences several Binding Sites, and if they insert inside a promoter they can modify its functioning [25, 52].

Transposons can be classified in families, on the basis of their structure and evolutionary origin. The main families in mammals are:

- LINEs (Long Interspersed Nuclear Repeats): The transposons of this family encode a Reverse Transcriptase allowing them to copy themselves to transpose elsewhere. In mammals two kinds of LINEs have been identified: LINE2, which stopped transposing before the mammalian radiation, and LINE1, many of which were inserted after mammalian radiation (and are still active). The typical length of LINE elements is of the order of few kilo-bases.

- SINEs (Short Interspersed Nuclear Repeats): These elements (typical length < 500 bases) are not capable to transpose themselves, but rely on LINEs for transposition. In fact, LINEs and SINEs have evolved together after the mammalian radiation. Among SINEs we find Alu elements (in primates) and B1, B2 in mice. Alus and B1, B2 have evolved independently from a common ancestor, originated by 7SL RNA, a component of the cell signal recognition particle.

- LTR (Long Terminal Repeats) retrotransposons: These elements have a structure similar to the RNA of some retroviruses, and are labeled as Endogenous Retroviruses (ERV). Such elements may have originated as an insertion by a retrovirus or, on the contrary, they may have been the source for the retroviruses they resemble [54].

- DNA transposons: They encode a protein (transposase) that allow them to be removed from one position and inserted in another.

The observation of the reverse complementarity of regular sequences in C2 and C3 suggested a correlation with transposons, since their typical

feature is that they can indifferently intrude on both of the DNA strands. It is worth to recall here that the promoters in the database of *H. sapiens* belong to a specific strand, the one that, in the gene, is transcribed by RNA polymerase. This means that in the sequences of our database we can find one strand or another of the transposon, depending on how it inserted in that promoter.

First of all we have identified (via the RepeatMasker software [55]) all of the transposons present in the promoters of the four clusters of *H. sapiens*. We have found that C2 and C3 contain a large number of transposons, with a majority of Alu ones. On the contrary, C1 and C4 contain a smaller number of transposons, where Alus are quite rare (see Fig. 2.5).

The overabundance of Alu elements found in C2 and C3 could be read as a straightforward consequence of the fact that the Alu family is the most frequent dispersed repeat of the human genome: over one million copies of repeat elements, with a non–uniform distribution [52]. Nevertheless, our results have the merit of identifying the biases in their distributions among the different clusters of promoters.

A similar scenario is observed also for *P. troglodytes* and *M. musculus*, while in *D. rerio* and *A. thaliana* the transposon content is approximately the same in all clusters (see Fig.s 2.10-2.13 in Supplementary Materials).

In order to disclose the conjectured relation between the most expressed regular sequences in C2 and C3 and transposons, we performed the following analysis. First, we computed the percentage of times each sequence belongs to a transposon (reported in the right column of Figure 2.4). Then, we compared this result with the percentage of the cluster covered by transposons, which represents an estimation of the percentage we would expect if the sequence were equally distributed inside and outside the transposons. We have found that in C2 and C3 all the most expressed regular sequences appear in transposons with frequency much higher than the fraction of the cluster covered by transposons (that amounts to $\simeq$ 44-45%). Therefore, such sequences are much more likely to be located inside than outside a transposon: in some cases the probability is actually close to 1. In particular, the sequences with the highest probabilities (e.g. CTAATTTTTGTATTTTTAG) belong to the aforementioned Alu family. This is a strong indication that Alus are responsible of the enrichment of C2 and C3 with these specific sequences. On the other hand, the same analysis performed on clusters C1 and C4 shows that the most frequent regular sequences appear essentially equally distributed inside or outside the transposons. Altogether, we have obtained evidence that such distinctive features are strongly related to the discrimination of the different clusters in *H. sapiens*. Moreover, according to this observation, C2 and C3 should be considered as a unique cluster: as already mentioned, their apparently different features are the mere consequence of the insertion of transposons in different promoter strands, that yields the reverse complementarity characterizing regular sequences in these

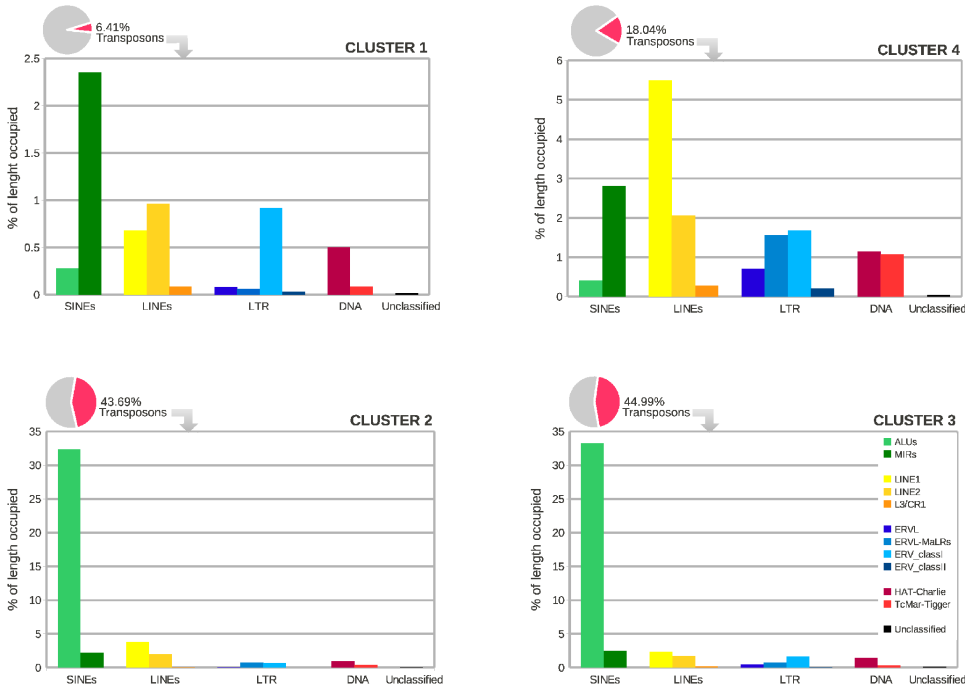


Figure 2.5 Distribution of the different families of transposons in the four clusters of *H. sapiens*. We report the total percentage of nucleotides in the cluster covered by transposons (pie chart) and the percentage of nucleotides covered by each family of transposons (histogram). Note the different scales in the histograms.

clusters.

In section 2.3.3 of Supplementary Materials we have reported also the results obtained via the RepeatMasker software [55] for the other species considered in this paper, namely *P. troglodytes*, *M. musculus*, *D. rerio* and *A. thaliana*. For the first two species we observe very similar features with respect to *H. sapiens*: in Fig. 2.8 we show that the correlation between the most expressed regular sequences in C2 and C3 and transposons is preserved. In accordance with the known divergence of transposable elements between primates and mice [56, 57], the regular sequences in C2 and C3 of *M. musculus* are in most cases different from those of *H. sapiens* and *P. troglodytes*. In the two other species transposons are equally distributed in all clusters. There is still a correlation between some regular sequences and transposons in *D. rerio*, while such a correlation is absent in *A. thaliana*.

2.3 Supplementary Materials: other species

2.3.1 Clustering and BCA of other species

We report here the results obtained for *P. troglodytes*, *M. musculus*, *D. rerio* and *A. thaliana*, making use of the same clustering procedure adopted for *H. sapiens*. The first two mammalian species exhibit four clusters as found for *H. sapiens*. The BCA of samples made by 2880 promoters for both species are reported in Figure 2.6. The similarity with *H. sapiens* is evident, thus suggesting that the clustering procedure singles out quite universal global properties of mammalian promoters. On the other hand, this clustering procedure does not work for *D. rerio* and *A. thaliana*: the alignment algorithm applied to the entire promoter extension does not allow to identify different clusters. In fact, according to the results reported in [10,26], this is not an unexpected result. BCA as well as entropic indicators show that in these species the region affected by some functional constraints of promoters reduces to a portion of the whole sequence close to the TSS, typically extending over 100 nucleotides. Accordingly, for *D. rerio* and *A. thaliana* we have applied the alignment algorithm to this shorter and certainly more specialized region of the promoter. With such a recipe we have obtained again clear signatures of different promoter clusters. For instance, in Figure 2.7 we show that a sample of 2880 promoters of *D. rerio* yields four clusters of almost equal size, dominated in the last 100 nucleotides by A,T, C and G nucleotides, respectively. For what concerns *A. thaliana*, we have obtained just two clusters (see Figure 2.7) whose BCA exhibits significantly different features only in the region close to the TSS, where either A or T nucleotides dominate.

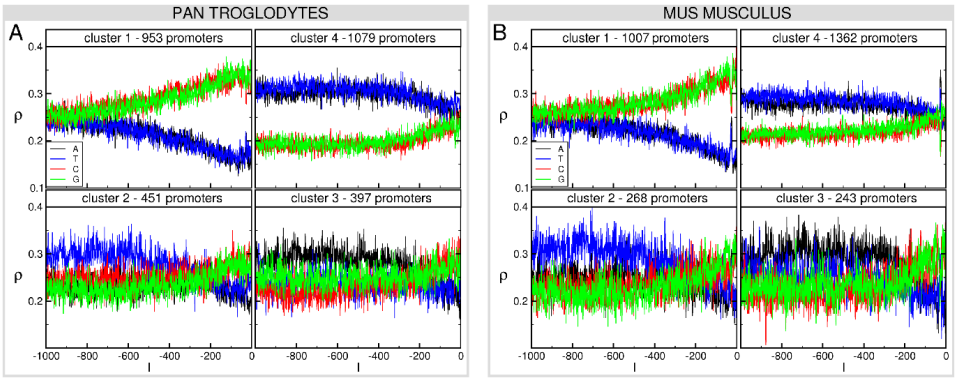


Figure 2.6 BCA of each of the clusters obtained with the clustering algorithm for *P. troglodytes* (panel A) and *M. musculus* (panel B). We report the frequency ρ of each of the four nucleotides A (black), T (blue), C (red) and G (green) as a function of the position l along the promoter (0 corresponds to the TSS).

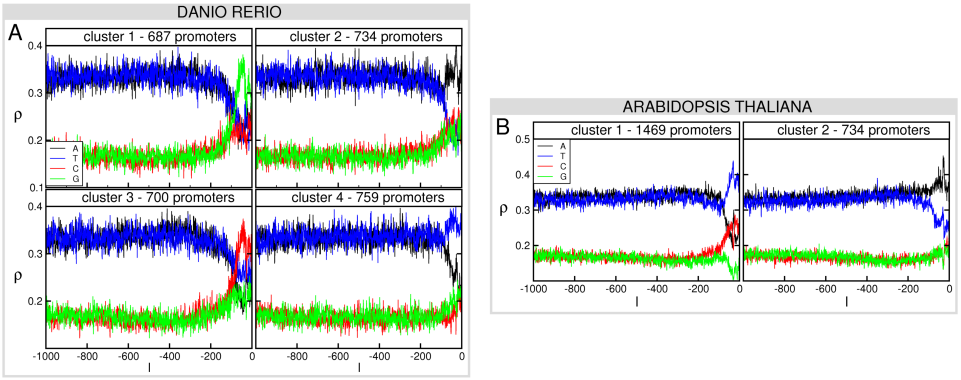


Figure 2.7 BCA of each of the clusters obtained with the clustering algorithm for *D. rerio* **(panel A) and** *A. thaliana* **(panel B).** We report the frequency ρ of each of the four nucleotides A (black), T (blue), C (red) and G (green) as a function of the position l along the promoter (0 corresponds to the TSS). Note that alignment and clustering are performed taking into account only 100 nucleotides before the TSS.

2.3.2 Most frequent regular sequences in other species

Figure 2.8 contains the list of the 15 most frequent regular sequences found in each cluster of *P. troglodytes* and *M. musculus*. The left column shows the percentage of promoters in each cluster that contain the sequence at least once. We have observed that, in most cases, any sequence is found inside a promoter only once. Accordingly, the sequences contained in a large percentage of promoters are also the most frequent ones. The large majority of the regular sequences in *P. troglodytes* coincide with those of *H. sapiens*, while the most frequent sequences of *M. musculus* are quite different from those of these two primates. After having computed how many times each regular sequence appears in all the promoters of each cluster, in the right column of Figure 2.8 we report the fraction of times it is contained inside a transposon. We find evidence of a strong correlation between the most frequent regular sequences of C2 and C3 and transposons. In *D. rerio* and *A. thaliana*, the search for regular sequences has been performed in all of the 1000 nucleotides of each promoter, even if the clusters differentiate only in the 100 nucleotides upstream the TSS. We have found that, at variance with mammals, the most common regular sequences are typically the same in all the clusters. Accordingly, in Figure 2.9 we report the data of the 15 most common regular sequences found in the whole sample of 2880 promoters.

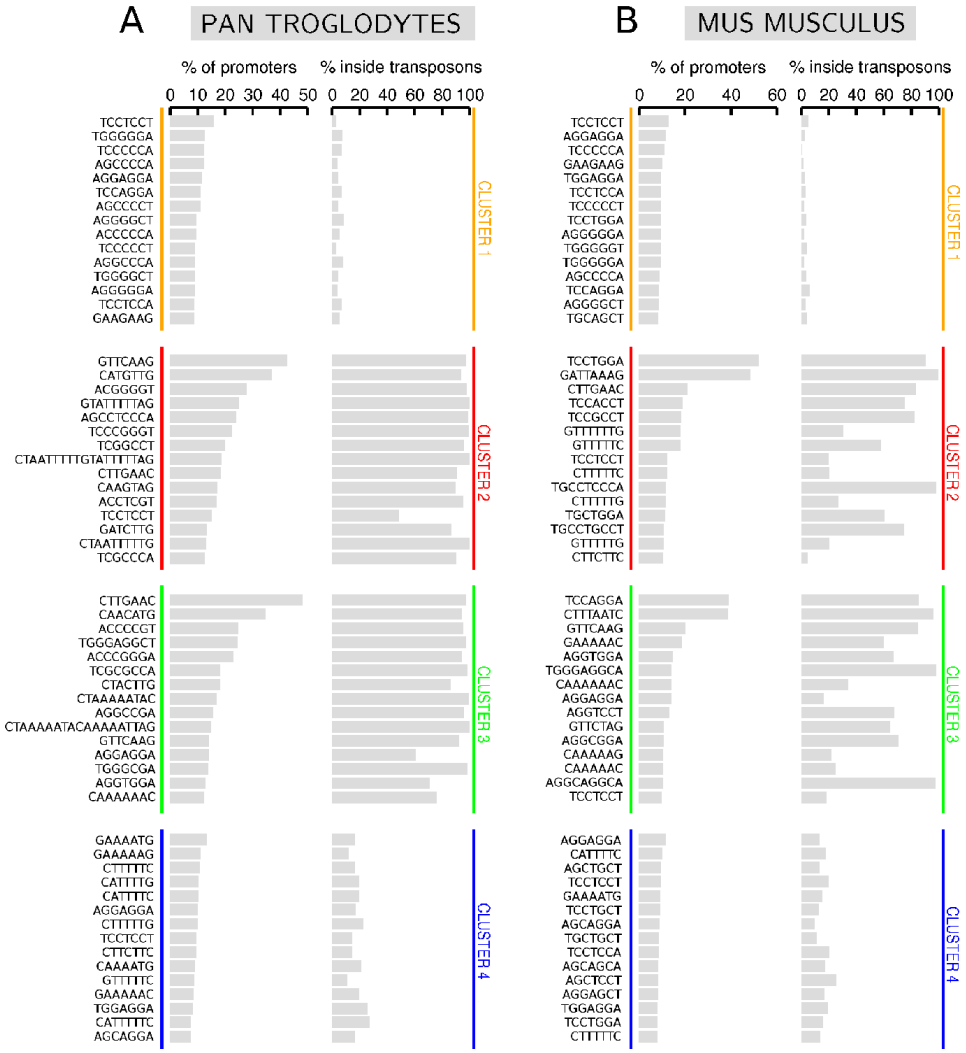


Figure 2.8 The most frequent regular sequences found in the clusters of *P. troglodytes* **(panel A) and** *M. musculus* **(panel B).** We report the percentage of promoters of the cluster in which the sequence appears at least once (left column), and the percentage of times the sequence is found inside a transposon (right column): it is calculated dividing the number of times it appears in a transposon by the total number of times it appears in the cluster.

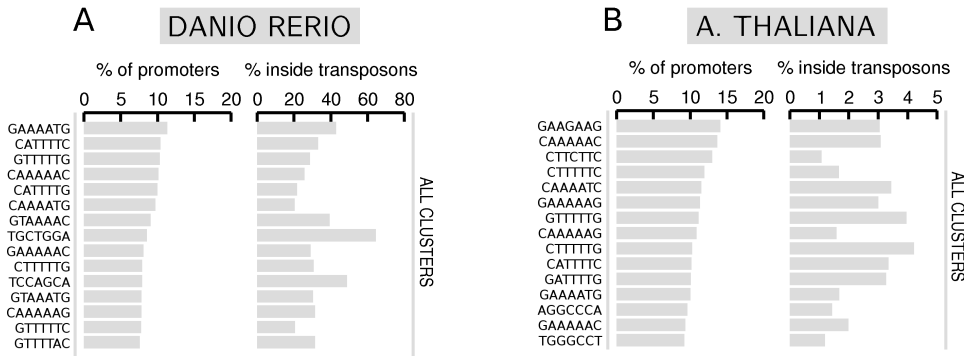


Figure 2.9 The most frequent regular sequences found in the entire sample of 2880 promoters of *D. rerio* (panel A) and *A. thaliana* (panel B). We report the percentage of promoters in which the sequence appears at least once (left column), and the percentage of times the sequence is found inside a transposon (right column): it is calculated dividing the number of times it appears in a transposon by the total number of times it appears in the cluster.

2.3.3 Transposons

In this section we report the results obtained with the Repeat Masker software [67] (see Methods), that screens DNA sequences for transposons. We have identified transposons in the clusters of *P. troglodytes*, *M. musculus*, *D. rerio* and *A. thaliana*. The transposon content of each cluster and the percentage of the most frequent family of transposons for *P. troglodytes* are very close to those obtained for *H. sapiens* (Figure 2.10). Some similarities with *H. sapiens* and *P. troglodytes* still emerge in *M. musculus* (Figure 2.11). As in *H. sapiens*, we have observed that the regular sequences in C2 and C3 of *P. troglodytes* (*M. musculus*) are mostly related with Alu elements (B1 elements). This aspect reflects both the conserved features of the old Alu families which spread among the mammalian genome before the primate-rodent split about 80 million year ago and the more recent primate-specific and murine-specific features acquired after their divergence [76]. On the other hand, in *D. rerio* and *A. thaliana* we do not observe differences in transposon content among the clusters (Figure 2.12 and Figure 2.13). Nonetheless, we find that regular sequences and transposons are definitely less correlated in *D. rerio* than in mammals. As far as *A. thaliana* is concerned such a correlation is even weaker (see Figure 2.9)).

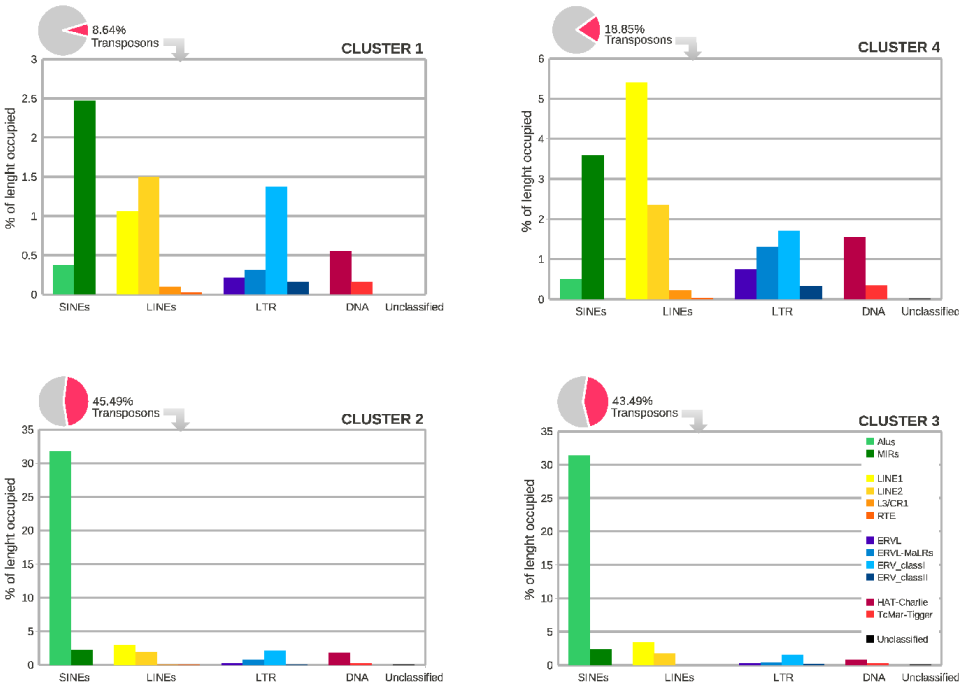


Figure 2.10 Distribution of the different families of transposons in the four clusters of *P. troglodytes*. We report the total percentage of nucleotides in the cluster covered by transposons (pie chart) and the percentage of nucleotides covered by each family of transposons (histogram). Note the different scales in the histograms.

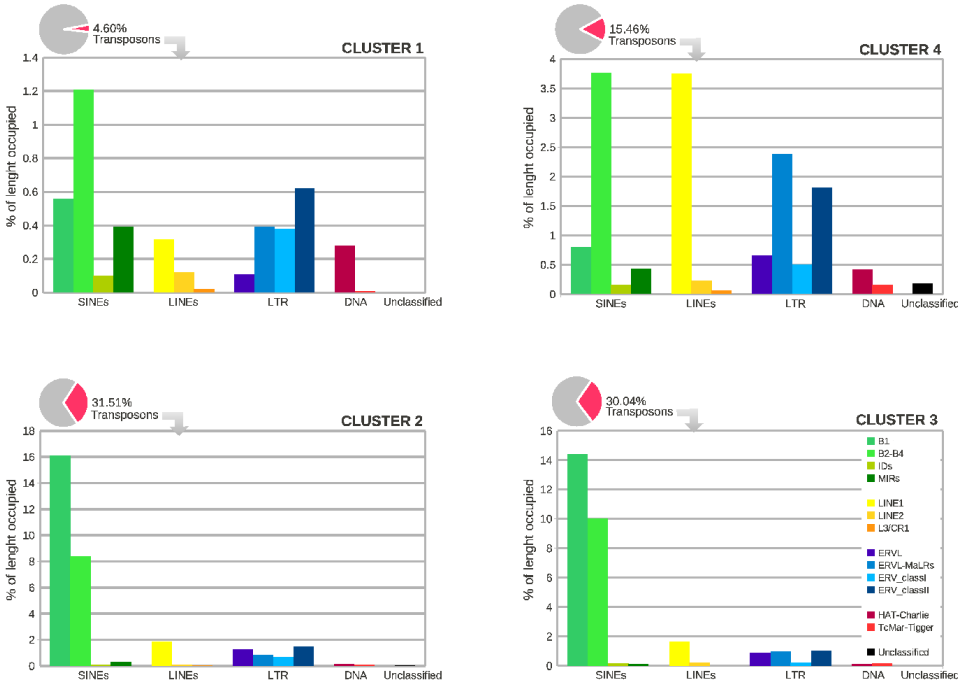


Figure 2.11 Distribution of the different families of transposons in the four clusters of *M. musculus*. It is shown the total percentage of nucleotides in the cluster covered by transposons (pie chart) and the percentage of nucleotides covered by each family of transposons (histogram). Note the different scales in the histograms.

