Universidade de Aveiro Departamento de Matemática
2018

Ana Helena Marques de Pinho Tavares

Análise de distribuições de distâncias entre palavras genómicas

Analysis of inter genomic word distance distributions

Universidade de Aveiro Departamento de Matemática
2018

Ana Helena Marques de Pinho Tavares

Análise de distribuições de distâncias entre palavras genómicas

Analysis of inter genomic word distance distributions

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Matemática, realizada sob orientação científica da Doutora Vera Mónica Almeida Afreixo, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro e da Doutora Maria Paula de Pinho de Brito Duarte Silva, Professora Associada da Faculdade de Economia da Universidade do Porto.

Apoio financeiro da Fundação para a Ciência e a Tecnologia (FCT) através da bolsa de investigação com a referência PD/BD/105729/2014.

dedicatória

Aos que intercetam a minha vida,
fruto da *aleatoriedade* da condição humana,
e que nela permanecem,
em resultado de características *excecionais* que os tornam únicos
e, por isso, *atípicos* entre a multidão.

Em particular,
meus pais, Ana e Efrem
meu marido, João
meus filhos, Iris, Diana e João
meus amigos, Andreia e Vera.

o júri / the jury

presidente / president

Doutora Silvina Maria Vagos Santana
Professora Catedrática da Universidade de Aveiro

vogais / examiners committee

Doutora Lisete Maria Ribeiro de Sousa
Professora Associada da Universidade de Lisboa

Doutor Jacobo de Uña Álvarez
Professor Catedrático da Universidade de Vigo

Doutor Miguel Francisco Almeida Pereira Rocha
Professor Associado da Universidade do Minho

Doutor Pedro Filipe Pessoa Macedo
Professor Auxiliar da Universidade de Aveiro

Doutora Vera Mónica Almeida Afreixo
Professora Auxiliar da Universidade de Aveiro (orientadora)

agradecimentos / acknowledgements

The work underpinning this PhD thesis was a team effort, built on respect and reciprocity between people and organizations. The challenging road towards a PhD degree is long and challenging, where inspiring ideas typically come to light only after a long and sometimes hard research work.

I would like to recognize the Center for Research & Development in Mathematics and Applications (CIDMA), the Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and the Institute for Biomedicine (iBiMED) for financial and technological support and for providing me a rich work environment. I also acknowledge the Portuguese Foundation for Science and Technology (FCT) for funding this research.

I would like to offer a special appreciation to my supervisor Vera Afreixo for her support always providing guidance and constructive feedback, and for all the knowledge sharing. Her sustained encouragement and companionship were crucial to the initial embracing of this project and for a positive outcome after 36 months. I would like to thank to my co-supervisor Paula Brito for all the knowledge sharing and for her vision to build international cooperation with other prestigious researchers.

This achievement was only made possible due to the collaborative work of such enlightening people as: Armando Pinho, Carlos Bastos, João Rodrigues and Paulo Ferreira, researchers at IEETA; Raquel Silva, researcher at IEETA and iBiMED; and Peter Rousseeuw and Jakob Raymaekers, researchers at the University of Leuven. Such collaborative work has enhanced the multidisciplinary approach of this thesis.

Lastly, to João, for being part of my life, always motivating me to look for new challenges and giving me strength to overcome the trials of life when it matters most – thank you.

palavras-chave

Distâncias entre palavras genómicas, Distribuições de distâncias, Dissimilaridade, Classificação, Deteção de entidades atípicas, Reconhecimento de padrões

resumo

A investigação do ADN é uma das áreas mais desenvolvidas neste e no último século. O crescente aumento do número de genomas sequenciados tem exigido técnicas quantitativas mais eficientes para a identificação de características gerais e específicas das sequências genómicas, os métodos matemáticos desempenham um papel importante na resposta a essa necessidade.

Uma característica com particular interesse no estudo de palavras genómicas é a sua distribuição espacial ao longo de sequências de ADN, podendo esta ser caracterizada pelas distâncias entre palavras. A contagem dessas distâncias fornece distribuições discretas passíveis de análise estatística. Neste trabalho, exploramos as distâncias entre palavras como um descritor matemático das sequências de ADN, tendo como objetivo delinear e desenvolver procedimentos estatísticos especialmente concebidos para o estudo das suas distribuições.
A caracterização das distribuições de distâncias empíricas entre palavras genómicas envolve o problema do crescimento exponencial do número de distribuições com o aumento do comprimento da palavra, gerando a necessidade de redução dos dados. Além disso, se os dados puderem ser validamente agrupados em classes então os representantes de classe fornecem informação relevante sobre semelhanças e diferenças entre cada grupo de distribuições. Assim, exploramos o potencial das distribuições de distâncias na obtenção de um agrupamento de palavras, que agrupe padrões de distâncias semelhantes e que coloque em evidência as características de cada grupo. Com vista ao estudo comparativo de sequências genómicas e à definição de assinaturas de espécies, focamo-nos no desenvolvimento de modelos teóricos que descrevam distribuições de distâncias entre palavras em cenários aleatórios. Esses modelos são utilizados na definição de assinaturas genómicas, capazes de discriminar entre espécies e de recuperar relações evolutivas entre estas. Presumimos que o estudo de semelhanças e a análise de agrupamento das distribuições permite identificar palavras cuja distribuição se afasta fortemente de uma distribuição de referência ou do comportamento global das maioria das palavras. Um dos principais tópicos de investigação foca-se na deteção de distribuições com comportamentos anormais, aqui referidas como distribuições atípicas.

No contexto genómico, palavras com distribuições de distâncias atípicas poderão estar relacionadas com alguma função biológica (motivos). Esperamos que os resultados obtidos possam ser utilizados para fornecer algum tipo de classificação de sequências, identificando padrões evolutivos e permitindo a previsão das propriedades funcionais, representando assim um passo adicional na criação de conhecimento sobre sequências de ADN.

keywords Genomic word distances, Distance distributions, Dissimilarity, Clustering, Outlier detection, Pattern recognition

abstract

The investigation of DNA has been one of the most developed areas of research in this and in the last century. However, there is a long way to go to fully understand the DNA code. With the increasing of DNA sequenced data, mathematical methods play an important role in addressing the need for efficient quantitative techniques for the detection of regions of interest and overall characteristics in these sequences.

A feature of interest in the study of genomic words is their spatial distribution along a DNA sequence, which can be characterized by the distances between words. Counting such distances provides discrete distributions that may be analyzed from a statistical point of view. In this work we explore the distances between genomic words as a mathematical descriptor of DNA sequences. The main goal is to design, develop and apply statistical methods specially designed for their distributions, in order to capture information about the primary and secondary structure of DNA.

The characterization of empirical inter-word distance distributions involves the problem of the exponential increasing of the number of distributions as the word length increases, leading to the need of data reduction. Moreover, if the data can be validly clustered, the class labels may provide a meaningful description of similarities and differences between sets of distributions. Therefore, we explore the inter-word distance distributions potential to obtain a word clustering, able to highlight similar patterns of word distributions as well as summarized characteristics of each set of distributions.
With the aim of performing comparative studies between genomic sequences and defining species signatures, we deduce exact distributions of inter-word distances under random scenarios. Based on these theoretical distributions, we define genomic signatures of species able to discriminate between species and to capture their evolutionary relation. We presume that the study of distributions similarities and the clustering procedure allow identifying words whose distance distribution strongly differs from a reference distribution or from the global behaviour of the majority of the words. One of the key topics of our research focuses on the establishment of procedures that capture distance distributions with atypical behaviours, herein referred to as atypical distributions.

In the genomic context, words with an atypical distance distribution may be related with some biological function (motifs). We expect that our results may be used to provide some sort of classification of sequences, identifying evolutionary patterns and allowing for the prediction of functional properties, thereby contributing to the advancement of knowledge about DNA sequences.

Contents

List of Figures

Part I

Introduction

Chapter 1

Introduction

The investigation on DNA has been one of the most explored areas of research in the last century. Since its first description, over 60 years ago, to the first complete sequence of the human genome, much has been discovered, but there is still a long way to go to fully understand DNA.

The Human Genome Project achieved a milestone with the record of the complete sequence of the three trillion nucleotides that make up the human genome and the identification of individual genes within this sequence, in April 2003 [103]. However, genome sequencing is not an end in itself. A major challenge is yet to be achieved: to understand *what* the genome contains and *how* it works.

Mathematical methods and computer science play an important role in the development of efficient quantitative techniques, for the overall characterization of DNA sequences and for the extraction of relevant information contained in it. Understanding how the genome functions is a different endeavor. It implies to locate genes and to determine their function, which may be a very hard endeavor. While in the past the attention has been directed at the expression of single genes, now the question has become more general and relates to the expression of sets of genes, due to their interactions.

The large quantities of data produced by DNA sequencing have required the development of new methods for sequence analysis. The description and classification of sequences is heavily dependent on mathematical and statistical models. This thesis is in line with the study of genome composition, applying and proposing new statistical methodologies with this objective.

1.1 Motivation

There are two complementary approaches to the study of nucleotide sequences: structural and functional. The former starts with traditional statistical analyses of the sequences, detection of their non-randomness, search for patterns with unusual occurrence or unusual structure. After the features are found the question of what might be their functions has to be addressed.

The functional approach, by contrast, starts from the other end and asks what would be a characteristic sequence structure to serve a given biological function [195]. It is important to realize, however, that while every biological function is realized via some particular sequence structure, a sequence peculiarity might not necessarily be involved in anything of biological importance.

In both cases, the interpretation of results requires biological knowledge. Sequences with peculiar features may be a good reason for a deep study from a biological point of view. For instance, early analysis of sequences revealed a striking repetition of some dinucleotides every three steps or multiples thereof. This observation remained a peculiarity until the discovery of the so-called frameshift mutations (mutations produced by insertion or deletion of one base during DNA synthesis) [195].

This thesis is in line with the structural study of nucleotide sequences, from a mathematical point of view. Our goal is to develop new methods able to detect unusual characteristics or overall characteristics of interest of the nucleotide sequences. After features are found, biologists and experts in the area could question for “what might be their functions?”. Nevertheless, our motivation comes from the detection of sequences or features that could potentially be related to some functional element, e.g. cruciform structures (see Chapter 7).

The statistical analysis of DNA symbolic sequences may require the conversion into numerical format. Obviously, the choice of the numerical transformation of a DNA sequence affects how the mathematical properties are revealed and the capability of highlighting the biological properties of the sequence.

One of the mappings’schemes which has been of interest in the last decade is the inter-word distance distribution [5; 6; 29]. The inter-word distance is defined as the difference between the positions of the first symbol of consecutive occurrences of a word, considering that the sequence is read through a sliding window (of length equal to the word length). Procedures based on inter-word distances have already been found useful to study genomic sequences [5; 6; 29]. We propose to explore the distribution of inter-word distance distributions as a mathematical descriptor of DNA sequences.

1.2 Main objectives

The main goal of our project is to develop statistical methods for genomic data, able to capture essential information about the primary or secondary structure of DNA, using, mainly, distance distributions as input data.

After an exploratory study of the distance distributions, we will focus on the detection of unusual features. Indeed, one of the key topics of our research is the establishment of procedures that capture atypical distributions. To achieve this goal we will develop methods for the comparative study of genomic sequences, which involves obviously exploring the advantages and disadvantages of distinct dissimilarity measures.

The identification and characterization of overall features of distance distributions, for different word lengths, is also one of our objectives. This topic of study is affected by an additional problem: the possible huge number of distributions to analyse (there are 4^k words of length k). The exponential increasing of the word's number as word length increases, creates the need for organizing distributions into clusters and for dealing with a class label distribution. In a word-by-word analysis, we explore the distance distributions potential to obtain a classification of the words in groups. If the data can validly be clustered, then the class labels may provide a meaningful description of similarities and differences in the data.

Still within the scope of clustering analysis, we intend to explore the distance distributions potential to obtain genomic signatures of species.

Some authors argue the contribution of selective evolution of genomes may be highlighted by the subtraction of the random background (independent nucleotide placement assumption) from the counting result [62]. Following this view, we will explore discrepancies between real sequences and the random background to construct species signatures. By confirming that those signatures are able, not only to discriminate between species, but also to capture some of the evolutionary relation between species, a link is created between distance distribution features and their potential biological interest.

Motivated by the symmetry phenomenon (in a single strand of DNA, sufficiently long, the number of occurrences of a word is similar to that of its reversed complement) observed in several organisms, including the human genome, we question ourselves whether a somehow similar phenomenon is observable in word's distribution patterns. So, we survey similarities between (the distance distribution of) words that are reversed complements and between (the distance distribution of) words that have similar composition.

In a more functional approach, we investigate distance distributions between pairs of reversed complementary words (in contrast to the same word). In particular, we address the problem of discovering pairs of reversed complementary words both occurring at distances that are over-represented and with "clusters" of over-represented distances. The reasoning behind is related with the occurrence of cruciform structures in DNA (four-armed structures that can be formed at sites containing reversed complementary words), instead of the classical double helix structure.

Throughout this research work three major statistical concepts/methods are covered, discussed and applied, namely, similarity measures, outlier detection and cluster analysis. New methodologies are proposed, by recycling existing concepts or entailing new concepts (such as the peak dissimilarity measure), with the purpose of applying them in the study of distance distributions. We expect that our methods and results can be used to provide some sort of knowledge about DNA sequences. Nevertheless, the domain of application of the proposed methods should not be restricted to the study of genomic word distance distributions.

1.3 Thesis organization

This thesis contains ten chapters, and is organized into three parts - Part I, comprising Chapters 1-3, being mainly concerned with the introductory background; Part II, incorporating Chapters 4-9, brings together the main scientific papers developed through this journey; and Part III, consisting of Chapters 10 and 11, being primarily concerned with the critical discussion of the scientific work developed.

Chapter 1 is an introductory chapter that aims to be succinct. It presents, in a very general way, the questions that motivated this research study and the general objectives that we set ourselves to achieve.

Chapter 2 is primarily concerned with background knowledge. Some useful biological concepts are introduced, facilitating the biological contextualization of this research work and its discussion. A brief review is made on the usual statistical techniques applied in the study of genomic sequences, drawing special attention to the mathematical background on word frequencies (methods that study under- or over- representation of words) and their distribution patterns (study of waiting times). Several mapping schemes, to convert DNA sequences into numerical sequences, are summarily described, including the “inter-word distances” mapping that leads to the inter-word distance distributions, a main concept on this work.

Chapter 3 details the motivation behind each of the topics under study, pointing the state of the art, and revealing the initial questions that guided this research.

Chapters 4 – 9 present six selected research papers, resulting from the work developed during this research project. Of these, four were published in scientific reviews indexed in Scopus (Articles I, II, IV and V). Article III was published in a conference proceedings. Article VI is, to date, submitted to a refereed journal that covers the interface between the statistical and computing sciences.

Chapter 10 is mainly concerned with a critical analysis of the research work presented in the previous chapters. Results are discussed and links between the six papers are highlighted. As a concluding chapter, it points some open questions that remain to answer and some others that emerge from the developed work. Future perspectives of research are suggested herein and Chapter 11 concludes.

The scientific communications and publications arising from this research study are enumerated in an Appendix.

Chapter 2

Background: genomic information and its analysis

The first section of this chapter introduces some basic concepts from biology, intending to make a biological contextualization of the dissertation. Starting from a general description of the cell structure and its subcellular components related to the storage and synthesis of DNA, some main functions of DNA and RNA are explained. The usual double-helix structure of DNA is described, and other unusual DNA structures are referred to. The concept of genome arises naturally, unraveling DNA organization into chromosomes and how they are folded to fit inside the cell. Looking deeper into chromosomes, the relation between gene and coding region is made, as well as the distinction between intron and exon. Then, by increasing some orders of magnitude, complete genomes, their sizes and the so-called reference genome are addressed. A remark on the history of DNA discovery and its structure ends the section. Admittedly, understanding these biological concepts falls outside the scope of this thesis in applied mathematics. Nevertheless, the reading of the first section likely facilitates the understanding of the motivation behind the proposed procedures and the discussion of their potentialities.

The second section focuses on methods for DNA sequence analysis. Mapping schemes commonly used to convert DNA symbolic sequences into numerical format are described, including the inter-word distance. It is followed with a brief description of some statistical methods commonly employed in the analysis of genomic sequences, and the insights they can provide about function (gene expression), structure (folding) and evolutionary patterns (phylogenetic relatedness) of DNA sequences.

This thesis emphasizes the development of new methods able to extract genomic information from the distribution of distances between pairs of words (see Chapters 5–9) and, to a lesser extent, from the word frequencies (see Chapter 4). Thus, background knowledge about expected word frequencies and expected distributions of distances in random sequences are of great interest. The third section draws attention to models on expected word frequencies and waiting times between words in random texts.

Section 4 presents the inter-word distance distribution, a core concept in this research.

2.1 Biological concepts

The genetic information of all living organisms is stored in their DNA - the molecule that contains the instructions an organism needs to develop, live and reproduce. These instructions are found inside every cell, and are passed down from progenitors to their descendants.

DNA is the abbreviation for deoxyribonucleic acid. In a very simplified way, it can be described as a long sequence of letters – the nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T) – where words can be identified. With only four letters, the genomic book – the genome – encodes a myriad of information necessary to create a specific organism with all its particularities. The same four letters, ordered in different ways, give rise to the impressive variability between and within species existing on our planet. The size of these genomic books ranges from 144 thousands of base pairs[1] in some bacteria [131], to an astonishing 1.5×10^{11} base pairs, in the rare flower *Paris japonica* [144]. However, the size of those books does not correlate with the complexity of the organism [47]. For example, Paris japonica has a genome 50 times bigger than that of humans. The human genome has *only* 3.2 billion of base pairs.

This section introduces some basic concepts from the cell to the genome, focusing on the DNA molecule, its discovery, structure and function.

2.1.1 From cell to DNA

All living organisms share a common set of characteristics, such as being constituted by cells, need energy to survive, respond to environmental stimuli, reproduce and evolve. According to their number of cells, they can be classified as unicellular or multicellular. All species of animals and land plants are multicellular organisms, while some fungi and some algae are unicellular. In addition, the structure of the cells also divides living organisms in eukaryotic and prokaryotic.

Prokaryotic cells are cells with a relatively simple structure, having no cell nucleus nor any, meaning that the genetic material DNA is not bound within a nucleus. The prokaryotic organisms comprise two domains of the three domains of life, namely, *Archaea* and *Bacteria*. Most of them are unicellular but a few are multicellular, like some cyanobacteria.

By contrast, eukaryotic organisms have more complex cells, in which the genetic material is organized into a membrane-bound nucleus. There is a wide range of eukaryotic organisms (domain *Eukarya*), including all animals, plants and fungi, as well as most algae. Eukaryotes may be either multicellular or unicellular.

[1] Just for now, consider that a base pair is a letter in the genomic text. Later, this concept will become clear.

Viruses are conceptually separated from the remain biological entities (the three domains of life), because they cannot live or reproduce without a host organism. Without cells of other organisms, viruses are not be able to multiply. Therefore, some biologists do not consider viruses as living things. While not inside an infected cell or in the process of infecting a cell, viruses exist in the form of independent particles, that contain genetic material made from either DNA or RNA.

Eukaryotic cell A typical eukaryotic cell contains a nucleus, ribosomes and several organelles individualized by cell membranes, called organelles, in which specific activities take place. The nucleus is surrounded by a nuclear envelope which separates it from the cytoplasm, a watery jelly-like liquid where many reactions of the cell take place. The ribosome is a complex molecular machine, where proteins are made using information from the nucleus and raw materials from the cytoplasm. One example of organelle is mitochondrion, which produces cellular energy through the process of cellular respiration. Mitochondria are found in nearly all eukaryotic cells. A diagram of a typical eukaryotic cell with some of its subcellular components, is depicted in Figure 2.1.

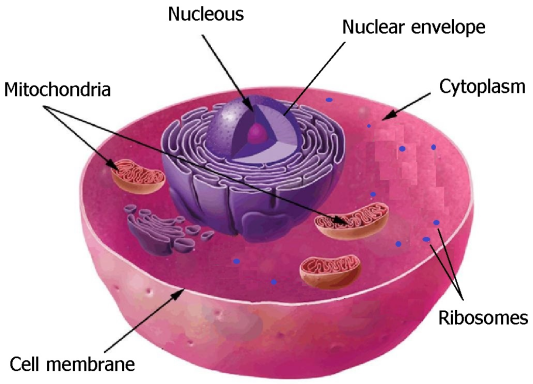


Image adapted from: Wikimedia Commons

Figure 2.1: Simplified structure of a typical animal cell revealing some of its subcellular components.

The instructions for the eukaryotic cell functioning and its genetic material are stored in nucleic acids. Two types of nucleic acids occur in cells, playing complementary roles. Deoxyribonucleic acid (DNA) acts as a carrier of genetic information, while ribonucleic acid (RNA) molecules play an important role in protein synthesis.

Most of DNA molecules of eukaryotic organisms are stored in the cell nucleus (in contrast with prokaryotic organism whose DNA floats freely around the cell). In addition, DNA is also found in some organelles: in mitochondria, present in all plant and animals cells, as well as in chloroplasts, present in plant cells. Both nuclear and organellar genomes constantly interact.

The complex interplay between these two types of DNA, environment, and lifestyle is most likely involved in phenomena such as ageing and longevity [82; 193].

RNA molecules are found inside and outside the cell nucleus, according to their function. The three main types of RNA are messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA): rRNA molecules are part of the structure of the ribosome, while the other two types of RNA are used to create new proteins.

The synthesis of proteins is preceded by the copy of information contained in DNA through the process of transcription. Transcription begins when an enzyme (the RNA polymerase enzyme) attaches to the DNA strand and starts assembling a new chain of nucleotides to produce a complementary RNA strand. In some cases, the newly created molecule is itself a finished product (tRNA, rRNA or other non-coding RNA), and it serves an important function within the cell. The major end product of transcription is mRNA, which carries information that is translated into proteins by ribosomes, as depicted in Figure 2.2.

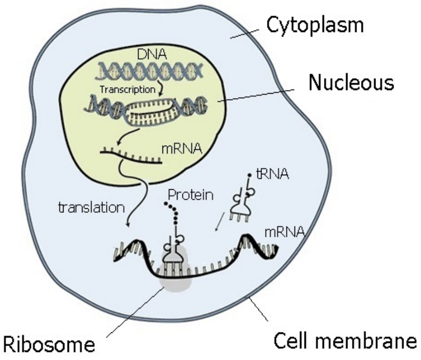


Image adapted from: National Human Genome Research institute

Figure 2.2: Processes of transcription and translation of the genetic information. RNA is the link between nuclear DNA and protein synthesis. The final product, protein, is a linked chain of amino acids.

Summarizing, all the biological information that gives rise to an individual, with all its characteristics and specificities, is embedded in its DNA. The pathways that bind DNA to its expression, through protein synthesis, are mediated by RNA.

2.1.2 DNA

The components of DNA From a chemical point of view, DNA is a macromolecule composed of small units called nucleotides. Each nucleotide is formed by a sugar (deoxyribose), a phosphate group and a nitrogenous base. DNA nucleotides assemble in chains linked by covalent bonds, formed between the sugar of one nucleotide and the phosphate group of the next. In turn, bases are attached to the sugar units, as shown in Figure 2.3. The alternating sequence of sugar-phosphate group is referred to as backbone structure.

The four nitrogenous bases found in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). They are classified according to their structure into purines and pyrimidines: adenine and guanine are purines, since they have a structure with two rings; cytosine and thymine are pyrimidines, as they have a single ring structure. The whole polynucleotide sequence is referred to as DNA chain or DNA strand.

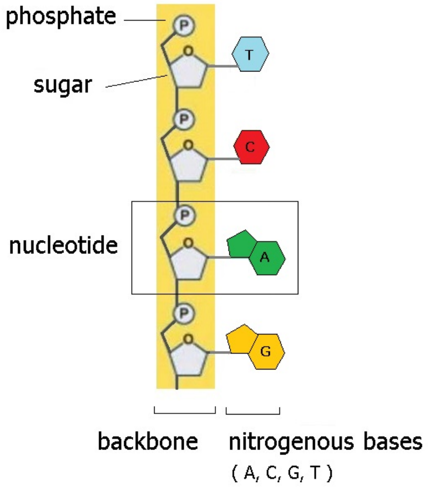


Image adapted from: Encyclopædia Britannica, Inc.

Figure 2.3: DNA strand and its block units, the nucleotides. Each nucleotide is composed of three distinctive sub-units: a sugar molecule, a nitrogenous base and one phosphate group. The alternating sequence of sugar-phosphate group is referred to as backbone structure.

It is worth noting that, the four nitrogenous bases found in DNA are not the same as those found in RNA molecules. In RNA molecules, the thymine base is substituted by the uracil (U) base. Thus, RNA sequences encloses A's, C's, G's and U's.

DNA structure The currently accepted structure of DNA molecule is the one described by James Watson and Francis Crick, in an article published in 1953 in the scientific journal Nature [206]. They described the structure of the DNA molecule as being constituted by two helical chains coiled around the same axis, where each small component of these chains, the nucleotides, would be made of a sugar, a phosphate group and one nitrogenous base. Bases are located near the center of the helix, while sugar and phosphate groups are in the outside of the helix, forming a backbone resistant to cleavage (see Figure 2.4).

The two strands of DNA are held together by hydrogen bonds that join a purine to a pyrimidine, but only specific pairs of bases can bond together: adenine always pairs with thymine (A–T), and guanine always pairs with cytosine (G–C). Since each linked pair consists of a two ring base and a one ring base, the size of each pair is almost identical. The arrangement of two nucleotides binding together on opposite complementary DNA is called a base pair, often abbreviated *bp*. Due to the nitrogenous bases complementarity, A–T and C–G, the nucleotide sequence of a strand fully determines the nucleotide sequence of the other strand.

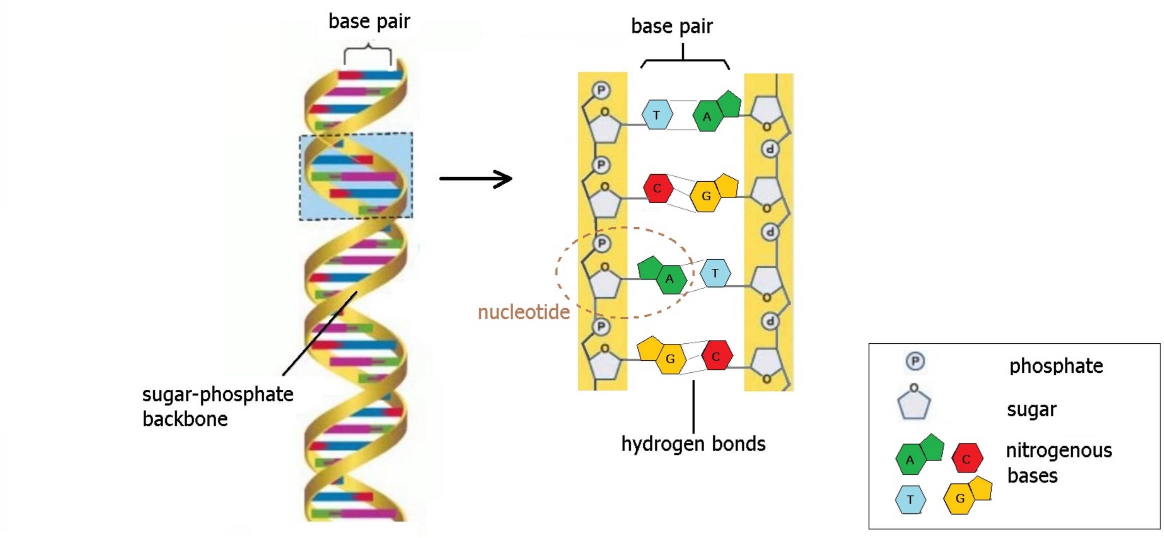


Image adapted from: Encyclopædia Britannica, Inc.

Figure 2.4: DNA double helical structure. Nucleotides from opposite strands pair according to their nitrogenous bases: A–T and G–C.

The double-helix structure physically protects the important atoms of the bases from chemical modifications [177]. The bases, located near the center of the helix, are attached to the strands by strong bonds (phosphodiester bonds), whereas the base-pairing interactions involve weak bonds (hydrogen bonds). The two types of base pairs form different numbers of hydrogen bonds. In general, A–T are linked by two hydrogen bonds and G–C are linked with three hydrogen bonds. For this reason, DNA with high GC-content (the percentage of nitrogenous bases that are either G or C) is more stable than DNA with low GC-content.

In processes involving DNA replication or transcription, such as cell duplication or protein syntheses, it is necessary to provide access to the genetic information protected in the center of the double helix. In these situations, the weak hydrogen bonds between base pairs are broken and each strand may serve as a model for the creation of new DNA or RNA molecules complementary to the original.

Coding and non-coding regions Similar to the way that the order of alphabet letters is used to form words, the order of nucleotides form genes which, in the language of the cell, dictate how to make proteins or other molecules. Usually, genes are long stretches of DNA that code for a molecule that has a function.

The view of gene as a blueprint for a protein, stated as “one gene, one protein”, is very reductive. With the development of genomics it became evident that some genes do not specify proteins; rather, the end-products are functional molecules, such as rRNA and tRNA [85]. In early 70's, scientists coined the term "junk DNA" to describe sections of genome that do not code for proteins (which means, about 98% of the human genome!), but the term has fallen out of favor. It is now clear that at least some of it is associated with the function of cells, particularly to the control of gene expression.

In multicellular organisms, like plants and mammals, the coding regions of most genes are not continuous. Rather, genes are split into several coding regions, called *exons*, in between non-coding regions called *introns*. Exons carry the code for the production of proteins, they are called protein-coding regions. Introns correspond to those stretches of DNA that are transcribed (into mRNA), but spliced out before the translation into protein (see Figure 2.5). Although they are removed before a protein is made, it appears that introns inclose regulatory elements [129; 149].

The parts of the genome that lie between genes, intergenic regions, are also considered as noncoding DNA. Most large genomes are filled with intergenic regions and the bulk of it is made up of repeated sequences [41].

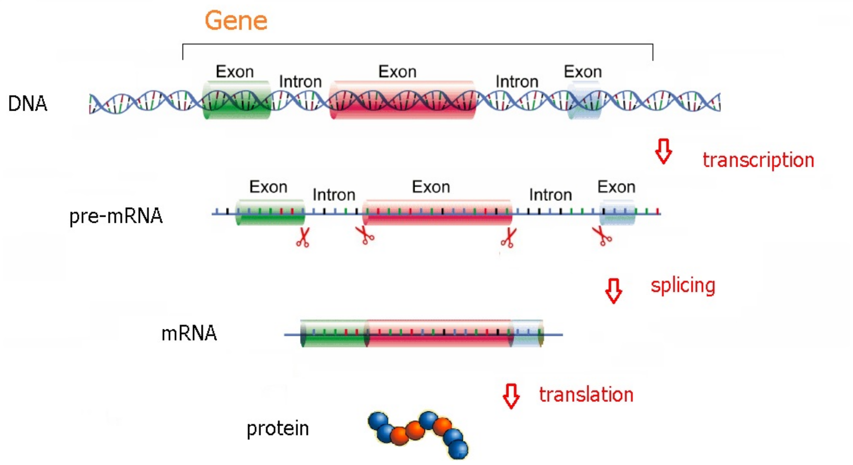


Image adapted from: National Human Genome Research Institute

Figure 2.5: Gene, introns and exons. Representation of a gene composed by three coding regions (exons) and two non-coding regions (introns), which are spliced out before translation.

The amount of non-coding DNA varies greatly among species. Often, only a small percentage of the genome is responsible for coding proteins, but a rising percentage of it is being shown to have regulatory functions. Because of the number of new discoveries resulting from genome research over the last few years, it is a mistake pointing out that any part of the genome should be unimportant simply because we do not currently know what its function might be. The expression "junk DNA" should be totally removed from the lexicon.

The genetic code A typical protein is made up of several hundred amino acids, linked together in chains. These molecules do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs. For instance, *ferritin* protein has the function of transport and storage of some molecules; *immunoglobulin G* act as an antibody; and *keratin* is a structural material making up hair and nails.

The *Central Dogma* of molecular biology is an explanation of the flow of genetic information between DNA, RNA and proteins [56; 57]. In particular, the direction in which information from genes flows to the formation of proteins can be performed by the two-step process, transcription and translation, previously mentioned: DNA → RNA → protein (see Figures 2.2 and 2.5). Messenger RNA, whose sequence originates from DNA, is decoded to produce a specific amino acid chain. This chain later folds into an active protein and performs its functions in the cell.

In the translation process, mRNA molecules are read in sequences of three nucleotides, called *codons*. Each nucleotide triplet codes for a single amino acid, e.g. the codon GUA is translated into the amino acid *valine*. There are 64 different codons ($4^3 = 64$) and only 20 distinct amino acids are used in the construction of proteins. Consequently, more than one codon can be used to specify a particular amino acid. For example, codons GUC, GUG and GUU all specify the same amino acid. This is why the genetic code is said to be degenerated.

The set of rules that define how the nucleotide triplets are translated into amino acids is called the genetic code (Figure 2.6). This code is configured in a very sophisticated way, minimizing the effect of some copying errors. For instance, any error in the third base of GUU, GUC, GUA, GUG will still result in the correct amino acid valine. Similarly, errors in codons specifying other amino acids will often (but no always) still result in the correct amino acid being used. Moreover, even if an error results in an incorrect amino acid being selected, the one selected will often have similar physico-chemical properties and is likely to be a good substitute. For example, the codon GUG, which specifies the amino acid valine, is hydrophobic. An error resulting in its second letter changing it to C leads to codon GCG which specifies *alanine*, and an error resulting in changing the first letter of valine into C leads to codon CUG which specifies *leucine*. Both alanine and leucine are also hydrophobic amino acids.

Translation starts with a chain-initiation codon or start codon, which also allows coding an amino acid (the most common start codon is AUG). Often this first amino acid will be removed in later processing of the protein. There are also stop codons, which signal that the translation should stop. Depending on where the reading starts, the same mRNA molecule can be read in multiple ways. For example, if the base sequence is CC$\underline{\text{AUG}}$CA$\underline{\text{AUG}}$GA$\underline{\text{AUG}}$UUGGC, reading could start from the first AUG and there will be five codons more. If reading starts at the second AUG, the string will have three other codons, and so on, as depicted in Figure 2.7. Mutations that disrupt the reading frame sequence by insertions or deletions of a non-multiple

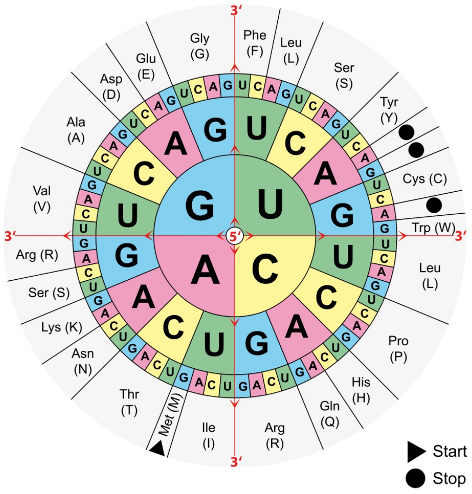


Image credits: Wikimedia Commons

Figure 2.6: Representation of the genetic code. In this mapping scheme between codons and amino acids, the codons should be read from the innermost to the outermost ring. Stop codons do not code amino acids.

of three nucleotide bases usually result in a completely different translation from the original, and likely cause a stop codon to be read, which truncates the protein.

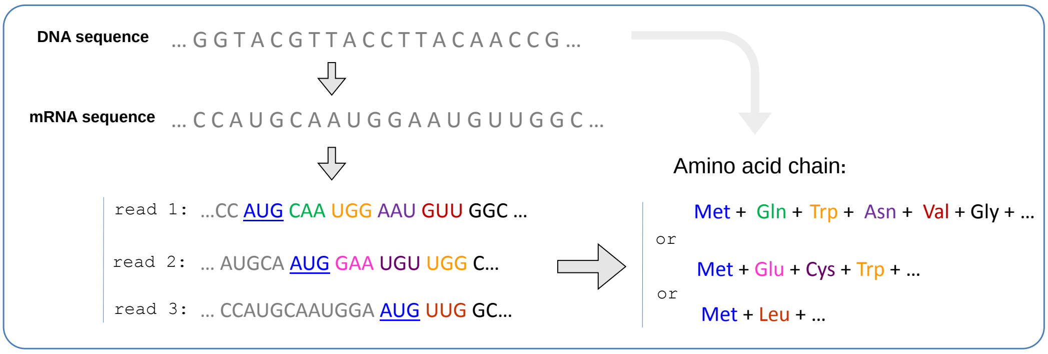

DNA sequence
... G G T A C G T T A C C T T A C A A C C G ...
mRNA sequence
... C C A U G C A A U G G A A U G U U G G C ...
read 1: ...CC AUG CAA UGG AAU GUU GGC ...
read 2: ... AUGCA AUG GAA UGU UGG C...
read 3: ... CCAUGCAAUGGA AUG UUG GC...
Amino acid chain:
Met + Gln + Trp + Asn + Val + Gly + ...
or
Met + Glu + Cys + Trp + ...
or
Met + Leu + ...

Figure 2.7: Reading frames. Illustration of three readings of the same mRNA sequence, and corresponding amino acid chains. Start codon sets the reading frame.

By choosing which amino acids are placed in which positions along the chain, different proteins can be made, each having very different functions. In that sense, the sequence of DNA nucleotides is like a language, where different combinations of letters have different meanings. Nevertheless, as depicted in Figure 2.7, a very peculiar feature distinguishes DNA language from natural languages: the same sequence of letters can be translated into distinct messages!

Primary, Secondary and High-order structures Under the current knowledge, it is accepted that the DNA molecule carries more information than the linear combination of its nucleotides. For instance, the three-dimensional structure that DNA assumes at each moment, which is not the same throughout the whole molecule, affects its replication and transcription.

The specific order of the four nucleotides in a DNA molecule is referred to as its primary structure. When a DNA strand forms a duplex by pairing with a complementary single strand it adopts another level of organization.

Structurally, DNA is a flexible molecule, and this flexibility is well pronounced in its polymorphic nature. The secondary structure of DNA molecules refers to the interactions between base pairs, close to each other, within a single or more DNA strands [27]. Such interactions are commonly represented by a structure graph, which puts in evidence the nucleotide sequence in each strand, and the base pairs interactions (see Figure 2.8). Tertiary structure refers to the final form the molecule takes, once the different secondary structures have all folded into a three dimensional structure, which involves steric relationship of bases. Secondary and tertiary structures are stabilized by a number of factors like hydrogen bonding, ionic interactions, etc.

Most DNA duplexes are generally considered to adopt the classical conformation described by Watson and Crick, known as B-DNA. Apart from this double helical structure, they can also adopt unusual structures like cruciforms, triplexes, quadruplexes, junctions, etc. [106]. Such structures are obtained in the presence of particular nucleotide sequences or through interactions with various proteins. For instance, a cruciform structure only arises from a combination of supercoiling and inverted repeats. An inverted repeat is a sequence of nucleotides followed downstream by its reverse complement. Cruciforms are formed when interstrand base pairing, in duplex DNA, with inverted repeats convert to intrastrand base pairing. A schematic representation of base pairs interactions in a cruciform structure formation and its tridimensional shape, is shown in Figure 2.8.

Investigation points that the DNA shape plays a crucial role in gene regulation, genome organization and integrity. Some unusual DNA conformations are hypothesized, or even known, to have functional roles in living organisms [34; 106; 110; 166; 213]. The formation of such unusual structures strongly depends on the DNA nucleotides sequence, also referred to as structural sequence motifs [95].

DNA discovery This introductory description of the “molecule of life” could not end without mentioning the credits for DNA discovery and for its structure’s discovery.

The history of DNA discovery dates back more than 150 years ago. The now called DNA was first identified inside the nuclei of human white blood cells, in 1869, by Friedrich Miescher, calling it “nuclein”. In 1866, Gregor Mendel described the actions of invisible factors (now called genes) in predictably of traits of an organism, based on his investigation of inheritance patterns in pea plants. However, Mendel’s breakthroughs were not recognized until more than three decades later. Early in the 20th century, Phoebus Levene was the first to characterize

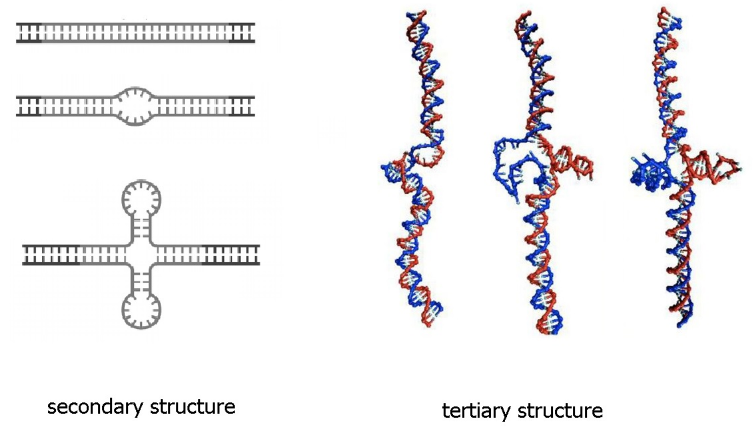


Image adapted from: Microbiology and Molecular Biology Reviews, 2010 [32]
and Proc. World Congress of the International Union of Physiological Sciences, 2013 [130]

Figure 2.8: Unusual DNA structure. Schematic representation of a cruciform structure formation, from a combination of supercoiling and inverted repeats. Cruciform base pairs interactions (left) and three-dimensional shape (right).

the different forms of nucleic acids, DNA and RNA, and to discover the order of the three major components of a single nucleotide (phosphate-sugar-base). In 1950, Erwin Chargaff ascertained that almost all DNA maintains a certain proportion between nucleotides. In particular, the amount of adenine is similar to the amount of thymine, and the amount of guanine is similar to that of cytosine. This conclusion is now known as *Chargaff's parity rule*. In 1952, the X-ray diffraction images produced by Rosalind Franklin suggested that DNA was a double helix [153]. Right after, in 1953, Watson and Crick brought together the scientific foundation provided by these pioneers to assemble their groundbreaking conclusion about the structure of the DNA molecule: a 3D double helix. For a nice review about the history of genetics and DNA discovery see [115; 153; 205].

2.1.3 Genomes

What is a genome? The genome is often described as the entire hereditary information of a living organism. It is encoded in each cell of the organism, through DNA molecules, and provides the basic code that tells each cell how to grow, function, and reproduce. The genome includes both the genes and the non-coding sequences of the DNA.

Most eukaryotic cells are endowed with different types of DNA, namely, nuclear DNA and mitochondrial DNA (plus chloroplast DNA, in plants). For this reason, mitochondria are said to have their own genome, often referred to as the "mitochondrial genome". In relation to eukaryotic organisms, the phrase "genome" is commonly used to refer only to the nuclear genome, whereas the remaining genomes are designated by their specific name. Genomes are organized into discrete structures, tightly packaged in a complex series of coils and loops, called chromosomes. The chromosomes vary widely between different organisms, and different types of chromosomes may be found within the same cell. Eukaryotic cells have several linear chromosomes and they are larger than that of prokaryotic cells. In contrast, the genetic material in prokaryotic cells typically exists in the form of a small circular chromosome.

Genome organization in eukaryotes In addition to nucleus chromosomes, eukaryotic organisms may also carry extra genetic elements such as mitochondrial chromosomes. Being located in different regions of the cell, nucleus genome and mitochondrial genome have different structural and functional properties.

Nucleus chromosomes are often observed and depicted as X-shaped structures, during the process of cell division. They present multiple levels of packaging that are only possible due to the existence of special proteins termed histones. DNA packaging is crucial, because it allows that large amounts of DNA are able to fit nicely in a cell that is many times smaller. The packaging process can be described, in a very simplified way, as: DNA wraps around histones, which form loops of DNA called nucleosomes; these nucleosomes coil and stack together to form fibers, called chromatin; and chromatin, in turn, forms larger loops and coils to form chromosomes (Figure 2.9). As a result, chromatin can be packaged into a much smaller volume than DNA alone [21].

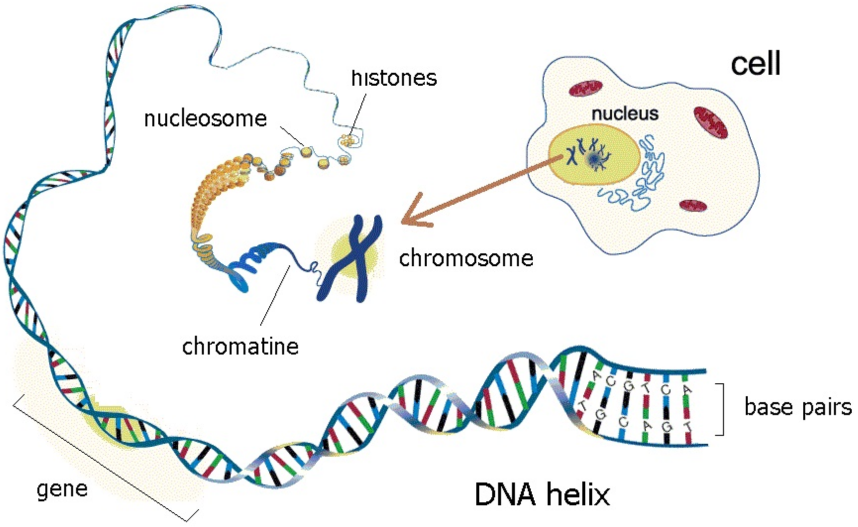


Image adapted from: National Human Genome Research Institute

Figure 2.9: **Organization of DNA in an eukaryotic cell.** Most DNA is found inside the nucleus of a cell, where, together with histone proteins, it forms the chromosomes.

The mitochondrial DNA (mtDNA) is only a small portion of the total DNA of a eukaryotic cell, and its description falls outside the scope of this work. However, it is noteworthy presenting some characteristics in which they differ most: mtDNA has usually a circular structure, whereas nuclear DNA has a linear open-ended structure; in most species mtDNA is solely inherited from the mother; this extranuclear DNA is prone to mutations which are not inherited, due to the lack of error checking capability that nuclear DNA has, thereafter mtDNA molecules can differ from one another within a single cell.

Apart from the efficient wrapping and protection that chromosomes structure provide to DNA, they are also highly dynamic, allowing the cell access to the genetic information stored in the center of the double helix.

Genome sizes A complete genome sequence lists the order of every DNA base that make up all the chromosomes of a organism, forming very long sequences of A's, C's, G's and T's. The genome size of an organism is usually expressed in terms of the number of base pairs in one copy of each chromossome.

Genome sizes are well known to vary enormously among species, and are poorly correlated with the organism complexity. On average, genome sizes are significantly larger in eukaryotic than in prokaryotic organisms. Among the eukaryotes, fungi have relatively small genome sizes when compared to those of animals and plants. Some examples of genome sizes are highlighted in Figure 2.10.

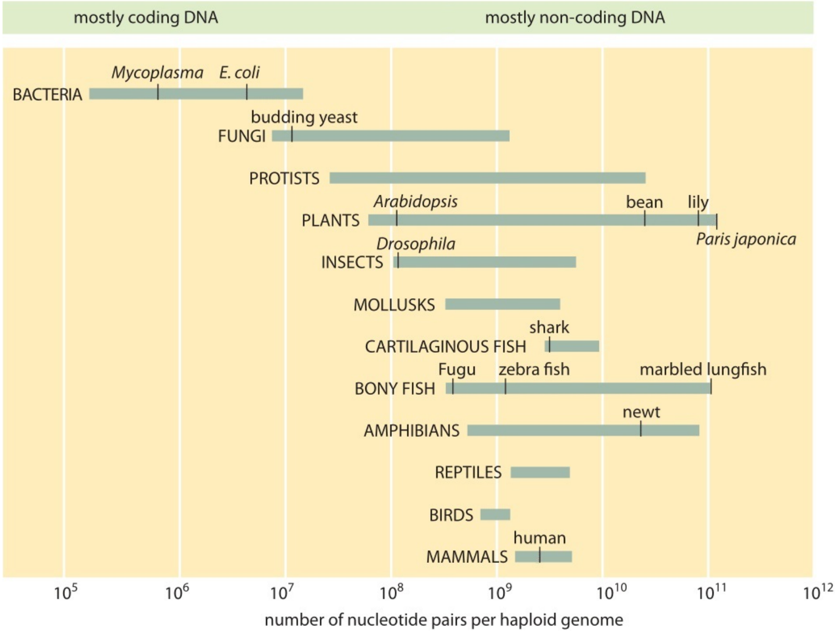
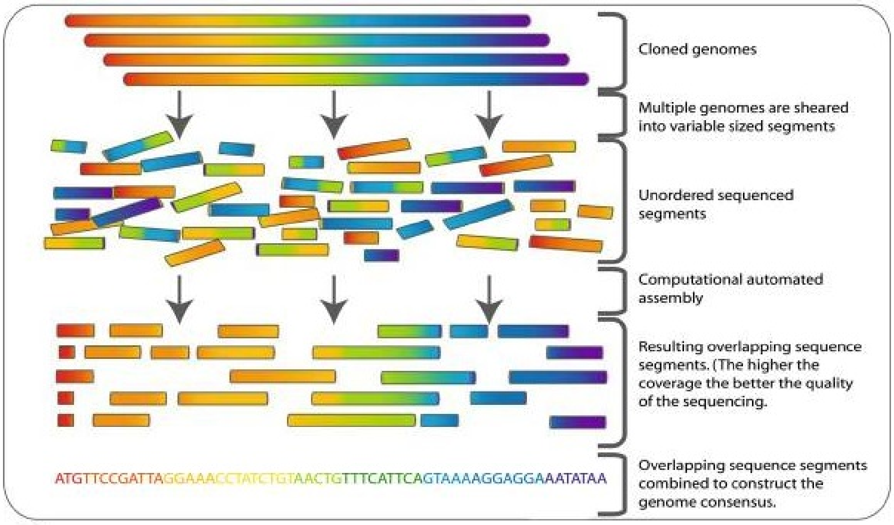


Image from: BioNumbers database (http://www.bionumbers.hms.harvard.edu)

Figure 2.10: Genome sizes. The range of genome sizes runs from 0.14 Mbp for bacterium *Carsonella ruddii*, to 150 Gbp for plant *Paris japonica*.

Sequencing genomes The development of new technologies has made genome sequencing dramatically cheaper and easier, and therefore the number of complete genome sequences is growing rapidly. Within a species, the vast majority of the genomic sequences are identical between individuals, but sequencing multiple individuals is necessary to understand the genetic diversity of a species. Comparative analysis of genomes provides an unprecedented opportunity to investigate what makes a given species unique.

One of the major challenges in sequencing eukaryotic chromosomes is their size, making it impossible to obtain an uninterrupted chromosome sequence end-to-end, using current technology. To overcome this problem, researchers "cut" several cloned chromosomes into small segments that can be read at once. Then, the DNA reads are aligned by matching the overlapping parts. The reconstruction of chromosome sequences can be compared to the assembly of a large puzzle, whose overlapping fragments give clues about their joining; nonetheless, to arrive at the correct solution of the puzzle is a hard task. Figure 2.11 provides a snapshot of a commonly assembly process schematized by Commins and others [55].

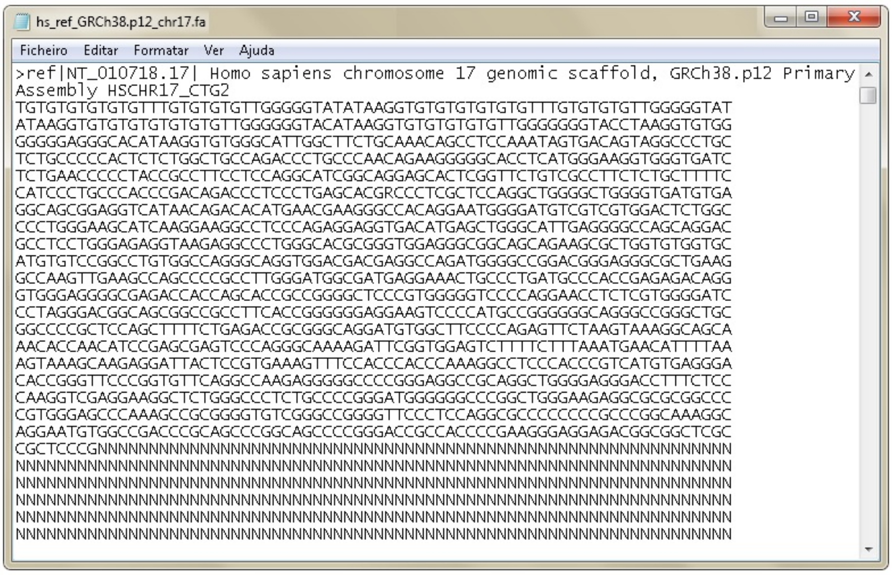

Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequence segments combined to construct the genome consensus.

Source: Commins et al. (2009) Biological procedures online 11(1), 52.

Figure 2.11: **Assembly process.** A series of different cuts is used to generate overlapping DNA fragments. The sequence reads are assembled into a series of overlapping fragments, to generate the complete sequence of a chromosome.

Due to uncertainties in the position identification of read sequences, genome assembly may be punctuated by gaps. Those regions are annotated with an "N" instead of the normal bases (A, C , G or T), meaning that any possible base could be found on that region. The insertion of N's allows obtaining a continuous record for an incompletely assembled chromosome, while also clearly shows the failure to assemble specific regions. Figure 2.12 shows a small portion of a DNA sequence with both recognized and unspecified segments, from human chromosome 17.

