Michel Eduardo Beleza Yamagishi

Mathematical Grammar of Biology

Springer

Michel Eduardo Beleza Yamagishi
Laboratório de Bioinformática Aplicada
Embrapa Informática Agropecuária
Campinas, SP, Brazil

ISSN 2191-8198 ISSN 2191-8201 (electronic)
SpringerBriefs in Mathematics
ISBN 978-3-319-62688-8 ISBN 978-3-319-62689-5 (eBook)
DOI 10.1007/978-3-319-62689-5

Library of Congress Control Number: 2017946683

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

> Few philosophers would now dispute this: the nihilist or irrationalist nature of Popperite philosophy of science is by now pretty much an open secret. (David Stove [88])

I used to be uncomfortable with the post-modernist outlook on science until I read one of the books [89] by Australian philosopher David Stove (1927–1994), and figured out the source of my uneasiness. Since then, I've realized that I'd rather cling to the classical view, in which, as Erwin Chargaff wrote

> Science is the attempt to learn the truth about those parts of nature that are explorable.

Man has always explored the natural world, building what I call knowledge heritage, without putting under question his own intellectual power to do so. Knowledge belongs to humanity, and for this reason, scientists[1] must try their best to disseminate their findings. Spreading new ideas seems to be an easy task, but it cannot be forgotten that human communication is susceptible to deficiencies. Depending on how a message is conveyed, its content may be completely obfuscated. Therefore, the scientific work is not restrained to reveal nature's secrets; it also demands a proper skill to share them. Otherwise, no matter how important one finding may be, it will remain unnoticed until somebody else rediscovers it and does a better job of presenting it. The history of science has innumerable poignant and sad cases like these. However there is an even more important reason to broadly communicate new results: science, like all human activities, is not an error-free enterprise. Its outcomes must be independently double checked by peers in order to identify mistakes and correct them.

Breaking new ground is usually difficult. Arguably, it is easier to present a consolidated matter than it is to introduce a new one. Indeed, novel findings ought to first be deeply understood before there is any serious attempt to publish them. The results I will present are quite seminal, and on top of that, they are inherently multidisciplinary, which makes their presentation even harder. For this reason,

[1] And whoever explores nature.

long before writing this book, I spent a considerable amount of time reflecting on the results themselves and on the best way to present them intelligibly. It is well established that creation is not a linear process. Most scientific books show a sequence of events that does not correspond to the facts. This artifice, however, is necessary in order to make it easier for others to understand the main ideas. Albeit reluctantly, I was forced[2] to adopt the same procedure, but I do regret the unavoidable side effect of misrepresenting the actual process of scientific discovery, which is full of deadlocks, retreats, and just few true advances.

Just to give a glimpse of how involved human understanding is, as awkward as it may sound, it is possible for someone to fail to fully grasp all of the subtleties of his *own* work. Based on my experience, I used to believe that I was the only one to recognize this embarrassing truth, but, fortunately for me, there are other researchers[3] who have publicly admitted this as well. For instance, a few years ago, I read the book *A Universe from Nothing* [60]. Its subtitle *Why there is something rather than nothing?* is an old philosophical question that has been addressed by humankind's greatest minds. Lamentably, no satisfactory answer has been proposed thus far, and because Philosophy[4] is one of my deepest interests, I could not help but buy the book. The book was enjoyable and at times even funny, which allowed me to keep reading, even after "nothing" was redefined as "empty space," and "why" was replaced by "how" in the ninth chapter, which had the following title: "Nothing Is Something."[5] I suppose if Chargaff could revise it, he would produce a piece entitled *"A quick descent from Mount Olympus."* [12] Though, in my humble opinion, the book did not deliver what it promised, its reading was important to me because I unexpectedly stumbled upon the following disconcerting confession:

> Indeed, there are several of my own most important papers that I only fully understood well after the fact.

Even in Mathematics, where reason plays a major role, human understanding is not a completely rational and conscious process. Commenting on mathematician modes of thought, G.H. Hardy (1877–1947) wrote

> …that unconscious activity often plays a decisive part in discovery; that periods of ineffective effort are often followed, after intervals of rest and distraction, by moments of sudden illumination; that these flashes of inspiration are explicable only as the result of activities of which the agent has been unaware – the evidence for all this seems overwhelming. [45]

In other words, contrary to the popular belief, scientific insights may arise from pure intuition, as in the *arts*. Both Henri Poincaré (1854–1912) and Jacques

[2] I've tried several alternatives, but none were successful.