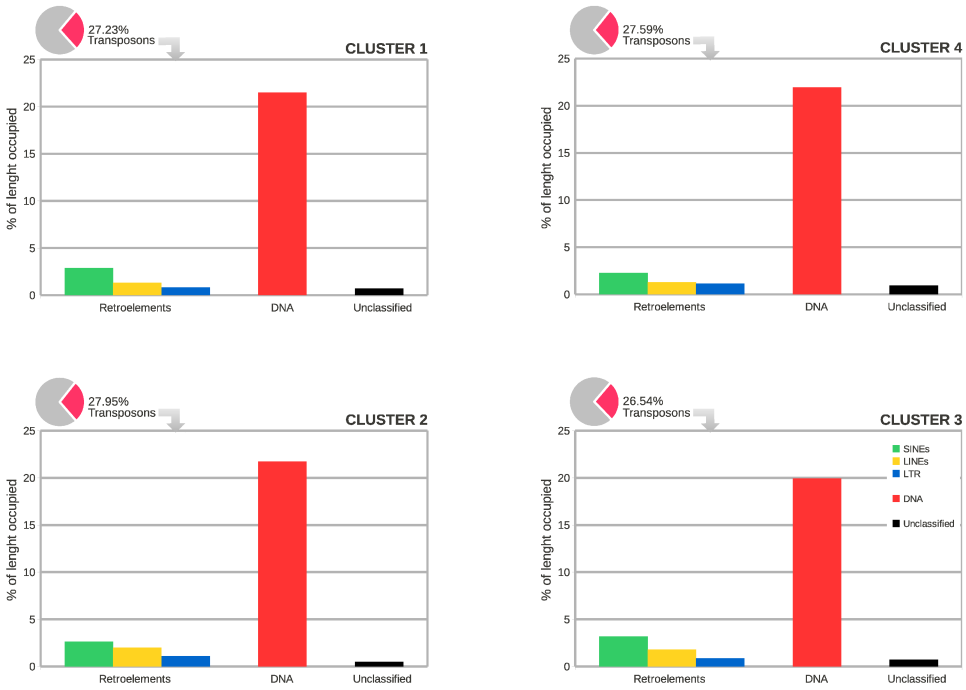


Figure 2.12 Distribution of the different families of transposons in the four clusters of *D. rerio*. It is shown the total percentage of nucleotides in the cluster covered by transposons (pie chart) and the percentage of nucleotides covered by each family of transposons (histogram).

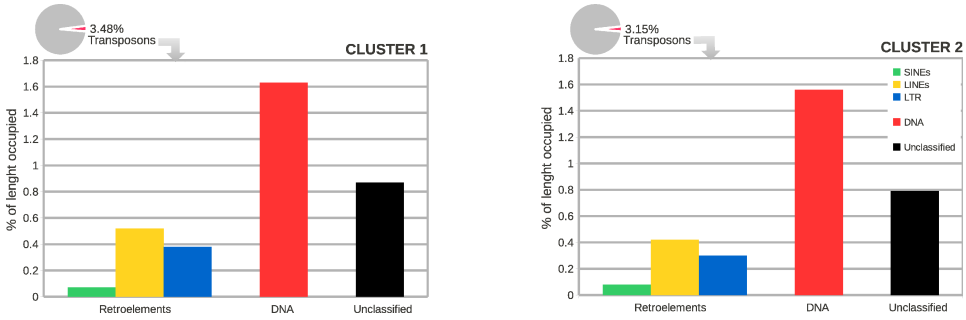


Figure 2.13 Distribution of the different families of transposons in the two clusters of *A. thaliana*. It is shown the total percentage of nucleotides in the cluster covered by transposons (pie chart) and the percentage of nucleotides covered by each family of transposons (histogram).

2.3.4 CpG dinucleotide analysis

A common explanation of the GC rise in mammalian promoters is the presence of the so-called CpG islands, i.e. GC-rich regions of DNA (typically 0.5-2 kb in length) that are relatively enriched in CpG dinucleotides with respect to the rest of the genome [58, 59]. In a previous work [37] it was addressed the question if the patterns observed in TATA-less sequences (those corresponding to C1 in this paper) could derive from this mechanism of CpG enrichment at promoter level. It was found that CpG dinucleotides increase towards the TSS with the same rate of the other three dinucleotides, i.e. GpC, CpC and GpG, despite they are still relatively under-expressed. In particular, all the four dinucleotides provide comparable contributions to the increase in GC content close to the TSS (see Fig. 2.14). We want to point out that this finding is coherent with the recent novel evolutionary model for the origin of CpG islands in promoter regions [60]. The absence of indications in favor of the selection on CpG densities suggests that the CpG increase at promoter level may be the result of the GC enrichment and not viceversa. This notwithstanding, the functional involvement of CpG is not excluded. In fact, the regulation through methylation of CpG could be the indirect cause of CpG hypomethylation and slow decay in a large number of promoter regions [60].

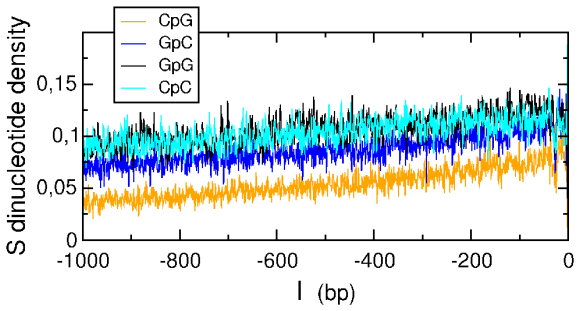


Figure 2.14 CG content and CpG islands. We report dinucleotide density S as a function of the position along the promoter (0 corresponds to the TSS). Data are obtained analysing the promoters of C1 of *H. sapiens*.

2.4 Methods

2.4.1 Databases

The promoters of *H. sapiens*, *M. musculus* and *D. rerio* have been downloaded from DBTSS (Version 6.0), a database of TSSs, obtained from a collection of experimentally determined 5'-end sequences of full-length cDNAs [61]. *P. troglodytes* promoters have been downloaded from ECRbase, a database which provides a comprehensive collection of promoters generated by using expressed sequence tag (EST) and mRNA data [62]. The promoters of *A. thaliana* have been downloaded from a database where annotation of genes is largely based on sequenced cDNAs and ESTs alignments with the genome, TAIR (The Arabidopsis Information Resource) web site [63] (released in March 2008).

2.4.2 TATA–box

Following Yang et al. [11], the TATA-box consensus sequence has been searched from position $l = -80$ to $l = -1$ in the top strand of each promoter by an exact–match search. It corresponds to the degenerate sequence HWHWWWWR (coded according to IUPAC nomenclature), which identifies 576 sequences (in the nucleotide quaternary code). In order to fit the structural definition of the interaction with the TATA–binding protein, 44 specific strings have been excluded, so that the actually employed sequences reduce to 532 elements. Each promoter is called TATA if a TATA box consensus sequence is found at least once, otherwise it is called TATA-less. We have searched all the same degenerate boxes in the sets of promoter sequences of all the investigated species.

2.4.3 Spectral Clustering

The aim of the procedure described in this section is to divide the collected promoters into groups depending on the similarity between the sequences. The method is structured into three main steps: the first one consists in aligning each sequence with all the others (pairwise alignment) so as to obtain a matrix of similarity scores. Then, the analysis of the spectral properties of the Laplacian matrix calculated from the similarity matrix enables one to determine the appropriate number of groups for the clustering procedure. The last step, based on the k–means algorithm, associates each sequence to one of the clusters.

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 evolution-

ary 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 allow to insert gaps in the sequences in order to align similar regions of the two sequences - clearly also gap insertion is associated to a score penalty. An example of alignment is reported in Figure 2.15. These algorithms, in general, fall into two categories: global and local techniques. 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 different overall. In this paper we have made use of the two most popular alignment methods, the Needleman–Wunsch global algorithm [28] and the Smith–Waterman local algorithm [29] implemented in the EMBOSS package version 6.3.1 [30].

```
----tctattgcagatttg----tgtgaccaagc
    ||    |||..||||    ||..|||
agcgtc----gcacgtttgaatttggcacc----
```

Figure 2.15 Example of alignment. In this figure we report an example of alignment of two sequences. The gap insertions in both sequences are shown, as well as matches and mismatches between sequences.

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 (GAPOPEN) and the value for each consecutive gap (GAPEXTEND), because the scoring matrix for the nucleotide match and mismatch has been taken equal to the standard EDNAfull matrix for both methods. 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 20) and a low GAPEXTEND penalty (0.5 or 1) in order not to penalize long gap sequences. This setting favors the scores of very similar sequences yielding an easier detection of the correct number of clusters (see section *The normalized Laplacian matrix*). 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 promoters gives the same similarity matrix S in both cases.

The normalized Laplacian matrix

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 sequences, while the entry s_{ij} represents the weighted link between sequence i and j. For the purpose of our work, however, dealing with a full 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 promoters, 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 that a certain threshold s^*, namely $w_{ij} = s_{ij}$ if $s_{ij} \geqslant s^*$ and $w_{ij} = 0$ otherwise. To estimate s^*, we have associated to each set of N analyzed promoters, the corresponding N reshuffled sequences, namely the sequences obtained randomly rearranging the nucleotides of each promoters. Then we have performed the alignment, and calculated s^* as the arithmetic mean of s_{ij}. To check the correctness of s^*, we have monitored s^* as a function of N and we have observed the convergence of s^* to a constant value for N approaching the values used in our simulations ($N = 1440$, $N = 2880$; the choice $N = 2880$ is due to the constraints on both the computational time and the size of the matrix to be stored). Finally, in order to manage a set of more homogeneous data, we have operated the normalization $d_{ij} \to w_{ij}/max\{w_{ij}\}$.

Following [34], once an appropriate adjacency 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 [64]. Indeed, given the spectrum $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_N$ 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. Fig. 2.16 helps to understand this approach. Both panels show the first part of the spectrum of L_{sym} associated to the alignment of 2880 *H. sapiens* promoters with the global algorithm. The first value is zero, and then three consecutive eigenvalues, located far from the others, follow.

Accordingly, the resulting number of clusters is 4. The distance from the fourth eigenvalue to the fifth one is larger in panel B where we used a higher GAPEXTEND value.

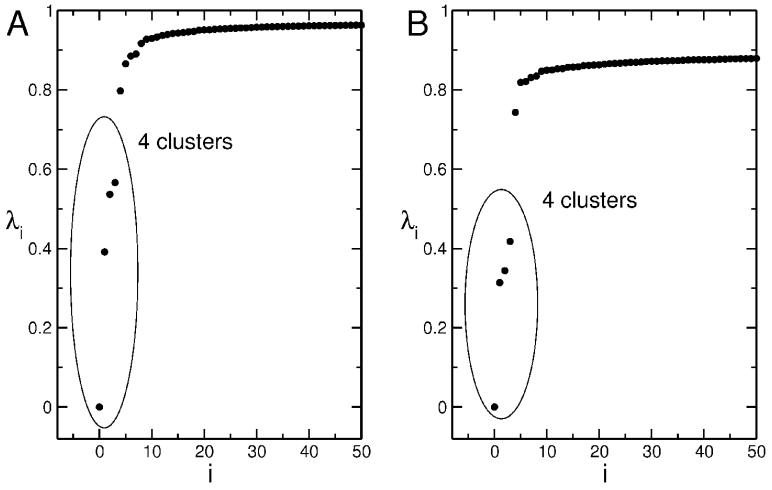


Figure 2.16 Eigenvalues of the Laplacian matrix. First 50 eigenvalues in ascending order of the normalized Laplacian matrix relative to the alignment of 2880 *H. sapiens* promoters. The method used is the Needleman–Wunsch with GAPOPEN = 20 and GAPEXTEND = 0.5 for panel **A**, GAPEXTEND = 1.0 for panel **B**.

Clustering algorithm: K-means

We are now able to apply the spectral clustering algorithm in order to assign each promoter to one of the clusters. The starting point is the computation of the first k eigenvectors $u_1, \ldots, u_k$ of L_{sym}, so as to form a new matrix $U \in R^{N \times k}$ containing the vectors $u_1, \ldots, u_k$ as columns. Let $T \in R^{N \times k}$ be the matrix obtained from U by normalizing the rows to norm 1, namely, $t_{i,j} = u_{i,j} / \left(\sum_k u_{i,k}^2 \right)^{1/2}$. For $i = 1, \ldots, N$ we denote by $y_i \in R^k$ 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 $A_1, \ldots, A_k$ 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 panel A of Fig. 2.1 this tolerance was fixed to 10^{-5}.

2.4.4 Spectral method for identification of regular sequences

Nucleotide sequences in promoters are characterized by the alternation of regular and disordered regions of different length. In particular, the regular ones exhibit various structures, ranging from homogeneous to periodic and palindromic. In this section we describe a method for the identification of all these regular sequences starting from the properties of a mechanical model of the DNA chain. It is worth pointing out that the method is based on a definition of *regularity* of finite–length regions in a promoter, that combines suitable quantitative indicators.

In practice, we adopt the model introduced by Peyrard and Bishop [31–33] (see section *Peyrard-Bishop model*). This model simplifies the molecular structure of the DNA by considering only one strand and neglecting the double-helix structure. It takes explicitly into account the nonlinear interactions between the nucleotides and, despite its apparent simplicity, it is quite effective for reproducing the dynamics of DNA at physiological temperatures. For our purposes, it is sufficient to consider the *harmonic* approximation of this model, that is valid in the low–temperature limit. In this sense, what remains of the information contained in the Peyrard-Bishop model are the presence of nearest–neighbor and on–site harmonic interactions and the phenomenological parameters defining their strength (see Eq. (2.3)). In section *Normal modes* we show that the properties of the chain in the *harmonic* regime are completely determined by the features of the Hessian matrix of the model.
Finally, in section *Determination of regular sequences*, we describe the procedure for the determination of the regular sequences using the eigenvectors of the Hessian matrix.

At variance with the notation adopted for labeling the position of nucleotides in a promoter (namely, $l = -1000, \cdots, -1$), in what follows we adopt the standard numeration for the index i of the sites in an oscillator chain, namely $i = 1, \cdots, L$ (with $L = 1000$ for promoters).

Peyrard-Bishop model

In the Peyrard-Bishop model each nucleotide $i = 1, ..., L$ is associated with one degree of freedom y_i, that corresponds to the displacement of the nucleotide from its equilibrium position. This displacement is in the direction of the hydrogen bonds connecting a nucleotide to its complementary in the opposite strand. The state of the chain is completely determined by the vector $\vec{y} = (y_1, ..., y_L)$. The interaction due to the hydrogen bonds is modeled by a Morse potential. Moreover, the model contains a stacking interaction between nearest neighbor nucleotides: the strength of this interaction decreases when the complementary nucleotides are farther. The total potential energy $U(\vec{y})$ is given by

$$\sum_i \left[\frac{K}{2}(1+\rho \mathrm{e}^{-\alpha(y_{i+1}+y_i)})(y_{i+1}-y_i)^2 + d_i(\mathrm{e}^{-a_i y_i}-1)^2 \right] \qquad (2.1)$$

The parameters K, ρ and α refer to the stacking interactions between two consecutive nucleotides; while the parameters d_i and a_i define the depth and the width of the Morse potential, respectively. In order to model heterogeneous DNA sequences two different values for the couple (d_i, a_i) are considered according to the two possible kind of nucleotides, weak (W) and strong (S). The former has two hydrogen bonds, while the latter has three hydrogen bonds. Therefore, the depth for the S Morse potential is chosen 1.5 times the one of the W Morse potential. The model is characterized by a *dichotomic* disorder along the chain: every nucleotide can be associated to the couple of values (d_W, a_W) or (d_S, a_S). The ground state of the model (i.e., the state of minimal energy) corresponds to a configuration of the chain with $\vec{y} = \vec{0}$. For the promoters analyzed in this paper we have $L = 1000$, while the parameter set is the one adopted in [65] (in order to avoid convergence problems in the algorithm for the diagonalization of the Hessian matrix of the potential U we chose $K = 0.030$ eV/Å^2 instead of 0.025 eV/Å^2).

Normal modes

The normal modes of the Peyrard–Bishop model of the DNA chain represent small oscillations around the ground state. In order to fully characterize them we need to know the frequencies and the amplitudes of oscillations of every nucleotide (that is equivalent to a harmonic oscillator). A normal mode is in fact a collective motion where every nucleotide vibrates with the same frequency but with a different amplitude. As the chain has L degrees of freedom there are L different ways of oscillation.

Approximation of the potential energy. From a mathematical point of view the normal–mode approach corresponds to consider a Taylor series expansion of the potential energy around the minimum $\vec{y} = \vec{0}$. At the second order it reads

$$U(\vec{y}) \simeq U(\vec{0}) + \nabla U(\vec{0})^T \vec{y} + \frac{1}{2}\vec{y}^T H(\vec{0})\vec{y} \qquad (2.2)$$

where $H_{ij} = \frac{\partial^2 U}{\partial y_i \partial y_j}$ is the symmetric Hessian matrix of the potential energy. Since in the minimum of the potential $U(\vec{0}) = \vec{0}$ and $\nabla U(\vec{0}) = \vec{0}$, Eq.(2.2) reduces to

$$U(\vec{y}) \simeq \frac{1}{2}\vec{y}^T H(\vec{0})\vec{y} = \frac{1}{2}\Big(\sum_{i=1}^{L} A_i y_i^2 + 2B \sum_{i=1}^{L-1} y_i y_{i+1} \Big) \qquad (2.3)$$

where: $A_i = [2d_i a_i^2 + 2(1+\rho)K]$ for $i = 2, \cdots, L-1$, $A_i = [2d_i a_i^2 + (1+\rho)K]$ for $i = 1, L$ and $B = -(1+\rho)K$. This amounts to the *harmonic* approximation, where the properties of the potential energy are summarized in the Hessian matrix evaluated in the minimum of the potential.

Hessian matrix: eigenvalues and eigenvectors. By a suitable change of coordinates $\vec{y} \to \vec{x}$, the quadratic form (2.3) can be rewritten in a diagonal form by a standard procedure (this is done by solving the spectral problem for the Hessian matrix, i.e., $H_d = A^T H A$ where A is an orthogonal matrix $A^T = A^{-1}$, H_d is the Hessian matrix in diagonal form and by setting $\vec{y} = A\vec{x}$). In the new variables, U reads as the energy associated to L harmonic springs

$$U(\vec{x}) \simeq \frac{1}{2}\vec{x}^T H_d \vec{x} = \frac{1}{2}\sum_{k=1}^{L} \lambda_k {x_k}^2 \tag{2.4}$$

where H_d is the diagonal form of the Hessian matrix and λ_k are the eigenvalues.

The eigenvectors $e_k(i)$ of the Hessian matrix (where $i = 1, ..., L$ is the nucleotide index relative to the TSS) are the eigenmodes of the DNA chain.

Properties of the eigenvectors. Regular sequences in the promoters are recovered by looking at eigenvectors of the Hessian matrix with suitable features of delocalization according to the method described in section *Determination of regular sequences*. In order to apply this procedure, the following indicators have been used to fully characterize the eigenvectors.

1. the *eigenvector center of mass*, x_k^{cm}, signals the position of the center of the eigenvector along the promoter chain and it is defined by

$$x_k^{cm} = \frac{\sum_{i=1}^{L} |e_k(i)| i}{\sum_{i=1}^{L} |e_k(i)|}, \tag{2.5}$$

2. the *eigenvector extension* along the chain is quantified by

$$\Delta_k = 2\sqrt{\Big(\frac{\sum_{i=1}^{L} |e_k(i)| i^2}{\sum_{i=1}^{L} |e_k(i)|}\Big) - (x_k^{cm})^2} \tag{2.6}$$

3. the *eigenvector participation number*, ξ_k, is a measure of the degree of delocalization of the eigenvector and it is defined by

$$\xi_k = \Big(\sum_{i=1}^{L} |e_k(i)|^4\Big)^{-1}; \tag{2.7}$$

for an eigenvector localized on a single site $\xi \simeq 1$, while for a completely delocalized eigenvector $\xi \simeq L$ (the eigenvectors are normalized to unity, i.e. $\sum_{i=1}^{L} |e_k(i)|^2 = 1$).

We want to point out that both the extension and the participation number are necessary to define the properties of the eigenvectors, because the two indicators are not always positively correlated (see Fig. 2.17). In fact, for some eigenvectors the degree of delocalization essentially coincides with the extension of the eigenvector (see panel A of Fig. 2.18). On the other hand, there are eigenvectors having very small participation number despite the very large extension, and this is typically due to the presence of very large components on a few sites and much smaller components on many sites in between (see panel B of Fig. 2.18).

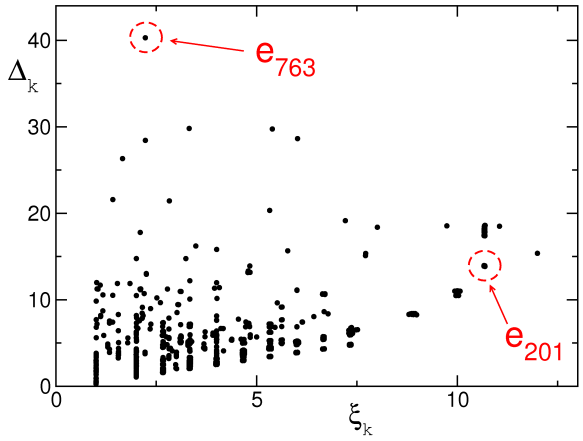


Figure 2.17 Eigenvector extension, Δ_k, as a function of the participation number, ξ_k. The (red) dashed circles refer to eigenvectors with different properties of localization. The eigenvector $e_{201}(i)$ (see Fig. 2.18 in Methods) has comparable values of ξ and Δ. While $e_{763}(i)$ (see Fig. 2.18 in Methods) has a small participation number, $\xi \simeq 2$, but large extension ($\Delta \simeq 40$). Data refer to the promoter of *H. sapiens* with Entrez GeneID 9542.

Determination of regular sequences

By regular sequence we mean a region of a promoter that exhibits any spatial regularity in the *weak-strong* binary code. Eigenvectors with large enough degree of delocalization, determined by the participation number ξ_k, generally extend over regular regions. Accordingly, the method for the identification of the regular sequences, that we are going to describe in detail, needs from the very beginning a conventional definition of *delocalized* eigenvectors and of their *effective* extension along the sequence (criteria I and II).

I. We consider *delocalized* those eigenvectors with participation number exceeding a fixed threshold value, i.e. $\xi_k \geq 3.9$, that typically correspond to a region of at least 7 nucleotides. This *heuristic* choice

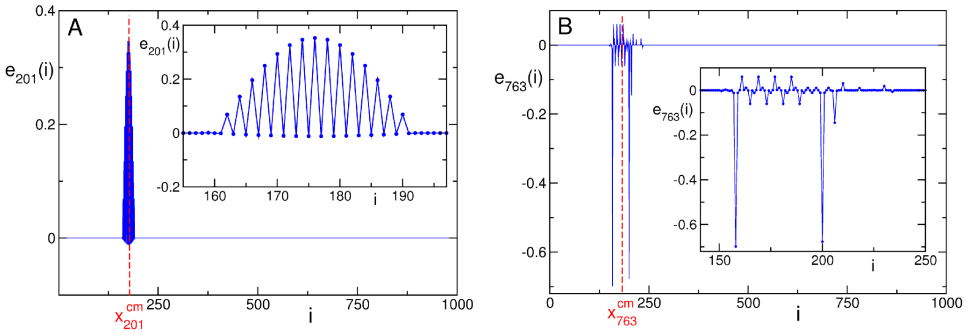


Figure 2.18 Eigenvectors of the Hessian matrix with different properties of delocalization. The eigenvector $e_{201}(i)$, in panel **A**, has comparable values of participation number and extension ($\xi \simeq 11$ and $\Delta \simeq 14$), while the eigenvector $e_{763}(i)$, in panel **B**, has a small participation number, $\xi \simeq 2$, but very large extension ($\Delta \simeq 40$). In the insets an enlargement of the region of delocalization is shown. Data refer to the promoter of *H. sapiens* with Entrez GeneID 9542 (the promoter of the neuregulin-2 gene). Entrez Gene is the gene-specific database at the National Center of Biotechnology Information (NCBI) [66].

is justified by the fact that many regular motifs of biological interest correspond to such a size (e.g, the TATA–box, that contains 8 nucleotides).