Genomes are now being sequenced at such a rapid rate that it is becoming routine. With the advent of new techniques in DNA analysis, scientists are able to look at the chromosome


hs_ref_GRCh38.p12_chr17.fa
Ficheiro Editar Formatar Ver Ajuda

```
>ref|NT_010718.17| Homo sapiens chromosome 17 genomic scaffold, GRCh38.p12 Primary
Assembly HSCHR17_CTG2
TGTGTGTGTGTGTTTGTGTGTGTTGGGGGTATATAAGGTGTGTGTGTGTGTTTGTGTGTGTTGGGGGTAT
ATAAGGTGTGTGTGTGTGTGTTGGGGGGTACATAAGGTGTGTGTGTGTTGGGGGGGTACCTAAGGTGTGG
GGGGGAGGGCACATAAGGTGTGGGCATTGGCTTCTGCAAACAGCCTCCAAATAGTGACAGTAGGCCCTGC
TCTGCCCCCACTCTCTGGCTGCCAGACCCTGCCCAACAGAAGGGGGCACCTCATGGGAAGGTGGGTGATC
TCTGAACCCCCTACCGCCTTCCTCCAGGCATCGGCAGGAGCACTCGGTTCTGTCGCCTTCTCTGCTTTTC
CATCCCTGCCCACCCGACAGACCCTCCCTGAGCACGRCCCTCGCTCCAGGCTGGGGCTGGGGTGATGTGA
GGCAGCGGAGGTCATAACAGACACATGAACGAAGGGCCACAGGAATGGGGATGTCGTCGTGGACTCTGGC
CCCTGGGAAGCATCAAGGAAGGCCTCCCAGAGGAGGTGACATGAGCTGGGCATTGAGGGGCCAGCAGGAC
GCCTCCTGGGAGAGGTAAGAGGCCCTGGGCACGCGGGTGGAGGGCGGCAGCAGAAGCGCTGGTGTGGTGC
ATGTGTCCGGCCTGTGGCCAGGGCAGGTGGACGACGAGGCCAGATGGGGCCGGACGGGAGGGCGCTGAAG
GCCAAGTTGAAGCCAGCCCCGCCTTGGGATGGCGATGAGGAAACTGCCCTGATGCCCACCGAGAGACAGG
GTGGGAGGGGCGAGACCACCAGCACCGCCGGGGCTCCCGTGGGGGTCCCCAGGAACCTCTCGTGGGGATC
CCTAGGGACGGCAGCGGCCGCCTTCACCGGGGGGAGGAAGTCCCCATGCCGGGGGGCAGGGCCGGGCTGC
GGCCCCGCTCCAGCTTTTCTGAGACCGCGGGCAGGATGTGGCTTCCCCAGAGTTCTAAGTAAAGGCAGCA
AACACCAACATCCGAGCGAGTCCCAGGGCAAAAGATTCGGTGGAGTCTTTTCTTTAAATGAACATTTTAA
AGTAAAGCAAGAGGATTACTCCGTGAAAGTTTCCACCCACCCAAAGGCCTCCCACCCGTCATGTGAGGGA
CACCGGGTTCCCGGTGTTCAGGCCAAGAGGGGGCCCCGGGAGGCCGCAGGCTGGGGAGGGACCTTTCTCC
CAAGGTCGAGGAAGGCTCTGGGCCCTCTGCCCCGGGATGGGGGGCCCGGCTGGGAAGAGGCGCGCGGCCC
CGTGGGAGCCCAAAGCCGCGGGGTGTCGGGCCGGGGTTCCCTCCAGGCGCCCCCCCCGCCCGGCAAAGGC
AGGAATGTGGCCGACCCGCAGCCCGGCAGCCCCGGGACCGCCACCCCGAAGGGAGGAGACGGCGGCTCGC
CGCTCCCGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Figure 2.12: Fragment of a DNA sequence with a gap - human chromosome 17 from the GRCh37.p13 assembly.

in much greater detail. There are several repositories that collect genome sequences, others annotate and analyze them, and provide public access to that information. The most well known nucleotide sequence database is GenBank, maintained by the National Center for Biotechnology Information, USA. The other two main databases are the European Nucleotide Archive, from the European Bioinformatics Institute, and the DNA Data Bank of Japan, from the National Institute of Genetics.

Human Genome Human genetic material stored in the cellular nucleus is organized into 23 pairs of chromosomes that vary widely in size and shape. The set of chromosomes is formed by 22 pairs of autosomes and a 23rd pair called allosome.

Autosomes are labeled with numbers from 1 to 22, roughly in decreasing order of their sizes. In each pair of chromosomes, one inherited from the mother and one from the father, the chromosomes are homologous, i.e. they are similar in length and gene's position. The 23rd pair of chromosomes are two sex-determining chromosomes, X and Y. Usually, females carry two X chromosomes (XX), whereas males have one X and one Y chromosome (XY). The X chromosome is significantly longer than the Y chromosome.

Each chromosome is a very long molecule. The shortest human chromosome has around 50 000 000 nucleotides in length and the longest 260 000 000 nucleotides (human mitochondrial DNA has a modest number of 16 000 base pairs).

Most human cells contain paired chromosomes (diploid cells). By contrast, germ line cells, which go on to produce egg or sperm cells, are called haploid because they contain half the chromosomes of diploid cells, i.e. only 23 unpaired chromosomes.

The human (haploid) genome contains approximately 3.2×10^9 base pairs. This generous number refers to the total amount of base pairs confined in a single set of chromosomes. However, most of human cells are diploid. They contain around 6 billion base pairs, each one 0.34 nanometers long, that makes a total of approximately 2 meters of DNA if stretched end-to-end; yet the nucleus of a human cell is only about 6 micrometres in diameter [18]. DNA packaging inside nucleus is an awesome process, allowing these enormously long DNA strands to fit perfectly into a cell.

Reference human genome In February 2001, rough drafts of the human genome sequence were published simultaneously by two independent groups in separated scientific papers. This landmark was handled by the Human Genome Project, a group of publicly funded researchers, and by the Celera Genomics, private company, using different approaches and independent data sets [102; 199]. In 2003, exactly 50 years after the description of the double helix, the final sequencing of the human genome was announced by the Human Genome Project [103]. The project aimed to assembly a representative example of the nucleotide sequences in an human genome.

Mapping the *reference human genome* involved sequencing a small number of different donors, and then assembling these together to get a consensus built for each chromosome. Therefore, the reference genome, commonly called assembly or build, provides a good approximation of the DNA of any single individual, but does not represent any one individually.

Since the completion of the first human genome reference, successive builds have been released. To distinguish between them, builds of each species are released with an identifier name, and subsequent assemblies maintain the identifier name increasing the code number. For instance, the latest build of the human reference genome is officially named GRCh38, which replaced GRCh37. The identifier name of the human reference genome comes from Genome Research Consortium human build 38, also termed as hg38.

Updating human reference genome The landmark sequencing of the human genome [103] surprised many with the small number of protein-coding genes that sequence annotators could identify. It revealed that humans have only 20,000–25,000 protein-coding genes. That estimate has continued to fall. Humans actually seem to have as few as 19,000 such genes, which correspond to a mere 1% of the genome size [66].

So, what is the rest of our DNA doing? The key to human complexity lies in how these genes are regulated by the remaining 99% of DNA. A natural follow-up research to the sequencing of the human genome was to identify all functional elements in the DNA, especially of those 99% which had non-protein-coding functions.

The ENCODE (Encyclopedia of DNA Elements) project, launched in 2003, has looked deeper into this non-protein-coding DNA than ever before, leading to the evidence that at least 80 percent of the genome is biologically active, with much DNA regulating nearby genes. Some recent studies have indicated that a majority of DNA regions associated with a particular phenotypic trait, including ones that contribute to human diseases, lie outside protein-coding regions [108; 172]. Many of the noncoding sequences are repeated elements; for example, nearly half of the human genome is covered by repeats [194]. Although some repeats appear to be nonfunctional, others have played a part in human evolution [40].

Another landmark project was launched in 2008 with the objective of establishing a most detailed catalog of human genetic variation: 1000 Genomes Project. By reconstructing the sequence of 2,504 genomes (1,092 in its first phase) from different populations around the world, it described human genetic variation and pointed out its implications for common diseases studies [1; 2].

Since the completion of the Human Genome Project, describing "linear" genome sequences, the human reference genome has continued to be updated and refined by the Genome Reference Consortium [83]. When genome sequencing initially started it was thought that the predominant form of variation was single nucleotide that occur at specific positions – single nucleotide polymorphisms (or SNP). However, subsequent research, such as the 1000 Genomes Project, showed that large-scale structural variations are more prevalent than originally thought [2; 54; 104].

Recognizing that some highly polymorphic regions of the genome were insufficiently represented in a single reference sequence, led to the need for a reference that includes common variation. To better represent human diversity, a new graph-like assembly model was introduced starting with GRCh37 (released in Feb. 2009), which includes three regions (called loci) with several alternate sequences. The most recent assembly, GRCh38 (released in Dec. 2013), has a total of 178 loci. We might picture these builds as a primary assembly, anchored with alternative sequences in regions with complex structural variation [53].

The evolution of the assembly models, with the interaction of several scientific fields, allows improving our understanding of genomic architecture and its impacts in human development and disease [54].

2.2 Analysis of genomic sequences

The goal of completely sequencing the human genome sparked a race in the development of more efficient sequencing techniques (in accuracy and speed). Improved sequencing techniques coupled with a continued decline in cost, led to increased throughput of genomic sequenced data in the last fifteen years. It has motivated and fostered the development of new techniques for rapid viewing and analysis of the data.

Because DNA sequences contain a large and complex amount of information, the purpose of holistically understanding its working requires skills from many diverse fields. Some of the disciplines that emerged, and have grown, in the last years are: genome annotation, which marks biological features in a DNA sequence; comparative genomics, that establish relationships between genes in otherwise unrelated organisms; and structural bioinformatics, that predicts the three-dimensional structure of biological macromolecules from the genetic sequence; just to name a few.

To handle the large quantities of data and probe the complex dynamics observed in genomic data, computers have become indispensable. Bioinformatics is an interdisciplinary science that combines biology and computational sciences, in order to conceptualize biology in terms of the physical-chemistry structures of macromolecules (such as proteins, RNA and DNA). Apart from dealing with aspects of data management (storage and retrieval), it applies computational techniques to analyze, and interpret the information in those biomolecules [122]. Such computational techniques are derived from computer science, information science, applied mathematics and statistics.

Statistics being the bulk science of this thesis, and genomic sequences the data around which the proposed methods are developed, this section is dedicated to DNA mappings and statistical techniques commonly used in the analysis of genomic sequences.

2.2.1 Numerical mapping of DNA

To overcome restrictions imposed on genomic sequences analysis due to their symbolic nature, several mapping schemes are found in the literature. Most of them convert DNA sequences onto numerical or vector sequences. However, there is a wide variety of DNA mappings, ranging from conventional integer sequences to graphical representations, as well as digital signal processing approaches, just to name a few. Some of the mappings used in DNA processing do not have a simple numerical interpretation and others do not have biological motivation. Obviously, the choice of the numerical transformation of a DNA sequence affects how the mathematical properties are revealed, and the capability of highlighting the biological properties of the sequence.

The mapping of DNA sequences into numerical sequences is done by assigning a numeral to each of the four nucleotides that compose the DNA sequence. The basic idea behind numerical characterization is that specific gene sequences are generally unique and, therefore, possess a characteristic signature in its composition and in the nucleotides distribution that compose it. The analysis of these numerical sequences, when proper transformations are used, often sheds light on the properties of the original symbolic sequences.

For mathematical purposes, any DNA sequence may be thought of as a sequence of integers in the set $\{1,2,3,4\}$. A transformation of this kind was used by Tsonis and Tsonis [196] in an attempt to clarify differences between coding and non coding sequences. The researchers applied Fourier analysis over transformed coding (using $A = 1$, $G = 2$, $C = 3$, $T = 4$) coding, non-coding and random sequences, confirming that DNA coding sequences exhibits a three-base periodicity. The conversion of symbolic sequences into numerical sequences is, clearly, not unique. Typically, mappings have a biological interpretation, and should preserve the specific structure of the DNA sequence under study. For instance, the complex number mapping suggested by Anastassiou [20] reflects the complementary nature of A–T and C–G pairs, as $\mathrm{A} = 1+i$, $\mathrm{T} = 1-i$, $\mathrm{C} = -1+i$ and $\mathrm{G} = -1-i$, where i denotes the the imaginary number. Via this assignment, the palindromes yield conjugate symmetric numerical sequences that have interesting mathematical properties. Both representations transform the DNA sequence into a single numerical sequence.

One of the most popular DNA numerical mapping is the Voss representation [203]. In this mapping scheme four indicator sequences are considered to describe the presence or absence of a specific nucleotide, with 1 indicating the presence of a nucleotide and 0 its absence. So, a single DNA sequence is transformed into four binary sequences (see Table 2.1). This representation has been widely used in many signal processing based methods, for finding hidden periodicity in DNA sequences.

Also, graphical methods has been proposed to visually map DNA sequences, using biophysical and biochemical properties of DNA molecules. Correlations, clustering of repeats, palindromes and GC–skews can be spotted using those visualization approaches. For instance, a one dimensional DNA walk, in which a step is taken upwards if the nucleotide is pyrimidine (C or T) or downwards if it is purine (A or G), was used to study long-range correlations [145]. This simple plot summarizes the relative occurrences of purine and pyrimidine nucleotides along a DNA sequence. Variations of purine/pyrimidine and variations in strong/weak bonds are also visualized in the Z-curve [211], a three-dimensional representation of genomic sequences. The tetrahedron mapping, in which the four nucleotides are assigned to the four corners of a regular tetrahedron, was used to study DNA periodicity [176]. The main application of tetrahedron representation is obtaining spectrograms of DNA sequences, that can be used to locate repeating DNA sections. Five distinct numerical representations of the short DNA sequence X = AACTGT are presented on Table 2.1 .

Another kind of DNA maps are statistical properties based, like the distance between nucleotides which provides a DNA numerical methodology to explore the correlation structure of DNA. The inter-nucleotide distances, introduced by Nair and Mahalakshmi [136], converts any DNA sequence into a unique integer sequence with the same length, where each base symbol is replaced by the distance to the next occurrence of the same symbol. In case such a symbol is not found then the in sequence value of that base is the length of the remaining sequence. The same authors also proposed the binucleotide distance representation, in which every base is replaced by the distance to the next occurrence of its complementary base [135].

Several other mapping schemes have been proposed and, as expected, no representation

Table 2.1: Some numerical representations of a short DNA sequence.

Name	Numeric representation of X = AACTGT	Reference
Integer	$S = (1, 1, 3, 4, 2, 4)$	[196]
Complex	$S = (1 + i, 1 + i, -1 + i, 1 - i, -1 - i, 1 - i)$	[20]
Voss	$S_A = (1, 1, 0, 0, 0, 0)$ $S_C = (0, 0, 1, 0, 0, 0)$ $S_G = (0, 0, 0, 0, 1, 0)$ $S_T = (0, 0, 0, 1, 0, 1)$	[203]
DNA walk	$S = (-1, -1, 1, 1, -1, 1)$	[145]
Z-curve	A tridimensional curve linking points given by $S_X = (1, 2, 1, 0, 1, 0)$ $S_Y = (1, 2, 3, 2, 1, 0)$ $S_Z = (1, 2, 1, 2, 1, 2)$	[211]
inter-nucleotide distance	$S = (1, 4, 3, 2, 1, 0)$	[136]
binucleotide distance	$S = (3, 2, 2, 3, 3, 1, 1, 0)$	[135]

can be considered as the "gold standard". The literature on this topic is vast. For a review of some of existing encoding schemes for genomic data representation, see e.g. references [7; 17; 23; 113; 168; 209].

The nucleotide sequence databases, as GenBank, made DNA sequences available in symbolic format. Prior to applying the statistical analysis techniques, mapping of DNA alphabet into numerical sequences is necessary using approaches, as those described above. Some statistical techniques widely used in a DNA analysis are referred to in the next subsection.

2.2.2 DNA as a statistical universe

In common language (not statistical) the word "universe" can be used with a figurative purpose. In such case, it is interesting to regard DNA as an "universe" given the enormous number of studies developed around it, and of methodologies proposed for its analysis.

In statistics, the universe (or population) represents the entire group of units which is the focus of the study. The universe is called finite or countable if it is possible to count its individuals, and is called infinite if it is not possible to count the units comprised therein. A statistical population can also be a conceptualized as hypothetical infinite population conceived as a generalization from experiences [123]. From this point of view, the set of "all" DNA sequences is an infinite statistical universe. The use of powerful statistical methods, provided by computer science algorithms, improve the likelihood of understanding the quaternary code printed in DNA sequences, and allows for the biological interpretation of the collected data.

Statistical analysis of DNA sequences is a prime procedure, immediately after the obtention of the structured assembly of the sequencing. The following are examples of common initial analysis: characterization of nucleotide frequencies, of chemical families (purines and pyrimidines), of codons, or of specific patterns of biological interest (motifs); analysis of the distribution of nucleotides or words of interest [147]; characterization of correlations [31; 105; 119] and periodicities [80]; the GC–content, i.e. the percentage of nucleotides that are either guanine or cytosine [93; 142] and statistical modeling [154]; statistical differentiation between coding and non-coding regions [112]. Many applications resort to statistical methodologies combined with other methodologies of applied mathematics, such as Information Theory (mutual information, entropy), stochastic processes (Markov chains), or spectral analysis (Fourier transform), among others.

The works of Afreixo [3] and Deusdado [61] are excellent references, in Portuguese language, for an overview of the large number of mathematical methodologies that have been commonly applied in DNA sequence analysis. Some of them are referred below.

Linguistic analysis Quantitative linguistic analysis for DNA sequences has been studied since the early days of biological sequencing. The goal is to obtain an indication of the relative frequency or magnitude of a linguistic form [158]. Considering that DNA sequences represent texts of a largely unknown language, and defining words (oligonucleotides) as strings with strong internal correlations, Brendel and others [39] concluded that linguistic analysis has power to identify words with biological meaning in nucleotide sequences. In literary texts, keywords are more correlated and clustered than common words. Some genome elements behave similarly, forming clusters. This is the case of genes and CpG islands, which are shown to be clustered through the genome [43; 64; 93].

An early systematic analysis of statistical properties of coding and noncoding DNA sequences has been performed by Mantegna and others [124], by adapting two approaches developed for the analysis of natural languages, the Zipf's law, and that of symbolic sequences, the Shannon entropy. In a recent paper [133], the level of importance of words in DNA sequences is defined using the q-entropy, a key concept of nonextensive statistical mechanics.

The statistical models of language analysis estimate the probabilistic distribution of the components (words or symbols) that integrate a sequence of a given language. Nowadays they are used by all of us when we type an SMS message in the mobile phone taking advantage of the suggestions provided by the software to finish the words of the message [61].

In spite of the fact that several studies point to analogies between DNA and verbal languages, it is postulated that the language of DNA is not comparable to natural human languages [114; 197]. For instance, the same sequence of letters can be translated into distinct messages, depending on the reading frame (see Figure 2.7). Thus, DNA is a peculiar "language" whose intrinsic properties are progressively captured by research.

Word frequencies The study of word frequencies is one of the topics covered in (quantitative) linguistic analysis. In 1932, George Zipf made an empirical observation on some statistical regularities of human writings [214]. Zipf's distribution describes a number of phenomena that are distributed in a very skewed way. It not only approximates word frequencies in texts, but also letter frequencies, city sizes, income ranks, and many other rank versus frequency graphs [186]. In fact, Zipf suggested that this law translates a natural phenomenon of the human behavior, from the individual and social point of view, which he named the *Principle of Least Effort* [215].

In the particular case of languages or long texts, Zipf's law states that the frequency of any word is inversely proportional to (an exponential form of) its rank: $f_r \propto 1/r^{\alpha}$, where f_r is the frequency of the r-ranked word and α is a positive parameter. Regression analysis is applied to discover exactly the coefficient $alpha$ for a particular language. In the case of $\alpha = 1$, the rule states that the most frequent word occurs twice as often as the second most frequent work, three times as often as the subsequent word, and so on until the least frequent word. There is an obviously relation between Zipf's principle and the Shannon entropy: for a finite number of words, a uniform distribution over word frequencies ($\alpha = 0$) leads to the largest entropy; the more skewed the word frequencies (higher values of α), the lower the entropy.

In spite of the fact that Zipf's model had become the most prominent statement of statistical linguistic, researchers have suggested that this model is unsatisfactory to describe the frequency of words distribution in a text.

Two early works that studied statistical features of words in DNA sequences suggested, on the basis of Zipf rank frequency data, that noncoding DNA sequence regions are more like natural languages than coding regions [124; 125]. These statements were highly contested by some researchers. For instance, Niyogi and Berwick [139] point that an empirical fit to Zipf's law cannot be used as a criterion for similarity to natural languages; Konopka and Martindale [109] argue that a reasonable conclusion would be that both coding and noncoding regions fit Zipf's law rather poorly.

Several studies supported the idea that other distributions describe better the frequency of DNA words than Zipf's law. For instance, Borodovsky and McIninch [36] have suggested a logarithmic function to model the rank frequency of codons (words with 3 symbols) in coding DNA sequences; Martindale and Konopka [127] advocated that oligonucleotide rank frequencies in both protein-coding and non-coding regions from several genomes follow a Yule distribution.

The number of word occurrences in a random text has been intensively studied, which consequently brings new approaches to the study of genomic words frequency. Some studies carried out in the context of probability theory are referred in to Section 2.3.

DNA word frequencies are simple, yet effective, statistical tools to capture information about structural patterns. The DNA sequences are not homogeneous and their heterogeneity is reflected at various levels, such as the GC–content, the CpG islands, the asymmetry of A-T. Statistics of the counts of such elements are usually involved in the prediction of genes or coding regions and, within these, in distinguishing exons from introns. Word frequency analysis has potential to reveal biologically significant features in DNA sequences, such as detecting the structural signature of a genome, as well as identifying phylogenetic relationships among different species [48].

Statistical analysis of the (ordered) symbols that make up a gene or a genome can also be used to construct probabilistic models of prediction. A classical gene-finding tool, GeneMark [35], incorporates concepts and algorithms from computational statistics, such as stochastic modeling of sequences using Markov models and Bayesian statistics.

Correlation studies The existence of periodicity in nucleotide sequences is recurrent in natural DNA, especially in eukaryotic genomes. The characterization of correlation structures in DNA sequences has been the subject of many studies. The correlations range in size, from codons, with 3 bp, that encodes the amino acids comprised in proteins, to long correlations, of about 106 bp, that occur in non-coding regions [97; 99]. Different techniques including mutual information functions [90; 111], autocorrelation functions [60; 98], power spectra [120; 208] or Zipf analysis [124; 179] are used for the statistical analysis of correlations in DNA sequences. The general result that emerges from these studies is that DNA statistics is characterised by short-range and long range correlations which are linked to the functional role of the sequences. Specifically, while coding sequences seem to be almost uncorrelated, noncoding sequences show long-range power-law correlations typical of scale invariant systems.

Sequence comparisons In the field of molecular biology, many consensus sequences are known. The comparison of sequences and the measurement of recurrent patterns are fundamental for phylogenetic inference, biological information compression, sequence segmentation, or motif discovery. Thus, it is important to have tools that allow finding occurrences of known patterns in new sequences.

Methodologies used in sequence comparison are oriented to two main cases: searching for exact patterns and searching for approximate patterns. The latter intends to identify segments of the biological sequences that present imperfect copies in other regions of the genome or in different genomes. There are several degenerations that are acceptable in molecular biology, since motifs may show some sequence variation without loss of function. Therefore, the search for approximate patterns is of extreme relevance in the functional analysis of sequences.

Tradicional methods for computing the similarity scores between sequences consist of applying sequence alignment methods. A sequence alignment is a way of organizing primary structures of DNA (or RNA or protein) to identify portions of successive nucleotide that are common in a pair of sequences or more than two sequences. The former is known as pairwise sequence alignment and the latter as multiple sequence alignment. Some uses of alignments are: detecting relationships between sequences; pinpointing conserved regions; defining functional motifs; prediction of protein structure; evolutionary biology, both within and between population comparisons.

Aligned sequences of nucleotides are typically represented as rows within a matrix. Gaps are inserted between the "residues" so that identical or similar characters are aligned in successive columns. There are two basic aspects to consider in this approach: the alignment itself and the scoring used to produce it. Several types of scores may be defined for a pattern. Deterministic scores look for either sequence-match or non-match, while probabilistic models give a probability between 0 and 1, and it is necessary to set some threshold on what should be considered a match or to include these matching probabilities in the score of the pattern. Scoring functions are sometimes based on statistical significance [38]. Comprehensive reviews on alignment methods may be found in [30; 49; 65; 204].

An alternative to the alignment methods are alignment-free methods that can be divided into two main categories: methods based in word frequency (numerical similarity), and those that do not require resolving the sequence with fixed word-length segments (graphical similarity) [202]. The word frequency analysis, correlation analysis, and study of GC–content, referred to above, belong to the former category.

The first group of alignment-free methods includes procedures based on metrics defined in coordinate space of word-count vectors, i.e. L-dimensional vectors defined by the $L = 4^k$ counts of each word of length k (called L-tuples) [202]. These methods quantify the dissimilarity between sequences by measuring the dissimilarity between the corresponding L-tuples. Some of the measures applied between the count vectors are Euclidean distance, correlations, Kullback–Leibler discrepancy, or the angle between them. An early method that does not rely on frequency vectors, and that was shown to be an effective alternative to alignment methods, is based on the distance between transition matrices, after sequences have been modeled as Markov chains.

On the contrary, the second category corresponds to techniques that do not involve counting segments of fixed length, being scale-independent. They include the use of Kolmogorov complexity theory and chaos game representation [202].

Alignment-free sequence analysis have been shown to be an effective alternative to alignment methods, and a growing number of successful applications have been reported on functional annotation and phylogenetic studies.

Identifying patterns There are two fundamentally different tasks related to identifying new patterns in biological sequences. One is called pattern matching and the other is called pattern discovery.

Pattern matching involves finding occurrences of a known pattern. Many consensus sequences (a common sequence shared among a family of similar proteins) are known in biology and it is important to have tools to find occurrences of known patterns in new sequences. This topic is closely related with “Sequence comparisons” previously discussed.

In pattern discovery, the task is to identify a new pattern in a set of several sequences. Identification of significant patterns and classification of sequences are two main issues addressed by pattern discovery. The former consists in discovering patterns that are unlikely to occur by chance and would therefore probably have functional or structural significance, and the latter consists in identifying motifs that characterize members of some sequence family and distinguish them from non-members [38]. After the motif characteristics of a family is found, it could be used as a classifier for new sequences. Pattern discovery and multiple sequence alignment are closely related tasks.

The detection of patterns in genomic data is commonly performed using methods of supervised and unsupervised learning (e.g. neural networks and support vector machines), clustering (e.g. self-organizing maps and k-means) and association rule mining (e.g. distance based association rules), among others. In a review article, Sandve and Drablos [171] survey more than 100 published algorithms for motif discovery.

2.3 Random counts and waiting times

The number of word occurrences in a random text has been intensively studied, with many concurrent approaches. From the statistical point of view, studying the distribution of the random count of a pattern may be a difficult task. Problems connected with waiting times and intersite distances between patterns are also very popular in the classical theory of probability. They can be formulated with no need of difficult notions or technical terms. However, their solutions are far from being trivial. Useful reviews of different approaches on random word occurrences can be found in [121; 140; 159; 161; 165], and reviews on waiting times can be found in [26; 79; 89].

Some of the solutions that have been suggested are reported below.

2.3.1 Word frequency

The comparison between frequencies observed in real sequences and in random sequences allows evaluating the exceptionality of a given word. Finding over- or under-represented words in biological sequences, to discover “relevant” words, is a common task in genomics (see e.g. [126]). A fundamental question that naturally arises is how to define the expected frequency of occurrence of the word in a random scenario. Almost one century ago, Yule [210] conjectured that the correct distribution for word frequencies would be a compound Poisson model, but not propose any mathematical distribution law. Fifty years latter, Sichel [175] did has introduced families of compound Poisson distribution functions to fit word frequencies.

This question is addressed in several studies, where various types of patterns, and various assumptions on the counting overlaps, are considered. Among a wide range of possible models, a popular choice consists in considering homogeneous Markov models of fixed order. This choice is motivated both by the fact that the statistical properties of such models are well known, and that it is a very natural way to take into account the sequence bias in letters or words [141].

Let $S = X_1 X_2 \dots X_n$ be a sequence of letters taken from the alphabet $\{A, C, G, T\}$ with the four letters corresponding to the four bases of DNA, and w be a word in S, i.e. a subsequence of S. The sequence may be generated either according to the Bernoulli model or the Markov model. In the case of the Bernoulli model, every symbol X_i in the sequence is generated independently of the other symbols, i.e. the symbols are generated randomly by a memoryless source. Rather, in a Markov model the probability of each symbol occurrence depends on the previous symbols. The expected frequency of a word, $N(w)$, in case of i.i.d. variables X_i, has been widely considered in the literature. In case of rare words, i.e. if the expectation of $N(w)$ is bounded when n increases, the number of non-overlapping occurrences of a word approximates the Poisson distribution (see e.g. [51; 75]). This result was extended to the more general case in which the sequence S is generated by a Markov chain, showing that the occurrences of rare words can be approximated by a compound Poisson variable [173]. In particular, it reduces to a Poisson variable if the word w cannot overlap itself. Several asymptotic approximations are reported in the literature, the most popular of which are Gaussian approximations [146; 155] and Poisson approximations [86; 88; 160].

Exact methods are based on a wide range of techniques like Markov chain embeeding, generating functions, combinatorial methods, or exponential families [22; 46; 76; 180; 183]. Guibas and Odlyzko [92] were the first to define a generating function to determine the exact probability that S contains at least one occurrence of w. Such functions were able to deal with the presence of substitutions, insertions, and deletions, if the characters of S are generated independently.

2.3.2 Distance between words

The statistics of separations between consecutive words is a very useful tool to isolate relevant terms from generic ones. In general, the former will tend to cluster themselves, as a consequence of their high specificity (attraction or repulsion), while the latter ones will have a tendency to be evenly distributed across the whole text [147].

Repeats, distances between consecutive patterns and waiting times, are closely related topics. The law distribution of waiting times between successive occurrences has been studied by Li [118] and Gerber and Li [84] in the independence case, using martingale methods. Blom and Thorburn [33] also study the problem in the independence case, making connections with Markov renewal theory. The combinatorial methods of Guibas and Odlyzko [92] are particularly effective, having been extended by Chrysaphinou and Papastavridis [50] by considering sequences generated by a markovian process.

There are numerous treatments of the pattern matching problem by probabilistic techniques. For instance, Robin and Daudin [162; 163] and Stefanov [182] provided the exact distribution of the distance between occurrences in Markov chains. There also exist Poisson processes or compound Poisson processes (for overlapping occurrences) which are more efficient than exact approaches in providing the significance of inter-word distance (see for instance [164]).

One of the most general techniques in waiting times studies is the Markov chain embedding method introduced by Fu [74], which has been further developed by Fu and Koutras [73], Antzoulakos [22], Fu [77] and Chang [78]. In this approach, the process of reaching a pattern is modelled by a Markov chain, whose states record the progress towards achieving it (with actually reaching the pattern being an absorbing state). The approaches of Stefanov and Pakes [180] and Stefanov [181] also use Markov chain embedding, though their method differs from Fu's by introducing the exponential family methodology. In Glaz and others [87] and Pozdnyakov [151] one can find an application of the gambling team method to the investigation of occurrences of patterns in Markov chains.

The statistical and probabilistic properties of words in sequences were systematized and reviewed in [159], with emphasis on the deduction of exact distributions and the evaluation of its asymptotic approximations.

2.4 The inter-word distance distribution

The global inter-nucleotide distance mapping, proposed by Nair and Mahalakshmi [136], converts any DNA sequence into a numerical sequence, where each number represents the distance of a symbol to the next occurrence of the same symbol.

Recognizing the potential of inter-nucleotide distances, but realizing that they do not explore the individual behavior of each nucleotide, Afreixo and others [5] have proposed a new methodology based on this mapping transformation. They split the global distance sequence into four inter-nucleotide distance sequences, introducing a new approach to explore DNA correlation structures that simultaneously allows exploring the individual behavior of each nucleotide. The inter-nucleotide distance mapping was then extended to the case of oligonucleotides [29], obtaining the inter-word distances.

Inter-word distance sequence The inter-word (inter-w) distance sequence, d^w, is the sequence of differences between the positions of the first symbols of two consecutive occurrences of that word. For instance, in the short DNA segment

$$s = AAACGTCGATCCGTGCGCG$$

the inter-CG distance sequence is $d^{CG} = (3, 5, 4, 2)$. Note that the concept makes sense only if the number of occurrences of w is at least two. When discussing distance sequences, this will be tacitly assumed.

For a given word w, the inter-word distance sequence, d^w, transforms the nucleotide sequence into a numerical sequence with length equal to the number of occurrences of w minus one. The set of all inter-w distance sequences, for all possible words w of a given fixed length, provides a reversible numerical representation of the original DNA sequence.

Inter-word distance distribution For some practical purposes, distance sequences seem somewhat redundant. In many cases, we care for how often a certain value occurs in the sequence, but not for the specific positions that the value occupies in it. This leads to the inter-word distance distribution.

To exemplify how this distribution translates different behaviors of a word, let's consider a word w and a sequence where it occurs eleven times. The length of the inter-w distance sequence will be ten. If the word occurrences are evenly distributed (say, every 9 nucleotides), the inter-w sequence will consist of the integer 9 repeated ten times:

$$d^w = (9, 9, 9, 9, 9, 9, 9, 9, 9, 9).$$

The plot of d^w will of course have only one peak. However, if the distances between the occurrences of w are randomly but uniformly distributed, one could have inter-w distance sequences such as

$$d^w = (7, 6, 6, 7, 8, 9, 8, 9, 7, 8)$$

which tend to have flatter plots. Finally, a situation where w appears in small clusters separated by a relatively large distance could yield distance sequences such as

$$d^w = (9, 9, 9, 50, 9, 9, 9, 50, 9, 9)$$

which has multiple peaks. It is easy to produce many other distance sequences and interpret them in terms of their empirical distributions. Despite being more concise than the original sequence, this distribution yields valuable insights about the original sequence and can moreover be used to compare sequences.

Distance distributions are related with the linear distribution of words along the genomic sequence, and are therefore related with its primary structure. Under the current knowledge about DNA structure, it is acceptable to hypothesize that some features of the distance distributions might be related to the secondary structure of DNA.

2.4.1 Initial exploratory analysis

Every researcher who is confronted with the observation of a large amount of data is familiar with the oppressive need to synthesize or condense the results. Hence the importance of applying statistical methods, to express relevant information contained in large amounts of data. This assertion was already mentioned by Fisher in 1925 in its hugely influential text [69]. Some decades latter, Tuckey published a book describing a set of new statistical techniques for flexible probing of data, which he coined as *Exploratory Data Analysis* [198].

Some exploratory data analysis techniques can be traced back to earlier authors, for example, to Francis Galton who emphasized order statistics and quantiles [81]; however, there is no doubt that it was Tukey who greatly developed and promoted the use of exploratory data analysis. Today, almost all data analyses are performed with the help of computers, and one of the most common uses of any statistical package is data exploration.

Exploratory data analysis, or EDA for short, employs a variety of techniques: non-graphical methods, that generally involve calculation of summary statistics; graphical data visualization methods, such as boxplots, histograms, scatter plots, and residual plots; dimensionality reduction methods, such as principal component analysis and cluster analysis; and also self-organised maps, just to mention a few. A common goal of all these EDA methods is to find insights that were not evident or are worth investigating.

An initial exploratory analysis of inter-word distance (*iw*D for short) distributions computed from the reference human genome, reveals some global features that are herein described.

A large variation exists between the maximum *iw*D of words of equal length. For instance, the maximum distance observed between consecutive TGG's words is around 5 thousands, while the maximum distance observed between consecutive GGC's is around 37 thousands. Overall, as the word length increases, the larger are the maximum *iw*D values. Since the maximum *iw*D value corresponds to the largest domain value of the distribution, longer tails are expected as the word length increases. Figure 2.13 displays plots of f_D^w for three selected words, namely, $w = AT$, $w = ATC$ and $w = ATCG$.

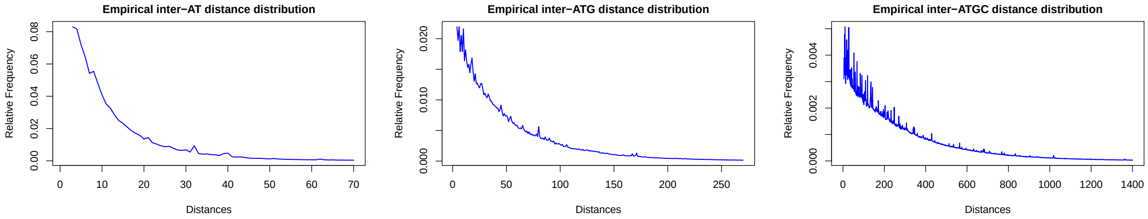


Figure 2.13: Plots of inter-word distance distributions ($k < 5$). The maximum distance displayed correspond to the 99th percentile of the corresponding distribution.

The *iw*D observations have extremely skewed distributions, with short distances being much more frequent than longer distances. In the case of longer words, however, it may occur that specific longer distances have higher frequency of occurrence far exceeding that of neighboring distances. In those cases we refer to those distances/frequencies as a *peak*. Indeed, while *iw*D distributions associated with words of length $k < 4$ have frequencies not too far from a decreasing curve, those of longer words display a most pronounced spiked behaviour (see figure 2.14).

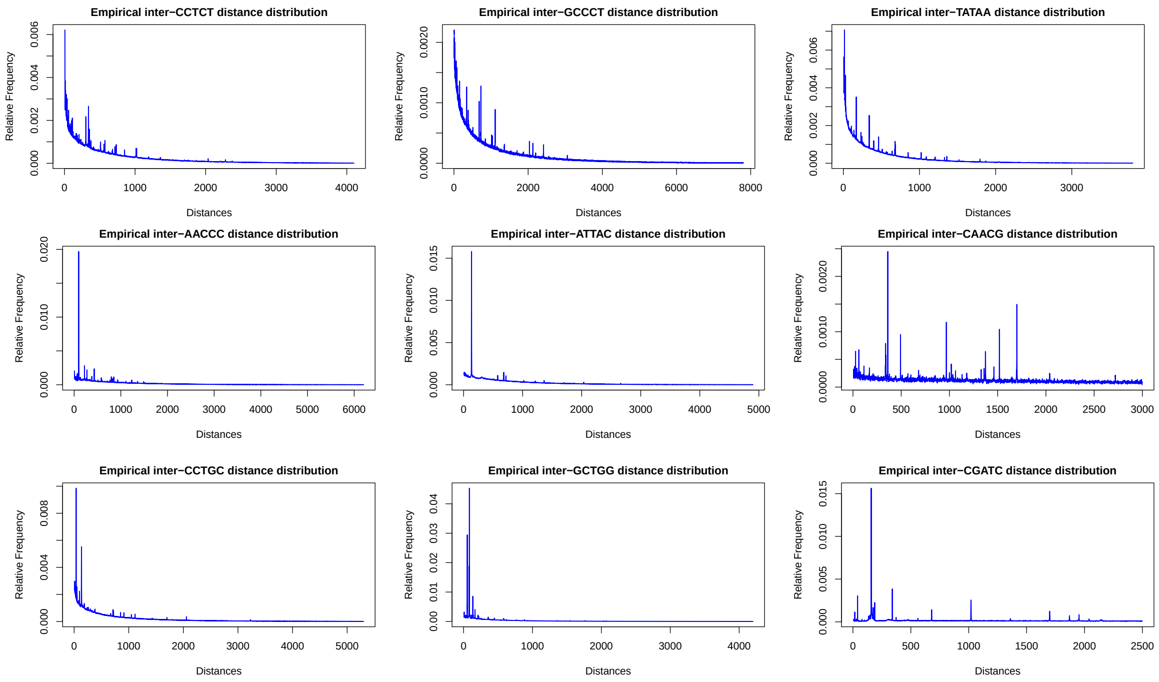


Figure 2.14: Plots of inter-word distance distributions ($k = 5$). A pronounced spiked behaviour is observed in some distributions. The maximum distance displayed corresponds to the 99th percentile, except for $w = CAACG$ and $w = CGATC$ (whose corresponding distances are around 40000).

The distributions of distances between words, together with the word frequencies, are the object of study of this thesis. The next Chapter describes the research questions put forward for this research study.

Chapter 3

Research pathways and questions

DNA word analysis topics include, but are not limited to, the testing of second Chargaff's parity rule and its extensions, the detection and prediction of patterns, motifs, regulatory elements, functional and structural elements. The investigation of such topics is carried out through a variety of methods, such as pattern recognition and alignment-free methods, as mention in Section 2.2.

The identification of words with an abnormal distribution pattern throughout the genome may facilitate the deciphering of the DNA code, as pointed out by Trifonov [195]. For instance, the discovery of the striking repetition of some dinucleotides with a period of about 10.5 nucleotides, remained a peculiarity until its meaning became clear. Later, the prevalence of that structure was associated and ascribed to the formation of nucleosomes, a DNA packaging structure in the chromatin (see Section 2.1). Procedures based on inter-word distances have already been found useful to study genomic sequences, e.g. to distinguish between coding and non-coding regions [29], to detect CpG islands [6] and to distinguishing between species [5]. These studies have also showed that the information contained in the distance distributions of the genomes of different organisms can be traced back to evolutionary patterns.

In view of the above, we consider that exploring the inter genomic word distance distributions, as a mathematical descriptor of DNA sequences, is a relevant topic of study. Our motivation is to uncover pattern (ir)regularities in the DNA, regarding the linear distribution of words along the sequence. Our strategy will be based on the comparison of distributions.

After an exploratory study we will focus on the development of models that are able to adjust the distribution of distances, providing a tool for comparative study of genomic sequences (word-by-word or globally). In a word-by-word analysis, we will explore the distance distributions potential to obtain a classification of the words in groups. For this purpose, we will explore the advantages and disadvantages of using different measures of dissimilarity. Based on inter-word distances, we also intend to construct a method that should allow obtaining genomic signatures, which have the capacity to discriminate between species and to highlight their evolutionary relation. One of the key topics of our research is the establishment of procedures that capture atypical distributions.

Summing up, throughout this thesis three major topics will be discussed: evaluation of dissimilarities, identification of outliers and classification. These topics, clearly interrelated, will be developed in the context of probability distributions, having as application problem the distributions of distances between genomic words. Existing methodologies will be explored and new procedures will be proposed.

3.1 Outlier detection

3.1.1 State of the art

A highly explored topic in genomic word analysis is the identification of words that show an "important" deviation between the observed frequency and the frequency predicted by a fixed model. Such words are usually called *contrast* words, *meaningful* words, *anomalous* words or *exceptional* words [173]. To refer to them as *atypical* or *outlier* words is equally appropriated. A standard approach to detect such outlier genomic words relies on the study of their frequencies.

There is no rigid mathematical definition of what constitutes an outlier. Roughly speaking, an outlier is an observation that appears apart from the bulk of the observations or from the expected value predicted by a model. Assessing the outlyingness of an observation depends both on the feature that is under study and on the dissimilarity measure applied in such study. Obviously, an observation pointed as outlier in relation to a feature, may exhibit a perfectly usual behavior as relates another characteristic.

The identification of outliers is an important aspect of any statistical analysis of data, see, for instance, Barnett and Lewis [28] for a general review on the topic. Outlier detection is a primary step in many scientific research studies, due to the strong impact that they may have on the results. In other cases, however, outliers are sought not for the purpose of eliminating them, but for obtaining critical information. And this is, in general, the motivation behind the outlier detection in the context of genomic words.

In DNA word analysis, outliers themselves may have some interesting biological properties [39], and it may be useful to detect outlying sequences [148]. Many relevant studies show that these words are of interest, due to their possible link with positive or negative selection pressures during evolution [42; 116].

A particular feature of interest, in the study of genomic words, is their spatial distribution along a DNA sequence. The distribution pattern of a word may be characterized by the inter-word distances. The search for outlying words in genomic data by inter-word distances and by word frequency are obviously related. However, a plain distinction between the two kinds of exceptionality exists. The next paragraph puts in evidence distinctions that are worth highlighting.

Over-represented words are generally related to repetitive elements, which may or not have some known biological function. The disposition of repetitive elements, found in genomes, consists either in tandem repeats (arrays of copies which lie adjacent to each other) or in repeats dispersed throughout the genome. This latter case is related with words that are over-represented, but not necessarily at the same distance from each other. Thus, their distance distributions may not point to any strong irregularity. Conversely, a word with a perfectly ordinary overall frequency, may display a preference to repeat itself at an exact distance. This behavior may not be detected by word frequency procedures alone. Indeed, two words with the same frequency may exhibit very distinct patterns of distribution. Consequently, word frequency and distance between words must be considered distinct issues, deserving separated research.

If the concept of multivariate outlier is not strict, the concept of outlier distribution (as a curve) could be even more flexible. The development of methods for the detection of outlying curves is challenging and has been a hot topic under study, in the last ten years, e.g. [24; 68; 100; 101; 167; 185]. It is intricate and nontrivial. Distributions may display outlyingness over a small part of their domain, or be outlying on a substantial part of their domain. Moreover, when dealing with a set of distributions, an outlying distribution may either lie outside the range of the vast majority of the data, or may be within the range of the rest of the data but having a very different behaviour. Thus, the detection of atypical word distributions is a challenging task.

3.1.2 Our questions

To explore the word-distances distribution as a mathematical descriptor of DNA sequences, and motivated by the detection of words with an atypical distribution along the genome, we address the following issues:

(Q1) Outlying distances Do some words reveal a preference for occurring at a given distance from each other? If yes, what words and what distances are those?

Words that present such preference are not necessarily over-represented words. We focus on the problem of identifying words that show a preference for occurring at a exact distance (single distance), or around a distance (cluster of distances) from each other. To evaluate the exceptionality of the frequency of a given distance it is necessary to have a reference value to compare it with. Two fundamental questions that arise are then the definition of a background model and that of a threshold value to assess the difference between the observed frequency and its predicted value.

As a remark, note that the occurrence of favored distances is related with distance distributions which display outlyingness over a small part of their domain.

(Q2) Reference distributions: Which model to use to set the expected frequency of inter-word distances?

Random backgrounds could be used as references to distinguish distances that are found “much more” frequently than expected merely by chance. From the perspective of molecular evolution, DNA sequences reflect both the results of random mutation and of selective evolution. It has been proposed that the subtraction of the random background from the simple counting result highlights the contribution of selective evolution [157]. Hence, a subsequent question concerns the definition of such reference distribution (e.g. Bernoulli and Markov models).

Another possible approach is to consider a set of distributions (for example, those related with all words of the same length) and define a kind of “mean distribution” that reflects the global behaviour of the data set. In this case, the distance frequency of a word is compared with the mean frequency (of that distance) of all words of the same size.

(Q3) Outlying distributions: Which criteria should be used to flag a distance distribution of a given word as exceptional?

To address the problem of identifying genomic words with not only one atypical frequency, but an atypical distribution, a concept of outlying distribution is needed. Should distance distributions with only a spiked frequency be considered outlying? Or should it be mandatory to have an abnormal behaviour on a substantial part of its domain? In either case, such identification requires a reference distribution to compare with. Words with an outlying distribution could be called exceptional words.

Since there is a large quantity of genomic words (for each word length k there are 4^k distinct words), a statistical procedure to identify exceptional words automatically is of utmost importance.

(Q4) Potentialities of identifying exceptional words: Could exceptional words be informative enough to distinguish between species? Could they point out unusual-structural conformations of DNA?