[3] Chargaff once said: "When I began to realize how unique were the regularities we had discovered, I tried, of course, to understand what it all meant, but I did not get very far." [14]

[4] Philosophy is not dead. Whoever believes the contrary has not kept up with authentic philosophy.

[5] I admit that if the book's title were "A Universe from *empty space*. *How* there is something rather than *something*?," I would have never bought it.

Hadamard (1865–1963) classified mathematicians as having either "logical" or "intuitive" minds. I'd rather represent any scientist as a *convex combination* of three components: Reason (R), Intuition (I), and Luck (L). Scientific achievements depend on all three, but I usually tell my students that only the latter is necessary and sufficient. Therefore, in my opinion, this is the main reason why, with the exception of geniuses, scientists in general should not be too proud. I would dare to represent three great mathematicians as follows:

$\mathit{Ramanujan} = 0.2R + 0.7I + 0.1L$[6];
$\mathit{Hilbert} = 0.7R + 0.29I + 0.01L$;
$\mathit{G\ddot{o}del} = 0.9R + 0.099I + 0.001L$.[7]

The weight of the R component is correlated with the skill to *consciously* grasp complex problems. I do not have the ridiculous pretension of comparing myself to these giants; nevertheless, given that every scientist may be represented in that way, I would say that $\mathit{Yamagishi} = 0.005R + 0.1I + 0.895L$.

Despite all my earnest attempts to equally address both mathematicians and non-mathematicians in this book, I'm afraid that my inborn capacities are not enough to accomplish this feat. Therefore, my fellow mathematician will probably yearn for a better exposition of some biological aspects, while the biologists reading will find the book wanting for clarity on several mathematical concepts. No writer is completely immune to this sort of compromise when addressing two or more different audiences. However, for the interested reader, the relevant bibliography on the main topics may be found in the reference section. In an attempt to facilitate comprehension, I've tried to exhibit only the aspects of the fields that have truly contributed to my scientific work.

I believe that publications like the *SBMAC SpringerBriefs* are very important. There is always trouble when it comes to bringing interdisciplinary works to press. Manuscripts on applied mathematics may be either too mathematical to be accepted by applied periodicals or too applied to be published in Mathematics journals. There is a whole spectrum within the field of Applied Mathematics, ranging from Pure Mathematics to any other natural sciences, that lack appropriate means of publication. The work presented in this book is a good example of this, for there is no *"Mathematical Grammar of Biology Journal"* yet. In the absence of interdisciplinary compendiums similar to the *SBMAC SpringerBriefs*, several original applied works are likely to have been condemned to oblivion.

Finally, I would like to pay homage to the great scientist Erwin Chargaff. I first learned of him through his scientific papers, but I soon realized that he also left

[6] Without that 0.1 of *Luck*, Hardy would have never read and realized the importance of Ramanujan's former works, and they would have never worked together to reach even more interesting results. Maybe a little bit more *Luck* would have prevented him from dying so young.

[7] G.K. Chesterton (1874–1936) once said that *"Imagination does not breed insanity. Exactly what does breed insanity is reason. Poets do not go mad; but chess-players do. Mathematicians go mad."* [17] Indeed, Gödel paid a high price for being extremely rational: he starved to death.

behind an equally important humanistic opus. Unfortunately, his books in English[8] are prematurely out of print. But I've managed to read all of the publications I could buy from used and antique booksellers, and they were worth every penny. Chargaff had a classic erudition that made him a *sui generis* scholar, as can be observed in his memoirs [14].

> I cannot serve as an example for younger scientists to follow. What I can teach cannot be learned. I have never been a "100 percent scientist". My reading has always been shamefully nonprofessional.

Highly influenced by his "former high-school teacher"[9] Karl Krauss (1874–1936), Chargaff was often satirical (a delicious example may be found in [12]), which earned him some strains in his relationships.

> Nevertheless, if at one time or another I have brushed a few colleagues the wrong way, I must apologize: I had not realized that they were covered with fur.

However, even at the risk of incomprehension, Chargaff did not hesitate to express deep thoughts through simple sentences, as in his statement:

> We posit intelligence where we deny it. We humanize things, but we reify man.