II. The start-site, i_{start}, and end-site, i_{end}, of a delocalized eigenvector are identified according to the following conditions,

i_{start} : $|e_k(i_{start}-1)| \leq \theta$ and $|e_k(i_{start})| > \theta$,

i_{end} : $|e_k(i_{end}-1)| \geq \theta$ and $|e_k(i_{end})| < \theta$,

with $\theta = 0.05$. The heuristic choice of the value of the threshold θ allows to remove the ambiguity that can be introduced by very small components of the eigenvectors (see Fig. 2.19).

Moreover, we use the property that the eigenvectors of an isolated regular region overlap with the eigenvectors of the whole promoter in that region.

A regular region of the promoter composed of n nucleotides has exactly n eigenvectors and if we could ideally neglect border effects also the whole promoter would have n eigenvectors extending over the regular region. Actually, in practical cases this condition on the number of the eigenvectors of the whole promoter can be only approximatively satisfied.

Therefore, the procedure for the determination of regular sequences is summarized in the following steps:

1. identification of the start-site and of the end-site for all the *delocalized* eigenvectors (see criteria I and II);

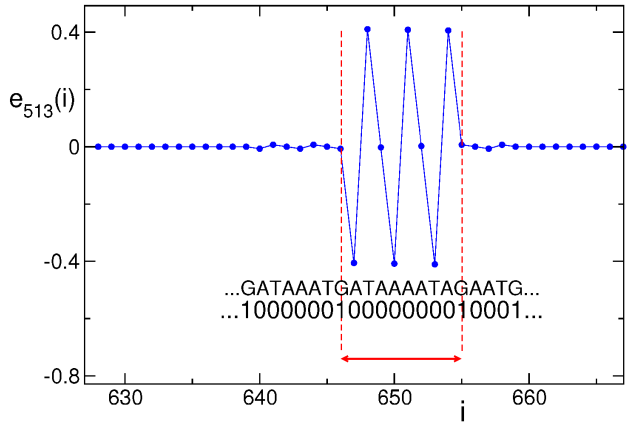


Figure 2.19 Start site and end site of an eigenvector. Determination of the effective extension (region in between the dashed lines) of a *delocalized* eigenvector overlying regular sequences. Notice the very small components of the eigenvectors aside the regular region. A portion of the sequence is reported both in quaternary and in binary code. Data refer to the promoter of *H. sapiens* with Entrez GeneID 54808 (the promoter of the dymeclin gene).

2. determination of the number of eigenvectors between the start-site and the end-site and comparison with the number of nucleotides contained in the same region: these quantities are assumed to be equivalent within a 30% tolerance.

In Fig. 2.20 we show some examples of regular sequences determined by delocalized eigenvectors. Following this procedure we were able to rule out false identifications.

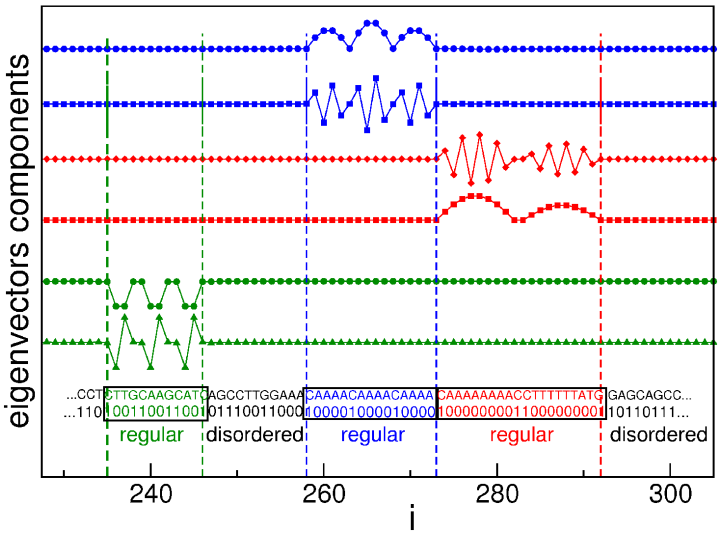


Figure 2.20 Regular and disordered sequences of a promoter. The regular sequences (highlighted in the black frames) are determined by the *delocalized* eigenvectors of the Hessian matrix. For the sake of clarity, for each of the three examples shown here we report just two of the eigenvectors, whose total number is 10 (green case), 16 (blue case) and 16 (red case). The sequence of the promoter is reported both in quaternary and in binary code. The curves refer to eigenvectors n. 988, 577, 567, 998, 946, 627 (resp. from the top to the bottom) of promoter with Entrez GeneID 9542 of *H. sapiens*.

2.4.5 Repeat Masker

Trasposons were identified by RepeatMasker [55], version 3.3.0, a program that screens DNA sequences for interspersed repeats. The output of the program is a detailed annotation of the repeats that are present in the query sequence. The options were chosen as follows:

Search engine: abblast
Speed/sensitivity: Default
DNA source: Human for *H. sapiens*, Mammal for *P. troglodytes*, Mouse for *M. musculus*, Danio for *D. rerio*, Arabidopsis thaliana for *A. thaliana*.
Comparison species: none
Alignment options: no alignments returned
Masking options: Repetitive sequences in lower case
Contamination check: No contamination check
Repeat options: Don't mask simple repeats or low complexity DNA
Artifact check: Report E. coli IS artifacts
Matrix: RepeatMasker choice
Divergence cutoff: none

2.5 Conclusion

In this chapter we performed a genome-wide analysis of *H. sapiens* promoters by exploiting a fully general mathematical procedure based on the combination of two spectral methods. The first one is a clustering algorithm that allows us to classify promoters according to global similarities. The second spectral method is capable of detecting any regular sequence in each promoter, without imposing any preliminary constraint. The clustering analysis showed that *H. sapiens* promoters can be pooled into four main groups. Two of the clusters are distinguished by the prevalence of weak or strong nucleotides and are characterized by short compositionally biased sequences. In the two remaining clusters regular regions are found to be correlated with transposons, that are known to play a major role in favoring evolutionary changes in cis-regulatory regions, as conjectured by some authors [25, 52, 56, 57]. A posteriori, we are therefore led to conclude that these two clusters actually represent a single one.

In summary, the main biologically relevant findings consist in the following:

- Promoters can be classified according to common global properties of the whole sequence and not on the basis of the presence of specific patterns in specific positions (as for example in the usual TATA/TATA-less classification or other specific short regulatory motifs);
- Promoters with the highest content of transposons group together in C2 and C3;
- The most expressed regular sequences of these clusters are essentially located inside transposons;
- Conversely, in clusters C1 and C4 (where strong and weak nucleotides are respectively dominant) the most expressed regular sequences appear equally distributed along the promoters without any specific relation with transposons.
- The different compositional properties of C1 and C4 reflect their different functional role, since promoters in C1 are mostly associated to housekeeping genes while those in C4 are associated to tissue-specific genes.

Moreover, the generality of the unbiased methods presented here allowed us to extend them to the investigation of promoter databases of other species. In Supplementary Materials we show that the comparison of *H. sapiens* with other mammalian species points out that such species seem to be generally characterized by the presence of the same cluster organization. On the other hand, while the promoter structural properties of *H. sapiens* and

P. troglodytes are almost identical, we find that *M. musculus* exhibits some differentiation in the most frequent regular sequences as well as in the correlation with transposons. An even more pronounced differentiation with respect to mammalian species is found in the promoters of a fish, *D. rerio*, and of a plant *A. thaliana*. At variance with mammalian promoters, where the information content spreads all over the promoter length, we have found that the clustering of promoters in these latter species is associated with a relatively short region (≈ 100 nucleotides) close to the TSS. Such a sharp differentiation of promoters structure in different species indicates that these DNA components are suitable candidates also for investigating the effects of evolutionary selection on DNA.

In the next Chapters, we will present the directions we decided to undertake in this PhD Thesis. They aim to deepen the analysis of the results presented in this paragraph and in [27], to study the biological relevance of the classification we obtained in determining the promoters' properties and role in different contexts.

Chapter 3

Entropic analysis of promoter sequences

3.1 Introduction

In the previous chapter we have described the clustering method that identifies three groups of promoters in *H. sapiens*, and we have pointed out the main features of each group.

Now we want to tackle promoter characterization in a different way, in order to deepen the analysis of promoters from a different point of view: in fact, we perform here an analysis of human promoters treating them as strings of a text. A procedure for characterizing promoter sequences can be worked out by suitable entropic indicators. In [10] the positional Shannon entropy proved to be quite a useful tool for identifying differences between TATA and TATA-less promoters. In this chapter we perform a more careful analysis of human promoters deepening the entropic analysis of promoter sequences.

For this purpose we use the method introduced in [67], where the standard Shannon entropy is combined with a specific entropic indicator that allows one to

1. Study the variability of a sample of measures.

2. Identify the keywords in a text.

Our purpose here is to apply these entropic indicators to the analysis of promoter sequences. The analysis of variability and the identification of the keywords will be applied to the entire set of human promoters and to the clusters obtained in 2. Regarding point 1, we aim to study the variability across different promoters as a function of the position along the sequence. This is discussed in Sec.3.2. This takes inspiration from [10] whose aim was to use Positional Shannon Entropy to investigate the correlations between

structure and function in gene promoter sequences and to obtain an insight on putative selective constraints to randomness at promoter level. Selective constraints in promoter sequences may be present due to the optimization of the interactions with the transcription factors, or in general because of functional constraints acting on the structure and composition of promoters: regions subject to such constraints can be identified as those that exhibit a biased variability across promoters with respect to the nearby regions, as well as those showing a selection on the sequences content (i.e. some of the possible words appear with a high frequency while other are suppressed). A good indicator of the variability across promoters is Positional Shannon Entropy, while another entropic indicator (introduced in [67]) allows to deepen the analysis about the selection on sequences content. Moreover, the analysis aims at characterizing the different features of the clusters obtained in Chapter 2.

Regarding point 2, we apply again the two entropic indicators for a different purpose. We treat promoters as strings of text to identify relevant information about promoter sequences via text mining techniques. This is reported in Sec. 3.3. Text mining refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents [68]. Most of its applications belong to language processing techniques, namely text analysis, information extraction, and summarization. For instance, a useful way to analyze a text is to identify its keywords. Such keywords are among the most frequent words that appear in the text, but at the same time they are very specific of the subject treated. Thus, a method that identifies keywords must be able to filter out generic words (i.e. "and", "that") that are just frequent but not significant. We will treat our promoter set as a text and we will apply this keywords analysis technique in order to identify putative important sequences. We expect to find the keywords (i.e. short nucleotide sequences) characterizing promoters. This analysis results in a set of short DNA sequences that are good candidates to have an important biological role for promoter functions. This set of sequences will undergo an in-depth survey on their biological role searching for their correlation with biologically relevant functions.

3.2 Positional Entropy

We have analyzed the text made up of the whole set of $M = 34170$ human promoter sequences [1]. This text is naturally divided in sections of 1000 nucleotides, i.e. the promoter sequences. Our analysis consists in measuring

[1] The database was downloaded from DBTSS (Version 8.0), a database of TSSs obtained from a collection of experimentally-determined 5'-end sequences of full-length cDNAs. Note that the number of promoters in this database is different from the previous chapter because the database was updated to a more recent release.

the variability across the M different promoters at each position along the whole promoter sequences, i.e. we study the sample $\{w_1^n(l); \cdots ; w_M^n(l)\}$ of M words of length n observed at position l in each promoter. In [10] Positional Shannon Entropy $H[w]$ has been employed to quantify the positional information content in promoter sequences. In [67] it has been introduced another entropic indicator, $H[K]$ that integrates the information given by $H[w]$. This new entropic indicator can be useful for extracting information from sampling a complex system. Here we want to apply both of these indicators to promoter analysis. For a fixed length n of word w and for a fixed position l along the promoters, K_w is the number of times word w is sampled at position l in the M promoters and $m_k = \sum_w \delta_{k,K_w}$ is the number of different words w that are sampled exactly k times. Entropies are defined as

$$H[w] = -\sum_w \frac{K_w}{M} \log \frac{K_w}{M} = -\sum_k \frac{km_k}{M} \log \frac{k}{M} \tag{3.1}$$

$$H[K] = -\sum_k \frac{km_k}{M} \log \frac{km_k}{M} \tag{3.2}$$

where $\sum_w$ denotes the sum over all the words of length n at position l, $\sum_k$ denotes the sum over all the measured occurrences k of the words of length n at position l. Note that it is implicit that *both entropies depend on l and n*.

In order to explain how $H[k]$ completes the information given by $H[w]$ on the variability of the sample $\{w_1^n(l); \cdots ; w_M^n(l)\}$, it is useful to consider some examples. In particular, we evaluate $H[w]$ and $H[K]$ in two opposite extreme cases: the uniform distribution case, in which all words have the same occurrence frequency, and the oversampled case, where the same word w_0 is repeated in the sample. This two cases help us to understand the kind of information provided by these two entropies, highlighting how $H[w]$ measures the variability of the sample and $H[K]$ measures the variability of occurrence frequencies.

Oversampled distribution: In the oversampled case the same word w_0 is observed M times, i.e. $K_w = M\delta_{w,w_0}$, $k = M$ and $m_k = 1$. Thus, $H[w] = 0$ and $H[K] = 0$.

$$H[w] = -\frac{M}{M} \log \frac{M}{M} = 0 \tag{3.3}$$

$$H[K] = -\frac{M}{M} \log \frac{M}{M} = 0 \tag{3.4}$$

Uniform distribution: This is the opposite extreme with respect to the previous case. Suppose we have a uniform distribution of all the different

words, i.e. all the words have the same occurrence frequency $K_w = M/N$ in the sample (N is the number of possible different words, it is implicit that $M > N$). Thus, $k = M/N$ and $m_k = N$. In this case we have a maximum of $H[w]$ since all the words are present with the same frequency and the sample variability is maximum. Nevertheless, $H[K]$ is zero.

$$H[w] = -\sum_{w} \frac{M}{M\ N} \log \frac{M}{M\ N} = \log N \tag{3.5}$$

$$H[K] = -\sum_{k} \frac{M\ N}{M\ N} \log \frac{M\ N}{M\ N} = 0 \tag{3.6}$$

In the oversampled case $H[K]$ is trivially zero since there is no variability at all in the sample. In the uniform case we have the maximum variability - and maximum $H[w]$ - but $H[K] = 0$ means that there is no "selection" on the words, since they all appear with the same frequency. Notice that if $K_w = K_{w'}$ the sample does not allow one to distinguish the two words w and w'. In this sense, $H[K]$ is not a mere measure of the variability of the sample (like $H[w]$) but it measures the significance of such variability. In intermediate cases, $H[w]$ will take an intermediate value in the interval $[0; H_{max}[w]]$ and we expect that different distributions are possible, which might provide a positive amount of information $H[K] > 0$ on the analyzed set of promoter sequences.

3.2.1 Results and discussion: positional entropy of the entire promoter set

We have computed $H[w]$ and $H[K]$ as a function of the position l along the promoter sequences and for several values of n (word length). In order to increase the statistics available, we analyze the sequences in the binary coding that originates naturally from the weak-strong classification of the nucleotides. C and G form 3 hydrogen bonds between the DNA strands and are classified as strong bases, while A and T form only 2 hydrogen bonds and are classified as weak bases. Thus, this coding allows to improve the statistics with a limited loss of information since part of the biological significance of the code is maintained. In practice, the binary coding allows to extend the analysis up to $n = 12$, since the number of possible words, 2^n, remains significantly smaller than the number of promoters M, avoiding intrinsic undersampling. For each position l along the promoter we take all the words $w^n(l)$ of length n that appear in all the promoters at position l. Then we compute the entropies $H[w]$ and $H[K]$ of the sample $\{w_1^n(l); \cdots ; w_M^n(l)\}$.

The results of this analysis are reported in Figure 3.1, 3.2 and 3.3. The trend of $H[w]$ is to decrease when the position l approaches the TSS, denoted with $l = 1000$ (in Figure 3.1 we report, as an example, the data for

$n = 10$): this is a signature of a smaller variability across promoters (for a detailed discussion of the meaning of the peak at about 30 bases upstream the TSS see next paragraph, especially the discussion about the results on C1). Moreover, the increase of $H[K]$ approaching the TSS observed for $n \geq 8$ tells us that this decrease in variability along the sequence corresponds also to the prevalence of some words with respect to the others. This trend can be attributed to the GC enrichment observed in the whole sample of human promoters. We can argue that such regular features should emerge from the assembly of specific sub-sequences that carry important biological information, which is conserved, as as a constraint to randomness, across several promoter sequences. For a more specific discussion on this topic see also the next paragraph, where the classification in clusters allows a more specific characterization of the causes of such variability features.

The comparison of the results of $H[w]$ for different values of n (Figure 3.2) shows an increasing variability for larger n. This is trivially due to the higher number of possible words. On the other hand, $H[K]$ shows a non-trivial trend since it grows up to its maximum in $n = 8$ and then it decreases again (see Figure 3.2 and 3.3). $H[K]$ reveals that the the sample of sequences with $n = 8$ is the most informative sample since it maximizes the occurrence variability. This suggest that sequences of length $n = 8$ are important and selected. The sequences of this length are very important in biology since this is the typical length of the TF binding sites. Our analysis gives a rigorous validation to this heuristic consideration. TF binding sites need a high information content for the specificity of their signal. The high variability in the frequency of sequences of length $n = 8$ may account for the difference between TF binding sites (highly underexpressed) from other sequences that must have a very different structure in order not to be confused with TFs and are scattered all over the promoter length (thus are highly expressed). For further discussion about this topic, see paragraph 3.2.3.

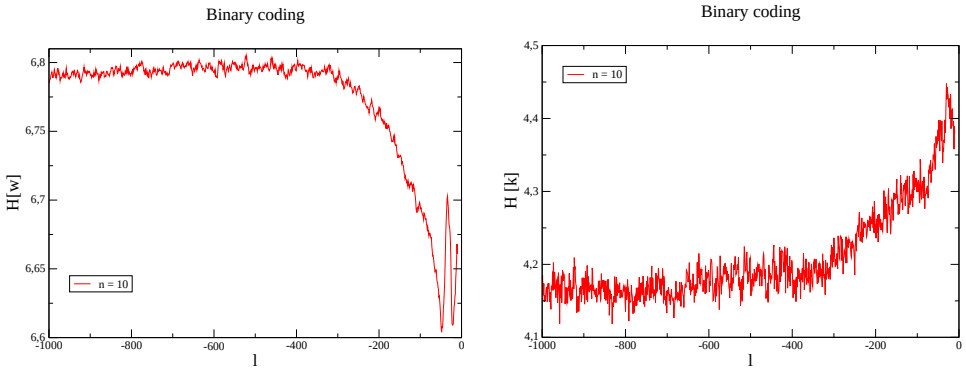


Figure 3.1 $H[w]$ (left) and $H[K]$ (right) as a function of the position l along the promoter for word length $n = 10$. $l = 0$ corresponds to the TSS.

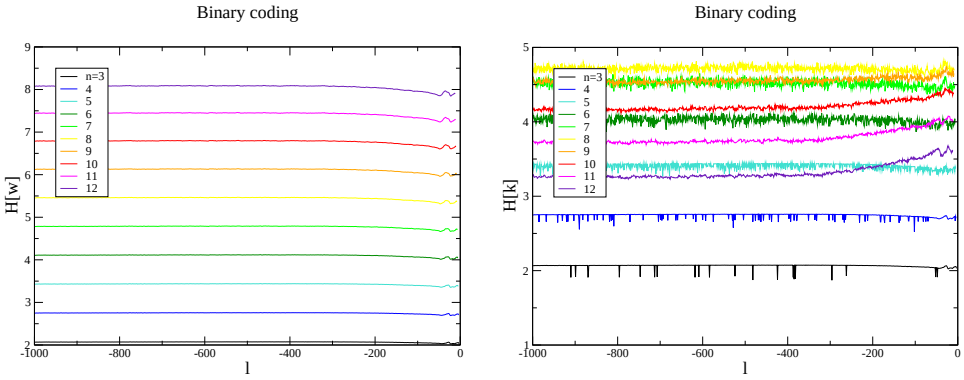


Figure 3.2 $H[w]$ (left) and $H[K]$ (right) as a function of the position l along the promoter ($l = 0$ corresponds to the TSS), for different values of n=word length.

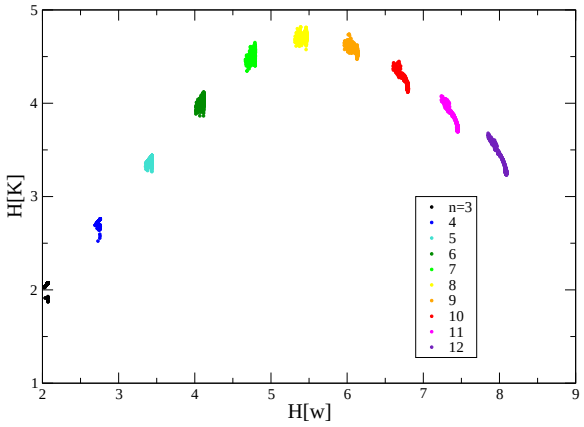


Figure 3.3 $H[K]$ as a function of $H[w]$ for several values of n=word length.

3.2.2 Results and discussion: positional entropy of the clusters

In the previous section the entropic analysis was applied to the whole set of promoters. In this section we want to apply the same entropic analysis to each cluster of promoters. We refer here to the clusters identified with the same method described in Chapter 2 and in [27]. In this case the clustering algorithm has been applied to the entire database in order to classify all the 34170 promoters. This allows to obtain a cluster size sufficient to avoid undersampling: we obtain 12605 promoters in C1, 5156 in C2, 4725 in C3, 11684 in C4. The larger size of the clusters allows to obtain a less noisy BCA, as shown in Figure 3.4.

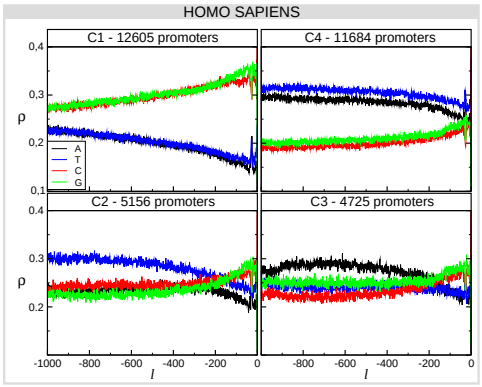


Figure 3.4 BCA of each of the clusters obtained with the clustering algorithm for the entire set of promoters of *H. sapiens*. We report the frequency ρ of each of the four nucleotides A (black), T (blue), C (red) and G (green) as a function of the position l along the promoter (0 corresponds to the TSS).

Discussion: $H[w]$

It is interesting to note how the behavior of $H[w]$ and $H[K]$ highlight different structural features of the four clusters (see Figure 3.5). In agreement with what has been observed for the analysis of the entire sample, also in the clusters $H[w]$ grows for larger n, and also in this case we can argue that it is due to the higher number of possible words, since the maximum value of $H[w]$ is $\log N$, where N is the number of possible words. Nevertheless, the behavior of $H[w]$ as a function of the position along the promoter is different in the different clusters. The most striking difference is between C1 and C4. In C1 we observe a decrease of $H[w]$ in the direction of the TSS, with a clear peak at about 30 bases from the TSS. The trend of $H[w]$ is quite the opposite for C4, where there is a slight increase of $H[w]$ in the direction of the TSS, with a dip at about 30 bases upstream the TSS (see

Figure 3.5 and also Figure 3.6 for a clearer comparison between clusters). This was already observed in [10] for the groups of TATA-less and TATA promoters, that are analogous to our C1 and C4 respectively.