The distance between a real genomic sequence and a random sequence constrained to the Chargaff’s first parity rule, may reflect natural evolution of the species. Indeed, Afreixo and others [5] used the inter-nucleotide distance to build dendrograms, which could be interpreted as phylogenetic trees, and concluded that they are in accordance with the expected similarities between species. Motivated by this result, we wonder to which extent the exceptional words could characterise species. Therefore, the definition of species signatures based on exceptional word profiles will be explored.

The distance distributions capture inherent features of the primary structure of the DNA. We question ourselves if they would have potential to capture features also related to the secondary structure of DNA, in particular unusual-structural conformations of DNA. Thus, we will explore features of distance distributions that might be related with cruciform structures.

The over-representation of a certain distance between a pair of reversed complementary words means a preference for such distance, which, in turn, may point to an increased likelihood of cruciform occurrence at the corresponding locations. Thus, the identification of exceptional words, regarding the distance distribution between reversed complements, will be addressed.

3.2 Similarities: Exploration of Symmetries

3.2.1 State of the art

Chargaff's first parity rule states that in a portion of double stranded DNA the amount of different types of nucleotides follows a specific ratio, namely, the number of adenines is exactly equal to the number of thymines, just as the number of cytosines equals that of guanines, that is, %A = %T and %C = %G. This phenomenon is fully explained by the Watson and Crick double helix model. Interestingly, the relationship between A's and T's and between C's and G's remains almost unchanged in each of the DNA single strands, in a long enough strand. Traces of strand symmetry were first discovered by Chargaff and colleagues [170]. Therefore, strand symmetry is often referred to as *Chargaff's second parity rule*, in the literature.

Some studies focusing on bacteria and eukaryotic genomes have confirmed this symmetry, not only between complementary nucleotides, but also between short oligonucleotides, meaning e.g. that AGC trinucleotide1 tends to be as frequent as its reversed complement GCT along a DNA strand [19; 25; 70; 152; 156]. The marked similarity of the frequencies of any oligonucleotide to those of the corresponding reversed complement, within a single strand, is the so-called *strand symmetry phenomenon*.

Strand symmetry is strongly supported by di-, tri- or higher order oligonucleotides, for sufficiently long genomic sequences. However, it may drop sharply with the increasing of oligonucleotide length; and the smaller the genomic sequence the more likely it may disrupt the rules at some degree [138]. The literature contains several examples where those rules are broken, such as some mitochondrial DNA and also many viruses [132; 138; 207]. The existence of *local* violations of the Chargaff's second parity rule is also well-known, meaning that genomes may be locally less compliant with the parity rule than the average genome, e.g. there exists an excess of G over C and T over A on the coding strand within most genes.

Despite the ubiquity of the strand parity, there is no consensual explanation for the occurrence of this phenomenon. Several theories have emerged evidencing two types of models: the conservation of DNA patterns and the evolution of DNA. In this second model, parity is attributed to mechanisms such as stem-loops [71], duplication followed by inversion [25], invertions and inverted transpositions [19], and statistical mechanics equilibrium [94].

The characterization of the symmetry phenomenon has been object of study of several researchers. Powdel and others [150] locally analyzed the frequency of oligonucleotides along a single strand of DNA, and concluded that the differences between the frequency of reverse complementary words are not statistically significant. Afreixo and others [4] study the single strand symmetry in the human genome by means of an overall frequency framework; they found evidence that the strand symmetry is statistical significant for words of length up to six nucleotides. Further analysis of their results, suggested that, although strand symmetry partially ceases for longer words, it would persist for words of length up to nine in the human genome [212]. Shporer and others [174] empirically observed that the value k up to which the single parity rule holds true in a sequence of length n, is about $0.7\ln(n)$ (for the human genome, this value is ten). These authors also showed that the strand symmetry only holds for reversed complement words pairs and not for complementary words or any random word pair.

In several real sequences, the similarity between the frequency of each word and that of the corresponding reversed complement is larger than the similarity with the frequency of any other word [9; 174]. If a random generator is constrained to respect the strand symmetry phenomenon, then not only reversed complements will have the same prevalence. Rather, every word comprising the same total number of A's or T's, e.g. AAGC and GTAG, will also be equally prevalent. So, it would be interesting to analyse the similarity between the frequency of words with the same composition, in terms of $A+T$, in real genomes. The set of all words with the same length that satisfy such condition is called an equivalent composition group (ECG).

The symmetry between words with equivalent composition was study by Afreixo and others [8]. It was observed that the human genome presents a kind of exceptional symmetry phenomenon, i.e. genomic words frequency is generally more similar to the frequency of its reversed complement than to the frequencies of other words in the same ECG. This trend was not observed in some words with longer lengths (≥ 5).

The existence of words whose frequency is more similar to the frequency of its reversed complement than to the frequencies of any other word in the same ECG, is conceptualized as an exceptional symmetry phenomenon. Further results about this phenomenon in other species and the potentiality of measures of exceptional symmetry measures to compare the evolutionary relation between species were published in [9; 10; 11; 12].

3.2.2 Our questions

To characterize the similarity between the distance distribution of pairs of words of the same length and, in particular between words that are reversed complements, we address the following issues:

(Q5) Exceptional word symmetry: How to measure the contribution of each word to the global strand symmetry effect?

The literature contains several measures to evaluate the symmetry phenomenon in nucleotide sequences. To the best of our knowledge only global measures were addressed. However, the analysis of exceptional symmetry by word was not studied in detail. We intend to define a measure to evaluate the effect of symmetry phenomenon in a word (and in its reversed complement) that takes into account the average frequency deviation between any two words in the same ECG.

(Q6) Distributions similarity: How to assess distributions similarity?

To assess the similarity between two distance distributions, homogeneity and effect size measures could be applied. A simple way to explore the degree of dissimilarity between all pairs of distributions is to use a dissimilarity matrix and apply an hierarchical clustering method. Indeed, the first levels of clustering will be formed by the most similar word pairs. This involves two implicit questions: what dissimilarity measure and what linkage criterion to use? Two earlier dissimilarity measures are the Euclidean distance and the Kullback-Leibler divergence. Are they adequate measures for the detection of discrepancies between words?

(Q7) Word symmetry, in distributions: How similar is the distance distribution of a word and that of its reversed complement?

Due to the close relationship between the frequency of words that are reversed complements (strand symmetry phenomenon), we are particulary interested in studying (dis)similarities among the distribution pattern of such words. In random sequences, generated under an independent symbol model where complementary nucleotides have equal occurrence probabilities, it is expected that reversed complements have similar inter-word distance distributions. Is that still true in real sequences? If "not" or "not always", another question emerges: if a word has an atypical distance distribution, would its reversed complement have an atypical distribution as well? This is one of the linking topics between the detection of outliers and the study of word symmetries in distance distributions.

To evaluate the similarity between de distribution patterns of reversed complementary words, a measure of dissimilarity is needed, which leads us to the previous question.

(Q8) Exceptional word symmetry, in distributions: Is there homogeneity between distance distributions of words in the same ECG?

In random sequences, generated under an independent symbol model where complementary nucleotides have equal occurrence probabilities, it is expected that reversed complements have similar distance distributions, and the same is expected between the distance distribution of words belonging to the same ECG, if restricted to distances greater than the word length (due to word-structural dependencies). However, the prevalence of exceptional strand symmetry phenomena in several real genomic sequences, including the human genome, leads us to suspect that the answer will be "no". Then, a natural follow-up question arises: is the distribution pattern of a word more similar to that of its reversed complement, than to that of any other word in the same ECG? Moreover, are distance distributions of words in the same ECG more similar to each other than to those in other ECG?

Again, a measure of dissimilarity is needed, which conducts the research back again to question Q6.

3.3 Clustering: genomic words

3.3.1 State of the art

Cluster analysis is of considerable interest and importance in the field of bioinformatics, either by clustering proteins or by clustering genes (expression profiles). A related task, which has been fuelled by the challenge of interpreting genome sequences, is to use the resulting information for sequence classification [96].

Historically, protein families have been identified based on multiple alignment. In this case, input data is composed by nucleotide sequences and the task is to find the most concise pattern which is present in all or most sequences. Most of the clustering approaches adopt a similarity value based on sequence alignment scores.

Clustering is the process of grouping similar entities together. The goal of this unsupervised machine learning technique is to find similarities in the data point and group similar data points together. The clustering of genes is used to identify groups of genes with similar patterns of expression, aiming at answering questions about how gene expression is affected by various diseases and which genes are responsible for specific diseases [107]. Gene expression data is usually represented by a matrix with rows representing genes, columns representing samples (e.g. various tissues, developmental stages and treatments), and each cell containing a number characterizing the expression level of the particular gene in the particular sample [37].

In clustering, no predefined reference vectors are used. On the other hand, in supervised classification, vectors are classified with respect to known reference vectors. Since the a *priori* knowledge on the gene expression data is reduced, unsupervised methods have been favored. The hierarchical and k-mean clustering algorithms as well as self-organizing maps have all been used for clustering expression profiles [37].

In biological sequences classification it is usual to look for patterns (exact or approximations) in the different sequences, taking into account the location of these patterns. Indeed, most of the clustering approaches referred above adopt a similarity score based on sequence alignments. Reviews on protein clustering and classification are found in [37; 96].

In spite of the fact that many successful algorithms deal with similarity scores based on sequence alignments, other approaches have been proposed to overcome some disadvantages.

For instance, the main drawback of clustering proteins by algorithms that require a similarity matrix is related with the functional nature of proteins. Proteins rarely act alone. Instead, they must interact with other biomolecular units to execute their function. A natural basis for organizing protein data is to group together proteins with similar protein-protein interactions [184], but there is no straightforward definition for measuring the similarity between protein-protein interactions. Recently, many research works focused on the problem of clustering protein by means of interaction networks.

To compare sequences based on the frequency of occurrence of a given pattern and the way it spreads along the sequence, without considering the specific positions of such pattern, the information contained in the distribution of distances of this pattern could be used. Patterns in distance distributions form an interesting research topic due to their link with structural features of DNA, such as CpG islands.

A different task is to explore how different patterns are spread along the same sequence. Patterns that exhibit more similar distributions could be related to a same biological element, and patterns that exhibit very dissimilar distributions from to the majority of the patterns could be related with some exceptional features.

3.3.2 Our questions

Words of the same size do not behave in the same way, as concerns their distribution along genome sequences. It is likely that distinct patterns coexist in a set composed of distance distributions of all words of the same length, and that such heterogeneity increases as the word length increases. However, it is also likely that the set encloses subsets of similar distributions.

The exponential increasing of the word's number as word length increases, creates the need for organizing the distributions into clusters. The organization of distance distributions in validly clusters fulfills two purposes: to reduce the size of the data, maintaining its representativeness, and to uncover groups of words with similar distribution patterns. The identification of such clusters may provide useful applications in DNA sequence characterization, such as sequence classification and function prediction.

Cluster analysis fits the purpose of finding groups of similar observations (segmentation), and of reducing the dimension of the dataset. To obtain a segmentation of distance distribution datasets, we address the following issues:

(Q9) Relevant features: What features are relevant for clustering distance distributions?

Clustering methods could be able to organize distance distributions according to relevant features of the data. An initial exploratory analysis of distance distributions datasets is necessary to highlight some of such characteristics. To select properly the features on which clustering is to be performed, encoding as much information as possible is a crucial step of any clustering method.

We will be primarily interested in identifying relevant features in sets of distributions of words of length greater than two. Such sets have an interesting number of distributions to cluster, for instance, there are $4^3 = 64$ words of length three, $4^4 = 256$ words of length four, $4^5 = 1024$ words of length five, and so on.

Some statistical characteristics will be analysed, such as the mean distance or its mode, the quartiles of the distribution and skewness. Hypothesizing that the mode could be a feature of interest, lead us to the exploration of a related feature: the presence of spiked peaks of frequencies (i.e. baseline changes or regions of high frequencies with sharp falls on either side). The potential of spiked peaks for outlier detection will be explored, due to its relation with the over-representation of distances, as outlined in section (3.1). Are spiked peaks also useful for clustering distributions? Another possible approach is based on comparisons between empirical distributions and corresponding reference distributions. For instance, the parameter estimates of a distribution fitted to the empirical one could be used as features of interest, if such parametric model exists.

After properly selecting the features on which clustering is to be performed, it is necessary to understand how to encode as much information as possible concerning the features of interest.

The clustering process may result in different partitionings of a data set, depending on the specific criterion used for clustering. According to this perspective, the encoding of features of interest to be taken into account for the clustering method is intrinsically related to the choice of the proximity measure, the clustering criterion, and the clustering algorithm itself. These topics cannot be assessed autonomously and independently. Therefore, a follow-up question closely related to the one we have just presented, is: How to perform clustering based on the selected features?

We will address the development of clustering procedures to group genomic words with similar distance distribution behaviour.

(Q10) Clustering validation: How to assess how well a clustering method performs, and the validity of the results?

An important issue of the clustering process is to assess the validity of the clustering results. Several methods for the quantitative evaluation of the clustering results, known as cluster validity methods, are found in the literature. However, it is important to remark that these methods only give an indication of the quality of the resulting partitioning. Internal validation measures rely on information in the data only, that is, the characteristics of the clusters themselves, such as compactness and separation. A slightly different approach is to assess the suitability of the clustering algorithm by testing how sensitive it is to perturbations in the input data.

On the other hand, when a new algorithm is designed, how to demonstrate the effectiveness of the clustering procedure? Simulation studies are often used to assess how well a clustering method performs. Again, a subsequent question arises when trying to answer the previous question: how to generate proper synthetic data to perform this simulation study?

3.4 Disclosing the research pathways

During the development of this research project, as a barometer, we submited the work developed at the different stages to the appreciation of the scientific community.

Oral and poster communications were made in several scientific meetings. At the same time, manuscripts were submitted and published in indexed international journals or in proceedings of international conferences, as listed in the Appendix.

The questions reported in Sections 3.1 to 3.3 are dealt with in the selected six articles that comprise the second part of the thesis.

Article I mainly concerns the DNA exceptional symmetry phenomenon. A new measure is proposed, to evaluate the difference between the number of occurrences of a word and that of its reversed complement. The word symmetry effect is evaluated in several species, by means of this new measure, and clusters of related species are formed. In this study, question Q5 is addressed.

Article II delves into the “symmetry phenomenon”, by comparing the distance distribution of words that are reversed complements, and words with the same $(A+T)$ content, for word lengths up to five. For each word length, distance distributions are used to build hierarchical clusters of words. The homogeneity between distance distributions is evaluated for specific sets of words: reversed complementary word pairs, and words with equivalent composition (in the same ECG). The existence of reversed complementary word pairs with very similar distance distributions, even when both distributions are irregular and contain strong peaks, is reported. An ECG weighted distribution is defined, and its performance as a profile for distance distributions of words in such ECG is evaluated. Most of the article deals with questions Q6, Q7 and Q8.

Article III is primarily focused on the identification of outlying distributions and their use as a potential species profile. To emphasize the contribution of genomes' selective evolution, occurrences that are likely to occur by chance are usually subtracted from the counting results. Based on this biologic perspective, criteria to identify exceptional words are introduced, grounding on local and global misfits between the observed and the expected distance distributions. To compute the expected frequency of each distance, under a nucleotide independence model, an algorithm is proposed. Then, and regarding the human genome, word-rankings are obtained by sorting words according to the global misfit of the corresponding distribution. Also, exceptional words are identified in the complete genome of 30 species. Dichotomic vectors of exceptional words (from length 1 to 5) are used as a genomic signature of species. In turn, genomic signatures are used to build dendrograms, and perform evolutionary comparisons. Questions Q1, Q2, Q3, Q4 and Q6 are addressed in this article.

Article IV concerns the identification of symmetric word pairs whose pattern of distances between them presents unlikely over-favoured distances. Non-standard DNA conformations, such as cruciform structures are formed at sites that contain reversed complementary words. For this reason, their study naturally leads to the study of the symmetry properties of the sequences. In this research, distances of nearest reversed complements (DNRC) distributions are explored. In order to assess whether or not there is an overall trend towards over-representation of short distances in the human genome, an overall DNRC is taken into account (for each word length). The Global DNRC distribution, a weighted sum of the DNRC distributions of all symmetric word pairs with words of length k, is compared with the corresponding distribution expected under a k-order markovian dependency, and a residual analysis is performed. Procedures to identify symmetric word pairs with uncommon empirical DNRC distribution and with clusters of over-represented short distances are developed, and applied over words of length up to seven in the human genome. Results are shown for words of lengths 6 and 7. These novel procedures for genomic word detection, have potential to uncover words with unexpected features in the DNRC distribution, which could not be detected by word frequency procedures alone. The issues investigated in this article are related with questions Q2, Q3 and Q4.

Article V explores differences between the distance distribution of a given word and that of its reversed complement (such words define a symmetric word pair) combined with differences in their frequency. In a previous study (Article II) it was reported that the distance distribution nearest to the distance distribution of a given word is most often that of its reversed complement (of length $k \leq 5$). Furthermore, that similarity could be surprisingly high in spite of the presence of irregular patterns or some strong peaks. Conjecturing that symmetric word pairs with different patterns could point to evolutionary features, it is interesting to study the differences between distance distributions of symmetric word pairs. Thus, a new dissimilarity measure [188], termed *peak dissimilarity*, was proposed. In the present study, the peak dissimilarity is used to identify words of lengths 5, 6 and 7 with highly distinct distributions, in the human genome, and compared to two well-known measures (Euclidean distance and the symmetrized Kullback-Leibler divergence). By combining low and high values of peak dissimilarity and frequency discrepancy, several behaviors could result and some of them could be of interest. For instance, symmetric pairs that preserve strand symmetry (similar frequency) but have dissimilar distance distributions; and symmetric pairs with dissimilar frequencies and similar distance distributions, could be of interest. Thus, the association between peak dissimilarity values and frequency discrepancy is analyzed. This article contemplates questions Q3, Q6 and Q7.

Article VI describes a clustering procedure designed to seek clusters of genomic words in human DNA by studying their distance distributions. From the extensive treatment carried out during this research project, it became clear that distance distributions display two central characteristics: the decreasing trend and the propensity to present strong peaks of frequencies. The former is associated with behaviours like faster decays with shorter tails, or smoother decays with longer tails. The latter, compresses several behaviours according to the number of peaks, their intensity (stronger or weaker) and their spread (isolated, clustered, dispersed, . . .). Indeed, our previous studies reported a particularly spiked nature of the distance distributions. In the present article, a procedure for decomposing the distance distribution of a word into the sum of a baseline distribution and a peak function is proposed. The baseline, a parametric Gamma distribution, is fitted to the empirical distributions using an outlier-robust fitting technique. The peak function, a peak structure on top of that baseline, is characterized by assessing the extremity of the observed frequencies. The proposed clustering procedure first decomposes each distribution into a baseline and a peak distribution. Then, the set of baselines and the set of peak functions are processed by principal component analysis, in order to create uncorrelated variables. Finally a k-means algorithm is applied to identify the clusters. Moreover, the user has the choice whether to use only the baseline information, only the peak information, or both. The performance of this approach is evaluated by a simulation study. It is also applied to the data set of all genomic trinucleotides in human DNA, as well as the set of all pentanucleotides ($k = 3$ and $k = 5$). The article pinpoints questions Q2, Q6, Q9 and Q10.

Intentionally blank page.

Part II

Research

Chapter 4

Article I

Exceptional symmetry by genomic word

Published:

Afreixo, V., Rodrigues, J. M., Bastos, C. A., Tavares, A. H., Silva, R. M. (2017). Exceptional Symmetry by Genomic Word. Interdisciplinary Sciences: Computational Life Sciences, 9(1), 14–23.

DOI: 10.1007/s12539-016-0200-9

Interdiscip Sci Comput Life Sci (2017) 9:14–23
DOI 10.1007/s12539-016-0200-9

ORIGINAL RESEARCH ARTICLE

Exceptional Symmetry by Genomic Word

A Statistical Analysis

Vera Afreixo[1] · **João M. O. S. Rodrigues**[2] · **Carlos A. C. Bastos**[2] · **Ana H. M. P. Tavares**[3] · **Raquel M. Silva**[4]

Received: 20 July 2016 / Revised: 2 November 2016 / Accepted: 4 November 2016 / Published online: 19 November 2016

Abstract Single-strand DNA symmetry is pointed as a universal law observed in the genomes from all living organisms. It is a somewhat broadly defined concept, which has been refined into some more specific measurable effects. Here we discuss the exceptional symmetry effect. Exceptional symmetry is the symmetry effect beyond that expected in independence contexts, and it can be measured for each word, for each equivalent composition group, or globally, combining the effects of all possible words of a given length. Global exceptional symmetry was found in several species, but there are genomic words with no exceptional symmetry effect, whereas others show a very high exceptional symmetry effect. In this work, we discuss a measure to evaluate the exceptional symmetry effect by symmetric word pair, and compare it with others. We present a detailed study of the exceptional symmetry by symmetric pairs and take the CG content into account. We also introduce and discuss the exceptional symmetry profile for the DNA of each organism, and we perform a multiple comparison for 31 genomes: 7 viruses; 5 archaea; 5 bacteria; 14 eukaryotes.

Keywords Single-strand symmetry · Exceptional symmetry · Multiple organism comparison · Genomic word analysis

1 Introduction

Erwin Chargaff was a biochemist that discovered a set of intriguing rules about the composition of DNA from the analysis of bacterial genomes [1]. The first rule states that the total percentage of complementary nucleotides (A-T and C-G) in double-stranded DNA must be equal. Of course, this is now known to result from the double helix structure of DNA [2]. The second rule sates that the percentage of complementary nucleotides is also identical in each strand [3–5], [6, chap. 4].

A natural extension of Chargaff's second parity rule is that, in each DNA strand, the number of occurrences of a given word (oligonucleotide or *k*-mer) should match that of its reversed complement [6]. The extension to the second parity rule is also known as the single-strand symmetry phenomenon. This symmetry phenomenon refers to the distributions of symmetric pairs, i.e., the distribution of occurrences of all words and the distribution of occurrences of the corresponding reversed complements.

Presently, there is not a generally accepted justification for the need of single-strand parity in DNA sequences, and there is no consensual explanation for the occurrence of the single-strand phenomenon. There are some attempts to explain the phenomenon, which could be classified in two

✉ Vera Afreixo
vera@ua.pt

[1] iBiMED-Institute of Biomedicine, IEETA-Institute of Electronic Engineering and Informatics of Aveiro, CIDMA-Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[2] IEETA-Institute of Electronic Engineering and Informatics of Aveiro, Department of Electronics, Telecommunications and Informatics, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[3] iBiMED-Institute of Biomedicine, Department of Mathematics, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[4] iBiMED-Institute of Biomedicine, IEETA-Institute of Electronic Engineering and Informatics of Aveiro, Department of Medical Sciences, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

groups: the conserved patterns model [7–9], and the evolutive models. Evolutive models can further be classified according to several underlying hypothesis, for example: the stem-loops hypothesis [10]; the duplication followed by inversion hypothesis [11]; the inversions and inverted transpositions hypothesis [12, 13]; the non-uniform substitutions hypothesis [14]; and the statistical mechanics equilibrium hypothesis [15].

To characterize the symmetry phenomenon, Powdel and others [16] analyzed the frequency distributions of oligonucleotides in localized windows along a single strand of DNA. They found that the differences between the frequency distributions of reverse complementary oligonucleotides are not statistically significant. Afreixo et al. [17] noted that the frequency of an oligonucleotide is more similar to the frequency of its reversed complement than to the frequencies of other words of equivalent composition (equal-length oligonucleotides with equal CG content). They called this phenomenon exceptional symmetry, defined measures to evaluate it, and identified several word groups with strong exceptional symmetry in the human genome. More recently, a different measure was introduced to overcome a disadvantage of the previous measure of exceptional symmetry by word [18]. This measure evaluates the difference between the number of occurrences of a word and its reversed complement and relates it with the dissimilarities of the number of occurrences in the corresponding equivalent composition group.

Here, we introduce an improved exceptional symmetry measure and use it to obtain the word symmetry effects in 31 complete genomes stratified by equivalent composition group for word lengths up to 14. Results confirm that measures of word exceptional symmetry can be used to form clusters of related species. Also, we identify words that show high symmetry effect across the 31 species, and across the 9 animal species studied.

2 Materials

The genomes analyzed here are available from the website of the National Center for Biotechnology Information (NCBI; ftp://ftp.ncbi.nih.gov/genomes/). The complete list of species is indicated in Table 1. We selected genomes of species representative of the major taxonomic groups across the tree of life. These include vertebrates, invertebrates, protozoans, fungi, plants, bacteria (gram-positive and gram-negative), archaea and viruses (both double-stranded and single-stranded DNA and RNA viruses).

All non-sequenced or ambiguous nucleotides (mostly N symbols in the sequence file) were discarded from the analysis. For genomes composed by several chromosomes, the chromosomes were processed as separate sequences. All genome sequences used under this study were processed to obtain the word counts, considering overlap between successive words. We obtained the word counts for word lengths from 1 to 14 nucleotides.

3 Methods

In a previous work [17], we called equivalent composition group (ECG) to a set of words with length k that contain a given number m of nucleotides a or t [17]. For example, for $k = 2$ there are three ECGs:

$$G_0 = \{cc, cg, gc, gg\};$$
$$G_1 = \{ac, ag, ca, ct, ga, gt, tc, tg\};$$
$$G_2 = \{aa, at, ta, tt\}.$$

The words division created by ECGs is also called a binary partition [19]. Consider the binary classification of nucleotides in two types, $T_1 = \{a, t\}$ and $T_2 = \{c, g\}$, and let G_m^k (or simply, G_m) be the ECG with words of length k where each word has m symbols of type T_1 and $k - m$ symbols of type T_2, with $m \in \{0, 1, ..., k\}$. Taking into account the combinatorial results (permutations with repetition of indistinguishable objects), it can be concluded that G_m has N_m distinct words,

$$N_m = 2^k \times \frac{k!}{m!(k-m)!}.$$

Note that, for k-mers there are $k + 1$ ECGs with a total of 4^k words.

For even values of k, some words are equal to their reversed complement. We denote these as self symmetric words (SSW). We also define a symmetric word pair as the set composed by one word w and the corresponding reversed complement word w', with $(w')' = w$ (for example, cca and tgg make a symmetric word pair).

We proposed in a previous work [17] one exceptional genomic word symmetry measure evaluated for ECGs and globally. Here, we highlight the exceptional genomic symmetry evaluated for each word, discussing the potentialities of the T measure (symmetric word pair effect, Eq. 1), an improvement of the S measure recently proposed in [18].

Let n_w be the total number of occurrences of word w in the sequence, and n_m be the total number of occurrences of words in the ECG G_m, which contains words composed by m nucleotides a or t. The symmetric word pair effect, for $w \in G_m = \{w_1, w_2, w_3, ..., w_{N_m}\}$, was given by,

$$T(w) = T(w') = \ln \frac{\sqrt{\frac{\sum_{i=1}^{N_m} \sum_{j=1}^{N_m} (n_{w_i} - n_{w_j})^2}{N_m^2 - N_m}} + 1}{|n_w - n_{w'}| + 1}. \quad (1)$$

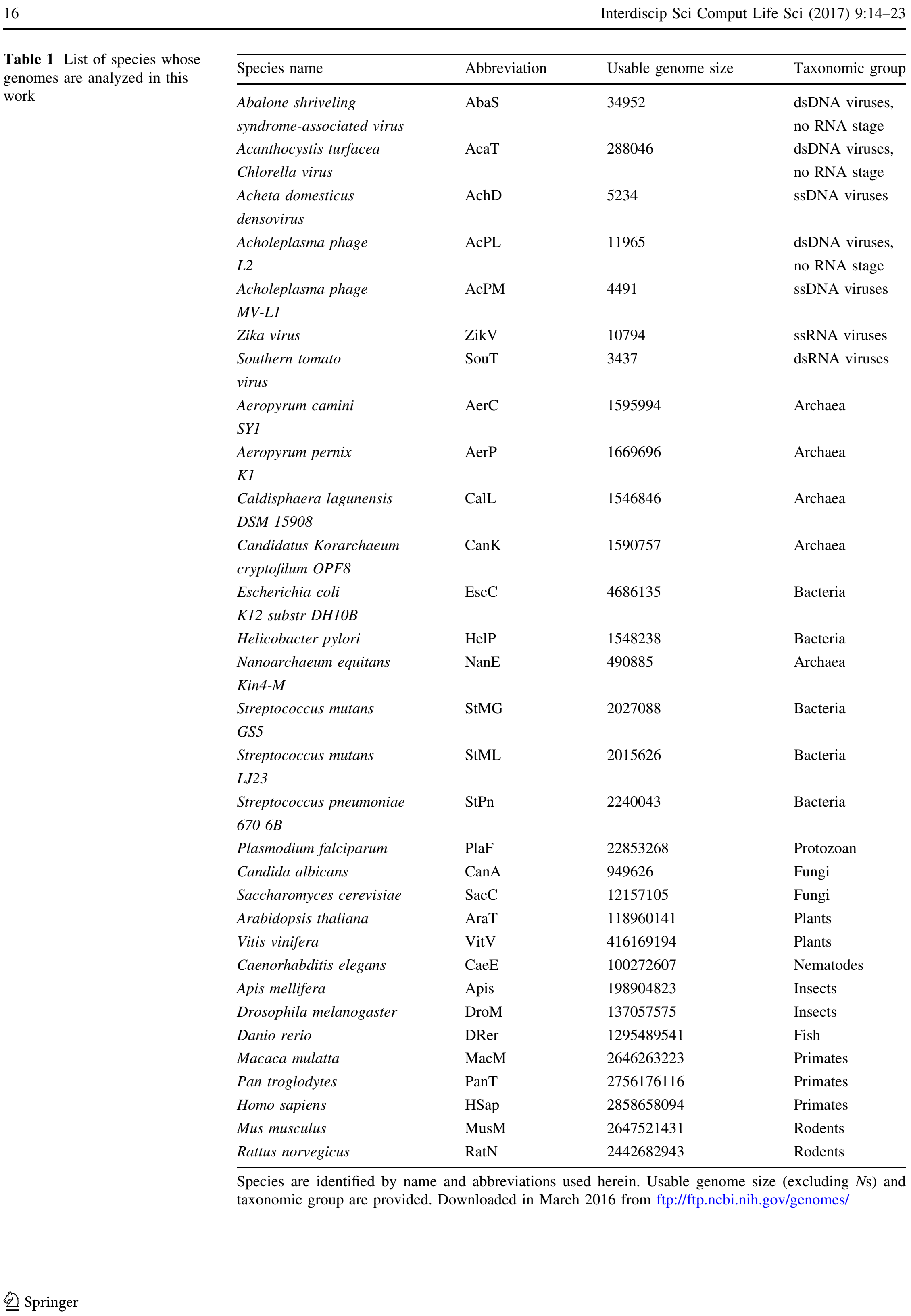


Table 1 List of species whose genomes are analyzed in this work

Species name	Abbreviation	Usable genome size	Taxonomic group
Abalone shriveling syndrome-associated virus	AbaS	34952	dsDNA viruses, no RNA stage
Acanthocystis turfacea Chlorella virus	AcaT	288046	dsDNA viruses, no RNA stage
Acheta domesticus densovirus	AchD	5234	ssDNA viruses
Acholeplasma phage L2	AcPL	11965	dsDNA viruses, no RNA stage
Acholeplasma phage MV-L1	AcPM	4491	ssDNA viruses
Zika virus	ZikV	10794	ssRNA viruses
Southern tomato virus	SouT	3437	dsRNA viruses
Aeropyrum camini SY1	AerC	1595994	Archaea
Aeropyrum pernix K1	AerP	1669696	Archaea
Caldisphaera lagunensis DSM 15908	CalL	1546846	Archaea
Candidatus Korarchaeum cryptofilum OPF8	CanK	1590757	Archaea
Escherichia coli K12 substr DH10B	EscC	4686135	Bacteria
Helicobacter pylori	HelP	1548238	Bacteria
Nanoarchaeum equitans Kin4-M	NanE	490885	Archaea
Streptococcus mutans GS5	StMG	2027088	Bacteria
Streptococcus mutans LJ23	StML	2015626	Bacteria
Streptococcus pneumoniae 670 6B	StPn	2240043	Bacteria
Plasmodium falciparum	PlaF	22853268	Protozoan
Candida albicans	CanA	949626	Fungi
Saccharomyces cerevisiae	SacC	12157105	Fungi
Arabidopsis thaliana	AraT	118960141	Plants
Vitis vinifera	VitV	416169194	Plants
Caenorhabditis elegans	CaeE	100272607	Nematodes
Apis mellifera	Apis	198904823	Insects
Drosophila melanogaster	DroM	137057575	Insects
Danio rerio	DRer	1295489541	Fish
Macaca mulatta	MacM	2646263223	Primates
Pan troglodytes	PanT	2756176116	Primates
Homo sapiens	HSap	2858658094	Primates
Mus musculus	MusM	2647521431	Rodents
Rattus norvegicus	RatN	2442682943	Rodents

Species are identified by name and abbreviations used herein. Usable genome size (excluding *N*s) and taxonomic group are provided. Downloaded in March 2016 from ftp://ftp.ncbi.nih.gov/genomes/

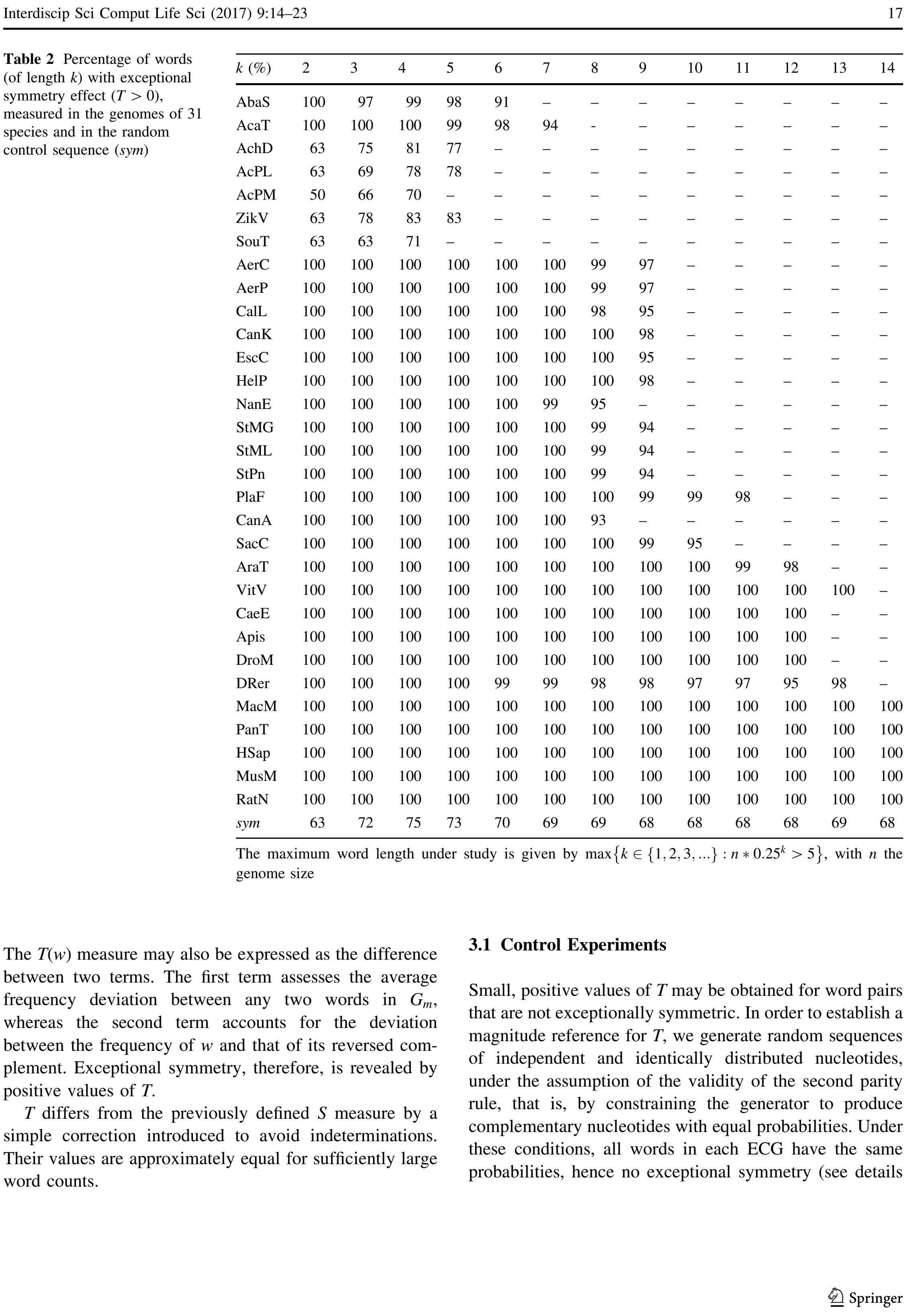


Table 2 Percentage of words (of length k) with exceptional symmetry effect ($T > 0$), measured in the genomes of 31 species and in the random control sequence (*sym*)

k (%)	2	3	4	5	6	7	8	9	10	11	12	13	14
AbaS	100	97	99	98	91	–	–	–	–	–	–	–	–
AcaT	100	100	100	99	98	94	-	–	–	–	–	–	–
AchD	63	75	81	77	–	–	–	–	–	–	–	–	–
AcPL	63	69	78	78	–	–	–	–	–	–	–	–	–
AcPM	50	66	70	–	–	–	–	–	–	–	–	–	–
ZikV	63	78	83	83	–	–	–	–	–	–	–	–	–
SouT	63	63	71	–	–	–	–	–	–	–	–	–	–
AerC	100	100	100	100	100	100	99	97	–	–	–	–	–
AerP	100	100	100	100	100	100	99	97	–	–	–	–	–
CalL	100	100	100	100	100	100	98	95	–	–	–	–	–
CanK	100	100	100	100	100	100	100	98	–	–	–	–	–
EscC	100	100	100	100	100	100	100	95	–	–	–	–	–
HelP	100	100	100	100	100	100	100	98	–	–	–	–	–
NanE	100	100	100	100	100	99	95	–	–	–	–	–	–
StMG	100	100	100	100	100	100	99	94	–	–	–	–	–
StML	100	100	100	100	100	100	99	94	–	–	–	–	–
StPn	100	100	100	100	100	100	99	94	–	–	–	–	–
PlaF	100	100	100	100	100	100	100	99	99	98	–	–	–
CanA	100	100	100	100	100	100	93	–	–	–	–	–	–
SacC	100	100	100	100	100	100	100	99	95	–	–	–	–
AraT	100	100	100	100	100	100	100	100	100	99	98	–	–
VitV	100	100	100	100	100	100	100	100	100	100	100	100	–
CaeE	100	100	100	100	100	100	100	100	100	100	100	–	–
Apis	100	100	100	100	100	100	100	100	100	100	100	–	–
DroM	100	100	100	100	100	100	100	100	100	100	100	–	–
DRer	100	100	100	100	99	99	98	98	97	97	95	98	–
MacM	100	100	100	100	100	100	100	100	100	100	100	100	100
PanT	100	100	100	100	100	100	100	100	100	100	100	100	100
HSap	100	100	100	100	100	100	100	100	100	100	100	100	100
MusM	100	100	100	100	100	100	100	100	100	100	100	100	100
RatN	100	100	100	100	100	100	100	100	100	100	100	100	100
sym	63	72	75	73	70	69	69	68	68	68	68	69	68

The maximum word length under study is given by $\max\{k \in \{1,2,3,\ldots\} : n * 0.25^k > 5\}$, with n the genome size

The $T(w)$ measure may also be expressed as the difference between two terms. The first term assesses the average frequency deviation between any two words in G_m, whereas the second term accounts for the deviation between the frequency of w and that of its reversed complement. Exceptional symmetry, therefore, is revealed by positive values of T.

T differs from the previously defined S measure by a simple correction introduced to avoid indeterminations. Their values are approximately equal for sufficiently large word counts.

3.1 Control Experiments

Small, positive values of T may be obtained for word pairs that are not exceptionally symmetric. In order to establish a magnitude reference for T, we generate random sequences of independent and identically distributed nucleotides, under the assumption of the validity of the second parity rule, that is, by constraining the generator to produce complementary nucleotides with equal probabilities. Under these conditions, all words in each ECG have the same probabilities, hence no exceptional symmetry (see details in [20]). The label *sym* is used to denote these random sequences in the remainder of the document.

3.2 Word Analysis Procedure

A word is declared as exceptionally symmetrical when its T value surpasses the critical value, which is defined as the 95th percentile of the T values obtained from the control experiments. To complement this analysis, we compute the percentage of words with $T \leq 0$ for each word length.

To identify groups of genomes with similar exceptional symmetry profiles ($T(w)$ values), we use a hierarchical clustering procedure, using the UPGMA aggregation criterion with Euclidean distance. A similar clustering procedure is used to identify words with similar exceptional symmetry profiles across species.

4 Results and Discussion

For the set of 31 genomes, the word counts were obtained for all word lengths between 1 and 14 nucleotides, and the symmetric word pair effect was obtained for each genomic word. However, for given genome, we only consider the genomic words with lengths k ($k \in \{1, \ldots, k_{\max}\}$), with

$$k_{\max} = \max\{k \in \{1, 2, 3, \ldots\} : n * 0.25^k > 5\}$$

and n the genome size. This threshold motivation is the count representability and the protection of the T measure to the sensitivity of rare counts occurrences.

Obviously, for $k = 1$, each ECG contains only one symmetric word pair, and so $T(w) = 0$, for all nucleotides. Almost all words in eukaryote genomes show significant exceptional symmetry effect (above the critical values obtained in the control experiments). Table 2 shows the percentage of words with $T > 0$ for each species and word length of this study. A high percentage of words in viruses show no exceptional symmetry. This result agrees with a previous work [20], which used a different measure and procedure.

Table 2 includes the *sym* row corresponding to one control scenario (sequence with length equal to the length of the human genome). This may be used as a reference of non-exceptional symmetry results.

4.1 Human Genome

A word analysis in the context of exceptional symmetry for the human genome was carried out.

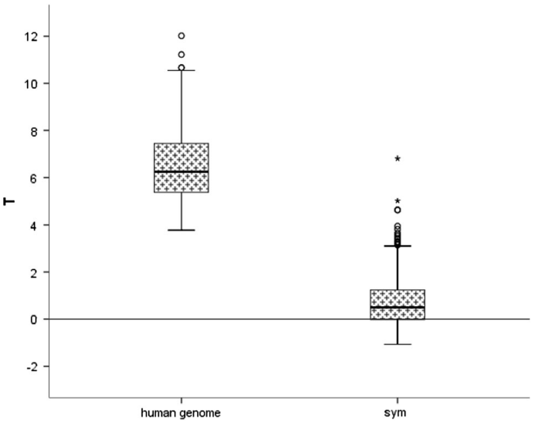


Fig. 1 Boxplots for T values in the human genome and in a random control sequence realization (*sym*) for word length 5

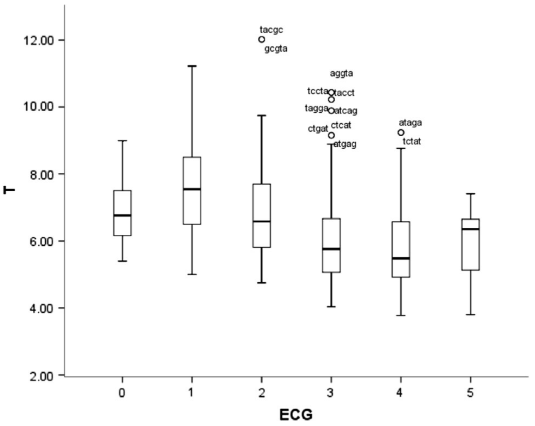


Fig. 2 Boxplots for T values in each ECG for word length 5, in the human genome

Figure 1 shows boxplots of the T values for $k = 5$ in the human genome and in the corresponding random realization *sym*. The boxplot for the human genome shows high and significant symmetric word pair effects. The most exceptionally symmetric word pairs, corresponding to the right outliers, detected in the human T boxplot are: (gcgta, tacgc), (accgg, ccggt), (gccac, gtggc), (gccca, tgggc), (cggga, tcccg).

Figure 2 shows the T values in each ECG for $k = 5$ in the human genome. We observe that as the CG content varies (decreases along the x-axis), the T median values have a non-monotonous behavior. The ECG G_1 has the highest T median value. In general, for the word lengths under study and for the human genome, the T median in ECG G_0 is lower than in G_1, and the T median for G_k is higher than for G_{k-1}. For the control scenario, on the other hand, we observed that the T median values remained essentially constant across all ECGs.

Table 3 The six symmetric word pairs (represented by a single word of the pair) that have the highest (−h) $T(w)$ values, and the six symmetric word pairs that have the lowest (−l) $T(w)$ values for each k, in the human genome

Rank	Word length (k)							
	2	3	4	5	6	7	8	9
1st -h	gg	ccg	cggg	gcgta	taatca	atgttag	atattaac	aaaaatata
	6.9	10.8	9.8	12.0	11.9	12.1	11.7	11.5
2nd -h	tc	ctc	gcga	accgg	acaccg	cagttgg	cgagtggc	atatattta
	6.3	9.7	9.6	11.2	11.9	12.0	11.0	10.2
3rd -h	ct	atg	gggc	tgggc	ggatcg	cttaggc	gagcacgc	aaattaaat
	5.9	9.2	9.2	10.7	11.9	12.0	11.0	10.2
4th -h		gcg	agcc	gtggc	gtcacg	agatcgg	accggcgt	ttaaatata
		8.2	9.2	10.7	11.7	12.0	11.0	10.1
5th -h		tcg	tccg	tcccg	ccgtga	acgaatg	gtcgcgga	aaattttat
		7.6	9.1	10.5	11.5	12.0	11.0	10.1
6th -h		gct	aggg	aggta	ttatct	gcgtacg	cgggtcga	ttattaata
		7.4	9.1	10.4	11.5	11.4	11.0	10.1
6th -l		tct	tctt	gtttt	ttttgt	gtgtgtg	atggaatg	tggaatgga
		4.9	4.4	4.0	3.6	3.3	2.5	1.5
5th -l		ctt	gttt	ttttg	tttttt	aatggaa	tggaatgg	ccaggctgg
		4.8	4.3	3.9	3.6	3.3	2.5	1.5
4th -l		gtt	tttg	tgttt	ttgttt	tttgttt	tgtgtgtg	aatggaatg
		4.4	4.1	3.9	3.6	3.1	2.4	1.4
3rd -l	gt	ttt	tttt	tttgt	tttttg	atggaat	gaatggaa	gtgtgtgtg
	4.8	4.3	4.1	3.8	3.6	3.1	2.3	1.4
2nd -l	tg	tgt	ttgt	ttttt	tttgtt	tgtgtgt	aatggaat	tgtgtgtgt
	4.7	4.2	4.0	3.8	3.5	3.0	2.3	1.3
1st -l	tt	ttg	tgtt	ttgtt	tgtttt	gaatgga	gtgtgtgt	gaatggaat
	4.4	4.2	4.0	3.8	3.5	2.9	2.3	1.2

Table 3 continued

Rank	Word length (k)				
	10	11	12	13	14
1st -h	tttaatataa	attttttataa	aattaaaaatat	atttattttttta	aaatatataaataa
	12.0	11.6	11.1	10.6	10.0
2nd -h	atatattttt	taaaaataatt	ataaaattaaat	taaaatattttaa	aaaaaataatttta
	12.0	10.9	11.1	10.6	10.0
3rd -h	aaatatataa	taaaaatttta	attaaaataaat	aatatatataatt	ataaatatttttaa
	11.6	10.9	11.1	10.6	10.0
4th -h	tacaataaaa	ttaattattaa	aattatatttta	aattaattaaaat	atataaataatata
	11.6	10.9	11.1	10.6	10.0
5th -h	aaattttgta	attaattaatt	aattaaaatata	ataatttaaaata	aattaaaaatatat
	11.1	10.5	11.1	10.6	10.0
6th -h	ttatttagaa	atttaaatttt	aaattaatttaa	aaaatatttatat	atattatttttaaa
	11.1	10.5	11.1	10.6	10.0
6th -l	cgaatggaat	ttgtgatgtgt	tttgtgatgtgt	ctggctaattttt	tatgtccagagttt
	0.8	0.0	−0.7	−1.4	−2.0
5th -l	gaatggaatg	ctggctaattt	tgtccagagttt	ctttgtgatgtgt	tagagcagttttga
	0.7	0.0	−0.8	−1.4	−2.0
4th -l	atggaatgga	tcgaatggaat	ctggctaatttt	tttgtgatgtgtg	agagcagttttgaa
	0.6	−0.2	−0.9	−1.5	−2.1
3rd -l	tgtgtgtgtg	gtgtgtgtgtg	gaatggaatgga	cctggctaatttt	cctggctaattttt
	0.5	−0.2	−0.9	−1.6	−2.3
2nd -l	gtgtgtgtgt	aatggaatgga	tgtgtgtgtgtg	gtgtgtgtgtgtg	tgtgtgtgtgtgtg
	0.5	−0.5	−1.1	−1.8	−2.7
1st -l	aatggaatgg	tgtgtgtgtgt	gtgtgtgtgtgt	tgtgtgtgtgtgt	gtgtgtgtgtgtgt
	0.5	−0.6	−1.1	−2.1	−2.7

In case of a tie, the words are sorted (largest to smallest) by frequency of occurrence. Note that for $k \geq 7$ there are (w, w') pairs where $n_w = n'_w$, exactly

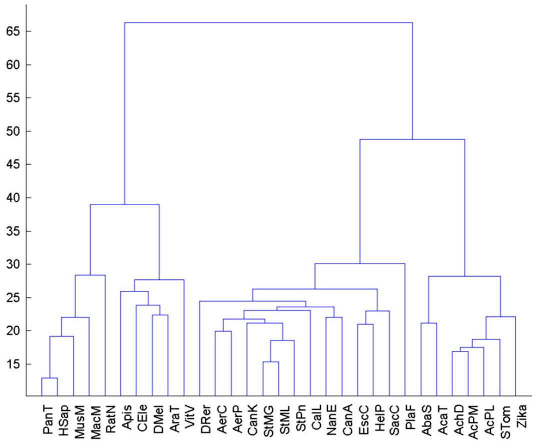


Fig. 3 Dendrogram obtained from the T values for all species under study, word length 4

Table 4 Word pairs with exceptional symmetry effect above the third quartile, which are most common across species, and most common across animal species

k	Word	% of species	Word	% of animals
2	*ca*	42	*cc*	67
3	*acg*	58	*ccg*	78
4	*cgac*	52	*aacg*	67
	cgga		*agcg*	
			cgac	
			cgcc	
			tccg	
5	*attcg*	61	*aacgg*	78
			cgatc	
			gcgcc	
6	*acgcgt*	74	*acggat*	89
			ccgtac	
			gacgta	
7	*tacgtaa*	74	*cgtacga*	100

Each pair is represented by a single word of the pair

Table 3 presents, for the word lengths under study, the twelve words with the six highest and the six lowest $T(w)$ values. Some of these extreme words could have some biological interest, e.g., regulatory elements, functional elements, motifs.

Based on the results of the effect size measure, we may conclude that the human genome presents exceptional symmetry. The human genome shows exceptional symmetry for the thirteen different word lengths ($k = 2, \ldots, 14$) used in this study.

Although the existence of global exceptional symmetry in the human genome was verified, there are distinct profiles for each chromosome. Consequently, the exceptional symmetry profile may be used as a signature of each chromosome. Preliminary results also suggest that exceptional symmetry profiles are distinct between species, which will be presented in the next section.

It may be also concluded that in the human genome there are ECGs that are more exceptionally symmetric than others. And a large percentage of the genomic words present some exceptional symmetry. However, for longer word lengths ($k \geq 5$), there are some words without any exceptional symmetry. With this analysis, it was identified that words rich in CG content behave differently from words rich in AT content, in terms of exceptional symmetry.

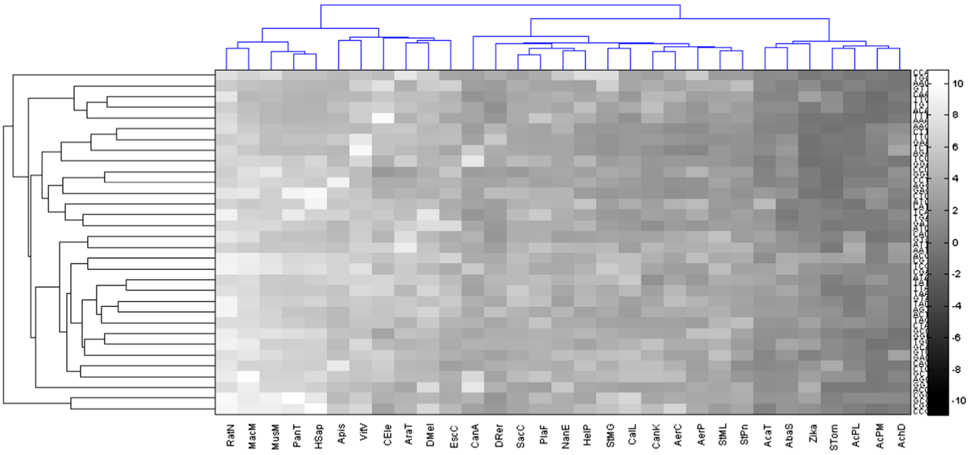


Fig. 4 Heatmap with biclustering organization of the T values for words of length 3 and for all species under study

4.2 Species Comparison

Figure 3 shows the dendrogram obtained with the hierarchical clustering procedure, for $k = 4$. Four distinct groups can be observed in Figure 3: mammalian (on the left); viruses (on the right); a group including the plants and the other animals (except *Danio rerio*); and a group with the unicellular species, plus *Danio rerio*. For other word lengths, the resulting dendrograms essentially maintain the same structure (the dendrogram for $k = 3$ is also included in Figure 4).

Figure 4 shows the heatmap with biclustering organization for trinucleotides. Species are shown on the horizontal axis, and words are shown on the vertical axis. The symmetric word pair effect is stronger on the left side of the heatmap, corresponding to multicellular organisms, and weaker on the right side. The word clustering highlights the group formed by two symmetric word pairs: (ccg, cgg), (gcg, cgc).

We identified the word pairs with high exceptional symmetry (T above the third quartile) in every species under study. From these, we selected the pairs that are highly symmetric across the most species under study, and those that are highly symmetric across the most animal species under study. The results are shown in Table 4. No word pair is considered highly symmetric across all the species under study. However, $T(cgtacga) = T(tcgtacg)$ is above the third quartile in all the animal species under study. The strongest symmetric word pair effect is observed in words composed by CpG dinucleotides.

The results presented in Table 4 are restricted to word lengths between 2 and 7 because for longer word lengths the number of most common symmetric word pair above the third quartile is high. The strongest symmetric word pair effect is observed in words composed by CpG dinucleotides.

5 Conclusions

We evaluated the exceptional symmetry effect in several species, with particular emphasis in the human genome. The word exceptional symmetry values contain information specific to the species and seem to contain information about the species evolution. Taking into account the species in this study, the primates and rodents species have the highest exceptional symmetry values and form a subgroup distinct from all the other species under study. Globally, the eukaryote group showed the highest word exceptional symmetry values, while viruses showed the lowest values. We reinforce that some viruses show a behavior opposite to the exceptional symmetry ($T < 0$) in almost all words under study.

Exceptional symmetry effect was found in a high percentage of words in all cellular organisms under study. Therefore, we conjecture that exceptional symmetry results from some universal law imposed on cellular organisms. Still, the exceptional symmetry profiles are species specific.

Acknowledgements This work was supported by Portuguese funds through the iBiMED-Institute of Biomedicine, IEETA-Institute of Electronics and Informatics Engineering of Aveiro, CIDMA - Center for Research and Development in Mathematics and Applications and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within projects: UID/BIM/04501/2013, PEst-OE/EEI/UI0127/2014 and UID/MAT/04106/2013.

References

1. Chargaff E (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia 6(6):201–209
2. Watson J, Crick F (1953) A structure for deoxyribose nucleic acid. Nature 171:737–738
3. Karkas JD, Rudner R, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase. Proc Natl Acad Sci USA 60(3):915–920
4. Rudner R, Karkas JD, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands, I. Biological properties. Proc Natl Acad Sci USA 60(2):630–635
5. Rudner R, Karkas JD, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands. III. Direct analysis. Proc Natl Acad Sci USA 60(3):921–922
6. Forsdyke DR (2011) Evolutionary bioinformatics. Springer, New York
7. Sobottka M, Hart AG (2011) A model capturing novel strand symmetries in bacterial DNA. Biochemical and biophysical research communications 410(4):823–828. doi:10.1016/j.bbrc.2011.06.072. http://www.sciencedirect.com/science/article/pii/S0006291X1101045X
8. Zhang SH, Huang YZ (2008) Characteristics of oligonucleotide frequencies across genomes: conservation versus variation, strand symmetry, and evolutionary implications. Nat Precedings:1–28.
9. Zhang SH, Huang YZ (2010) Strand symmetry: characteristics and origins. In: Fourth international conference on bioinformatics and biomedical engineering (iCBBE) 2010. pp. 1–4 (2010). doi:10.1109/ICBBE.2010.5517388
10. Forsdyke DR, Bell SJ (2004) Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. Appl Bioinform 3(1):3–8
11. Baisnée PF, Hampson S, Baldi P (2002) Why are complementary DNA strands symmetric? Bioinformatics 18(8):1021–1033
12. Albrecht-Buehler G (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc Natl Acad Sci USA 103(47):17,828–17,833
13. Albrecht-Buehler G (2007) Inversions and inverted transpositions as the basis for an almost universal "format" of genome sequences. Genomics 90:297–305
14. Lobry TH (1995) Properties of a general model of DNA evolution under no-strand-bias condition. J Mol Evol 40:326–330

15. Hart A, Martnez S, Olmos F (2012) A gibbs approach to Chargaff's second parity rule. J Stat Phys 146:408–422

16. Powdel B, Satapathy S, Kumar A, Jha P, Buragohain A, Borah M, Ray S (2009) A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (chargaff's second parity rule). DNA Res 16:325–343

17. Afreixo V, Rodrigues JMOS, Bastos CAC (2015) Analysis of single-strand exceptional word symmetry in the human genome: new measures. Biostatistics 16(2):209–221

18. Afreixo V, Rodrigues JMOS, Bastos CAC, Silva RM (2016) Exceptional symmetry profile: A genomic word analysis. In: PACBB

19. Kong SG, Fan WL, Chen HD, Hsu ZT, Zhou N, Zheng B, Lee HC (2009) Inverse symmetry in complete genomes and whole-genome inverse duplication. PLoS ONE 4(11):e7553

20. Afreixo V, Rodrigues JMOS, Bastos CAC (2014) Exceptional single strand DNA word symmetry: analysis of evolutionary potentialities. J Integr Bioinform 11(3):250

Intentionally blank page.

Chapter 5

Article II

The symmetry of oligonucleotide distance distributions in the Human Genome

Published:

Tavares, A. H., Afreixo, V., Rodrigues, J. M., Bastos, C. A. (2015). The Symmetry of Oligonucleotide Distance Distributions in the Human Genome. In Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM-2015) vol.2, 256–263.

DOI: 10.5220/0005223102560263

The Symmetry of Oligonucleotide Distance Distributions in the Human Genome

Ana Helena Tavares[1], Vera Afreixo[1,2], João M. O. S. Rodrigues[3,4] and Carlos A. C. Bastos[3,4]
[1]*Department of Mathematic, University of Aveiro, 3810-193, Aveiro, Portugal*
[2]*Center for Research and Development in Mathematics and Applications (CIDMA), Aveiro, Portugal*
[3]*Department of Electronics Telecommunications and Informatics, University of Aveiro, 3810-193, Aveiro, Portugal*
[4]*Institute of Electronics and Telematics Engineering of Aveiro (IEETA), Aveiro, Portugal*

