1

What Is Biology?

Life is a wondrous phenomenon to anyone who experiences, observes, or contemplates it. The science of *biology* aims to encompass and understand the phenomenon of life on Earth, not only as it exists and could be observed today but, perhaps more importantly, also as it has existed in the Earth's past, simply because past life may provide insights to help understand what it is today and what it will become in the future. Furthermore, it includes the much harder "cosmic" question of its origin. *How does life that occurs in ephemeral time for units of life (e.g. a living cell or a mayfly for one day) manage to persist as long as billions of years (even in ensembles such as species), despite ever-changing conditions (sometimes radically) for organisms and their environment?* Given the physical nature of time in limited human minds, questions about the past are difficult; questions about the future are even harder. In the absence of a time machine that would allow travel into the past to ascertain what actually happened, *what methods and tools has biology used or could use to attempt to provide answers about the past? How can it anticipate the range of possible outcomes through predictive approaches?* Some specific questions of interest are:

- How did life come to exist? What are its origins? How did the most basic unit of life, a living cell, come about? Is life possible on other planets? If so, can they evolve? If so, what is their mode of "evolution"?
- What patterns shape *macroevolution*, the process that explains the diversity of life forms? What leads to the origins of major groups (taxa)? What is the cause of mass extinctions? Can these patterns be used to predict evolutionary changes?
- What processes lead to the formation of new species (*speciation*)? What mechanisms shape *microevolution* in general?
- How do complex traits such as behavior and culture evolve? Is there a major "force" leading to a trait? Is it just the synthesized expression of multigenic effects combined with environmental factors?
- Is the microbiome's effect on human health, disease, and the environment larger or smaller than suggested to date?

Where to begin to address these questions was not ideal (it never is in science), but a lot of knowledge about them has been gained. Progress in every science usually depends on progress in others that can be used as new tools (like telescopes in physics and microarrays in biotechnology). That knowledge can be used to rethink and refine approaches to the questions for better answers, as will be further illustrated in every chapter of this book.

Predictive Methods for Genomics and Evolution, First Edition. Max Garzon, Luis F. Garcia and Alexander Colorado.

As it happens in most sciences, biology has started with a globally observable macrophenomenon that is evident around us, (living) organisms. *How can these organisms exist, i.e. function in and interact with others and with their environments?* (Other objects like rocks and planets do not behave in the same dynamic fashion on the same scales of time). Progress in science is made by attempting to explain the reasons and causal chains why this phenomenology occurs as it does, from more fundamental forces in the universe. Physics has demonstrated that most phenomena (physical and even metaphysical) are a consequence of a few far more primitive and fundamental forces, namely global gravitational and more local electric and magnetic forces. However, the gap between them and biological phenomena has remained wide because of the complexity of the interactions and their cumulative effect over space and time. Developments in biology (e.g. genetics and systematics) and in other sciences as well (data science, machine learning) have now made it possible to tackle this problem, as is already evident in the field of bioinformatics.

Fortunately for biology, there is a pervasive feature across all organisms (living or extinct), both in space and time: *deoxyribonucleic acid* (DNA). It is known now that its structure has remained essentially unchanged for billions of years (despite drastic changes in the physical conditions on Earth), resulting in significant changes in the existing biodiversity it supports. DNA enables replication (the ability to produce an identical copy of itself), the (re)generation of the entire gene expression machinery, and its self-regulation in a complex organization of life at different levels, including unicellular to multicellular organisms, generation after generation. However, different levels of complexity in the molecule play a fundamental role in the structure and function of different biological groups. *Therefore, a deep examination of the nature and complexity of DNA probably offers a novel and the best chance to make inroads into these questions. That is the approach taken in this book to the science of biology.* (Answers of a divine origin are a logical possibility and, although briefly discussed in Chapter 9, fall outside the scope of this book. A more extensive discussion of the relationship between science and religion can be found in Runehow et al. (2013).)

The goal of this first chapter is to describe the nature and the role of DNA as responsible for life's continuity and the transmission of genetic information across generations. This role is fundamental to the survival and persistence of all known organisms and forms of life. A major objective is to describe the local physical interactions among basic biomolecules that make this role possible and give rise to life. In particular, recent results about its structure (the so-called deep structure of DNA) enable the extraction of the information contained in DNA that can eventually be leveraged to *make fairly accurate predictions at different biological levels about the past and the future.* Third, the relatively straightforward novel ways to sequence genomic DNA (e.g. *next-generation sequencing* – NGS) have recently enabled biologists, as never before, to explore further questions about how species evolve and investigate the evolutionary relationships among organisms as more species are described and others have become extinct. Thus, traditional methods in biology have used comparison through so-called *sequence alignments* to make inferences about major questions in biology (e.g. to define species and to infer phylogenetic relationships through the past and the origin of evolutionary novelties, among others). However, they do not come without limitations. Therefore, this chapter also reviews available alternative alignment-free methods that can address the same questions but may offer better or different answers to these major and intriguing questions. Both approaches offer their own advantages and disadvantages, so they must be explored comparatively throughout the book.

1.1 DNA: Nature, Role, and Function

Life has existed and persisted on a changing planet for about 4.5 billion years. It is diverse and can be found in a variety of conditions (e.g. extreme hot or cold, light or dark, basic or acidic environments). *How does the complex organization of life work to bring about the variations that exist for all kinds of organisms to survive in these environments?* Several major landmarks in evolution have allowed these increased levels of complexity in the organization of life (see Figure 1.1), including:

- Self-replication that allows the genetic material, with the aid of other molecules, to produce a (nearly) identical copy of itself.
- A nuclear membrane that holds the genetic material inside it and isolates it from the rest of the cell. Many organisms on the planet are single-celled, but the formation of tissues (ensembles of cells) makes possible the existence of multicellular organisms.
- Tissues that make up organs, which in turn organize themselves into organ systems and give rise to complex organisms.
- Organisms that interact and have adaptations to organize into populations (individuals of the same species living in the same area), communities (populations of different species living in the same area), and ecosystems (communities of organisms in a given environment, with both living and nonliving elements).

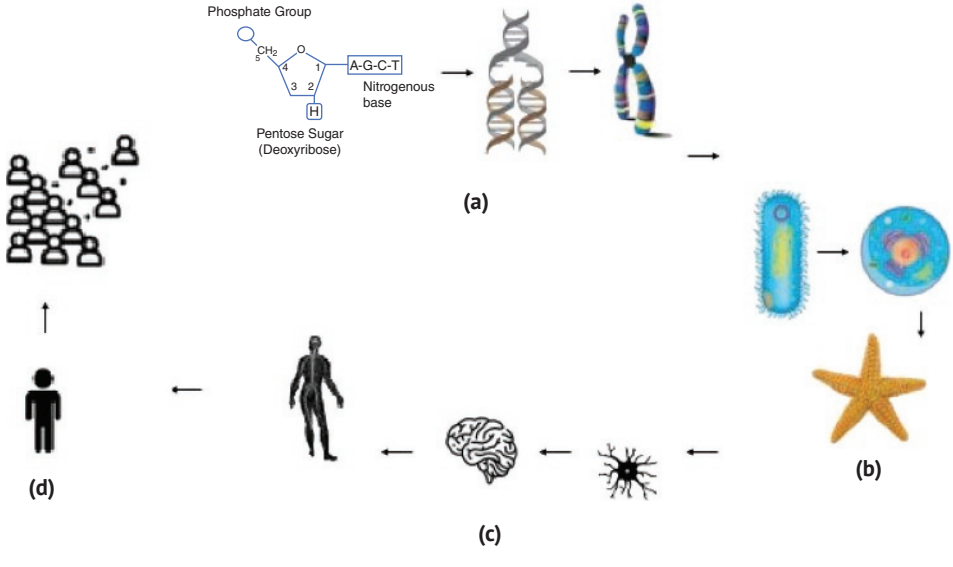

Figure 1.1 Life exhibits a very complex organization. Different disciplines in biology approach biodiversity at various scales in a hierarchical way. (a) RNA and DNA, the first self-replicating molecules, and DNA condensation. (b) Combinations of cells may form multicellular individual organisms (at first with no nuclear membrane), such as membranes and tissues. (c) Tissues make up organs and systems, and organs form organisms. (d) Organisms self-organize into populations, communities, and ecosystems.

The goal of this section is to summarize basic biological knowledge about DNA, the most important molecule that every single organism contains in abundance and is responsible for the maintenance of life and the transmission of genetic information to offspring through generations.

The self-replicating property of DNA is practically identical in all organisms, although the complexity of how it occurs may vary, particularly among divergent groups. Without these properties, life is unlikely to exist. *What are the processes and mechanisms that bring about these properties?*

DNA goes through *replication*, only certain regions with particular functions (called *genes*) undergo a process of *transcription* to a similar, shorter-lived molecule, RNA that acts as a messenger for the *translation* of the original DNA into its expression, a *protein*. *Gene regulation* controls the timing at which genes are expressed, the location in a given cell where the process takes place, and the amounts of protein produced. This way, organisms self-regulate with complex mechanisms that biologists are still trying to understand at various levels of organization, such as the maintenance of energy and metabolism, the response to stimuli, the strategies for reproduction and development, and how organisms adapt to changing environmental and ecological conditions.

1.1.1 Nucleotides, Ligation, and Hybridization

The goal of this section is to describe essential *local interactions at molecular levels* within a living cell that maintain an organism's functioning in an environment. It is a well-established fact that DNA encodes for most of the physically evident (phenotypic) features of an organism. What may not be well known is that, furthermore, recent works demonstrate that DNA sequences encode enough information about an organism so that features about phenotype, taxonomic group, environmental conditions of the natural habitat where an organism lived and so on, could be predicted to a large extent (Mainali et al. 2020a,b) from this information, somehow contained and hidden in its DNA and expressed through these local interactions. Some of these implications will be described in the following Section 1.2. They require a precise description of the essential facts about DNA's physical chemistry to enable a more careful analysis of more pervasive properties in interaction with other DNA molecules in later sections and chapters.

DNA itself was discovered/identified (nearly 90years before the discovery of its precise structure) by the Swiss chemist (Miescher 1871), who referred to it as *nuclein*, at a time proteins were thought to be the primary carriers of heredity's information due to their wide variability, consistent with the great diversity of life. It took nearly a century for biologists to get evidence that perhaps protein is not the carrier (Avery et al. 1944) and nearly another decade to turn the page on protein (Hershey and Chase 1952) and ascertain that DNA is the true carrier. The building blocks of DNA formation are relatively simple chemical molecules called *nucleotides* or *bases*. Figure 1.2a shows the two kinds, *purines* and *pyrimidynes*. Nucleotides can form chemical bonds to produce two kinds of biomolecules, *Ribonucleic Acid* (RNA) and DNA, two particular kinds of polymers herein simply referred to as *strands* throughout. Each can be a single or a double strand. *Single strands* are obtained when nucleotides are joined by covalent bonds into longer and longer strands, herein referred to as *n*-mers if *n* nucleotides are involved, in a chemical reaction called *ligation*.