The decrease of $H[w]$ in **C1** shows that the variability among promoters in this cluster decreases in the region close to the TSS. This trend is analogous to what we observe in the whole sample of human promoters. We believe that the trend observed in the whole sample is due to the contribution of the promoters of C1, that are a significant fraction (about 37% of the sample) and, when we average on the whole sample, they mask the features of the other clusters both on the BCA and on the entropic analysis. The decrease of $H[w]$ in C1 can be attributed to the GC enrichment (see also BCA of this cluster in Figure 3.4), that reduces variability across promoters in the region close to the TSS. We argue that such smaller variability points out that this region is subject to some constraints, meaning that specific biological information is stored in this promoter region, and that such information is shared across promoters in C1. For instance, we expect that such conservation in proximity of the TSS is related to sequence elements necessary for a correct transcription initiation. Anyway, we will see in the next paragraph that the analysis of $H[K]$ gives further insight on the features of this region in C1.

An interesting result is the presence in C1 of the peak of $H[w]$ at about 30 bases upstream the TSS (see Figure 3.5). An analogous result was observed in [10] about the set of TATA-less promoters. This peak indicates a more pronounced variability of words in this region of the promoter than in the GC-rich surroundings, and is confirmed by the peak in A and T densities that we observe 30 bases upstream the TSS in the BCA of C1 (see Figure 3.4). Such peak of A and T bases at that position re-establishes similar W/S nucleotide densities, causing the peak in $H[w]$. This result is quite puzzling: we do not expect in C1 the presence of AT-reach motifs in the typical position of the TATA box (that is a sequence typically composed of weak bases), since C1 contains a large majority of TATA-less promoters. It is interesting to note from our results that the AT enrichment in this region emerges as a very widespread feature common to many promoters of this cluster, and it is not a feature of some particular promoters. We hypothesize that even if such motifs are not recognized as TATA boxes (at least according to the definition here adopted), they might correspond to the TATA-like motifs, that have been recently found to be associated with functional roles in yeasts [69, 70]. It is unlikely that the TATA-like motifs are just unrecognized TATA-boxes, even if this could be possible for some of them. Actually, it has been observed that promoters with a TATA-like motifs are not simply a slight variation of the TATA promoters, since their regulation mechanism and expression patterns have been proved to be different - at least in yeast [69, 71–74]. Another hypothesis about the role of such AT enrichment can be related to DNA dynamics. In fact, there is

an interest in DNA dynamics especially in breathing dynamics and bubble formation in promoters. This interest is justified since, in order to allow transcription initiation, the two DNA strands must separate, i.e. they form a so called bubble. Some results show that the region around the TSS is prone to form stable bubbles especially in CG rich promoters [75]. Other studies highlight how structural and dynamic properties are encoded in the promoter sequences and are a distinguishing feature of promoter sequences near the TSS, and that besides transcription factor binding, DNA dynamics and spontaneous bubble opening is a necessary conditions for transcription initiation [76, 77]. An interesting development of the results found here about the promoters in C1 can be the investigation on how the AT peak can condition the dynamical properties of the DNA in that region.

An opposite behavior with respect to C1 is observed in **C4**, a cluster with a high content of TATA promoters (see Figure 3.5 and also Figure 3.6 for a direct comparison between C1 and C4). In this cluster we observe a slight increase of $H[w]$ in the direction of the TSS. This suggests a higher variability among promoters in this region, that may be related to the fact that these promoters are associated to tissue specific genes, because even if they belong to the same cluster, they are involved in many different and specific regulatory patterns. So, it is not surprising that they exhibit different structures and probably different binding sites. At the same time, a dip of $H[w]$ is observed around 30 bases upstream the TSS, according to what is reported in [10] about TATA promoters. This dip is a consequence of the presence of the TATA box in that position. Thus, all promoters in this class are very similar to each other in this very small region, and this explains the dip in $H[w]$.

Altogether, the results about $H[w]$ in C1 and C4 suggest that some biological constraints, favoring the transcription process, determine a relative abundance of AT bases in the postion 30 bases upstream the TSS in C1 and C4 promoter subsets. This yields an increase of redundancy in promoters of C4 and of variability in promoters of C1. This testifies the importance of this region associated with fundamental functional properties in both C1 and C4.

The trend of $H[w]$ in **C2** and **C3** is not as clear as in C1 and C4 and does not allow similar conjectures (see Figure 3.5). Anyway, it is coherent with the Base Composition properties of these clusters (see Figure 3.4)). It is possible that the growth and the subsequent drop of $H[w]$ is due to a small fraction of promoters with a GC growth gradient in a very small region near the TSS. Such promoters have a structure similar to those of C1 but also have a high transposon content in the rest of the sequence, and are classified in C2 and C3 by our algorithm. This can be also the reason of the peak at 30 bases from the TSS we observe also here in C2 and C3 (see Figure 3.5 and Figure 3.6). We hypothesize that promoters in C2 and C3 are C1-like promoters that have been modified by the insertion

of transposons. This hypothesis is supported by the observation that Alu transposons (that are overabundant in C2 and C3) preferentially insert near housekeeping genes [78].

Discussion: $H[K]$

The results obtained so far about $H[w]$ confirm what was yet observed in [10], even if here the analysis has been specifically performed for the four clusters. The results about $H[K]$ allow a development and a further deepening of the analysis started with $H[w]$.

As already observed in the analysis of the whole sample, $H[K]$ grows when n grows in all clusters until $n = 7$ (C2, C3) or $n = 8$ (C1, C4) (see Figure 3.5 and Figure 3.8).

In **C1** $H[K]$ has a very interesting trend. We can see in Figure 3.5 that the behavior of $H[K]$ is quite different according to the word length n. If $n < 8$ we notice that $H[K]$ decreases in the direction of the TSS, while it grows for $n > 8$. We remind here that $H[K]$ is maximum when every word has a different occurrence from the others. $H[K]$ tells us that short words tend to appear more or less with the same occurrence frequency near the TSS, while there is a selection on longer words, since some of them are more rare and some more frequent, according to the growth of $H[K]$. Hence, short and long words exhibit a different trend of $H[K]$. On the other hand, as observed in the previous paragraph, $H[w]$ decreases in the direction of the TSS for both short and long words, indicating a smaller variability in that part of the promoter. On short words this smaller variability causes a drop also in $H[K]$. For longer words it is possible that the trend of $H[K]$ inverts because there are many different possible words but only some of them are found in that region due to the small variability. Indeed, we have verified that, for a given position along the sequence, all of the possible 2^n words with $n \leq 8$ are found at least once in all the sample (data not shown). This is not true for $n > 8$: in this case we observe that there is a fraction of all the possible words that is not present in the region near the TSS, with an exception near 30 bases upstream the TSS where the number of distinct words found has a peak (see Figure 3.7). $H[K]$ grows for long words because some words become very rare or do not appear at all while others are still frequent. Note that the number of possible words is at most 2^{12}, which is definitely smaller than the number of promoters: thus, words are really selected and this is not a problem of undersampling.

Overall, the results obtained for $H[K]$ highlight a very interesting and non-trivial feature of C1: the selection in this region of promoter sequences appears to act on long words. Short words are more or less equally present but they combine to form longer words with a non-trivial rule, avoiding to form some long words that are suppressed.

At the same time, also the behavior of $H[K]$ at 30 bases upstream the TSS is quite intriguing. The peak in $H[w]$ testifies a high variability among promoters in this position, i.e. in this position almost all possible words are found. Yet, $H[K]$ tells us that for $n < 8$ there is a selection on this words since some are more frequent than others (peak of $H[K]$). On the other hand, for $n > 10$ the peak turns into a dip, showing that some of the possible words appear with the same occurrence frequency. This is related to the peak we observe in the number of distinct words found shown in Figure 3.7: that peak suggests that some of the long words that were suppressed in the region near the TSS actually are back in this region at -30 bases. It appears that at that position the selection tends to act on words shorter than $n = 8$, consistently with the length of that "TATA-like" region that is about 8 bases [69].

A quite different scenario emerges for **C4**, where a peak in $H[K]$ clearly emerges 30 bases upstream the TSS for $n \geq 8$, while the overall trend shows a decrease in $H[K]$ in the direction of the TSS. This suggests that there is a strong selection on the words that appear 30 bases upstream the TSS, and this is not surprising since that is the position of the TATA box. Nevertheless, it is quite puzzling to note that the peak is not present for length $n < 8$, considering that the length of the TATA box is 8 bases. If we compare $H[K]$ in that position for C1 and C4 we see a quite opposite behavior, even if in both cases we observe an AT enrichment at that position. This might be a clue about different features of the TATA-box in C4 and the TATA-like region in C1.

For **C2** and **C3**, the same considerations reported for $H[w]$ apply for $H[K]$. Since the main feature of these clusters is the high transposon content, the fact that $H[w]$ and $H[K]$ do not show a clear trend as a function of the position along the promoter suggests that the transposon insertion has not privileged a specific location in the promoter, otherwise we would observe regions of low variability in a specific position.

It is interesting to note that in all clusters $H[K]$ has a more pronounced trend as a function of position for $n > 8$.

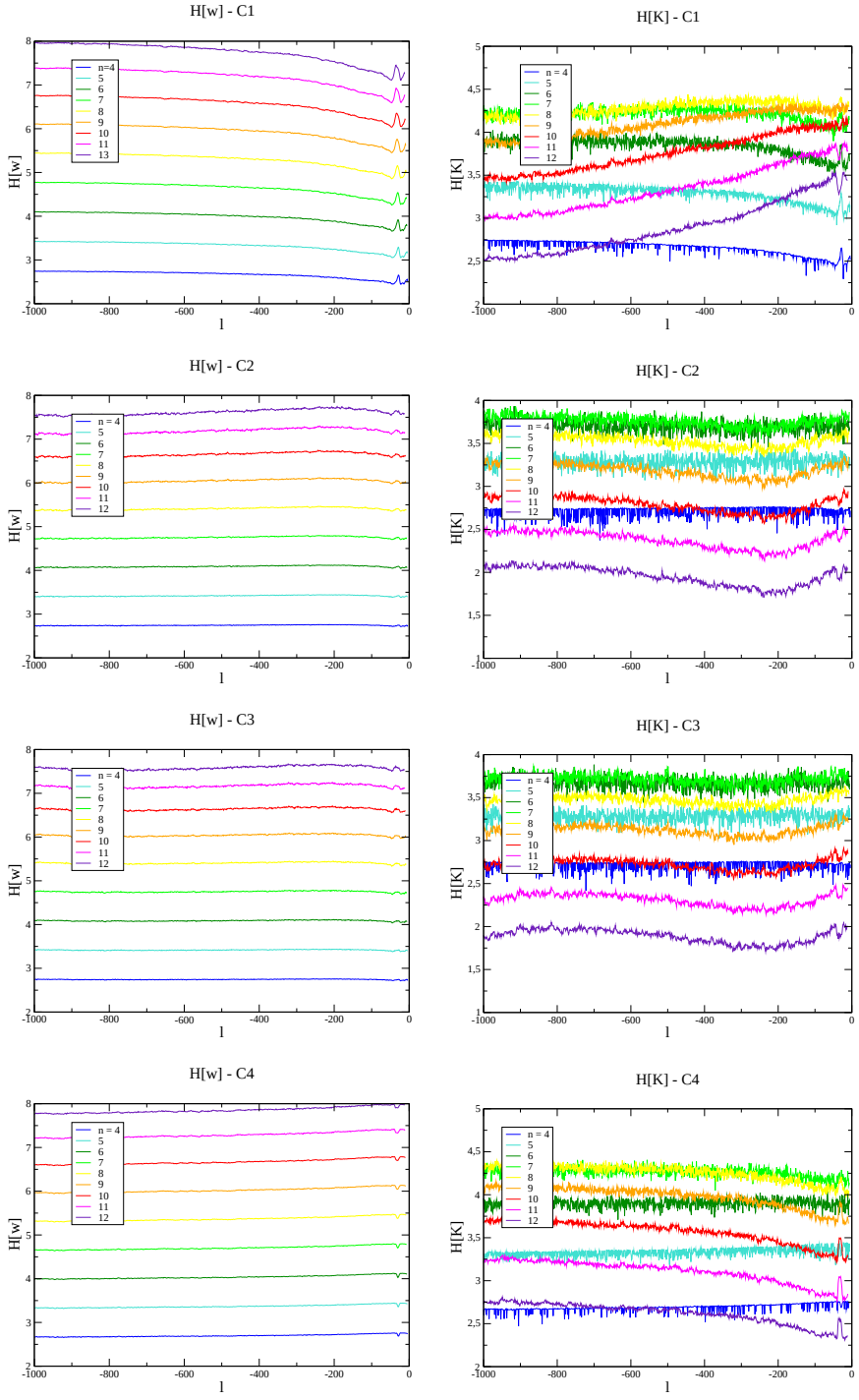


Figure 3.5 $H[w]$ (left column) e $H[k]$ (right column) as a function of position l ($l = 0$ corresponds to the TSS), for different values of n. Data refer to the four clusters obtained with the method described in [27] and in Chapter 2. The promoters have been analyzed in the binary coding of weak (A,T) and strong (C,G) bases.

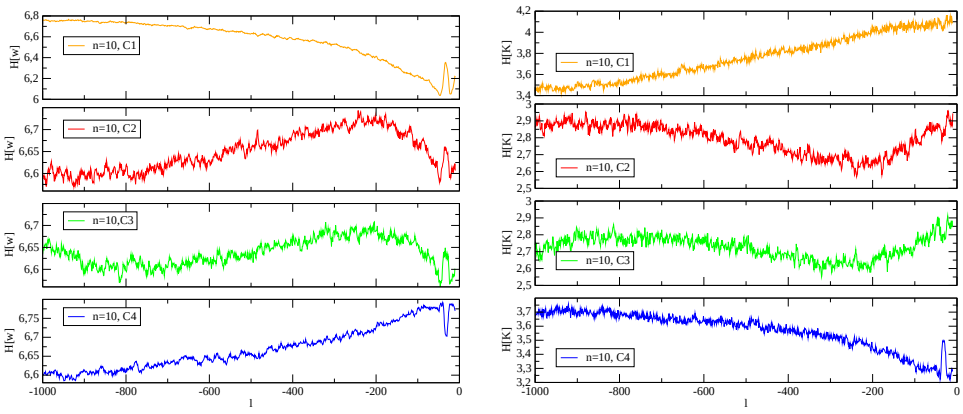


Figure 3.6 $H[w]$ (left) and $H[K]$ (right) as a function of position l ($l = 0$ corresponds to the TSS). Data refer to $n = 10$. Comparison between clusters.

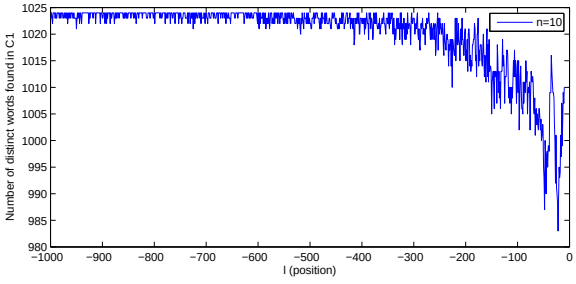


Figure 3.7 Number of distinct words found at least once in the promoters of C1 as a function of the position along the promoter sequence. Data refer to $n = 10$. The same graph for $n \leq 8$ is flat to the value of 2^n (not shown).

3.2.3 Supplementary discussion

In this paragraph we report some supplementary discussion about the data. The supplementary discussion presented in this paragraph arises from further analysis performed by E. Calistri. In Figure 3.8 we report $H[K]$ as a function of $H[w]$ for the different values of n. The data reported in this figure are the same data of Figure 3.5, but plotted in a different way. For each n we have a cloud of points corresponding to the different positions along the promoter. We report the data in this form because it is easier to notice that there is a maximum of $H[K]$ corresponding to $n = 7, 8$. This has been noticed also in paragraph 3.2.1. Here we want to remark again that the data highlight that the sample of words of length $n = 7, 8$ is the sample with the maximum occurrence variability. In fact, this is the typical length of TF binding sites, and it is likely that the sequences of this length are highly subject to selection mechanisms, explaining the occurrence variability testified by $H[K]$.

In order to better understand the meaning of this data, we have reported in Figure 3.9 the histograms of m_k as a function of k for several word lengths. m_k and k are specified as in the previous paragraph in the calculations of $H[w]$ and $H[K]$, i.e. k is an occurrence and m_k is the number of words with occurrence k. However, k and m_k are computed for words in 4-bases coding searched in the whole promoter sequence. Hence, the data in Figure 3.9 refer to a slightly different analysis with respect to Figure 3.8, where we had a positional analysis in a binary coding. Nevertheless, the results in Figure 3.9 reveal very peculiar properties of the distribution of m_k as a function of k: in the case $n = 8$ the histograms referring to C2, C3, C4 and to the whole sample clearly show a double peak structure that is much sharper than in the data referring to $n = 4$ and $n = 10$ (but a double peak is clearly visible in C4, $n = 4$).

We have focused our analysis on C4: the histogram of $n = 8$ identifies two groups of words corresponding to the two peaks. In order to characterize the features discriminating these two groups of words, we have performed a dinucleotide analysis of the words. We have found that in the words corresponding to the second peak ($k = 180 - 185$) the CpG dinucleotide[2] is almost absent. CpG is absent also in very frequent words (i.e. those with $k > 500$). On the other hand, there is a higher CpG content in the less frequent words corresponding to the first peak ($k = 25 - 30$). It is worth to point out here that CpG dinucleotide occurs with a lower frequency in vertebrate genomes than would be expected due to random chance. Such CpG suppression is probably due to due to mutational biases (methylation). Nevertheless, it has been observed that promoters are relatively enriched in CpG dinucleotides with respect to the rest of the genome [58, 59]. The

[2]CpG stands for cytosine and guanine separated by a phosphate, which links the two bases together in a DNA single strand, i.e. it represents C followed by G in a DNA strand.

analysis of C4 shows that CpG enrichment is not widespread in all promoters since in C4 seems to be underexpressed. This does not mean that those underexpressed words may play an important functional role.

Note that no double peak structure is observed in C1, and we can argue that in C1 the words containing CpG are not selected. This is not surprising, considering the high CG content of C1. Nevertheless, we have observed that in C1 CpG dinucleotides increase towards the TSS with the same rate of the other three dinucleotides, i.e. GpC, CpC and GpG, but CpG remain still relatively underexpressed (see Figure 2.14). In particular, all the four dinucleotides provide comparable contributions to the increase in GC content close to the TSS. The absence of indications in favor of the selection on CpG densities suggests that the CpG increase at promoter level may be the result of the GC enrichment and not viceversa. In our opinion, the actual role of CpG dinucleotide in promoters (especially in C1) remains an open problem.

Overall, this analysis clearly discloses interesting properties on the selection mechanisms on words, while at the same time leaves several open questions.

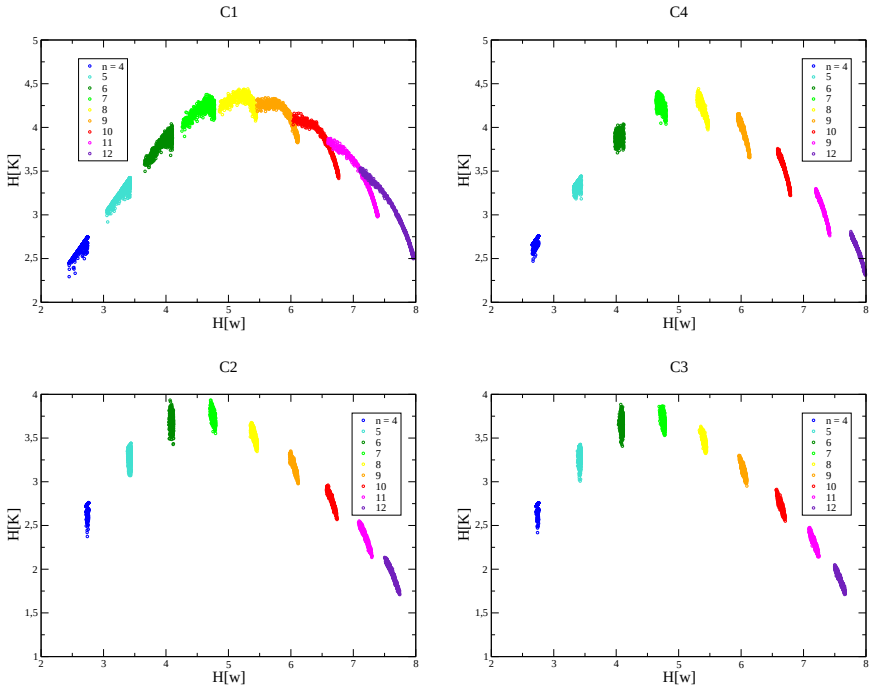
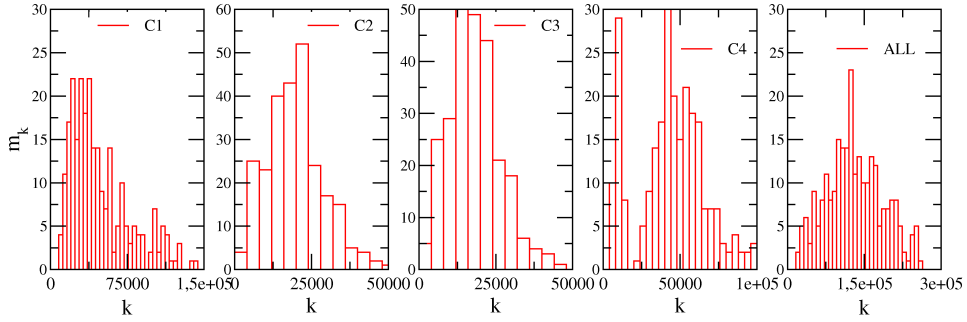
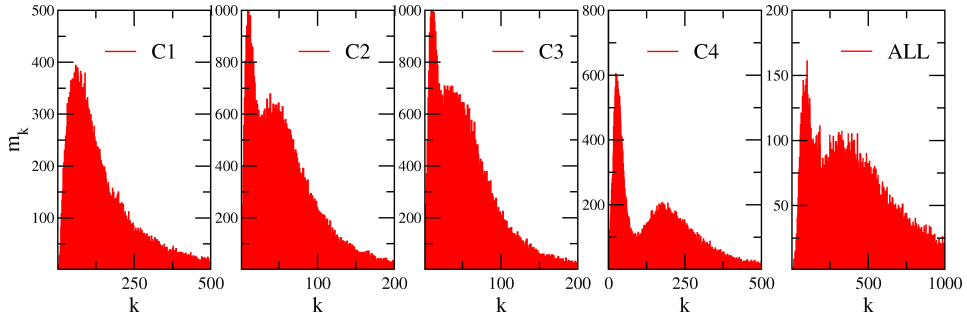


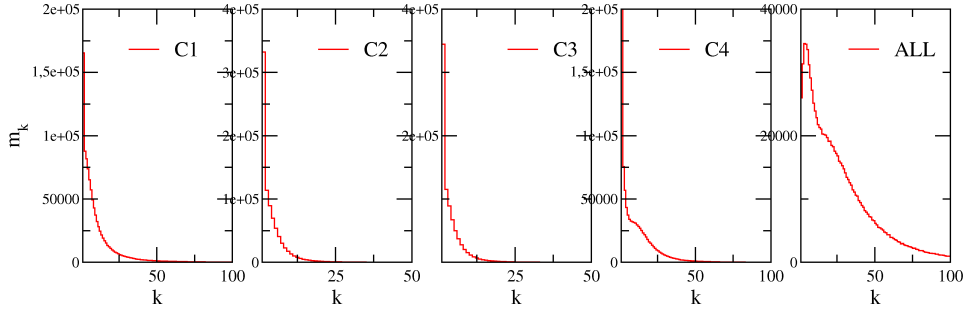
Figure 3.8 $H[w]$ as a function of $H[k]$. Data refer to the four clusters obtained of Figure 3.4. The promoters have been analyzed in the binary coding of weak (A,T) and strong (C,G) bases.