Keywords: Chargaff's Second Parity Rule, Single Strand Symmetry, Oligonucleotide Distance Distribution, Human Genome.

Abstract: The inter-oligonucleotide distance is defined as the distance to the next occurrence of the same oligonucleotide. In this work, using the inter-oligonucleotide distance concept, we develop new methods to evaluate the lack of homogeneity in symmetric word pairs (pairs of reversed complement oligonucleotides), in equivalent composition groups. We apply the developed methods to the human genome and we conclude that a strong similarity exists between the distance distributions of symmetric oligonucleotides. We also conclude that exceptional distance symmetry is present in several equivalent composition groups, that is, there is a strong lack of homogeneity in the group and a strong homogeneity in the included symmetric word pairs. This suggests a stronger parity rule than Chargaff's: in the human genome, symmetric oligonucleotides have equivalent occurrence frequency and, additionally, they present similar distance distributions.

1 INTRODUCTION

Chargaff's first parity rule states that, in any sequence of double-stranded DNA molecules, the total number of complementary nucleotides is exactly equal (%A=%T and %C=%G). Clearly, this is an inevitable consequence of the complementarity of opposing nucleotides in the two strands of the DNA molecule. Chargaff's second parity rule states that %A≅%T and %C≅%G in a single strand of DNA (Forsdyke and Mortimer, 2000). The extensions to second parity rule state that, in each DNA strand, the proportion of an oligonucleotide (a subsequence of adjacent nucleotides) should be similar to that of its reversed complement (the oligonucleotide obtained reversing its letters and interchanging complementary nucleotides). Unlike the first rule, there is no single accepted reason that justifies this single strand symmetry. However, the relative ubiquity of this phenomenon suggests a relationship with genomic evolution (Forsdyke 2010, ch. 4).

Several works discuss the prevalence of Chargaff's second parity rules for several oligonucleotide lengths, and in different organisms (Afreixo et al., 2013b; Albrecht-Buehler, 2006; Baisnée, Hampson and Baldi, 2002). However, the universality of Chargaff's second parity rule has been questioned for organellar DNA and some viral genomes (Mitchell and Bridge, 2006). Powdel and others (2009) studied the symmetry phenomenon from an interesting new perspective, by defining and analysing the frequency distributions of the local abundance of mono/oligonucleotides along a single strand of DNA. They found that the frequency distributions of reverse complementary mono/oligonucleotides tend to be statistically similar. Afreixo et al. (2014) introduced a new symmetry measure, which emphasizes that the frequency of an oligonucleotide is more similar to the frequency of its reversed complement than to the frequencies of other equivalent composition oligonucleotides.

The inter-nucleotide distances introduced by Nair and Mahalakshmi (2005) convert any DNA sequence into a unique numerical sequence, where each number represents the distance of a symbol to the next occurrence of the same symbol. Afreixo et al (2009) explored the global inter-nucleotide representation and proposed the extraction of four

Tavares A., Afreixo V., M. O. S. Rodrigues J. and A. C. Bastos C..
The Symmetry of Oligonucleotide Distance Distributions in the Human Genome.
DOI: 10.5220/0005223102560263
In *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2015), pages 256-263
ISBN: 978-989-758-077-2

sequences, one for each nucleotide, to represent the inter-nucleotide distances. This methodology allows to perform comparative analysis between the behaviour of the four nucleotides. Bastos et al (2011) proposed an inter-dinucleotide distance distribution and compared the distance distributions of all dinucleotides in the human genome. Moreover, evolutionary patterns have been recognized from information contained in the distance distributions of the genomes of different organisms (Afreixo et al., 2009).

In this work we explore the symmetry of distance distributions by comparing each inter-oligonucleotide distance distribution to the distance distribution of its reversed complement, using homogeneity discrepancy measures. We also characterize the discrepancy in equivalent composition groups (ECGs), and compare ECGs results for different oligonucleotide lengths.

We focus our study in the human genome as an example of a typical genome exhibiting single strand symmetry.

2 MATERIALS AND METHODS

2.1 Materials

We analyse the whole human genome, reference assembly build 37.3, available from the website of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/), discarding all ambiguous or non-sequenced nucleotides from the analysis, that is, all non-ACGT symbols.

In our data processing, the chromosomes of the human genome were processed as separate sequences, words were counted with overlap and non-ACGT symbols were considered as sequence separators.

2.2 Methods

Let $\mathcal{A}$ be the alphabet formed by the four nucleotides $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and let $s = s_1, s_2, \dots, s_N$ be a symbolic sequence defined in $\mathcal{A}$. A genomic word, or oligonucleotide, w, is a sequence of length k.

Assuming that the sequence is read through a sliding window of length k, we can define the inter-oligonucleotide (inter-w) distance sequence, d^w, as the sequence of differences between the positions of the first symbol of consecutive occurrences of that oligonucleotide. For instance, in the DNA segment $s = \mathrm{AAACGTCGATCCGTGCGCG}$, the inter-CG distance sequence is

$$d^{CG} = (3,5,4,2).$$

The inter-w distance distribution (or word distance distribution), denoted as f_w, gives the relative frequency of each inter-w distance. For each k, there are 4^k distance distributions.

The reversed complement of a genomic word is a sequence obtained by reversing the order of the letters in the word, interchanging A and T and interchanging C and G. For instance, the reversed complement of ACTGG is CCAGT. A symmetric word pair is defined as the set composed by one word, w, and the corresponding reversed complement word, w', with $w'' = w$ (Afreixo, Garcia and Rodrigues, 2013a; Afreixo *et al*., 2014).

In this work we compare the inter-w distance distribution, f_w, of symmetric word pairs. To the set formed by the distance distributions of symmetric word pairs we will call complementary distributions.

An equivalent composition group (ECG), of words with length k, is a set composed by all the words with the same total number of As or Ts. For instance, the four dinucleotides AA, AT, TA and TT form an ECG. For words of length k there are $k + 1$ equivalent composition groups and the group formed by words comprising m As or Ts is denoted as G_m, with $0 \leq m \leq k$. The number of words of length k in G_m is given by

$$\#G_m = \binom{k}{m} 2^k$$

Every symmetric word pair is a subset of an ECG, which contains several distinct symmetric word pairs (Afreixo, Bastos and Rodrigues, 2014).

We will call equivalent composition distributions to the distance distributions of words in the same ECG.

Under the second parity rule, and under a scenario of nucleotide independence, it is expected that reversed complements have similar frequency and similar inter-distance distribution (homogeneous distributions), but so do all words in the same ECG. The similarity between the frequencies of occurrence of reversed complements and of other equivalent composition words is described by Afreixo *et al*. (2014).

We assess homogeneity in symmetric word pairs and in ECGs, using the distance distributions of words of length up to five, in the complete human genome.

Using empirical data from the contingency table, whose columns are filled with the absolute frequency of inter-w distances of a set, $\mathcal{S}$, of words, we find the expected frequency of each distance for each word (dividing the product of the row total and the column total by the total sum). The chi-squared statistic is defined as

$$X_{\mathcal{S}}^2 = \sum_{i,j} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}},$$

where n_{ij} is the observed frequency count in word i for distance j, and e_{ij} is the corresponding expected frequency, in the homogeneity context.

2.2.1 Symmetric Word Pair Measures

To evaluate the dissimilarity between the inter-words distributions of symmetric word pairs (w and w') we use an effect size measure based on a chi-square statistic to measure the discrepancy between the distance distributions of reversed complement words: the phi coefficient given by

$$\varphi_{w,w'} = \sqrt{\frac{X^2_{\{w,w'\}}}{N^w + N^{w'}}},$$

where N^w and $N^{w'}$ are the number of occurrences of w and w' in the sequence, respectively. Equal distributions will result in $\varphi_{w,w'} \cong 0$ and an increase in dissimilarity will be reflected in an increased $\varphi_{w,w'}$.

For interpreting the phi coefficient, we consider a value above 0.10 as a descriptor for small effect size, above 0.30 for medium effect size, above 0.50 for large effect size (Cohen, 1988), above 0.60 for strong effect size and above 0.80 for a very strong effect size (Rea and Parker, 1992).

We define the weighted distribution of the complementary distributions (w,w'), denoted as $f_{w,w'}$, the following distribution

$$f_{w,w'} = \frac{f_w \times N^w + f_{w'} \times N^{w'}}{N^w + N^{w'}}$$

and the distance corresponding to the 99th percentile of the weighted distribution of the symmetric pair is denoted as $d_{0.99}^{w,w'}$.

2.2.2 ECG Measures

We define an ECG distribution profile as the weighted distribution of the equivalent composition distributions. The G_m distribution, denoted as f_{G_m}, is given by

$$f_{G_m} = \sum\nolimits_{w \in G_m} \frac{f_w \times N^w}{n_{Gm}},$$

where n_{Gm} is the total number of occurrences of words that belong to G_m, in the sequence. The 99th percentile of this weighted distribution is denoted as $d_{0.99}^{G_m}$

Since different ECGs may contain distinct numbers of elements, to evaluate the dissimilarity between the inter-word distance distributions of each ECG we use the Cramér's V coefficient given by

$$V_{G_m} = \frac{\varphi_{G_m}}{\sqrt{\#G_m - 1}},$$

which takes into account the degrees of freedom of the chi-square distribution (under the homogeneity hypothesis) to normalise the phi coefficient,

$$\varphi_{G_m} = \sqrt{\frac{X^2_{G_m}}{n_{Gm}}}.$$

3 RESULTS AND DISCUSSION

With the increase of the oligonucleotide length, we observe a large variation in basic descriptive statistics of the distance distributions. For example, for $k = 5$, the maximum recorded distance of the distributions ranges from 27,800 to 1,355,000, approximately. Unsurprisingly, the distributions that reach the greatest maximum distance contain larger percentages of distances with null frequencies. Figure 1 displays box plots (organized by word length) of the maximum recorded distance of each distribution, d_{max}^w, and the 99th percentile of each distribution, $d_{0.99}^w$. Figure 2 displays a box plot of the percentage of distances, from 99th percentile to maximum recorded distance, with null frequencies, of each distribution.

The differences in the length of distance distribution, the amount of longer distances with null frequencies, and the sensitivity of the chi-square statistic to low frequencies that occur for longer distances, lead us to define a cutoff that ensures an adequate representation of the distributions, without introducing the long tails of low density. To incorporate the contribution of the tail in our calculations, we also group all the remaining distances in one extra residual class.

Thus, we compute $X_{\mathcal{S}}^2$ and $V_{\mathcal{S}}$, making a cutoff in the 99th percentile of the weighted distribution of $\mathcal{S}$, where $\mathcal{S}$ is one of the following sets: (w, w'), ECG, and $\mathcal{K}$. We use the weighted average of the distributions of the elements in $\mathcal{S}$ because that leads to a low mean squared error unbiased estimate of the cutoff point under the null hypothesis assumption (homogeneity in $\mathcal{S}$).

Since the structure of words with overlap (words with a suffix that matches with one proper prefix of the word) prevents some short distances from occurring, we also expected a large variability, in the first k distances, between distributions of the same ECG. Therefore, we also explore the similarity between the equivalent composition distributions, excluding the first k distances of the empirical distributions in the calculations of $X^2_{G_m}$ and V_{G_m}.

3.1 Inter-word Distance Analysis for Symmetric Pairs

For each word length (from 1 to 5), we use the inter-w distance distributions to build dendrograms that show hierarchical clusters. The inter-word distance distributions in the same cluster are more similar to each other than to those in other groups.

We use the complete linkage clustering and the average linkage clustering to build the dendrograms, and compute the similarity matrix with the Euclidian distance. We performed several cluster analysis varying the dimension of the similarity matrix.

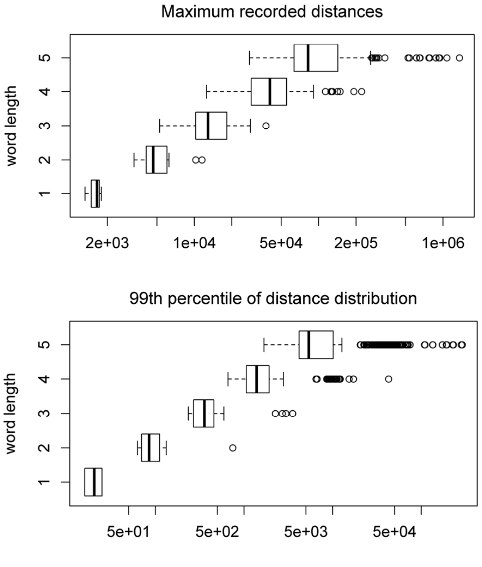


Figure 1: Box plots of the: maximum recorded distance of each distribution, d^w_{max}(**top**); the 99th percentile of each distribution, $d^w_{0.99}$(**bottom**). Organized by word length.

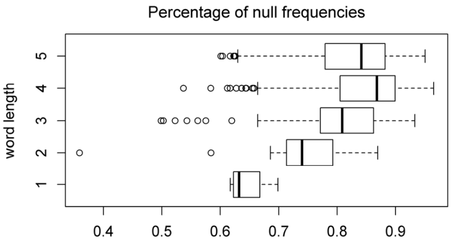


Figure 2: Box plot of the percentage of distances, from 99th percentile to maximum recorded distance, with null frequencies.

To compute the cluster analysis of all the distance distributions, of words of the same length, we had to define a cut point in the distributions. We use the maximum 99th percentile of the ECG distributions and consider a residual class containing the remaining distances.

Since some distances from 1 to k may be absent due to the structure of the words, we also perform the cluster analysis removing the first k distances and normalizing the distributions.

In all the obtained dendrograms, we observe that the first similarity levels are formed by complementary distributions. This indicates that inter-word distance distributions of symmetric word pairs are the most similar, over all the words of the same length. Figure 3 shows one dendrogram of distance distributions of trinucleotides using distances from 1 to the maximum of the 99th percentile of the ECG distributions and a residual class.

These results motivated us to compare and evaluate the similarity between the inter-word distance distributions of symmetric word pairs. Thus, we compute the phi coefficient, $\varphi_{w,w'}$ and sort the symmetric pairs according to that value.

In general, we obtained very low values of phi. Table 1 presents the maximum recorded phi for each word length. We found that, for $1 \leq k \leq 4$, all the symmetric pairs have low values of $\varphi_{w,w'}$, meaning similarity between the complementary distributions. However, for $k = 5$, we detected 16 symmetric pairs with medium effect size ($0.3 \leq \varphi_{w,w'} < 0.5$), 2 symmetric pairs with large effect size ($0.5 \leq \varphi_{w,w'} < 0.6$) and 1 pair with strong effect size ($0.6 \leq \varphi_{w,w'} < 0.8$). All of these distance distributions belong to oligonucleotides comprising one or more CGs.

Another result that stands out for $k \geq 3$, is that the distributions that reach the highest values of phi coefficient are always distributions of CG-rich words (i.e., oligonucleotides comprising one or more CG). On the other hand, the distributions that reach the lowest values of phi coefficient are distributions of words rich in Ts or As. Table 2 displays the symmetric word pairs whose distance distributions have the 6 highest and the 6 lowest $\varphi_{w,w'}$, organized by word length.

Table 1: Maximum and 90th percentile of phi coefficient.

k	1	2	3	4	5
$\max(\varphi_{w,w'})$	0.001	0.001	0.016	0.094	0.662
90th percentile of $\varphi_{w,w'}$	0.001	0.001	0.008	0.055	0.019

Table 2: Symmetric word pairs with the 6 highest and the 6 lowest effect size.

$\varphi_{w,w'}$	Word length				
	1	2	3	4	5
max1	A/T	AC/GT	CGC/GCG	CGAC/GTCG	CGACG/GTCG
max2		AA/TT	CGA/TCG	CGTA/TACG	ACGCG/CGCGT
max3		CA/TG	CCG/CGG	ACCG/CGGT	CGCGA/TCGCG
max4			ACG/CGT	GCGA/TCGC	CGCCG/CGGCG
max5			GAC/GTC	CGTC/GACG	CGTAC/GTACG
max6			GCC/GGC	CGCA/TGCG	CCGCG/CGCGG
…					
min6			CAG/CTG	AATA/TATT	AAATT/AATTT
min5			TAA/TTA	GAAA/TTTC	AGAAA/TTTCT
min4			AGA/TCT	TAAA/TTTA	AAATA/TATTT
min3		CC/GG	TCA/TGA	AGAA/TTCT	TAAAA/TTTTA
min2		GA/TC	AAA/TTT	AAAT/ATTT	AAAAT/ATTTT
min1	C/G	AG/CT	AAT/ATT	AAAA/TTTT	AAAAA/TTTTT

The similarity between the complementary distributions is clearly observable in histograms. An extraordinary observation that comes out of this study is the conservation of the similarity in the unexpected spikes of the symmetric distributions, for the generality of the symmetric pairs. Figure 4 displays three word distance distributions of symmetric pairs (the first 150 distances). The similarity between the distributions of symmetric pairs is remarkable even when the distributions are so irregular as those of GCTA/TAGC or ATCAC/GTGAT. All of these cases present negligible effect sizes.

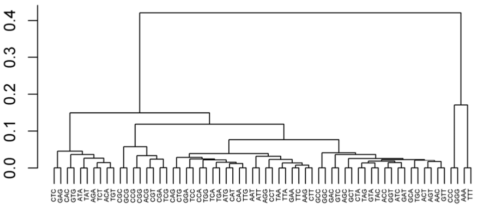


Figure 3: Dendrogram using Euclidean distance and complete linkage clustering for inter-word distance distributions of trinucleotides.

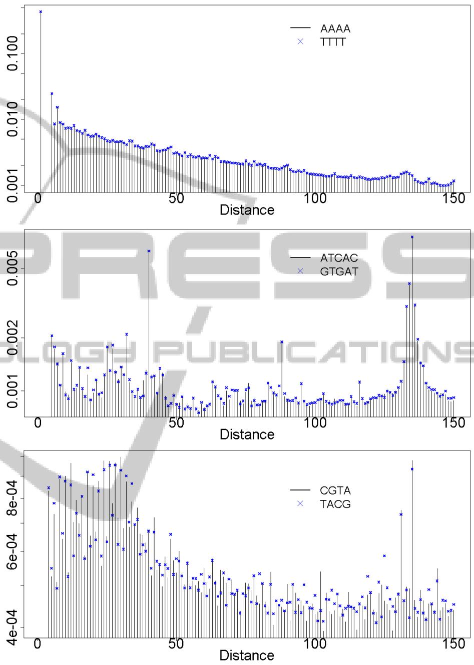


Figure 4: Inter-word distance distributions of the first 150 distances for symmetric pairs, in log-scale: AAAA vs TTTT), $\varphi_{\mathrm{AAAA,TTTT}} \cong 0.003$(top); GCTA vs TAGC, $\varphi_{GCTA/TAGC} \cong 0.011$(middle); ATCAC vs GTGAT, $\varphi_{\mathrm{ATCAC,GTGAT}} \cong 0.03$(bottom).

3.2 Inter-word Distance Analysis for ECG

To find the ECG groups with stronger exceptional symmetry, we compute the Cramér's V values obtained for each group of equivalent composition distributions.

As already mentioned, to compute V_{G_m}, we set a distance cut point and create a residual class with the remaining distances. The V_{G_m} is calculated for distances from 1 to $d_{0.99}^{G_m}$. Since the word structure of some words prevents some distances from 1 to k from occurring, because of the word overlap, we also explore the similarity between the equivalent composition distributions, excluding the first k distances of the empirical distributions.

For nucleotides, $k = 1$, we conclude that there is no significant dissimilarity between the distance distributions in each ECG. In fact, those equivalent composition distributions get effect size values much less than 0.1.

Considering all distances up to the cut point and a residual class, we observe that, for $k > 1$, the minimum effect size is associated to G_{k-1}. Moreover, G_k tends to reach one of the highest effect sizes (see Table 3a).

With the removal of the first k distances, and due to the non-existence of some distances in a few distributions in the same ECG, we expect a decrease in the effect sizes. In fact, for $k > 1$, this decrease occurs and the existence of homogeneity in G_{k-1} holds true. Moreover, the most homogeneous ECG is G_k, that is, the group of words comprising only As and Ts.

The lack of exceptional symmetry in G_k, with the removal of the first distances, may be related to the extraordinary spikes that poly-A and poly-T distributions reach at distance one (see Figure 4, top). For $k \geq 4$, some groups present strong φ_G effect sizes (which can be computed from the V_G values in Table 3b).

Table 3: Cramér's V effect size of each ECG, organized by word length. (a) distances from 1 to $d_{0.99}^{G_m}$, with a residual class. (b) distances from k+1 to $d_{0.99}^{G_m}$, with a residual class.

(a)

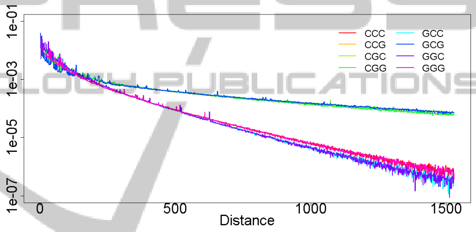
ECG	Word length				
	1	2	3	4	5
G_0	0,0003	0,31	0,23	0,17	0,14
G_1	0,0003	0,31	0,23	0,17	0,14
G_2	--	0,29	0,07	0,07	0,05
G_3	--	--	0,24	0,04	0,04
G_4	--	--	--	0,20	0,02
G_5	--	--	--	--	0.16

(b)

ECG	Word length				
	1	2	3	4	5
G_0	0,0004	0,24	0,17	0,12	0,10
G_1	0,0007	0,06	0,12	0,08	0,06
G_2	--	0,06	0,04	0,06	0,04
G_3	--	--	0,05	0,03	0,03
G_4	--	--	--	0,05	0,03
G_5	--	--	--	--	0,05

In general, ECG discrepancies (Table 3a) are higher than symmetric pair discrepancies (Table 2), suggesting the existence of an exceptional symmetry of distance distributions.

Figure 5 displays the equivalent composition distributions of trinucleotides in G_0 and in G_3. In G_0 the irregularity of the distributions is clearly visible in the first 100 distances. Furthermore, for distances higher than 500, we also observe a huge divergence between two groups of distributions. The combination of these behaviours results in a dissimilarity between the distance distributions related to this ECG. In Figure 5 (bottom) we observe that the distributions have a more homogeneous behaviour, which results in a smaller Cramér's V effect size (Table 3).



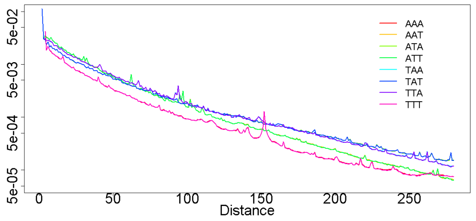


Figure 5: Inter-word distance distribution of trinucleotides in G_0, $d_{0.99}^{G_0} = 1526$(**top**); Inter-word distance distribution of trinucelotides in G_3, $d_{0.99}^{G_3} = 280$(**bottom**).

To evaluate the variability inside each ECG we use the standard deviation of the Euclidean distance between the word distribution and its ECG profile. The Euclidian distance was computed considering distances from $k + 1$ to $d_{0.99}^{w}$ and a residual class. We conclude that, in general, G_0 is the ECG with one of the highest variations (Table 4). The only exception is verified for $k = 3$, in which G_0 presents the lowest dispersion. We also observe that G_1 reaches one of the highest variations.

We extend our study to the evaluation of the ECG weighted distribution as a profile of the inter-word distance distributions. For each word w, we want to analyse if the most similar ECG distribution, in relation to f_w, is the G_w distribution.

Let G_w denote the ECG of the word w and $\bar{G}_w$ denote any of the other ECGs. To assess the similarity between the word distribution, f_w, and each of the ECG weighted distributions, we compute the Euclidean distance between f_w and f_{G_i}, for $i = 0, \dots, k$, considering word distances from $k+1$ to $d^w_{0.99}$ and a residual class. Then, we sort the Euclidean distances and extract the ECG distribution most similar to f_w.

We found that the lowest Euclidean distance is not always associated to the G_w distribution, meaning that the most similar ECG weighted distribution in relation to f_w is not always G_w. For example, only 38% of tetranucleotides have distance distributions closer to G_w distribution than to any of the other ECGs, and all the distance distributions of tetranucleotides in G_1 are closer to some $\bar{G}_1$ distribution than to the G_1 distribution. Table 5 summarize the percentage of distance distributions, f_w, that are closest to G_w distribution (than to $\bar{G}_w$), over all k-mer distributions. It also presents this percentage, over all the equivalent composition distributions.

These results may give evidence that, even inside an ECG group, the words could not follow the same profile, which is agreement with the exceptional distance symmetry of some ECG. As an example, recall the distance distributions of trinucleotides inside G_0 (Figure 5, top), which suggest the existence of two distinct distribution profiles. These results are in agreement with previously related exceptional distance symmetry of some ECG, and with the hierarchical clustering performed in subsection 3.1, where distributions of words in the same ECG were not grouped in the same cluster.

To assess similarities between the ECG weighted distributions (the ECG profiles), we build dendrograms for each word length. We used the complete linkage clustering and the average linkage clustering to build the dendrograms, and we computed the similarity matrix with the Euclidian distance. To perform the hierarchical clustering we set a cutoff in all the distributions and create a residual class. To ensure an adequate representation of all the ECG we define the distance cut point at the maximum of the 99th percentile of the ECG profiles, that is, $max\{d^{G_i}_{0.99} : i = 0, \dots, k\}$.

We observe that, for $k \leq 4$, the G_k and the G_0 profile distributions are grouped in the same cluster. Moreover, for $k > 2$, the G_{k-1} and the G_1 profile distributions are also grouped in the same cluster. Figure 6 display some of the obtained dendrograms.

Table 4: Standard deviation of the Euclidean distance between the word distance distribution and its ECG distribution.

ECG	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
G_0	3.70E-08	**0.029**	0.004	**0.017**	**0,017**
G_1	1.25E-07	0.011	**0.017**	**0.010**	**0,023**
G_2		0.001	0.006	**0.010**	**0,014**
G_3			0.006	0.007	**0,012**
G_4				0.008	0,010
G_5					0,008

Table 5: Percentage of words of length k, %k, whose distance distribution is closest to the $G_w$distribution than to a $\bar{G}_w$distribution. Percentage of words in each ECG, %G_i, whose distance distribution is closest to $G_w$distribution than to $\bar{G}_w$.

k	1	2	3	4	5
% k	100	88	63	38	29
% G_0	100	100	50	69	65
% G_1	100	75	42	0	3
% G_2		100	75	43	8
% G_3			100	50	42
% G_4				88	53
% G_5					75

4 CONCLUSIONS

In this work, we contribute with a new method to evaluate one refinement of Chargaff's second parity rules: symmetry of word distance distributions. For each word length, we propose measures of symmetry in symmetric word pairs based on the comparison of the inter word distance distributions. We also compare the homogeneity of symmetric words with the homogeneity inside an ECG. In general, we conclude that the lack of homogeneity between symmetric words is negligible. In some ECGs the discrepancy in word distance distributions is negligible but in other ECGs it is very strong. These results led us to identify the exceptional words in the context of the symmetry of distance distributions: mostly CG-rich words.

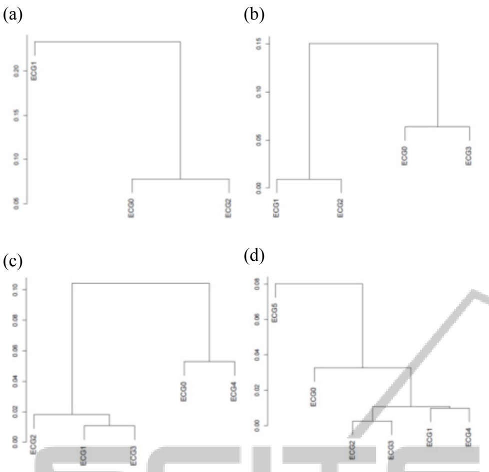


Figure 6: Dendrogram using Euclidean distance and complete linkage clustering for ECG weighted distributions, distances from 1 to the maximum of the 99th percentile of the ECG distributions and a residual class.

(a) $k = 2$; (b) $k = 3$; (c) $k = 4$; (b) $k = 5$.

ACKNOWLEDGEMENTS

This work was supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, IEETA - Institute of Electronics and Telematics Engineering of Aveiro and the Portuguese Foundation for Science and Technology (\FCT {Fundação para a Ciência e a Tecnologia"), within projects PEst-OE/MAT/UI4106/2014 and PEst-OE/EEI/UI0127/2014. Ana H M Tavares is supported by the project EXPL/MAT-STA/1674/ 2013.

REFERENCES

Afreixo, V., Bastos, C. A., Rodrigues, J. M., (2014), 'Analysis of exceptional word symmetry in single strand DNA: new measures', doi: 10.1093/biostatistics/kxu041.

Afreixo, V., Garcia, S. P. and Rodrigues, J. M. (2013a), 'The breakdown of symmetry in word pairs in 1,092 human genomes', *Jurnal Teknologi*, 63(3).

Afreixo, V., Bastos, C. A., Garcia, S. P., Rodrigues, J. M., Pinho, A. J., & Ferreira, P. J. (2013b), 'The breakdown of the word symmetry in the human genome'. *Journal of theoretical biology*, 335, pp.153-159.

Afreixo, V., Bastos, C. A., Pinho, A. J., Garcia, S. P. and Ferreira, P. J. (2009), 'Genome analysis with inter-nucleotide distances', *Bioinformatics*, 25(23), pp. 3064-3070.

Albrecht-Buehler, G. (2006). 'Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions', *Proceedings of the National Academy of Sciences*, 103(47), pp.17828-17833.

Baisnée, P. F., Hampson, S. and Baldi, P. (2002). 'Why are complementary DNA strands symmetric?', *Bioinformatics*, 18(8), pp.1021-1033.

Bastos, C. A., Afreixo, V., Pinho, A. J., Garcia, S. P., Rodrigues, J. M. O. S. and Ferreira, P. J. (2011), 'Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions', *Journal of Integrative Bioinformatics*, 8(3), pp.172.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* , 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Forsdyke, D. R. and Mortimer, J. R. (2000), 'Chargaff's legacy', *Gene*, *261*(1), pp.127-137.

Forsdyke, D. R. (2010). Evolutionary Bioinformatics. Springer, Berlin.

Mitchell, D. and Bridge, R. (2006), 'A test of Chargaff's second rule', *Biochemical and Biophysical Research Communications*, *340*(1), pp.90-94.

Nair, A. S. S. and Mahalakshmi, T. (2005), 'Visualization of genomic data using inter-nucleotide distance signals', *Proceedings of IEEE Genomic Signal Processing*, 408. Bucharest, Romania.

Powdel, B. R., Satapathy, S. S., Kumar, A., Jha, P. K., Buragohain, A. K., Borah, M., & Ray, S. K. (2009). 'A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's second parity rule)', *DNA Research*, 16(6), pp.325-343.

Rea, L. M. and Parker, R. A. (1992) '*Designing and conducting survey research'*, San Francisco, CA: Jossey–Bass.

Intentionally blank page.

Chapter 6

Article III

Detection of exceptional genomic words: a comparison between species

Published:

Tavares, A., Rodrigues, J., Bastos, C., Ferreira, P., Brito, P., Afreixo, V. (2016). Detection of exceptional genomic words: a comparison between species. In Proceedings of the 22nd International Conference on Computational Statistics (COMPSTAT 2016), Oviedo, Spain.

DOI: 10.29007/jvg4

Detection of exceptional genomic words: a comparison between species

Ana Tavares, *University of Aveiro*, ahtavares@ua.pt
João Rodrigues, *University of Aveiro*, jmr@ua.pt
Carlos Bastos, *University of Aveiro*, cbastos@ua.pt
Armando Pinho, *University of Aveiro*, ap@ua.pt
Paulo Ferreira, *University of Aveiro*, pjf@ua.pt
Paula Brito, *University of Porto*, mpbrito@fep.up.pt
Vera Afreixo, *University of Aveiro*, vera@ua.pt

Abstract. In this study we explore the potentialities of the inter-word distances to detect exceptional genomic words (oligonucleotides) in several species, using whole-genome analysis. We confront the empirical results obtained from the complete genomes with the corresponding results obtained from the random background. We develop a procedure, based on some statistical properties of the global distance distributions in DNA sequences, to discriminate words with exceptional inter-word distance distribution and to identify distances with exceptional frequency of occurrence. We identify the statistically exceptional words in whole-genomes, i.e., words with unexpected inter-word distance distributions, and we suggest species signatures based on exceptional word profiles.

Keywords. inter-oligonucleotide distances, DNA sequence, exceptional genomic word, stochastic model, goodness of fit.

1 Introduction

Several authors tried to identify exceptional words using different statistical criteria. A standard approach to detect exceptional words relies on their frequency. For example, based on genomic word frequencies and on comparisons between those frequencies and the random background (e.g. [10, 15, 16]).

The distance between two successive occurrences of a pattern in strings has been thoroughly studied and theoretical results have been deduced, in particular the generating functions of the waiting times to return to a specific pattern (e.g., [14, 18]). The probability mass function of the waiting times to return for the first time to a specific genomic word, or inter-word distance distribution, can be obtained by the Markov chain embedding technique, first developed by Fu (see, for example, [6]).

There are some interesting and counter-intuitive relations between frequency and distance distributions. Thus, the two perspectives are worth of separate investigation.

The inter-nucleotide distance (i.e., the distance between successive occurrences of the same nucleotide) has been previously explored to compare the complete genomes of several organisms; this comparison was based on genome distance distributions explored by [2]. The inter-nucleotide distance was also explored in the context of genome annotation by [11]. In [3], the inter-dinucleotide distance distribution was proposed and a comparison between all dinucleotide distributions in the human genome was performed. Note that in [3] overlapping dinucleotides were excluded from analysis, so that the expected distance distribution under an independent nucleotide model is a geometric distribution. Based on an inter-CpG distance, a CpG-island detection algorithm was proposed by [8], where a geometric distribution was used as a reference for comparison.

In this paper, we describe a procedure to highlight exceptional words that is based on inter-word distance distributions, rather than word frequencies. The subtraction of the random background from the counting result (under an independent nucleotide placement assumption) has been suggested as a way of emphasizing the contribution of selective evolution ([12, 5]). Based on this biologic perspective, we take a nucleotide independent model as the departing point and evaluate the discrepancy between real sequences and random background.

2 Materials and methods

Materials

In this study, we used the complete DNA sequences of 30 species, listed in Table 1, downloaded from the website of the National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov/genomes`). For each species, we processed the available assembled chromosomes as separate sequences. In each sequence, we studied every word formed by k consecutive unambiguous nucleotides, with $1 < k \leq 5$. The analysis included words partially overlapping preceding or succeeding words. All ambiguous or unsequenced nucleotides, i.e., all non-ACGT symbols, are considered word delimiters.

Methods

Inter-word distance

Consider the alphabet formed by the four nucleotides $\mathcal{A} = \{A, C, G, T\}$, and let s be a symbolic sequence of length N defined in $\mathcal{A}$. For each nucleotide $x \in \mathcal{A}$, consider a numerical sequence, d^x (or simply d), that represents the inter-nucleotide distances between each occurrence of symbol x and the previous occurrence of the same symbol, i.e., the differences between the positions occupied by successive occurrences of symbol x. As an example, we show the four inter-nucleotide distance sequences for $s = AAACGTCGATCCGTG$:

$$d^A = (1, 1, 6),\ d^C = (3, 4, 1),\ d^G = (3, 5, 2),\ d^T = (4, 4).$$

A genomic word, or oligonucleotide (w), is a sequence of length k defined in $\mathcal{A}$. We can extend the notion of inter-nucleotide distance to the case of oligonucleotides. Assuming that the

Table 1: List of DNA builds used for each species

Species	Biological taxonomy	Abbr.
Homo sapiens (human)	animalia	H.sapiens
Macaca mulatta (Rhesus macaque)	animalia	M.mulatta
Pan troglodytes (chimpanzee)	animalia	P.troglodytes
Mus musculus (mouse)	animalia	M.musculus
Rattus norvegicus (brown rat)	animalia	R.norvegicus
Eqqus caballus (horse)	animalia	E.caballus
Cannis lupus familiaris (dog)	animalia	C.lupus
Bos taurus (cow)	animalia	B.taurus
Monodelphis domesticus (opossum)	animalia	M.domesticus
Ornithorhynchus anatinus (platypus)	animalia	O.anatinus
Danio rerio (zebrafish)	animalia	D.rerio
Apis mellifera (honey bee)	animalia	A.mellifera
Arabidopsis thaliana (thale cress)	plantae	A.thaliana
Vitis vinifera (grape vine)	plantae	V.vinifera
Saccharomyces cerevisiae str	fungi	S.cerevisiae
Schizosaccharomyces pombe	fungi	C.pombe
Escherichia coli	bacteria	E.coli
Helicobacter pylori	bacteria	H.pylori
Streptococcus pneumoniae	bacteria	S.pneumoniae
Streptococcus mutans LJ23	bacteria	S.mutansLJ
Streptococcus mutans GS	bacteria	S.mutansGS
Aeropyrum pernix str.K1	archaea	A.pernix
Nanoarchaeum equitans	archaea	N.equitans
Candidatus korarchaeum	archaea	C.korarchaeum
Caldisphaera lagunensis	archaea	C.lagunensis
Aeropyrum camini	archaea	A.camini
NC001341 virus	virus	vir.001341 virus
NC001447 virus	virus	vir.001447 virus
NC004290 virus	virus	vir.004290 virus
NC011646 virus	virus	vir.011646 virus

sequence is read through a sliding window of length k, we can define the inter-oligonucleotide (inter-w) distance sequence d^w as the differences between the positions of the first symbol of consecutive occurrences of that oligonucleotide. For example, the inter-CG distance sequence for the short DNA segment above is $d^{CG} = (3, 5)$.

Reference distribution under a nucleotide independence model

Let $w = x_1x_2x_3 \dots x_k \in \mathcal{A}^k$ be a generic oligonucleotide and D be the random variable that represents the inter-oligonucleotide distance, from a sequence whose nucleotides are independently generated.

The reference distribution of inter-w distances can be deduced using a state diagram, which represents the progress made towards identifying w as each symbol is read from the sequence. The state diagram has $k + 1$ states. The first k states, $S_0, S_1, \dots, S_{k-1}$, represent intermediate points in the process and state S_k is the final, absorbing state. In the diagram, being in state S_i means that the last i symbols read from the sequence match a prefix of w. As each new symbol is read, a transition occurs from S_i to a new state S_j, until the final, or absorbing, state S_k is reached, meaning that a new occurrence of w has just been identified in the sequence.

We define the distance to the next occurrence of w, starting from an initial state S_I ($I < k$), as the number of steps (transitions) it takes to walk through the diagram from S_I until the final state S_k is reached. The initial state is given by the longest word overlap of w, different from w.

To illustrate this procedure, we present the state diagram for inter-ACG distances in Figure 1. In this specific case, the probability of transition between two non-absorbing states, S_i to S_j, is given by element m_{ij} $(0 \leq i, j \leq 2)$ of the the transition matrix

$$M_{ACG} = \begin{bmatrix} 1-p_A & p_A & 0 \\ 1-p_A-p_C & p_A & p_C \\ 1-p_A-p_G & p_A & 0 \end{bmatrix}.$$

where p_x denotes the nucleotide probability $(x \in \mathcal{A})$. Distance one between two occurrences of ACG is only possible from state S_2. Thus, the probabilities of distance one, from each non-absorbing state are

$$P(D=1) = \begin{bmatrix} P(D=1|S_0) \\ P(D=1|S_1) \\ P(D=1|S_2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ p_G \end{bmatrix}.$$

For higher distances, $d > 1$, the probabilities can be found by combining the transition probabilities for the first step with the probabilities for distance $d-1$, which leads to the recurrence relation

$$\begin{bmatrix} P(D=d|S_0) \\ P(D=d|S_1) \\ P(D=d|S_2) \end{bmatrix} = M_{ACG} \times \begin{bmatrix} P(D=d-1|S_0) \\ P(D=d-1|S_1) \\ P(D=d-1|S_2) \end{bmatrix}$$

where M_{ACG} is the transition matrix of non-absorbing states. Since ACG has only null word overlap besides itself, we must consider S_0 as the initial state. Therefore, under an independent symbol model, the reference probability distribution of inter-ACG distances is given by

$$f(d) = P(D=d|S_0).$$

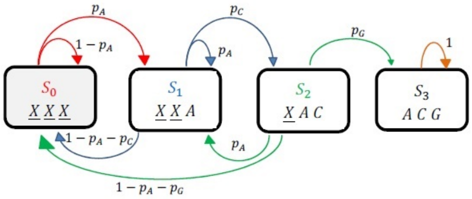


Figure 1: State diagram associated to inter-ACG distances (initial state S_0).

For the generic word w, the reference distance distribution under the independent nucleotide model is given by $f(d) = P(D=d|S_I)$, with

$$\begin{bmatrix} P(D=d|S_0) \\ \vdots \\ P(D=d|S_{k-1}) \end{bmatrix} = M^{d-1} \times \begin{bmatrix} P(D=1|S_0) \\ \vdots \\ P(D=1|S_{k-1}) \end{bmatrix},$$

and

$$P(D=1) = \begin{bmatrix} 0 & \dots & 0 & p_{x_k} \end{bmatrix}^T.$$

where p_{x_k} is the occurrence probability of nucleotide x_k and M is the transition matrix of non-absorbing states.

Our approach to obtain the exact distribution of inter-word distances is a special case of Fu's procedure based on finite Markov chain embedding [6, 7]. To find the transition matrix for a given word requires "a deep understanding of the structure of the specified pattern" [6]. Next, we propose a general expression to compute the transition matrix of non-absorbing states $M = [m_{ij}]$, with $i, j = 0, \ldots, k-1$, based on the concept of word overlap.

Let us denote by $\mathcal{L}(w_1, w_2)$ the length of the longest overlap (a suffix of w_1 that matches with a prefix of w_2) between words w_1 and w_2. Being in S_i means we have just read symbols that match w^i. The next symbol x, appended to w^i, determines the next state. A transition from S_i to S_j with $j > 0$ is only possible if $\mathcal{L}(w^i x_j, w) = j$, so its probability is

$$\text{for } j > 0, \quad m_{ij} = \begin{cases} p_{x_j} & , \quad \mathcal{L}(w^i x_j, w) = j \\ 0 & , \quad \text{otherwise} \end{cases} .$$

And the probability of a transition from S_i to S_0 ($j = 0$) is given by the complementary probability

$$m_{i0} = \begin{cases} 1 - p_{x_{i+1}} - \sum_{s=1}^{i} m_{is} & , \quad i \geq 1 \\ 1 - p_{x_{i+1}} & , \quad i = 0 \end{cases} .$$

The reference distribution under independent nucleotide structure, that we just described, can easily be computed for any whole-genome and for any genomic word, using only four input parameters: the nucleotide frequencies in the sequence.

Measures

To evaluate the goodness of fit between the inter-oligonucleotide distance distribution and the corresponding reference distribution we used the chi-square statistic and the phi coefficient. We also used an effect size measure, Cohen's d, to identify the existence of exceptional distances inside the distribution of a single word.

Due to the sensitivity of these measures to low frequencies that occur for longer distances, we made a cutoff at the 99th percentile of the empirical distribution, $d_{0.99}$. Then, we grouped all distances larger than $d_{0.99}$ in one residual class, $\widetilde{d} = d_{0.99} + 1$.

The empirical distance distribution is given by

$$q_i = \frac{n_i}{N'}, \text{ for } i = 1, \ldots, d_{0.99}$$

and the remaining frequency, $q_{\widetilde{d}}$, where n_i is the number of occurrences of distance i and N' is the total number of inter-w distances. In order to match the size of the reference distribution to the empirical distribution we also made a cutoff in the reference distribution, at $d_{0.99}$.

To extract the exceptional words of each species, we compare the empirical distribution to the corresponding reference distribution under the nucleotide independence (model I). A word is considered exceptional if the empirical inter-word distance and the reference distribution are distinct in a statistically precise way. There are two cases to consider: either the two distributions show a global misfit or there is at least one distance value that deviates significantly from the reference distribution. In the first case, the empirical distribution shows a global misfit to the random background; in the second case, the misfit is more noticeable for specific distances.

To test the goodness of the fit between the empirical and the reference distributions, for each oligonucleotide w, we can use a chi-square statistic, denoted by X_w^2,

$$X_w^2 = \sum_{i=1}^{d} \frac{(n_i - f_i \cdot N')^2}{f_i \cdot N'}.$$

To obtain an effect size measure to evaluate the lack of goodness of fit, we use the phi coefficient, denoted by φ_w,

$$\varphi_w = \sqrt{\frac{X_w^2}{N'}}.$$

A perfect fit between the distributions corresponds to $\varphi_w = 0$. We consider a value above 0.10 as a descriptor for small effect size, above 0.30 for medium effect size, above 0.50 for large effect size ([4]), above 0.60 for strong effect size and above 0.80 for a very strong effect size ([13])

For each inter-w distance distribution we are interested in identifying and evaluating the existence of exceptional distances, i.e., distances that occur with a frequency much higher than the expected value. In order to obtain a standard score able to compare how exceptional a distance is over all oligonucleotides of the same length, we use Cohen's d given by

$$CD_i = \frac{q_i - f_i}{\sqrt{f_i(1 - f_i)}}.$$

For reporting and interpreting Cohen's d, we considered a value above 0.20 as a descriptor for small effect size, above 0.50 for medium effect size and above 0.80 for large effect size ([4]). We established those acceptance thresholds as the levels above which the distance is considered exceptional or very exceptional, respectively.

To identify the most exceptional distance inside a distribution, if there is one, we use Cohen's d effect size. After computing Cohen's d for all distances up to the 99th percentile, we identify the distance d for which the maximum Cohen's d is attained and consider it the candidate to the most exceptional distance of the distribution, i.e., $C_d = \max\{CD_i : i = 1, \ldots, d_{0.99}\}$.

The expected values for distances less than or equal to k (the word length) can be null for certain words. For example, the distances between the word AAA in the text $AAAAAAA \cdots$ can never be 2 or 3. Such zero distances were not considered in the computation of the mentioned measures.

3 Results and discussion

Exceptional distance distributions in human genome

We are interested in exceptional distributions, i.e., empirical distributions that either show a significant global misfit to the reference distribution or that exhibit frequencies much higher than expected for specific distances. For all words, we observe the existence of statistical significant differences between empirical and reference distributions ($p\text{-}value < 0.001$).

In order to evaluate the lack of fit phenomenon over all words of the same length, we computed the phi coefficient, φ_w, and sorted the word distance distributions according to the value of φ_w. We observe that CG-rich words (i.e., words comprising one or more CG) and words with long word overlap lead to the poorest goodness of fit, in relation to the reference model (see Table 2).

This means that these word distributions have a global misfit or a few distances with exceptional misfit to the reference distribution, in a whole-genome analysis. Let us note that the top-two dinucleotides correspond to well known local motifs (recurrent CG pairs in CpG islands and the TATA binding boxes on transcription start sites). Other high-scoring words may be related to biological motifs.

Conversely, we observe that words with no overlap and without CGs attained the lowest divergences.

Table 2: Phi coefficient between empirical and reference distributions, in the *Homo Sapiens* genome. The maximum and minimum φ_w, the words distributions which present the ten largest and the ten smallest values of φ_w, organized by word length (k).

k	1	2	3	4	5
max(φ_w)	0.191	1.72e+05	3.11e+05	3.84e+12	8.84e+19
min(φ_w)	0.136	0.209	0.116	0.101	0.127
highest φ_w	C	CG	CGA	CGCG	ACGCG
2nd highest	G	TA	TCG	CGAC	CGCGT
3rd highest	-	CC	CGC	GTCG	CGTCG
4th highest	-	GG	GCG	ATCG	CGACG
5th highest	-	GC	ACG	TACG	CGCGA
6th highest	-	AT	CGT	CGTA	TCGCG
7th highest	-	AC	CCG	TCGA	CGGCG
8th highest	-	GT	CGG	TTCG	CGCCG
9th highest	-	-	ATA	CGAA	CGATA
10th highest	-	-	TAT	TCGT	TATCG
⋮					
10th lowest	-	-	TGT	ACTT	CTCTA
9th lowest	-	-	ACA	AAGT	TAGAG
8th lowest	-	AA	CAA	GACA	TCAGT
7th lowest	-	TT	TTG	TGTC	TGACT
6th lowest	-	AG	ACT	ATCT	AGTCA
5th lowest	-	CT	AGT	AGAT	ACTGA
4th lowest	-	TC	TCA	ATGC	AAGCT
3rd lowest	-	GA	TGA	GCAT	AGCTT
2nd lowest	T	CA	ATG	GCTT	AGAGT
lowest φ_w	A	TG	CAT	AAGC	ACTCT

It is known that the human genome has low CG content ([9]). For inter-oligonucleotide distances, the information about CG content ($k = 2$) or CG-rich word ($k > 2$) contents in the sequence is not included in model I. Under this assumption, CG-rich words reach higher phi coefficients and, as a consequence, these words will be identified as exceptional words.

Using Cohen's d, we explored the existence of exceptional distances inside a single distribution, i.e., specific distances with an occurrence probability much higher than expected. Consider, for example, the unexpected spike at distance 24 in the inter-TGCA distance distribution, $C_{24} = 0.616$ (Figure 2).

Note that a high Cohen's d could result from a generalized misfit between the empirical and the reference distribution, rather than from a genuine exceptionality of that distance. Thus, we suggest a practical decision based on the goodness of fit between empirical and reference distance distributions: for one empirical distance distribution that presents moderate to strong discrepancy ($0.2 < \varphi_w < 0.8$) we use 0.5 as the cut point on Cohen's d to identify exceptional distances. For the human genome, only eleven inter-word distributions have been identified as comprising exceptional distances. We do not observe the presence of exceptional distances in

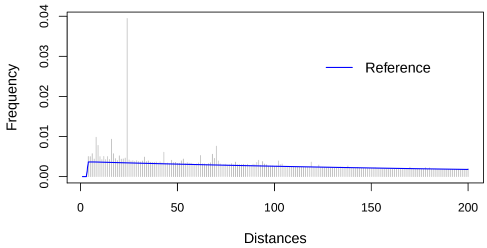


Figure 2: Empirical distance distribution *vs* reference distribution: w =TGCA, $\varphi_w = 0.694$, $C_{24} = 0.616$.

distance distributions for word lengths less than 4 (see Table 3). Figure 3 shows two inter-word distance distributions that comprise an exceptional distance, by our criteria. This procedure detects exceptional words based on their atypical distance distribution along the sequence and not on their frequency of occurrence.

Table 3: Number of distance distributions with moderate or strong lack of fit ($0.2 < \varphi_w < 0.8$) that present an exceptional distance, organized by strength of effect size and word length.

Strength of Cohen's d maximum	word length			
	2	3	4	5
medium effect size ($0.5 \leq C_d < 0.8$)	0	0	1	10
large effect size ($C_d \geq 0.8$)	0	0	0	0

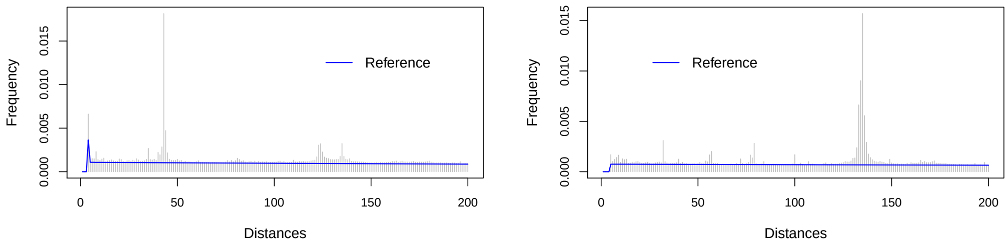


Figure 3: Empirical distance distribution *vs* reference distribution: w =TCACT, $\varphi_w = 0.633$, $C_{43} = 0.533$ (left); w =ATCCC, $\varphi_w = 0.791$, $C_{135} = 0.577$ (right).

This procedure may lead to the identification of new motifs. For example, a word with a perfectly ordinary overall frequency of occurrence may exhibit an abnormal "preference" for occurring at a distance d from the previous occurrence and a slightly decreased preference for occurring at other distances.

Analysis of multiple organisms

Taking into account the empirical distance behaviour and the random background (model I), we introduce exceptionality word criteria and define dichotomic vectors, that may be used as a genomic signature of species.

Consider the following exceptionality word criteria:

- Misfit criterion: the word shows a very strong dissimilarity effect between distributions, $\varphi_w > 0.8$, highlighting the contribution of selective evolution [12];
- Peak criterion: the word has a small or medium dissimilarity effect between distributions and presents a peak with medium or large effect size, $0.2 < \varphi_w < 0.8 \land C_d > 0.5$.

Consider, for each specie, a dichotomic vector that marks as nonzero the words identified as exceptional accordingly to one of the criteria. These vectors allows to build dendrograms, which could then be interpreted as phylogenetic trees.

We performed a hierarchical analysis of the 30 species listed in Table 1, considering each one of the exceptionality criteria. The dendrograms were build using the average linkage method. The similarity matrix was computed using the Euclidean distance. In the case of the *misfit criterion*, the dendrogram displays a first branching between eukaryotes and non-eukaryotes (Figure 4a). Inside the eukaryote cluster, we observe that some related species are grouped in the same branch. For instance, primates (*H.sapiens*, *P.troglodytes* and *M.mulatta*), the rodentia (*M.musculus* and *R.norvegicus*) and the fungi (*S.cerevisiae* and *C.pombe*). In the second branch it is observed that, in general, bacteria and archaeotas are closer to each other and separated from the virus. We also notice that the bacteria *S.mutansLJ*, *S.mutansSG* and *S.pneumoniae* are in the same cluster. We emphasize that only the animal organisms reveal distance distributions that verify the *peak criterion.* Restricting the analysis to animal organisms, we obtain a dendrogram which reveals the group of primates and the group of rodentia (Figure 4b).

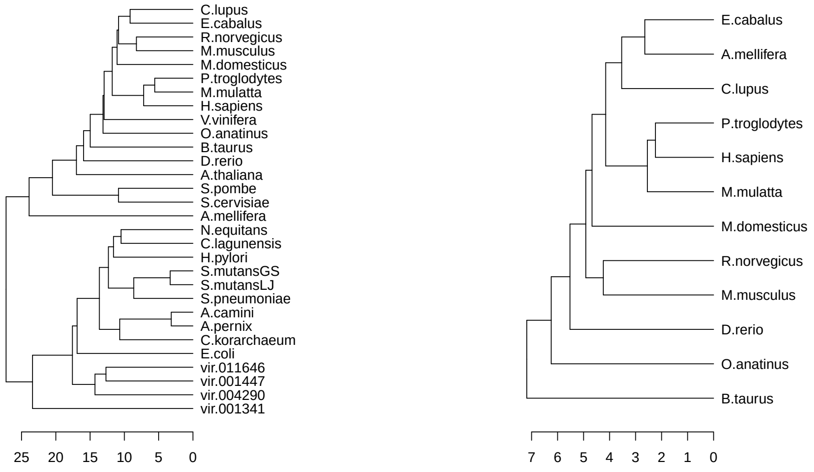
Thus, the binary vector of exceptional words defined by the *misfit criterion* may be used as a genomic signature in all the studied species, while the *peak criterion* can only be used as genomic signature in animal species.

We also constructed dendrograms for the 10 mammal species, using both criteria separately. The obtained dendrograms present some similarities (the split distance between dendrograms is 0.43). We observe that primates are clustered together, as well as the rodentia (Figure 5). These dendrograms support several evolutionary relationships between species. For example, the split distance between our dendrograms and those presented in [17], based in alignment and non-alignment algorithms, is around 50%, which is lower than in random scenarios (see [1]).

4 Conclusions and future research

In this work we studied the inter-word distances in the complete genomes of up to 30 species, for word length k varying between 1 and 5.

We intended to detect exceptional words by comparing the empirical distribution of the inter-word distances with the theoretical one under independent nucleotide model, taking the word overlap structure into account. We evaluated the discrepancy between real sequences and the random background, as a way of emphasizing the contribution of selective evolution. The comparison of the empirical distance frequencies with those that would be observed if the



(a) *misfit criterion*

(b) *peak criterion*

Figure 4: Dendrogram of the 30 organisms, with binary vector of exceptional words defined by all words of length 2 to 5 by *misfit criterion* (left); and of the animals by *peak criterion* (rigth).

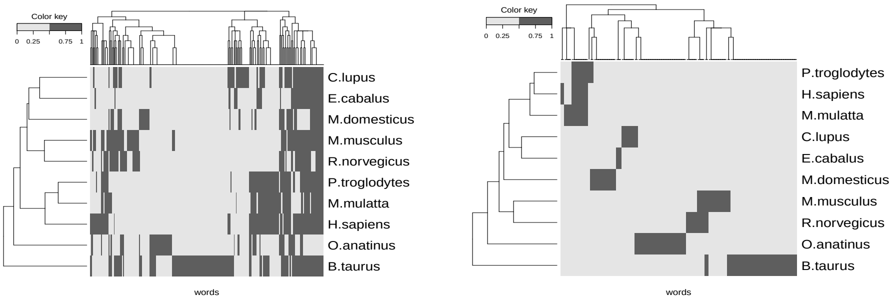


Figure 5: Heat map of mammal species *vs* exceptional words. Binary vectors of exceptional words defined by *misfit criterion* (left) and by *peak criterion* (rigth), considering only vectors with variation.

random background model were valid, allowed us to highlight distinct distance distributions for classes of genomic words.

We introduced a statistical procedure to automatically identify genomic words whose distance distributions show a significant discrepancy from the random background. Our procedure allows to detect some words with a very high lack of fit. These were, in general, words with CG-rich content (as expected). Moreover, we found words with a moderate to strong lack of fit and an unexpected strong spike. Only less than 1 percent of the words of length 4 and 5 show this kind of exceptional distance distribution.

We believe that this procedure, which detects statistically exceptional distributions, may lead to the identification of new motifs. For example, a word with a perfectly ordinary overall frequency of occurrence may exhibit an abnormal "preference" for occurring at a distance d from the previous occurrence and a slightly decreased preference for occurring at other distances.

We also found that the differences mimic, to a certain extent, the evolutionary relationships between the species, which were used to construct dendrograms and perform evolutionary comparisons. In the mammalian organisms, we found matching word dissimilarity values.