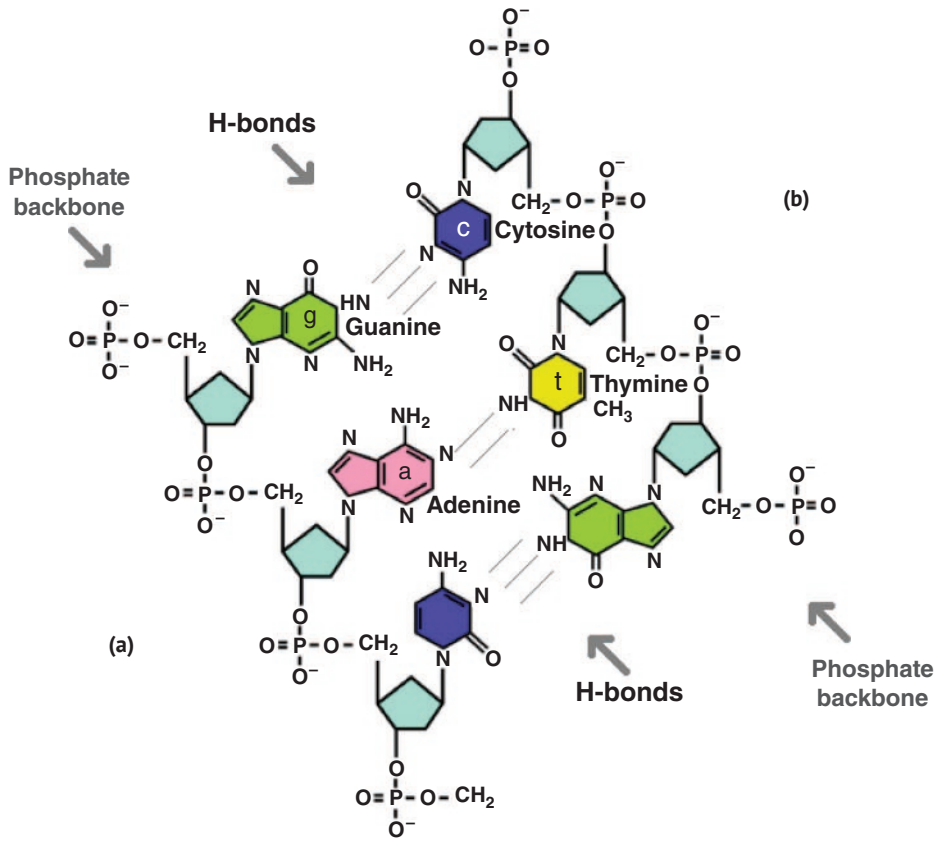

Figure 1.2 The essential molecules for life on Earth include single nucleotides of DNA *a* (adenine), *c* (cytosine), *g* (guanine), *t* (thymine), strands of these, and double helices. Nucleotides are classified into *purines* (two carbon rings) and *pyrimidines* (one carbon ring). (a) They ligate through covalent bonds to form single-stranded DNA molecules (lower left) with dangling hydrogen (H–) atoms. (b) Single strands bond with other corresponding (2 or 3) H-atoms in Watson–Crick (WC) complementary molecules (upper right) to form double helices. RNA behaves likewise, except that *t* is replaced by *u* (uracil).

A single molecule is an ordered structure with a "head" and a "tail" determined by the positions where they attach to the carbon atom in the backbone (the 5′ or 3′ end). They will be described by strings (or "words") of nucleotide characters over the alphabet $\{a, c, g, t\}$, e.g. $5'-acgt-3'$ and *cgta* are two very different strands chemically. (For simplicity in this book, they will always be written in lower case letters beginning with 5′— toward the 3′—end [so the end references can be omitted] and will be further explored in Section 9.1.1). Ligation can happen freely among nucleotides, so that the exact number of possible strands of a given length n grows exponentially with n, as 4^n.

Furthermore, two single strands can then bond through a process called *hybridization* to form the familiar double helix (or *duplex*) proposed by Watson and Crick (1953) to be the basic structure of DNA, as illustrated in Figures 1.2 and 1.3. The simplest example consists of two nucleotides bonded by hydrogen (H—) bonds. The bonds are very stable when the

nucleotides are WC complementary, i.e. when they form a WC-pair of either $c - g$ (3 H-bonds), or $a - t$ (2 H-bonds), but mix-ups ($a - c$, or $a - g$, or $c - t$ or $g - t$) may be forced when they are a part of longer single strands and next to other neighboring nucleotides that are perfect WC-complements and favor hybridization. For every single strand, there is another perfectly matched DNA molecule that it may hybridize with its *WC complement*, obtained by reversing it and swapping the nucleotides for their WC-complements. The hydrogen bonds between a mismatched pair of nucleotides are not very stable by themselves. Two WC-complementary molecules have perfect *hybridization affinity*, but their bond may be stable even with less than perfect affinity (more details below). Thus, single strands can actually hybridize even under less than ideal conditions of perfectly WC-complementary. The chemistry of hybridization requires the 5′- and 3′-ends to face each other, so one of the strands has to reverse direction as well to form the helix. *Hybridization* of two single DNA strands to form a double helix (also called a *duplex*), is an essential process, for example for self-reproduction. This most fundamental and powerful property drives the *exquisite discriminating ability in forming double strands* (*helices*) in most other cell functions, as discovered by subsequent research following up on the elucidation of the nature and structure of DNA by Franklin and Gosling (1953) and Watson and Crick (1953). The physical chemistry of hybridization of specific pairs is not fully understood, and it is very hard to predict whether and how it will actually happen between two given strands x and y when they come in close proximity (within nanometers), even in the case where they are identical. Segments of the strand actually may hybridize to other WC-complementary segments to form what is called its *secondary* structure (illustrated in Figure 1.3). Very long double strands (e.g. in human chromosomes with millions of nucleotides) develop even a *tertiary* structure by coiling and twisting about themselves, as in the natural state of DNA chromosomes (more in Section 8.2.3).

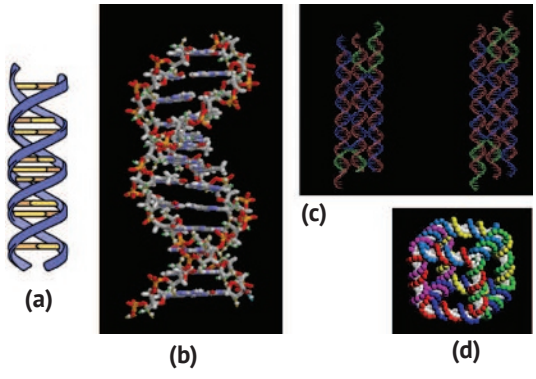

Figure 1.3 DNA strands bond in endless ways. (a and b) Double strands of DNA are formed by *hybridization* of two single strands. (c) Several double helices (red, blue, and green) can have dangling "sticky" ends that can hybridize to analogous complementary sticky ends (of the same color) in other molecules (called "tiles") to form *DNA complexes*, e.g. periodic structures in a process of DNA self-assembly (Seeman 2003). (d) Artistic rendition of a DNA macromolecule with the connectivity of a geometric cube synthesized in a test tube by using copies of the molecular tiles in (c) to build artificial nanostructures through a process of self-assembly (Chen and Seeman) (more in Chapter 8).

Chemically, hybridization is governed by the *Gibbs energy* of hybridization, i.e. the physical energy released (i.e. negative) when two single strands of DNA bond together to form a more stable double helix in an exothermic reaction, akin to potential energy in physics. The process can be reversed by an endothermic reaction that supplies the energy back (e.g. heating the molecule to about 90 °F) so that the H-bonds dissolve, and the duplex breaks back into two single strands. When the Gibbs energy drops below a certain threshold, affinity is enough to hold the single strands together in a stable double helix. The precise threshold can be pinpointed by experimental measurements using high-resolution calorimeters, originally obtained by statistical averages in populations of like molecules with the same composition, but more recently with individual molecules and pairs using optical tweezers (see Gieseler et al. 2021). A threshold commonly used for hybridization is −6 kcal/M for short oligonucleotides (length under 60-mers) and it will be used in the following sections.

Gibbs energies depend on physical parameters (such as the internal energy, pressure, volume, temperature, and entropy) of the environment in which the duplex is formed, in addition to the predominant specific sequence (composition) of the strands. The more negative the Gibbs energy, the more stable the duplex formed. Unfortunately, the available models in biochemistry provide no gold standard to determine Gibbs energies exactly, other than just accepted empirical approximations (Wetmur 1999), so models have been developed to estimate it. The most popular is the *nearest-neighbor model (NNM)* (SantaLucia 1998), although there are more sophisticated models (e.g. the staggered-zipper model) that account for stacking effects, e.g. a bonding pair $c - g$ makes it more likely that neighboring pairs will hybridize even if they are not WC-complementary. NN is an additive model that adds up the Gibbs energies of facing pairs in specific alignments of the strands (including single nucleotides without a matching nucleotide) to determine the total Gibbs energy released. As usual, the molecules will optimize the physical process and hybridize in the frameshift, releasing the most Gibbs energy (something akin to a ball dropping to the lowest possible point when released in a gravitational field to release the most potential energy).

Given the enormous abundance of many different DNA strands (e.g. a human has in the order of 22 000 genes), understanding the behavior of an organism's genomics requires fine discrimination of which DNA molecules one will hybridize to among the many possible competitors available in the pond of the surroundings, offering different affinities. They present a so-called *hybridization landscape* to solve an impossibly difficult minimization problem. *How to approach it, even if only an approximate solution is possible?* These questions will be addressed in Section 1.2 after describing the accumulated larger role that DNA plays in life self-organization at the macro-level as a result of these local interactions at the micro-level.

1.1.2 DNA: Its Role and Function

Biology has developed numerous approaches and tools for the study of biodiversity from different viewpoints, including ecological, systematic, evolutionary, and genetic. Empirical and theoretical approaches afford an understanding of any taxonomic group and its dynamic diversity, as to why it either persists or reaches extinction. The traditional disciplines of biology have been strengthened with tools for handling enormous amounts of information, particularly in disciplines such as genomics and bioinformatics. Today, biological science seeks to explain the reason for the evident diversity, anticipate possible changes resulting from global climate change, and predict possible consequences in the face

of a changing planet. Yet, it is also remarkable that our knowledge of the essential DNA molecule has not deepened in its own right after its discovery, an issue to be addressed later in Section 1.2, after summarizing first its deep impact on other aspects of the diversity of life.

Organisms are broadly classified into two categories: *procaryotes* and *eukaryotes*, based on the presence or absence of a nucleus within the cell. The first one includes organisms, which do not have a nuclear membrane, so genetic material remains within the cytoplasm, where all processes related to the function of the cell take place. In the second category, organisms do have a nucleus and a membrane that separates its contents from the rest of the cell. Some processes take place within the nucleus, whereas others occur in the cytoplasm. The *genome* is the whole set of genetic material within a cell that may include just a few hundred genes, or in many cases thousands of genes. Despite differences in genome size, DNA is similar in symbiotic bacteria, the smallest genome size found so far (139–250 Kbps) encoding about 121–127 proteins (McCutcheon and Moran 2012) to the largest genome reported for a eukaryote in the angiosperm *Paris japoniza* with a genome of ca. 148 000 Mbps (Pellicer et al. 2018). Viruses also contain genetic material (DNA or RNA) even though they are not considered living organisms because they do not require a substrate to feed and survive, they do not communicate through quorum sensing like bacteria do, and they do not reproduce asexually or sexually. Instead, they replicate by using the replication machinery of a host (Sanjuán and Domingo-Calap 2016).

A *gene* is a stretch of DNA encoding for a protein. These long stretches of (21) amino acids provide enormous diversity in structure and function. Genes may vary in length, from small sizes as 76 bps as in human histones (Alberts et al. 2002), a protein responsible for the structural support for a chromosome. Others may be very long, such as the dystrophin gene that spans 2200 bps in humans, a protein important for both skeletal and cardiac muscle movement (Monco et al. 1986). At any rate, the flow of information starts with DNA encoding the genetic information; DNA is then copied to messenger RNA (mRNA) for some genes, through a process called *transcription* (each triplet or 3-mer of DNA constitutes a *codon* that basically encodes an amino acid). This mRNA participates in the process of *translation* where a new protein is produced. This flow of information in one direction, from DNA to RNA to protein, or RNA directly to protein, is referred to as *the Central Dogma of molecular biology*, a topic that will be discussed in detail in the following chapters.

Proteins perform a wide variety of functions, including building and repairing tissues, transporting molecules, and regulating the body's metabolism. Despite the differences that can occur in different cellular processes in organisms with different levels of complexity, it is surprising how highly similar the processes of information transmission from DNA to RNA and translation from mRNA are. However, the biggest differences come from gene regulation processes. As the understanding of DNA as the blueprint of life continues to deepen, it is necessary to broaden and open up new approaches and disciplines to better understand its mechanisms of action and modes of evolution.

1.1.2.1 Replication

A notable feature of DNA is its ability to replicate itself, an essential process for the survival and reproduction of all organisms on the planet. When cells divide, the DNA is replicated to ensure that each new cell receives a complete set of genetic information

(Kornberg 2000). The DNA replication process is extremely precise, ensuring that genetic information is passed down faithfully from one generation to the next, and surprisingly, it is very well conserved in all organisms, although there are some differences between prokaryotes and eukaryotes (Bell and Dutta 2002), reflecting the differences in their organization, complexity, and evolutionary history. For example, in prokaryotes, DNA replication occurs in a single circular chromosome located in the cytoplasm, with a single origin of replication. Replication proceeds in both directions along the chromosome until the two replication lines meet. On the other hand, eukaryotic replication is more complex, with multiple origins and replication occurring simultaneously on many linear chromosomes located in the nucleus. The replication machinery in prokaryotes involves only a few proteins, such as DNA polymerase, helicase, and a few single-stranded binding proteins, while in eukaryotes it is more complex and involves a larger number of proteins, including DNA polymerases, helicases, topoisomerases, and other related proteins. The replication process is also different in terms of time, since prokaryotes take as little as 20 minutes due to their relatively small and less complex genomes, while eukaryotes may take several hours to replicate during a cell division process (Bell and Dutta 2002; O'Donnell and Kuriyan 2006).