(a) n=4



(b) n=8



(c) n=10

Figure 3.9 Histograms of m_k as a function of k for $n = 4, 8, 10$. The data refer to words in the ATCG coding searched in the entire promoter sequence.

3.3 Keywords analysis

In this section, we show the results obtained applying a keywords identification technique to the set of human promoters, treating them as strings of text. This is not a positional analysis as the one presented in the previous section, since here we take into account the promoter sequence as a whole and we try to find small sequences that characterize it, independently of their position with respect of the TSS.

The choice and frequency of the words employed to build a text is constrained by syntax and semantics. Typically, the frequency distribution in a text is highly peaked on relatively few words. In [67] $H[w]$ and $H[K]$ are employed to identify the keywords in a specific text, namely Darwin's *On The Origin of Species*. The method consists in dividing the text in M sections of length L. Then, for each word w one computes K_w^i (the number of times the word appears in the i-th section) and $m_{K_w}^i$, in order to evaluate $H[K]$ as a function of $H[w]$ for different values of L. The keywords of the text will be found among the most frequent ones, but one has to distinguish generic words (such as "and" and "that") from really significant words. The main idea is that generic words are uniformly spread across the text, therefore their occurrence does not vary much across the samples in which the text has been split; on the other hand, the distribution of the keywords follows the constraints of the complex design and meaning of the text. Thus we expect that among the most frequent words the keywords will have higher values of $H[K]$. For instance, the graph of $H[K]$ as a function of $H[w]$ in Figure 7 of [67] shows how keywords like 'generation', 'seed', 'bird' have a higher peak of $H[K]$ allowing to distinguish them from other generic words.

Here we want to employ this method to identify the keywords in the text made by the set of the human promoters. In this case there is a natural division in sections of length $L = 1000$ nucleotides. When we let L vary, we consider $L \leq 1000$, otherwise the sections would glue together parts of different promoters that are not sequential in the DNA strand, thus introducing spurious effects that are not inherent in the DNA structure. In particular, we have analyzed the cases $L = 10, 20, 50, 100, 200, 500, 1000$. We fix a word length n and we compute K_w^i and $m_{K_w}^i$ for each of the 2^n (in the binary case) or 4^n (in the ATCG alphabet case) words w of length n. Each word is characterized by a curve of $H[K]$ as a function of $H[w]$ (each point of the curve, labeled by index i, corresponds to a different value of L). The keywords will be characterized by the highest values of $H[k]$ and are in fact identified computing the area under the curve $H[K]$ vs. $H[w]$ and taking the words with maximum area.

In order to avoid any confusion, it is useful to point out that in the previous paragraph $H[w]$ and $H[K]$ depended on l, since they were computed from the frequency of all possible words w in a specific position along the promoters; here they depend on L since they involve the frequency of a

word w in the different sections in which our text was split. As we have just explained, a curve of $H[K]$ as a function of $H[w]$ is associated to each word. Every point of the curve is obtained for a different value of L.

Considering the definition and the identification method of the keywords, with the method presented here we expect to deepen the analysis performed on the positional entropy. The keywords are relatively frequent words that are not scattered uniformly along the promoter sequence but tend to concentrate in specific regions. Thus, even if the analysis presented here is different from the positional analysis of the previous section, it can help to shed more light on the results obtained with the positional entropies, especially on which words characterize the low variability regions.

The keywords analysis has been applied to the whole promoter set (results reported in sec. 3.3.1) and also to the clusters identified in the previous chapter (results reported in sec. 3.3.2)

3.3.1 Results and Discussion: Keywords analysis of the entire promoter set

We have applied the keyword identification procedure to the whole promoter set. The results in the binary case, performed for $n = 4, 8, 10$ identify as keywords the homogeneous sequences reported in Table 3.1. This could be due to the method used to search words. While a real text is naturally split in words, in promoters words are searched shifting in the text of one nucleotide at a time, so that there is an overlap between the words found. Long homogeneous strings in promoters cause an overestimation of the frequency of the homogeneous words, that appear to concentrate in specific portions of the text and are thus identified as keywords. Nevertheless, beyond this *caveat*, the results are not trivial at all, since they suggest that long homogeneous tracts are quite significant in promoters: this enforces the conjecture, often reported in the specialized literature, that assigns those tracts an important role in facilitating the the TFs in retrieving the corresponding binding site [14–18], as discussed in paragraph 1.2.2 in Chapter 1.

In order to obtain more significant results, the analysis has been repeated in the natural ATCG alphabet. In Table 3.2 we have reported the 12 keywords with larger area. Among the keywords of length $n = 4$ we find again homogeneous sequences, and this can be due to the same mechanism mentioned for the binary case. We also identify several keywords made only of strong (C,G) nucleotides. The dinucleotide CpG appears in several of the identified keywords: as already stated in par. 3.2.3, it is supposed to play an important role in the regulatory functions of promoters, since it is suppressed everywhere in the genome except for some promoter regions (called CpG islands) [58, 59]. On the other hand, the real significance of CpG islands is still quite a debated problem in the literature [60].

Concerning the keywords of length $n = 8$ the results are quite different

from what we find for $n = 4$. In this case homogeneous sequences are made of weak nucleotides A and T, only. It can be argued that in this case there has been an overestimation of the number of homogeneous sequences due to the fact that the words are searched shifting of one nucleotide at a time. Anyway, it is a matter of fact that homogeneous sequences of C or G are not found, and this highlights an asymmetry of nucleotide distributions in promoters. This overestimation may be also the cause of the identification of keywords made of an alternation of two different nucleotides, but also in this case only some of the possible patterns are identified (notice also that GTGTGTGT is the reverse complement of ACACACAC). This kind of sequences can be related to DNA elements called *micro-* and *mini-satellites*, but this will be discussed in more details in the next paragraph.

The keywords of length $n = 10$ show a more interesting and nontrivial structure. They have a non-homogeneous and non-periodic structure, and as highlighted by the color boxes, and some are the same word, just shifted, as if that word belonged to a broader structure made by the repetition of this non-trivial pattern.

Keywords - binary

n=4	n=8	n=10
0000	00000000	0000000000
1111	00000001	0000000001
	00000010	0000000010
	00000100	0000000100
	00001000	0000001000
	00010000	0000010000
	00100000	0000100000
	01000000	0001000000
	10000000	0010000000
	01111111	0100000000
	10111111	1000000000
	11011111	0111111111
	11101111	1011111111
	11110111	1101111111
	11111011	1110111111
	11111101	1111011111
	11111110	1111101111
	11111111	1111110111
		1111111011
		1111111101
		1111111110
		1111111111

Table 3.1 Keywords found in the binary code. The binary code is 0 for A,T and 1 for C,G.

Keywords - atcg

n=4		n=8		n=10	
word	area	word	area	word	area
tttt	5.67	gtgtgtgt	1.75	acctccataa	2.37
aaaa	5.15	tgtgtgtg	1.66	cctccataac	2.37
cccc	4.54	acacacac	1.63	tggtctggtc	2.20
gggg	4.51	cacacaca	1.54	ggtctggtct	2.17
gcgg	4.39	tatatata	1.04	gtgaactgcg	2.15
gggc	4.31	atatatat	1.01	tgtcacaccg	2.11
ccgc	4.28	tttttttt	0.79	agacggtggt	2.08
gccc	4.23	aaaaaaaa	0.74	cggtggtgaa	2.07
ggcg	4.21	ggcggcgg	0.42	acggtggtga	2.01
cgcc	4.11	gcggcggc	0.37	ccataacctc	2.00
ctcc	4.02	gatggatg	0.36	gtctggtctg	1.97
ggag	4.03	ctctctct	0.34	cgtggctcgc	1.96

Table 3.2 Keywords found in the atcg code.

3.3.2 Results and Discussion: Keywords analysis of the clusters

In this section we want to deepen the cluster characterization by identifying the distinctive keywords of each cluster. The analysis has been performed for word length $n = 4, 8, 10$. The results are reported in Table 3.3 where we report the 12 keywords with larger area for each cluster.

The four cluster exhibit different features in their keywords even for very short words ($n = 4$). On such short sequences it is hard to identify interesting patterns, nevertheless we can argue that the keywords of this length follow the base composition properties of the corresponding cluster. In C1 we find keywords composed entirely of strong bases while in C4 weak bases prevail. Many keywords display the CpG dinucleotide.

With $n = 8$ interesting patterns start to emerge. In all the clusters we find keywords with periodic composition, made by the repetition of a dinucleotide (namely, AC, TG, TA, with the remark of the complementarity of AC and TG). Such periodic keywords are highlighted in gray. It is interesting to note that, as already stated, our method of keyword identification can overestimate the contribution of homogeneous or periodic sequences; nevertheless, only this three kinds of periodic keywords are found (while homogeneous sequences are not present, apart from one exception). Such periodic sequences are probably related to DNA micro- and mini-satellites.

Microsatellites are DNA tracts characterized by several iterations of (indicatively) 1–6 bp nucleotide motifs. They are widespread in all organisms, but it is an open controversy whether they are evolutionarily neutral elements or have a functional role. Their origin relies mainly in two mutational mechanisms: DNA slippage during DNA replication and recombination between DNA strands [79]. Their abundance is associated with their very high mutation rate, manifested as changes in the number of motif repeats. Microsatellites are widespread across the genome, even if they are relatively rare in protein-coding regions. Accordingly, there is great abundance of hypotheses about the functional role of microsatellite DNA. We focus here on their role in gene regulation activity. Several studies have highlighted that microsatellites located in promoter regions may affect gene regulation [79–82], for instance it has been observed that the deletion of a repetitive tract from the promoter region can reduce its transcriptional activity. It is interesting to note that, in many cases, the composition and structure of the microsatellites identified in such studies correspond to the keywords identified with our method. For example TG repeats close to the promoter sequence can effectively enhance transcription [83]. Other works [84–86] report that CA/TG repeats and their repeats number are in some cases pivotal for the transcription activity . Moreover, some poly(GA)- and poly(GT)- sequences found in promoter sequences serve as binding sites for a variety of regulatory proteins [87]. An important point highlighted by

several studies is that the repeats number has a effect on the functional role of microsatellites: the high variability rate among individuals of the length of AC/GT repeats has been linked with variations in human phenotypes and also to disease onset [80]. AC/GT repeats may also act on gene regulation by influencing DNA structure forming Z-DNA (left-handed spin double helix) [88].

Minisatellites are usually defined as the repetition in tandem of a short (6- to 100-bp) motif spanning 0.5 kb to several kilobases. [89]. In addition to the fact that they represent genome hypervariable regions and are thus used in fingerprinting studies, some of them have been considered in specific association studies, finding out that they are associated with an increased risk of acquiring a pathological condition.

Regarding the other more complex keywords, in many cases we find that different keywords are the same sequence, just shifted, pointing out that the keywords are probably originated by a longer, repetitive, yet complex structure.

The same considerations apply to the data with $n = 10$. In this case the repetitive keywords are almost absent, giving way to more complex and repetitive structures, that in some cases is the same of some keywords with $n = 8$. We have tried to highlight with the same color the keywords sharing the same structure inside each cluster. Further analyses, performed by E. Calistri, allowed to determine that such keywords belong to large structures in some promoters made by the repetition of a nontrivial pattern. Such structures may be micro- or minisatellites again. Moreover, some of them may fall in the category of G-quadruplexes, whose functional role within regulatory regions is well established (see for example the recent review on G-quadruplexes and their emerging role in neurodegenerative diseases [90]).

Keywords - atcg

n=4

C1		C2		C3		C4	
word	area	word	area	word	area	word	area
cccc	5.49	tttt	6.50	tttt	4.42	tttt	5.90
gggg	5.43	attt	3.94	gagg	3.95	attt	4.52
gcgg	5.14	cctc	3.88	ggag	3.91	aaat	4.21
gggc	5.07	ctcc	3.83	aaat	3.77	ttta	4.16
ccgc	5.02	ctgg	3.63	cctg	3.62	taaa	3.83
gccc	5.00	ttta	3.56	ccag	3.52	ttct	3.81
ggcg	4.96	cagg	3.53	tggg	3.48	tatt	3.75
cgcc	4.83	ttct	3.48	agaa	3.40	tttc	3.74
cggg	4.65	tttc	3.46	taaa	3.35	atat	3.72
cccg	4.63	cctg	3.45	cagg	3.35	cttt	3.71
ggag	4.63	tttg	3.43	ctgg	3.34	tttg	3.65
ctcc	4.62	gcct	3.39	gaga	3.34	ttaa	3.61

n=8

C1		C2		C3		C4	
word	area	word	area	word	area	word	area
acacacac	1.92	catccatc	1.94	acacacac	2.18	gtgtgtgt	1.76
cacacaca	1.83	ccatccat	1.72	cacacaca	2.01	tgtgtgtg	1.69
gtgtgtgt	1.82	gtgtgtgt	1.70	gtgtgtgt	1.10	acacacac	1.09
tgtgtgtg	1.74	tgtgtgtg	1.67	atatatat	0.93	cacacaca	1.00
gatggatg	0.88	tatatata	1.53	tgtgtgtg	0.91	taacctcc	0.86
tatatata	0.86	atatatat	1.42	tatatata	0.90	atatatat	0.82
ggatggat	0.81	tttttttt	1.28	ggaaggaa	0.86	tatatata	0.81
aagacggt	0.80	acacacac	0.99	aaggaagg	0.72	cgttaggc	0.73
gtctggtc	0.75	cacacaca	0.93	aagaaaga	0.66	agtgcgat	0.59
atggatgg	0.72	atatacac	0.85	gaaggaag	0.64	gcgatgtc	0.56
ggtctggt	0.70	atccatcc	0.77	aaagaaag	0.59	tgactcgt	0.53
ctctctct	0.69	tccatcca	0.74	aggaagga	0.58	tactcgat	0.48

n=10

C1		C2		C3		C4	
word	area	word	area	word	area	word	area
ggtctggtct	3.17	tccatccatc	3.29	gtggctttgt	1.79	cctccataac	2.84
tggtctggtc	2.96	catccatcca	2.18	ggtgtgaggt	1.76	acctccataa	2.69
gtctggtctg	2.94	atccatccat	1.80	tggctttgtc	1.75	acctccctaa	2.59
ttggtctggt	2.75	atatatacac	1.69	cggtgtgagg	1.59	cctaacctcc	2.48
gtgaactgcg	2.61	tatatacaca	1.67	gctttgtctt	1.55	ccataacctc	2.45
agacggtggt	2.42	ccatccatcc	1.58	acacacacac	1.53	taacctccat	2.42
gtggtgaact	2.42	tgggataatc	1.58	ggctttgtct	1.47	cataacctcc	2.42
tgaagacggt	2.38	atgggataat	1.58	ctttgtcttc	1.39	ctccataacc	2.29
gacggtggtg	2.36	ggtaatggga	1.41	gtcttcctga	1.39	aacctccata	2.28
acggtggtga	2.35	ggataatcca	1.39	cacacacaca	1.28	tccctaacct	2.27
tggtgaactg	2.31	gtaatgggat	1.36	tgaggtgtgg	1.27	tccataacct	2.26
ggtgaactgc	2.29	taatgggata	1.34	tttgtcttcc	1.25	taacctccct	2.23

Table 3.3 Keywords found in the atcg code in each of the clusters.

3.4 Conclusion

In this chapter we have reported the results obtained employing two different entropic indicators, $H[w]$ and $H[K]$. Such indicators allowed to perform two different kind of analysis. The first is a positional analysis highlighting variability properties across promoters and thus identifying promoter regions subject to some constraint and selection. The second is a text mining technique that allows to identify the keywords in a text and that was applied to the promoter sequences treated as strings of text. The aim was to identify relevant small sequences (i.e. the keywords of the text) that are good candidate to play an important role in promoter functioning from a biological point of view.

The positional analysis allowed to highlight very different features of the clusters, especially between C1 and C4. In C1, $H[w]$ shows that the variability across promoters decreases near the TSS. $H[K]$ reveals that such decrease is not trivially due to the CG enrichment in this promoter region, but it is probably related to a constraint on the possible words appearing in this region: there is a selection mechanism that suppresses some words, and such suppression seem to act on longer words ($n > 8$), while $H[K]$ gives evidence that there is no such frequency variability among shorter words.

In C4, $H[w]$ and $H[K]$ reveal quite an opposite scenario with respect to C1. They both testify that there is an increase of sequence variability in the direction of the TSS (increase of $H[w]$), and such increase does not correspond to a increase of frequency variability, i.e. to a selection on possible words (decrease of $H[K]$). These features are in good agreement with the fact that the promoters in C4 mainly belong to tissue specific genes. This means that the promoters in this cluster are involved in a variety of very specific and different regulation mechanisms, and such variety is reflected in the variability among the promoters in this cluster near the TSS.

The entropy analysis also highlights the important role of the small region at -30 bases upstream the TSS: such region exhibit peculiar entropic properties in all the clusters.

Overall, the positional entropy analysis reveals that the differences between C1 and C4 go beyond the mere base composition properties, but regard the mechanisms that shaped the evolution of such promoters, and favored selection in C1 and variability in C4. Concerning C2 and C3, the entropic analysis suggests that the transposon insertion has not privileged a specific location in the promoter, otherwise we would observe regions of low variability in a specific position.

The keywords analysis gives different indications depending on the word length. For $n = 4$ the keywords identified reflect the base composition properties of the clusters. It is interesting to note that in the keywords of C1 we

often find the dinucleotide CpG, that is supposed to play an important role in gene regulation, since it is suppressed everywhere in the genome except in promoters. For $n = 8$ we do not find a great differentiation among the keywords in the different clusters: it is interesting to note that the keywords found, characterized by a peculiar structure formed by the repetition of the same dinucleotide (e.g. acacacac, tgtgtgtg, etc), are probably related to microsatellite DNA. Our method draws attention on this peculiar structures, signaling their putative importance in promoter functioning, in agreement with other studies that discussed the role of microsatellites in gene regulation. For $n = 10$, it is interesting to note that the keywords emerging from the analysis have a different and more complex structure with respect to those found for $n = 8$. Further analyses, performed by E. Calistri, allowed to determine that such keywords belong to large structures in some promoters made by the repetition of a nontrivial pattern. Such structures may be microsatellites again, or even more complex structures whose role is currently unknown.

Overall, the keywords analysis is able to shed light on interesting structures in promoters, that in many cases are probably related to functional elements. A detailed biological survey of the putative functional role of the keywords identified here goes beyond the scope of this thesis. Our aim here was to apply original methods to promoter studies, and show that non-trivial results emerge.

Chapter 4

Network

4.1 Introduction

In this section we present a further characterization of the clusters obtained in Chapter 2. In the work presented here we broaden the field of analysis beyond the mere nucleotide sequence of the promoter, taking into account biological information regarding interactions among promoters. To be more specific, we define an interaction between promoters and we analyze the network obtained. The pivot of the analysis presented here takes inspiration from previous works [91, 92], where notions of Random Matrix Theory (RMT) are applied to the study of a gene co-expression network. Nevertheless, this work is not just a mere repeat of an analysis that was yet performed by others: thanks to the cluster identification, we are able to isolate the properties of each subnetwork of the clusters, to highlight the differences between each other and with the entire network. The work presented here regards the analysis of two different network of promoters, defined in a completely different way.

In section 4.2 we give some very essential notions of RMT, focusing on the concepts that will be applied to our analysis and on the main idea about what we expect to deduce about our networks from RMT.

In section 4.3 we present the analysis of a network where specific biological information about genetic interaction is included in the network construction. The interactions between promoters are built starting from the information extracted from http://thebiogrid.org [93], a public database that archives genetic and protein interaction data. We highlight the different structural properties of the subnetworks of the different clusters. Moreover, in this case the tools of RMT allow to build a procedure that identifies "important" (in a sense that will be specified) nodes in the network, in order to extract biologically relevant information in the frame of gene prioritization problems.

In section 4.4 we take into account a completely different way of build-

ing the links between promoters: we build a similarity network. This is a weighted network where the links between promoters have a score depending on the similarity between promoter sequences. The adjacency matrix of this network is none other than the similarity matrix used for the clustering of promoters and is obtained as described in section 2.4.3. In this case the biological information introduced in defining the links does not regard the functional interactions between genes but just a similarity at a sequence level. We analyze this network with the tools of RMT in order to characterize the similarity relation between promoters inside each cluster.

4.2 Random Matrix Theory

Taking inspiration from previous works [91, 92], we employ the results of RMT to analyze the adjacency matrix of our networks in order to identify some of its relevant properties. In this paragraph we will give some essential notions of RMT.

RMT deals with the statistical properties of matrices with independent random entries. It was initially proposed by Wigner and Dyson in the 1960s as mathematical tool to study the spectrum of complex nuclei [94,95]. Nowadays, it is a powerful approach for modeling various physical systems. It has been successfully used to study the behavior of complex systems, such as spectral properties of large atoms, chaotic systems, and the stock market, as well as gene networks [91,92]. Recent analyses of complex networks under RMT framework [96–98] show that various network models and real world networks also follow RMT predictions.

One of the essential statistical properties in RMT is the correlation among eigenvalues. Real and symmetric matrices with independent random entries are described by the universal law of the Gaussian Orthogonal Ensemble (GOE) [99]. Random matrices are characterized by a strong correlation among eigenvalues. In this case the distribution of the spacing between nearest neighbor eigenvalues is described by Wigner distribution, where we observe the so called "level repulsion", i.e. arbitrarily small spacing between eigenvalues are highly improbable. An adjacency matrix with such features pertains to a random network, where each link is uncorrelated with the others. On the other hand, consider a regular network (where each node has the same number k of neighbors), described by a band matrix with 1 in a diagonal band of width k. In this case the theory argues that there is a high correlation between the links. This causes an uncorrelated spectrum, and the spacing between nearest neighbor eigenvalues follows Poisson distribution (see also [97,98] where also intermediate cases between the two distributions are shown).

We employ here the indicators of the RMT to analyze the two networks we take into account in this chapter, the similarity network and the interac-

tion network. We will compare the properties of the adjacency matrix of our network with the predictions of RMT about Gaussian Orthogonal Ensemble, since our adjacency matrix is real and symmetric. We study the eigenvalue correlations in order to investigate the structure of the whole network and to compare it to that of the colored subnetworks. The main idea behind our attempt relies on the following consideration. RMT predictions represent an average over all possible random connections between the nodes of our network. Deviations from the universal predictions of RMT identify system-specific, nonrandom properties of the system under consideration, allowing to discern relevant information from random noise [99–101]. Therefore we will focus our attention on the deviation of results from RMT, as this is a footprint of a non-randomness in the network due to functional constraints on the gene network. As we will show in more details in par. 4.3.4, RMT also provides a method that allows to identify the nodes and the modules that are the main responsible of such deviation from randomness. Therefore, this kind of analysis allows to extract information about important nodes in the network, and in this sense, it is a powerful tool to identify important genes in the network functioning. Clearly, our analysis here can only identify putative important genes. For a strict validation of the results a deep experimental survey from a biological point of view will be necessary.

4.2.1 Spectral rigidity

In the following, we introduce the spectral rigidity Δ_3. This parameter is a measure of the long range correlation of eigenvalues. More generally, Δ_3 gives information about the fluctuations between the spectrum and its fluctuation-free form. In the following we define Δ_3 and we give the theoretical predictions in the GOE and in the Poisson cases.