In future we intend to extend our procedure to longer words, and evaluate if the method allow to point out known patterns with biological significance. Furthermore, since whole genome are highly heterogeneous, we also expect to perform analysis for detection of regions with exceptional inter-nucleotide distances.

5 Acknowledgements

This work was supported by Portuguese funds through the iBiMED - Institute for Biomedicine, IEETA - Institute of Electronics and Informatics Engineering of Aveiro and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within projects UID/BIM/04501/2013 and UID/CEC/00127/2013.

Bibliography

[1] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, and P. J. Ferreira. Genome analysis with distance to the nearest dissimilar nucleotide. *Journal of theoretical biology*, 275(1):52–58, 2011.

[2] V. Afreixo, C. A. C. Bastos, A. J. Pinho, S. P. Garcia, and P. J. S. G. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23):3064–3070, Dec. 2009.

[3] C. A. C. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. a. M. O. S. Rodrigues, and P. J. S. G. Ferreira. Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions. *Journal of Integrative Bioinformatics*, 8(3):172, 2011.

[4] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum, 1988.

[5] S. Ding, Q. Dai, H. Liu, and T. Wang. A simple feature representation vector for phylogenetic analysis of dna sequences. *Journal of Theoretical Biology*, 265(4):618–623, Aug. 2010.

[6] J. C. Fu. Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica*, 6(4):957–974, 1996.

[7] J. C. Fu and W. W. Lou. *Distribution theory of runs and patterns and its applications: a finite Markov chain imbedding approach.* World Scientific, 2003.

[8] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, and J. L. Oliver. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, 7(1):446, 2006.

[9] J. Karro, M. Peifer, R. Hardison, M. Kollmann, and H. von Grünberg. Exponential decay of GC content detected by strand-symmetric substitution rates influences the evolution of isochore structure. *Molecular biology and evolution*, 25(2):362–374, 2008.

[10] M. Lothaire. *Applied combinatorics on words*, volume 105. Cambridge University Press, 2005.

[11] A. S. S. Nair and T. Mahalakshmi. Visualization of genomic data using inter-nucleotide distance signals. In *Proceedings of IEEE Genomic Signal Processing*, 2005.

[12] J. Qi, B. Wang, and B.-I. Hao. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *Journal of molecular evolution*, 58:1–11, 2004.

[13] L. Rea and R. Parker. *Designing and conducting survey research: a comprehensive guide.* Public Administration Series. Jossey-Bass Publishers, 1992.

[14] S. Robin. A compound Poisson model for word occurrences in DNA sequences. *Applied Statistics*, 51, Part 4:437–451, Aug. 2002.

[15] S. Robin, F. Rodolphe, and S. Schbath. *DNA, Words and Models: Statistics of Exceptional Words.* Cambridge University Press, 2005.

[16] S. Robin, S. Schbath, and V. Vandewalle. Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics*, 8(1):84, 2007.

[17] G. E. Sims, S.-R. Jun, G. A. Wu, and S.-H. Kim. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences*, 106(40):17077–17082, 2009.

[18] T. V. Stefanov. The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach. *Journal of Applied Probability*, 40:881–892, 2003.

Intentionally blank page.

Chapter 7

Article IV

DNA word analysis based on the distribution of the distances between symmetric words

Published:

Tavares, A. H., Pinho, A. J., Silva, R. M., Rodrigues, J. M., Bastos, C. A., Ferreira, P. J., Afreixo, V. (2017). DNA word analysis based on the distribution of the distances between symmetric words. Scientific reports, 7(1), 728.

DOI: 10.1038/s41598-017-00646-2

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 11 August 2016
Accepted: 2 March 2017
Published online: 07 April 2017

DNA word analysis based on the distribution of the distances between symmetric words

Ana H. M. P. Tavares[1,2], Armando J. Pinho[3,4], Raquel M. Silva[2,4], João M. O. S. Rodrigues[3,4], Carlos A. C. Bastos[3,4], Paulo J. S. G. Ferreira[3,4] & Vera Afreixo[1,2,4]

We address the problem of discovering pairs of symmetric genomic words (i.e., words and the corresponding reversed complements) occurring at distances that are overrepresented. For this purpose, we developed new procedures to identify symmetric word pairs with uncommon empirical distance distribution and with clusters of overrepresented short distances. We speculate that patterns of overrepresentation of short distances between symmetric word pairs may allow the occurrence of non-standard DNA conformations, such as hairpin/cruciform structures. We focused on the human genome, and analysed both the complete genome as well as a version with known repetitive sequences masked out. We reported several well-defined features in the distributions of distances, which can be classified into three different profiles, showing enrichment in distinct distance ranges. We analysed in greater detail certain pairs of symmetric words of length seven, found by our procedure, characterised by the surprising fact that they occur at single distances more frequently than expected.

The similarity between the frequency of complementary nucleotides in a single strand of DNA is known as Chargaff's second parity rule[1]. An extension to this parity rule suggests that, for each DNA strand, the proportion of an oligonucleotide (a sequence of adjacent nucleotides, also referred to as a genomic word) should be similar to that of its reversed complement, a property that has been studied both for prokaryotes and eukaryotes[2,3].

The origin of single strand symmetry is a topic of great interest, because it can contribute to the study of the origin and evolution of genomes. Currently, there is no single accepted justification for the intra-strand symmetry, although several hypotheses about its origin have been proposed[4]. It has been suggested that the occurrence of secondary DNA structures, such as stem-loops and cruciforms, is associated with the DNA symmetry phenomenon. Cruciforms are structures with four arms that can be formed at sites containing reversed complementary words. They are relevant in biological processes, including those of replication and transcription, recombination and translocation[5]. Because these structures are associated with genome instability, the determination of their occurrence in the human genome and the identification of the corresponding sequence motifs is of paramount importance, both in the context of disease development and evolutionary events[6,7].

Here, we address the distance distribution of symmetric word pairs and investigate the different distance profiles in the human genome. In particular, we develop a procedure to identify genomic words with patterns of overrepresented short distances (<1000 bp). Overrepresented distances are those that have observed frequency higher than the expected frequency predicted by an adequate model, in a statistically significant way. We suggest that patterns of overrepresentation of short distances between reversed complements may be related to the occurrence of cruciform structures, and we evaluate this hypothesis in the human genome. We study the distance distribution between reversed complements, in order to provide knowledge about the words that are strong candidates to the formation of cruciform structures in human DNA. Procedures based on inter-word distances have already been found useful to study genomic sequences, e.g., to detect CpG islands[8] and to compare species[9]. The study addressed in this paper shows yet another use of inter-word distances and distance distributions, which may lead to a deeper understanding of intra-strand symmetry and its connection with secondary DNA structures.

[1]Department of Mathematics & CIDMA, University of Aveiro, Aveiro, Portugal. [2]Department of Medical Sciences & iBiMED, University of Aveiro, Aveiro, Portugal. [3]Department of Electronics, Telecommunications and Informatics, University of Aveiro, Aveiro, Portugal. [4]IEETA, University of Aveiro, Aveiro, Portugal. Correspondence and requests for materials should be addressed to V.A. (email: vera@ua.pt)

Materials and Methods

Materials. We used the complete DNA sequences of the human genome, downloaded from the website of the National Center for Biotechnology Information (NCBI). We processed the available assembled chromosomes (GRCh38.p2) as separate sequences. All ambiguous or unsequenced nucleotides, i.e., all non-ACGT symbols, were considered sequence delimiters.

We also used pre-masked sequences[10] available from the UCSC Genome Browser (http://genome.ucsc.edu) downloads page. These files contain the same GRCh38 assembly sequences, but with repeats reported by RepeatMasker[11] and Tandem Repeats Finder[12] masked by Ns.

To address the problem of possible assembly artefacts, we also used the whole-genome shotgun assembly (WGSA, to which we refer as "Celera") of the human genome generated at Celera in December 2001[13], and the May 2007 HuRef genome of J. Craig Venter, sequenced with capillary-based whole-genome shotgun technologies using the Applied Biosystems 3730xl DNA analyser, and de novo assembled with the Celera Assembler[14], to which we refer as "HuRef".

Distance between symmetric word pairs. Consider the alphabet $\mathcal{A} = \{A,\ C,\ G,\ T\}$ and let w be a symbolic sequence (word) defined in $\mathcal{A}^k$, where k is the length of w. In this work, the pair composed by one word, w, and the corresponding reversed complement word, w', is called a symmetric word pair. For example, (AC, GT) is a symmetric word pair.

We are interested in finding the distance between a given w and w', with no w or w' between them. As an example, consider $w = AC$ and the sequence $\underline{AC}T\underline{AC}TCC\overline{GT}\underline{AC}TATA\overline{GT}C\overline{GT}$. In this example, there are three occurrences of the word AC (underlined), but only the 2nd and the 3rd occurrences are considered for the calculation of distances to their nearest reversed complements (overlined), since between the 1st and the 2nd occurrences of w there are no occurrences of w'. Distances are measured between the start positions of the words, so a distance d between reversed complements of length k implies that the words are separated by $(d - k)$ intervening nucleotides. In this example, $d = 5$ for the first (AC, GT) pair and $d = 6$ for the second.

Distances $d < k$ may only occur if a suffix of w matches a prefix of w'. On the other hand, $d = k$ is impossible for words such as $CGCG$. To avoid this dependence on the specific composition of w, distances $d \leq k$ are not considered for analysis.

The distribution of the *distances of nearest reversed complements* (*DNRC*) is denoted as $f_{w,w'}$. Note that $f_{w,w'}$ may be different from $f_{w',w}$.

For a fixed word length, k, we are also interested in the overall DNRC distribution across all the symmetric word pairs. We define the *global DNRC distribution*, f_k, as a weighted sum of the DNRC distributions of all symmetric word pairs with words of length k,

$$f_k(d) = \sum_{w,w' \in \mathcal{A}^k} \frac{n_{w,w'}}{n} f_{w,w'}(d), \quad d > k, \tag{1}$$

where $n_{w,w'}$ is the number of observations of nearest pairs of symmetric words (w, w') of length k, and n is the total number of such distances. Only the analysed distances $(d > k)$ are counted in n and $n_{w,w'}$.

For generating the symmetric words, we used a simple algorithm that, for each position in the DNA sequence, i, and associated word of size k, w, searches for the first occurrence of w'. If w is found before w', the algorithm skips to the next position i. For practical reasons, a maximum searching distance is specified by the user, allowing the program to maintain in memory a table with all possible words w and the corresponding number of occurrences at each distance.

In order to study the behaviour of the empirical global DNRC distributions of the human genome, f_k, we carried out comparisons with the DNRC distributions obtained from nucleotide sequences generated by a k-order Markov process (random background). The expected global DNRC distribution under k-order Markov dependence, f_k^e, can be deduced using the transition probabilities and a state diagram that represents the progress made towards identifying w or w' as each symbol is read from the sequence. The algorithm used to find this exact distribution[15] is a special case of Fu's procedure based on finite Markov chain embedding[16].

Parameter assumptions. The stem and loop lengths of hairpin/cruciform structures seem to vary over a wide range. According to different authors, the stem length varies between 6 and 100 nucleotides, while loop lengths may range from 0 to 2000 nucleotides[6, 17, 18].

Since this study intends to characterise the short distances between symmetric words, but avoiding the direct word dependencies, a range of distances from $(k + 1)$ to 1000 was considered for computing all the DNRC distributions. Taking into account computational limitations and the possible stem length of cruciform structures, the histograms of the DNRC were computed for all symmetric word pairs of lengths up to seven, for all human chromosomes. For each of these sequences, a global DNRC distribution, comprising all symmetric word pairs of the same length, was also determined.

Chromosome homogeneity. To assess the homogeneity of the global DNRC distribution, for a fixed k, among all chromosomes of the genome, we used the phi coefficient,

$$\varphi_k = \sqrt{\frac{\chi_k^2}{n}}, \tag{2}$$

where n is the total number of DNRC counts, as defined in (1), and χ_k^2 is the Pearson's chi-squared statistic,

$$\chi_k^2 = \sum_{w,j} \frac{(O_{w,j} - E_{w,j})^2}{E_{w,j}}, \tag{3}$$

where $O_{w,j}$ is the observed frequency count of distances from w to w' in chromosome j, and $E_{w,j}$ is the expected frequency count under homogeneity, with $w \in \mathcal{A}^k$ and $j \in \{1, \ldots, 22, X, Y\}$.

The assumption of homogeneity of the distance distributions of the chromosomes allows us to discuss the statistical properties of the complete genome based on a sequence with all chromosomes concatenated.

Residual analysis. From the perspective of molecular evolution, DNA sequences may reflect both the results of random mutation and of selective evolution. In order to highlight the contribution of selective evolution, one should subtract the random background from the simple counting result[19, 20]. To this purpose, the global DNRC distributions expected under the k-order Markov dependence, f_k^e, were obtained and the goodness-of-fit was evaluated by the φ measure ($\varphi = 0$ reveals a perfect fit between the distributions). To explore the differences between the empirical and the expected distributions, a residual analysis was carried out through the calculation of standardised residuals for a given distance d, are given by

$$r(d) = \frac{f_k(d) - f_k^e(d)}{\sigma}, \quad d > k, \tag{4}$$

where n is the total number of observed distances between symmetric pairs of length k and $\sigma = \sqrt{f_k^e(d)\left(1 - \frac{f_k^e(d)}{n}\right)}$ is the standard deviation of a binomial distribution. These standardised residuals are used to highlight the contribution of the selective evolution on the relative position of the symmetric word pairs.

We recall that, under k-order Markov dependence assumption, each standardised residual has an asymptotic standard normal distribution[21].

The focus of this study is mainly in the short distances between symmetric word pairs, thus we fixed a maximal distance to 1000. The global Type I error was fixed to $\alpha = 5\%$ and, for each distance comparison test, it was correct to $0.05\,(1000 - k)$. So, absolute residuals greater than four are considered to be significant residuals.

Short distances between reversed complements may be related with the occurrence of cruciform structures, with maximum loop length of twenty nucleotides[6]. To identify a thresholding distance which may discriminate the overrepresented short distances from the underrepresented, we assumed that short distances up to the threshold are overrepresented and the others are underrepresented (this assumption makes sense under the hypothesis of enrichment of words able to form cruciform structures).

We determined the thresholding distance, d, as the distance that maximises the sum of the number of significant positive residues less than d and the number of significant negative residues greater than d. We defined a *discriminator function* as a sum of indicator functions (for example, $\mathbb{1}_{SP(i)} = 1$, if $r(d) > 4$)

$$R(d) = \sum_{i=k+1}^{d-1} \mathbb{1}_{SP(i)} + \sum_{i=d+1}^{1000} \mathbb{1}_{SN(i)}, \quad d > k, \tag{5}$$

where $SP(d) = \{d \mid r(d) > 4\}$ and $SN(d) = \{d \mid r(d) < -4\}$ and r is defined in Equation 4. The value d that maximises the discriminator function, R, was considered as the thresholding distance.

Results and Discussion

The global DNRC distribution was determined for each of the 24 human chromosomes and each word length ($k = 1, \ldots, 7$). These distributions have heavy tails, strongly affecting the chi-square statistic. To avoid this problem, for each k, a cutoff distance was defined as the 99th percentile of the DNRCs observed in the complete genome (all chromosomes). Distances larger than this cutoff were lumped together into a residual class, in each distribution. Naturally, the DNRCs and hence the cutoff distances were found to increase with word length in the human genome, as would be expected even in a sequence of randomly generated nucleotides.

We measured the degree of homogeneity (φ effect sizes measure) between the human chromosomes, for the global DNRC distributions. According to the obtained φ values ($\varphi < 0.04$), we conclude that the homogeneity effect is weak. Thus, we consider that there is homogeneity between the global DNRC distributions of the several chromosomes. This chromosome homogeneity in the global DNRC distributions points to a general feature of the complete human genome, which may be due to genomic architecture constrains.

Global DNRC distributions for the complete genome. The discrepancies between the global DNRC distribution in the human genome and in the k-order Markov process were measured by φ effect size measure. Although the misfit effect is not strong, it is nevertheless non-negligible. The φ values are always greater than 0.05 and the p-values smaller than 0.05.

Figure 1 shows the global DNRC distributions of the human genome and the global DNRC distributions of the k-order Markov random sequence, for $k = 6$ and $k = 7$. The misfit between the human distance distributions and the corresponding k-order Markov process is clear. Analysing the residuals between the empirical distribution and the distribution of this random background, we observe a tendency of overrepresentation of short distances in the human genome, for all analysed values of k.

Figure 2 presents the results of the residual discriminator function (the R profile), for the global DNRC distribution of the complete human genome (between the observed and the corresponding k-order Markov process), for $k = 6$ (left) and $k = 7$ (right). The discriminator functions increase for $d \lesssim 260$, showing an evident favouring

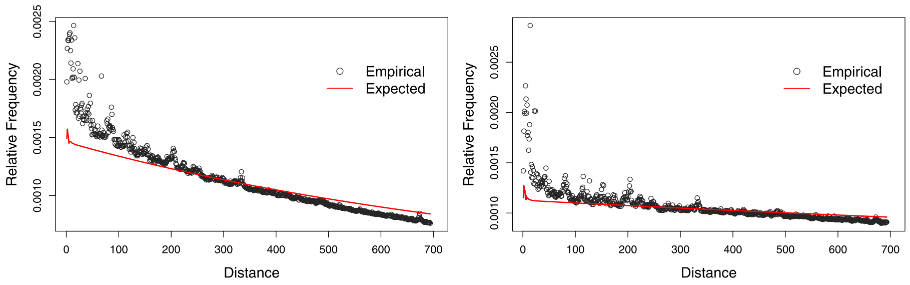


Figure 1. Empirical and expected global DNRC distributions, for the complete human genome, for $k=6$ (left) and $k=7$ (right). The expected distributions were obtained under the k-order Markov dependence assumption.

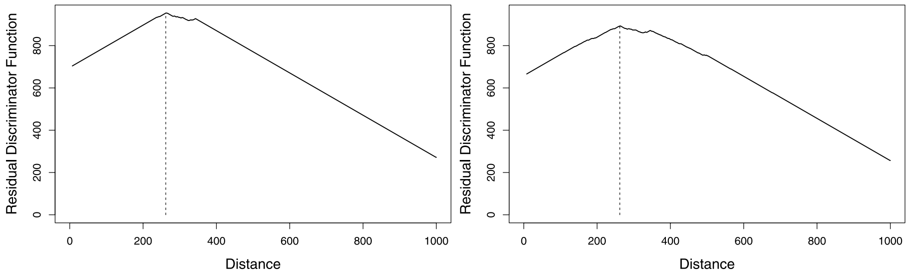


Figure 2. Residual discriminator function (R) for global DNRC distributions, relatively to the complete human genome, for $k=6$ (left) and $k=7$ (right). Both reach their maximum at distance 262.

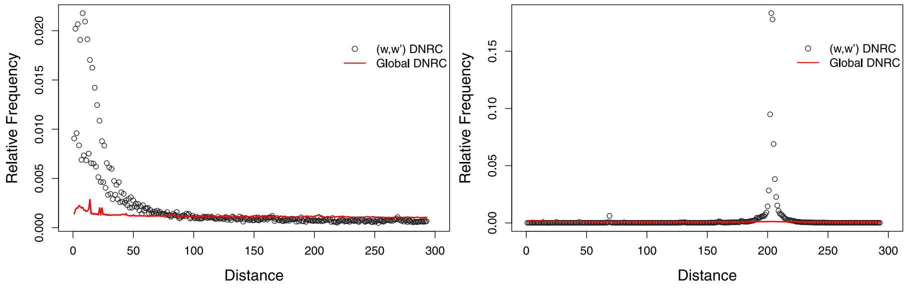


Figure 3. DNRC distribution, $f_{w,w'}$, and global DNRC distribution, $f_{k=7}$, for complete human genome. Overrepresentation of short distances in different ranges: $w = ATATATG$ (left), $w = GGCTCAC$ (right).

of short distances, and decrease for $d \gtrsim 350$. Both functions reach their maximum at distance 262. In fact, the human genome seems to favour the occurrence of shorter distances.

Islands of favoured distances. In this analysis, we computed all 4^k DNRC distributions, $f_{w,w'}$. The plots of all the empirical DNRC distributions, for $k=6$ and $k=7$, are available in the Supplementary Material. Comparing each DNRC distribution with the global DNRC distribution, f_k, it is possible to identify words which surpass the global behaviour observed for short distances (see, for example, Fig. 3).

We fixed $[k+1, d_M]$ as the interval of interest, where d_M is the distance where R reaches the maximum value in the global distance distribution. We detected a subset of symmetric word pairs having DNRC distributions with an enrichment of distances in the interval of interest, when compared to the global DNRC distribution. Those distributions display a non-negligible misfit in relation to f_k, for distances in $[k+1, d_M]$. However, not all words are significantly enriched (see the Supplementary Material).

Although some words may be visually identified as having enrichment of short distances, it is not always possible to perform meaningful statistical analysis, due to the small number of occurrences of the DNRC. Moreover, the runs of significant positive residuals ($r > 4$) are related with the ranges of overrepresented short distances.

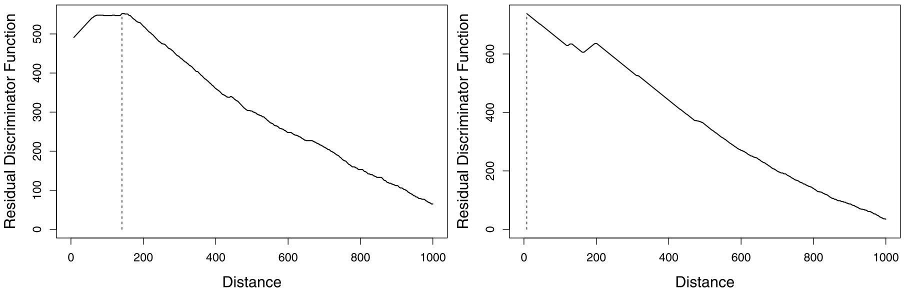


Figure 4. R profile for residuals between $f_{w,w'}$ and $f_{k=7}$, for the complete human genome. Different types of patterns: type T_1 for $w = TATATAC$ (left), and type T_2 for $w = TCACGCC$ (right).

The following procedure was developed to identify the DNRC distributions containing islands of favoured short distances, for a given k:

- We exclude the symmetric word pairs with occurrence frequency lower that 0.0001;
- We exclude the pairs (w, w') such that $f_{w,w'}(d) = 0$ for more than 5% of the distances d in $[k + 1, 1000]$;
- The misfit between $f_{w,w'}$ and f_k is evaluated by the phi coefficient. The symmetric word pairs with $\varphi > 0.80$ are considered to have a very strong effect size[22]. Symmetric pairs with phi coefficient below 0.8 are removed;
- For distances up to d_M, the lengths of the longest run of significant positive residuals are considered. Symmetric word pairs with the longest positive run less than 25 are removed.

The four successive filters of the procedure above reduce the initial set of words to 17, for $k = 6$, and to 48, for $k = 7$. Note that other thresholds could have been used in the procedure, which would result in the selection of different subsets of words (see the Supplementary Material).

In order to classify the shape of the DNRC distribution of each symmetric word pair, a residual discriminator function R was obtained for each word pair, based on adjusted Pearson residuals, r_a, computed from the contingency table of all words of length k and distances between $(k + 1)$ and 1000, instead of the standardised residuals (eq. 4). The adjusted Pearson's residuals are given by

$$r_a(d) = \frac{f_{w,w'}(d) - f_k(d)}{\sqrt{f_k(d)}}\sqrt{n}, \tag{6}$$

where n is the total number of DNRC counts for a given word length k, as defined in (1).

The symmetric word pairs were classified in three different types, according to the R graphical profile:

T_1 - A profile showing a marked initial increase, reaching its maximum, and stabilising or decreasing after it; see, for example, Fig. 4 (left);

T_2 - A profile showing an initial decrease, comprising smooth or strong inverted peaks; see, for example, Fig. 4 (right).

T_3 - Other profiles, not matching previous criteria.

The pairs of type T_1 are characterised by high residual discriminator values (max$(R) > 50$), and their DNRC distributions show an enrichment for short distances ($d < 100$). See, for example, Fig. 3, left. Pairs of type T_2 also have high residual discriminator values, but their DNRC distributions show an overrepresentation for distances $d > 100$, with localised bell-shaped peaks. See, for example, Fig. 3, right. All pairs of type T_3 have irregular low-R profiles (max$(R) \leq 50$).

Table 1 presents the subset of symmetric word pairs obtained by our procedure, for $k = 7$. The table also contains the maximum DNRC frequency and the corresponding distance, the max(R) values, the distribution type, and the distance peak location class. It was observed that type T_1 is the largest group and is formed by TA-rich words. Most DNRC distributions of this type reach their maxima for $d < 100$ (C1). Curiously, it was reported that, in *E. coli*, cruciform formation is enhanced by TA-rich sequences and may correlate with transcriptionally-active promoters[23,24]. Also, the cruciform-binding protein PARP-1 (Poly(ADP-ribose) polymerase-1), which is involved in DNA recombination and repair, was shown to interact with promoter-localised cruciforms[25], and promoters are frequently enriched with TA elements[26]. Thus, the overrepresentation of short distances of TA-rich symmetric word pairs, detected by the procedure that we propose, may point to the occurrence of hairpin/cruciform structures.

The proposed procedure also identifies the T_2 group. DNRC distributions in this group have localised bell-shaped peaks for $d > 100$, forming marked islands of favoured distances. The occurrence of peaks in the short distance region of the DNRC distribution could signal the formation of hairpin/cruciform structures. However, DNRC distribution peaks for $d > 100$ could be associated to other structural or functional DNA functions.

Single over-favoured distance. Apart from the words that have clear islands of favoured distances, in the complete list of words of length six and seven (see Supplementary Material) several words can be observed with a

w	max($f_{w,w'}$)	*arg* max($f_{w,w'}$)	max(R)	Type	Class
ATATATA	0.04	9	932	T_1	C1
GTGTATA	0.05	9	57	T_1	C1
TATATAT	0.03	10	881	T_1	C1
TATATAC	0.03	15	552	T_1	C1
GTATATA	0.05	9	76	T_1	C1
ATATATG	0.02	15	453	T_1	C1
TATATGT	0.02	13	301	T_1	C1
TATATAA	0.03	13	566	T_1	C1
ATATACA	0.03	13	248	T_1	C1
TGTGTAT	0.03	9	60	T_1	C1
TTATATA	0.03	9	83	T_1	C1
TATATTA	0.04	11	109	T_1	C1
TATACAC	0.02	294	62	T_1	C4
TATAATA	0.04	9	96	T_1	C1
CATATAT	0.03	9	68	T_1	C1
TGTATAT	0.03	9	118	T_1	C1
TATACAT	0.03	9	92	T_1	C1
AATATAT	0.02	9	109	T_1	C1
ATATAAT	0.03	11	134	T_1	C1
ATATTAT	0.04	9	90	T_1	C1
ATTTTAT	0.03	107	271	T_1	C2
ATACATA	0.03	9	77	T_1	C1
TATGTAT	0.02	9	87	T_1	C1
TATGTGT	0.02	15	56	T_1	C1
ATATGTA	0.02	11	114	T_1	C1
ATGTATA	0.02	15	71	T_1	C1
ATATATT	0.02	13	486	T_1	C1
ACATATA	0.02	11	71	T_1	C1
TATTATA	0.02	9	63	T_1	C1
TACATAT	0.02	13	75	T_1	C1
ATACACA	0.01	15	59	T_1	C1
TAATATA	0.02	13	73	T_1	C1
TCACGCC	0.33	179	738	T_2	C3
GTTCAAG	0.27	122	913	T_2	C3
GGCTCAC	0.18	210	935	T_2	C4
TGGCTCA	0.14	213	911	T_2	C4
TTGAGAC	0.14	199	857	T_2	C3
CAGTGGC	0.12	230	820	T_2	C4
GCAGTGG	0.10	232	700	T_2	C4
TTTGAGA	0.12	201	776	T_2	C3
GTGCAGT	0.08	236	347	T_2	C4
ATCATGG	0.04	148	116	T_2	C3
ATCTCAT	0.04	121	54	T_2	C3
CCTGGGC	0.03	115	91	T_2	C3

Table 1. Words of length seven with DNRC overrepresentation of short distances, identified by our procedure, with indication of DNRC distribution maximum and its argument value, discriminator R function maximum, group type ($T1$ and $T2$) and class of peak distance. Distance peak classes: C_1 ($d < 100$), C_2 ($d \approx 100$), C_3 ($100 < d < 200$) and C_4 ($d > 200$).

single distance very highlighted, due to its high frequency. In order to perform an automatic selection of this kind of words, we defined the procedure:

- We start with the complete set of symmetric words of a fixed length k;
- We exclude the symmetric word pairs with occurrence frequency lower that 0.0001;
- We exclude the pairs (w, w') such that $f_{w,w'}(d) = 0$ for more than 5% of the distances d in $[k + 1, 1000]$;
- The remaining words are sorted by the maximum frequency, max($f_{w,w'}$), of the distances under analysis $d = k + 1, \ldots, 1000$);

Peak type	w	$\max(f_{w,w'})$	$\arg\max(f_{w,w'})$
$d \leq 30$	CATTAGG	0.78	14
	TGCAGTG	0.77	21
	CATGTCC	0.71	14
	TCAACTC	0.71	10
	TTCAACT	0.66	12
$30 < d \leq 200$	TAGCTGG	0.67	31
	GTTGAAC	0.60	157
	TGTTCTC	0.46	31
	CCACAAT	0.45	133
	GAGTTGA	0.43	161
$d > 200$	CCATGCT	0.28	251
	TCCCCAT	0.25	292
	GAATTCT	0.22	339
	TGAATGG	0.22	344
	ATGGGAT	0.21	490

Table 2. Words with highest $f_{w,w'}$ maximum, with indication of maximising distance, for word length 7, organised by peak type: $d \leq 30$, $30 < d \leq 200$ and $d > 200$.

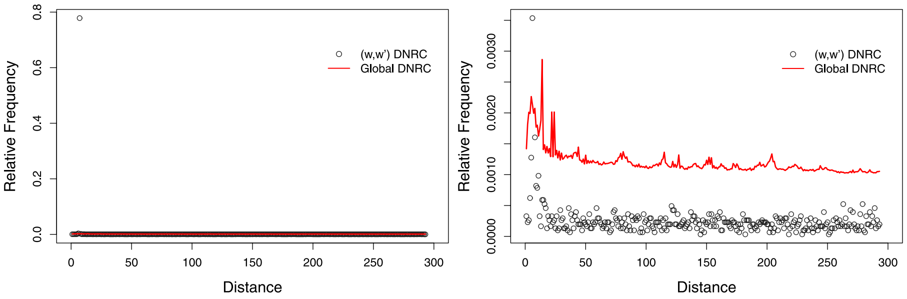


Figure 5. DNRC distribution $f_{w,w'}$ for $w = CATTAGG$, and global DNRC distribution, $f_{k=7}$, for the complete human genome. Very strong enrichment for distance 14: $f_{w,w'}(14) = 0.778$ (left). The right plot is a zoom of y axis.

The first words obtained by these criteria identify words with a single over-favoured distance. To the purpose of classifying the words with relation to single favoured distance, we defined three subsets: peak in distances $d \leq 30$, peak in distances $30 < d \leq 200$ and peak in distances $d > 200$. Table 2 shows the first five words obtained by the previous procedure, for each peak interval type. Taking into account the expected decrease of the distribution, the peak in distances $d > 200$ is an obvious unexpected behaviour. It is noteworthy that for some words a single distance accounts for about 70% of occurrences in a total of $(1000 - k)$ distances.

Figure 5 presents the DNRC distribution of *CATTAGG* (first word in Table 2). This symmetric word pair shows a single over-favoured distance $d = 14$ ($f_{w,w'}(14) \approx 0.8$). In a y-axis zoom (right), a local island of favoured distances is observed. However, in general, these frequencies do not surpass the global distance distribution behaviour. Figure 6 shows another example of a symmetric word pair with a single over-favoured distance at $d = 133$.

In the absence of obvious biological motivation for the occurrence of these single over-favoured distances, we conducted further analyses for some word pairs that have these features. To address the possibility that the reported behaviour may result from a sequencing procedure artefact, we studied the pair (*CCACAAT*, *ATTGTGG*) in detail. Using three independently sequenced and assembled genomes (Celera, HuRef, GRCh38.p2), we computed the distance distributions and found a similar peak in each (see Table 3, displaying the frequencies around distance 133), ruling out the hypothesis that the observed distance peaks are sequencing or assembly artefacts.

We further analysed the sequences comprised between *CCACAAT* and *ATTGTGG*. Taking into account the sequence direction, the distance 133 was only enriched for *CCACAAT* to *ATTGTGG* (15599 occurrences) but not for *ATTGTGG* to *CCACAAT* (11 occurrences, which is in the expected range). The sequence logo (not shown) for the 15599 sequences of the GRCh38.p2 human genome, obtained using WebLogo 3.4[27], shows a significant degree of conservation, suggesting that these sequences may be part of repetitive DNA segments. Using the genomic coordinates for the *CCACAAT* words which are at distance 133 from *ATTGTGG* words, we searched the RepeatMasker annotations available from the UCSC Table Browser. From the 15599 occurrences, 15586 locate within Long INterspersed Elements (LINEs), specifically from the L1 retrotransposon family. L1 are

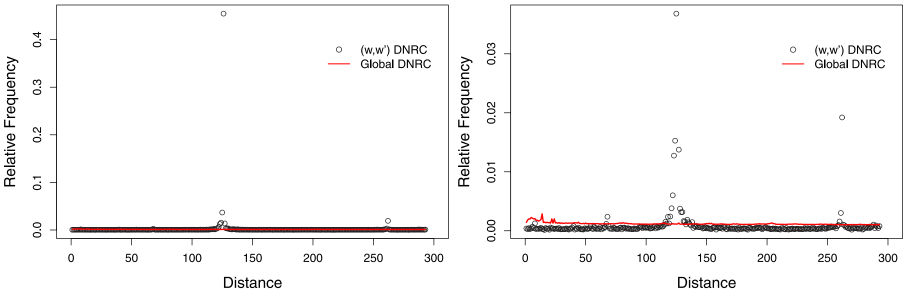


Figure 6. DNRC distribution $f_{w,w'}$ for $w = CCACAAT$, and global DNRC distribution, $f_{k=7}$, for the complete human genome. Very strong enrichment of distance 133: $f_{w,w'}(133) = 0.455$ (left). The right plot is a zoom of y axis.

Distance	Celera	HuRef	GRCh38.p2
125	76	81	81
126	35	40	43
127	78	82	83
128	117	125	131
129	196	205	206
130	402	426	437
131	469	512	524
132	1103	1235	1263
133	12962	14938	15599
134	528	472	472
135	131	126	129
136	98	109	108
137	98	101	109
138	46	56	56
139	52	51	55

Table 3. DNRC partial distribution of *CCACAAT*, around distance 133, for three distinct human genome assemblies (Celera, HuRef, GRCh38.p2).

active transposable elements that also mobilise non-autonomous elements, such as Alu sequences, thus shaping the genome landscape and variation, with implications in evolution and disease[28, 29].

Masked Sequences. To reduce bias from known repetitive sequences in the original genome assembly, we also analysed a pre-masked version of the genome (as reported by RepeatMasker and Tandem Repeats Finder). Masked sequences exclude major known classes of repeats[30], such as long and short interspersed nuclear elements (LINEs and SINEs), long terminal repeat elements (LTRs), Satellite repeats or Simple repeats (micro-satellites).

As expected, masking eliminates distance peaks in several DNRC distributions. For instance, the DNRC distribution of $w = CCACAAT$ (Fig. 6) loses the strong peak observed for the complete genome, because the enrichment of distance 133 is due to LINEs repeats. However, the peaks are preserved in several other distributions.

To select words with single over-favoured distances, in these masked sequences, we applied the procedure described in the previous section. As before, results were classified in three subsets.

The highest-ranking words in the $d \leq 30$ group are *TA*-rich words. Also, we observed that the shape of the DNRC distributions for these words remain unchanged by the masking of repeats. These distributions preserve their characteristic islands of enriched short distances. The distributions of highest-ranking words in the other two groups do not show islands of favouring distances. They display just one or a few strong peaks in the repeat-masked genome.

Globally, the distance peak of the DNRC distributions in the repeat-masked genome correspond to a local maximum in the original genome (80% of the distributions).

Table 4 shows the first five words obtained for each subset, for $k = 7$. The words reported in Table 4 do not show up in the top-5 list of the complete genome sequence (Table 2). Nevertheless, the peaks detected in their distributions are also local maxima, or even global, in the corresponding non-masked distribution. As an example, Fig. 7 shows the DNRC distribution of *TATGAAT* in the complete genome (left) and in the repeat-masked genome (right). Distance 63 is the distribution mode (peak) in the repeat-masked genome, and also a local maximum in the distribution extracted from the complete genome.

Peak type	w	$\max(f_{w,w'})$	$arg\max(f_{w,w'})$	chr
$d \leq 30$	TGTGTAT	0.033	9	several
	TATATAT	0.031	13	several
	AATATAT	0.027	9	several
	TGTGTGC	0.027	9	several
	ATATACA	0.027	13	several
$30 < d \leq 200$	GGGCCCA	0.033	101	chr13
	CAGGCTC	0.023	31	chr1
	AAGCTTT	0.020	83	chr19
	TATGAAT	0.018	63	chr19
	GCCACAG	0.013	115	chr1
$d > 200$	GTTTTCC	0.010	425	chr1
	TGAAATC	0.010	555	chr1
	GGCTCAG	0.009	401	chr1
	TGAGAGA	0.009	502	chr1
	TTTTGTC	0.009	256	chr1

Table 4. Words with highest $f_{w,w'}$ maximum, with indication of the maximising distance, in masked sequences, organised by peak type: $d \leq 30$, $30 < d \leq 200$ and $d > 200$.

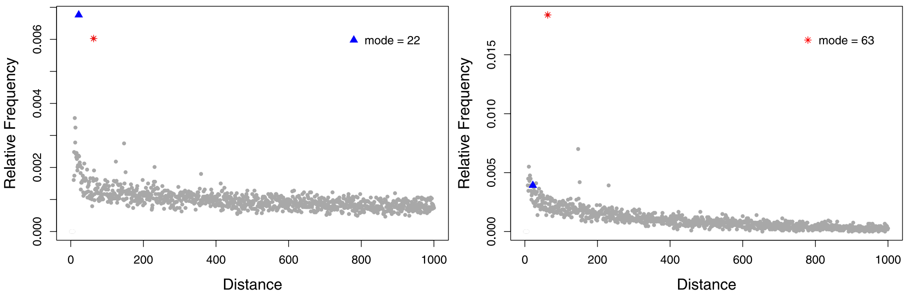


Figure 7. DNRC distribution $f_{w,w'}$ for $w = TATGAAT$ in the complete genome (left) and in the repeat-masked genome (right). The triangle symbol identifies the mode in the complete genome ($d = 22$) and the asterisk symbol is the mode in the masked genome ($d = 63$).

The peaks of DNRC distributions of words in Table 4 were analysed in order to assess the existence of biological features. The peak distances in the $d \leq 30$ subset arise from the overall contribution of several chromosomes. For the words in the other subsets, there is clearly a chromosome that is the main contributor to the single distance peak (see Table 4). Annotations within genomic coordinates for the words listed in Table 4 were retrieved from UCSC GENCODE v24 (https://genome.ucsc.edu/cgi-bin/hgTables) and the resulting gene lists were analysed with the functional annotation tool in DAVID[31, 32]. Overall, word pairs with peaks at distances $d > 30$ are enriched in genes with several and well-defined protein domains, namely, DNA-binding Zinc-finger proteins and members from the neuroblastoma breakpoint family (NBPF). These are duplicated genes with extreme copy number expansion that are associated with brain development and pathology, and are located in a human-specific pericentric inversion in chromosome 1[33, 34]. Word pairs with distance peaks at $d \leq 30$ are scattered throughout the genome, and show enrichment in genes associated with the membrane, which also display a conserved protein topology. As in the T_1 group of the complete genome, the words of this subset are *TA*-rich which may be associated with cruciforme structure occurrence.

Conclusions

We developed new procedures to describe some characteristics of genomic words. In particular, the relative position and distance between reverse complemented word pairs was addressed, using the notion of distance to the nearest reversed complement (DNRC). Under this framework, we studied the DNRC distribution of each word in comparison with the global DNRC distribution and verified the homogeneity of the global DNRC distribution across human chromosomes, for word sizes $1 \leq k \leq 7$.

Using these novel procedures for genomic word detection, we were able to find words with unexpected features in the DNRC distribution, which could not be detected by word frequency procedures alone. The detection of pairs of symmetric words that occur very often at a fixed distance (e.g., the pair (*CCACAAT*, *ATTGTGG*) at distance 133) suggests structural characteristics of the DNA. Some of these are already known but some others may be new.

We explored the global DNRC distributions of words of lengths $k = 6$ and $k = 7$ in the human genome, comparing them with the expected distributions obtained under k-order Markov dependence. A lack of fit was globally detected. The global DNRC distributions show a strong overrepresention of distances up to 350, a feature that may be associated with the occurrence of cruciform structures.

The DNRC distributions of some word pairs display significantly overrepresented distances. In the complete genome, those distributions fall into one of several distinct patterns: distributions with islands of favoured distances $d < 100$ (typically TA-enriched words); distributions with islands of favoured distances between 50 and 350; distributions with a single overrepresented distance. In the masked genome version, distributions with islands of favoured distances for $d \leq 30$ (typically TA-enriched words) and distributions with single over-favoured distance for $d > 30$, were observed. Some of these peaks are present in both complete and masked genomes, thus they are not related to the major known classes of repeats.

DNA structures such as stem-loops and cruciforms are formed at sites that contain reversed complementary words. For this reason, their study naturally leads to the study of the symmetry properties of the sequences, and in particular to the study of the distribution of distances between nearest reversed complements. We performed an exhaustive study of these distance distributions and identified words that are strong candidates to the formation of cruciform structures in human DNA. We are convinced that the new procedures defined and proposed in this work are relevant for a better understanding of the structure of DNA.

References

1. Forsdyke, D. R. & Mortimer, J. R. Chargaff's legacy. *Gene* **261**, 127–137 (2000).
2. Powdel, B. *et al.* A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's second parity rule). *DNA Research* **16**, 325–343 (2009).
3. Afreixo, V., Rodrigues, J. M. & Bastos, C. A. C. Analysis of single-strand exceptional word symmetry in the human genome: new measures. *Biostatistics* **16**, 209–221 (2015).
4. Zhang, H., Zhong, H.-S. & Zhang, S.-H. Conservation vs. variation of dinucleotide frequencies across bacterial and archaeal genomes: evolutionary implications. *Frontiers in Microbiology* **4**, 269 (2013).
5. Brázda, V., Laister, R. C., Jagelská, E. B. & Arrowsmith, C. Cruciform structures are a common dna feature important for regulating biological processes. *BMC Molecular Biology* **12**, 33 (2011).
6. Kolb, J. *et al.* Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. *Chromosome Research* **17**, 469–483 (2009).
7. Inagaki, H. *et al.* Palindrome-mediated translocations in humans: A new mechanistic model for gross chromosomal rearrangements. *Frontiers in Genetics* **7**, 125 (2016).
8. Hackenberg, M. *et al.* CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* **7**, 446 (2006).
9. Afreixo, V., Bastos, C. A. C., Pinho, A. J., Garcia, S. P. & Ferreira, P. J. S. G. Genome analysis with inter-nucleotide distances. *Bioinformatics* **25**, 3064–3070 (2009).
10. Genome Browser team. GRCh38/hg38 assembly of the human genome, masked, one file per chromosome. URL http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.chromFaMasked.tar.gz.
11. Smit, A. F. A., Hubley, R. M. & Green, P. RepeatMasker Open – 4.0. 2013–2015 (http://repeatmasker.org). URL http://repeatmasker.org.
12. Benson, G. *et al.* Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
13. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
14. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, 1–32 (2007).
15. Tavares, A. H. M. P. *et al.* Detection of exceptional genomic words: A comparison between species. In *Proceedings of 22nd International Conference on Computational Statistics* (*COMPSTAT*), 255–264 (2016).
16. Fu, J. C. Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica* **6**, 957–974 (1996).
17. Wang, Y. & Leung, F. C. Long inverted repeats in eukaryotic genomes: Recombinogenicmotifs determine genomic plasticity. *FEBS Letters* **580**, 1277–1284 (2006).
18. Cer, R. Z. *et al.* Non-b db: a database of predicted non-b dna-forming motifs in mammalian genomes. *Nucleic Acids Research* **39**, D383–D391 (2011).
19. Qi, J., Wang, B. & Hao, B.-I. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *Journal of Molecular Evolution* **58**, 1–11 (2004).
20. Ding, S., Dai, Q., Liu, H. & Wang, T. A simple feature representation vector for phylogenetic analysis of DNA sequences. *Journal of Theoretical Biology* **265**, 618–623 (2010).
21. Agresti, A. *An Introduction to Categorical Data Analysis* (Wiley, 2007).
22. Rea, L. M. & Parker, R. A. *Designing and Conducting Survey Research* (Jossey-Boss, San Francisco, 1992).
23. Dayn, A., Malkhosyan, S. & Mirkin, S. M. Transcriptionally driven cruciform formation *in vivo*. *Nucleic Acids Research* **20**, 5991–5997 (1992).
24. Haniford, D. B. & Pulleyblank, D. E. Transition of a cloned d(AT)n-d(AT)n tract to a cruciform *in vivo*. *Nucleic Acids Research* **13**, 4343–4363 (1985).
25. Potaman, V. N., Shlyakhtenko, L. S., Oussatcheva, E. A., Lyubchenko, Y. L. & Soldatenkov, V. A. Specific binding of poly(ADP-ribose) polymerase-1 to cruciform hairpins. *Journal of Molecular Biology* **348**, 609–615 (2005).
26. Lubliner, S., Keren, L. & Segal, E. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Research* (2013).
27. Crooks, G., Hon, G., Chandonia, J. & Brenner, S. WebLogo: A sequence logo generator. *Genome Research* **14**, 1188–1190 (2004).
28. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351** (2016).
29. Teixeira-Silva, A., Silva, R. M., Carneiro, J., Amorim, A. & Azevedo, L. The role of recombination in the origin and evolution of alu subfamilies. *Plos One* **8**, e64884 (2013).
30. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
31. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**, 1–13 (2009).
32. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).
33. Pratas, D., Silva, R. M., Pinho, A. J. & Ferreira, P. J. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. *Scientific Reports* **5** (2015).
34. O'Bleness, M. S. *et al.* Evolutionary history and genome organization of duf1220 protein domains. *G3: Genes— Genomes— Genetics* **2**, 977–986 (2012).

Acknowledgements

This work was supported by FEDER (Programa Operacional Fatores de Competitividade – COMPETE) and FCT (Fundação para a Ciência e Tecnologia), within the projects UID/MAT/04106/2013 to CIDMA (Center for Research & Development in Mathematics and Applications), UID/BIM/04501/2013 to iBiMED (Institute of Biomedicine) and UID/CEC/00127/2013 to IEETA (Institute of Electronics and Informatics Engineering of Aveiro), and the grants PD/BD/105729/2014 to A.T. and SFRH/BPD/111148/2015 to R.S.

Author Contributions

A.T. and V.A. designed the study. A.P., C.B., P.F. and J.R. wrote the programs and collected data. A.T. carried out statistical analysis of the data and prepared the figures. A.T., V.A., A.P., R.S., J.R., C.B. and P.F. discussed the results and contributed to the development of the study. R.S. performed analyses for the biological interpretation of the results. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-00646-2

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 8

Article V

Comparing reverse complementary genomic words based on their distance distributions and frequencies

Published:

Tavares, A. H., Raymaekers, J., Rousseeuw, P. J., Silva, R. M., Bastos, C. A., Pinho, A., Brito, P., Afreixo, V. (2018). Comparing reverse complementary genomic words based on their distance distributions and frequencies. Interdisciplinary Sciences: Computational Life Sciences, 10(1), 1–11.

DOI: 10.1007/s12539-017-0273-0

Interdisciplinary Sciences: Computational Life Sciences
https://doi.org/10.1007/s12539-017-0273-0

ORIGINAL RESEARCH ARTICLE

Comparing Reverse Complementary Genomic Words Based on Their Distance Distributions and Frequencies

Ana Helena Tavares[1] · Jakob Raymaekers[2] · Peter J. Rousseeuw[2] · Raquel M. Silva[3] · Carlos A. C. Bastos[4] · Armando Pinho[4] · Paula Brito[5] · Vera Afreixo[6]

Received: 24 August 2017 / Revised: 4 October 2017 / Accepted: 8 November 2017

Abstract
In this work, we study reverse complementary genomic word pairs in the human DNA, by comparing both the distance distribution and the frequency of a word to those of its reverse complement. Several measures of dissimilarity between distance distributions are considered, and it is found that the peak dissimilarity works best in this setting. We report the existence of reverse complementary word pairs with very dissimilar distance distributions, as well as word pairs with very similar distance distributions even when both distributions are irregular and contain strong peaks. The association between distribution dissimilarity and frequency discrepancy is also explored, and it is speculated that symmetric pairs combining low and high values of each measure may uncover features of interest. Taken together, our results suggest that some asymmetries in the human genome go far beyond Chargaff's rules. This study uses both the complete human genome and its repeat-masked version.

Keywords Chargaff's rules · Human genome · Distance distribution · Peak dissimilarity · Symmetric word pairs

1 Introduction

The analysis of DNA sequences is an extremely broad research domain which has seen several new approaches over the last years. One of these newer approaches is the study of distance distributions of genomic words. A genomic word, also called an oligonucleotide, is a sequence of nucleotides which are represented by the letters $\{A, C, G, T\}$. In DNA segments, the inter-word distance is defined as the number of nucleotides between the first symbol of consecutive occurrences of that word [1, 2]. For instance, in the DNA segment $A\underline{CG}T\underline{CG}ATC\underline{CG}TG\underline{CG}\,\underline{CG}$ the inter-CG distances are $(3, 5, 4, 2)$. For each word, all of its inter-word distances in the genome sequence can be counted and aggregated into a *distance distribution*, which contains the frequency of each distance. These distributions provide a characterization of genomic words which can be studied using statistical techniques for probability density functions.

In this paper, we are particularly interested in the study of symmetric word pairs. A symmetric word pair is formed by a word w and its reverse complement $\bar{w}$, which is the word obtained by reversing the order of the letters and interchanging the complementary nucleotides $A \leftrightarrow T$ and $C \leftrightarrow G$. For instance, the reverse complement of $w = AAGT$ is $\bar{w} = ACTT$, and together they form the symmetric pair $\{w, \bar{w}\}$. The interest in these pairs stems from Chargaff's second parity rule which implies that within a strand of DNA the number of complementary nucleotides is similar [3]. One potential explanation postulates that this phenomenon would be an original feature of the primordial genome, the most primitive nucleic acid genome, and the preservation of strand symmetry would rely on evolutionary mechanisms [4]. Symmetric word pairs can occur in a genome through recombination events such as duplications, inversions and inverted transpositions [5, 6]. These segments have been

✉ Ana Helena Tavares
ahtavares@ua.pt

[1] Department of Mathematics and CIDMA and iBiMED, University of Aveiro, Aveiro, Portugal

[2] Department of Mathematics, KU Leuven, Leuven, Belgium

[3] Department of Medical Sciences and iBiMED and IEETA, University of Aveiro, Aveiro, Portugal

[4] Department of Electronics Telecommunications and Informatics and IEETA, University of Aveiro, Aveiro, Portugal

[5] Faculty of Economics and LIAAD-INESC TEC, University of Porto, Porto, Portugal

[6] Department of Mathematics and CIDMA and iBiMED and IEETA, University of Aveiro, Aveiro, Portugal

Published online: 06 December 2017

associated with specific biological functions, namely, replication and transcription, and major evolutionary events including recombination and translocations. Also, the potential to form secondary DNA structures can cause the genome instability observed in some diseases [7].

Chargaff's second parity rule has led to the natural question whether this also holds for symmetric word pairs. This question has been answered to a certain extent in the existing literature [6, 8–10], as it has been observed that even for long DNA words in several organisms, including the human genome, the frequency of a word is typically (but not always) similar to that of its reverse complement. However, two words with the same frequency in a sequence may exhibit very distinct distance distributions along that sequence. This leads to the natural follow-up question: do symmetric word pairs have similar distance distributions?

Tavares et al. [2] addressed this question for words of length $k \leq 5$ in the human genome. Adopting a whole-genome analysis approach, the discrepancy between distance distributions was evaluated using an effect size measure. The authors concluded that the dissimilarity between the distributions of symmetric word pairs of this length was negligible. The authors also reported that for each word w, the distance distribution nearest to the distance distribution of w is most often that of $\bar{w}$, the reverse complement of w.

As an example, Fig. 1 shows the distance distribution of the word $w = GGGAGGC$ in the human genome. Its peaks correspond to three distances that occur much more often than others. In this example, the distance distribution of the reverse complement $\bar{w} = GCCTCCC$ is extremely similar.

In order to study differences between distance distributions, a new dissimilarity measure was proposed by Tavares et al. [11]. Based on the gaps between the locations of their peaks and the difference between the sizes of these peaks, the peak dissimilarity becomes high when the distributions have very different peaks, or when one distribution has strong peaks and the other does not. In this article, we extend their work in two ways. First, we compare the peak dissimilarity with two earlier dissimilarity measures and argue for its superiority in the analysis of distance distributions between symmetric word pairs. Secondly, we combine the peak dissimilarity with information about the frequencies of the word and its reverse complement to improve the identification of atypical genomic word pairs. We also draw a comparison between the observed distribution and the expected distribution under randomness. Using these techniques we detect several atypical word pairs, which we annotate by identifying the chromosomes and genes where their differences are most pronounced.

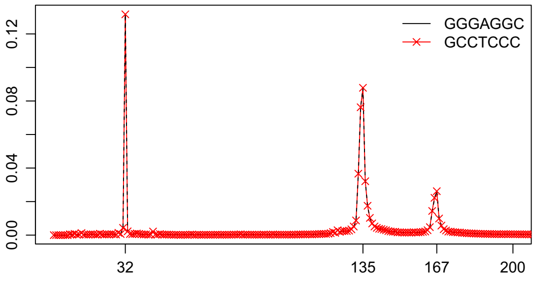


Fig. 1 Distance distribution of the genomic word $w = GGGAGGC$ and of its reverse complement $\bar{w} = GCCTCCC$ in the human genome. Adapted from [2]

The paper is organized as follows. In Sect. 2, we describe measures of the discrepancy between frequencies and distance distributions, including the peak dissimilarity. Section 3 compares the behavior of these dissimilarity measures in our particular research problem. Section 4 identifies and investigates the symmetric word pairs that are most and least dissimilar, using both their frequencies and their distance distributions. It also explores how well the results hold up in a masked sequence. Section 5 concludes.

2 Measures of Dissimilarity

2.1 Discrepancy Between Word Frequencies

To measure the discrepancy between the total absolute frequencies of reverse complementary words w and $\bar{w}$, we count all occurrences of each word along the DNA sequence. The number of times w occurs is denoted as n^w, and that of $\bar{w}$ is $n^{\bar{w}}$. Under the null hypothesis that the true underlying probabilities of w and $\bar{w}$ are equal, the expected frequency of w is $e = (n^w + n^{\bar{w}})/2$. The Pearson residual [12] of w is then given by $(n^w - e)/\sqrt{e}$. The absolute Pearson residual (APR) of w is thus

$$\mathrm{APR}(w) = \frac{|n^w - e|}{\sqrt{e}} = \frac{|n^w - n^{\bar{w}}|}{\sqrt{2(n^w + n^{\bar{w}})}}. \tag{1}$$

Note that $\mathrm{APR}(w) = \mathrm{APR}(\bar{w})$ and that $2\mathrm{APR}^2(w)$ equals the usual chi-squared statistic for testing the equality of the underlying probabilities.

2.2 Dissimilarity Measures for Distance Distributions

Assuming that the DNA sequence is read through a sliding window of word length k, the inter-word distance sequence is defined as the differences between the positions of the first symbol of consecutive occurrences of that word. For instance, the inter-CG distances sequence in the DNA segment $CGTACGCGACG$ is (4, 2, 3). The distance distribution of w, denoted by f^w, gives the relative frequency of each distance, i.e., the number of times a certain distance occurs divided by the total number of occurrences of the word w.

The word structure influences the distance distribution, as some distances from 1 to k may be absent. As an example, note that the inter-*AAA* distance can be equal to one, but cannot be two or three. So, for words of length k we will only consider distances greater than k.

We now wish to compare the distance distribution of each word w with the distance distribution of $\bar{w}$. For this we describe three dissimilarity measures, two of which have been used for a long time and one is new.