Finally, eukaryote replication differs from prokaryotes in that eukaryotes possess telomeres, specialized structures located at the ends of linear chromosomes that protect the genetic material from degradation and maintain genome stability (Olovnikov 1971). The telomeres are shortened during the process of replication, which eventually leads to cellular death or senescence (Lundblad and Szostak 2019; Ait Saada and Lambert 2021). In contrast, prokaryotes do not have telomeres and genome replication occurs without genome shortening (Watson et al. 2013).

1.1.2.2 Transcription

Copying DNA into mRNA is catalyzed by the enzyme RNA polymerase, which recognizes a specific sequence of nucleotides on the DNA molecule called the promoter, located upstream of the coding region. In eukaryotes, transcription is initiated by the binding of several transcription factors to the promoter sequence. The RNA polymerase unwinds the DNA double helix and synthesizes a new mRNA molecule complementary to one of the DNA strands carrying the genetic information used to synthesize proteins (Alberts et al. 2002). This complex process guarantees the expression for the proper functioning of cells and organisms, and it is regulated by different factors, including the presence of specific transcription factors and the activity of other regulatory proteins. Additionally, different types of RNA molecules are produced by transcription, including ribosomal RNA and transfer RNA, which are involved in the synthesis of proteins (Lodish et al. 2000). Another notable difference between prokaryotes and eukaryotes relates to post-transcriptional modifications. While in prokaryotes, mRNA is immediately available for translation; in eukaryotes, newly synthesized RNA molecules undergo extensive post-transcriptional modifications, including the addition of a 5′ cap and a poly(A) tail, as well as alternative splicing of introns (i.e. DNA segments that get transcribed, but do not make part of the final mature mRNA). This allows the production of multiple protein isoforms from a single gene, which enhances the functional complexity of organisms (Black 2003; Alberts et al. 2002).

Interestingly, organisms can regulate or control protein synthesis through post-transcriptional control; in prokaryotes, this process is relatively simple since it only involves

the binding of regulatory proteins to the promoter region. A more complex control expression is observed in eukaryotes and involves chromatin (a complex of DNA and proteins) remodeling, where the structure of the molecule is modified to regulate the access of transcription factors and RNA polymerase to the DNA (Li et al. 2007). Chemical histone modification (acetylation, methylation, phosphorylation, and ubiquitination) with different effects on gene expression (Liu et al. 2021) and binding of transcription factors to enhancers and silencing sequences (Hnisz and Young 2013) may promote or inhibit the activity of a gene and often are cell-type specific. This specificity permits accurate control of gene expression throughout the developmental process, cellular differentiation, and response to external stimuli (Villar et al. 2015).

1.1.2.3 Translation

For translation, a structure inside the cell is made of RNA and proteins, where protein synthesis (called the *ribosome*) binds to an mRNA molecule carrying the genetic information that is used to synthesize a protein. The process consists of three main steps: initiation, elongation, and termination. In initiation, the ribosome identifies and attaches to the mRNA molecule at a specific nucleotide sequence known as the start codon, which triggers the protein synthesis process. During elongation, the ribosome reads the genetic information encoded in the sequence of codons on the mRNA molecule, recruiting specific amino acids and linking them together to form a polypeptide chain (Lodish et al. 2012). This process continues until the ribosome encounters a stop codon, which indicates the end of protein synthesis. Once the ribosome releases the newly formed protein and the mRNA molecule, the protein folds into a characteristic shape that allows it to perform its functions. Like replication and transcription, translation is a complex process that is strictly regulated by various factors, such as specific enzymes and regulatory proteins (Pestova et al. 2001). For example, translational control may involve a complex network of regulatory factors, including RNA-binding proteins and microRNAs, that can either promote or inhibit protein synthesis. Regulatory factors bind to specific targets of mRNA molecules, changing their translation when altering the stability, localization, and access to the mRNA. In some cases, regulatory proteins bind to the 5′ untranslated region (UTR) and 3′ UTR of mRNA molecules, which can either promote or disrupt translation. These events are context-dependent, including the presence of other regulatory factors (Sonenberg and Hinnebusch 2009).

Since DNA and its products resulting from transcription and translation are required for every cellular function of an organism, DNA can be used as a molecular fingerprint that allows for a variety of studies, such as the detection of genetic variation to determine and track genetic diversity for population structure or phylogenetic inferences, forensic analysis, paternity testing, plant breeding to identify genetic markers associated with desirable traits, genomic studies involving the entire genome, and gene expression analysis to evaluate differences in expression under different variables or treatments. Significant advancements in high-throughput technology have allowed the generation of impressive amounts of data, which calls for new approaches to analyzing it and has revolutionized the way scientists interpret gene expression and the evolution of organisms. These advancements include RNA sequencing (RNA-seq), which allows for high-throughput sequencing of RNA molecules and can be used to quantify gene expression levels, identify alternative splicing

events, and detect novel transcripts (Wang et al. 2009); single-cell RNA-sequencing (scRNA-seq), allowing the analysis of gene expression in a single-cell, providing insight into cell heterogeneity and cell-specific gene expression patterns (Stuart et al. 2019); microarrays, which allow for the simultaneous analysis of thousands of genes and are used to quantify gene expression levels and identify differentially expressed genes (Brazma et al. 2001); reverse transcription quantitative polymerase chain reaction (RT-qPCR), which is used to quantify gene expression levels (Higuchi et al. 1996); NanoString, a digital gene expression technology that uses color-coded molecular barcodes to detect and quantify gene expression levels, particularly useful in low-abundance transcripts (Geiss et al. 2008); and spatial transcriptomics, which allows for the analysis of gene expression *in situ*, providing spatially resolved information (Ståhl et al. 2016).

However, efficient analysis of this large and complex information requires novel approaches that combine multiple disciplines and technologies for studies at the level of organisms and their evolution. This approach integrates molecular biology, genetics, genomics, bioinformatics, ecology, and evolutionary biology to develop a more comprehensive understanding of biological systems. Computational methods are particularly useful in analyzing the large datasets generated by genomics and other molecular biology techniques. These methods help biologists and bioinformaticians identify patterns and relationships within genomic data, leading to new insights into how organisms have evolved over time. An integrated approach that combines multiple disciplines and technologies is increasingly necessary for a more holistic understanding of biological systems and their evolution. Such approaches have the potential to revolutionize our understanding of biology and are likely to have numerous applications in fields such as medicine, agriculture, and conservation biology. By utilizing contemporary analytical methods, one can gain a comprehensive understanding of biological processes and systems.

1.1.3 Alignment-based Methods

The sequencing era started in 1977 when the Sanger method allowed the possibility of obtaining the sequence of small DNA fragments (Sanger et al. 1977). In the 1980s, it became possible to sequence small genomes with the Sanger method or modifications of it. In the early 2000s, the first draft of the human genome was launched and became a significant milestone in a new area, "genomics," which soon revolutionized biology with high-throughput sequencing technologies with reduced costs and time required to obtain the precise composition of the sequences in text form. *How can such a diverse amount of information be efficiently analyzed? What kinds of questions in biology and evolution (only dreamed of just a few decades before) could be explored now?* Computational biologists soon started to develop tools to compare several sequences at once to make biological inferences about groups of organisms. Most methods place sequences side by side in comparative blocks (called *alignments*) of nucleotide bases of DNA or amino acids and use some criterion of similarity or correspondence based on the WC-complementarity of nucleotide bases or protein products. While these methods provide invaluable insights into the evolution of molecules and organisms, they also present significant challenges and limitations that became evident with their use: considerable variations in DNA sequences, particularly of

divergent groups, genome size, appropriate thresholds for similarity criteria, and computational effort (more in Section 1.2). Thus, alignment-free approaches arose as an alternative to alignment-based methods for exploring biological questions.

The goal of this section is to summarize the most important alignment-based approaches and the most commonly used methods available. Alignment-free methods will then be summarized in the following Section 1.1.4.

Alignment-based methods search for identical matches at each position in two sequences of nucleotides or amino acids under comparison. Methods usually aim to find an alignment that optimizes the *homology* of the sequences, that is, the existence of a particular nucleotide a, c, g, t for two or more sequences resulting from the existence of a common ancestor. However, it is possible that such a character with the same state (e.g. a matching nucleotide) has evolved independently in the two sequences. Both patterns can occur in related species (*parallelism*) or distantly related species (*convergence*). Homology is commonly confused with simple similarity of sequences, e.g. some transcription factor binding sites (Stormo 2000) are confused with regulatory motifs (Roth et al. 1998). To complicate matters further, homologous sequences may appear to be very different (Goodman et al. 1975; Ferrier and Holland 2001).

Most algorithms incorporate gaps to consider insertions and deletions in sequences presumed to have occurred over time. There is a system that optimizes the presence of gaps for the definition of homologous positions. However, it also penalizes them when establishing blocks of correspondence between all the sequences under comparison (see Figure 1.4). Alignments are more straightforward when the sequences have similar or identical lengths, since gaps are not required to optimize the correspondence between character states for the different sequences. However, when the sequences are of different lengths, various outputs can result when changing parameters for the gap size and penalty values.

An explicit evolutionary model in the construction of alignments assumes that the observed changes obey a specific pattern of change according to one model and are different

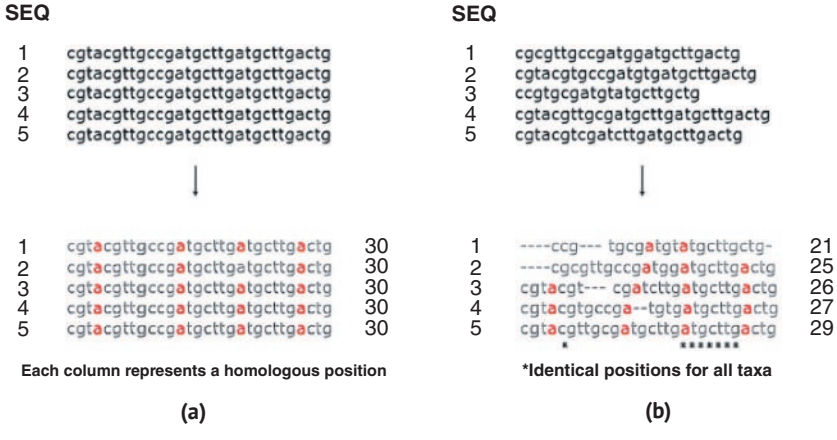

Figure 1.4 In a typical alignment-based protocol: (a) Very similar sequences of equal length do not require gaps, so homologous positions are easier to find. (b) Divergent sequences of different lengths will require gaps. The definition of homologous characters can generate alternatives depending on different parameters in the analysis or refinements of the alignment by the researcher.

from others. For example, the rate of substitutions: transitions (interchanges of purines or pyrimidines) or transversions (interchanges of purine for pyrimidine bases) is different, and the nucleotide frequencies can be the same or different. Since the variables between models are different between sequences, the choice of model can influence the analyses carried out and produce different conclusions. In the case of proteins, there are substitution matrices (scores for aligning nucleotides or amino acids, for example point accepted mutation (PAM) (Dayhoff 1978) or BLOSSUM (Henikoff and Henikoff 1992) that also require interpreting a specific evolutionary model that does not necessarily correspond to the real one for all protein sequences (Dayhoff 1978).

Finally, many of the methods assume that the substitutions are made stochastically and independently. That is, one substitution does not affect the probability that another occurs. However, there is evidence that this is not necessarily the case since the existence of a change (for example, transitions $a-g$, $c-t$) has a clear effect on the changes that can occur in the opposite chain.