Given the spectrum of the adjacency matrix, $\lambda_1 < \lambda_2 < \cdots < \lambda_N$, the cumulative spectral function $N(\lambda)$ is defined as

$$N(\lambda) = \int_{\lambda_1}^{\lambda} \sum_{i=1}^{N} \delta(\lambda - \lambda_i)\, d\lambda \tag{4.1}$$

$N(\lambda)$ is a staircase function (it is the number of eigenvalues between λ_1 and λ): the information about the eigenvalue correlation lies in the fluctuations of $N(\lambda)$ with respect to a smooth function. The procedure for the computation of the spectral rigidity aims to isolate the fluctuations from the smooth trend. So, the first step is to determine the smooth function. In absence of theoretical assumptions about the analytical form of this smooth function, it can be determined empirically via a fit procedure. Nevertheless, finding the right fitting function is an absolutely non trivial task, and it is a crucial point for the success of the procedure. Anyway, the fit of $N(\lambda)$ is the smooth function $\xi(\lambda)$, necessary to perform the so called spectrum unfolding (see [100] for details). $\xi(\lambda)$ maps the series of eigenvalues

$\{\lambda_1, \cdots, \lambda_N\}$ onto a series $\{\xi_1, \cdots, \xi_N\}$. The unfolded staircase function $\widehat{N}(\xi)$ is the staircase computed as a function of these new variables. After unfolding, average spacings is unity, independent of the system, and the derivative of the smooth part with respect to ξ is unity: the unfolding procedure removes from the data the system-specific mean level density. See for instance Figure 4.7 or Figure 4.11 for an example of unfolded staircase. The spectral rigidity Δ_3 characterizes long range correlation among eigenvalues: it is defined as the least-squares deviation of the staircase function $\widehat{N}(\xi)$ from the best local fit to a straight line, averaged over different parts of the spectrum:

$$\Delta_3(L) = \frac{1}{L} \langle \min_{A,B} \int_x^{x+L} (\hat{N}(\xi) - A\xi - B)^2 \, d\xi \rangle \tag{4.2}$$

where $\langle . \rangle$ denotes the average over $x \in [\xi(\lambda_1), \xi(\lambda_N) - L]$ and A, B are the parameters of the best local linear fit.

The theoretical prediction for the GOE (i.e. a random graph) is

$$\Delta_3(L) = \frac{1}{\pi^2} \left[\ln(2\pi L) + \gamma - \frac{5}{4} - \frac{\pi^2}{8} \right] \tag{4.3}$$

where $\gamma \simeq 0.5772$ is the Euler's constant.

For uncorrelated eigenvalues (i.e. a regular graph), instead

$$\Delta_3(L) = \frac{L}{15} \tag{4.4}$$

4.3 Interaction Network of Promoters: InterNet-Pro

Nowadays there is a growing interest in identifying all molecules involved in cellular processes and especially in characterizing their interactions. This may help to understand how these molecules and the network of interactions between them determine the function of the complex cell machinery. Network biology offers a new conceptual framework that could give an essential contribution to our understanding of both biological processes as well as of disease onset [102]. The cell biochemical processes are shaped by the interactions between molecules and between genes, in a complex feedback process where proteins involved in major cellular processes (metabolism and signaling to other cells) form an interaction network that is in turn controlled by the genetic regulatory network [103]. Thus we expect that promoters, regulating the expression of the corresponding genes, are major players in determining network functioning and interactions among genes. This is the reason why we decided to use the information about clustering presented in [27] in studying a gene network topology. The main idea is to color

the nodes of the network on the basis of the cluster the promoter belongs to: then we can study the topology of this node-colored network in order to identify whether nodes of a specific color have specific features, and to study also the topology of single color subnetworks on themselves and in relation to the whole network.

Another analysis we perform on our network is *gene prioritization*. When we take into account genetic disorders, one of the hardest tasks is the identification of the genes responsible of the disease onset, in order to clarify the causes of the disease at a molecular level. This is a non trivial task, especially if one considers that in many cases the disorder is caused by many genes or by faulty interactions among them. Identifying the genes underlying the disorder is the first step for the development of an effective treatment. [104, 105].

With the methods of the RMT we are able to identify *important nodes* in terms of functionality of the gene network related to a disease. Our analysis aims at helping biologists who need to select the most promising genes from large gene lists related to a specific disease, so as to focus the analysis on biological relevance, validation experiments or functional studies only on the most promising candidates.

The first step in network analysis is choosing the network we want to analyze. This passage is definitely non trivial, since the term "gene network" is quite vague and undefined, since many different kind of interaction are possible [106]. In other words, it is necessary to choose the kind of relationship between genes we want to take into account. Many experimental techniques have been developed in order to identify different kinds of interaction between genes or between their corresponding proteins. For instance, two genes can be considered related (thus linked in the network) if the corresponding proteins physically interact in some biological process, or if the proteins are involved in the same metabolic pathway, or if they share the same protein domain. Other kinds of interaction can involve genes instead of their products, so it is possible to consider networks of genetic interactions where two genes are linked if they if their expression levels are similar across conditions in a gene expression study (co-expression network), or if the effects of perturbing one gene were found to be modified by perturbations to a second gene (genetic interaction), or if they are both expressed in the same tissue/the same cellular location (co-localization network).

With such a variety of different data, it is hard to identify the kind of network that is most suitable to one's needs. We chose to use the data obtained with the experimental technique "Two Hybrid" because it is one of the most reliable. This results in a protein-protein interaction network. Moreover, the entire gene network must be somehow truncated: due to the computational limits in the diagonalization of very large matrices, taking into account the network of all human genes is a very hard task. Thus, we chose to take into account the gene network related to cardiomyopathy in

H. sapiens.

4.3.1 Methods: how to build the network

First of all, we download the gene network data from http://thebiogrid.org [93] (file BIOGRID-ORGANISM-Homo_sapiens-3.2.110.tab.txt). We select "Two hybrid" as experimental system in which the interaction was shown. This results in a protein–protein interaction (PPI) network. Each protein is associated to the corresponding gene.

Then, we download from http://www.ncbi.nlm.nih.gov/gene the list of genes related to the keyword "Cardiomyopathy". We use thise gene list as a "seed" of our gene network. To build our network we select these genes and all of their first neighbors in the BioGRID network.

To associate a color to the nodes, we take all the promoters of the genes in the network, and we cluster them with the method reported in [27] and in Chapter 2. Note that since a gene can have several alternative promoters, there is not a straightforward way to associate a single color to a node. Thus, we decided to split each node (representing a gene) in different nodes on the basis of the clusters of its promoters. At most, each node splits in four nodes, when it has promoters belonging to each the four clusters. We do not take into account how many promoters belong to each cluster, i.e. if a gene has two promoters in C1 and three in C4 its node will split in two nodes of different colors. If two genes interact in the BioGRID network, we connect all the alternative nodes of one gene to all the nodes of the other. There is no link among nodes representing the same gene. The largest connected component of this network is a network of interaction between promoters of different classes, which we call InterNetPro. This is made by 2309 nodes (corresp. to 1374 genes) and 14746 links. In this network there are 235 genes associated to cardiomyopathies. This is an undirected network, i.e. the edges have no orientation. We define the adjacency matrix A of this network as $A_{ij} = 1$ if there is a link between node i and node j, $A_{ij} = 0$ otherwise. The adjacency matrix is symmetric, because the network is undirected. In the original BioGrid network each edge has a correspondent adjacency score, but, in order to simplify the analysis, we discard this information and we transform our network in an unweighted graph, where all nonzero entries are equal to 1.

4.3.2 Basic properties of InterNetPro

In order to have a preliminary idea of the network features, we analyze basic network properties such as number of nodes and number of links, shown in Table 4.1 and Table 4.2. We also focus on the properties of the subnetworks of a given color. It interesting to note from Table 4.2 that nodes in C2 and C3 have a high propensity to form links with nodes belonging to other

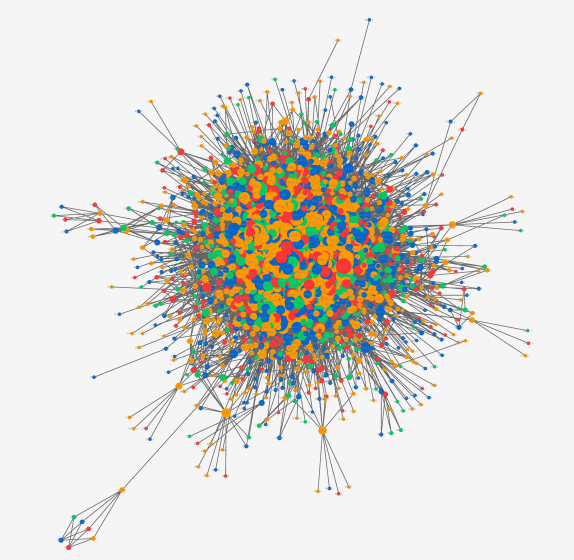


Figure 4.1 InterNetPro. Image of the network obtained with the method described in the text. Nodes belonging to C1, C2, C3 and C4 are colored in orange, red, green and blue respectively. The size of the nodes is related to their connectivity.

subnetworks instead of forming links inside their network.

Fist of all, we compute the degree of the nodes. The degree of a node is defined as the number of edges belonging to that node: $d_i = \sum_j A_{ij}$. In Figure 4.2 we report the connectivity as a function of a node index, where nodes are indexed on the basis of their color. This figure shows that the nodes with high connectivity are equally distributed among all the clusters. Regarding the distribution of the connectivity (see Figure 4.3), we observe that the distribution is similar to that of scale free networks [107,108], even if there are some deviations especially for small degree. Such deviation may be due to the node splitting in nodes of different colors described in the previous paragraph. Anyway, it is hard to identify a clear trend in degree distribution. The degree distribution of each subnetwork (Figure 4.4) has a slightly clearer trend that is more similar to a scale-free behavior.

In addition, we compute the betweenness of the nodes. The communication of two non-adjacent nodes, say j and k, depends on the nodes belonging to the paths connecting j and k. Consequently, a measure of the relevance of a given node can be obtained by counting the number of geodesics going through it, and defining the so-called node betweenness. More precisely, the betweenness b_i of a node i, is defined as [109]

$$b_i = \sum_{j,k \in N, j \neq k} \frac{n_{jk}(i)}{n_{jk}} \tag{4.5}$$

where n_{jk} is the number of shortest paths connecting j and k, while $n_{jk}(i)$ is the number of shortest paths connecting j and k and passing through i. The results are shown in Figure 4.5. It is interesting to note that the nodes with the highest betweenness are in C1. Note also that among the shortest paths connecting any couple of nodes, the longest we find is 10 steps long. That means that, starting from a node, we can reach any other node in the network with 10 steps or less.

Finally, in order to check if links between nodes of the same color are somehow privileged, we also compared the size of the largest connected component of each color with 500 random realizations. We take all the nodes of each color and all the links between them and we compute the size of the largest connected component of this network. Then we take again the whole network, we color the nodes randomly and we compute again the size of the largest connected component for each color. We repeat for 500 random realizations. The results are shown in Figure 4.6. It is very interesting to note that the largest connected component of C1 positively deviates from the random case. This cluster appears to be much more self-linked than the others.

Overall, the results of this preliminary analysis on the network highlight that C1 may have a peculiar structure that differentiates it from the subnetwork of the other clusters.

InterNetPro

InterNetPro				
Nodes	links	Avg. links per node	Inter-subnet links	In-subnet links
2309	14746	6.4	10542	4204

Table 4.1

Layers			
Layer C1			
Nodes	Avg. links per node	Avg. In-subnet links per node	Avg. Inter-subnet links per node
887	7.9	2.6	5.2
Layer C2			
Nodes	Avg. links per node	Avg. In-subnet links per node	Avg. Inter-subnet links per node
406	8.9	1.2	7.7
Layer C3			
Nodes	Avg. links per node	Avg. In-subnet links per node	Avg. Inter-subnet links per node
384	6.4	1.0	5.4
Layer C4			
Nodes	Avg. links per node	Avg. In-subnet links per node	Avg. Inter-subnet links per node
632	5.7	1.6	4.2

Table 4.2

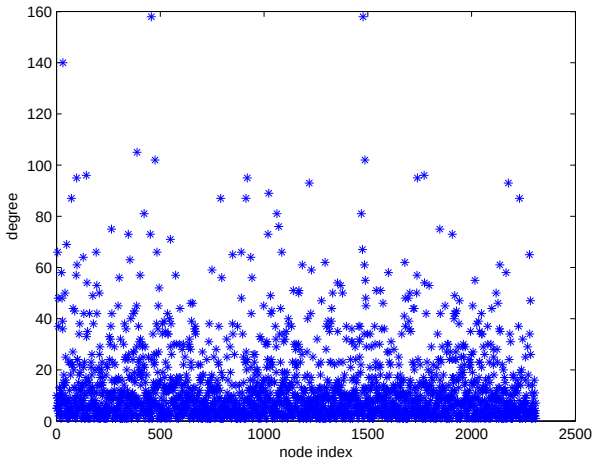


Figure 4.2 Degree as a function of the node index (1 to 887 C1, 888 to 1293 C2, 1294 to 1677 C3, 1678 to 2309 C4).

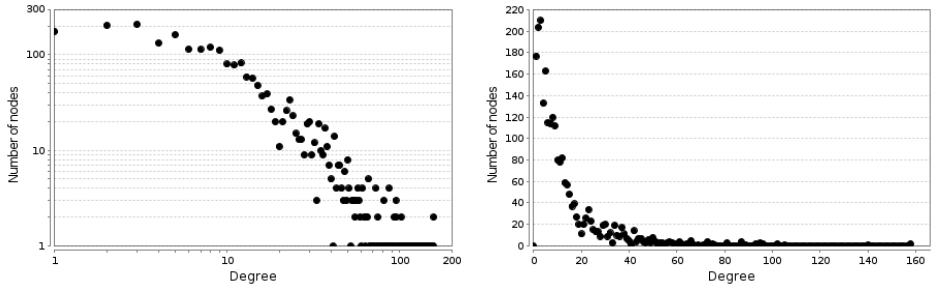
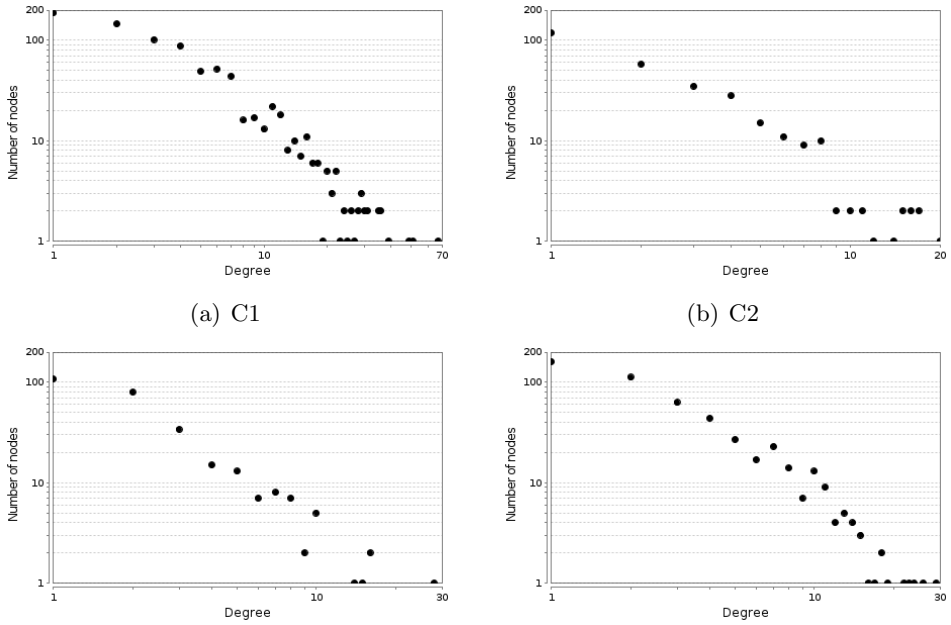


Figure 4.3 Node degree distribution of the whole network InterNetPro. Log-log scale (left) e linear scale (right).



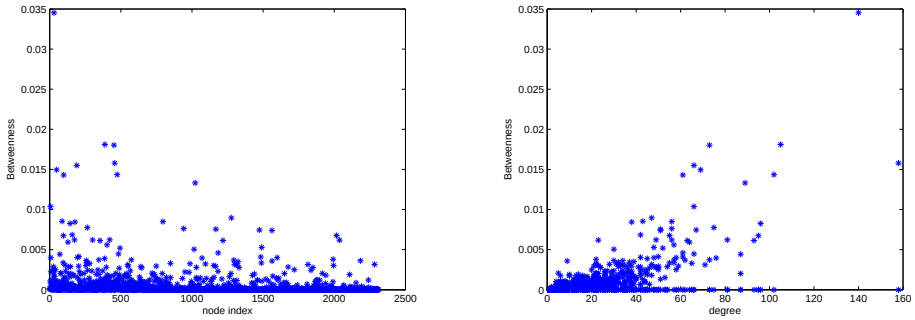
(a) C1

(b) C2

(c) C3

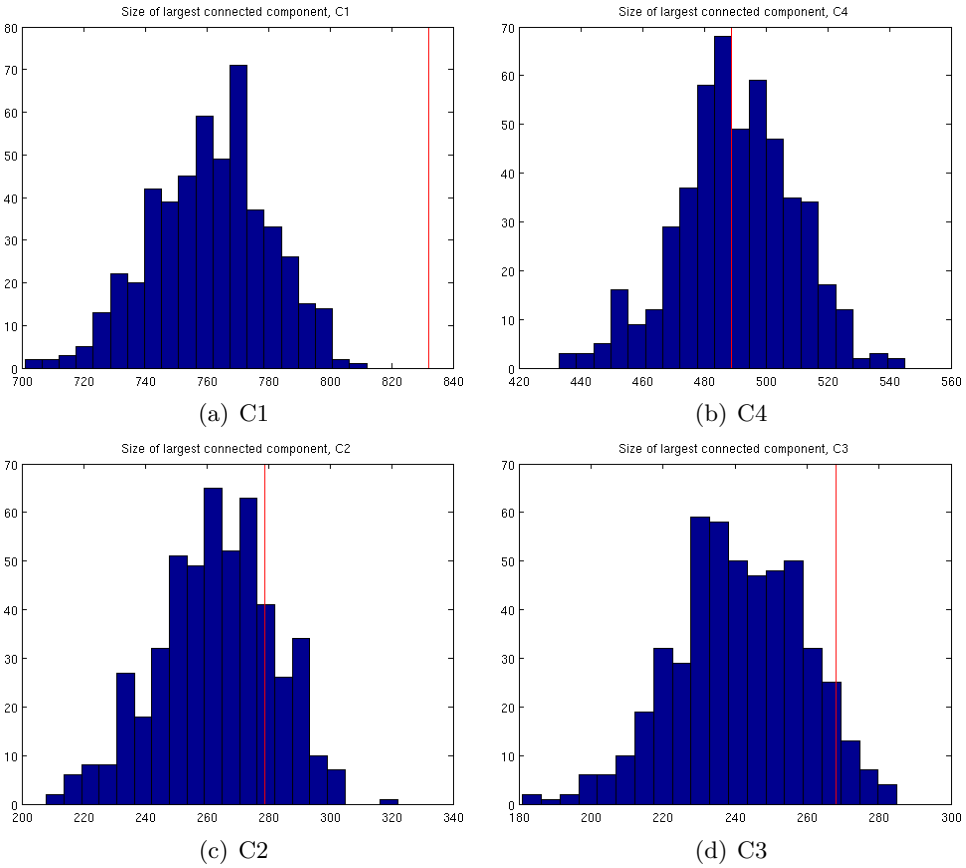
(d) C4

Figure 4.4 Node degree distribution of each of the subnets. Only links inside the subnet are taken into account.



(a) (b)

Figure 4.5 (a) Betweenness as a function of the node index (1 to 887 C1, 888 to 1293 C2, 1294 to 1677 C3, 1678 to 2309 C4). (b) Betweenness as a function of node degree.



(a) C1 (b) C4

(c) C2 (d) C3

Figure 4.6 Histograms of the size of the largest connected component of the subnetwork of each color, obtained with 500 random realizations. The red line is the size obtained with the real colors.

4.3.3 RMT of InterNetPro: spectral rigidity

Following the procedure described in par. 4.2.1, we computed the spectral rigidity of the entire network and of each subnetwork of the clusters. We noted that the procedure poses some issues in the staircase fitting procedure. As a matter of fact, the spectrum of the adjacency matrix is highly degenerate in $\lambda = 0$. This is a problem when we build the staircase function $N(\lambda)$, because the function is ill-defined in $\lambda = 0$. While it is perfectly normal to have some degeneracy in the eigenvalues, such a huge degeneracy in $\lambda = 0$ renders the fitting with a unique function an impossible task.

It is quite surprising to note that such degeneracy problem does not emerge in the main works we referred to [91, 92], because when we tried to repeat the analysis with several different networks (e.g. FANTOM5 [110]), the outcome was always characterized by a very high eigenvalue degeneracy. This suggests that this is a feature of many protein interaction networks. We have several hypotheses about the origin of such degeneracy. To each null eigenvalue correspond two non linearly independent lines of the adjacency matrix. This means that there are many nodes with few connections, and they are all linked to the same nodes. This is not surprising in our network since we observe that there are few hub nodes with many links, while many nodes are scarcely connected - preferentially to the hubs. Moreover, the node duplication process occurred during node coloring surely sharpened this effect. This point was also discussed in [111], where it was related to the particular scale-free structure of protein networks. Anyway, there is another very important point to consider, which is all too often omitted in the discussion of biological networks. Due to both experimental and computational limits, one always considers a subnetwork of all possible interactions between all possible proteins in an organism. This means that the network presented here, and the networks usually analyzed in the literature, are just a portion cut out from a larger network. This implies that many nodes in the network taken into account result to have very few connections, or even just one. This is particularly evident when one considers the peripheral nodes of our network and star-like structures are observed, where many nodes are liked to a single central node. Some of these star-like structures are visible at the borders of the network in Figure 4.1. Such structures are not a peculiarity of our network but are constantly found in many biological networks (e.g. [112, 113]). They surely give a contribution to the spectrum degeneracy: nevertheless, one must bear in mind that they may even be artifacts, since many of the nodes that appear to have just one link may be connected to other nodes that are not taken into account in our limited network.

In our opinion, it still an open problem whether such degeneracy is due to intrinsic properties of the network or it is an artifact due to the limitedness of the approach where just a part of the entire human protein network is

considered.

Let's go back now to the fitting procedure of the staircase function. In order to proceed with the computation of the spectral rigidity Δ_3, we had no choice but ignoring the degenerate part and fitting separately the two remaining branches of $N(\lambda)$ (see Figure 4.7). Note that the same problem arises in the analysis of the subnetworks of the clusters. The two branches were fitted with exponential functions ($f(\lambda) = a + b \cdot e^{c\lambda}$). We have found that the same approach was adopted in an analogous case [114].

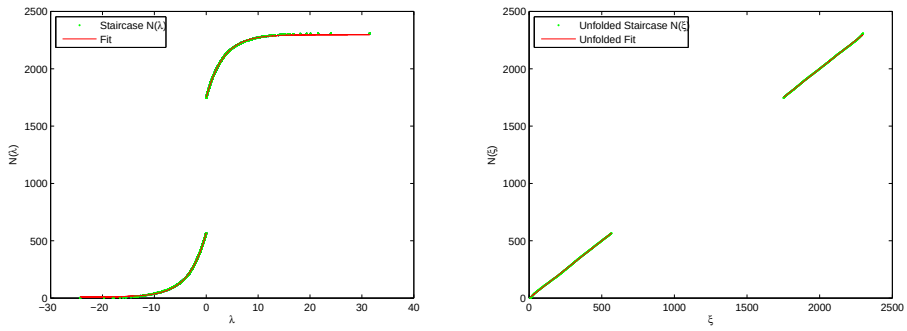


Figure 4.7 Left panel: Staircase $N(\lambda)$ and its fit. Right panel: unfolded staircase $N(\xi)$ and its unfolded fit. Data refer to the entire network.