2.2.1 Euclidean Distance

The Euclidean distance is a standard tool which is also used between distributions. In our situation, the discrete probability distributions f^w and $f^{\bar{w}}$ have the same domain. The word 'discrete' refers to the domain, as the distances are always integers. The probabilities (i.e., frequencies) of a distance i are denoted as $p_i = f^w(i)$ and $q_i = f^{\bar{w}}(i)$. Then the Euclidean distance $D_{\mathrm{E}}(f^w,f^{\bar{w}})$ is obtained by summing the squares of the frequency differences:

$$D_{\mathrm{E}}(f^w,f^{\bar{w}}) = \sqrt{\sum_i (p_i - q_i)^2}. \tag{2}$$

2.2.2 Jeffreys Divergence

The Kullback–Leibler divergence [13] between f^w and $f^{\bar{w}}$ is given by

$$D_{\mathrm{KL}}(f^w,f^{\bar{w}}) = \sum_i p_i \log(p_i/q_i),$$

where the $0 \log 0 = 0$ convention is adopted. The Kullback–Leibler divergence stems from information theory. It is always nonnegative and becomes zero when the distributions are equal, and it is widely used as a divergence measure between distributions. But it is not symmetric, as $D_{\mathrm{KL}}(f^w,f^{\bar{w}})$ need not equal $D_{\mathrm{KL}}(f^{\bar{w}},f^w)$. Therefore, we will use a symmetrized version called the Jeffreys divergence [14]:

$$D_{\mathrm{J}}(f^w,f^{\bar{w}}) = D_{\mathrm{KL}}(f^w,f^{\bar{w}}) + D_{\mathrm{KL}}(f^{\bar{w}},f^w). \tag{3}$$

Note that D_{J} is not well defined if some p_i or q_i are zero. In practice this can be avoided by replacing the zero values by a small positive value. The Jeffreys divergence D_{J} is a semimetric, meaning that it is symmetric, nonnegative, and reduces to zero when the two distributions are identical.

2.2.3 Peak Dissimilarity

The distance distributions f^w and $f^{\bar{w}}$ may present several peaks, i.e., distances with frequencies much higher than the global tendency of the distribution, as we saw in Fig. 1. To describe the recently proposed peak dissimilarity [11] we go through three steps.

1. Identifying peaks To determine peaks we slide a window of fixed width h along the domain of the distribution. In each such interval of width h we average the absolute values of the differences between successive frequencies, and call the result the *size* of the peak on that interval. The peak's location is defined as the midpoint of the interval. The strongest peak is then determined by the interval with the highest size. For the second strongest peak we only consider intervals that do not overlap with the first one, and so on.

The bandwidth h is a tuning parameter which controls the number of consecutive frequencies that are aggregated in a region. There is no best bandwidth, and different bandwidths can reveal different features of the data. To illustrate the effect of h on peak identification, consider the distance distribution of the word $w = GGGAGGC$ in Fig. 1 which has a local maximum at distance 135. When $h \leq 3$ the region around distance 135 gives rise to two intervals with high peak size. However, when $h \geq 4$ these high frequencies are combined into a single peak.

2. Dissimilarity between two peaks To measure the dissimilarity between two peaks, we take into account the difference between their sizes and between their locations. Consider the distance distributions f^w and $f^{\bar{w}}$ which are defined on the same domain with length R. Let t_i^w be a peak of f^w with location l_i and size v_i and let $t_j^{\bar{w}}$ be a peak of $f^{\bar{w}}$ with location $\bar{l}_j$ and size $\bar{v}_j$. To measure the dissimilarity between these peaks we propose to use

$$d(t_i^w, t_j^{\bar{w}}) = \left(\frac{|l_i - \bar{l}_j|}{R} + 1\right)\left(\frac{|v_i - \bar{v}_j|}{\min\{v, \bar{v}\}} + 1\right) - 1, \tag{4}$$

where v and $\bar{v}$ are the highest peak sizes observed in each distribution. If the peaks have the same location the dissimilarity is reduced to a relative size difference $|v_i - \bar{v}_j| / \min\{v, \bar{v}\}$, and if they have the same size it is reduced to a relative location difference $|l_i - \bar{l}_j|/R$. The denominator $\min\{v, \bar{v}\}$ yields a high dissimilarity when one distribution has strong peaks and the other does not.

3. Peak dissimilarity between two distributions To measure the dissimilarity between two distributions, we compare their n strongest peaks, for fixed n. We propose

$$D_{\mathrm{P}}(f^w,f^{\bar{w}}) = \min_{\pi \in \mathcal{P}_n} \left\{ \sum_{i=1}^{n} d(t_i^w, t_{\pi(i)}^{\bar{w}}) \right\}, \tag{5}$$

where π is a permutation of the indices $i = 1, \ldots, n$, meaning that $\pi(i)$ is the image of i. The minimum is taken over the set $\mathcal{P}_n$ of all permutations π of n elements. In Fig. 1, the minimum in (5) is attained for the simple permutation $\pi(1) = 1$,

$\pi(2) = 2$, $\pi(3) = 3$ yielding a tiny dissimilarity. In general the proposed measure (5) depends on n, the number of peaks considered, and on the bandwidth h used in the peak search. Like D_J also D_P is a semimetric, which is why we call it a 'dissimilarity' rather than a 'distance'.

2.3 Data and Data Preprocessing

In this study, we used the complete genome assembly, build GRCh38.p2, downloaded from the website of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/genome). We also used pre-masked data available from the UCSG Genome Browser (http://genome.ucsc.edu), in which the repeats determined by Repeat Masker [15] and Tandem Repeats Finder [16] were replaced by Ns.

The chromosomes were processed as separate sequences and non-ACGT symbols were used as sequence separators. The counts of word distances were generated using the C language, taking overlap between successive words into account and setting the maximal distance to 1000. The R language was used to compute the distance distributions, the dissimilarity measures and to perform the statistical analysis.

3 Comparison of Dissimilarity Measures

In this section, we will compare the dissimilarity measures of Sect. 2 on the data under study, consisting of all words of lengths 5, 6, and 7 in the human genome. In particular, the peak dissimilarity is computed with bandwidth $h = 5$ which revealed the essential peak structure of the data, by capturing both "isolated" and "grouped" high frequencies. The results are not overly sensitive to this choice, and in fact very similar results were obtained for $h = 4, 5, 6$. Also, we used the $n = 3$ strongest peaks (for $n = 4, \ldots, 7$ we obtained similar results in much higher computation time).

3.1 Correlation Analysis

For every symmetric word pair $\{w, \bar{w}\}$, each of the four dissimilarity measures provides a value. These are the frequency discrepancy APR, Euclidean distance D_E, Jeffreys divergence D_J, and peak dissimilarity D_P. To evaluate the agreement between these four measures we compute Spearman's rank correlation coefficient r_S between each pair. For instance, to compare APR and D_E we rank the values of each of them, and then compute the product-moment correlation between these two vectors of ranks. Comparing each pair of measures yields the Spearman correlation matrices in Table 1, one for each word length $k = 5, 6, 7$.

Overall the correlations decrease with increasing word length, with D_E and D_J remaining the most correlated ($r_S > 0.90$). The rather high correlation between D_E and D_J may perhaps be explained by the formal analogy between $D_E^2 = \sum_i (p_i - q_i)^2$ and $D_J = \sum_i (p_i - q_i)(\log p_i - \log q_i)$. By comparison D_P is less correlated with either of them, especially for $k = 7$. The correlation between APR and the measures D_E, D_J and D_P lies in between. We may conclude that the various measures yield complementary information, with the possible exception of D_E and D_J. Therefore, the adopted measure(s) should take into account the features that are considered important for the subject matter. In the next subsection, we will argue which dissimilarity measures are the most useful in the context of the present research problem.

3.2 Comparing Top-Ranked Sets

For each distance distribution dissimilarity measure (D_E, D_J and D_P), we now rank the dissimilarity values from smallest to largest. The highest ranks correspond to the most dissimilar word pairs for that particular dissimilarity measure. For instance, the top 10% ranked set for D_E consists of the word pairs whose Euclidean distance exceeds the 90th percentile of D_E. As discussed earlier, the ranks of D_E and D_J are more correlated than those of D_P and D_J (see Table 1). One way to assess whether the most dissimilar distributions are the same in each top-ranked set (regardless of their position within that set) is to count the number of common word pairs in those sets. In particular, Table 2 records the fraction of common elements in the top 1% ranked sets for D_E and D_J (under the heading $R_{E,J}$), etc. The top 1% ranked sets for D_E and D_J indeed have the largest overlap, whereas those of D_J and D_P have the least in common, especially for $k = 6$ and $k = 7$. The results for the top 10% ranked sets are similar.

Looking at the top-ranked sets for $k = 7$ in more detail shows specific differences. In Fig. 2a, we see that the 1%

Table 1 Spearman rank correlation matrices for frequency discrepancy APR and distance distribution dissimilarities D_E, D_J, and D_P, by word length

	$k = 5$				$k = 6$				$k = 7$			
	APR	D_E	D_J	D_P	APR	D_E	D_J	D_P	APR	D_E	D_J	D_P
APR	1				1				1			
D_E	0.635	1			0.551	1			0.283	1		
D_J	0.573	**0.988**	1		0.403	**0.962**	1		0.029	**0.904**	1	
D_P	0.663	0.836	0.800	1	0.622	0.784	0.678	1	0.457	0.641	**0.427**	1

Bolded numbers refers to values referred in the text. D_E and D_J are the most correlated (r_S >0.9)

Table 2 Comparison between the rankings for D_E, D_J and D_P: fraction of common elements in the top 1% and top 10% ranked sets

k	Overlap in top-ranked sets					
	Top 1%			Top 10%		
	$R_{E,J}$	$R_{E,P}$	$R_{J,P}$	$R_{E,J}$	$R_{E,P}$	$R_{J,P}$
5	0.98	0.49	0.49	0.89	0.66	0.63
6	0.12	0.24	**0.05**	0.63	0.61	**0.38**
7	0.23	0.03	**0.00**	0.58	0.47	**0.18**

Bolded numbers refers to values referred in the text. D_J and D_P have the least in common, especially for k>=6

top-ranked word pairs for D_J and D_E consist of words with low word frequencies, whereas the 1% top-ranked word pairs for D_P are composed of words with much higher frequencies. In Fig. 2b, we note that the top-ranked word pairs for D_P also have higher frequency discrepancy values (absolute Pearson residuals).

A visual inspection of the distance distributions in word pairs with high-ranked D_J reveals that there are many sparse distributions among them. By sparse we mean that there are many zero frequencies $f^w(i)$, and we already saw that these words have a low total absolute frequency. Indeed, the dissimilarity measures D_J and D_E may be overstating the disagreement between distance distributions with local differences. In fact, D_J is quite sensitive to small frequencies, while D_E is sensitive to the presence of a few high frequencies. It should be noted that in the presence of sparse distributions both low and high relative frequency values are expected, which strongly affect the results of D_E and D_J. On the other hand, D_P ignores small frequencies and evaluates the disagreement between the sizes of the three strongest peaks, which are taken into account even when their locations do not precisely coincide. Moreover, the peak size differences are scaled by the highest peak sizes observed in each distribution.

In view of these results, in what follows we will focus on the dissimilarity measures D_P and APR for the detection of discrepancies between symmetric word pairs.

4 Detection of Atypical Symmetric Word Pairs

In this section, we focus on symmetric word pairs consisting of words with length $k = 5, 6$, and 7, both in the complete human genome assembly and in a masked version.

In order to identify atypical words, we will use three approaches. First, we will consider the peak dissimilarity between the distance distributions. Second, we will combine this information with the frequency discrepancy. Finally, we will study the deviations between the observed distance distributions and the distance distributions under the assumption of randomness and Chargaff's parity rule.

4.1 Analyzing the Observed Peak Dissimilarities

As before, the peak dissimilarity is computed with bandwidth $h = 5$ and the $n = 3$ strongest peaks. To capture the most dissimilar distance distributions we select those symmetric word pairs with peak dissimilarity above the 99th percentile of D_P values. This procedure captured 6 word pairs of

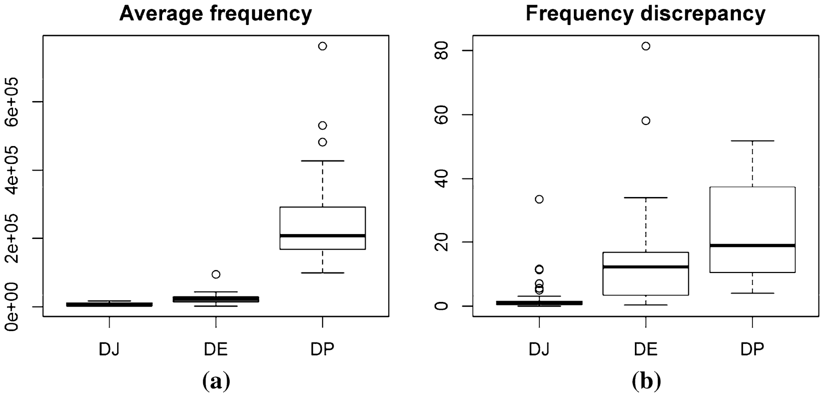


Fig. 2 Statistics of symmetric pairs $\{w, \bar{w}\}$ in the 1% top-ranked set of each divergence measure, for $k = 7$: **a** average word pair frequency $(n^w + n^{\bar{w}})/2$ and **b** frequency discrepancy APR. Complete genome

length $k = 5$, 21 of length $k = 6$ and 82 of length $k = 7$. Next, these words were sorted by decreasing peak dissimilarity value. The results are listed in Table 3 (for $k = 6$ and $k = 7$ only the first 20 results are shown).

Looking at these distributions, it turns out that these high peak dissimilarities are often caused by one distribution with strong peak(s) and another displaying low variability or small peaks, as illustrated in Fig. 3.

The symmetric pairs with low values of D_p have very similar distributions. For some words, this dissimilarity is surprisingly low in spite of their distance distributions having irregular patterns and/or some strong peaks. Some of

Table 3 Symmetric word pairs with peak dissimilarity above the 99th percentile of D_P values, by word length (only the first 20 results are shown)

$k = 5$		$k = 6$				$k = 7$			
w	D_P	w	D_P	w	D_P	w	D_P	w	D_P
CGAAG	127.9	AGTATC	91.0	GAAATC	58.7	AAATTCC	178.8	AGGTTAA	106.0
ACGAA	87.2	AGTTAC	86.4	AAGGCC	46.3	ACTTTAC	145.4	AACAATC	105.2
TACGA	43.5	GGTTAA	84.5	CCTTCG	46.3	GCTTGAA	138.9	AAACTTA	102.5
AACGG	37.0	AGTAAC	80.7	ATACGA	45.8	CTGTCAA	123.8	GCAGTTA	102.3
GAAAC	25.8	GTTGGA	80.6	GTCACA	45.1	AACACAA	120.4	CTTGACA	100.1
TCCAA	22.1	ACCCGT	69.1	CTTCGA	44.6	AGTTTAA	116.1	GTAGAAC	97.1
		AGGTTA	68.2	AAGTTA	43.6	GGGAAGA	110.4	AAATCCT	96.8
		AAATCG	65.9	ACGAAG	42.3	GATGCCA	107.7	CGGGTTC	96.3
		GAATAC	61.2	AGTCAC	41.6	CACTAAG	107.5	AAGGTTA	95.0
		AGTCGA	60.1	CGGGTA	39.4	AACAGTA	106.8	ATTGGAG	91.7

For each word w its $D_P(w, \bar{w})$ value is given. Complete genome

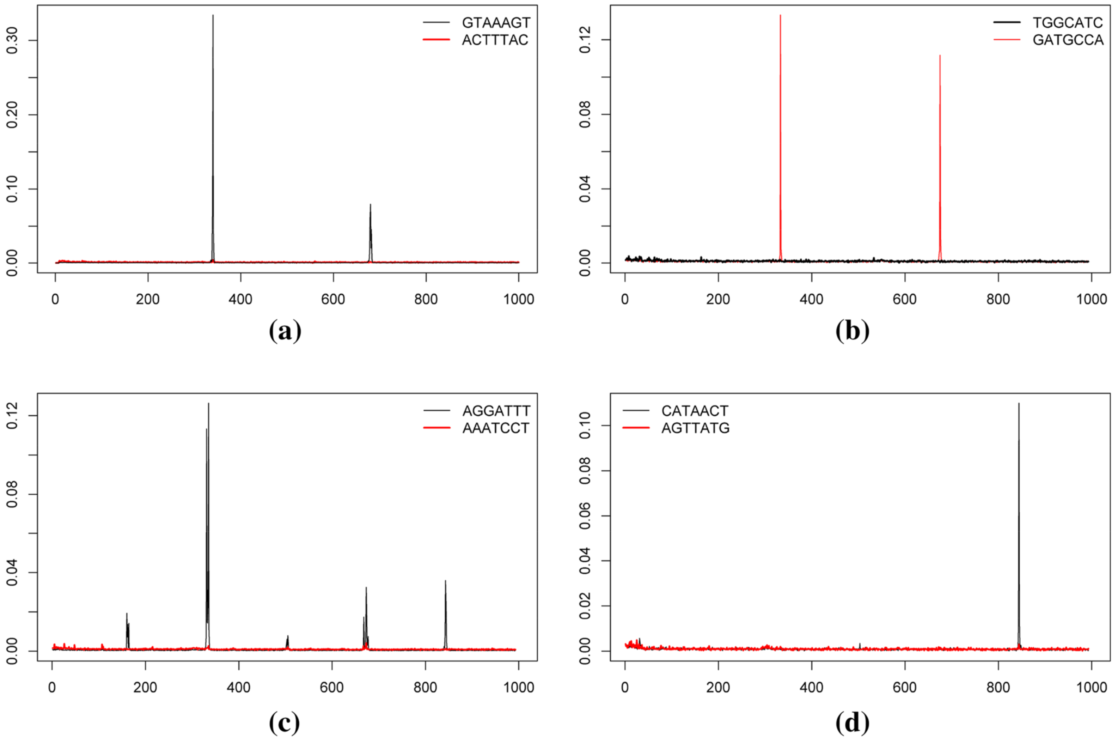


Fig. 3 Distance distributions of some reverse complements, f^w and $f^{\bar{w}}$, with high peak dissimilarity values: **a** $D_P = 145.4$, APR = 37.0; **b** $D_P = 107.6$, APR = 4.9; **c** $D_P = 96.8$, APR = 50.9; **d** $D_P = 55.75$, APR = 2.0. Complete genome

those distributions, with peak dissimilarities below the 10th percentile of D_P, are illustrated in Fig. 4.

4.2 Combining Peak Dissimilarity and Frequency Discrepancy

In order to explore the (dis)similarity between reverse complements, we also combine the peak dissimilarity D_P with the frequency discrepancy APR. Figure 5 plots D_P against APR for each word length, with lines indicating the 90th and 99th percentile of both. Whereas there is a kind of positive relation between D_P and APR for short words, this becomes less clear for longer words, where we know that the rank correlation between these measures decreases (see Table 1).

Several combinations of APR and D_P are observed in Fig. 5: similar word frequency with similar distance distribution (call this case c1, which is common); dissimilar word frequency with similar distance distribution (c2); and similar word frequency with dissimilar distance distribution (c3). (A fourth combination, dissimilar word frequency and dissimilar distance distribution, becomes increasingly rare for longer words.)

The interesting cases are (c2) and (c3), which may reveal features of interest and should be further studied. In case (c2), words have similar distance distributions but their frequencies of occurrence are quite different, which corresponds to points at the upper left of Fig. 5. To illustrate, consider the symmetric pair with $w = CCGTCCG$ (Fig. 4c), which has peak dissimilarity below the 10th percentile of D_P and frequency discrepancy around the 90th percentile of APR. Conversely, in case (c3) strand symmetry holds but the words have distinct distance distributions along the genome. This corresponds to points at the bottom right of the plot. For instance, the symmetric pair with $w = AGTTATG$ (Fig. 3d) has peak dissimilarity above the 90th percentile of D_P and frequency discrepancy around the median of APR. Observe that all word pairs listed in Table 3 are located on the right side of the scatter plot.

These results indicate that some asymmetries in the human genome go far beyond Chargaff's parity rule.

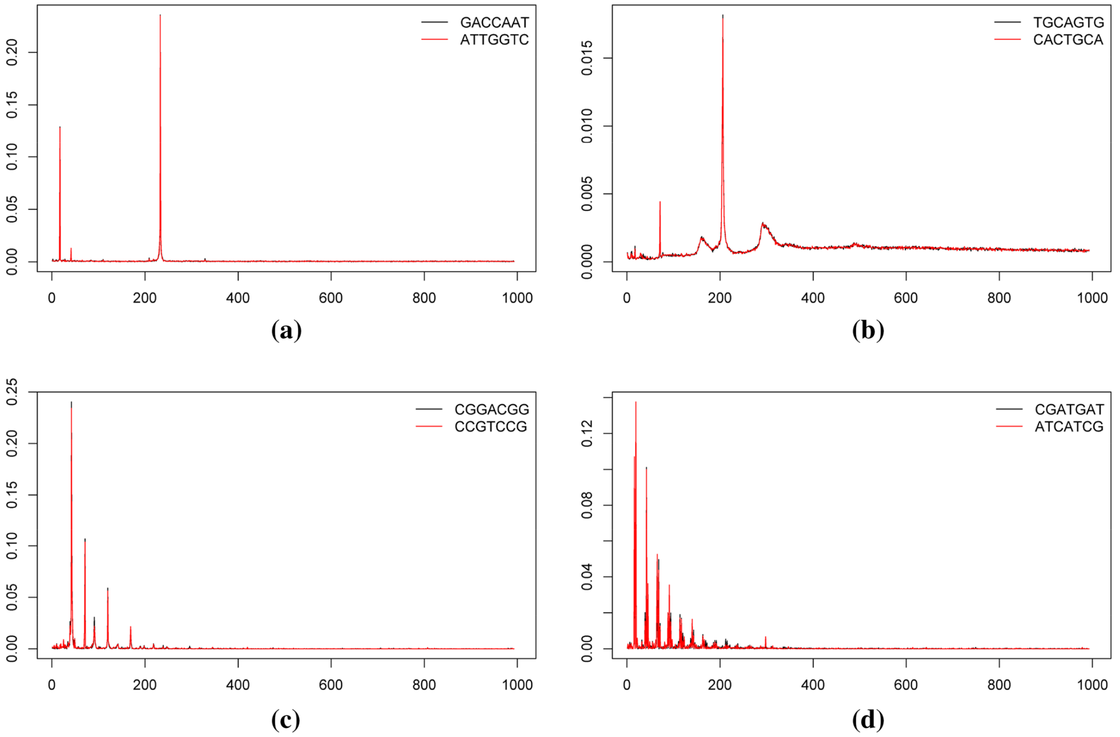


Fig. 4 Distance distributions of some reverse complements, f^w and $f^{\bar{w}}$, with low peak dissimilarity values: **a** $D_P = 0.012$, APR = 0.70; **b** $D_P = 0.026$, APR = 0.73; **c** $D_P = 0.060$, APR = 11.1; **d** $D_P = 0.116$, APR = 4.04. Complete genome

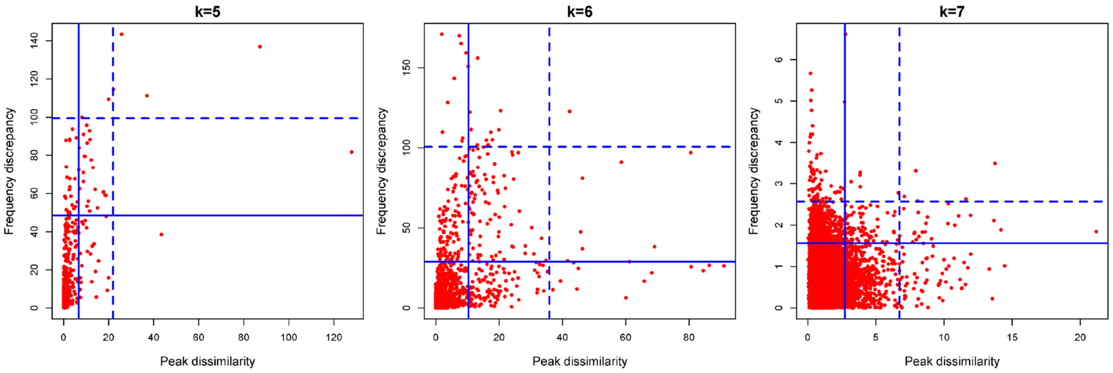


Fig. 5 Frequency discrepancy (APR) versus peak dissimilarity, for word lengths 5, 6 and 7. Solid and dashed lines indicate the 90th and the 99th percentile of each measure, respectively. Complete genome

4.3 Deviations from Randomness

It is intriguing that the distance distributions of a symmetric pair can be very similar even when their pattern is unexpected. If genomic sequences were generated from independent symbols only subject to Chargaff's parity rule ($\%A = \%T$ and $\%C = \%G$), the inter-word distance distributions would be close to an exponential distribution. We are interested in investigating how dissimilar distance distributions from such symmetric pairs can be from the pattern under the random scenario. For that purpose, we compute the peak dissimilarity between the averaged distance distribution of the symmetric pair, $(f^w + f^{\bar{w}})/2$, and the corresponding averaged reference distribution. The expected distance distribution can be deduced using a state diagram, which represents the progress made towards identifying w as each symbol is read from the sequence. The input parameters are the nucleotide frequencies in the sequence. The algorithm used to construct those reference distributions is a special case of Fu's procedure based on finite Markov chain embedding [17].

We select all symmetric pairs with intra-pair peak dissimilarity below the 10th percentile of D_{P}, and ranked them according to the peak dissimilarity between their average distribution and their average reference distribution (denoted as rs). This yields a list of symmetric pairs with similar but unexpected distance distributions. For each word length the top 20 results are listed in Table 4. To illustrate some distance distribution of symmetric word pairs with this behavior, consider the pairs associated with the words $w = CCGTCCG$ (Fig. 4c) and $w = ATCATCG$ (Fig. 4d), which are listed in this table under $k = 7$. The symmetric pairs have very similar distance distributions and their strong peaks make them very dissimilar from the expected distributions in the random scenario.

4.4 Masked Genome Assembly

To reduce the effect of repetitive sequences in the original genome assembly, we also analyze a masked version of the genome which excludes major known classes of repeats [18], such as long and short interspersed nuclear elements (LINE and SINE), long terminal repeat elements (LTR), satellite repeats or simple repeats (micro-satellites). All distributions and measures in this subsection are from the masked sequence and for $k = 7$.

Masking the genome sequence markedly affects the shape of the distance distributions. Several strong peaks observed in the complete genome are eliminated by masking, as described in [11]. It also greatly reduces the frequency discrepancy between reverse complements. To visually inspect those discrepancies, we plot the word frequencies against those observed for the reverse complement. We observe that, for the masked genome, the points are located much closer to the diagonal line than in the complete genome (Fig. 6a, b).

To select symmetric pairs with similar and dissimilar distance distributions, the authors in [11] retained word pairs with peak dissimilarity below the 10th percentile of D_{P} values and those above the 90th percentile of D_{P} values, after filtering out words with low total absolute frequency. They distinguish between two groups of word pairs with low peak dissimilarity: those where both distributions have strong peaks at short distances, and on those where neither distribution has strong peaks. These patterns are illustrated in Fig. 7a, b. Interestingly, the unusual pattern of $w = ATCATCG$ in the complete sequence (Fig. 4d) remains in the masked sequence (Fig. 7b). Symmetric pairs with high dissimilarity usually have one distribution with one or more strong peaks at short distances (< 200), whereas the other presents low variability. Some very dissimilar pairs are shown in Fig. 7c, d.

Table 4 Symmetric pairs with intra-pair peak dissimilarity below the 10th percentile of D_P, sorted by decreasing dissimilarity to the random scenario (only the first 20 results are shown) and organized by word length

$k = 5$			$k = 6$			$k = 7$		
w	D_P	rs	w	D_P	rs	w	D_P	rs
CGCCC	0.009	213.80	CGCCCG	0.029	583.44	ACGCGTA	0.141	1621.58
CCTCC	0.015	207.89	CGGGAG	0.018	443.79	CAACGAG	0.122	1556.41
CGGCC	0.014	206.40	GCCTCC	0.005	418.84	CTCGAGA	0.160	1481.80
CCAGC	0.009	190.02	AGGCCG	0.014	360.64	ATCGCCA	0.082	1350.15
CCTCG	0.025	184.80	CAGACG	0.012	354.04	CGTCTGA	0.130	1292.38
CGCCA	0.014	174.63	CAGGAG	0.012	339.94	ACGCAAA	0.056	1257.21
CCGCC	0.014	153.47	GGTCTA	0.034	332.90	GTTCGGA	0.120	1097.62
CAGGC	0.008	136.91	AGATCG	0.024	326.56	ATCATCG	0.116	1040.96
GCCGA	0.024	136.10	CGAGAC	0.025	291.41	CATCGAA	0.111	1038.82
CCCGG	0.021	133.17	CACGCC	0.038	289.29	TCATCGA	0.143	1031.44
CCACC	0.023	115.13	CCCGTC	0.037	276.62	AGGAGCG	0.099	995.72
CTCCC	0.018	103.37	ACGGGG	0.041	267.93	CAGACGA	0.120	957.98
CCCAG	0.011	95.68	CGTCTC	0.009	266.46	TCCCGGA	0.025	904.82
AGGAG	0.011	88.48	GAGGCA	0.018	265.75	GGATCTA	0.138	893.08
GGCCA	0.014	87.62	CCTCCC	0.015	260.13	CCGGACG	0.099	892.40
CAGGA	0.013	83.81	CTCGGC	0.021	258.12	ACGCTCC	0.096	891.33
CCGAG	0.024	78.98	CCCGGC	0.030	246.31	AGACGCT	0.064	886.83
CCAGG	0.027	74.37	CCGGGC	0.029	242.70	CCGTCCG	0.060	866.16
CTGCC	0.021	66.48	CCCGGA	0.042	242.56	CAGACGG	0.009	855.86
AGTAG	0.005	64.42	CGCCTC	0.034	231.77	CGGGCGC	0.030	840.74

For each word w its $D_P(w, \bar{w})$ value is given and dissimilarity to the random scenario (rs). Complete genome

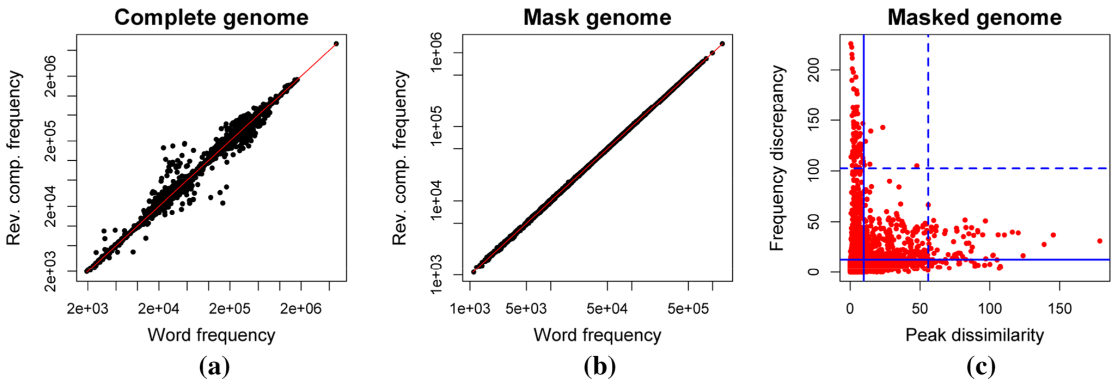


Fig. 6 **a** Word frequencies (n^w) in the entire genome against those observed for the reverse complements ($n^{\bar{w}}$) with both axis in log scale, all for $k = 7$; **b** same for the masked genome; **c** frequency discrepancy versus peak dissimilarity for $k = 7$ in the masked genome, where solid lines indicate the 90th percentile of each quantity

4.4.1 Annotation Analysis

To investigate whether an association exists between dissimilar reverse complements and functional DNA elements, we perform an annotation analysis for the 15 most dissimilar symmetric pairs. For each such pair we list the word with the strongest peaks. Then we look for the 'favored' distance(s), i.e., those where the strongest peak(s) are located. These peaks are often concentrated in one chromosome rather than being spread over the entire genome sequence. Table 5 lists the chromosome in which the favored distances are most pronounced, for each of the 15 pairs. The positions of the words occurring at that distance from each other are recorded. Then,

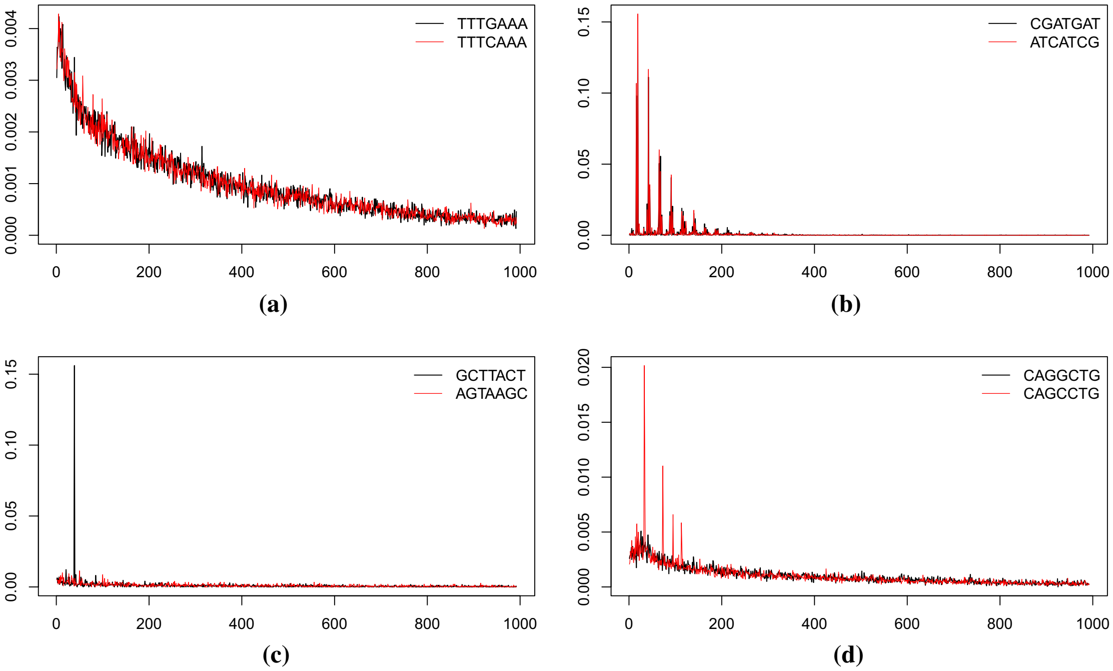


Fig. 7 Distance distributions of some reverse complements with low dissimilarity values: 0.144 (**a**), 0.125 (**b**); and with high dissimilarity values: 11.74 (**c**), 6.49 (**d**). Masked genome

Table 5 The 15 most dissimilar symmetric pairs with $k = 7$, characterized by their word with the strongest peaks

Chromosome	13		1	4	3	8
Word w	*ACCATTC*	*GGTAAGC*	*AGCATCT*	*GTTGGTA*	*TGGTATG*	*GCTTACT*
	CTTCAGG	*TAAGCAT*	*GAGCATC*	*TGGTAGA*		
	GACCATT	*TCAGGAT*	*TGAGCAT*			
	TCCTTCA	*TTCAGGA*				

The chromosome on which these peaks are prominent is indicated. Masked sequence

we retrieve annotations within these genomic coordinates from UCSC GENCODE v24. Interestingly, the words we obtain that are located on chromosome 13 all fall within the gene LINC01043 (long intergenic non-protein coding RNA 1043) and all of our words on chromosome 1 are contained in the gene TTC34 (tetratricopeptide repeat domain 34). These results suggest that the most dissimilar distributions may be related to repetitive regions associated with RNA or protein structure.

A deeper investigation into the biological meaning of these words is necessary to investigate whether the observed dissimilarities reflect the selective evolutionary process of the DNA sequence.

5 Conclusions

In this work, we explore the DNA symmetry phenomenon in the human genome, by comparing each inter-word distance distribution to the distance distribution of its reverse complement, for word lengths $k = 5, 6$ and 7.

We use the peak dissimilarity to evaluate the dissimilarity between the distance distributions of reverse complements and compare it to two well-known measures. Our results suggest that peak dissimilarity achieves its intended purpose in the detection of highly dissimilar distance distributions.

In the complete human genome, we confirm the existence of symmetric word pairs with quite distinct distance distributions. In such cases, one of the distance distributions typically has well-defined peaks and the other has low variability. We also report symmetric pairs with very similar distance distributions even though these distributions are themselves unexpected with strong peaks.

The association between distance distribution dissimilarity and frequency discrepancy is analyzed. In general, the correlation between those measures is moderate. Several behaviors are observed in symmetric pairs, by combining low and high values of both measures. In particular, there are symmetric pairs that preserve strand symmetry (similar frequency) but have dissimilar distance distributions; and symmetric pairs with dissimilar frequencies and similar distance distributions. Symmetric pairs with either behavior may uncover features of interest.

We also investigate how well our results hold up in a masked sequence, which excludes major known classes of repeats. Even though masking generally reduces the dissimilarity between distance distributions of symmetric pairs, there remain quite a few word pairs with high dissimilarity, which in our study are mainly localized on a specific chromosome and even a specific gene. A question worth investigating is to what extent the high dissimilarities may be linked to evolutionary processes.

Taken together, our results suggest that some asymmetries in the human genome go far beyond Chargaff's rules. Of particular note are some symmetric pairs with a perfectly ordinary frequency similarity and distribution similarity, that exhibit a strong preference for occurring at some particular distances.

Acknowledgements This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), Center for Research and Development in Mathematics and Applications (CIDMA), Institute of Biomedicine (iBiMED) and Institute of Electronics and Telematics Engineering of Aveiro (IEETA), within projects UID/MAT/04106/2013, UID/BIM/04501/2013 and UID/CEC/00127/2013. A. Tavares acknowledges the Ph.D. Grant PD/BD/105729/2014 from the FCT. The research of P. Brito was financed by the ERDF—European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization—COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by the FCT as part of project UID/EEA/50014/2013. The research of J. Raymaekers and P. J. Rousseeuw was supported by projects of Internal Funds KU Leuven.

References

1. Afreixo V, Bastos CAC, Pinho AJ, Garcia SP, Ferreira PJSG (2009) Genome analysis with inter-nucleotide distances. Bioinformatics 25(23):3064–3070
2. Tavares AH, Afreixo V, Rodrigues JMOS, Bastos CAC (2015) The symmetry of oligonucleotide distance distributions in the human genome. Proc ICPRAM 2:256–263
3. Forsdyke DR, Mortimer JR (2000) Chargaff's legacy. Gene 261(1):127–137
4. Zhang SH, Huang YZ (2010) Strand symmetry: characteristics and origins. In: 2010 4th international conference on bioinformatics and biomedical engineering (iCBBE). IEEE, pp 1–4
5. Albrecht-Buehler G (2007) Inversions and inverted transpositions as the basis for an almost universal 'format' of genome sequences. Genomics 90(3):297–305
6. Baisnée PF, Hampson S, Baldi P (2002) Why are complementary DNA strands symmetric? Bioinformatics 18(8):1021–1033
7. Inagaki H, Kato T, Tsutsumi M, Ouchi Y, Ohye T, Kurahashi H (2016) Palindrome-mediated translocations in humans: a new mechanistic model for gross chromosomal rearrangements. Front Genet 7:125
8. Afreixo V, Bastos CAC, Garcia SP, Rodrigues JMOS, Pinho AJ, Ferreira PJSG (2013) The breakdown of the word symmetry in the human genome. J Theor Biol 335:153–159
9. Afreixo V, Rodrigues JMOS, Bastos CAC (2015) Analysis of single-strand exceptional word symmetry in the human genome: new measures. Biostatistics 16(2):209–221
10. Albrecht-Buehler G (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc Natl Acad Sci 103(47):17828–17833
11. Tavares AH, Raymaekers J, Rousseeuw PJ, Silva RM, Bastos CAC, Pinho AJ, Brito P, Afreixo V (2017) Dissimilar symmetric word pairs in the human genome. In: Fdez-Riverola F, Mohamad M, Rocha M, De Paz J, Pinto T (eds) 11th International Conference on Practical Applications of Computational Biology & Bioinformatics. PACBB 2017. Advances in Intelligent Systems and Computing, vol 161. Springer, Cham, pp 248–256
12. Agresti A (2007) An introduction to categorical data analysis. Wiley series in probability and statistics. Wiley, New York
13. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
14. Jeffreys H (1946) An invariant form for the prior probability in estimation problems. In: Proceedings of the Royal Society of London. Series A, Mathematical and physical sciences, vol 186. The Royal Society, London, pp 453–461
15. Smit AFA, Hubley RM, Green P (2013) RepeatMasker open-4.0. 2013–2015. http://repeatmasker.org
16. Benson G et al (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580
17. Fu JC (1996) Distribution theory of runs and patterns associated with a sequence of multi-state trials. Stat Sin 6:957–974
18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860–921

Chapter 9

Article VI

Clustering genomic words in human DNA using peaks and trends of distributions

Preprint:

Tavares, A. H., Raymaekers, J., Rousseeuw, Brito, P., Afreixo, V.. Clustering genomic words in human DNA using peaks and trends of distributions. *Preprint submitted.*

manuscript No.
(will be inserted by the editor)

Clustering genomic words in human DNA using peaks and trends of distributions

Ana Helena Tavares · Jakob Raymaekers · Peter J. Rousseeuw · Paula Brito · Vera Afreixo

Received: date / Accepted: date

Abstract In this work we seek clusters of genomic words in human DNA by studying their inter-word lag distributions. Due to the particularly spiked nature of these histograms, a clustering procedure is proposed that first decomposes each distribution into a baseline and a peak distribution. An outlier-robust fitting method is used to estimate the baseline distribution (the 'trend'), and a sparse vector of detrended data captures the peak structure. A simulation study demonstrates the effectiveness of the clustering procedure in grouping distributions with similar peak behavior and/or baseline features. The procedure is applied to investigate similarities between the distribution patterns of genomic words of lengths 3 and 5 in the human genome. These experiments demonstrate the potential of the new method for identifying words with similar distance patterns.

Keywords Classification · Pattern Recognition · Robustness · Word distances

Ana Helena Tavares
Department of Mathematics & CIDMA & iBiMED, University of Aveiro, Aveiro, Portugal
E-mail: ahtavares@ua.pt
orcid.org/0000-0003-4632-3561

Jakob Raymaekers
Department of Mathematics, KU Leuven, Belgium
orcid.org/0000-0002-2093-3137

Peter J. Rousseeuw
Department of Mathematics, KU Leuven, Belgium
orcid.org/0000-0002-3807-5353

Paula Brito
Faculty of Economics & LIAAD - INESC TEC, University of Porto, Porto, Portugal
orcid.org/0000-0002-2593-8818

Vera Afreixo
Department of Mathematics & CIDMA & iBiMED & IEETA, University of Aveiro, Aveiro, Portugal
orcid.org/0000-0003-1051-8084

1 Introduction

Genomes encode and store information that defines any living organism. They may be represented as sequences of symbols from the nucleotide alphabet $\{A, C, G, T\}$. A segment of k consecutive nucleotides is called a *genomic word* of length k. For each length k there are 4^k distinct words.

Some words have a well-defined biological function, and several functionally important regions of the genome can be recognized by searching for sequence patterns, also called 'motifs' [21]. For instance, the trinucleotide ATG serves as an initiation site in coding regions, i.e. a marker where translation into proteins begins [25]. Also the word CG is interesting. Although CG dinucleotides are under-represented in the human genome, clusters of CG dinucleotides ('CpG islands') are used to help in the prediction and annotation of genes [3]. Furthermore, CpG islands are known to be associated with the silencing of genes [8,16,36]. These examples illustrate the importance of identifying word patterns in genomic data.

Finding over- or under-represented words in biological sequences, to discover "relevant" words, is a common task in genomics (see e.g. [22]). The comparison between frequencies observed in real sequences and in random sequences allows evaluating the exceptionality of a given word. The number of word occurrences in a random text has been intensively studied, with many concurrent approaches. Useful reviews of different approaches on random word occurrences can be found in [28,29,20,30,26].

A particular characteristic of a genomic word is its distribution pattern. For some practical purposes, we may care for how a certain word spreads out in the sequence, but not for the specific positions that the value occupies in it. The distribution pattern of a word along a genomic sequence can be characterized by the distances between the positions of the first symbol of consecutive occurrences of that word. The *distance distribution* of the word is the frequency of each lag in the DNA sequence. Patterns in distance distributions of genomic words have been studied through several approaches (see e.g. [2,42,43]) and form an interesting research topic due to their link with positive or negative selection pressures during evolution [5,18].

The search for features in genomic data by inter-word distances and by word frequency are obviously related. However, a plain distinction between the two kinds of features' study exists. Over-represented words are generally related to repetitive elements, which may or not have some known biological function. The disposition of repetitive elements, found in genomes, consists either in tandem repeats (arrays of copies which lie adjacent to each other) or in repeats dispersed throughout the genome. This latter case is related with words that are over-represented, but not necessarily at the same distance from each other. Thus, their distance distributions may not point to any strong irregularity. Conversely, a word with a perfectly ordinary overall frequency, may display a preference to repeat itself at an exact distance. This behavior may not be detected by word frequency procedures alone. Indeed, two words with the same frequency may exhibit very distinct patterns of distribution. Consequently, word frequency and distance between words must be considered distinct issues, deserving separated research.

In this paper we look for clusters of genomic word distance distributions. Because of the particularly spiked nature of these distributions, we have developed a 3-step procedure. First, we fit a smooth baseline distribution using an outlier-robust fitting technique. Secondly, we identify and characterize the peak structure on top of that baseline. Finally, a clustering procedure is applied to the characterization obtained in the first two steps.

The paper is organized as follows. Section 2 describes distance distributions and the proposed clustering procedure. Section 3 describes and reports the results of a simulation study which measures the performance of the proposed method. Section 4 clusters real data, consisting of distance distributions of words in the human genome. Section 5 concludes and outlines future research directions.

2 Methodology

2.1 Word distance distributions

Distance between words and waiting times are closely related topics. One of the most general techniques in waiting times studies is the Markov chain embedding method introduced by Fu [9] and further developed by several authors (for a review see [10,4]). Exact distributions of the distance between occurrences of words are obtained by probabilistic techniques in [31,32,39] and their approximations thereof by compound Poisson processes are given in [33]. The approach of Stefanov *et al.* [37,38] combines Markov chain embedding with an exponential family methodology.

In a simple random sequence with words generated independently from an identical distribution, the distance distribution of a word (without overlap structure) follows a geometric distribution [27], whose continuous approximation is an exponential distribution. By adding some correlation structure between a symbol and the symbols at preceding positions, a more refined DNA model is obtained. This can be achieved by assuming a k-th order Markov model, as in [42].

However, real genomic sequences are more complex and do not follow the simple models mentioned above. Many unexpected patterns occur in the distance distributions of genomic words. For instance, Figure 1 shows the distance distributions of the words $w = TACT$ and $w = ACGG$ in the human genome assembly. They have strong peaks, which correspond to distances that occur much more often than others.

The distance between neighboring occurrences of a certain word translates local behaviors of the word (associated with low or long range) in a global way, since those behaviours are not confined to a specific location. When it matters most, positions associated with features of interest (such as very frequent distances) may be extracted and other analyzes carried out.

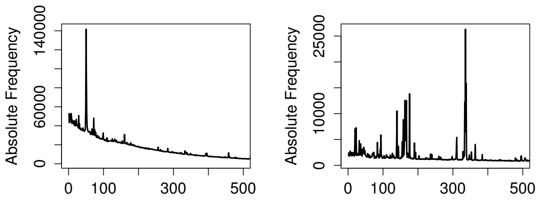


Figure 1 Distance distribution of the genomic words $w = TACT$ (left) and $w = ACGG$ (right) in the human genome. Both distributions exhibit over-favored distances (peaks). The strongest peaks correspond to distances 54 (left) and 340 (right).

2.2 Decomposition of distance distributions

In this study we decompose a distance distribution into a smooth underlying distribution (the 'trend') and a peak function. This decomposition allows us to separate the two essential properties of a distribution.

Consider a genomic word w of length k and denote its relative frequency (histogram) by f, observed on a domain consisting of lags $\{k+1, k+2, \ldots, L\}$. Note that $\sum_{j=k+1}^{L} f(j) = 1$. Such a distribution typically consists of an overall trend and some upward peaks. Therefore, we model the distribution as a mixture of a baseline distribution f_b and a peak function f_{pk} :

$$f = f_b + f_{pk} \quad . \tag{1}$$

We will denote the mass of the baseline component as $m_b = \sum_{j=k+1}^{L} f_b(j)$ and that of the peak function as $m_{pk} = \sum_{j=k+1}^{L} f_{pk}(j)$. Both f_b and f_{pk} are nonnegative hence $0 \leq m_b \leq 1$ and $0 \leq m_{pk} \leq 1$, with $m_b + m_{pk} = 1$.

From many trial fits on distance distributions of genomic words we concluded that a properly scaled gamma density function provides a good fit of the underlying trend. Therefore we set $f_b = \alpha f_\gamma$ with $\alpha \geq 0$ and

$$f_\gamma(x; \theta, \lambda) = \frac{\lambda^\theta x^{\theta-1} e^{-\lambda x}}{\Gamma(\theta)} I(x > 0) \tag{2}$$

where $\theta > 0$ is the shape parameter, $\lambda > 0$ is the rate parameter (note that $1/\lambda$ is a scale parameter), and $\Gamma(.)$ is Euler's gamma function [1]. The gamma distribution includes the exponential distribution as a special case (with $\theta = 1$) and can therefore be seen as an extension of the model in [27].

The peak function f_{pk} describes the mass excess above the baseline. If there is a peak at lag j it follows that $f_{pk}(j) = f(j) - f_b(j)$, and if there is no peak $f_{pk}(j) = 0$.

Figure 2 illustrates the decomposition of the distance distribution of the word $w = ACGG$ shown in Figure 1 into a smooth baseline function f_b and a peak function f_{pk}.

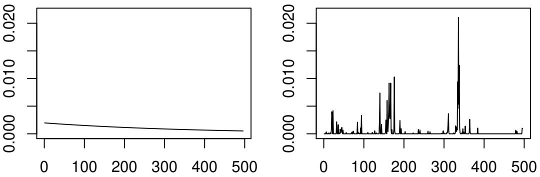


Figure 2 Decomposition of the distance distribution of the genomic word $w = ACGG$ into f_b (left) and f_{pk} (right).

2.3 Estimating the baseline

To estimate the baseline distribution f_b we need to fit a scaled gamma curve αf_γ to the points $(j, f(j))$ of the observed histogram, where $j = k+1, k+2, \ldots, L$. Note that f_b is defined by three parameters: α, θ and λ, so we have to estimate all three together.

A first thought would be to work with the residuals $f(j) - \widehat{f}_b(j)$, but these suffer from heteroskedasticity as the variability in $f(j)$ is larger for low j than for high j. In fact, if we generate n data points from the model (1) the observed *absolute* frequency $f_{ob}(j)$ at a lag j in which there is no peak follows a binomial distribution with n experiments and success probability $f_b(j)$. (Note that in the real data n is the total number of times the word w occurs in the genome.) When the success probability is low and n is high the binomial distribution can be well approximated by a Poisson distribution with mean and variance $n\, f_b(j)$. The standard deviation of that Poisson distribution is thus $\sqrt{n\, f_b(j)}$ and therefore decreasing in j, which implies heteroskedasticity of $f_{ob}(j) - n\widehat{f}_b(j)$. On the other hand, it is known that the square root of a Poisson variable has a nearly constant standard deviation. Therefore, we will fit the function $\sqrt{n\, f_b}$ to the transformed data $\sqrt{f_{ob}}$. We thus use the square root as a variance-stabilizing transform for the Poisson distribution. In practice, we will consider the residuals

$$r(j) = \sqrt{f_{ob}(j)} - \sqrt{n\, \widehat{f}_b(j)} \tag{3}$$

whose standard deviation is roughly constant at those j in which there is no peak, so we are in the usual homoskedastic setting.

The next question is how to combine these residuals in an objective function to be minimized. The standard approach for this is the least squares (LS) objective, which is simply the sum of all squared residuals $\sum_{j=k+1}^{L} r^2(j)$. However, this does not work in our case because of the peaks in the data, which are outliers. Minimizing the LS objective would assign very high weight to the outliers, which do not come from the baseline f_b . Instead we apply the least trimmed squares (LTS) approach of [34]. This method minimizes the sum of the h smallest squared residuals, so that

$$(\hat{\alpha}, \hat{\theta}, \hat{\lambda}) \quad \text{minimizes} \quad \sum_{i=1}^{h} (r^2)_{(i)} \tag{4}$$

where $(r^2)_{(1)} \leqslant (r^2)_{(2)} \leqslant \ldots$ are the ordered squared residuals. In this application we set h equal to 95% of the number of values j in the domain. By using only the 95% smallest squared residuals, the LTS method does not fit the peaks of the distribution and focuses only on the trend. To avoid overemphasizing the high lags j where the fit is close to zero and to get a more accurate fit for the lower lags, we carry out the LTS fit on a shorter set $j \in \{k+1, \ldots, L^*\}$ with $L^* < L$.

2.4 Estimating the peak function

We now want to flag the peaks in the observed absolute frequencies $f_{ob}(j)$, noting that even in lags j without a peak we do not expect $f_{ob}(j)$ to be exactly equal to $n\widehat{f}_b(j)$ because $f_{ob}(j)$ exhibits natural Poisson variability with mean and variance $n\widehat{f}_b(j)$. Therefore we assess the extremity of the observed frequency $f_{ob}(j)$ by comparing it with a high quantile $Q(j)$ (e.g. with probability 0.99) of the Poisson distribution with mean $n\widehat{f}_b(j)$. That is, we flag a peak at the lag j if and only if

$$f_{ob}(j) > Q(j) \quad . \tag{5}$$

At any lag j that is flagged we set the peak function value equal to the difference between the observed and the expected relative frequencies, i.e. $f_{pk}(j) = f(j) - \widehat{f}_b(j) > 0$. At all the other lags we set $f_{pk}(j) = 0$.

2.5 Dimension reduction

Suppose now that we wish to analyze m genomic words, where m could be the number of words of length k in the genome. The raw data is then a matrix of size $m \times (L-k)$ containing the m observed lag distributions. Each row corresponds to a discrete distribution (a vector of length $L-k$), denoted by f, which sums to one. In the preceding subsections we have seen how each row f can be decomposed into the sum of a baseline and a peak function.

First consider the baseline functions. In what follows we are interested in computing a kind of distance between such functions. Since each baseline function f_b is characterized by a triplet of parameters $(\alpha, \theta, \lambda)$, a simple idea would be to compute the Euclidean distance between such triplets. However, the three parameters have different scales, and triplets with relatively high Euclidean distance can describe similar-looking curves and vice versa. To remedy this, we first construct the cumulative distribution function (CDF) of each baseline, given by $F_b(j) = \sum_{i=k+1}^{j} f_b(i)$ for $j = k+1, \ldots, L$. The left panel of Figure 3 illustrates this for the word $w = ACGG$, the lag distribution of which was shown in Figure 1 and decomposed in Figure 2. Note that $F_b(L) = m_b < 1$ when $m_{pk} > 0$.

We can then think of the Euclidean distance between two CDFs F_b and G_b as a way to measure their dissimilarity. Note that these CDFs still have $L-k$ dimensions, which is usually very high. Therefore, in the second step we apply a principal component analysis (PCA) to these m high-dimensional vectors. This operation preserves much of the Euclidean distances. The number of components we retain, q_b, is selected such that at least a given percentage of the variance is explained. Typically $q_b \ll L-k$ so the dimension is reduced substantially. The scores associated to the first q_b components yield a data matrix of much smaller size $m \times q_b$. Note that these scores are uncorrelated with each other by construction.

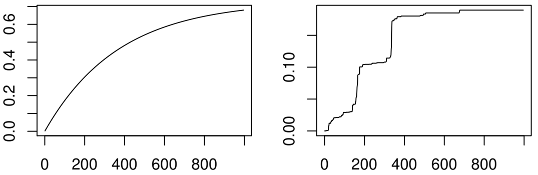


Figure 3 Cumulative distribution functions of the baseline (left) and the peak function (right) of the genomic word $w =$ *ACGG*.

For the peak functions, stacking the m rows on top of each other also yields a matrix of size $m \times (L-k)$. This data matrix is sparse in the sense that few of its elements are nonzero. We then follow the same strategy to that used for the baseline functions: first we convert the peak functions to CDFs as illustrated in the right panel of Figure 3, and then we apply PCA yielding q_{pk} components, where q_{pk} is selected to attain at least a given explained variance. The resulting score matrix has size $m \times q_{pk}$.

2.6 Clustering

Clustering, also known as unsupervised classification, aims to find groups in a dataset (see e.g. [17]). Here our dataset is a matrix of size $m \times (q_b + q_p)$ obtained by applying the above preprocessing to all of the m frequency distributions. We explore clustering based on only the peak component *(Method 1)*, only the baseline component *(Method 2)*, and based on both *(Method 3)*. To each of these datasets we apply the k-means method, in which k stands for the number of clusters which is specified in advance. (The letter 'k' in the name of this method differs from the word length k used elsewhere in this paper.) This approach defines the center of a cluster as its mean, and assigns each object to the cluster with the nearest center. Its goal is to find a partition such that the sum of squared distances of all objects to their center is as small as possible. The algorithm starts from a random initialization of cluster centers and then iterates from there to a local minimum of the objective function. This is not necessarily the global minimum.

As a remedy for this problem, multiple initial configurations are generated and iterations are applied to them, after which the final solution with the lowest objective is retained.

Since k-means looks for spherical clusters, it works best when the input variables are uncorrelated and have similar scales. The preprocessing by PCA in the previous step has created uncorrelated variables, and in our experiments their scales were of the same order of magnitude. However, in other data sets it is necessary to take into account the possibility of observing nonspherical clusters and clusters of unequal volume after preprocessing.

2.7 Selecting the number of clusters

The result of k-means clustering depends on the number of clusters k, which is often hard to choose a priori. Therefore it is common practice to run the method for several values of k, and then select the 'best' value of k as the one which optimizes a certain criterion called a validity index. Many such indices have been proposed in the literature. Here we will focus on three of them: the Calinski-Harabasz (CH) index, the C index, and the silhouette (S) index.

The CH index [6] evaluates the clustering based on the average between- and within-cluster sums of squares. The approach selects the number of clusters with the highest CH index.

The C index reviewed in [15] relates the sum of distances over all pairs of points from the same cluster (say there are N such pairs) to the sum of the N smallest and the sum of the N largest distances between pairs of points in the entire data set. It ranges from 0 to 1 and should be minimized. To compute the C index all pairwise distances have to be computed and stored, which can make this index prohibitive for large datasets.

The S index [35] is the average silhouette width over all points in the dataset. The silhouette width of a point relates its average distance to points of its own cluster to the average distance to points in the 'neighboring' cluster. The silhouette index ranges from -1 to $+1$ and large values indicate a good clustering.

The performance of these measures depends on various data characteristics. An early reference for comparing clustering indices is [23], which concludes that CH and C exhibit excellent recovery characteristics in clean data (the S index was not yet proposed at that time). More recent works evaluate clustering indices also in datasets with outliers and noise, see e.g. [12, 19]). Guerra et al. [12] rank CH and S in top positions, and report poor performance of the C index in that situation.

Rather than choosing one of these indices we will compute all three in our study, and plot each of them against the number of clusters. The local extrema in these curves can be quite informative.

3 Simulation study

To better understand the behavior of the proposed procedure, a simulation study is performed. To assess how well a clustering method performs, we compute a measure of agreement between the resulting partition and the true one.

3.1 Study design

Experiments are performed on datasets consisting of three distinct groups of discrete distributions, denoted by G_1, G_2 and G_3, whose characteristics are defined by a five factor factorial design. The factors and levels used in the study are listed in Table 1. They have the following meaning.

- Trend (T) is defined by the Gamma parameters θ (shape) and λ (rate). When T is 'same' the distributions in all groups have the same baseline parameters.
- Number of peaks (NP) gives the number of peaks generated in each distribution. When NP is 'same' all distributions exhibit the same number of peaks, np, set as 10. In case NP is 'distinct' the number of peaks is set to 20 in G_1, 10 in G_2, and 5 in G_3.
- Peak locations (PL). In each group the 'mean locations' (ml) are generated uniformly on the domain. For each member of that group the peak locations are generated around the mean locations of that group ($ml \pm h$, with $h = 3$). When PL is 'similar' all groups have the same mean locations.
- Peak mass (PM) corresponds to the amount of mass m_p in the peaks of the distribution, so the mass of the baseline is $1 - m_p$. Three levels are considered: distributions of all groups have the same $m_p > 0$; distributions of distinct groups have different $m_p > 0$; distributions of G_1 and G_2 have different $m_p > 0$ and distributions from G_3 have $m_p = 0$. Note that the factors NP and PM are not independent, as $NP = 1$ implies $PM \neq 3$, and $PM = 3$ implies that the distributions in $G3$ have no peaks ($np = 0$).
- Sample size (SS) describes the number of elements in each group. In the 'balanced' setting all groups have the same number of distributions.

Each simulated distribution is constructed from a baseline function and a peak function. All distributions belonging to the same group have the same factor levels.

Table 1 Factors of the experimental study and corresponding levels. Factors: trend, T; number of peaks, NP; peak locations, PL; peak mass, PM; sample size per group, SS.

Factor	Level	Para- meters	Groups G1	 G2	 G3
Trend (T)	1. same	θ	0.8	0.8	0.8
		λ	0.0005	0.0005	0.0005
	2. distinct	θ	0.6	0.8	0.95
		λ	0.0001	0.0005	0.001
Number of Peaks (NP)	1. same	np	10	10	10^*
	2. distinct	np	20	10	5^*
Peak Locations (PL)	1. similar	-	-	-	-
	2. distinct	-	-	-	-
Peak Mass (PM)	1. same	m_p	0.05	0.05	0.05
	2. distinct	m_p	0.1	0.05	0.02
	3. distinct with 0	m_p	0.1	0.05	0
Sample Size (SS)	1. balanced		200	200	200
	2. not balanced		50	150	400

*These values are replaced by 0 in case factor PM takes level 3.

Note that for the baseline function (2) only the parameters θ and λ are user-defined, while α is not. This is because α is determined from the peak mass m_p by

$$\alpha = (1 - m_p) / \sum_{j=k+1}^{L} f_\gamma(j; \theta, \lambda) \quad . \tag{6}$$

Therefore the baseline functions are determined by the trend T and the total peak mass PM. Since the baseline construction depends on PM, it is required that the peak mass takes the same value in all groups (PM=1) in order to obtain similar baselines (T=1).

We will say that groups have *similar baselines* when their T is 'same' and peak mass PM is 'same', and that they have *distinct baselines* when T is 'distinct'. Also, when number of peaks NP is 'same' and peak location PL is 'similar', we will say that the groups have *similar peak functions*, and when PL is 'distinct' they are said to have *distinct peak functions*.