Although alignment-based methods have been frequently used in biology, they present some difficulties, particularly in their application to large and highly divergent datasets. Some of these disadvantages include:

- Homology is a key point when performing alignments. If the sequences share a common ancestor, analyzing the corresponding fragments from different organisms is ideal. If the fragment is short and identical in all the sequences, the alignment can be established relatively easily. However, as sequences become more divergent, significant changes can result from evolutionary processes such as genetic recombination, horizontal gene transfer (HGT), and gene loss or gain, making the sequences less likely to be homologous.
- In DNA sequences with only four nucleotides, sequences can align by chance, even if they are not related. The same can occur even in amino acid sequences, although with lower probability. This is particularly true when the gene sequence is unknown or when their lengths are significantly different.
- The amount of computer memory and the time required to compute the alignment increase as the number of sequences or their length increases, as in the case of large genomic data, because the number of possible alignments increases exponentially. This means exhaustive searches are impractical or impossible to obtain optimal alignments, so that typical alignments in practice are just approximations obtained using simplifying assumptions and heuristic approximations. This means that the conclusions are uncertain, since it is highly likely that the resulting outcome will not be close enough to the actual evolutionary trajectory.
- The alignments are based on many *a priori* variables (gaps, insertions, deletions, evolutionary models), and their choice is therefore subjective and dependent on the researcher. Since the choice of variables for analysis impacts the results, incorrect alignments are likely to lead to conclusions that do not correspond to biological reality.
- In assumptions made for multiple alignments where insertions and deletions (indels) are expected, it is common to specify gap penalties and gap size specifications to obtain better scores, which again generate a subjective dependency on the researchers' biases. That can lead to different results depending on the initial assumptions and incorporate

biases in the analysis, which make it difficult to determine whether they correspond to biological reality or not.
- Models that assume a mode of evolution may not correspond to biological reality. For example, assuming differences in mutation rates may be a simplification of the way molecules evolve over time.

Classical alignment methods are divided into three main categories based on their approach to establishing the alignment blocks of the sequences under study, as described in Table 1.1.

1.1.4 A Review of Common Alignment-free Methods

The goal of this section is to summarize the most common and significant alignment-free methods, including frequency of *k*-mers, genomic signatures on genomes, randomness and complexity of DNA sequences, chaos game representation, and machine learning approaches.

Table 1.1 Major alignment methods.

Sub/type	How it works	Use
Pair alignments		
Global alignments	Analyzes a pair of sequences from start to finish	Sequences of equal size and closely related, Needleman–Wunsch algorithm (Needleman and Wunsch 1970)
Local alignments	Find the best alignment between subsequences	Sequences of different sizes, highly similar (Smith and Waterman 1981)
Multiple sequence alignments		
Progressive alignments	Aligns sequences step by step, creating a guide tree to do progressive alignment	Useful for multiple sequences of different sizes, ClustalW (Thompson et al. 1994), Muscle (Edgar 2004)
Iterative alignment	Performs alignments iteratively to improve accuracy	Divergent and large sequence data from phylogenetic diverse groups, MAFFT (Katoh et al. 2002), T-Coffee (Notredame et al. 2000)
Consistency method (ProbCons)	Incorporates a measure of consistency from pairwise alignments, followed by multiple alignments	Highly divergent sequences or moderate levels of similarity, or when high accuracy is required, ProbCons (Do et al. 2005)
Hidden Markov Models (HMMs)		
Model the probability of sequences given a particular alignment	Uses statistical methods to represent groups of sequences (families) and, based on observed patterns, predict alignments	High variability or when searching for remote homologs in very large genome datasets, HMMER3 (Eddy 2011).

In contrast with global alignment-based methods, alignment-free methods make comparisons between small fragments of the sequences (e.g. k-mers) to evaluate the similarity of sequences. Alignment-free methods are used in biology and bioinformatics to assess the similarity of biological sequences such as DNA, RNA, and proteins of organisms, without the typical sequence alignments (Zielezinski et al. 2017). They are also used when the data set is large or contains very divergent sequences and when insertions or deletions are very frequent. While sequence-alignment methods identify similar regions, alignment-free methods bypass that step, making it faster and more applicable to complete genomes and metagenomics datasets, for which alignment-based methods are very challenging or misleading. The methods have proved efficient, scalable, and robust, making them a viable alternative to classical alignments. Alignment-free methods have a well-supported conceptual framework for linear algebra, information theory, and statistics.

The advantages of alignment-free methods can be traced back to the ways similarity searches are performed or the kinds of assumptions of the methods.

- **Homology:** This key aspect of the alignment-based methods in Section 1.1.3 is not considered in alignment-free methods. Regardless of their size, fragments come from different sites within the genome, and there is no presumption that they belong to the same site in the organism's chromosome or genome.
- **Speed:** Alignment-free methods are generally faster and less computationally intensive than traditional alignment-based methods. As the lengths of the sequences increase, the number of possible alignments grows exponentially with conventional approaches, and getting a nearly optimal alignment becomes very difficult. For example, for two sequences of length n, there are $(2n)!/(n!)^2$ different gap alignments (Lange 2002; Zielezinski et al. 2017).
- **Capacity:** They are helpful in the analysis of large datasets obtained from NGS data that can be produced in relatively short periods of time. The data volume of samples sequenced up to 2011 was estimated to be only $10 - 20\%$ of the total DNA on Earth (Microbiology 2011), which illustrates the complexity of the amount of data yet to come and the need for computational power and alternative approaches to analyze that data.
- **Robustness:** Given the high variation in DNA sequences, the methods have been demonstrated to be helpful with divergent sequences. They seem less sensitive to possible genome sequence arrangements (insertions and deletions), which are usually problematic in alignment-based methods.
- Some of these methods examine physical, local interactions in both space and time among the basic molecules of DNA, just as they occur in nature. This allows for the extraction and inference of information as it accumulates over space and time to be used in a predictive fashion.

Although the methods seem to be very effective, particularly in large and divergent sequences, they also present some disadvantages:

- Alignment-free methods are based on similarity queries based on features of the sequences and on measurements of similarity or dissimilarity, which are used to construct trees. Although similar, trees may be like evolutionary trees; that is not always the case. Additionally, once the tree is built from a distance matrix, original sequences cannot be used as characters to trace back the evolutionary history of those sequences along the tree.

- The method analyzes global patterns (k-mers), which can cause localized variations to be ignored. Additionally, the size of the k-mers can generate biases. If k is very small, the method may not capture that variation, but if it is very large, it may overemphasize small differences.

Two major methods exist in alignment-free approaches (Vinga and Almeida 2003). First, *word-based methods* analyze frequencies of subsequences of different lengths, usually defined by the researcher. Second, *information-theoretic methods* measure the informational content between full-length sequences (Zielezinski et al. 2017), as shown in Table 1.2.

1.1.4.1 Word-based Methods

Instead of comparisons based on typical residue correspondence when comparing two sequences, like aa, gc, cc, tt in complete segments of DNA, the sequence representation of alignment-free approaches is based on the comparison of sequences with small fragments (k-mers) of different lengths. For example, the sequence *attgc* can be represented as k-mers of length $k = 2$: at, tt, tg, gc, or as 3-mers: att, ttg, tgc. The shorter the k, the more likely it is that the k-mer will appear randomly in the sequence and longer k-mers should be used for very similar sequences (Zielezinski et al. 2017). Long k-mers (in the range $k = 2 - 6$) are

Table 1.2 Major alignment-free methods.

Method	Criterion	Distance	Software
Word count	Frequency of k-mers of different lengths K	Euclidean, Manhattan, Pearson correlation coefficient	KAT (Mapleson et al. 2017), JellyFish (Marcais and Kingsford 2011), Kraken (Wood and Salzberg 2014).
Information content	The amount of information shared between biological sequences	Lempel-Ziv complexity estimation (Lempel and Ziv 1976)	(Liu et al. 2012)
Entropy	"Uncommon words"	Shannon Entropy H, Kullback–Leibler divergence	FSC-Q (Haubold et al. 2015)
Chaos game representation	Biological sequences are transformed into unique fractal patterns to be compared in terms of similarity.	Euclidean, Manhattan, Pearson correlation	CHAOS/DIALIGN (Brudno et al. 2004)
Iterative maps	Mathematical transformations are applied to biological sequences.	Euclidean, Hausdorff	Matlab and Python scripts
Graphical representation	Transformations of sequences to plots representing features such as GC content, k-mer frequency, others	Manhattan or other metrics of similarity in graphical representations	GraphDNA (Thomas et al. 2007), MatGAT (Campanella et al. 2003)

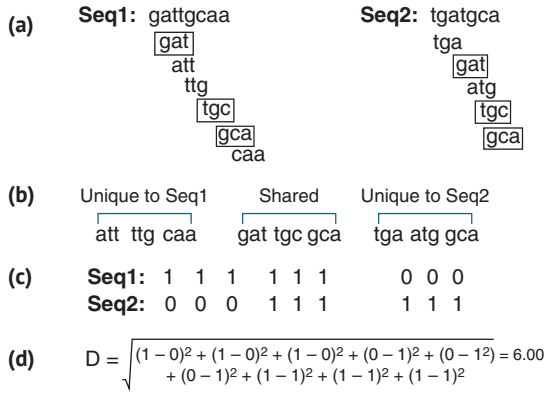

Figure 1.5 Word-based methods used in the alignment-free approach.

useful for a wide range of pylogenetic divergences (Chan et al. 2014), and up to 25-mers for very similar sequences as isolate comparisons of bacterial species (Bernard et al. 2019).

Several steps are included in the process, as shown in Figure 1.5.

- Original sequences to compare, seq1 and seq2, are broken up into different fragments of size $k = 3$. There will be unique and shared fragments (in squares).
- Unique and different fragments are joined together; there will be unique sequences to each original sequence as well as fragments common to both.
- Each sequence is transformed into a vector (array of numbers). Word count consists of determining the presence or absence of each fragment. Zero times when a sequence is unique to one of the query sequences, 1 when it is present one time, 2 when it is present two times, and so forth. This step will allow distance determination.
- Some distance is calculated, e.g. the ordinary Euclidean distance, but other metrics include Hamming and the Jaccard Index (Tang and Gaut 2010). The higher the number, the more dissimilar the sequences, whereas identical sequences will be at a distance 0, as should be the case for any metric. (The concept of distance is defined precisely in Section 1.2.)

1.1.4.2 Information Content

These methods evaluate the informational content of the sequences at full length by recognizing and computing the amount of information shared between biological sequences. The method compares two sequences at a time; if they are identical, the information content is 0, whereas the content increases as differences between the sequences increase. Different measurements evaluate the distance among sequences based on the information content. For example, when comparing a pair of sequences, it is possible to assess the number of different subsequences in each one.

Table 1.3 Major results with alignment-free approaches.

Application	Advances	Examples and references
Phylogenetics	Simulated data (nucleotides and amino acids) show accuracy and robustness in different empirical sets.	Chan et al. 2014
Genomic analysis	Phylogenetic analyses in hierarchical (vertical) and reticulate (lateral) aspects of genome evolution.	Bernard et al. 2019
Metagenomics	In simulated microbial genomes, the sensitivity of methods is evaluated under lateral gene transfer and genome rearrangement scenarios.	Bernard et al. 2016
Virus evolution	New method that avoids the computational complexity of multiple alignments. Applied to several viruses, such as SARS-CoV-2, Dengue virus, Hepatitis B virus, and human rhinovirus, with similar results to those obtained by alignment-based methods.	He et al. 2021

1.1.4.3 Entropy

A measure of uncertainty can be defined without resorting to thermodynamics. The Shannon entropy index was introduced by Shannon (1948) to quantify the uncertainty (entropy) in the values of an RV (see Section 1.2). Other authors have developed ways to measure the entropy to analyze sequences based on the frequency of symbols a, c, g, t's.

1.1.5 Major Results by Alignment-free Methods

The goal of this section is to summarize the major results obtained with alignment-free methods, including large-scale genomic comparisons, microbial diversity and metagenomics, and comparative genomics and phylogenetics.

Since alignment-free approaches allow comparing whole genomes, metagenomes, or very large datasets, their applications are versatile, and each identifies some type of similarity to address specific questions in biology. Applications include identifying similarities and differences among organisms, inferring evolutionary relationships, and identifying evolutionary patterns, as summarized in Table 1.3.

- A common approach is to choose a distance metric to compare sequences of organisms to infer *phylogenetic relationships* between them (e.g. which one came first in the course of evolution). The results can be compared with those obtained with alignment-based methods. There is also the possibility of generating phylogenies based on whole genomes.
- In areas such as epidemiology and genomics, pathogen outbreaks can be tracked by comparing their sequences, which is very useful for understanding the dynamics of transmission and possible evolutionary changes in pathogens.