Certainly, he belonged to a rare class of Scientists. There were few before him, and there have been none like him since his death.[10]

This is the last time that the pronoun "I" will appear in this manuscript. Although I am the only author of every subjective opinion or philosophical digression in the book, part of the scientific results reported within these pages were obtained in partnership with other colleagues, namely *Alex I. Shimabukuro* and *Roberto H. Herai*. Nevertheless, all inaccuracies or gross mistakes that may have reached the final revision should be credited to me alone.

Campinas, SP, Brazil
July 29, 2017

Michel Eduardo Beleza Yamagishi

[8] "The reason why I stopped publishing in English is very simple, because I couldn't find a publisher." [47]

[9] Actually, Karl Krauss was an editor and writer. Chargaff read his texts and attended his lectures.

[10] In Chargaff's own words:

> I would say that most of the great scientists of the past could not have arisen, that, in fact, most sciences could not have been founded, if the present utility-drunk and goal directed attitude had prevailed [14].

Culture is to human beings as soil is to plants, geniuses need a rich and fecund "soil" to flourish. Perhaps, in the West, the first historical example of such a rare soil occurred in the classical Athens that nourished Socrates, Plato, and Aristotle. The last instance took place in Vienna, between the late nineteenth and early twentieth century, where great minds either were born or lived for a while. Paradoxically, those "Viennese" giants ushered *the-man-without-qualities* era. Chargaff was one of them.

Contents

Chapter 1
Introduction

Abstract We do not master DNA's language yet, but the recent biotechnological advances have enable us to edit genome sequences pretty much the way a text editor modifies written language. In this Chapter, we contextualize our results, and argue that they should be taken in account whenever Synthetic Biology creates artificial DNA sequences.

> Is the universe Gödelian in the sense that there is no end to the discovery of its laws? Perhaps. It may be that no matter how deeply science probes there will always be laws uncaptured by the theories, an endless sequence of wheels within wheels. (Martin Gardner [38])

> As the field of existence is limited and pre-occupied, it is only the hardier, more robust, better suited to circumstance individuals, who are able to struggle forward to maturity, these inhabiting only the situations to which they have superior adaptation and greater power of occupancy than any other kind; the weaker, less circumstance-suited, being prematurely destroyed. This principle is in constant action, it regulates the colour, the figure, the capacities, and instincts; those individuals of each species, whose colour and covering are best suited to concealment or protection from enemies, or defense from vicissitude and inclemencies of climate, whose figure is best accommodated to health, strength, defense, and support; whose capacities and instincts can best regulate the physical energies to self-advantage according to circumstances – in such immense waste of primary and youthful life, those only come forward to maturity from strict ordeal by which Nature tests their adaptation to her standard of perfection and fitness to continue their kind by reproduction. (Patrick Matthew [67])

> The unity of life is no less remarkable than its diversity. Most forms of life are similar in many respects. The universal biologic similarities are particularly striking in the biochemical dimension. From viruses to man, heredity is coded in just two, chemically related substances:DNA and RNA. The genetic code is as simple as universal. There are only four genetic "letters" in DNA: adenine, guanine, thymine, and cytosine. Uracil replaces thymine in RNA. The entire evolutionary development of the living world has taken place not by invention of new "letters" in the genetic "alphabet" but by elaboration of ever-new combinations of these letters. (Theodosius Dobzhansky [23])

> Never lose a holy curiosity. (Albert Einstein)

What is life? This is one of the oldest questions humanity has struggled to answer. It is likely that no other species on Earth is aware of being alive and simultaneously wonders about it. Man is definitely a different kind of animal. His intellectual powers know no limit. He has even created a science called biology to investigate life itself. However, biology's main object of study lacks definition, as Erwin Chargaff (1905–2002) noted in his statement:

> No other science deals in its very name with a subject that it cannot define.

Even worse, he pointed out that

> The realm of life has no boundaries recognizable to us except death itself.

Thus life has neither definition nor recognizable limits. Yet, man has not been intimidated by the undefinable boundlessness of life. He humbly started searching for patterns and discovered, as Theodosius Dobzhansky (1900–1975) noted, that the most recurrent pattern is the alphabet used in life's cookbook, for all *known* life forms share it. This alphabet is made up of only four letters: A, C, G and T—the nucleotides Adenine, Cytosine, Guanine and Thymine, respectively. The DNA[1] is just a long string composed of those monomers. Enclosed within cells, DNA stores the instructions necessary for every organism to thrive. The parallel with computer science is almost irresistible: every living being is a hardware[2] whose software is written in an unknown language that uses a quaternary code.