Once that the unfolded staircase is obtained, we compute the spectral rigidity Δ_3 using the formula 4.2. There will be two contributions from the two branches. The results show that the values of Δ_3 computed for the whole network follow the theoretical prediction of GOE up to $L \sim 30$ (see Figure 4.8 (a)). This behavior has been observed for small-world and scale-free networks (see [98, 115]), as well as for a real world co-expression network [91]. According to the RMT, this implies that besides randomness, the network has some specific features that are probably correlated to functional constraints on the corresponding genes. In the next paragraph, the analysis of the eigenvectors aims at identifying the deviations from randomness within the network: it will become clear that the causes of the deviation from randomness are modules of nodes with a non-trivial and non-random configuration of links among the nodes.

The same analysis has been performed for each of the colored subnetworks. Each subnetwork is represented by an adjacency matrix built extracting from the adjacency matrix of the entire network only the entries corresponding links between nodes of a given color. The analysis of the subnetworks shows interesting results, since C1 follows a behavior similar to small world networks while the other clusters show a completely different structure, similar to Poisson statistics (see Figure 4.8 (b)). These results require a careful explanation; in the analysis of the subnetwork of a single

color all the links with nodes of different colors are cut out, so that the structure of the subnetwork is modified. Indeed, in Table 4.2 we note that these clusters (especially C2 and C3) have a tendency to form links with nodes of different color. It is possible that the removal of such links is the cause of the different behavior observed for these clusters (see also [97, 98] and especially [92], where a transition from GOE to Poisson statistics takes place as some links are removed). Anyway, the results suggest that C2, C3 and C4 have a very different internal link structure with respect to C1.

Overall, we think that the results obtained are quite intriguing. They highlight that our clustering procedure is able to select sets of genes that inside this network form some substructures (namely C2, C3 and C4) with very different properties from the entire network. In other words, the clustering procedure, that is just based on the similarity between promoter sequences, identifies clusters of promoters whose corresponding genes seem to have different roles inside the network we considered.

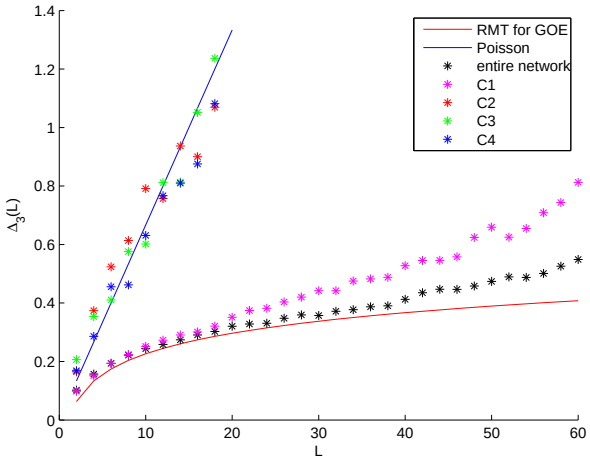


Figure 4.8 $\Delta_3(L)$ as a function of L for the whole network (black stars), and for C1, C2, C3 and C4 (magenta, red, green and blue stars resp.). It is also reported the theoretical prediction for GOE (red line) and for uncorrelated eigenvalues (Poisson, blue line).

4.3.4 RMT of InterNetPro: gene prioritization

We have investigated so far the properties of the eigenvalues of the adjacency matrix. The eigenvector analysis will allow us to identify the network components mainly responsible of the deviation from RMT we observe for Δ_3 (see also [91]).

We define the Inverse Participation Ratio (IPR) I_k and the Shannon Entropy H_k of the eigenvector u^k as follows (l denotes the components of

the k-th eigenvector, N is the size of the network; it is understood that the eigenvectors are normalized).

$$I^k = \sum_{l=1}^{N} (u_l^k)^4 \tag{4.6}$$

$$H^k = \sum_{l=1}^{N} (u_l^k)^2 \ln(u_l^k)^2 \tag{4.7}$$

I^k gives information about the number of components of the eigenvector that are significantly different from zero ($I^k = 1$ if there is only one nonzero component, $I^k \simeq 1/N$ if all components are equal). H^k accounts for the variability of the components. The relevant eigenvectors are those whose I^k and H^k deviate significantly from the random matrix predictions. They are localized on the important nodes of the network, i.e. the nodes mostly responsible of the non-random component. Therefore, the important nodes correspond to the significantly non-zero components of important eigenvectors. Such nodes may have a key functional role in the network and they are good candidate for further analysis on their biological role in cardiomyopathies. The random matrix theory predicts $I^k \sim 1/N$ and $H^k \sim \ln(N/2)$ with $N = 2309$ in our case[1]. In Figure 4.9 we indicate the most relevant eigenvectors, selected with the criterion $H^k < 4$ and $I^k > 0.07$.

In Table 4.3 we report the selected eigenvectors with the corresponding indicated nodes. Note that among the nodes indicated by each eigenvectors we often find nodes of different colors corresponding to the same gene. More generally, in every case the eigenvector indicates all the alternative nodes corresponding to a gene. This suggests that this method identifies the important information at gene level, instead of promoter level.

In Figure 4.10 we report the genes corresponding to the important nodes selected by each eigenvector. In order to highlight the local network structure we have reported, besides the important genes, their nearest neighbors and the corresponding links. Such modules are connected to the rest of the network with links departing from grey or red genes. Such links are not shown in the figure. Instead, all the links of important genes (blue)

[1]The eigenvector components of a GOE random matrix are Gaussian distributed random variables. In the limit of large N, the distribution of $r = |u_k^l|^2$, is the Porter-Thomas distribution [116]

$$P(r) = \frac{N}{\sqrt{2\pi r}} e^{-\frac{Nr}{2}} \tag{4.8}$$

Shannon entropy for the state whose components are described by the above distribution in large N limit is (see also [91])

$$H_s \sim -N \int_0^{\inf} r \ln(r) P(r) dr \sim \ln(\frac{N}{2}) \tag{4.9}$$

are shown. First of all, we notice that the different genes pointed out by an eigenvector are strictly related in the network structure. Moreover, they form local modules with peculiar structures, that are very unlikely in random networks. The important genes appear to play the role of bottlenecks in this modules. Overall, these observation confirm that the nodes selected by our method are probably very important in the functioning of the gene network. The investigation of the role in cardiomyopathy of the genes selected by our method (reported in Table 4.4) will require a biological analysis. For the moment, we can point out that among important genes TP53 emerges. This is a protein involved in many basic cell processes, and while it is not surprising that our method recognizes it as an important node in the network (in fact, it is the largest hub), probably its role is not specific in cardiomyopathies onset but it is to be considered important for cellular life in a broader sense. Other genes tagged as important in the network are GAB1 and GAB2, which form a very peculiar module together with PTPN11. Even if GAB1 and GAB2 have not been directly connected with cardiomyopathies, they are involved in cardiac muscle development and functioning. Connections with cardiac processes has been found also for ADH6, KCTD1,OGDHL,PCBD2, MYH7B and SMYD1. A description of the genes identified is reported in Appendix A.

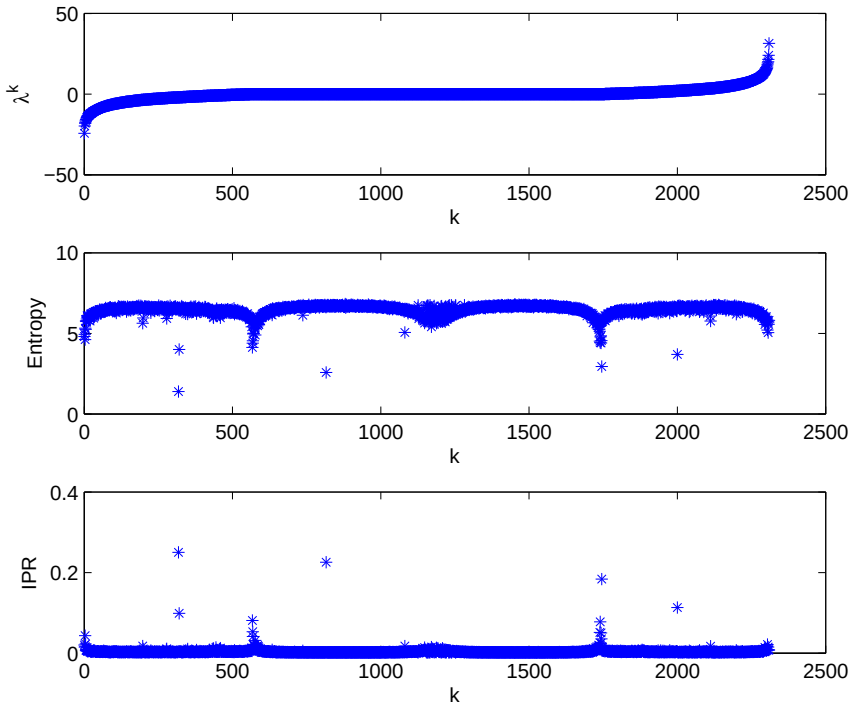


Figure 4.9 First panel: eigenvalues as a function of the eigenumber k. Second panel: entropy H^k as a function of the eigenumber k. Third panel: IPR I^k as a function of the eigenumber k.

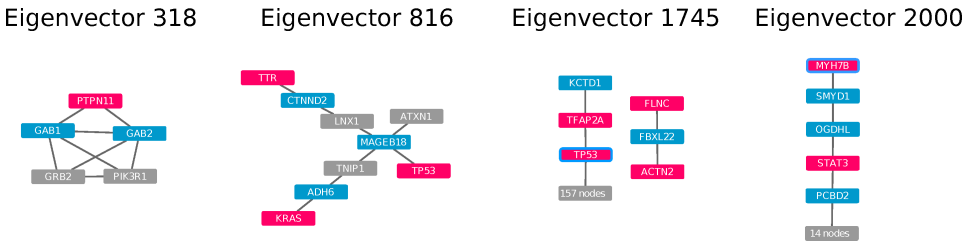


Figure 4.10 Subnetwork made of the nearest neighbors of the important genes pointed out by the most relevant eigenvectors. Important genes are blue, cardiomyopathy genes are red, cardiomyopathy genes identified as important are red with a blue edge, other genes are gray. Each subnetwork is cut out taking only important nodes and their first neighbors.

List of selected important eigenvectors

$H < 4\ IPR > 0.07$

eigenumber k	1/IPR	H	nodes	cluster	connectivity	gene ID
318	4	1.39	177 *	C1	6	2549
			584 +	C1	6	9846
			1788 *	C4	6	
			2086 +	C4	6	
816	4.44	2.57	102 *	C1	2	1501
			1685	C4	4	130
			1744 *	C4	2	
			2309	C4	7	286514
1745	5.44	2.95	457 *	C1	158	7157
			886 +	C1	1	284252
			1477 *	C3	158	
			2307	C4	4	283807
			2308 +	C4	1	
2000	8.86	3.69	790	C1	5	55753
			807	C1	2	57644
			1290 *	C2	2	150572
			1661	C3	15	84105
			2297 *	C4	2	

Table 4.3 Relevant eigenvectors and nodes corresponding to the most significant components. The symbols * and +, for each eigenvector, indicate alternative nodes, i.e. nodes of different color referring to the same gene.

List of top 12 important genes

$H < 4\ IPR > 0.07$

Gene ID	Gene Symbol	Official full name
2549	GAB1	GRB2-associated binding protein 1
9846	GAB2	GRB2-associated binding protein 2
1501	CTNND2	catenin (cadherin-associated protein), delta 2
130	ADH6	alcohol dehydrogenase 6 (class V)
286514	MAGEB18	melanoma antigen family B, 18
7157	TP53	tumor protein p53
284252	KCTD1	potassium channel tetramerization domain containing 1
283807	FBXL22	F-box and leucine-rich repeat protein 22
55753	OGDHL	oxoglutarate dehydrogenase-like
57644	MYH7B	myosin, heavy chain 7B, cardiac muscle, beta
150572	SMYD1	SET and MYND domain containing 1
84105	PCBD2	pterin-4 alpha-carbinolamine dehydratase/ dimerization cofactor of hepatocyte nuclear factor 1 alpha (TCF1)

Table 4.4 Important genes corresponding to the nodes indicated by the relevant eigenvectors.

4.4 Alignment Network

In this section we apply the RMT tools to a completely different network. We describe the results obtained from the analysis of the alignment network. The promoters of this network are the same of the previous one. This means that the nodes of the network are the same, but the links are built in a different way[2]. This network is not affected by the eigenvalue degeneracy problem, so it is a good example of a case in which the staircase function $N(\lambda)$ can be fitted by a unique function.

4.4.1 How to build the network

We build the similarity network where each link is weighted with a similarity score. The similarity score is computed as described in paragraph 2.4.3, with the Needleman-Wunsch algorithm. In other words, the adjacency matrix of the network is the similarity matrix computed in the first step of the clustering algorithm described in Chapter 2. This results in a symmetric matrix. With the analysis presented here, we aim to highlight the topological properties underlying the similarities *within* each cluster.

4.4.2 Results

We computed the spectral rigidity for the entire network and for each of the clusters. In this case we fit the staircase function with the following sigmoid function:

$$f(\lambda) = a + \frac{b-a}{1+10^{d*(c-\lambda)}} \tag{4.10}$$

where a, b, c, d are the fit parameters. In Figure 4.11 we report the staircase with its fit and the unfolded spectrum are shown for the entire network. Analogous results are obtained for the subnetworks corresponding to the clusters.

The results obtained for Δ_3 are reported in Figure 4.12. If we compare the results of this alignment network with the spectral rigidity obtained in the previous paragraph, we notice that in this case the differences between the clusters are less pronounced. Even in this case, data suggest that there is a deviation from the prediction of the RMT, but such deviation occurs at higher values of L. Such deviation is greater for C4. We conclude that there are not significant differences in internal similarity properties of the clusters.

[2]Note that in the previous paragraph the network size was 2309 nodes, while here we have 3660 nodes. The network size does not correspond because in the previous paragraph the analysis was performed on the largest connected component of the network. This implies that some nodes were excludes (the excluded nodes were isolated or part of subnetworks of just two or three nodes).

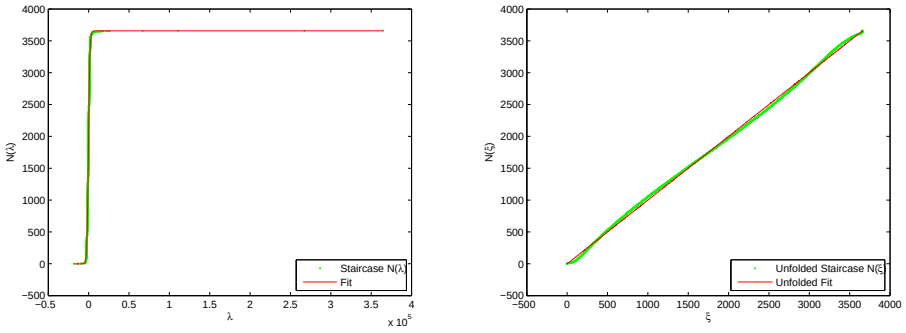


Figure 4.11 Left panel: Staircase $N(\lambda)$ and its fit. Right panel: unfolded staircase $N(\xi)$ and its unfolded fit. Data refer to the entire network.

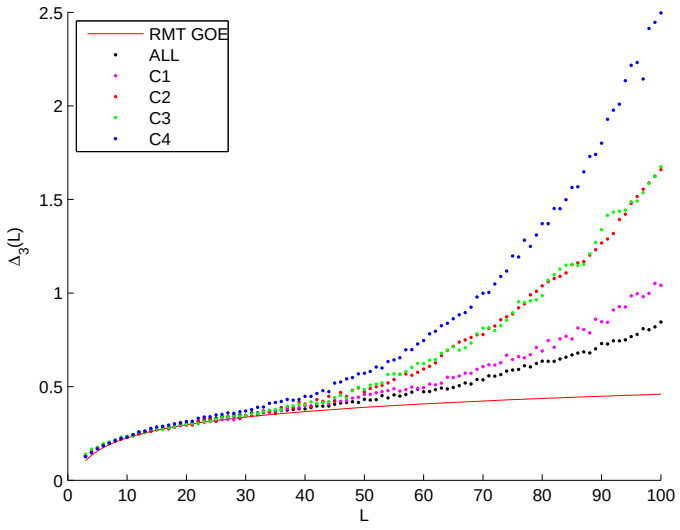


Figure 4.12 $\Delta_3(L)$ as a function of L for the whole network (black dots), and for C1, C2, C3 and C4 (magenta, red, green and blue dots resp.). It is also reported the theoretical prediction for GOE (red line).

4.5 Conclusion

In this chapter, we have applied the methods of the RMT to the study of the properties of two different genetic networks. The interactions between the elements of the network have been defined in two different ways in the two networks taken into account. This study was performed introducing in the network also the information about our clusters, coloring the nodes on the basis of the corresponding cluster, in order to identify the different properties of the subnetworks of different color. The goal was to compare the properties of the adjacency matrix of our network with the predictions of RMT: deviations from the universal predictions of RMT identify system-specific, nonrandom properties of the system under consideration, allowing to discern relevant information from random noise.

In the first case we analyze a protein-protein interaction network. The data about this network are downloaded from a public database [93]. We applied the tools of the RMT to this network, computing the Spectral rigidity for the entire network and for the subnetworks corresponding to the clusters. The comparison of the resulting spectral rigidity with the theoretical predictions allows to highlight peculiar properties of the entire networks and also of the subnetworks of the clusters. We find out that:

1. There are clues of non-randomness in the network, since the data do not follow exactly the prediction of RMT, at least for high values of L.

2. The subnetworks corresponding to the clusters exhibit very different properties, namely C2, C3 and C4 look better described by Poisson theory of regular networks while C1 is more similar to RMT - even with a clear deviation from the theory.

3. The analysis of the eigenvectors of the adjacency matrix allows to identify important nodes in the network and highlight the presence of highly non-random modules in the network that are probably related to the deviations from randomness identified for the entire network.

The results suggest that the four clusters classification we developed, that classified promoters on the basis of their sequence, seems to be meaningful also from a functional point of view, since the proteins encoded by the genes corresponding to the promoters of our clusters exhibit different patterns of interaction in the different clusters. While the subnetwork of C1 is more similar to the theoretical predictions of random networks (even with some deviations), C2, C3 and C4 are very similar to the predictions for regular networks. Regarding the entire network, our interpretation is that there is a mixture of randomness and regular structures in the network. An

hypothesis is that there are highly regular modules, crucial for the functioning of the network, interconnected by more random structures. Overall, we think that our analysis is a good starting point for a biological survey of the data, especially for an experimental validation of our results about the important nodes.

In the second case we analyze a similarity network. We notice that the results deviate from the random predictions more remarkably for C4, and also for C2 and C3.

Chapter 5

Conclusion

First of all we want to remark here the spirit that guided all the work presented in this thesis: the goal was to tackle a typically biological problem with different tools from the mainstream approach. After all, I think that my contribution to the study of promoters would be mainly to offer additional tools and methods with respect to those usually applied to the problem. The very essence of inter-disciplinarity is the mixture of points of view and methodological approaches complementing each other.

As already stated in the Introduction, the main idea behind this work was to gain information about *general* properties of promoters. We tried to characterize *globally* the entire set of promoters of a species (focusing on *H. sapiens*); and we took into account the entire promoter sequence as it is, instead of the just the binding sites. This is the main difference from the usual approach, where studies focus on analyzing the features of a single promoter of a specific gene.

The work starts with the development of a clustering algorithm (Chapter 2, also published in [27]). In accordance with the approach chosen, we search for properties characterizing groups of promoters, forgetting about the features of a single sequence. We apply our algorithm to the promoters of *H. sapiens*, but also a comparison with other species is presented. The clustering algorithm is based on the similarities between promoters, that are compared taking into account the entire sequence. This approach proved to be effective for the identification of biologically relevant features. First of all, we identify four clusters, with very different compositional properties. The promoters of C1 exhibit a CG bases gradient in the direction of the TSS and are associated with housekeeping genes, while promoters of C4 have a prevalence of AT bases and are generally associated with tissue specific genes. The analysis proceeds with the algorithm for the identification of the regular sequences (an interesting peculiarity of promoters). The regular sequences give important clues to understand that promoters of C2 and C3 have been subject to a heavy transposon insertion (of the Alu family),

and are to be considered a single cluster (the only difference is the insertion strand of the transposon). All these results show that evolutionary mechanisms have privileged a certain structure for the promoters of housekeeping genes (C1, CG gradient) that is quite opposite from the structure of tissue specific genes (C4, AT prevalence). Moreover, results highlight the important role in promoter evolution given by transposons. The comparison with other species shows the same cluster organization in mammals (*P. troglodytes* and *M. musculus*), again with a clear transposon footprint in C2 and C3. In more distant species, instead, like a fish (*D. rerio*) and a plant (*A. thaliana*) there is a less marked specificity of promoters: the clusters differentiate only in a small region near the TSS, and do not show specificities nor in regular sequences content nor in transposon content.

In the following chapters we deepen the cluster characterization. In Chapter 3 we make use of two entropic indicators in order to perform two different kind of analyses. The variability analysis of human promoters with positional entropies allows to better characterize compositional features of the clusters: we observe a higher selection near the TSS on promoters of C1, while there is a higher variability in promoters of C4 (according to the tissue specificity of these promoters, that need to perform very different tasks in different conditions). The other analysis performed is a text mining analysis: we apply a keywords identification algorithm to the set of human promoter sequences, treating them like strings of text. The keywords have different features in the different clusters, and some of them are related to mini- and micro-satellites, DNA elements that have been shown to have a role in promoter functions.

In Chapter 4 we broaden our analysis taking into account additional biological information about protein and genetic interactions. The aim is to study a protein-protein interaction network introducing in the network the information about our cluster classification, i.e. coloring the nodes on the basis of the cluster they belong to, in order to compare the properties of the subnetworks of each color with the properties of the entire network. Our analysis employs tools from the Random Matrix Theory. The comparison between the results obtained for our network and the theoretical predictions allows to identify footprints of non-randomness in the network due to functional constraints on the gene network. The methods applied also allow to identify the specific modules inside the network mainly responsible for the deviation from randomness, and allow to identify putative functionally important nodes in the network. Moreover, the analysis of the subnetworks shows that the subnetworks corresponding to C2, C3 and C4 exhibit a very different organization from C1 and from the entire network.

Appendix A

Description of genes identified by the network analysis

In this appendix we report a brief description of the genes identified by the network analysis and reported in Table 4.4.

GAB1 The protein encoded by this gene is a member of the IRS1-like multisubstrate docking protein family. It is an important mediator of branching tubulogenesis and plays a central role in cellular growth response, transformation and apoptosis. Two transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Aug 2008].

Function: SH3/SH2 adaptor activity, protein binding, signal transducer activity.

Process: Fc-epsilon receptor signaling pathway, activation of JUN kinase activity, cell proliferation, epidermal growth factor receptor signaling pathway, epidermis development, fibroblast growth factor receptor signaling pathway, *heart development*, innate immune response, insulin receptor signaling pathway, interleukin-6-mediated signaling pathway, labyrinthine layer development, neurotrophin TRK receptor signaling pathway, phosphatidylinositol-mediated signaling, platelet-derived growth factor receptor signaling pathway, regulation of cell migration, response to oxidative stress.

Component: cytosol.

The GAB1 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog. Orthologs from Annotation Pipeline: 164 organisms have orthologs with human gene GAB1. Location: 4q31.21

GAB2 This gene is a member of the GRB2-associated binding protein (GAB) gene family. These proteins contain pleckstrin homology (PH) do-

main, and bind SHP2 tyrosine phosphatase and GRB2 adapter protein. They act as adapters for transmitting various signals in response to stimuli through cytokine and growth factor receptors, and T- and B-cell antigen receptors. The protein encoded by this gene is the principal activator of phosphatidylinositol-3 kinase in response to activation of the high affinity IgE receptor. Two alternatively spliced transcripts encoding different isoforms have been described for this gene. [provided by RefSeq, Nov 2009]

Function: phosphatidylinositol-3,4,5-trisphosphate binding, phosphatidylinositol-3,4-bisphosphate binding, protein binding, transmembrane receptor protein tyrosine kinase adaptor activity.