- **Epidemiology and Genomics:** Pathogen outbreaks can be tracked by comparing their sequences. This is very useful for understanding the dynamics of transmission and possible evolutionary changes of pathogens as they invade the host and respond to its immune system.
- The field of *metagenomics* deals with interactions among different species of organisms in a range of biome. Samples contain a mix of different types of organisms (viruses, bacteria, and fungi). Their analysis with alignment-free methods reduces the problems associated with the lack of homology typically found in their alignments.
- A recent application consists of analyzing genomic signatures to detect deviations from genomic content that may indicate the presence of (recombination of) foreign DNA, typical of *HGTs* from other organisms.
- Identification of HGT and detection of recombination.

1.2 Hybridization Affinity

The alignment-based methods described in Section 1.1 leave several questions unresolved when addressing fundamental problems in biology. For example, many of the alignment-based methods are effective because of the use of statistical tools that make sense at the level of populations of organisms and establish correlation effects between DNA and the macro-observables (e.g. phylogenetic trees, defined precisely in Chapter 6). However, they leave a logical gap in bridging the connection between them (e.g. what are the physical, chemical, and/or environmental reasons why they happened so?) Furthermore, alignments impose a heavy computational burden that increases proportionally and jointly with the number of organisms involved and the length of the sequences used as proxies. On the other hand, alignment-free methods based on physical chemistry have the potential to address this gap because they are usually based on local interactions between molecules that reflect the physical nature of underlying biological processes; just as important, they also provide more causal chains of interaction connecting them to enable prediction *at the level of individual organisms.*

The goal of this section is to address the same problems, perhaps in more effective ways, using alignment-free methods, or where alignments prove impossible to use, either because of practical issues of efficiency or because of more principled issues.

1.2.1 Models of Gibbs Energy Landscapes

A predictive science requires precise definitions of the concepts involved in the predictions. For further data analyses in later chapters, the goal of this section is to provide more precise and quantitative definitions of models of DNA molecules and their representations on digital computers than are commonly given in biology. Of course, they are abstractions of the actual chemical molecules described in Section 9.1.1 that may appear to be gross oversimplifications to a biologist's mind, but they will suffice to deliver some new, interesting predictions in later sections that bear reasonable accuracy.

The goal of this section is to describe the physical chemistry of DNA molecule interactions that enable its functional properties in two fundamental ways. First, as a repository of basic information about the physical expression of an organism (phenotype), and second,

as the mechanism for its interaction with other DNA molecules to enable integration of this information across space and time to enable biological function.

As mentioned above, the fundamental physical quantity associated with DNA is *Gibbs energy* (Holde 2006). The simplest estimates of the Gibbs energy are given in the form of melting temperatures as averages of the energy (heat) that must be supplied to a population of duplex molecules of the same composition (sequence) in order to dissociate/denature/melt (break the H-bonds) about 50% of the molecules. An estimate is given by Wetmur (1999, 1976).

$$T_m = \frac{\Delta H^\circ}{\Delta S^\circ + R \ln([c_t]/4)},$$

where ΔH° is the change in enthalpy, ΔS° is the change in entropy, R is a gas constant in the entropy model, and $[c_t]$ is the molar concentration. Single-stranded molecules absorb UV radiation, and therefore, as the temperature increases, increased absorbance with melting can be detected in a photograph by gel electrophoresis (Holde 2006). For a biologist, the obvious criteria to decide hybridization are usually related to the GC content of the sequences, since it is a good indicator of the melting temperature of short oligonucleotides (Wetmur 1999). These estimates are not very useful when it comes to predicting the hybridization behavior of two specific sequences among a mixture of many different molecules of very different compositions.

The major factors intervening in the calculation of Gibbs energies of two specific sequences x and y are their actual compositions (i.e. sequences) and the Gibbs energies associated with various frameshifts (an example can be seen in Figure 1.6) to capture the optimization done by the actual physical process underlying their hybridization, according to the Second Law of Thermodynamics. Assuming that the physical factors in the environment (e.g. solvent in a test tube) are chosen appropriately (e.g. human body temperature 37 °C), a model can simply use experimental measurements of the various combinations of facing nucleotides in a given alignment and combine them into a total for the Gibbs energy. The simplest model, the *NNM* is a *linear* model that just adds up those pairwise energies from a table of pre-determined (usually experimental) estimates for all possibilities. The actual Gibbs energy for the pair is taken to be the minimum of all estimates *over all possible frameshifts*. The model produces reasonable estimates when there is no secondary structure in the strands x and y, for example, it rarely appears in short oligonucleotides (of length up to 60-meters) (SantaLucia 1998). More realistic approaches account for stacking effects, where the energies for a pair of facing nucleotide in a frameshift depend on the neighboring pairs, for example two pairs of consecutive nucleotides, or nucleotides within a given radius (the staggered-zipper model).

Even staggered-zipper models are insufficient to obtain a good idea of the range of the likelihood of hybridization across all possible pairs of DNA strands x and y, even of the same size, simply because the number of combinations grows at a high exponential rate to apply these methods pairwise exhaustively. To do so, more theoretical models of hybridization that afford analyses of the Gibbs energy landscapes have recently become available. At a fundamental level, this problem can be construed as a standard molecular problem in conventional physical chemistry. However, the solutions can only be arrived at by a more abstract definition of the concept of DNA strands and their entire ensemble, at least for

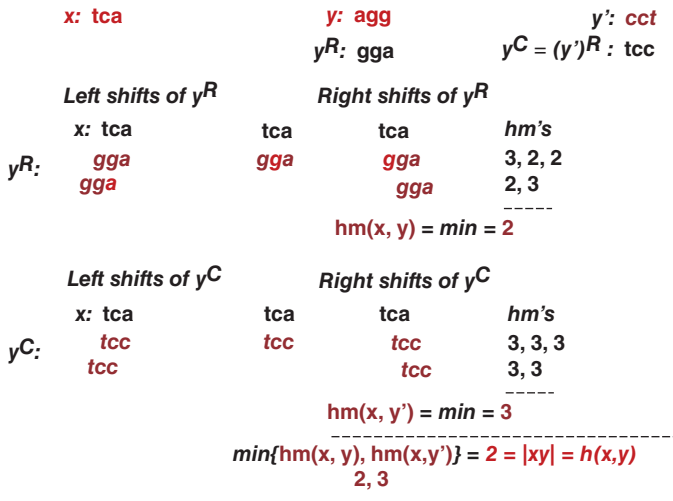

Figure 1.6 The *h*-measure between any pair of pmers x and y chooses the frameshift with most WC-complementary (min of WC-nonmatching) pairs of facing nucleotides in the *h*-measure s of the pairs x, y and x, y' (lone nucleotides are considered mismatches) and the *h*-distance chooses the minimum of the two. The value of the *h*-distance varies between 0 and $m = |x| = |y|$, their length. The smaller the *h*-distance, the more likely the oligos are to hybridize, e.g. $|xy| = 0$ means the oligos are perfectly complementary or identical (like $x = aaa/ttt$ and $y = ccc/ggg$), whereas $|xy| = m$ means the oligos at the maximum *h*-distance are very unlikely to hybridize (like $x = aaa/ccc$ and $y = ccc/ttt$.)

strands of a given length, affording the computational approaches described next. Their analysis leads to interesting new views of hidden structures in the world of DNA and RNA oligonucleotides.

Definition 1.1 DNA sequences and spaces (Garzon et al. 1997) A *DNA sequence* x is a string of symbols from the DNA alphabet $\Sigma = \{a, c, g, t\}$ of length a positive integer $m > 0$. They will also be referred to as polymers of length m, or just as m-mers. The string x' obtained after first taking the reverse of x (i.e. x^r) and replacing every a (c) by t (g, respectively) and vice versa, i.e. $x' = (x^r)^c$ is the *WC complement* of x. An (unordered) pair of two WC complementary pmers sequences $\{x, WC(x)\}$ (simply denoted x/x', or just x, the lexicographically first element in the pair) will be referred to as a *pmer* (or $|x|$-pmer). The set of all such m-pmers will be referred to as the *DNA space* $\mathbf{D}_m$ of length $m > 1$.

If $WC(x) = x' = x, x$ will be referred to as a WC-palindrome. If so, the corresponding pmer is really a set with a single strand. WC-palindromes will be excluded from consideration throughout for reasons that will become apparent below.

1.2.2 Deep Structure of DNA

This section describes computational *metric models* of hybridization between DNA strands of a fixed length that will reveal some hidden deep structure of DNA. This structure will be used and leveraged in the next section to obtain alignment-free methods to perform

exponential dimensionality reduction of arbitrary long DNA sequences to low-dimensional feature vectors, which in turn will make possible genomic analysis using machine learning in the following chapters. A metric model is one satisfying three properties (reflexivity, symmetry, and the triangle inequality) that have been identified in mathematics and the physical sciences to be essential to our abstractions of ordinary space to enable us to navigate the ordinary world with powerful intuitions (e.g. standard Euclidean geometry, calculus, and so forth).

Definition 1.2 Distance function (Garzon et al. 1997) An assignment of distance values $h(x, y)$ to every pair of points x, y in a set $\mathbf{D}_m$ is a *distance function* (or just a *metric*) if and only if every triple x, y, z of elements in $\mathbf{D}_m$ satisfies

- (Reflexive) $h(x, y)| = 0$ if and only if $x = y$
- (Symmetry – distance is adirectional) $h(x, y) = h(y, x)$
- (Triangle Inequality – a detour increases the distance) $h(x, z)| \leq h(x, y)| + h(y, z)|$.

The natural phenomenon of Gibbs energy and its current biochemical models, in particular the NNM, lack all the three properties of a *metric* for a good approximation that lends itself to deeper analysis and makes it very difficult to navigate the landscapes, for example to select a good set of probes to capture information from DNA sequences by hybridization in a microarray in a reproducible way. The question then arises: *Is it possible to find a systematic metric approximation of the Gibbs energy of hybridization?* Perhaps surprisingly, useful insights into the structure of Gibbs energy landscapes of DNA duplex formation have been revealed through such an approximation for DNA oligonucleotides of the same length, as discussed next.

A natural first step would be to consider familiar metric models in Shannon's information theory of error-correcting codes, for example the Hamming distance function between *binary strings* of a fixed length given by the number of mismatches (pairs $0-1$ or $1-0$). The Hamming distance between any two sequences counts, in a perfect alignment, the number of positions where the facing characters do not match. For a more appropriate model of hybridization, this concept must be modified so that *matching* now refers to WC complementary pairs, i.e. only $a-t$'s or $c-g$'s. Despite this is a step in the right direction, Hamming distance between two DNA sequences remains too crude as an estimate of the affinity of strands for hybridization, simply because it excludes the possibility of two strands hybridizing in shifted alignments, something occurring much more frequently with actual DNA. Hence, an alternative model was introduced in Garzon et al. (1997) and Phan and Garzon (2009), the *hybridization distance*, or just h-distance. This h-distance turns out to be a reasonable choice for an approximation of the Gibbs Energy because it satisfies metric properties that biochemical Gibbs energy models do not; more effectively, *because hybridization decisions made when the h-distance falls below an appropriately chosen threshold τ agree with those given by the NNM of the Gibbs Energy (close to the actual decisions made by the real oligomers of length below 60 or so) about* 80% *of the time* (Garzon and Bobba 2012; Garzon et al. 2009).