We are interested in the following three scenarios:

Scenario 1 - Groups have similar baselines and distinct peak functions;
Scenario 2 - Groups have similar peak functions and distinct baselines;
Scenario 3 - Groups have distinct baselines and distinct peak functions.

The remaining case where both the baselines and the peaks are similar is not of interest since its groups are basically the same.

The combination of the three scenarios of interest with the possible levels of the design factors leads to 20 possible data configurations: 4 cases for scenario 1, 4 cases for scenario 2 and 12 cases for scenario 3, as can be seen in Table 2. For each case 100 independent samples were generated, and the clustering methods described in section 2.6 were applied to each sample.

Table 2 Possible combinations of factor levels, leading to 20 data conditions, organized by scenario (1, 2 or 3).

		Peak Functions Similar	 Distinct
	Factor Levels	NP=PL=1	PL ≠1
Baselines Similar	$T = 1$ and $PM = 1$		**Scenario 1** $T = 1$; $NP \in \{1, 2\}$; $PL = 2$; $PM = 1$; $SS \in \{1, 2\}$
Baselines Distinct	$T = 2$	**Scenario 2** $T = 2$; $NP = 1$; $PL = 1$; $PM \in \{1, 2\}$; $SS \in \{1, 2\}$	**Scenario 3** $T = 2$; $NP \in \{1, 2\}$; $PL = 2$; $PM \in \{1, 2, 3\}$; $SS \in \{1, 2\}$

3.2 Data generation

The data sets were generated according to the corresponding levels of the factors T, NP, PL, PM and SS. All data sets consist of $m = 600$ discrete distributions on $L = 1500$ lags, with their peaks located in the first 1000 lags. The distributions are labeled by group (G_1, G_2 and G_3).

Baseline distribution. The baseline distributions f_b are given by α times the gamma density $f_\gamma(\theta, \lambda)$ of (2). The gamma parameters θ and λ are determined by the factor T with parameter values shown in Table 1, plus Gaussian noise. The formula is

$$f_b(j) = \alpha f_\gamma(j; \theta + \delta_\theta, \lambda + \delta_\lambda) \tag{7}$$

where $\delta_\theta \sim N(0, 0.01)$, $\delta_\lambda \sim N(0, 0.00001)$ and α is determined from the triplet $(\theta + \delta_\theta, \lambda + \delta_\lambda, m_p)$ according to (6).

Peak function. To define a peak function f_{pk} we first determine the peak locations from the factors PL and NP (as described above), and their magnitudes from PM and T. In all non-peak positions the peak function is set to zero.

Sampling variability. The generated baseline function and peak function together yield a discrete distribution f as in formula (1). We then sample a dataset with 50,000 observations from this population distribution, in a natural way. We first construct the CDF of f, given by $F(j) = \sum_{i \leqslant j} f(i)$ for all j in the domain. Then we consider the quantile function denoted as F^{-1}: for each value u in $]0,1[$ we set $F^{-1}(u) = \min\{j\,;\, F(j) \geqslant u\}$. This quantile function takes only a finite number of values. Now we draw 50,000 random values from the uniform distribution on $]0,1[$ and apply F^{-1} to each, which yields 50,000 lags in the domain that are a random sample from the distribution f given by (1). This sample forms an empirical probability function f_{ob}. We then apply the procedure of Section 2 to carry out a clustering on 600 such empirical distributions.

3.3 Performance evaluation

Each replication takes a set of 600 distributions and returns a partition of these data. To assess the performance of the method, a measure of agreement between the resulting partition and the true partition is needed. Milligan and Cooper [24] evaluated different indices for measuring the agreement between partitions and recommended the Adjusted Rand Index (ARI), introduced in [14]. The ARI has a maximum value of 1 for matching classifications and has an expected value of zero for random classifications [40]. For each case we report the mean and standard deviation of ARI over the 100 replications.

3.4 Results

Table 3 summarizes the results of the simulation. Each row in the table corresponds to a particular case, determined by the levels of the 5 factors (T, NP, PL, PM, SS). The rows are grouped by the 3 scenarios listed in Table 2. Scenario 1 has distinct peak functions, scenario 2 has distinct baselines, and scenario 3 has both.

The first columns of Table 3 describe the factor levels, followed by columns for each of the three methods. In each of those the mean and the standard deviation (in parentheses) of the Adjusted Rand Index over the 100 replications are listed. The final columns list the number of principal components retained for the baselines (b) and the peak functions (pk). These numbers were obtained by requiring that the percentage of explained variance is at least 90%. We see that the baselines require only 2 components. For the peaks the number is high when the peak masses are the same (PM=1) and low otherwise (in the latter case it requires few PCs to explain the larger peaks).

Method 1 The first method applies the clustering to the PCA scores obtained from the peak functions. Therefore, good performance is expected in scenarios with distinct peak locations between the groups (scenarios 1 and 3). Indeed, Method 1 performs very well in scenario 1 (ARI $\geqslant 0.821$) and scenario 3 (ARI $\geqslant 0.900$).

Table 3 Mean and standard deviation of the Adjusted Rand Index obtained from 100 replicas of each case. Results are organized by scenario and method. Each case is defined by a combination of five factors: trend, T; number of peaks, NP; peak locations, PL; peak mass, PM; and sample size per group, SS. The final columns list the number of principal components retained for the baselines (b) and the peak functions (pk).

Factors					**Method 1**	**Method 2**	**Method 3**	**#PC**	
T	NP	PL	PM	SS				b	pk
Scenario 1									
1	1	2	1	1	0.989 (0.046)	0.000 (0.003)	0.817 (0.255)	2	62
1	1	2	1	2	0.886 (0.224)	0.000 (0.007)	0.493 (0.245)	2	58
1	2	2	1	1	0.987 (0.052)	0.000 (0.002)	0.837 (0.245)	2	55
1	2	2	1	2	0.821 (0.245)	-0.002 (0.007)	0.530 (0.208)	2	39
Scenario 2									
2	1	1	1	1	0.082 (0.131)	0.966 (0.019)	0.969 (0.018)	2	42
2	1	1	1	2	0.043 (0.085)	0.934 (0.036)	0.940 (0.036)	2	45
2	1	1	2	1	1.000 (0.000)	0.987 (0.008)	1.000 (0.000)	2	3
2	1	1	2	2	1.000 (0.000)	0.989 (0.008)	1.000 (0.000)	2	2
Scenario 3									
2	1	2	1	1	0.976 (0.060)	0.965 (0.019)	0.999 (0.002)	2	58
2	1	2	1	2	0.919 (0.183)	0.988 (0.009)	1.000 (0.000)	2	58
2	1	2	2	1	1.000 (0.000)	0.992 (0.006)	1.000 (0.000)	2	5
2	1	2	2	2	1.000 (0.000)	0.998 (0.003)	1.000 (0.000)	2	5
2	1	2	3	1	1.000 (0.000)	0.933 (0.036)	0.999 (0.004)	2	3
2	1	2	3	2	1.000 (0.000)	0.988 (0.008)	1.000 (0.000)	2	2
2	2	2	1	1	0.989 (0.030)	0.989 (0.008)	1.000 (0.000)	2	56
2	2	2	1	2	0.900 (0.203)	0.992 (0.007)	1.000 (0.000)	2	40
2	2	2	2	1	1.000 (0.000)	0.997 (0.004)	1.000 (0.000)	2	6
2	2	2	2	2	1.000 (0.000)	0.964 (0.022)	0.999 (0.002)	2	4
2	2	2	3	1	1.000 (0.000)	0.929 (0.042)	0.999 (0.002)	2	3
2	2	2	3	2	1.000 (0.000)	0.988 (0.009)	1.000 (0.000)	2	2

In scenario 2 the peak locations are the same. In the first two cases the peak masses are similar and in the other two cases the peak masses are distinct. As expected, Method 1 recovers the peak differences in the latter cases, whereas there are no differences to recover in the former.

Method 2 This method clusters the PCA scores of the baselines, so it is expected to work well in scenarios 2 and 3 in which the trends are distinct, and not in scenario 1 in which the baselines are similar. The simulation results confirm this, as the groups are not recovered in scenario 1 (ARI ≈ 0) and are identified with high accuracy in scenarios 2 and 3 (ARI $\geqslant 0.929$).

Method 3 The input for Method 3 are the scores of the baselines as well as those of the peaks, and indeed it is the best performer in scenario 3 where the groups have distinct baselines combined with distinct peaks (ARI $\geqslant 0.999$). In that scenario it is also good at distinguishing groups with peaks from groups without peaks ($PM = 3$). Also in scenario 2 we see that Method

3 works well, in fact it even slightly outperforms the other methods in that situation. Only in scenario 1 does Method 3 perform less well. It is still fine when the groups have balanced sizes ($SS = 1$) but becomes weaker when the groups are unbalanced ($SS = 2$).

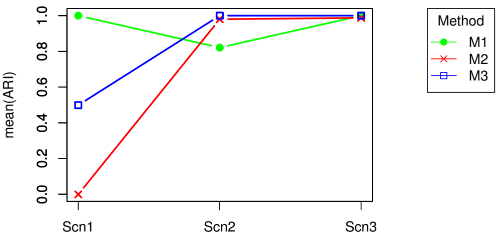


Figure 4 Performance of each method, given by the mean ARI of all replications from cases in each scenario: groups with similar trends and distinct peak locations (scenario 1); groups with distinct trends and similar peak locations (scenario 2); and groups with distinct trends and distinct peak locations (scenario 3). The clustering methods 1, 2 and 3 correspond to the three broken lines.

Figure 4 provides a rough summary of the simulation results by showing the ARI averaged over all cases of each scenario. The performance of a method is thus measured by three numbers. We note that no method is best in all scenarios. Method 2, which ignores the peak information, is never the best method. Method 1 is the best in scenario 1, and Method 3 is the best in scenarios 2 and 3. For a given dataset it is recommended to carry out a preliminary inspection to determine which scenario it corresponds to, before selecting the clustering method.

4 Application to real data

In this section we analyze two datasets, consisting of the lag distributions of all words of length $k = 3$ and $k = 5$ in the complete human genome. These datasets are denoted by DD_k where k identifies the word length. DD_3 consists of 64 distributions and DD_5 contains 1024 distributions. A preliminary visual inspection of these histograms revealed that there are substantial differences in both the trends and the peak structures, so in accordance with the conclusions of the simulation study we selected Method 3 (described in Section 2.6) for clustering the words in each dataset.

4.1 Data and data processing

We used the complete DNA sequence of the human genome assembly, downloaded from the website of the National Center for Biotechnology Information. The available assembled chromosomes (in version GRCh38.p2) were processed as separate sequences and all non-ACGT symbols were considered as sequence separators.

The counts of word lags were obtained by a dedicated C program able to handle large datasets (the haploid human genome has over 3 billion symbols). We analyzed the absolute frequences of the lags $j = k+1, \ldots, L$ where $L = 1000$ for $k = 3$ and $L = 4000$ for $k = 5$.

The R language was used to decompose the lag distributions, to perform the principal component analysis and the clustering and to carry out further statistical analysis.

4.2 Decomposing the lag distributions

In both datasets we first estimated the baseline distribution by LTS as described in Subsection 2.3, in which we set $L^* = 200$ for DD_3 and $L^* = 1500$ for DD_5. The peak functions were then estimated as described in Subsection 2.4.

4.3 Clustering words of length 3

Each distribution in DD_3 is summarized by 4 values, as the PCA retains 2 components for the peaks and 2 components for the baselines.

Figure 5 plots the validation indices against the number of clusters (< 10). The CH index has a local maximum at 3 clusters and is high again at 6 clusters or more, whereas the silhouette index is highest for 2 clusters and the C index is lowest (best) for 2 clusters and gets low again for over 6 clusters. From the 3 indices together it would appear natural to select 2 clusters, for which CH $= 108$, $S = 0.68$ and $C = 0.052$. The cluster C_1 has 8 elements, and cluster C_2 has 56.

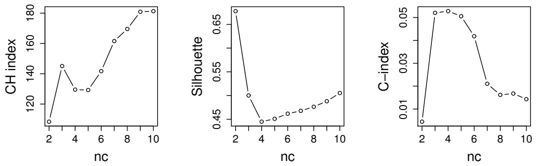


Figure 5 Validation indices for clustering DD_3 by the number of clusters nc: the Calinsky-Harabasz index (left), silhouette coefficient (center) and C-index (right).

To test the stability of this clustering we follow the approach of Hennig [13]. We draw a so-called bootstrap sample, which is a random sample with replacement from the 64 objects in the DD_3 dataset. This creates a different dataset with 64 objects, some of which coincide. We then apply the same clustering method to it, set to 2 clusters. Let us call the new clusters D_a and D_b. Then we compute the so-called Jaccard similarity coefficient of C_1 with the new clustering, defined as

$$J(C_1) = \max\left(\frac{|C_1 \cap D_a|}{|C_1 \cup D_a|}, \frac{|C_1 \cap D_b|}{|C_1 \cup D_b|}\right) \leqslant 1 \qquad (8)$$

where $|\ldots|$ stands for the number of elements. A high value $J(C_1)$ indicates that C_1 is similar to one of the clusters of the new partition. We compute $J(C_2)$ analogously. Then we repeat this whole procedure for a new bootstrap sample and so on, 200 times in all. The average of the 200 values of $J(C_1)$ equals 0.952, which means that the cluster C_1 is very stable. For cluster C_2 we attain the stability index 0.978 which is even higher.

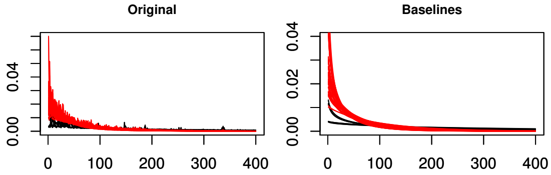


Figure 6 Clustering of the dataset DD_3 in clusters C_1 and C_2. The lag distributions are shown on the left, and the corresponding baselines on the right. Cluster C_1 is in black and $C2$ in red.

Figure 6 depicts the clusters C_1 and C_2. The lag distributions in C_1 are flatter than those in C_2. It turns out that all the words in C_1 contain the dinucleotide CG (known as CpG). In fact, C_1 consists exactly of the 8 words of length 3 that contain CG (i.e., ACG, CCG, GCG, TCG, CGA, CGC, CGG, CGT), so C_2 contains no words with CG. The special behaviour of the CG dinucleotide in the human genome is well reported in the literature. Although human DNA is generally depleted in the dinucleotide CpG (its occurrence is only 21% of what would be expected under randomness), the genome is punctuated by regions with a high frequency of CpG's relative to the bulk genome. This DNA characteristic is related to the CpG methylation [7,11]. We may conclude that the clustering of DD_3 has biological relevance.

It is worth noting that if one considers all k-means clusterings into 2 to 40 clusters, the second best silhouette coefficient is attained for 26 clusters, which also corresponds to the point where the CH index has a large increase and the C-index is very small ($CH = 436$, $S = 0.61$ and $C = 0.0046$). In this partition with 26 clusters, over half of the clusters are formed by pairs of words that are reversed complements of each other, i.e., obtained by reversing the order of the word's symbols and interchanging A-T and C-G. The similarity between lag patterns of reversed complements is a well-known feature described in the literature, see e.g. [41].

4.4 Clustering words of length 5

Also the lag distributions of DD_5 contain quite distinct baselines and peak structures. Figure 7 shows four lag distributions, with their corresponding estimated baselines.

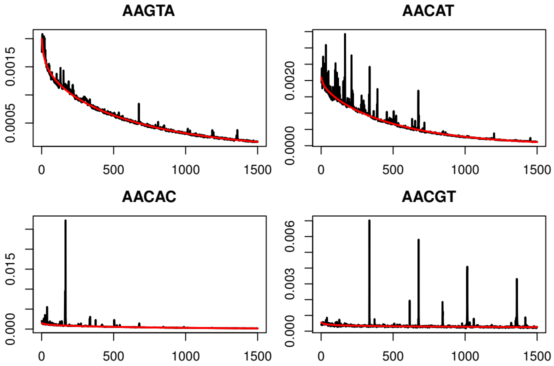


Figure 7 Lag distributions of some words of length $k = 5$, with the corresponding baselines indicated in red.

Our procedure retains 3 principal components for the peaks and 2 components for the baselines, so that each lag distribution is converted into 5 scores. Carrying out k-means clustering for different numbers of clusters yields the plots of validation indices in Figure 8. They do not all point to the same choice, however. The CH and S indices have local maxima at 2 and 6 clusters, while the C-index would support a choice of 5 or more clusters. It would appear that 2 or 6 clusters are appropriate.

When choosing 2 clusters we obtain clusters with 278 and 746 members, and when choosing 6 clusters they have sizes 19, 92, 166, 141, 367 and 239.

We verified that both these partitions are very stable. For this we again drew 200 bootstrap samples, and partitioned each of them followed by computing the Jaccard similarity coefficient of the original clusters. In the case of 2 clusters the average Jaccard (stability) indices were 0.94 and 0.97. In the case of 6 clusters they were 0.84, 0.91, 0.93, 0.92, 0.93 and 0.93. Since we aim to decompose the DD_5 dataset of 1024 distributions into smaller groups with similar patterns, we will focus on the solution with 6 clusters from here onward.

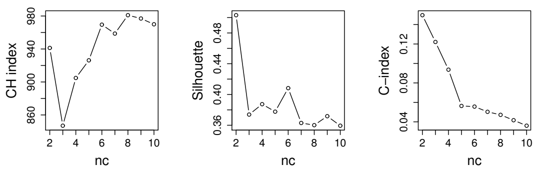


Figure 8 Validation indices for clustering DD_5 by the number of clusters nc: the Calinsky-Harabasz index (left), average silhouette width (center) and C-index (right).

The 6-cluster partition consists of two large clusters ($|C_5| = 367$ and $|C_6| = 239$), three middle-sized clusters ($|C_2| = 92$, $|C_3| = 166$ and $|C_4| = 141$), and the much smaller cluster C_3 with only 19 elements. Figure 9 shows the lag distributions of each cluster. As a graphical summary we also consider the *median function* of each cluster, which in each domain point (lag) equals the median of the cluster's function values in that point.

We see the most pronounced peaks in the clusters C_1, C_3 and C_4. Those in the small cluster C_1 are the strongest. Several of them occur in the same location for most of the cluster members, which explains why they remain visible in the median function. The words in C_1 are listed in Table 4.

Table 4 List of words in cluster C_1 of the partition of DD_5 in six clusters.

AAACG	AACGG	ACGGG	AGCGC	CGAGA	CGCTT
CGGGA	CGTTC	CGTTG	CTTCG	GAGGC	GCCTC
GCGCT	GCGTT	TCGTA	TCGTT	TCTCG	TTCGT
TTTCG					

The distributions in C_4 have most of their peaks before lag 500, with little going on after that. Cluster C_3 is quite different, as strong peaks occur over the whole domain. The distributions in clusters C_2, C_5 and C_6 have rather small peaks, so few major irregularities. Their main difference is in the baselines: those of C_2 have a high rate λ, whereas the baselines of C_6 are much flatter.

We also explore the composition of the words in each cluster, by computing the percentage of words that contain a given dinucleotide or trinucleotide. Clusters C_1, C_2 and C_3 stand out in this respect. Cluster C_2 contains the largest proportion of words with the dinucleotides AA (47%) and TT (49%), which is also reflected in the high frequency of AAA and TTT (25% and 26%, respectively). The clusters C_1 and C_3 have a lot of words containing the dinucleotide CG (89% and 98%). This is very different from the other clusters: only 9% of the words in C_2 contain CG, in C_4 this is 11%, in C_5 only 1%, and in C_6 16%. Even though both C_1 and C_3 have many CG dinucleotides, these occur in different trinucleotides: C_1 has many words containing CGT and TCG (both 32%), whereas in C_3 many words contain CGA (27%) and ACG (23%).

5 Summary and conclusions

In this work we have proposed a methodology for decomposing the lag distribution of a genomic word into the sum of a baseline distribution (the 'trend') and a peak function. The baseline component is estimated by robustly fitting a parametric function to the data distribution, in which the residuals are made homoskedastic and the robustness to outliers is essential. The peak function is then obtained by comparing the absolute frequency at each lag to a quantile of a Poisson distribution.

When analyzing a dataset consisting of many genomic words we can apply principal component analysis to the set of baselines and the set of peak functions, which greatly reduces the dimensionality. This lower-dimensional data set has uncorrelated scores and retains much of the original information, such as that in the Euclidean distances. This allows us to carry out k-means clustering, in which we have the choice whether to use only the baseline information, only the peak information, or both. The performance of this approach was evaluated by a simulation study, which concluded that in situations where both distinct baselines as well as distinct peak functions occur, the clustering procedure using the combined information performs very well.

This procedure was applied to the data set DD_3 of all genomic words of 3 symbols in human DNA, as well as the set DD_5 of all words of length 5. This resulted in clusters of words with specific distribution patterns. By looking at the composition of the words in each cluster we found connections with the frequency of certain trinucleotides and dinucleotides, such as CG which plays a particular biological role.

Topics for further research are the analysis of longer words, and the application of other statistical methods (such as classification) on genomic data after applying the decomposition technique developed here. In the classification task, genomic words may be classified in

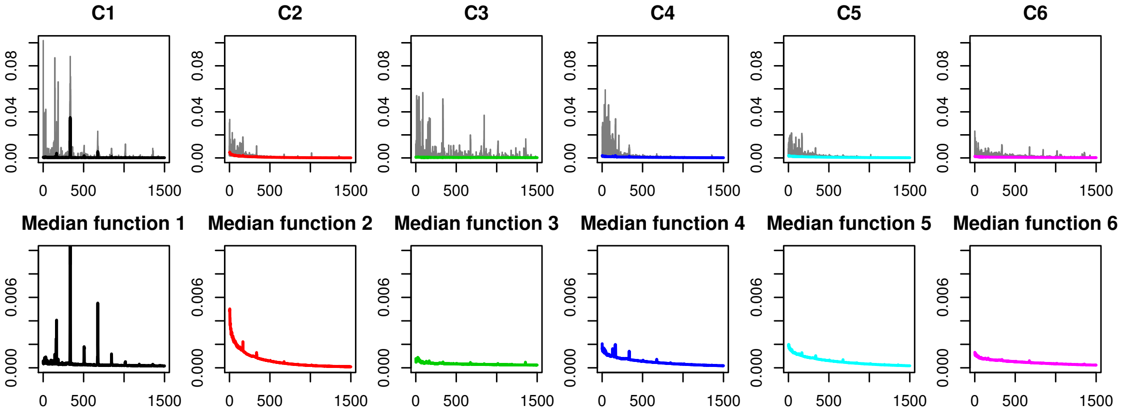


Figure 9 Clustering of DD_5 in six clusters. In each cluster the lag distributions are shown in grey, and the cluster's median function is in color (top). The median functions are also shown with a scaled vertical axis (bottom).

groups according to the similarities between their distributions patterns and, eventually, putting in evidence functional or structural biological relationships between those words.

Acknowledgements This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), Center for Research & Development in Mathematics and Applications (CIDMA) and Institute of Biomedicine (iBiMED), within projects UID/MAT/04106/2013 and UID/BIM/04501/2013. A. Tavares acknowledges the PhD grant PD/BD/105729/2014 from the FCT. The research of P. Brito was financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by the FCT as part of project UID/EEA/50014/2013. The research of J. Raymaekers and P. J. Rousseeuw was supported by projects of Internal Funds KU Leuven.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of mathematical functions: with formulas, graphs, and mathematical tables, vol. 55. Courier Corporation (1964)
2. Afreixo, V., Rodrigues, J.M., Bastos, C.A.: Analysis of single-strand exceptional word symmetry in the human genome: new measures. Biostatistics **16**(2), 209–221 (2014)
3. Bajic, V.B., Seah, S.H.: Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. Genome research **13**(8), 1923–1929 (2003)
4. Balakrishnan, N., Koutras, M.V.: Runs and scans with applications, vol. 764. John Wiley & Sons (2011)
5. Burge, C., Campbell, A.M., Karlin, S.: Over-and under-representation of short oligonucleotides in DNA sequences. Proceedings of the National Academy of Sciences **89**(4), 1358–1362 (1992)
6. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods **3**(1), 1–27 (1974)
7. Consortium, I.H.G.S., et al.: Initial sequencing and analysis of the human genome. Nature **409**(6822), 860 (2001)
8. Deaton, A.M., Bird, A.: CpG islands and the regulation of transcription. Genes & development **25**(10), 1010–1022 (2011)
9. Fu, J.C.: Distribution theory of runs and patterns associated with a sequence of multi-state trials. Statistica Sinica pp. 957–974 (1996)
10. Fu, J.C., Lou, W.W.: Distribution theory of runs and patterns and its applications: a finite Markov chain imbedding approach. World Scientific (2003)
11. Gardiner-Garden, M., Frommer, M.: CpG islands in vertebrate genomes. Journal of molecular biology **196**(2), 261–282 (1987)
12. Guerra, L., Robles, V., Bielza, C., Larrañaga, P.: A comparison of clustering quality indices using outliers and noise. Intelligent Data Analysis **16**(4), 703–715 (2012)
13. Hennig, C.: Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. Journal of Multivariate Analysis **99**(6), 1154–1176 (2008)
14. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1), 193–218 (1985)
15. Hubert, L.J., Levin, J.R.: A general statistical framework for assessing categorical clustering in free recall. Psychological bulletin **83**(6), 1072 (1976)
16. Jacinto, F.V., Esteller, M.: Mutator pathways unleashed by epigenetic silencing in human cancer. Mutagenesis **22**(4), 247–253 (2007)
17. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. Wiley-Interscience, NY (1990)
18. Leung, M.Y., Marsh, G.M., Speed, T.P.: Over-and under-representation of short DNA words in herpesvirus genomes. Journal of Computational Biology **3**(3), 345–360 (1996)
19. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on, pp. 911–916. IEEE (2010)
20. Lothaire, M.: Applied combinatorics on words, vol. 105. Cambridge University Press (2005)

21. MacIsaac, K.D., Fraenkel, E.: Practical strategies for discovering regulatory DNA sequence motifs. PLoS computational biology **2**(4), e36 (2006)
22. Marino-Ramrez, L., Spouge, J.L., Kanga, G.C., Landsman, D.: Statistical analysis of over-represented words in human promoter sequences. Nucleic Acids Research **32**(3), 949–958 (2004)
23. Milligan, G.W., Cooper, M.C.: An examination of procedures for determining the number of clusters in a data set. Psychometrika **50**(2), 159–179 (1985)
24. Milligan, G.W., Cooper, M.C.: A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research **21**(4), 441–458 (1986). DOI 10.1207/s15327906mbr2104_5. PMID: 26828221
25. Nakamoto, T.: Evolution and the universality of the mechanism of initiation of protein synthesis. Gene **432**(1), 1–6 (2009)
26. Nuel, G.: Numerical solutions for patterns statistics on markov chains. Statistical Applications in Genetics and Molecular Biology **5**(1) (2006)
27. Percus, J.K.: Mathematics of genome analysis, vol. 17. Cambridge University Press (2002)
28. Régnier, M.: A unified approach to word occurrence probabilities. Discrete Applied Mathematics **104**(1-3), 259–280 (2000)
29. Reinert, G., Schbath, S., Waterman, M.S.: Probabilistic and statistical properties of words: an overview. Journal of Computational Biology **7**(1-2), 1–46 (2000)
30. Robin, S., , Rodolphe, F., Schbath, S.: DNA, words and models: statistics of exceptional words. Cambridge University Press (2005)
31. Robin, S., Daudin, J.J.: Exact distribution of word occurrences in a random sequence of letters. Journal of Applied Probability **36**(1), 179–193 (1999)
32. Robin, S., Daudin, J.J.: Exact distribution of the distances between any occurrences of a set of words. Annals of the Institute of Statistical Mathematics **53**(4), 895–905 (2001)
33. Robin, S., Daudin, J.J., Richard, H., Sagot, M.F., Schbath, S.: Occurrence probability of structured motifs in random sequences. Journal of Computational Biology **9**(6), 761–773 (2002)
34. Rousseeuw, P.J.: Least median of squares regression. Journal of the American Statistical Association **79**, 871–880 (1984)
35. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
36. Saxonov, S., Berg, P., Brutlag, D.L.: A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences **103**(5), 1412–1417 (2006)
37. Stefanov, V., Pakes, A.G.: Explicit distributional results in pattern formation. The Annals of Applied Probability pp. 666–678 (1997)
38. Stefanov, V.T.: On some waiting time problems. Journal of Applied Probability **37**(3), 756–764 (2000)
39. Stefanov, V.T.: The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach. Journal of Applied Probability **40**(4), 881–892 (2003)
40. Steinley, D., Brusco, M.J., Hubert, L.: The variance of the adjusted rand index. Psychological methods **21**(2), 261 (2016)
41. Tavares, A.H., Afreixo, V., Rodrigues, J.a.M., Bastos, C.A.C.: The symmetry of oligonucleotide distance distributions in the human genome. In: ICPRAM (2), pp. 256–263 (2015)
42. Tavares, A.H.M.P., Afreixo, V., Rodrigues, J.a.M., Bastos, C.A.C., Pinho, A.J., Ferreira, P.J.S.G., Brito, P.: Detection of exceptional genomic words: A comparison between species. In: Proceedings of 22nd International Conference on Computational Statistics (COMPSTAT), pp. 255–264 (2016)
43. Tavares, A.H.M.P., Pinho, A.J., Silva, R.M., Rodrigues, J.M.O.S., Bastos, C.A.C., Ferreira, P.J.S.G., Afreixo, V.: DNA word analysis based on the distribution of the distances between symmetric words. Scientific reports **7**(1), 728 (2017)

Intentionally blank page.

Part III

Discussion and Conclusions

Chapter 10

Discussion

Throughout the present work, exploratory data analysis was performed on datasets of genomic sequences to maximize data acquaintance, bringing insights that were not evident beforehand or appeared to be worth investigating. The patterns found by exploring the data suggested hypothesis which led to interesting follow-up studies. Moreover, they give clues to properly infer what model would be appropriate to create knowledge (conclusions) about the dataset. As an example, let us mention the unexpected similarity found between the frequencies of inter-w distance of reversed complementary words. The identification of this behaviour in a few words of length four with irregular distribution patterns, led us to conduct a systematic study about it (Article II) and has influenced the definition of a dissimilarity measure to compare inter-word distance distributions (Article V) and on tracing relevant features for the clustering procedure of such distributions (Article VI).

The exploratory data analysis (EDA) procedures applied fall back on summary statistics, residual analysis, principal component analysis and cluster analysis. Having in mind that EDA is mostly a philosophy of data analysis, it is not restricted to techniques described in a *compendium*; sometimes it is necessary to trace new ways of looking at specific data. This was precisely the situation we have faced at some stages of our research study about genomic sequences.

The research questions put forward for this thesis met three main objectives, namely, similarity assessment, clustering and outlier detection. This chapter is dedicated to the integrated discussion of the major accomplishments along research pursued. The specific achievements that resulted from the research processes described within the set of original studies are now discussed together as a *continuum*, in articulation with the research questions presented above.

10.1 Random variables under study

To unify the terminology used in previous articles, improving the discussion of stated procedures and results, a standard nomenclature is introduced. DNA sequences are herein considered as finite totally ordered sets with elements belonging to the nucleotide alphabet $\mathcal{A} = \{A, C, G, T\}$. In such sequences, words of different sizes can be identified. By word of length k, we refer to a sub-sequence $w = x_1x_2\dots x_k$ with $x_i \in \mathcal{A}$. The reversed complement of a word w is obtained by reversing the order of the letters in the word interchanging letters $A - T$ and $C - T$, and is denoted by w'. Let Ω be a set of genomic sequences, and $\mathcal{A}^k$ be the set of all genomic words of length k. Considering an arbitrary, but fixed, $S \in \Omega$ and $w \in \mathcal{A}^k$, the following random variables are defined:

L_S^w - position (local index) of word w in S;

N_S^w - frequency of word w in S;

D_S^w - distance between consecutive occurrences of word w in S;

DR_S^w - distance between w and its reversed complement w', without w or w' between them, in S.

Note that both N_S^w and D_S^w can be defined as a function of L_S^w. In fact, if $\widehat{L} = (l_1, \dots, l_m)$, $\widehat{N}$ and $\widehat{D}$ denote a concretization of L_S^w, N_S^w and D_S^w, then

$$\widehat{N} = m$$

and

$$\widehat{D} = (l_2 - l_1, l_3 - l_2, \dots, l_m - l_{m-1}).$$

Random variable DR_S^w can also be defined as a function of L_S^w and $L_S^{w'}$, but its formula is not so straightforward. The discrete probability distribution functions of N_S^w, D_S^w and DR_S^w random variables are denoted as

$$f_N^{w,S}(t) = P(N_S^w = t) \tag{10.1}$$

$$f_D^{w,S}(n) = P(D_S^w = n) \tag{10.2}$$

$$f_{DR}^{w,S}(n) = P(DR_S^w = n) \tag{10.3}$$

with $t \in \mathbb{N}_0$ and $n \in \mathbb{N}$. Whenever it is not misunderstood which genomic sequence is being studied, the probability distribution functions are simply denoted as f_N^w, f_D^w and f_{DR}^w. The research project here described focus on the exploration of these three distributions, and on the development of statistical procedures to reveal relevant features about nucleotide sequences: f_N^w is the probability of occurrence of word w, dealt within Article I; f_{DR}^w is the probability distribution of distances between near reversed complementary words, tackled in Article IV; and f_D^w is the probabilitry distribution of the inter-word distances, explored trough the remaining four articles.

In some studies it may be opportune to study distances restricted to a subset of their domain. For example, to compare distributions that have distinct domains, or if, for some reason, only distances lower or greater than a given number (say $> k$ or $<$1000) are of interest. In these situations the studied distance distributions refer to a conditional inter-word distance distribution.

To define a distance distribution "profile" for a set of words, functions of the previous random variables may be considered. For instance, to explore the inter-word distances regarding words in the set W we could consider the random variable

$$D_S^W = \sum_{w \in W} D_S^w \tag{10.4}$$

whose corresponding probability distribution function, denoted as f_D^W, verifies the following relation

$$f_D^W(d) = \frac{1}{n^W} \sum_{w \in W} (n^w - 1) f_D^W(d) \tag{10.5}$$

where n^W is the number of words in that specific set, and n^w are the corresponding realizations of N^w. This reasoning could be applied to globally study patterns of inter-word distance distributions regarding symmetric word pairs, e.g. $W = \{CC, GG\}$, as well as to study global patterns regarding words belonging to the same equal composition group, e.g. $W = \{CC, CG, GC, GG\}$. The corresponding probability distribution functions are denoted as $f_D^{ww'}$ and f_D^{ECG}, respectively.

Throughout the remaining of the text, we refer to the <u>i</u>nter-<u>w</u>ord <u>d</u>istance simply as *iw*D, and to the near-<u>r</u>everse-<u>c</u>omplements <u>d</u>istance simply as *rc*D. Moreover, expressions f_N^w, f_D^w and f_{DR}^w interchangeably denote the probability distribution of variables N_S^w, D_S^w and DR_S^w (respectively), or empirical distributions obtained by a concretization of them.

A genomic sequence could be randomly generated through a stochastic process. Let S_R denote a random sequence $\{X_i\}_{i \in I}$, where X_i are random variables associated with the generation of one symbol from $\mathcal{A}$, I is an ordered and numerable set of indexes. If S_R is generated by a Bernoulli model, i.e. each symbol X_i is generated independently of the other symbols, then the following equality is verified for all $x_1 \ldots x_i \in S_R$,

$$P(X_i = x_i | X_{i-1} X_{i-2} \ldots X_1) = P(X_i = x_i)\,. \tag{10.6}$$

On the other hand, if the sequence is randomly generated by a m-order Markov model, i.e. if the probability of occurrence of X_i depends on the previous m symbols ($1 \leq m \leq i$), then the equality

$$P(X_i = x_i | X_{i-1} X_{i-2} \ldots X_1) = P(X_i = x_i | X_{i-1} X_{i-2} \ldots X_{i-m}) \tag{10.7}$$

is verified for all $x_1 \ldots x_i \in S_R$. Note that a memoryless source is interpretable both as a Bernoulli source and a zero-order Markov source.

Let S_R be a sequence generated by one of the previous models. In the case where the Bernoulli model is used, S_R is referred to as a random sequence under the nucleotide independence scenario; else, as a random sequence under a m-order Markov dependence scenario.

10.2 Word frequencies

Previous studies evaluated the prevalence of the symmetry phenomenon in several organisms, by analysing their genomes locally (stretches of DNA) or globally (complete genome), and confirming that the frequency distribution of words along nucleotide sequences tend to be statistically similar [150].

Random variables N_S^w and $N_S^{w'}$ are considered in this thesis in order to study the symmetry phenomenon. In particular, the dissimilarity between the frequency of a word and the frequency of its reversed complement is evaluated regarding the word frequency dissimilarities within the corresponding equivalent composition group (ECG).

To study the symmetry phenomenon, a measure of "exceptional symmetry" was proposed in [8], that evaluates the symmetry above that expected in independence contexts. The authors state that the frequency of a word is more similar to the frequency of its reversed complement than to that of any other word in the same ECG. A pair of words formed by a word and its reversed complement is called *symmetric word pair*. In spite of the fact that their R measure allows obtaining ranks of genomic words by exceptional symmetry, two undesirable features were observed in such ranks. First the R measure may produce distinct values for each word of the same symmetric pair. Second, two symmetric word pairs with identical dissimilarities between their frequency of occurrence, could present distinct R values. To exemplify, let us consider a toy example where n_{w_i} denotes the frequency of word w_i and $n_{w'_i}$ that of its reversed complement such that: $n_{w_1} = n_{w'_1} + 1 = 20$, $n_{w_2} = n_{w'_2} + 1 = 9$ and $n_{w_3} = n_{w'_3} + 1 = 1$. The symmetric word pair with occurrences nearest to the average number of occurrences (equal to 9.5) is ranked as less exceptional than the word pair whose number of occurrences is most distant.

In order to avoid the described disadvantage of the R measure we introduced a new measure, T, to assess the word pair symmetry effect (Article I). This new measure of exceptional symmetry overcomes the disadvantages detected in the previous measure. Namely, the effect for both words of one symmetry word pair is the same, $T(w) = T(w')$; and the global deviation between words of the same ECG is taken into account, instead of the deviation to the mean of the number of occurrences thereof (as in the previous measure). Moreover, measure T avoids indetermination cases produced with R by self symmetric words ($w = w'$).

To identify word pairs (w, w') as exceptional their $T(w)$ value is compared with a critical value obtained by control experiments. By control experiment we mean sequences generated under the nucleotide independence scenario and assuming the validity of the second parity rule ($\%A = \%T$ and $\%C = \%G$). Under this scenario all words in the same ECG have equal expected probability of occurrence.

Word pairs with T value higher than the third quartile of those obtained from random sequences are flagged as exceptional. Our results show that the symmetry effect value has the potential to discriminate between species groups. And that there are sets of words which present high symmetry effect in all species under study.

10.3 Word distances

10.3.1 Expected distance distributions

Let $w = x_1x_2 \ldots x_k \in \mathcal{A}^k$ be a generic word and S_R be a nucleotide random sequence generated under a Markov model. Considering that the sequence is read through a sliding window of length k, with word overlapping, what is the probability that consecutive occurrences of w are 10 symbols apart? And 12 symbols? And 1978 symbols?

The exact distribution of the *iw*D can be deduced using a state diagram, which represents the progress made towards identifying w as each symbol is read from the sequence. A general state diagram and the corresponding transition matrix of probabilities are described in Article III, from which $f^w_{S_R}$ is easily obtained. The state diagram is composed by k non-absorbing states and one absorbing state, reached when a new occurrence of w is identified in the sequence. Later, we became aware that our approach to obtain the exact distribution of inter-word distances is a special case of a procedure based on finite Markov chain embedding [76]. As pointed out by Fu in [76], to find the transition matrix for a given word requires "a deep understanding of the structure of the specified pattern". In Article III explicit general expressions are proposed both to identify the initial state and to compute the transition matrix between non-absorbing states, based on the concept of word overlap. The expected *iw*D distribution under an independent nucleotide generation hypothesis, can be easily computed for any whole-genome and for any genomic word, using only four input parameters: the nucleotide frequencies in the sequence.

The Markov chain approach described in Article III was also explored to obtain *iw*D distributions under k-order Markov dependence. In such scenario, the transition matrix depends on the 4^k frequencies of words of length k that occur in the sequence. In the development of the research described in Article IV this procedure was adjusted to generate the exact *rc*D distributions under k-order Markov dependence. The rationale behind is the same, however the state diagram accommodates not just one but two absorbing states (reached when a new occurrence of w or w' is identified in the sequence).

10.3.2 Dissimilarity measures for iwD distributions

Data collected through a concretization of the D_S^w random variable for two distinct words, say w_1 and w_2, give rise to two datasets of *iw*D. How to evaluate the similarity between the corresponding *iw*D distributions? When the question of dissimilarity measurement between distributions was initially posed in this study, a survey was made on this topic.

To compare two discrete probability distributions several dissimilarity measures may be used. A possible approach is to assess the homogeneity (or agreement) between the two empirical distributions, using the chi-square statistic, and then use an effect size measure to assess the dissimilarity between them.

Indeed, the chi-square statistic was used to determine whether *iw*D frequency counts are distributed identically across empirical and random sequences, corresponding to an agreement between $f_D^{w,S}$ and f_D^{w,S_R} (Article III); and similarly, for *rc*D distribution from empirical and random sequences, i.e. $f_{DR}^{w,S}$ and f_{DR}^{w,S_R}(Article IV). The same statistics was used to asses the similarity between pairs of distance distributions in a same genomic sequence (the human genome) associated with reversed complementary words, i.e. f_D^w and $f_D^{w'}$ (Article II). The dissimilarity between each pair of distributions was measured by an effect size measure.

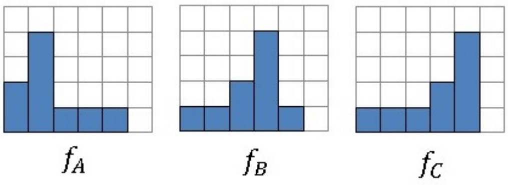
The similarity between the *iw*D distribution of words belonging to the same ECG was evaluated. Let $w_1, \ldots, w_{n^*}$ denote the words on the same ECG and n^* denote the number of words in that ECG. The chi-square statistic was used to assess whether $D^{w_1}, \ldots, D^{w_{n^*}}$ variables are identically distributed in the Human Genome (Article II). An effect size measure based on chi-square statistic was interpreted as a measure of the global disagreement between the distributions associated to a same ECG set of words. In Article II, we also assess the similarity between the *iw*D distribution, f_D^w, and the distance distribution profile of its ECG, f_D^{ECG}, as defined in Equation (10.5). To assess whether D^w and D^{ECG} variables are identically distributed in the Human Genome, the chi-square statistic was used. Then, an effect size measure based on its value was computed and interpreted as a measure of the global disagreement between two distributions.

The literature on dissimilarity measures is vast. Some common dissimilarity measures between probability distributions may be classified as bin-to-bin measures or cross-bin measures. The former means that given two probability vectors P and Q, the dissimilarity measure is based on differences between corresponding elements P_i and Q_i; the latter refers to measures that take into account cross-bin relationships [143]. From this perspective, the well known L_p norms and Kulback-Leibler divergence are bin-to-bin distances, while Mallows (or Wasserstein), Mahalanobis and Kolmogorov-Smirnov distances are cross-bin measures.

The decision on which measure of dissimilarity should be used should take into account the nature of the data. For instance, noticing that probability vectors represent the parts of a whole, carrying only relative information of it components, *iw*D distributions could be interpreted as compositional data (with the disadvantage of being a vector shuffle invariant approach). In the context of Compositional Data Analysis, some usual measures of the difference between two compositions are Mahalanobis, Minkowski, City Block and Aitchison distances [128]. The Earth Mover's Distance [169] is a cross-bin distance widely used for measuring dissimilarity between histograms, in the fiel of Computer Science. It can be interpreted as the least amount of work needed to transform one histogram into the other histogram; when used to measure the difference between two probability distributions it is exactly the same as the Mallows distance [117; 201].

The discussion around distribution dissimilarities is extensive and prolific. For example, Aherne *et al.* [15] highlights the advantageous properties of the Bhattacharyya metric over the chi-squared statistic for comparing frequency distributed data. According to Aitchison, an adequate measure for Compositional Data analysis should be invariant to scale, permutation and perturbation [16], which exclude the Euclidean distance. In the context of Histogram Data, Verde and Irpino [200] argue the superiority of the Mahalanobis-Wassertein distance. For a review on dissimilarity measures see [44; 45; 52; 143].

The dissimilarity measure adopted to perform a statistical study should take into account both the nature of the data and which features are important for the subject matter. This means that higher dissimilarity values should be obtained under a perceived feature dissimilarity between elements of the dataset; and lower dissimilarity values should be achieved under a perceived feature similarity in data elements. To clarify, consider f_A, f_B and f_C as the distributions represented in Figure 10.1. A quick visual perception will rank f_A and f_B as more similar than f_A and f_C, as it is for the Mallows distance (Earth Mover's Distance). However, according to the Euclidean distance, f_A is so similar to f_B as it is to f_C. As a remark, the three distributions could be considered as equidistant by a measure that assumes invariance of horizontal translations.



	$d(f_A, f_B)$	$d(f_B, f_C)$	$d(f_A, f_C)$
Euclidean	0.50	0.42	0.50
Mallows	0.89	0.44	1.33

Figure 10.1: Dissimilarity measure between two distributions. Let f_A, f_B and f_C be distributions associated with histograms represented at left. Euclidean and Mallows dissimilarity values are given at right.

Throughout the development of this research project we used the Mallows distance to measure the dissimilarity between pairs of *iw*D distributions, in order to take into account the similarity between non-overlapping parts of the distributions. Such values were then used to define a dissimilarity matrix and to perform hierarchical cluster analysis (not shown in this thesis).

To study the similarity between the *iw*D distribution of pair of reversed complementary words, w and w', we defined a dissimilarity measure that puts in evidence how consistent are the peaks of such distributions. The *peak dissimilarity measure* introduced in [188], relies precisely on the observed differences in location and magnitude of a given number of peaks. The rationale behind the definition of this measure is as follows.

An initial exploratory analysis of genomic datasets evidenced that the *iw*D distribution of some words is very similar to that of its reversed complement, despite their irregular pattern. The observed irregular patterns result, essentially, from the existence of peaks of frequency, and are not expected in random scenarios. The fact that pairs of reversed complementary words are associated with a "same" irregular pattern, led us to explore this topic (related with questions Q7). Therefore, a dissimilarity measure based on the observed differences between the peaks of the distributions, was assumed as adequate for assessing the similarity of *iw*D distributions of words that are reversed complements.

The peak dissimilarity is compared with two earlier dissimilarity measures, the Euclidean distance and the Jeffreys divergence, and we argue for its superiority in the analysis of distance distributions between pairs of reversed complementary words (Article V). While the Jeffreys divergence, d_J, is quite sensitive to small frequencies, the Euclidean distance, d_E, is sensitive to the presence of a few high frequencies. In the presence of sparse distributions both low and high relative frequency values are expected, which strongly affect the results of d_J and d_E. The proposed peak dissimilarity ignores small frequencies and evaluates the disagreement between the sizes of the n strongest peaks, which are taken into account even when their locations do not coincide. Moreover, the peak size differences are scaled by the highest peak sizes observed in each distribution.

10.3.3 Clustering procedure

The knowledge accumulated through the exploratory analysis performed along the various studies focusing on the *iw*D distributions, drew our attention to two main characteristics of their plots. Such characteristics are related with an overall trend and some upward peaks, i.e. distances presenting frequencies much higher than that presented by neighbor distances. Concerning the overall trend, some distribution plots present higher rates of decreasing and shorter tails, while others present an opposite baseline behaviour. Distance distributions also display different peak behaviors, which could be roughly described as: one or more isolated very spiked frequency values (isolated peaks); one or more consecutive distances with high frequency values (peak islands); smaller peaks spread along the function domain; or slightly pronounced peaking behavior. Examples of such behaviours are depicted in Figure 1 and

Figure 7 of Article V.

Given the specific characteristics of these data sets, we developed a clustering procedure that first decomposes each distribution into a baseline and a peak distribution, described in Article VI. By noticing that a properly scaled Gamma density function provides a good fit of the baseline, a least trimmed squares approach is used to estimate a baseline Gamma distribution. The obtained Gamma baseline is outlier-robust, since the objective function to minimize excludes a small percentage of residuals (the highest 5% residual were excluded). Then, the peak structure on top of that baseline is captured by assessing the extremity of the observed frequencies, which allows distinguishing peaks from sampling variability. Therefore, the initial dataset of *iw*D distributions is transformed into a dataset of baselines and a dataset of peak functions. The next step consists in applying a clustering algorithm. It could be applied only on the baseline datasets, only on the peak functions, or on both baseline and peak functions. In the first case, words (distributions) are clustered according to the similarity between their *iw*D trend; in the second case, they are clustered according to the peak behaviour of their *iw*D distribution; and, in third case, both trend and peak features are considered to form the clusters of words.

An advantage of k-mean over hierarchical methods is that it does not require the computation of the dissimilarity matrix between all pairs of elements, making it suitable to use on datasets with a large number of elements. The proposed procedure was constructed taking into account the application of k-means, which performs best when the input variables are uncorrelated. Therefore, the set of baselines and the set of peak functions are pre-processed before carrying out the clustering algorithm. First, peak and baseline functions are converted into cumulative functions. Then, each dataset of cumulative functions is transformed by principal component analysis to fulfill a twofold objective: to reduce the data dimension and to create uncorrelated variables.

To assess how well the clustering procedure performs, it was applied over controled data, i.e. synthetic data. The agreement between the resulting partition and the true partition could be assessed by the Adjusted Rand Index (ARI), whose maximum value is equal to one and indicates matching classifications. For each one of the considered scenarios of the simulation study, ARI values close to 1 were obtained (> 0.9) by selecting an appropriate data matrix (peak matrix, baseline matrix or both peak and baseline matrices).

The procedure was applied to real genomic data allowing obtaining a partition of the words of same length (lengths 3 and 5) according to the peak patterns and the baseline of the corresponding *iw*D distributions. In the particular case of trinucleotides, we observed that for a large number of clusters there is a tendency for the formation of clusters composed by a pair of reversed complementary words.

10.4 Outlier detection

The detection of phenomena and universal rules is often triggered by the observation of patterns that repeat themselves in all (or almost all) elements of a large dataset. The nature of such underlying rule, law or phenomenon can be deterministic or probabilistic, and an explanation for its occurrence must be sought. Elements that deviate strongly from such rule uncover a breakdown mechanism that could be of interest.

A key issue of statistics, is to develop ways of presenting the data that highlight interesting and important features, which may imply to investigate central characteristics, variability and the presence of outliers.

Throughout this study we point out some features of genomic sequences as deserving a more careful study, such as the total number of reversed complementary words or the distribution patterns thereof. Nucleotide sequences with similar number of complementary nucleotides obey to an extension of the symmetry phenomenon stated by Chargaff's second parity rule; in particular, reversed complementary words whose *iw*D distributions have a similar pattern are in compliance with a distinct parity "rule". The identification of atypical genomic words is a transversal theme throughout this thesis. However, no general procedure was developed for their identification. In fact, a survey on word outlier detection would imply a definition of atypical word or atypical distribution, which could be difficult to define.

Regarding a feature of interest, atypical words can be sought by measuring the difference between the observed behavior and a reference behavior. The latter may reflect a global behavior observed in the data set, or the expected behavior under some theoretical considerations. The definition of a reference, as well as the establishment of a measure of discrepancy between the reference and the observed values, seems to be unavoidable for outlier detection. The procedure of reducing data to a numerical value and then look for outlying values in the obtained dataset (e.g. reduce an *iw*D distribution to a discrepancy value) is sensitive to both the reference and the discrepancy measure adopted.