Process: Fc-epsilon receptor signaling pathway, cell migration, innate immune response, integrin-mediated signaling pathway, osteoclast differentiation, phosphatidyli-
nositol - mediated signaling, positive regulation of cell proliferation, positive regulation of mast cell degranulation, transmembrane receptor protein tyrosine kinase signaling pathway.

Component: cytoplasm, cytosol, plasma membrane.

The GAB2 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog. Orthologs from Annotation Pipeline: 159 organisms have orthologs with human gene GAB2. Location: 11q14.1 Ho trpvato che :PI3K-Akt pathway regulates cardiomyocyte size, survival, angiogenesis, and inflammation in both physiological and pathological cardiac hypertrophy.

CTNND2 This gene encodes an adhesive junction associated protein of the armadillo/beta-catenin superfamily and is implicated in brain and eye development and cancer formation. The protein encoded by this gene promotes the disruption of E-cadherin based adherens junction to favor cell spreading upon stimulation by hepatocyte growth factor. This gene is overexpressed in prostate adenocarcinomas and is associated with decreased expression of tumor suppressor E-cadherin in this tissue. This gene resides in a region of the short arm of chromosome 5 that is deleted in Cri du Chat syndrome. Alternative splicing results in multiple transcript variants encoding different isoforms. [provided by RefSeq, Dec 2013]

Function: beta-catenin binding, protein binding.

Process: Wnt signaling pathway, cell adhesion, dendritic spine morphogenesis, learning, morphogenesis of a branching structure, regulation of canonical Wnt signaling pathway, regulation of synaptic plasticity, regulation of transcription, DNA-templated, signal transduction, single organismal cell-cell adhesion, synapse organization, transcription, DNA-templated.

Component: adherens junction, cytoplasm, dendrite, nucleus, perikaryon, postsynaptic density.

The CTNND2 gene is conserved in chimpanzee, Rhesus monkey, dog,

cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, and frog. Orthologs from Annotation Pipeline: 126 organisms have orthologs with human gene CTNND2. Location: 5p15.2

ADH6 This gene encodes class V alcohol dehydrogenase, which is a member of the alcohol dehydrogenase family. Members of this family metabolize a wide variety of substrates, including ethanol, retinol, other aliphatic alcohols, hydroxysteroids, and lipid peroxidation products. This gene is expressed in the stomach as well as in the liver, and it contains a glucocorticoid response element upstream of its 5' UTR, which is a steroid hormone receptor binding site. Alternatively spliced transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]

Function: alcohol dehydrogenase (NAD) activity, zinc ion binding.

Process: ethanol oxidation, ethanol oxidation, response to ethanol, small molecule metabolic process, xenobiotic metabolic process.

Component: cytosol, extracellular exosome.

The ADH6 gene is conserved in chimpanzee, Rhesus monkey, cow, rat, chicken, and frog. Orthologs from Annotation Pipeline: 31 organisms have orthologs with human gene ADH6. Location: 4q23

MAGEB18 Function: protein binding

Component: cytoplasm

The MAGEB18 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, and rat. Orthologs from Annotation Pipeline: 31 organisms have orthologs with human gene MAGEB18. Location: Xp21.3

TP53 This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Mutations in this gene are associated with a variety of human cancers, including hereditary cancers such as Li-Fraumeni syndrome. Alternative splicing of this gene and the use of alternate promoters result in multiple transcript variants and isoforms. Additional isoforms have also been shown to result from the use of alternate translation initiation codons (PMIDs: 12032546, 20937277). [provided by RefSeq, Feb 2013] The TP53 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, zebrafish, and frog. Orthologs from Annotation Pipeline: 109 organisms have orthologs with human gene TP53. Location: 17p13.1 .

KCTD1 Function: protein binding, transcription corepressor activity, transcription factor binding.

Process: negative regulation of transcription, DNA-templated, protein homooligomerization, transcription, DNA-templated.

Component: nucleus.

The KCTD1 gene is conserved in chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, and frog. Orthologs from Annotation Pipeline: 140 organisms have orthologs with human gene KCTD1. Location: 18q11.2

FBXL22 This gene encodes a member of the F-box protein family. This F-box protein interacts with S-phase kinase-associated protein 1A and cullin in order to form SCF complexes which function as ubiquitin ligases.[provided by RefSeq, Sep 2010]

Function: contributes to ubiquitin protein ligase activity, ubiquitin protein ligase activity.

Process: SCF-dependent proteasomal ubiquitin-dependent protein catabolic process, proteasome-mediated ubiquitin-dependent protein catabolic process, protein ubiquitination.

Component: SCF ubiquitin ligase complex, Z disc, cytoplasm.

Orthologs from Annotation Pipeline: 161 organisms have orthologs with human gene FBXL22. Location: 15q22.31

OGDHL Function: metal ion binding, oxoglutarate dehydrogenase (succinyl-transferring) activity, protein binding, thiamine pyrophosphate binding.

Process: glycolytic process, tricarboxylic acid cycle.

Component: NOT cytosol, mitochondrial matrix, mitochondrion, oxoglutarate dehydrogenase complex.

The OGDHL gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog. Orthologs from Annotation Pipeline: 155 organisms have orthologs with human gene OGDHL. Location: 10q11.23

MYH7B The myosin II molecule is a multi-subunit complex consisting of two heavy chains and four light chains. This gene encodes a heavy chain of myosin II, which is a member of the motor-domain superfamily. The heavy chain includes a globular motor domain, which catalyzes ATP hydrolysis and interacts with actin, and a tail domain in which heptad repeat sequences promote dimerization by interacting to form a rod-like alpha-helical coiled coil. This heavy chain subunit is a slow-twitch myosin. Alternatively spliced transcript variants have been found, but the full-length nature of these variants is not determined. [provided by RefSeq, Mar 2010] The MYH7B gene is conserved in chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, C.elegans, and frog. Orthologs from Annotation Pipeline: 143 organisms have orthologs with human gene MYH7B. Function: ATP binding, actin binding, motor activity, protein binding.

Process: metabolic process.
Component: membrane, myosin filament. Location: 20q11.22

SMYD1 Function: DNA binding, histone-lysine N-methyltransferase activity, metal ion binding, protein binding, transcription corepressor activity.

Process: chromatin remodeling, *heart development*, histone lysine methylation, negative regulation of transcription, DNA-templated, positive regulation of myoblast differentiation, positive regulation of myotube differentiation, skeletal muscle cell differentiation, transcription, DNA-templated.

Component: cytoplasm, nucleus. The SMYD1 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog. Orthologs from Annotation Pipeline: 161 organisms have orthologs with human gene SMYD1. Location: 2p11.2

PCBD2 Function: 4-alpha-hydroxytetrahydrobiopterin dehydratase activity, phenylalanine 4-monooxygenase activity, protein binding, oxidation-reduction process, positive regulation of transcription, DNA-templated, protein heterooligomerization, protein homotetramerization, tetrahydrobiopterin biosynthetic process.

Component: cellular component, mitochondrion, nucleus.

The PCBD2 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, mosquito, C.elegans, S.pombe, and frog. Orthologs from Annotation Pipeline: 149 organisms have orthologs with human gene PCBD2. Location: 5q31.1

Acknowledgements

Finally, this thesis has come to a conclusion! Now, I look behind and think about the PhD years, and I see that this thesis is the result of a long and intense experience. It would not have been brought to completion without the contribution of many people who advised me and supported me in many different ways. I am very happy to have the chance to thank them all.

I want to thank my supervisor, Roberto Livi, who found a balance between advising me and letting me free to experience my own way. I wish to acknowledge the research group I worked in, Elisa Calistri and Francesca di Patti, for their help in the biological topics and for sharing with me their experience in many technical fields. I am especially grateful for the useful advices and discussions, and for their support and patience in the last year of my PhD (especially for the late, late afternoon meetings!).

I also want to thank Stefano Luccioli and Stefano Lepri: they always kindly and cheerily offered me support, as well as insightful discussions.

I want to acknowledge Matteo Marsili, for presenting me his entropic indicator and for his will to collaborate in applying it to promoters. I am grateful to Michele Caselle for the very stimulating discussions about DNA, promoters and especially transposons, and to Matteo Benelli, for his help in finding a way across the huge load of biological networks data.

I also wish to thank my colleagues in Altran and Selex ES for their warm encouragement to bring this thesis to an end. A special share of my gratitude goes to Enrico Fossati, who considerably supported me to find the necessary time to complete this thesis.

In questa lunga esperienza ho condiviso alti e bassi con tanti compagni che hanno reso le difficoltà più sopportabili e i bei momenti più piacevoli: il mio compagno di stanza Enrico, Elena, Davide, Edo, Alessandro, Lucio...

Questa tesi non sarebbe venuta alla luce senza il supporto delle persone che mi sono state vicino: ringrazio specialmente Irene e il Pupillo, che mi hanno sempre esortata a tenere duro nei momenti più faticosi. Senza il supporto della mia famiglia (specialmente dei miei genitori) che ha sempre creduto in me, non sarei sopravvissuta a questo ultimo anno di dottorato. Completare il percorso di dottorato ha anche richiesto molte rinunce, tutte condivise con Curzio: a lui un ringraziamento davvero speciale per essermi stato sempre vicino anche da lontano e per aver stretto i denti insieme a me.

Bibliography

[1] Franklin RE, Gosling RG (1953) Molecular configuration in sodium thymonucleate. Nature 171: 740–741.

[2] Crick FHC, Watson JD (1953) Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature 171: 737-738.

[3] Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, et al. (1965) RNA codewords and protein synthesis, VII. On the general nature of the RNA code. Proc Natl Acad Sci U S A 53: 1161–1168.

[4] Consortium IHGS (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

[5] The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

[6] King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188: 107-116.

[7] Carroll SB (2000) Endless forms: The evolution of gene regulation and morphological diversity. Cell 101: 577 - 580.

[8] Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8.

[9] Carroll SB (2005) Evolution at two levels: On genes and form. PLoS Biol 3: e245.

[10] Calistri E, Buiatti M, Livi R (2014) Variation and constraints in species-specific promoter sequences. Journal of Theoretical Biology 363: 357 - 366.

[11] Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E (2007) Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene 389: 52 - 65.

[12] Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, et al. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet 8: 424-436.

[13] Shelenkov A, Korotkov E (2009) Search of regular sequences in promoters from eukaryotic genomes. Computational Biology and Chemistry 33: 196 - 204.

[14] Sela I, Lukatsky D (2011) DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. Biophysical Journal 101: 160 - 166.

[15] Afek A, Lukatsky D (2013) Genome-wide organization of eukaryotic preinitiation complex is influenced by nonconsensus protein-DNA binding. Biophysical Journal 104: 1107 - 1115.

[16] Kolomeisky A (2013) Mechanisms of protein binding to DNA: Statistical interactions are important. Biophysical Journal 104: 966 - 967.

[17] Afek A, Schipper JL, Horton J, Gordân R, Lukatsky DB (2014) Protein−DNA binding in the absence of specific base-pair recognition. Proceedings of the National Academy of Sciences 111: 17140-17145.

[18] Afek A, Lukatsky D (2012) Nonspecific protein-DNA binding is widespread in the yeast genome. Biophysical Journal 102: 1881 - 1888.

[19] Tchernaenko V, Halvorson HR, Kashlev M, Lutter LC (2008) DNA bubble formation in transcription initiation†. Biochemistry 47: 1871-1884.

[20] Abeel T, Saeys Y, Bonnet E, Rouzé P, Van de Peer Y (2008) Generic eukaryotic core promoter prediction using structural features of DNA. Genome Research 18: 310-323.

[21] Cairns BR (2009) The logic of chromatin architecture and remodelling at promoters. Nature 461: 193-198.

[22] Bolshoy A, Nevo E (2000) Ecologic genomics of DNA: upstream bending in prokaryotic promoters. Genome research 10: 1185–1193.

[23] Segal E, Widom J (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. Current Opinion in Structural Biology 19: 65 - 71.

[24] Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual Review of Genetics 44: 445-477.

[25] Testori A, Caizzi L, Cutrupi S, Friard O, De Bortoli M, et al. (2012) The role of transposable elements in shaping the combinatorial interaction of transcription factors. BMC genomics 13: 400.

[26] Calistri E, Livi R, Buiatti M (2011) Evolutionary trends of GC/AT distribution patterns in promoters. Molecular Phylogenetics and Evolution 60: 228 - 235.

[27] Pettinato L, Calistri E, Di Patti F, Livi R, Luccioli S (2014) Genome-wide analysis of promoters: Clustering by alignment and analysis of regular patterns. PLoS ONE 9: e85260.

[28] Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48: 443–453.

[29] Smith TF, Waterman MS (1981) Identification of common molecular subsequences. Journal of Molecular Biology 147: 195–197.

[30] Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276–277.

[31] Peyrard M, Bishop AR (1989) Statistical mechanics of a nonlinear model for DNA denaturation. Phys Rev Lett 62: 2755–2758.

[32] Dauxois T, Peyrard M, Bishop AR (1993) Dynamics and thermodynamics of a nonlinear model for DNA denaturation. Phys Rev E 47: 684–695.

[33] Dauxois T, Peyrard M, Bishop AR (1993) Entropy-driven DNA denaturation. Phys Rev E 47: R44–R47.

[34] von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17: 395–416.

[35] Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B (2004) Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics 5: 34.

[36] Louie E, Ott J, Majewski J (2003) Nucleotide frequency variation across human genes. Genome Research 13: 2594-2601.

[37] Calistri E (2008) Variability and constraints in promoter evolution. Ph.D. thesis, Nonlinear dynamics and complex systems.

[38] Meader S, Ponting CP, Lunter G (2010) Massive turnover of functional sequence in human and other mammalian genomes. Genome Research 20: 1335-1343.

[39] Anderson PW (1958) Absence of diffusion in certain random lattices. Physical review 109: 1492.

[40] Shine J, Dalgarno L (1975) Determinant of cistron specificity in bacterial ribosomes. Nature 254: 34-38.

[41] Jaumot J, Gargallo R (2010) Using principal component analysis to find correlations between loop-related and thermodynamic variables for G-quadruplex-forming sequences. Biochimie 92: 1016 - 1023.

[42] Aviñó A, Cubero E, González C, Eritja R, Orozco M (2003) Antiparallel triple helices. Structural characteristics and stabilization by 8-amino derivatives. Journal of the American Chemical Society 125: 16127-16138.

[43] Yang Z, Engel JD (1994) Biochemical characterization of the developmental stage- and tissue-specific erythroid transcription factor, NF-E4. Journal of Biological Chemistry 269: 10079-87.

[44] Koch KA, Thiele DJ (1999) Functional analysis of a homopolymeric (dA-dT) element that provides nucleosomal access to yeast and mammalian transcription factors. Journal of Biological Chemistry 274: 23752-23760.

[45] Farnham PJ (2009) Insights from genomic profiling of transcription factors. Nat Rev Genet 10: 605-616.

[46] Grandi FC, Rosser JM, An W (2013) LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. Molecular Biology and Evolution 30: 503-512.

[47] Carey MF, Peterson CL, Smale ST (2012) Identifying cis-acting DNA elements within a control region. Cold Spring Harbor Protocols 2012: pdb.top068171.

[48] Kel A, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis O, et al. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Research 31: 3576-3579.

[49] Akiyama Y. TFSEARCH: Searching transcription factor binding sites. URL `http://www.rwcp.or.jp/papia/`. Accessed 2012 July.

[50] Bryne JC, Valen E, Tang MHE, Marstrand T, Winther O, et al. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Research 36: D102-D106.

[51] Wunderlich Z, Mirny LA (2009) Different gene regulation strategies revealed by analysis of binding motifs. Trends in Genetics 25: 434 - 440.

[52] Polak P, Domany E (2006) Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics 7: 133.

[53] Jacques PE, Jeyakani J, Bourque G (2013) The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet 9: e1003504.

[54] Cotton J (2001) Retroviruses from retrotransposons. Genome Biology 2: reports0006.

[55] Smit A, Hubley R, Green P. RepeatMasker. URL `http://repeatmasker.org`. Accessed 2012 September.

[56] Bourque G, Leong B, Vega VB, Chen X, Lee YL, et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Research 18: 1752-1762.

[57] Bourque G (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. Current Opinion in Genetics and Development 19: 607 - 612.

[58] Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321: 209-213.

[59] Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. Journal of Molecular Biology 196: 261 - 282.

[60] Cohen N, Kenigsberg E, Tanay A (2011) Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell 145: 773-786.

[61] Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, et al. DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. Nucleic Acids Res : D86–89.

[62] Loots G, Ovcharenko I (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. Bioinformatics 23: 122-124.

[63] The Arabidopsis Information Resource (TAIR), March 2008. on www.arabidopsis.org. URL `ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_-release/TAIR8_sequences/`.

[64] Stewart G, Sun J (1990) Matrix perturbation theory. Academic Press.

[65] Campa A, Giansanti A (1998) Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains. Phys Rev E 58: 3585–3588.

[66] Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez gene: gene-centered information at NCBI. Nucleic Acids Research 33: D54-D58.

[67] Marsili M, Mastromatteo I, Roudi Y (2013) On sampling and modeling complex systems. Journal of Statistical Mechanics: Theory and Experiment 2013: P09003.

[68] Tan AH, et al. (1999) Text mining: The state of the art and the challenges. In: Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. volume 8, pp. 65–70.

[69] Rhee HS, Pugh BF (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature 483: 295–301.

[70] Sugihara F, Kasahara K, Kokubo T (2010) Highly redundant function of multiple AT-rich sequences as core promoter elements in the TATA-less RPS5 promoter of Saccharomyces cerevisiae. Nucleic acids research : gkq741.

[71] Basehoar AD, Zanton SJ, Pugh B (2004) Identification and distinct regulation of yeast TATA box-containing genes. Cell 116: 699 - 709.

[72] Huisinga KL, Pugh BF (2004) A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Molecular Cell 13: 573 - 585.

[73] Lee TI, Causton HC, Holstege FC, Shen WC, Hannett N, et al. (2000) Redundant roles for the TFIID and SAGA complexes in global transcription. Nature 405: 701–704.

[74] Tirosh I, Barkai N (2008) Two strategies for gene regulation by promoter nucleosomes. Genome Research 18: 1084-1091.

[75] Dineen DG, Wilm A, Cunningham P, Higgins DG (2009) High DNA melting temperature predicts transcription start site location in human and mouse. Nucleic acids research 37: 7360–7367.

[76] Alexandrov BS, Gelev V, Yoo SW, Alexandrov LB, Fukuyo Y, et al. (2010) DNA dynamics play a role as a basal transcription factor in the positioning and regulation of gene transcription initiation. Nucleic acids research 38: 1790–1795.

[77] Alexandrov BS, Gelev V, Yoo SW, Bishop AR, Rasmussen KØ, et al. (2009) Toward a detailed description of the thermally induced dynamics of the core promoter. PLoS Comput Biol 5: e1000313.

[78] Eller CD, Regelson M, Merriman B, Nelson S, Horvath S, et al. (2007) Repetitive sequence environment distinguishes housekeeping genes. Gene 390: 153 - 165.

[79] Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular Ecology 11: 2453–2465.

[80] Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, et al. (2013) Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PloS one 8: e54710.

[81] Darvish H, Nabi MO, Firouzabadi SG, Karimlou M, Heidari A, et al. (2011) Exceptional human core promoter nucleotide compositions. Gene 475: 79–86.

[82] Ohadi M, Mohammadparast S, Darvish H (2012) Evolutionary trend of exceptionally long human core promoter short tandem repeats. Gene 507: 61–67.

[83] Stallings R, Ford A, Nelson D, Torney D, Hildebrand C, et al. (1991) Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. Genomics 10: 807 - 815.

[84] Gebhardt F, Zänker KS, Brandt B (1999) Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. Journal of Biological Chemistry 274: 13176-13180.

[85] Gendrel CG, Boulet A, Dutreix M (2000) (CA/GT) n microsatellites affect homologous recombination during yeast meiosis. Genes & development 14: 1261–1268.

[86] Okladnova O, Syagailo YV, Tranitz M, Stöber G, Riederer P, et al. (1998) A promoter-associated polymorphic repeat modulatespax-6expression in human brain. Biochemical and Biophysical Research Communications 248: 402 - 405.

[87] Aharoni A, Baran N, Manor H (1993) Characterization of a multisubunit human protein which selectively binds single stranded d(GA)n and d(GT)n sequence repeats in DNA. Nucleic Acids Research 21: 5221-5228.

[88] Wang G, Vasquez KM (2006) Z-DNA, an active element in the genome. Frontiers in bioscience: a journal and virtual library 12: 4424–4438.

[89] Vergnaud G, Denoeud F (2000) Minisatellites: Mutability and genome architecture. Genome Research 10: 899-907.

[90] Simone R, Fratta P, Neidle S, Parkinson GN, Isaacs AM (2015) G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome. {FEBS} Letters 589: 1653 - 1668.

[91] Jalan S, Solymosi N, Vattay G, Li B (2010) Random matrix analysis of localization properties of gene coexpression network. Phys Rev E 81: 046118.

[92] Luo F, Yang Y, Zhong J, Gao H, Khan L, et al. (2007) Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics 8: 299.

[93] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) Biogrid: a general repository for interaction datasets. Nucleic acids research 34: D535–D539.

[94] Dyson FJ (1962) Statistical theory of the energy levels of complex systems. i. Journal of Mathematical Physics 3: 140–156.

[95] Dyson FJ, Mehta ML (1963) Statistical theory of the energy levels of complex systems. iv. Journal of Mathematical Physics 4: 701–712.

[96] Bandyopadhyay JN, Jalan S (2007) Universality in complex networks: Random matrix analysis. Phys Rev E 76: 026109.

[97] Jalan S (2009) Spectral analysis of deformed random networks. Physical Review E 80: 046101.

[98] De Carvalho J, Jalan S, Hussein M (2009) Deformed gaussian-orthogonal-ensemble description of small-world networks. Physical Review E 79: 056222.

[99] Mehta ML (2004) Random matrices, volume 142. Academic press.

[100] Guhr T, Müller-Groeling A, Weidenmüller HA (1998) Random-matrix theories in quantum physics: common concepts. Physics Reports 299: 189–425.

[101] Brody TA, Flores J, French JB, Mello P, Pandey A, et al. (1981) Random-matrix physics: spectrum and strength fluctuations. Reviews of Modern Physics 53: 385.

[102] Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113.

[103] Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296: 910-913.

[104] Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nature Reviews Genetics 12: 56–68.

[105] Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics 13: 523–536.

[106] Weng G, Bhalla US, Iyengar R (1999) Complexity in biological signaling systems. Science 284: 92-96.

[107] Barabási AL, et al. (2009) Scale-free networks: a decade and beyond. science 325: 412.

[108] Barabási AL, Albert R, Jeong H (2000) Scale-free characteristics of random networks: the topology of the world-wide web. Physica A: Statistical Mechanics and its Applications 281: 69 - 77.

[109] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Physics Reports 424: 175 - 308.

[110] Consortium TF, et al. (2014) A promoter-level mammalian expression atlas. Nature 507: 462–470.

[111] Yadav A, Jalan S (2015) Origin and implications of zero degeneracy in networks spectra. Chaos 25: 043110.

[112] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, et al. (2012) Wisdom of crowds for robust gene network inference. Nature methods 9: 796–804.

[113] Thum KE, Shin MJ, Gutiérrez RA, Mukherjee I, Katari MS, et al. (2008) An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in arabidopsis. BMC systems biology 2: 31.

[114] Agrawal A, Sarkar C, Dwivedi SK, Dhasmana N, Jalan S (2014) Quantifying randomness in protein–protein interaction networks of different species: A random matrix approach. Physica A: Statistical Mechanics and its Applications 404: 359 - 367.

[115] Jalan S, Bandyopadhyay JN (2007) Random matrix analysis of complex networks. Physical Review E 76: 046107.

[116] Zyczkowski K (1991). Chapter in Quantum chaos, edited by H. A. Cerdeira, R. Ramaswami, M. C. Gutzwiller and G. Casati.