The h-distance reveals a natural geometric representation of sequences in DNA spaces of an arbitrary length in such a way that the physical distance between any two points in the

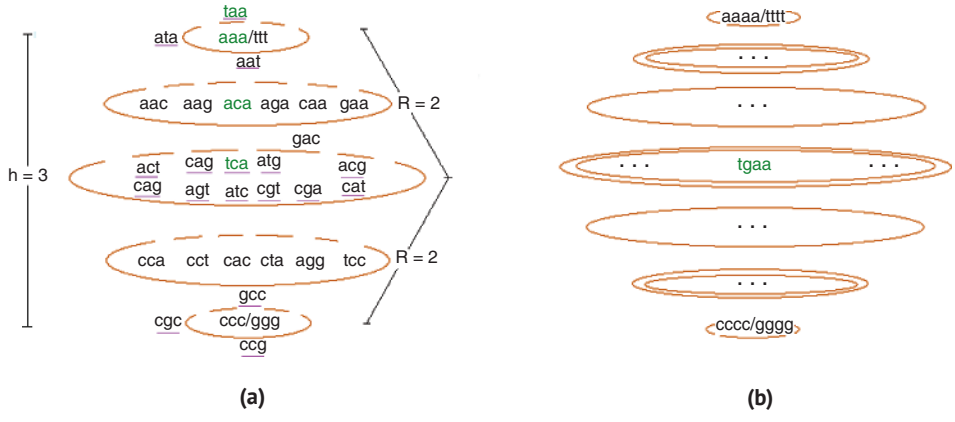

Figure 1.7 The Gibbs energies landscapes of the of DNA hybridization are not random; they exhibit interesting geometric and mathematical properties, as evidenced by the h-distance approximation in $\mathbf{D}_m$ for (a) $m = 3$; and (b) $m \geq 4$. The pmers are geometrically arranged into two isometric groups (the northern and southern hemispheres, with a North pole of a/t's, a South pole of c/g's and an equator E), reminiscent of the solar planet Saturn, but with equatorial rings not only around the equator, but alternate parallels as well. (b). This organization scales up essentially unaltered (except perhaps with more rings around more parallels) for spaces of larger length m. (Although the embeddings are not isometric, the distance in ordinary three-dimensional (3D) Euclidean space between the locations of m-pmers is roughly proportional to their actual h-distance (hybridization affinity), in the DNA spaces $\mathbf{D}_m$ (see Figure 1.6). WC-palindromes have been excluded to avoid cross(self)-hybridization.

space is indicative of their hybridization affinity. For example, a pair of WC complementary oligonucleotides (e.g. *aaa* and *ttt* for oligonucleotides of length 3) collapse into a single point (the pair *aaa*/*ttt* and so are at distance 0; likewise for *acg*/*cgt*). Such pairs are referred to as *paired mers*, or simply *pmers*. Moreover, this distance allows the study of hybridization landscapes with the geometric and spatial analogies that people use so easily to facilitate reasoning and ordinary life in the real world, as described in Section 1.2. For example, the 3-pmers *aaa*/*ttt* (*ccc*/*ggg*) are located as a sort of north N (south S, respectively) pole in the DNA space of 3-pmers, as shown in Figure 1.7a. These concepts afford a good model of the complex Gibbs energy landscapes of DNA hybridization that are fundamental to genomics and physiology in living organisms.

The precise definition of the h-distance follows the biochemistry of DNA much more closely than the Hamming distance, as shown in its operational definition 1.2 given in Figure 1.6. This h-distance is laborious to calculate manually, but can be computed for arbitrary pmers by a short computer program very quickly.

1.2.3 New Alignment-free Methods from Deep Structure

In this section, the deep structure described in Section 1.2.2 is used to introduce novel alignment-free methods to address new important biological problems in the following chapters.

1.2.3.1 Noncrosshybridizing Bases

The deep structure makes possible the selection of *universal biomarkers and genomic signatures of long DNA sequences (such as genes and even arbitrary biomarkers and genomes used in biology)* that will make possible genomic analyses of very long DNA sequences. This reduction is based on the pointwise hybridization pattern exhibited by a long sequence or a set of DNA sequences to a *common/universal judiciously selected set of DNA oligonucleotide probes located strategically on the full hybridization landscape* of a DNA space $\mathbf{D}_m$.

The standard approach in biology to extracting information from DNA sequences is a *microarray* (Schena 2003). They are planar substrates made of glass, mica, plastic, or silicon. cDNA strands (typically WC-complements of full genes or genome fragments) are attached to them to capture specific biosamples (usually referred to as probes) collected from an organism by hybridization binding. Since the 1990s, microarrays have been refined to enable analysis of genomic and metabolomic information in applications to various fields such as biology, medicine, and health.

However, microarrays have some serious drawbacks. First, analyses of their readouts give results that are *hardly reproducible* because of the high uncertainty of hybridization of probes to targets, since the sequences of the genes attached to them (probes) are unknown *a priori*. Second, the targets may cross-hybridize because they are not affixed to the chip sufficiently far apart, while the probes are floating in solution. No constraints are implemented in these chips to minimize cross-hybridization between targets (Garzon and Mainali 2017a). A second disadvantage is that they might miss target strands if they do not hybridize with any probe on the microarray and thus miss signals that could yield useful information. With the recent advances in NGS, this problem can be partially solved if genes are substituted by DNA fragments coding for single proteins. Currently, a number of NGS platforms using different sequencing technologies are available. These platforms perform sequencing of millions of small fragments of DNA in parallel. Some bioinformatic analyses join these fragments by mapping the individual reads to the human reference genome (Behjati and Tarpey 2013). However, analyzing the sequences generated using these platforms remains a big challenge due to the uncertainty of hybridization.

An alternative method has been proposed by Garzon and Mainali (2017a) that leverages the intuitive and geometric understanding of the full Gibbs energy landscapes for m-mers of a given length m afforded by its deep structure (as described in Section 1.2.2) to *obtain designs for a next-generation microarray* (NGMs), or DNA chips to address both limitations in conventional microarrays, as described next. In the DNA space $\mathbf{D}_m$, reactions conditions can be modeled by a hybridization threshold τ so that two m-mers hybridize if and only if their h-distance $|xy| < \tau$. One can use it as a set of m-mer probes B for a microarray design, a set of targets that do not cross-hybridize, i.e. they do *not* hybridize with one another under stringency τ, which amounts to a prudent separating distance in DNA space $\mathbf{D}_m$. In addition, one might position them strategically enough to select them so that *every other random m-mer does hybridize to at least one of the probes in B*, i.e. they capture the full set of m-mers by hybridization within stringency τ. Such a set of probes (if one is lucky enough to find one) will be referred to as a *noncross-hybridizing (nxh) basis* of the DNA space $\mathbf{D}_m$ of all m-pmers. One might have noticed that these pmers will be referred to as *probes* (opposite to the standard use in biology, where they are referred to as targets), while the subject of analysis (e.g. a gene, genome fragment, or possibly a full genome) will be referred to as a target.

Table 1.4 Centroidal nxh bases for larger DNA spaces.

Basis ID	τ	Length	Size p	Entropy	Probe h-census
3mE4b	1.1	3	4	0.45	[]
4mP3	2.1	4	3	0	[0 120 0 0]
5miC3Mg	3.1	5	3	0.34	[0 479 33]
7miC4Sa	4.1	7	4	0.17	[4 7997 191]
7miC4Sb	4.1	7	4	0.15	[4 8024 164]
8mP10	4.1	8	10	0.57	[]

An nxh basis could be used to achieve an exponential reduction of the dimension of very long DNA sequences to short feature vectors, the *genomic signatures* described next, in hopes that they will enable the extraction of more reliable and relevant information about long DNA sequences based on the knowledge of Gibbs energy landscapes. *This expectation is justified based on the principled design that hybridization is a local process based on interactions of short oligonucleotides DNA sequences and that the Gibbs energy is a fundamental biochemical factor in their physical process of hybridization.* Some such designs are readily available for short m-mers, but they are under further development. (How to obtain such nxh bases is a problem that will be discussed below after establishing the soundness of this approach. Readers with some knowledge of linear algebra in mathematics will not miss the resemblance between nxh basis and the concept of basis in a mathematical vector space that inspired the concept of nxh basis, with hybridization playing the role of the concept of linear combination of vectors in Euclidean spaces.)

Table 1.4 shows a number of such nxh bases along with the quantification of their quality. These bases were obtained using a judicious selection among the centroids of the parallels in $\mathbf{D}_m$ on the "earth" of DNA spaces (see Figure 1.7), a natural idea afforded by our physical intuition in ordinary Euclidean spaces.

Armed with just these few nxh bases, one can extract information as feature vectors, or genomic signatures, from target genomic sequences, as described next.

1.2.3.2 Genomic Signatures

Definition 1.3 Genomic signature (Garzon et al. 1997) For a DNA space $\mathbf{D}_m$, a nxh basis B with p probes, a h-distance threshold $\tau > 0$, the *genomic signature* of a DNA sequence x is obtained by as follows:

- Shred x to nonoverlapping fragments of size m; (any shorter leftover shreds are ignored)
- For each probe $z_i \in B$, compute the total number of shreds in x that hybridizes with z_i for the given threshold τ
- Normalize the pD vector thus obtained using the partition function (i.e. dividing each component by the total number of shreds)

Genomic signatures can be readily computed for a given DNA sequence by a short computer program (e.g. in the programming language Python). This affinity can be expressed as an pD vector, where p is the number of probes in the basis B. The simplest example of such a basis (as simplistic as it may sound) is given by $m = 1$ in DNA space $\mathbf{D}_1 = \{a/t, c/g\}$ with two 1-pmers and the nxh polar basis $1mP1 = a/t, g/c$. With $0 < \tau < 1$, the genomic signature of a long DNA sequence x is then a pair of numbers $(x_{a/t}, x_{c/g})$ counting the g, c's in x that hybridize to probes a/t and c/g, respectively, (at h-distance less than $\tau < 1$) divided by the number of nucleotides in x, i.e. the familiar AT- and GC-contents of sequence x. For longer m and larger bases B of size p (like 3mE4b2, 4mP3, or 8mP10 above), *the signatures can thus be interpreted as generalized AT/GC indices* capturing hybridization patterns in a sequence x through a strategically located array of sensors given by copies of the probes in basis B. The workflow to obtain genomic signatures (*in vitro* or *in silico*) is illustrated in Figure 1.8. Two important questions emerge. *Do these genomic signatures capture any significant biological information? Are these signatures of any practical use to address biological problems?* The following chapters will provide a resounding positive answer to this question. To begin with, one can be reassured that the design of these DNA chips relies on solid principles of physical chemistry (Gibbs energies) and on the results of quantitative assessments of the quality of an nxh basis to ensure they really resolve the two basic problems with conventional microarrays mentioned above.

There are two kinds of assessments of the quality of nxh bases. The first one is a principled inherent metric where the quality of the information extracted by these bases is quantified regardless of their application. The second one is by quantifying the quality of solution models (by standard quantitative metrics discussed below) for problems arising in applications based on the features extracted by nxh bases from genomic sequences.

The first metric quantifies the uncertainty of the hybridization process of a probe onto the basis to address the issue of unreproducibility of microarray readouts. It requires the

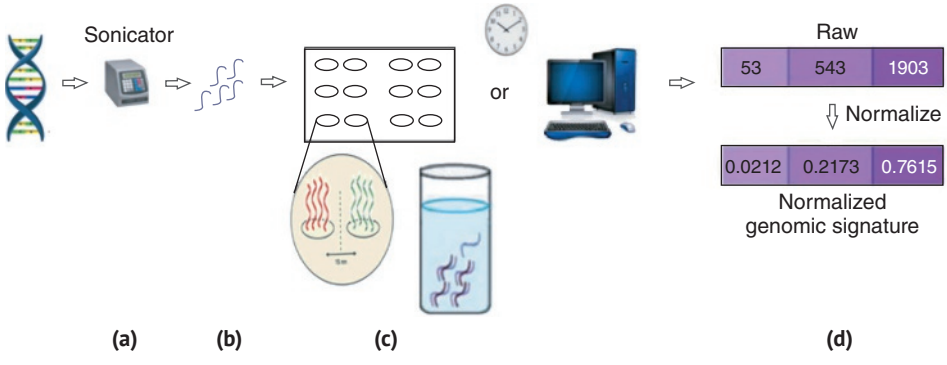

Figure 1.8 The workflow to compute a genomic signature on an nxh basis B consists of four steps: (a) Pass the DNA x through a sonicator (or computer program *in silico* to shred it into fragments (shreds) of about the same length as the probes in B; (b) Pour the shreds onto the DNA chip (next-generation microarray with sufficiently many copies of the probes and their WC-complements, properly separated on different spots to avoid cross-hybridization) and allow time for hybridization of the shreds to the probes; (c) take a census of the number x_i of hybridization to each probe i on the nxh chip (using either a gel *in vitro* or a computer program); (d) Normalize the raw counts to get the genomic signature vector of x with respect to the basis B.

standard concepts from probability theory, namely, a sample space Ω (the set of all possible outcomes of a random experiment, here drawing at random a m-pmer from space $\mathbf{D}_m$) (as a proxy for a gene fragment, for example), a (discrete) probability distribution on it, RVs (observations on all possible outcomes in Ω), and the expected value of a RV. (For background about machine learning concepts, see Appendix 10.2). The metric is the *(Shannon) entropy* quantifying the degree of uncertainty of the random process (Shannon 1948). To calculate it, it is necessary to choose an appropriate RV. *The most appropriate choice counts the number of probes in B that a random pmer hybridizes to under the given stringency* τ.