The identification of observations worthy of the epithet of outlier can be performed by evaluating the depth or the outlyingness of the discrepancy values. Some common methods are Standard Deviation Method, Z-score, modified Z-score or Tukey fence. The depth of a point is inversely related to its outlyingness and trimmed regions depth-based can be interpreted as quantiles. These concepts have proved extremely fruitful and a rich statistical methodology has developed around the concept of data depth (for a review see [134; 216], and [137] in portuguese language). Tukey depth measure for d-dimensional points generalizes the univariate median to the multivariate case (median is a point with maximum Tukey depth). The degree of outlyingness concept, independently proposed by Stahel [178] and by Donoho [63], relies on the assumption that a multivariate outlier should also be an univariate outlier in some projection. Thus, the measure of the outlyingness is based on one-dimensional projection of multivariate data.

The procedure of reducing data to a numerical value and then look for outlying values in the obtained dataset, although simple, involves loss of information. The most sensitive aspect of this approach lies in the establishment of the discrepancy measure to adopt, as it may or may not put in evidence deviations regarding features of interest.

Outlying words By comparing the observed value with the expected value, according to a theoretical model or to global values observed in the dataset, we identified atypical words regarding different characteristics:

- Assessing the dissimilarity between $f_N^{w,S}$, $f_N^{w',S}$ and $f_N^{w_i,S_R}$ for all w_i in the same ECG, led to the identification of symmetric word pairs whose similarity between the frequency of occurrence is above that expected under the independence scenario (Article I).
- Evaluation of the agreement bewteen $f_D^{w,S}$ and f_D^{w,S_R} led to the identification of word pairs of two types of exceptionality (Article III). One is related with words having iwD distributions highly dissimilar to the one expected under the independence scenario. The other set holds words whose dissimilarity is not too strong and, simultaneously, present strong local dissimilarity (local deviations evaluated by residual analysis).
- Determining whether iwD frequency counts are distributed identically across empirical and random sequence, symmetric word pairs were identified with very dissimilar f_D^w and $f_D^{w'}$ (Article II). Given that in an independence scenario it is expected that symmetric words have similar iwD distributions, strong dissimilarity is pointed out as an exceptional behaviour.
- Measuring the similarity between $f_{DR}^{w,S}$ and f_{DR}^{w,S_R} allows identifying words of length k whose rcD distribution is very dissimilar to that expected under the k-order dependence scenario.

In the design of the clustering procedure described in Article VI, the concept of atypical frequency is also present. In fact, when differentiating between frequencies that are peaks or a result of mere sample variability, an univariate analysis was performed at each point of the domain, evaluating the distance between the observed frequency value and the expected frequency according to the Poisson distribution. A frequency is determined to be an outlier if the observed value surpasses a given quantil.

Functional approaches In the context of finding outlying distributions, an alternative to the reduction of each distribution to a dissimilarity measure is the application of a univariate procedure, point by point. This procedure, which can be seen as a residual analysis, allows identifying curves that are outliers at a given domain. However this procedure does not allows identifying all type of atypical curves, since it does not consider dependencies along the domain.

A curve could be considered outlier by its global behavior, even when it displays an inner behavior at any domain point [67]. A functional approach allows overcoming the limitation inherent to the residual analysis. Fraiman and Muniz [72] were the first to introduce a depth measurement for functional data, whose central idea is to measure how long a curve lies in the middle of the group of all other trajectories. The generalization of the concept of depth to functional data allows for the evaluation of the centrality of a curve in relation to a set of curves. Furthermore, the ordering of the curves in ascending order of their depth measurement provides a rank from inner to outer distributions.

There is an increasing interest in developing procedures for detection of outliers in functional data. The procedure must be preceded by an adequate definition of the outlier concept inherent to the application context and by the development of methods for its identification.

For a fixed word length, the set of 4^k distance distributions can be seen as a sample of curves, which may be treated as (discrete) functional data. Throughout the development of this project we have experimented several functional approaches in the exploration of outlier distance distributions in our genomic datasets. For instance, we applied the method proposed in [59], which makes use of functional depths and trimming bootstrap to estimate the cutoff of depth values. We also applied the functional highest density region (HDR) boxplot proposed in [101], which makes use of the first two robust principal component scores, Tukey's depth and highest density regions (this work is not reported in Part II).

Let wD_3 denote the dataset consisting of the iwD distributions of all words of length $k = 3$ in the complete human genome, and $dmax$ denote the maximum distance of such distributions. Outlier detection by depth measures was performed in the wD_3 dataset with $dmax = 200$, considering several depth measures: Fraiman-Muniz depth, dpFM (see [72]); h-modal depth, dpMO, random projection depth, dpRP, and random projection depth using derivatives, dpRPD (see [59]); and random Tukey depth, dpRT (see [58]).

In general, it has been observed that the procedure does not necessarily detect the same curves at each iteration. When the procedure is based on dpFM, it tends to flag as outlier curves that reach high values of frequency in the first distances; based on dpMO, dpRT or dpRP, it identifies flatter curves; and based on dpRPD, it retrieves curves of the two types. The computational speed of the outlier detection method depends on the depth measure: dpRT is the fastest (2.78 min), closely followed by dpFM (4.25 min) and dpRP (4.77 min), while dpRPD (3.51 hrs) stands out for its slowness. By increasing the size of the dataset the procedure becomes considerably more time consuming. For example, considering the wD_5 with $dmax = 400$, dpFM takes 53.45 minutes and dpMO takes 3.77 days (mean values of 10 repetitions).

We also applied the functional highest density region (HDR) boxplot proposed in [101], which makes use of the first two robust principal component scores, Tukey's depth and highest density regions. Figure 10.2 shows the HDR boxplot related with the wD_3 dataset in the human genome with $dmax = 200$. The iwD distribution of the flagged outliers are outlined with distinct colors.

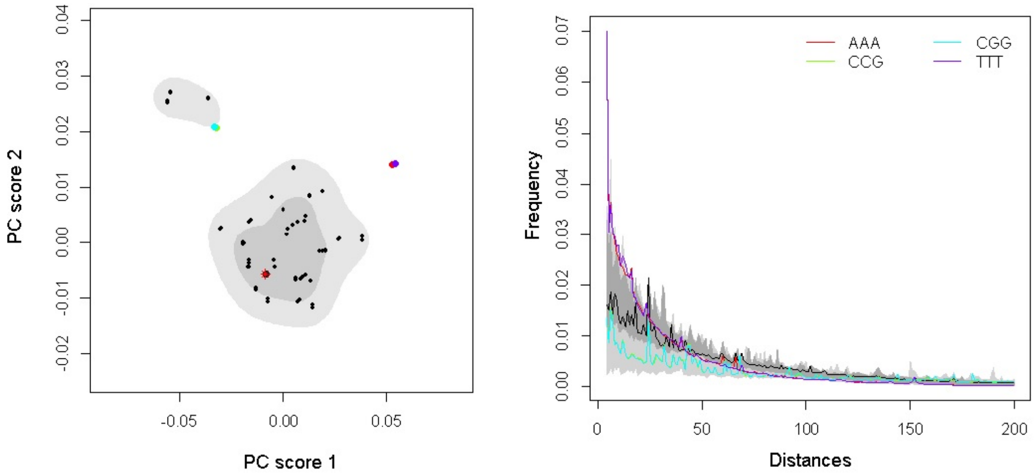


Figure 10.2: Highest density region boxplot for the iwD distributions of words of length $k = 3$ (left): the black line is the modal curve. The curves outside the outer region are outliers. The iwD distribution of flagged outliers are represented in color (right).

More recently, a new procedure to detect outlying functions was proposed by Rousseeuw *et al.* [167]. They introduced a measure of directional outlyingness (DO), based on the Stahel-Donoho outlyingness. By assigning a value of outlyingness to each gridpoint of the function domain, they designed a procedure that allows detecting outlying functions and outlying parts of a function. Such distinct types of outlying curves may be highlighted in an graphical tool that they call functional outlier map.

We apply the DO measure to identify atypical distance distributions between genomic words (see preliminary results in [191]). The results indicate that the DO procedure is promising for our problem. It allowed capturing iwD distributions whose shape strongly differs from the majority, distributions with several strong peaks, and distributions with peaks at subdomains where no other peaks occur.

However, the efficiency of the procedure falls as the word length increases. For words of length $k = 7$ there are many distributions flagged as exceptional. The functional outlier map does not show a cloud of points near the origin, on the contrary it is a cloud of points tracing a kind of two semi-arcs, as shown in Figure 10.3. Indeed our data is characterized by a strong peak behaviour, spread along the domain. The log transformation proposed in [167] to transform the right-skewed distribution of DO values into a closer gaussian distribution is not effective in our wD_7 dataset.

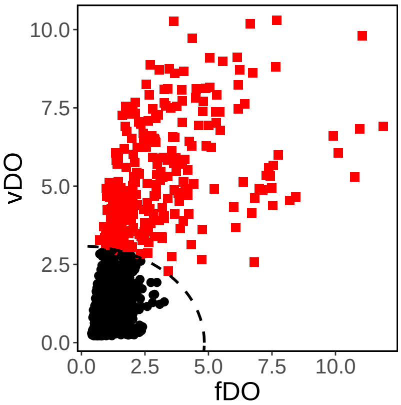


Figure 10.3: Functional outlier map of the *iw*D distributions of words of length $k = 7$, in the complete human genome.

The results obtained by applying such functional procedures for outlier detection on our genomic data lead us to question whether it will even be possible to identify a set of distributions that are somehow suspicious or surprising as they do not follow the same pattern as that of the rest of curves. Indeed such methods seem to be designed for sets of distributions that are generated by a same process (which is indeed the assumption in [59]). Faced with large datasets of distributions characterized by a strong peak behaviour, spread along the domain, we wonder if the detection of outlying curves in such heterogeneous datasets will even be possible by those functional approaches.

If large heterogeneous datasets where distinct patterns coexist can validly be clustered, then the class labels may provide a meaningful description of similarities and differences in the data. By representing each group by its class label, the inicial dataset is reduced to a given number of distributions. So, why not to use a functional outlier procedure on the set of (class label) distributions to evaluate the existence of outliers? Apart from that, if one of these distributions is flagged as atypical, why not to introduce a concept of *atypical group*?

An outlier definition widely reported in the scientific literature is the one proposed by Grubbs [91] and quoted in Barnett and Lewis [28]: *An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.* This general definition of outlier is vague and becomes meaningful only under a given context or application. As a result, outliers have been defined in various ways and a wide variety of outlier detection methods have been drawn from Computer Science and Statistics, and outlier detection is still considered an open-ended problem. To the best of our knowledge, the concept of outlier group is new and worth investigating.

10.5 Present and Future work

The work described in this thesis is organized into three research topics dedicated to the development of procedures for similarity assessment, outlier detection and clustering. During its development, several scientific results and observations gave origin to a reorganization of ongoing and future research topics.

This research project is clearly associated with the analysis of complete genomes, since our goal is to develop procedures able to perform comparisons between genomic sequences, within and between species. However, it is not inappropriate to apply the procedures herein proposed on small stretch of DNA to perform local analysis.

In the future we intend to continue this research on the study of similarities between distance distributions. Thus, one possible step is to assess how well our results hold up in a local analysis of genomic sequences.

Another topic for further research is the application of other statistical methods (such as classification) on genomic data after applying the decomposition technique developed in Article VI.

A feature that we detected and remained to investigate is the identification of a set of words with patterns of rcD distribution that are surprisingly different from those of the majority. Figure 10.4 displays the case of distances between $w = CACTGCA$ and its reversed complement, in the complete human genome. This distribution, $f_{DR}^{CACTGCA}$, will certainly be mentioned in future works that embrace the research on clustering distributions and outlier detection.

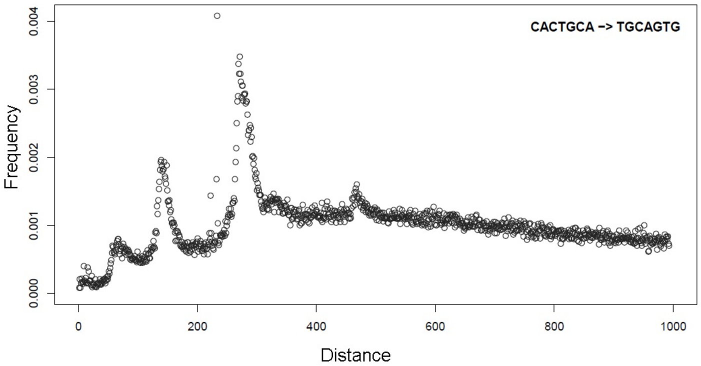


Figure 10.4: Distance distribution between $w = CACTGCA$ and its reversed complement, in the complete human genome.

Ongoing work is related with the issue of atypical group identification. Challenges that arise in atypical group detection first concern the identification of a segmentation of the data, and secondly flagging the atypical segments. To the best of our knowledge, this concept is new. We focus our work on the detection of atypical groups in data that can be represented by a function and, in particular, in distance distributions between genomic words. We are particularly interested in studying a method that recovers groups of words with similar distribution patterns and, in particular, those very small groups with a distribution pattern which is demarcated from the majority, here called an atypical group. The interest in atypical groups identification stems from the fact that word clusters may provide useful applications in DNA sequence characterization, such as sequence classification and function prediction. Moreover, atypical distribution patterns may be related with words that have specific biological meaning.

Chapter 11

Conclusions

The development of this thesis has allowed for the characterization of distance distributions between genomic words and the development of some procedures to effectively extract information from this type of data. Considering the research questions initially proposed the following milestones have been accomplished:

- The definition of a new measure of dissimilarity between distributions that focuses on the gaps between the locations of their peaks and the difference between the sizes of these peaks.

- The development of an innovative research tool for clustering distributions based on baseline and peak features.

An open question is the definition of atypical word-distance distribution. Throughout the work several meaningful criteria of exceptionality were explored regarding features of interest of the nucleotide sequences. We are convinced that the identification of a more general procedure for the identification of atypical distributions may involve the introduction of the *atypical group* concept. This is the goal of our ongoing work and, to the best of our knowledge, this concept is new.

In the more restricted context of the application of these methodologies in the field of genomic data, the following milestones have been accomplished:

- The definition of a new measure of exceptional symmetry to analyse exceptional symmetry phenomenon by word. The word exceptional symmetry values contain information specific to the species and seem to contain information about the species evolution.

- The detection of exceptional words based on the discrepancy between their iwD and the expected one under a random scenario. We evaluated the discrepancy between real sequences and the random background, as a way of emphasizing the contribution of selective evolution, and found that the differences mimic, to a certain extent, the evolutionary relations between the species.

- An unexpected similarity between the *iw*D distribution of some words and that of its reversed complement was found. After a systematic study, it was shown that the lack of homogeneity between symmetric words is negligible, for words of length up to five.

- The definition of a new distance between distributions based on the location and magnitude of their peaks, the peak distance. It was shown to improve existent dissimilarity measures in the detection of highly dissimilar symmetric word pairs. We report the existence of reverse complementary word pairs with very dissimilar distance distributions, as well as word pairs with very similar distance distributions even when both distributions are irregular and contain strong peaks.

- The design of new procedures to identify symmetric word pairs with uncommon empirical distance distribution and with clusters of overrepresented short distances. We performed an exhaustive study of these distance distributions and identified words that are strong candidates to the formation of cruciform structures in human DNA.

- The proposal of a new methodology for decomposing the distance distribution of a genomic word into the sum of a baseline distribution and a peak function.

This new understanding could contribute to the advancement of knowledge about DNA sequences. As expected, there are no definitive answers or knowledge produced in such fast evolving fields as Genomics. The answers provided to the research questions proposed herein, allowed for the development of methodologies raising new hypotheses for further studies.

Bibliography

[1] 1000 Genomes Project Consortium and others. An integrated map of genetic variation from 1,092 Human Genomes. *Nature*, 491(7422):56, 2012.

[2] 1000 Genomes Project Consortium and others. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[3] V. Afreixo. *Sinais simbólicos e Aplicações em genómica*. PhD thesis, University of Aveiro, Portugal, 2008.

[4] V. Afreixo, C. A. Bastos, S. P. Garcia, J. M. Rodrigues, A. J. Pinho, and P. J. Ferreira. The breakdown of the word symmetry in the Human Genome. *Journal of Theoretical Biology*, 335:153–159, 2013.

[5] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, and P. J. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23):3064–3070, 2009.

[6] V. Afreixo, C. A. Bastos, J. M. Rodrigues, and R. M. Silva. Identification of DNA CpG islands using inter-dinucleotide distances. In *EURO Mini-conference on Optimization in the Natural Sciences*, pages 162–172. Springer, 2014.

[7] V. Afreixo, P. J. Ferreira, and D. Santos. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, 14(6):523–530, 2004.

[8] V. Afreixo, J. M. Rodrigues, and C. A. Bastos. Analysis of single-strand exceptional word symmetry in the Human Genome: new measures. *Biostatistics*, 16(2):209–221, 2014.

[9] V. Afreixo, J. M. Rodrigues, and C. A. Bastos. Exceptional single strand DNA word symmetry: analysis of evolutionary potentialities. *Journal of Integrative Bioinformatics*, 11(3):48–59, 2014.

[10] V. Afreixo, J. M. Rodrigues, and C. A. Bastos. Analysis of single-strand exceptional word symmetry in the Human Genome: new measures. *Biostatistics*, 16(2):209–221, 2015.

[11] V. Afreixo, J. M. Rodrigues, C. A. Bastos, and R. M. Silva. The exceptional genomic word symmetry along DNA sequences. *BMC Bioinformatics*, 17(1):59, 2016.

[12] V. Afreixo, J. M. Rodrigues, C. A. Bastos, and R. M. Silva. Exceptional symmetry profile: A genomic word analysis. In *10th International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 151–159. Springer, 2016.

[13] V. Afreixo, J. M. O. S. Rodrigues, C. A. C. Bastos, A. H. Tavares, and R. M. Silva. Exceptional symmetry by genomic word. *Interdisciplinary Sciences: Computational Life Sciences*, 9(1):14–23, 2017.

[14] V. Afreixo and A. H. Tavares. Leis que governam a estrutura primária do ADN dos seres vivos. Boletim da SPE, 2015.

[15] F. J. Aherne, N. A. Thacker, and P. I. Rockett. The Bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4):[363]–368, 1998.

[16] J. Aitchison. On criteria for measures of compositional difference. *Mathematical Geology*, 24(4):365–379, 1992.

[17] M. Akhtar, J. Epps, and E. Ambikairajah. Signal processing in sequence analysis: advances in eukaryotic gene prediction. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):310–321, 2008.

[18] B. Alberts, A. Johnson, J. Lewis, P. Walter, M. Raff, and K. Roberts. *Molecular Biology of the Cell 4th Edition: International Student Edition*. Routledge, 2002.

[19] G. Albrecht-Buehler. Inversions and inverted transpositions as the basis for an almost universal ”format” of genome sequences. *Genomics*, 90(3):297–305, 2007.

[20] D. Anastassiou. Genomic signal processing. *IEEE Signal Processing Magazine*, 18(4):8–20, 2001.

[21] A. Annunziato. DNA packaging: nucleosomes and chromatin. *Nature Education*, 1(1):26, 2008.

[22] D. L. Antzoulakos. Waiting times for patterns in a sequence of multistate trials. *Journal of Applied Probability*, 38(2):508–518, 2001.

[23] S. B. Arniker. *Human promoter prediction using DNA numerical representation*. PhD thesis, University of Windsor, 2010.

[24] A. Arribas-Gil and J. Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619, 2014.

[25] P.-F. Baisnée, S. Hampson, and P. Baldi. Why are complementary DNA strands symmetric? *Bioinformatics*, 18(8):1021–1033, 2002.

[26] N. Balakrishnan and M. V. Koutras. *Runs and scans with applications*, volume 764. John Wiley & Sons, 2011.

[27] G. Banfalvi. Structural organization of DNA. *Biochemical Education*, 14(2):50–59, 1986.

[28] V. Barnett, P. Barnett, and T. Lewis. *Outliers in Statistical Data.* Wiley Series in Probability and Statistics. Wiley, 1994.

[29] C. A. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. M. Rodrigues, and P. J. Ferreira. Inter-dinucleotide distances in the Human Genome: an analysis of the whole-genome and protein-coding distributions. *Journal of Integrative Bioinformatics*, 8(3):31–42, 2011.

[30] S. Batzoglou. The many faces of sequence alignment. *Briefings in Bioinformatics*, 6(1):6–22, 2005.

[31] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. Oliver. Study of statistical correlations in DNA sequences. *Gene*, 300(1):105–115, 2002.

[32] D. Bikard, C. Loot, Z. Baharoglu, and D. Mazel. Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiology and Molecular Biology Reviews*, 74(4):570–588, 2010.

[33] G. Blom and D. Thorburn. How many random digits are required until given sequences are obtained? *Journal of Applied Probability*, 19(3):518–531, 1982.

[34] M. L. Bochman, K. Paeschke, and V. A. Zakian. DNA secondary structures: stability and function of G-quadruplex structures. *Nature Reviews Genetics*, 13(11):770, 2012.

[35] M. Borodovsky and J. McIninch. GENMARK: parallel gene recognition for both DNA strands. *Computers & Chemistry*, 17(2):123–133, 1993.

[36] M. Y. Borodovsky and S. Gusein-Zade. A general rule for ranged series of codon frequencies in different genomes. *Journal of Biomolecular Structure and Dynamics*, 6(5):1001–1012, 1989.

[37] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS Letters*, 480(1):17–24, 2000.

[38] B. Brejová, T. Vinar, and M. Li. Pattern discovery. In *Introduction to bioinformatics*, pages 491–521. Springer, 2003.

[39] V. Brendel, J. S. Beckmann, and E. N. Trifonov. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of Biomolecular Structure and Dynamics*, 4(1):11–21, 1986.

[40] R. J. Britten. Transposable element insertions have strongly affected human evolution. *Proceedings of the National Academy of Sciences*, 107(46):19945–19948, 2010.

[41] T. Brown, T. Brown, D. Brown, and L. Brown. *Genomes 3*. Taylor & Francis group, an informa business. Garland Science Pub., 2007.

[42] C. Burge, A. M. Campbell, and S. Karlin. Over-and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences*, 89(4):1358–1362, 1992.

[43] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. Coronado, and J. Oliver. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E*, 79(3):035102, 2009.

[44] S.-H. Cha. Taxonomy of nominal type histogram distance measures. *City*, 1(2):1, 2008.

[45] S.-H. Cha and S. N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6):1355 – 1370, 2002.

[46] Y.-M. Chang. Distribution of waiting time until the rth occurrence of a compound pattern. *Statistics & Probability Letters*, 75(1):29–38, 2005.

[47] B. Charlesworth and N. Barton. Genome size: does bigger mean worse? *Current Biology*, 14(6):R233–R235, 2004.

[48] P. Chaudhuri and S. Das. Statistical analysis of large DNA sequences using distribution of DNA words. *Current Science*, 80(9):1161–1166, 2001.

[49] B. Chowdhury and G. Garai. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 2017.

[50] O. Chrysaphinou and S. Papastavridis. The occurrence of sequence patterns in repeated dependent experiments. *Theory of Probability & its Applications*, 35(1):145–152, 1991.

[51] O. Chryssaphinou and S. Papastavridis. A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials. *Journal of Applied Probability*, 25(2):428–431, 1988.

[52] J. Chung, P. Kannappan, C. Ng, and P. Sahoo. Measures of distance between probability distributions. *Journal of Mathematical Analysis and Applications*, 138(1):280 – 292, 1989.

[53] D. M. Church, V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.-C. Chen, R. Agarwala, W. M. McLaren, G. R. Ritchie, et al. Modernizing reference genome assemblies. *PLoS Biology*, 9(7):e1001091, 2011.

[54] D. M. Church, V. A. Schneider, K. M. Steinberg, M. C. Schatz, A. R. Quinlan, C.-S. Chin, P. A. Kitts, B. Aken, G. T. Marth, M. M. Hoffman, et al. Extending reference assembly models. *Genome Biology*, 16(1):13, 2015.

[55] J. Commins, C. Toft, and M. A. Fares. Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures online*, 11(1):52, 2009.

[56] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.

[57] F. H. Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[58] J. A. Cuesta-Albertos and A. Nieto-Reyes. The random tukey depth. *Computational Statistics & Data Analysis*, 52(11):4979–4988, 2008.

[59] A. Cuevas, M. Febrero, and R. Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496, 2007.

[60] M. de Sousa Vieira. Statistics of DNA sequences: A low-frequency analysis. *Physical Review E*, 60(5):5932, 1999.

[61] S. Deusdado. *Análise e compressão de sequências genómicas.* PhD thesis, University of Minho, Portugal, 2008.

[62] S. Ding, Q. Dai, H. Liu, and T. Wang. A simple feature representation vector for phylogenetic analysis of DNA sequences. *Journal of Theoretical Biology*, 265(4):618–623, 2010.

[63] D. L. Donoho, M. Gasko, et al. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827, 1992.

[64] D. Durand and D. Sankoff. Tests for gene clustering. *Journal of Computational Biology*, 10(3-4):453–482, 2003.

[65] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge university press, 1998.

[66] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress. Multiple evidence strands suggest that there may be as few

as 19000 human protein-coding genes. *Human Molecular Genetics*, 23(22):5866–5878, 2014.

[67] M. Febrero, P. Galeano, and W. González-Manteiga. A functional analysis of nox levels: location and scale estimation and outlier detection. *Computational Statistics*, 22(3):411–427, 2007.

[68] M. Febrero, P. Galeano, and W. González-Manteiga. Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics*, 19(4):331–345, 2008.

[69] R. A. Fisher. *Statistical methods for research workers.* Edinburgh: Oliver and Boyd, 1925.

[70] D. Forsdyke. Relative roles of primary sequence and (G+C)% in euro mini-conferenceng the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. *Journal of Molecular Evolution*, 41(5):573–581, 1995.

[71] D. R. Forsdyke and S. J. Bell. Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. *Applied Bioinformatics*, 3(1):3–8, 2004.

[72] R. Fraiman and G. Muniz. Trimmed means for functional data. *Test*, 10(2):419–440, 2001.

[73] J. Fu and M. Koutras. Distribution theory of runs: a markov chain approach. *Journal of the American Statistical Association*, 89(427):1050–1058, 1994.

[74] J. C. Fu. Reliability of consecutive-k-out-of-n: F systems with (k-1)-step markov dependence. *IEEE Transactions on Reliability*, 35(5):602–606, 1986.

[75] J. C. Fu. Poisson convergence in reliability of a large linearly connected system as related to coin tossing. *Statistica Sinica*, pages 261–275, 1993.

[76] J. C. Fu. Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica*, pages 957–974, 1996.

[77] J. C. Fu. Distribution of the scan statistic for a sequence of bistate trials. *Journal of Applied Probability*, 38(4):908–916, 2001.

[78] J. C. Fu and Y. Chang. On probability generating functions for waiting time distributions of compound patterns in a sequence of multistate trials. *Journal of Applied Probability*, 39(1):70–80, 2002.

[79] J. C. Fu and W. W. Lou. *Distribution theory of runs and patterns and its applications: a finite Markov chain imbedding approach*. World Scientific, 2003.

[80] A. Fukushima, M. Kinouchi, S. Kanaya, Y. Kudo, and T. Ikemura. Statistical analysis of genomic information. *Genome Informatics*, 11:315–316, 2000.

[81] F. Galton. Iv. statistics by intercomparison, with remarks on the law of frequency of error. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 49(322):33–46, 1875.

[82] P. Garagnani, C. Pirazzini, C. Giuliani, M. Candela, P. Brigidi, F. Sevini, D. Luiselli, M. G. Bacalini, S. Salvioli, M. Capri, et al. The three genetics (nuclear DNA, mitochondrial DNA, and gut microbiome) of longevity in humans considered as metaorganisms. *BioMed Research International*, 2014, 2014.

[83] Genome Reference Consortium. http://www.ncbi.nlm.nih.gov/grc.

[84] H. U. Gerber and S.-Y. R. Li. The occurrence of sequence patterns in repeated experiments and hitting times in a markov chain. 1981.

[85] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-ENCODE? history and updated definition. *Genome Research*, 17(6):669–681, 2007.

[86] M. X. Geske, A. P. Godbole, A. A. Schaffner, A. M. Skolnick, and G. L. Wallstrom. Compound poisson approximations for word patterns under markovian hypotheses. *Journal of Applied Probability*, 32(4):877–892, 1995.

[87] J. Glaz, M. Kulldorff, V. Pozdnyakov, and J. M. Steele. Gambling teams and waiting times for patterns in two-state markov chains. *Journal of Applied Probability*, 43(1):127–140, 2006.

[88] A. P. Godbole. Poisson approximations for runs and patterns of rare events. *Advances in Applied Probability*, 23(4):851–865, 1991.

[89] A. P. Godbole and S. G. Papastavridis. *Runs and Patterns in Probability: Selected Papers: Selected Papers*, volume 283. Springer Science & Business Media, 1994.

[90] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley. Species independence of mutual information in coding and noncoding DNA. *Physical Review E*, 61(5):5624, 2000.

[91] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[92] L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, Series A*, 30(2):183–208, 1981.

[93] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, and J. L. Oliver. Cpgcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, 7(1):446, 2006.

[94] A. Hart, S. Martínez, and F. Olmos. A gibbs approach to chargaff's second parity rule. *Journal of Statistical Physics*, 146(2):408–422, 2012.

[95] S. Harteis and S. Schneider. Making the bend: DNA tertiary structure and protein-DNA interactions. *International Journal of Molecular Sciences*, 15(7):12335–12363, 2014.

[96] A. Heger and L. Holm. Towards a covering set of protein family profiles. *Progress in Biophysics and Molecular Biology*, 73(5):321–337, 2000.

[97] H. Herzel and I. Große. Correlations in DNA sequences: The role of protein coding segments. *Physical Review E*, 55(1):800, 1997.

[98] H. Herzel, E. Trifonov, O. Weiss, and I. Grosse. Interpreting correlations in biosequences. *Physica A: Statistical Mechanics and its Applications*, 249(1-4):449–459, 1998.

[99] D. Holste, I. Grosse, S. Beirer, P. Schieg, and H. Herzel. Repeats and correlations in human DNA sequences. *Physical Review E*, 67(6):061913, 2003.

[100] M. Hubert, P. J. Rousseeuw, and P. Segaert. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–246, 2015. (with discussion).

[101] R. J. Hyndman and H. L. Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45, 2010.

[102] International Human Genome Sequencing Consortium and others. Initial sequencing and analysis of the Human Genome. *Nature*, 409(6822):860, 2001.

[103] International Human Genome Sequencing Consortium and others. Finishing the euchromatic sequence of the Human Genome. *Nature*, 431(7011):931, 2004.

[104] M. Jäger, M. Schubach, T. Zemojtel, K. Reinert, D. M. Church, and P. N. Robinson. Alternate-locus aware variant calling in whole genome sequencing. *Genome Medicine*, 8(1):130, 2016.

[105] S. Karlin and V. Brendel. Patchiness and correlations in DNA sequences. *Science*, 259(5095):677–680, 1993.

[106] M. Kaushik, S. Kaushik, K. Roy, A. Singh, S. Mahendru, M. Kumar, S. Chaudhary, S. Ahmed, and S. Kukreti. A bouquet of DNA structures: Emerging diversity. *Biochemistry and Biophysics Reports*, 5:388 – 395, 2016.

[107] B. Kenidra, M. Benmohammed, A. Beghriche, and Z. Benmounah. A partitional approach for genomic-data clustering combined with k-means algorithm. In *Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on*, pages 114–121. IEEE, 2016.

[108] D. A. Kleinjan and V. van Heyningen. Long-range control of gene expression: emerging mechanisms and disruption in disease. *The American Journal of Human Genetics*, 76(1):8–32, 2005.

[109] A. K. Konopka, C. Martindale, et al. Noncoding DNA, Zipf's law, and language. *Science-AAAS-Weekly Paper Edition*, 268(5212):785–790, 1995.

[110] F. Kouzine, D. Wojtowicz, L. Baranello, A. Yamane, S. Nelson, W. Resch, K.-R. Kieffer-Kwon, C. J. Benham, R. Casellas, T. M. Przytycka, et al. Permanganate/S1 nuclease footprinting reveals non-B DNAstructures with regulatory potential across a mammalian genome. *Cell Systems*, 4(3):344–356, 2017.

[111] A. Krishnamachari, V. moy Mandal, et al. Study of DNA binding sites using the rényi parametric entropy measure. *Journal of Theoretical Biology*, 227(3):429–436, 2004.

[112] D. Kugiumtzis and A. Provata. Statistical analysis of gene and intergenic DNA sequences. *Physica A: Statistical Mechanics and its Applications*, 342(3-4):623–638, 2004.

[113] H. K. Kwan and S. B. Arniker. Numerical representation of DNA sequences. In *Electro/Information Technology, 2009. eit'09. IEEE International Conference on*, pages 307–310. IEEE, 2009.

[114] J. K. Lanctot. Estimating DNA sequence entropy. Citeseer.

[115] E. S. Lander and R. A. Weinberg. Journey to the center of biology. *Science*, 287(5459):1777–1782, 2000.

[116] M.-Y. Leung, G. M. Marsh, and T. P. Speed. Over-and underrepresentation of short DNA words in herpesvirus genomes. *Journal of Computational Biology*, 3(3):345–360, 1996.

[117] E. Levina and P. Bickel. The earth mover's distance is the Mallows distance: Some insights from statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 251–256. IEEE, 2001.

[118] S.-Y. R. Li. A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *The Annals of Probability*, pages 1171–1176, 1980.

[119] W. Li. The study of correlation structures of DNA sequences: a critical review. *Computers & Chemistry*, 21(4):257–271, 1997.

[120] W. Li, G. Stolovitzky, P. Bernaola-Galván, and J. L. Oliver. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Research*, 8(9):916–928, 1998.

[121] M. Lothaire. *Applied combinatorics on words*, volume 105. Cambridge University Press, 2005.

[122] N. M. Luscombe, D. Greenbaum, and M. Gerstein. What is bioinformatics? a proposed definition and overview of the field. *Methods of Information in Medicine*, 40(04):346–358, 2001.

[123] J. Mandel. *The statistical analysis of experimental data.* Courier Corporation, 2012.

[124] R. Mantegna, S. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. Stanley. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical Review E*, 52(3):2939, 1995.

[125] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley. Linguistic features of noncoding DNA sequences. *Physical Review Letters*, 73(23):3169, 1994.

[126] L. Marino-Ramírez, J. L. Spouge, G. C. Kanga, and D. Landsman. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Research*, 32(3):949–958, 2004.

[127] C. Martindale and A. K. Konopka. Oligonucleotide frequencies in DNA follow a Yule distribution. *Computers & Chemistry*, 20(1):35–38, 1996.

[128] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. Measures of difference for compositional data and hierarchical clustering methods. In A. Buccianti, G. Nardi, and R. Potenza, editors, *Proceedings of International Association for Mathematical Geosciences (IAMG)*, 1998.

[129] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the Human Genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

[130] C. Matek, T. E. Ouldridge, J. P. Doye, and A. A. Louis. Studying molecular physiology of DNA cruciforms with a coarse-grained computational model, 2013.

[131] J. P. McCutcheon, B. R. McDonald, and N. A. Moran. Origin of an alternative genetic code in the extremely small and gc–rich genome of a bacterial symbiont. *PLoS Genetics*, 5(7):e1000565, 2009.

[132] D. Mitchell and R. Bridge. A test of chargaff's second rule. *Biochemical and Biophysical Research Communications*, 340(1):90–94, 2006.

[133] H. Moghaddasi, K. Khalifeh, and A. H. Darooneh. Distinguishing functional DNA words; a method for measuring clustering levels. *Scientific Reports*, 7:41543, 2017.

[134] K. Mosler. Depth statistics. In *Robustness and complex data structures*, pages 17–34. Springer, 2013.

[135] A. S. S. Nair and T. Mahalakshmi. Gsp using bi-nucleotide distance signals. In *13th International Conference on Advanced Computing and Communications*, 2005.

[136] A. S. S. Nair and T. Mahalakshmi. Visualization of genomic data using inter-nucleotide distance signals. *Proceedings of IEEE Genomic Signal Processing*, 408, 2005.

[137] M. R. Neto. O conceito de profundidade em estatística, 2008.

[138] C. Nikolaou and Y. Almirantis. Deviations from Chargaff's second parity rule in organellar DNA: Insights into the evolution of organellar genomes. *Gene*, 381:34–41, 2006.

[139] P. Niyogi and R. C. Berwick. A note on Zipf's law, Natural Languages, and noncoding DNA regions. *Technical ReportMemo 118, M.I.T. CBCL. In the Computation and Language Archive.*, 1995.

[140] G. Nuel. Numerical solutions for patterns statistics on markov chains. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 2006.

[141] G. Nuel, L. Regad, J. Martin, and A.-C. Camproux. Exact distribution of a pattern in a set of random sequences generated by a markov source: applications to biological data. *Algorithms for Molecular Biology*, 5(1):15, 2010.

[142] J. L. Oliver and A. Marín. A relationship between GC content and coding-sequence length. *Journal of Molecular Evolution*, 43(3):216–223, 1996.

[143] O. Pele. *Distance functions: Theory, algorithms and applications.* Citeseer, 2011.

[144] J. Pellicer, M. F. Fay, and I. J. Leitch. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 164(1):10–15, 2010.

[145] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168, 1992.

[146] P. A. Pevzner, M. Y. Borodovsky, and A. A. Mironov. Linguistics of nucleotide sequences i: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *Journal of Biomolecular Structure and Dynamics*, 6(5):1013–1026, 1989.

[147] F. Piazza and P. Lio. Statistical analysis of simple repeats in the Human Genome. *Physica A: Statistical Mechanics and its Applications*, 347:472–488, 2005.

[148] S. Pietrokovski and E. N. Trifonov. Imported sequences in the mitochondrial yeast genome identified by nucleotide linguistics. *Gene*, 122(1):129–137, 1992.

[149] J. L. Plank and A. Dean. Enhancer function: mechanistic and genome-wide insights come together. *Molecular Cell*, 55(1):5–14, 2014.

[150] B. Powdel, S. S. Satapathy, A. Kumar, P. K. Jha, A. K. Buragohain, M. Borah, and S. K. Ray. A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's second parity rule). *DNA Research*, 16(6):325–343, 2009.

[151] V. Pozdnyakov. On occurrence of patterns in markov chains: Method of gambling teams. *Statistics & Probability Letters*, 78(16):2762–2767, 2008.

[152] V. V. Prabhu. Symmetry observations in long nucleotide sequences. *Nucleic Acids Research*, 21(12):2797, 1993.

[153] L. Pray. Discovery of DNA structure and function: Watson and Crick. *Nature Education*, 1(1):100, 2008.

[154] A. Provata and T. Oikonomou. Power law exponents characterizing human DNA. *Physical Review E*, 75(5):056102, 2007.

[155] B. Prum, F. Rodolphe, and É. de Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 205–220, 1995.

[156] D. Qi and A. J. Cuticchia. Compositional symmetries in complete genomes. *Bioinformatics*, 17(6):557–559, 2001.

[157] J. Qi, B. Wang, and B.-I. Hao. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of Molecular Evolution*, 58(1):1–11, 2004.

[158] D. W. Reed. A statistical approach to quantitative linguistic analysis. *Word*, 5(3):235–247, 1949.

[159] M. Régnier. A unified approach to word occurrence probabilities. *Discrete Applied Mathematics*, 104(1-3):259–280, 2000.

[160] G. Reinert and S. Schbath. Compound poisson and poisson process approximations for occurrences of multiple words in markov chains. *Journal of Computational Biology*, 5(2):223–253, 1998.

[161] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, 7(1-2):1–46, 2000.

[162] S. Robin and J.-J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability*, 36(1):179–193, 1999.

[163] S. Robin and J.-J. Daudin. Exact distribution of the distances between any occurrences of a set of words. *Annals of the Institute of Statistical Mathematics*, 53(4):895–905, 2001.

[164] S. Robin, J.-J. Daudin, H. Richard, M.-F. Sagot, and S. Schbath. Occurrence probability of structured motifs in random sequences. *Journal of Computational Biology*, 9(6):761–773, 2002.

[165] S. Robin, S. Robin, F. Rodolphe, and S. Schbath. *DNA, words and models: statistics of exceptional words*. Cambridge University Press, 2005.

[166] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig. The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248, 2009.

[167] P. J. Rousseeuw, J. Raymaekers, and M. Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, pages 1–15, 2018.

[168] A. Roy, C. Raychaudhury, and A. Nandy. Novel techniques of graphical representation and analysis of DNA sequences – a review. *Journal of Biosciences*, 23(1):55–71, 1998.

[169] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[170] R. Rudner, J. D. Karkas, and E. Chargaff. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proceedings of the National Academy of Sciences*, 60(3):921–922, 1968.

[171] G. K. Sandve and F. Drabløs. A survey of motif discovery methods in an integrated framework. *Biology direct*, 1(1):11, 2006.

[172] M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder. Linking disease associations with regulatory information in the Human Genome. *Genome Research*, 22(9):1748–1759, 2012.

[173] S. Schbath. Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics*, 1:1–16, 1997.

[174] S. Shporer, B. Chor, S. Rosset, and D. Horn. Inversion symmetry of DNA k-mer counts: validity and deviations. *BMC Genomics*, 17(1):696, 2016.

[175] H. S. Sichel. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547, 1975.

[176] B. Silverman and R. Linsker. A measure of DNA periodicity. *Journal of Theoretical Biology*, 118(3):295–300, 1986.

[177] R. R. Sinden. *DNA structure and function*. Elsevier, 2012.

[178] W. Stahel. *Robust estimation: Infinitesimal optimality and covariance matrix estimators (in German)*. PhD thesis, ETH, Zuerich, 1981.

[179] H. Stanley, S. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng, and M. Simons. Scaling features of noncoding DNA. *Physica A: Statistical Mechanics and its Applications*, 273(1-2):1–18, 1999.

[180] V. Stefanov and A. G. Pakes. Explicit distributional results in pattern formation. *The Annals of Applied Probability*, pages 666–678, 1997.

[181] V. T. Stefanov. On some waiting time problems. *Journal of Applied Probability*, 37(3):756–764, 2000.

[182] V. T. Stefanov. The intersite distances between pattern occurrences in strings generated by general discrete-and continuous-time models: an algorithmic approach. *Journal of Applied Probability*, 40(4):881–892, 2003.

[183] V. T. Stefanov and W. Szpankowski. Waiting time distributions for pattern occurrence in a constrained sequence. *Discrete Mathematics and Theoretical Computer Science*, 9(1):305–320, 2007.

[184] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[185] Y. Sun and M. G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011.

[186] Y. Tambovtsev and C. Martindale. Phoneme frequencies follow a Yule distribution.

[187] A. H. Tavares, V. Afreixo, J. M. O. S. Rodrigues, and C. A. C. Bastos. The symmetry of oligonucleotide distance distributions in the human genome. In *Proceedings of the 4th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2015*, pages 256–263, Lisbon, Portugal, 2015.

[188] A. H. Tavares, J. Raymaekers, P. J. Rousseeuw, R. M. Silva, C. A. C. Bastos, A. J. Pinho, P. Brito, and V. Afreixo. Dissimilar symmetric word pairs in the human genome. In *11th International Conference on Practical Applications of Computational Biology & Bioinformatics, PACBB 2017, Porto, Portugal, 21-23 June, 2017*, volume 616 of *Advances in Intelligent Systems and Computing*, pages 248–256. Springer, 2017.

[189] A. H. Tavares, J. Raymaekers, P. J. Rousseeuw, R. M. Silva, C. A. C. Bastos, A. J. Pinho, P. Brito, and V. Afreixo. Comparing reverse complementary genomic words based on their distance distributions and frequencies. *Interdisciplinary Sciences: Computational Life Sciences*, 10(1):1–11, 2018.

[190] A. H. Tavares, J. M. O. S. Rodrigues, C. A. C. Bastos, A. Pinho, P. Ferreira, P. Brito, and V. Afreixo. Detection of exceptional genomic words: a comparison between species. In *Proceedings of the 22nd International Conference on Computational Statistics, COMPSTAT 2016*, pages 255–264, Oviedo, Spain, 2016.

[191] A. H. M. P. Tavares, V. Afreixo, P. Brito, and P. Filzmoser. Directional outlyingness applied to distances between genomic words. In *Proceedings of the 22nd edition of the Portuguese Conference on Pattern Recognition, RECPAD 2016*, pages 108–110, Aveiro, Portugal, 2016.

[192] A. H. M. P. Tavares, A. J. Pinho, R. M. Silva, J. M. O. S. Rodrigues, C. A. C. Bastos, P. J. S. G. Ferreira, and V. Afreixo. DNA word analysis based on the distribution of the distances between symmetric words. *Scientific Reports*, 7(1), apr 2017.

[193] R. W. Taylor and D. M. Turnbull. Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics*, 6(5):389, 2005.

[194] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36, 2012.

[195] E. N. Trifonov. The multiple codes of nucleotide sequences. *Bulletin of Mathematical Biology*, 51(4):417–432, 1989.

[196] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis. Periodicity in DNA coding sequences: implications in gene evolution. *Journal of Theoretical Biology*, 151(3):323–331, 1991.

[197] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis. Is DNA a language? *Journal of Theoretical Biology*, 184(1):25–29, 1997.

[198] J. W. Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

[199] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the Human Genome. *Science*, 291(5507):1304–1351, 2001.

[200] R. Verde and A. Irpino. Comparing histogram data using a Mahalanobis–Wasserstein distance. In P. Brito, editor, *COMPSTAT 2008*, pages 77–89. Physica-Verlag HD, 2008.

[201] R. Verde, A. Irpino, and A. Balzanella. Dimension reduction techniques for distributional symbolic data. *IEEE Transactions on Cybernetics*, 46(2):344–355, 2016.

[202] S. Vinga and J. Almeida. Alignment-free sequence comparison—a review. *Bioinformatics*, 19(4):513–523, 2003.

[203] R. F. Voss. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters*, 68(25):3805, 1992.

[204] M. S. Waterman. *Introduction to computational biology: maps, sequences and genomes.* CRC Press, 1995.

[205] J. D. Watson. The secret of life. *New York: Alfred Knopf*, 2003.

[206] J. D. Watson, F. H. Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[207] S.-J. Wei, M. Shi, X.-X. Chen, M. J. Sharkey, C. van Achterberg, G.-Y. Ye, and J.-H. He. New views on strand asymmetry in insect mitochondrial genomes. *PLoS One*, 5(9):e12708, 2010.

[208] C. Yin and J. Wang. Periodic power spectrum with applications in detection of latent periodicities in DNA sequences. *Journal of Mathematical Biology*, 73(5):1053–1079, 2016.

[209] N. Yu, Z. Li, and Z. Yu. Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning. *Big Data Mining and Analytics*, 1(3):191–210, 2018.

[210] G. U. Yule et al. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis. *Philosophical Transactions of the Royal Society of London, Serie B*, 213(402-410):21–87, 1925.

[211] R. Zhang and C.-T. Zhang. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. *Journal of Biomolecular Structure and Dynamics*, 11(4):767–782, 1994.

[212] S.-H. Zhang. Persistence and breakdown of strand symmetry in the Human Genome. *Journal of Theoretical Biology*, 370:202–204, 2015.

[213] J. Zhao, A. Bacolla, G. Wang, and K. M. Vasquez. Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences*, 67(1):43–62, 2010.

[214] G. K. Zipf. Selected studies of the principle of relative frequency in language. 1932.

[215] G. K. Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[216] Y. Zuo and R. Serfling. General notions of statistical depth function. *Annals of Statistics*, pages 461–482, 2000.

Intentionally blank page.

Appendix

Intentionally blank page.

List of Publications

Papers in International Scientific Periodicals (with referees)

2018 A. H. Tavares, J. Raymaekers, P. J. Rousseeuw, R. M. Silva, C. A. C. Bastos, A. J. Pinho, P. Brito, and V. Afreixo. Comparing reverse complementary genomic words based on their distance distributions and frequencies. *Interdisciplinary Sciences: Computational Life Sciences*, 10(1):1-11, 2018. [189]

2017 A. H. M. P. Tavares, A. J. Pinho, R. M. Silva, J. M. O. S. Rodrigues, C. A. C. Bastos, P. J. S. G. Ferreira, and V. Afreixo. DNA word analysis based on the distribution of the distances between symmetric words. *Scientific Reports*, 7(1), Apr 2017. [192]

2017 V. Afreixo, J. M. O. S. Rodrigues, C. A. C. Bastos, A. H. Tavares, and R. M. Silva. Exceptional symmetry by genomic word. *Interdisciplinary Sciences: Computational Life Sciences*, 9(1):14?23, 2017. [13]

Papers in International Conferences Proceedings (with referees)

2017 A. H. Tavares, J. Raymaekers, P. J. Rousseeuw, R. M. Silva, C. A. C. Bastos, A. J. Pinho, P. Brito, and V. Afreixo. Dissimilar symmetric word pairs in the human genome. In *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, PACBB 2017, Porto, Portugal, 21-23 June, 2017, volume 616 of Advances in Intelligent Systems and Computing, pages 248-256. Springer, 2017. [188]

2016 A. H. Tavares, J. M. O. S. Rodrigues, C. A. C. Bastos, A. Pinho, P. Ferreira, P. Brito, and V. Afreixo. Detection of exceptional genomic words: a comparison between species. In *Proceedings of the 22nd International Conference on Computational Statistics*, COMPSTAT 2016, pages 255-264, Oviedo, Spain, 2016. [190]

2015 A. H. Tavares, V. Afreixo, J. M. O. S. Rodrigues, and C. A. C. Bastos. The symmetry of oligonucleotide distance distributions in the human genome. In *Proceedings of the 4th International Conference on Pattern Recognition Applications and Methods*, ICPRAM 2015, pages 256-263, Lisbon, Portugal, 2015. [187]

Papers in National Conferences Proceedings (with referees)

2016 A. H. M. P. Tavares, V. Afreixo, P. Brito, and P. Filzmoser. Directional outlyingness applied to distances between genomic words. In *Proceedings of the 22nd edition of the Portuguese Conference on Pattern Recognition*, RECPAD 2016, pages 108-110, Aveiro, Portugal, 2016. [191]

Other publications

2015 V. Afreixo and A. H. Tavares. Leis que governam a estrutura primária do ADN dos seres vivos. Boletim da SPE, 2015. [14]

List of Communications

Oral Communications in International Conferences

2018 III Encontro Luso-Galaico de Biometria, EBio2018. *Deteção de grupos de observações atípicas: uma aplicação em dados genómicos.* Ana Helena Tavares, Vera Afreixo, and Paula Brito. Aveiro, Portugal. June, 2018.
(Best Young Scientist oral presentation)

2017 1st Conference on Data Science, Statistics & Visualisation 2017, DSSV. *Clustering DNA words through distance distributions.* Ana Helena Tavares, Vera Afreixo, and Paula Brito. Lisbon, Portugal. July, 2017.

2017 11th International Conference on Practical Applications of Computational Biology & Bioinformatics, PACBB 2017. *Dissimilar symmetric word pairs in the human genome.* Ana Helena Tavares, Jakob Raymaekers, Peter Rousseeuw, Raquel M. Silva, Carlos A. C. Bastos, Armando Pinho, Paula Brito, and Vera Afreixo. Porto, Portugal. June, 2017.

2017 Fourth Workshop on Molecular Logic. *Detection of outlying distributions in DNA word sequences.* Ana Helena Tavares, Vera Afreixo, and Paula Brito. Aveiro, Portugal. June, 2017.

2016 22nd International Conference on Computational Statistics, COMPSTAT 2016. *Detection of exceptional genomic words: a comparison between species.* Ana Helena Tavares, João M. O. S. Rodrigues, Carlos A. C. Bastos, Armando Pinho, Paulo Ferreira, Paula Brito, and Vera Afreixo. Oviedo, Spain. August, 2016.

Oral Communications in National Conferences

2018 XXV Jornadas de Classificação e Análise de Dados, JOCLAD 2018. *Classificação de distribuições de distâncias genómicas.* Ana Helena Tavares, Vera Afreixo, and Paula Brito. Almada, Portugal. April, 2018.

2017 XXIII Congresso Anual da Sociedade Portuguesa de Estatística. *Classificação de distribuições no contexto de distâncias entre palavras genómicas.* Ana Helena Tavares, Vera Afreixo and Paula Brito. Lisboa, Portugal. October, 2017.

2017 XXIV Jornadas de Classificação e Análise de Dados, JOCLAD 2017. *Deteção de distribuições atípicas em sequências genómicas.* Ana Helena Tavares, Vera Afreixo, and Paula Brito. Porto, Portugal. April, 2017.

2015 XXII Congresso Anual da Sociedade Portuguesa de Estatística. *Palavras genómicas com distribuição excecional.* Ana Helena Tavares and Vera Afreixo. Olhão, Portugal. October, 2015.

Poster Communications

2017 Ciência 2017 - Encontro com a Ciência e Tecnologia em Portugal (Science and Technology in Portugal Summit). *Analysis of genomic words distance distributions.* Ana Helena Tavares, Vera Afreixo, and Paula Brito. Lisbon, Portugal. July, 2017.

2017 Research Day 2017 *Distance distributions between words: a mathematical descriptor of DNA sequences.* Ana Helena Tavares, Vera Afreixo, and Paula Brito. Aveiro, Portugal. June, 2017.

2016 22nd edition of the Portuguese Conference on Pattern Recognition, RECPAD 2016 *Directional outlyingness applied to distances between genomic words.* Ana Helena M. P. Tavares, Vera Afreixo, Paula Brito, and Peter Filzmoser. Aveiro, Portugal. October, 2016.

2015 4th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2015. *The symmetry of oligonucleotide distance distributions in the human genome.* Ana Helena Tavares, Vera Afreixo, João M. O. S. Rodrigues, and Carlos A. C. Bastos. Lisbon, Portugal. January, 2015.

Invited Seminars

2018 *Oligonucleotide clustering through distance distributions.* Sixth meeting of the Center for Research & Development in Mathematics and Applications (CIDMA). June 11, 2018. Department of Mathematics. University of Aveiro. Portugal.

2017 *Deteção de distribuições atípicas em sequências de ADN.* March 8, 2017. Department of Mathematics. University of Aveiro. Portugal.

2017 *Detection of exceptional genomic words.* January 20, 2017. Department of Mathematics. University of Leuven, Belgium.