Mastering the language of DNA is no easy task, because there is no written grammar to learn from. Like children imitating their parents, we are still in the babbling stage; now and then, we recognize the meaning of and reproduce some words. Yet the efforts to learn this language have proven encouraging. For instance, we know that several DNA regions have recognizable biological functions, and that some of them are interconnected in a way that forms complex networks.[3] Notwithstanding our still flawed and incomplete knowledge on the subject, recent technological advances have created tools with a level of sophistication that allows us to easily modify the DNA sequence, pretty much the way a text editor modifies written language.[4] Of course, such a powerful tool has already some unheard-of applications. For instance, this technology has evolved so rapidly that it is now possible to convert heterozygous mutations into homozygous mutations through a mutagenic chain reaction (MCR) [36], violating Mendelian's laws of inheritance. Translating the last sentence into plain English, MCRs may be either used to cure genetic diseases, or can be adapted to drive whole populations into extinction[5] in a short time [37].

[1] If we consider viruses as a life form, then we should include RNA as well, in which case Thymine is replaced by Uracil (U).

[2] The analogy is not fully applicable. Hardwares are supposed to be unchangeable regardless their softwares; yet it is possible to change one species to another just replacing its genome [63].

[3] Biological pathways.

[4] Deleting and inserting DNA sequences wherever we want. The latest technology behind this is called CRISPR-CAS9. Originally, CRISPR-CAS9 is a prokaryotic system of defense, and it may be considered a genuine Lamarckian mechanism of evolution [58]. After been ridiculed for so long, Lamarck (1744–1829) is back [16].

[5] Regulation for such godlike power is encouraged [1].

Chargaff once said that science had fallen under “the Devil’s doctrine,” which affirms that “What can be done must be done.” He was preemptively against “genetic engineering,” which, as he explained it,

> has set out not only to alleviate human destiny, but also to improve and supplement nature, to create new forms of life that nature had not brought forth heretofore.

Chargaff was extremely worried about the danger of meddling with nature, particularly when ignorance supersedes knowledge. We fully share his concern, yet, because some scientists seem to have already fallen into the *Devil’s doctrine*, it would be wise at least to reduce the risk by expanding of the knowledge available.

The technological progress is not limited to genome editing. Even a complete chemically synthesized genome has been assembled and inserted into a bacterial cell, which started to reproduce itself.[6] More recently, in 2016, this technology matured to the point that Dr. Craig Venter’s group asserted that:

> whole genomes can now be built from chemically synthesized oligonucleotides and brought to life by installation into a receptive cellular environment [50]

The above quotation may lead to the question “have we finally created life?” No, but we’ve managed to copy its recipe with minor alterations, and we have successfully *installed* it inside a living being without killing the living being. Although no life was actually created, this is a huge and unprecedented step in history.

1.1 Synthetic Biology

The twenty-first century has witnessed the dawn of the era of Synthetic Biology. Although there are different fields of research under the label, we are referring to that which has the objective to modify existing DNA sequences or to assemble artificial[7] genomic sequences using a set of known rules and some basic building blocks. Today, it would be no exaggeration to say that creating artificial sequences or editing existing ones is within reach of anyone [62, 64]. This used to be science-fiction, but it is no more.

In order to grasp the rationale underlining Synthetic Biology, we should use a LEGO® block analogy: using simple building blocks and following certain rules of assembly, it is possible to create a lot of different toys. The analogy is not too far-fetched, and the potential applications of the Synthetic Biology are as innumerable as LEGO® toys. The whole biosphere can benefit from this power. However, the

[6] On May 13 2016, the New York Times published the article: “Scientists talk privately about creating a synthetic human genome.”

[7] “Artificial” in the sense of “not found in nature.”

LEGO® block analogy does not apply when (un)intended mistakes occur. In the case of Synthetic Biology, the consequences are proportional to its potential (i.e., limitless.)