Definition 1.4 Shannon entropy Let X be RV taking on a finite number of values $x_1, x_2, \cdots, x_n$ with corresponding probabilities $P[X = x_i]$. The *Shannon entropy* $H(X)$ of X is the average uncertainty in the RV X taking any specific value given by

$$H(X) = -\Sigma_i P(X = x_i) \log(P(X = x_i)).$$

An ideal nxh basis (like B=4mP3 at stringency $\tau = 2.1$, obtained through an exhaustive search of $\mathbf{D}_4$ and shown in Table 1.4) will produce a noise-free genomic signature (Mainali et al. 2021) with $H(B) = 0$, i.e. there is no uncertainty for hybridization of any pmer to a probe. Table 1.4 shows the proposed new nxh bases having entropies less than 0.5. As the value of m increases, the value of the entropy may come closer to 0. Thus, as the length of the probes grows, the quantity of information extracted also increases for these nxh bases obtained by the centroid methods (Garzon and Mainali 2017a).

Finally, one can also perform a control for the quality of the process to obtain them, i.e. how nxh they are as bases (in terms of separation and coverage of $\mathbf{D}_n$ by all the balls of radius τ centered at them). For 6-pmers, one can choose a random set of pmers containing the same number and the length of pmers as in the nxh bases in Table 1.4, then repeat the same procedure 16 times for 6 and 7-pmers. The test is the comparison of their quality metrics (the average of the expected number of hybridizations to the given probes and their entropies) to the nxh basis. One can also perform two t-tests, each with a null hypothesis that the mean entropy of the sample is the same as the entropy of our corresponding nxh basis. In a typical run, for $\alpha = 0.05$ and one-tailed test, the critical value was 1.746. The computed t-values for two bases (11.748 for 7miC4Sb, 11.475 for 7miC4Sa, and 14.159 for 6miC4Sa) are greater than the critical value. Thus, the null hypotheses were rejected, i.e. the quality of the information extracted by these bases should be statistically significantly better than that by a random set of pmers.

Now, for these methods to be useful, one must have in hand some high-quality nxh bases. Random selection of probes is quite unlikely to produce useful such. In fact, it has been shown that the problem of finding nxh's is essentially reduced to a popular and well-researched problem in geometry, a sphere-packing problem (Garzon 2014; Garzon and Bobba 2012), already faced by Newton in the early 1700s when trying to solve the "sphere-kissing" gravitational problem of determining the maximum number of spheres of the same radius that can be placed in a kissing position in ordinary 3D space. The problem remains yet unresolved for higher-dimensional spaces today, and even for *proving* the optimality in two-dimensional space (the ordinary Euclidean plane) of arbitrary regular packings. (It took a genius of the caliber of Carl F. Gauss to give such a proof for the well-known hexagonal tesselation in the 1820s). In discrete spaces such as DNA spaces, the problem can be formulated precisely as a computational problem (the *CodeWord Design* problem – **CWD**), but it

has been established that obtaining nxh bases is very difficult in general because **CWD** has been proven to belong to a class of computational problems (**NP**-complete) that are very likely to be intractable by efficient algorithms (Garzon 2014; Phan and Garzon 2009; Garzon and Bobba 2012). Fortunately, the deep structural properties of DNA spaces (as discussed in Section 1.2.2) afford a method to obtain nxh bases of good enough quality (though not perfect) in some particular cases (for short probes), as discussed next.

1.2.3.3 Pmeric Signatures

Another family of Genomic Information Systems (GenISs) can be obtained by using a variant of genomic signatures called *pmeric (pmc) signatures* to extract information from DNA sequences. The coordinate systems, along with the means to assess their qualities, are described next. They attempt to emulate polar coordinates in DNA spaces, for which an origin (pole) is needed, namely the concept of a centroid of DNA spaces enabled by the h-metric structure. Pmeric coordinates will be alternative patterns of hybridization affinity to the set of all the h-centroids of a DNA spaces $\mathbf{D}_m$.

Since DNA spaces have a *geometric structure* determined by the h-distance metric, a question of particular interest to a biologist arises: *What pmer(or pmers) could claim the title of "origin" or "center" of* $\mathbf{D}_m$? In physics, there is the concept of the center of gravity, which is unique for a given mass distribution, such as the whole Earth. This would be a point that balances out the gravitational forces of all other point masses in the distribution. It is possible to give such a principled definition for DNA spaces by analogy, as shown next. This is another example of how a metric structure makes the analysis of DNA using known standard methods in physics possible, informative, and even easy.

An *h-centroid* of a DNA space $\mathbf{D}_m$ would be a point that is, *on average*, *closest to every other* pmer in the space than any other, in h-distance of course. Again, this average can be computed with a short computer program (perhaps in Python) for small values of m and then one can take the m-pmers that minimize the average. As it turns out, due to the symmetries of $\mathbf{D}_m$, there is no unique h-centroid minimizing this average distance, unlike on Earth, where all objects are attracted toward a unique center of mass due to gravitational forces. The h-centroids (up to $\mathbf{D}_8$) have been precomputed (not shown) by a brute-force search of the entire DNA space $\mathbf{D}_m$. As mentioned earlier, as the length of pmers increases, the size of the space explodes combinatorially. Thus, it is impossible to perform such a search beyond 8-pmers, even on a high-performance computer (HPC).

To obtain a genomic signature for a longer DNA sequence x (such as a biomarker like COI or COII serving as a proxy for an organism), a more sophisticated analysis is required. Shredding the sequence into fragments of size m will produce a "mass" distribution in a Euclidean spaces, where an ordinary centroid can then be computed as a pmeric signature. A number of interesting examples of genomic signatures are illustrated in Figure 1.8 and their distribution in Euclidean spaces in Figures 3.6–3.10 in Chapter 3.

Following the lead of genomic signatures, a Python script can be used to shred a long DNA sequence x into uniform length pmers of size m. Each pmer in $\mathbf{D}_m$ can be viewed as a point with certain weights given by its h-distances from *all* the h-centroids of $\mathbf{D}_m$. However, the number of occurrences of these pmers in genomic sequences of different organisms are likely to be different. Thus, placing masses at a pmer shred of size equal to the ratio of the total number of times the pmer occurs in x to the total number of m-pmers shreds occurring

in x will distinguish several organisms based on these vectors. These vectors will also be used as signatures for the respective organisms.

Definition 1.5 pmeric signature (Garzon and Mainali 2017b) For a DNA space $\mathbf{D}_m$ with k-centroids and a DNA sequence x of length $n > m$, the kD *pmeric signature* of x is a mD numeric vector obtained as follows:

- Shred x into nonoverlapping fragments of size m (any shorter leftover shreds are ignored)
- For each centroid $z_i \in \mathbf{D}_m$ and unique shred x_j, compute $y_{ij} = w_j|z_i x_j|$, where w_j is the fraction of the number of occurrences of x_j divided by the total number of shreds in x.
- The i^{th} component of the pmeric signature of x is given by the average of the y_{ij} across all shreds x_j.

Many examples of pmeric signatures can be found in the data shown in Chapter 3.

The same questions arise about pmeric signatures as they did above for genomic signatures. *Why would these signatures be of any use to solve biological problems?, one might ask.* Again, one could try to argue principled reasons for their effectiveness. However, an entropic quality assessment of pmeric coordinates is not possible because there is more than one centroid, and they are very close to each other in the expanse of the entire $\mathbf{D}_m$. An alternative would be to argue that pmeric coordinates determine a sequence x uniquely, at least where x is a pmer. However, this is not true even in $\mathbf{D}_3$, where two different 3-pmers exist that have identical pmeric coordinates. Thus, the question can only be resolved by trying the coordinate system to address specific problems. In fact, it will turn out that they are indeed very useful, as illustrated by a number of applications in Chapters 5 and 9. (Also, see Problems at the end of this chapter.)

1.2.3.4 Genomic Information Systems

A number of successful applications of just these two basic concepts of genomic and pmeric signatures in the following chapters naturally lead to the idea of a platform for computational solutions in biology. The idea of *GenIS* as an integrated software platform comprising a coordinate system (e.g. genomic or pmeric) was suggested by Garzon and Mainali (2017) and (Garraffoni 2019) to transform an arbitrary DNA sequence into a numeric vector and a library of conventional statistical or data science/machine learning models designed to solve biological problems using these coordinates as input features. These GenIS serves for the biome the role that an analogous Geographic Positioning System does to determine locations (e.g. cell phones) on planet Earth. Methods developed for computer networks (such as the internet and wireless communication) have enabled billions of people to communicate, e.g. using cell phones. This requires, in particular, the ability of the systems to determine a location anywhere on the planet so as to quickly establish paths to send messages through. That is similar to what biological organisms do (e.g. living cells and brains), where physical proximity, obstruction, and location amount to hard anchoring constraints that are exploited for biological function, such as cell membranes, organs, and organisms. Without them, biological reality, in particular organs and living organisms as they occur in life today, would be impossible. A GenIS offers a similar system for biological information processing

with planet Earth being replaced by the entire biome on it. The concept and development of these systems was initiated in Garzon and Mainali (2017) and has been given a number of applications in Mainali et al. (2020a,b) and Garzon et al. (2022). The following chapters will illustrate their wide applicability in biology, medicine, health, and even in other areas outside biology (e.g. image processing).

1.3 Summary

Chapter 1 has refocused the study of life to place a larger emphasis on its physical chemistry as a potential source of methods and tools to address major problems in biology. It has introduced novel tools from geometry and mathematics that enable deeper analyses of the Gibbs free-energy landscapes of hybridization, fundamental to the biochemistry of life. These models have unveiled a deep structure in these landscapes that has led to NGM designs (nxh bases and DNA chips) and will provide a handy tool to tackle key problems in biology (e.g. in genomics and evolution) through a more analytical approach. This approach will enable more predictive solutions to problems in biology in the coming chapters.

Problems

1.1 The Central Dogma (DNA-RNA-protein) is a key (though oversimplified) foundational element in molecular biology. The transcription and translation processes do not occur with 100% fidelity, so many questions remain unanswered in full.

- *What is the specific role and relative contribution of gene expression to epigenetic changes, and how precisely do these nongenetic features affect gene regulation across generations?*
- *What is the specific role and relative contribution to gene expression of microRNAs, and what molecular mechanisms underline their effect on boosting, delaying, or stopping gene expression?*
- *How can DNA repairs occur so effectively that in short periods during replication, DNA fidelity is very accurate across organisms, regardless of genome size?*
- *Will genetic advancements and breakthrough technologies make it possible (in the foreseeable future) to bring back species that have gone extinct recently or even thousands or millions of years ago, based on the remains of DNA in ancient samples? Would it require whole-genome reconstruction based on powerful predictive tools? What implications does that have for the development of science and the equilibrium of ecosystems if these species were released?*

1.2 Basic concepts such as lines and ellipses make sense in any metric space because these concepts admit definitions based on distance in Euclidean spaces. *What is the geometry of DNA spaces like?*

- *What does a "straight" line look like (by analogy to lines in Euclidean spaces where any three points are collinear)?*
  Hint: Figure 1.7a shows a full line in green color starting at *tca* going north and returning to *tca*, the codon for the amino acid Ser (*Serine*), like a diamond.

- *What do other familiar geometric figures (like ellipses and parabolas) look like in DNA spaces? Do they have any biological significance or implications?*
  Hint: Figure 1.7b shows the (largest) polar ellipse (underscored pmers in purple) defined by the north and south poles as foci and a constant distance $c = 2R = 4$ for the sum of the h-distances of a point on it from the poles.
- *Does the geometry of DNA spaces reveal anything about biological questions?*
  Hint: The concepts of malignancy of single nucleotide polymorphisms (Mainali et al. 2021) and pathogenicity of bacteria and fungi (Garzon et al. 2022) *in humans* have been directly related to this geometry.

1.3 *Is it possible to predict the pathogenicity of one organism to another, perhaps in the future, based on just their DNA?*
Hint: Section 9.1 provides references to strong evidence that this is indeed possible.

1.4 *Does the deep structure hold for RNA as well?*

Bibliography

A. Ait Saada and S.A. Lambert. Intrinsic and extrinsic determinants of telomere behavior in cells. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1868(7): 118989, 2021.

B. Alberts, A. Johnson, J. Lewis, et al. Introduction to pathogens. In *Molecular Biology of the Cell*. 4th edition. Garland Science, 2002.

O. Avery, C. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *The Journal of Experimental Medicine*, 79:137–58, 1944.

S. Behjati and P.S. Tarpey. What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, 98(6):236–238, 2013.

S.P. Bell and A. Dutta. DNA replication in eukaryotic cells. *Annual Review of Biochemistry*, 71(1):333–374, 2002.

G. Bernard, C.X. Chan, Yao-ban Chan, et al. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in Bioinformatics*, 20(2):426–435, Mar 2019. https://doi.org/10.1093/bib/bbx067.

G. Bernard, C.X. Chan, and M.A. Ragan. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Scientific Reports*, 6: 28970, 2016. https://doi.org/10.1038/srep28970.

D.L. Black. Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, 72(1):291–336, 2003.

A. Brazma, P. Hingamp, J. Quackenbush, et al. Minimum information about a microarray experiment (MIAME)–toward standards for microarray data. *Nature Genetics*, 29(4):365–371, 2001. https://doi.org/10.1038/ng1201-365.

M. Brudno, R. Steinkamp, and B. Morgenstern. The CHAOS/DIALIGN www server for multiple alignment of genomic sequences. *Nucleic Acids Research*, 32(Web Server issue): W41–W44, July 2004. https://doi.org/10.1093/nar/gkh361.

J.J. Campanella, L. Bitincka, and J. Smalley. Matgat: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics*, 4(29), 2003.

C.X. Chan, G. Bernard, O. Poirion, et al. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Scientific Reports*, 4: 6504, 2014. https://doi.org/10.1038/srep06504.

J. Chen and N.C. Seeman. DNA in a material world. *Nature*, 350(6921): 631–633, 1991.

M.O. Dayhoff. Atlas of protein sequence and structure. *National Biomedical Research Foundation*, 5:1–345, 1978. The Atlas includes detailed discussions on PAM matrices and the theoretical background of substitution matrices.

C.B. Do, M.S. Mahabhashyam, M. Brudno, et al. PROBCONS: probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330–340, 2005. https://doi.org/10.1101/gr.2821705.

S.R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*,7(10): e1002195, 2011. https://doi.org/10.1371/journal.pcbi.1002195.

R.C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

D.E. Ferrier and P.W. Holland. Ancient origin of the hox gene cluster. *Nature Reviews Genetics*, 2(1):33–38, January 2001. https://doi.org/10.1038/35047605.

R.E. Franklin and R.G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741, 1953.

M.H. Garzon. DNA codeword design: theory and applications. *Parallel Processing Letters*, 24(02): 1440001, 2014.

M.H. Garzon and K.C. Bobba. A geometric approach to Gibbs energy landscapes and optimal DNA codeword design. In *International Workshop on DNA-Based Computers*, pp. 73–85. Springer, 2012.

M.H. Garzon and S Mainali. Towards reliable microarray analysis and design. In *9th International Conference on Bioinformatics and Computational Biology, ISCA, 6p*, 2017a.

M.H. Garzon and S. Mainali. Towards a universal genomic positioning system: phylogenetics and species identification. In International Conference on Bioinformatics and Biomedical Engineering, pp. 469–479. Springer, 2017b.

M. Garzon, S. Minali, M.F. Chacon, et al. A computational approach to biological pathogenicity. Molecular Genetics and Genomics, 1(1):1741–1754, 2022.

M. Garzon, P. Neathery, R. Deaton, et al. A new metric for DNA computing. In Proceedings of the 2nd Genetic Programming Conference, vol. 32, pp. 636–638. Morgan Kaufman, 1997.

M.H. Garzon, V. Phan, and A. Neel. Optimal DNA codes for computing and self-assembly. *International Journal of Nanotechnology and Molecular Computation (IJNMC)*, 1(1):1–17, 2009.

G.K. Geiss, R.E. Bumgarner, B. Birditt, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology*, 26(3):317–325, 2008.

J. Gieseler, J.R. Gomez-Solano, A. Magazzù, et al. Optical tweezers—from calibration to applications: a tutorial. Advances in Optics and Photonics, 13(1):74–241, 2021.

M. Goodman, G.W. Moore, and G. Matsuda. Darwinian evolution in the genealogy of haemoglobin. *Nature*, 253(5493):603–608, February 1975. https://doi.org/10.1038/253603a0.

B. Haubold, F. Klötzl, and P. Pfaffelhuber. andq: a tool for alignment-free sequence comparison based on string subsampling. *PLOS ONE*,10(7): e0130143, 2015. https://doi.org/10.1371/journal.pone.0130143.

L. He, S. Sun, Q. Zhang, et al. Alignment-free sequence comparison for virus genomes based on location correlation coefficient. *Infection, Genetics and Evolution*, 96: 105106, Dec 2021. https://doi.org/10.1016/j.meegid.2021.105106.

S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89:10915–10919, 1992. https://doi.org/10.1073/pnas.89.22.10915.

A.D. Hershey and M. Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, 36:39–56, 1952.

R. Higuchi, C. Fockler, G. Dollinger, et al. Quantitative reverse transcription PCR using TaqMan and a fluorogenic probe. *Nucleic Acids Research*, 24(11):2267–2272, 1996.

D. Hnisz and R.A. Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155:934–947, 2013. https://doi.org/10.1016/j.cell.2013.09.053.

Holde. *Principles of Physical Chemistry*. Princeton: Princeton University Press, 2006.

K. Katoh, K. Misawa, K.-I. Kuma, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.

A. Kornberg. DNA replication. *Nature*, 404(6779):921–928, 2000.

K. Lange. *Mathematical and Statistical Methods for Genetic Analysis*. New York, NY: Springer New York, 2nd edition, 2002.

A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22(1):75–81, 1976.

B. Li, M. Carey, and J.L. Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, 2007. https://doi.org/10.1016/j.cell.2007.01.015.

W. Liu, Z. Fan, Y. Zhang, et al. Metagenomic next-generation sequencing for identifying pathogens in central nervous system complications after allogeneic hematopoietic stem cell transplantation. Bone Marrow Transplantation, pp. 1–6, 2021.

L. Liu, D. Li, and F. Bai. A relative Lempel-Ziv complexity: application to comparing biological sequences. *Chemical Physics Letters*, 530:107–112, March 2012. https://doi.org/10.1016/j.cplett.2012.01.061.Epub2012Feb 1.

H. Lodish, A. Berk, S.L. Zipursky, et al. *Molecular Cell Biology*. New York: W. H. Freeman, 4th edition, 2000.

H. Lodish, A. Berk, S.L. Zipursky, et al. *Molecular Cell Biology*. New York: W. H. Freeman, 7th edition, 2012.

V. Lundblad and J.W. Szostak. A mutant with a defect in telomere elongation leads to senescence in yeast. *Cell*, 57(4):633–643, 2019.

S. Mainali, M.H. Garzon, and F.A. Colorado. Profiling environmental conditions from DNA. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pp. 647–658. Springer, 2020a.

S. Mainali, M.H. Garzon, and F.A. Colorado. New genomic information systems (GenISs): species delimitation and identification. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pp. 163–174. Springer, 2020b.

S. Mainali, M. Garzon, D. Venugopal, et al. An information-theoretic approach to dimensionality reduction in data science. International Journal of Data Science and Analytics, 12(3):185–203, 2021.

D. Mapleson, G. Garcia Accinelli, G. Kettleborough, et al. KAT: a k-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, 33(4):574–576, 2017. https://doi.org/10.1093/bioinformatics/btw663.

G. Marcais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011. https://doi.org/10.1093/bioinformatics/btr011.

J.P. McCutcheon and N.A. Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1):13–26, 2012. ISSN 17401526.

F. Miescher. Ueber die chemische zusammensetzung der eiterzellen (on the chemical composition of pus cells). *Medicinisch-chemische Untersuchungen*, 4:441–460, 1871.

A.P. Monco, C.J. Bertelson, S. Liechti-Gallati, et al. An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus. *Genomics*, 2(1):90–95, 1986.

Nature Reviews Microbiology. Microbiology by numbers. *Nature Reviews Microbiology*, 9: 628, 2011. https://doi.org/10.1038/nrmicro2644.

S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970. https://doi.org/10.1016/0022-2836(70)90057-4.

C. Notredame, D.G. Higgins, and J. Heringa. T-COFFEE: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217, Sep 2000. https://doi.org/10.1006/jmbi.2000.4042.

M. O'Donnell and J. Kuriyan. DNA replication: machines at work. *Cell*, 126(5):837–840, 2006.

A.M. Olovnikov. Principle of marginotomy in template synthesis of polynucleotides. *Doklady Akademii Nauk SSSR*, 201(6):1496–1499, 1971.

J. Pellicer, O. Hidalgo, S. Dodsworth, et al. Genome size diversity and its impact on the evolution of land plants. *Genes*, 9(2), 2018. URL: https://www.mdpi.com/20734425/9/2/88.ISSN 2073-4425.

T.V. Pestova, V.G. Kolupaeva, I.B. Lomakin, et al. Molecular mechanisms of translation initiation in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 98(13):7029–7036, Jun 2001. https://doi.org/10.1073/pnas.111145798.

V. Phan and M.H. Garzon. On codeword design in metric DNA spaces. *Natural Computing*, 8(3): 571, 2009.

F.P. Roth, J.D. Hughes, P.W. Estep, et al. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, 16(10):939–945, October 1998. https://doi.org/10.1038/nbt1098-939.

A. Runehow, L. Oviedo, and N.P. Azari, editors. *Encyclopedia of Sciences and Religions*. Heidelberg: Springer, 2013.

F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, 1977.

R. Sanjuán and P. Domingo-Calap. Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, 73(23):4433–4448, 2016.

J. SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4):1460–1465, 1998.

M. Schena. *Microarray Analysis*. Wiley-Liss, 2003.

N.C. Seeman. DNA in a material world. *Nature*, 421(6921):427–431, 2003.

C.E. Shannon. A note on the concept of entropy. *Bell System Technical Journal*, 27(3):379–423, 1948.

T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981. https://doi.org/10.1016/0022-2836(81)90087-5.

N. Sonenberg and A.G. Hinnebusch. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–745, 2009. https://doi.org/10.1016/j.cell.2009.01.042.

P.L. Ståhl, F. Salmén, S. Vickovic, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. https://doi.org/10.1126/science.aaf2403.

G.D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, January 2000. https://doi.org/10.1093/bioinformatics/16.1.16.

T. Stuart, A. Butler, P. Hoffman, et al. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019. https://doi.org/10.1016/j.cell.2019.05.031.

H. Tang and B.S. Gaut. A fast implementation of pairwise statistical significance estimation for local sequence alignment. *Bioinformatics*, 26(22):2983–2985, 2010. https://doi.org/10.1093/bioinformatics/btq561.

J.M. Thomas, D. Horspool, G. Brown, et al. GraphDNA: a java program for graphical display of DNA composition analyses. *BMC Bioinformatics*, 8: 21, 2007. https://doi.org/10.1186/1471-2105-8-21.

J.D. Thompson, D.G. Higgins, and T.J. Gibson. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

D. Villar, P. Flicek, and D.T. Odom. Enhancer evolution across 20 mammalian species. *Cell*, 160:554–566, 2015. https://doi.org/10.1016/j.cell.2015.01.006.

S. Vinga and J. Almeida. Alignment-free sequence comparison—a review. *Bioinformatics*, 19(4):513–523, 2003.

Z. Wang, M. Gerstein, and M. Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. https://doi.org/10.1038/nrg2484.

J.D. Watson, T.A. Baker, S.P. Bell, et al. Molecular Biology of the Gene. Pearson, 2013.

J. Watson and F. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953. https://doi.org/10.1038/171737a0.

J.G. Wetmur. Hybridization and renaturation kinetics of nucleic acids. *Annual Review of Biophysics and Bioengineering*, 5:337–361, 1976.

J.G. Wetmur. Physical chemistry of nucleic acid hybridization. *DIMACS Series in Discret Mathematics and Theoretical Computer Science*, 48:1–23, 1999.

D.E. Wood and S.L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3): R46, 2014. https://doi.org/10.1186/gb-2014-15-3-r46.

A. Zielezinski, S. Vinga, J. Almeida, et al. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1), 2017. https://doi.org/10.1186/s13059-017-1319-7.