Synthetic Biology may bring substantial progress in all areas of life sciences; yet, meddling with life's codex without due knowledge may also be threatening [9]. For this reason, and also because of the importance of respecting strict ethical principles, it would be prudent to imitate nature as much as possible. For instance, we do not know all the rules that natural sequences should comply with. Among those that we do know of, we lack a complete understanding of their role in living organisms; therefore, until DNA rules are better understood, it would be commendable to avoid building non-compliant synthetic sequences, except to investigate their fitness under strict biosecurity protocols.

1.2 DNA's First Principles

In natural languages, grammar is a set of rules used to compose utterances. However the grammatical rules are very specific to the linguistic aspect that is being considered. For example, morphology has its own set of rules, which differ from those of syntax and semantics. Analogously, we've already learned of several of DNA's "biological" rules; yet, DNA sequences have also some first principles or "intrinsic" rules, such as Chargaff's second parity rule (CSPR), which little is known about, but with which almost every natural DNA sequence complies [71].

Unfortunately, just few such rules are known. Along with CSPR, which was discovered in 1951, there is Szybalski's transcription direction rule: in 1966, Szybalski realized that mRNA-template strand tends to be pyrimidine-rich. It is also important to cite the Symmetry Principle (SP), which was established by Prabhu in 1993. The temporal sparseness of these findings alone should be enough to prove how hard it is to discover them. It is obvious that, the greater number of rules there are, the more efficient Synthetic Biology becomes, and that, consequently, fewer unintended mistakes will occur.[8] Thus, scientists in the field must focus all of their efforts on deeply studying the language of DNA. Both the biological and intrinsic rules should be conquered. Though, this challenging task seems to be the responsibility of the life sciences, given its urgency and complexity, all other disciplines should be invited to contribute as well.

[8] Avoiding mistake is one way to prevent disaster; however, little can be done to prevent an intentional catastrophe.

1.3 The Science of Patterns

In his famous book, *What is Mathematics?* [21], Richard Courant (1888–1972) wrote:

> What points, lines, numbers "actually" *are* cannot and need not be discussed in mathematical science. What *matters* and what corresponds to "verifiable" fact is structure and relationship, that two points determine a line, that numbers combine according to certain rules to form other numbers, etc.

Indeed, mathematicians are trained to recognize structure and relationship. Keith Devlin has already said that:

> Mathematics, *the science of patterns*, is a way of looking at the world, both the physical, biological, and sociological world we inhabit, and the inner world of our minds and thoughts. [22]

There is no doubt that mathematicians look the world differently. Mathematicians love to figure out hidden *patterns*. In his book *A Mathematician's Apology* from 1940, G.H. Hardy wrote that the patterns that mathematicians seek *must* be beautiful; otherwise they will not do. He declared:

> The mathematician's patterns, like the painter's or the poet's, must be beautiful, the ideas, like the colours or the words, must fit together in a harmonious way. Beauty is the first test; there is no permanent place in the world for ugly mathematics. [46]

If DNA sequences have some structure based on hidden rules, then Mathematics is the science that will be able to reveal them, or, at least, deliver the conceptual theoretical framework necessary to do so. The very term "grammar of Biology," coined by Chargaff, implicitly implies that DNA has some set of laws that it must follow. The title of this book, *Mathematical Grammar of Biology*, intends to be an invitation to all mathematicians join this quest for new rules.

This book will show that the mathematical investigation of the language of DNA has already produced results. We have developed a conceptual theoretical framework that enabled us to derive four new intrinsic DNA rules that are valid for almost every publicly available genomic sequence. In 1831, Patrick Matthew (1790–1874) wrote that nature tests organisms "adaptation to her standard of perfection and fitness." Perhaps in doing so for more than 4.280 [24] billion years, nature has been positively selecting organisms whose DNA sequences follow those rules. We are tempted to conjecture that whatever *fitness* was conferred upon the organisms, mathematical signatures (i.e., *perfection*), were also left behind as a by-product at DNA level.[9]

Consequently, in order to mimic the natural evolution of DNA sequence, we strongly suggest that the rules that we've discovered should also be imposed on artificial sequences designed by Synthetic Biology. It is true that we are still in the dark as to *if* and *why* these rules are actually important for living beings. Only

[9] Or vice-versa.

when scientific community becomes aware of their existence will those questions be answered. For instance, Dr. Craig Venter's group recently published a study in which they were trying to minimize a cellular genome. In short, they applied whole-genome design and synthesis to probe which genes were essential for life. Their final conclusion was as follows:

> The minimal cell concept appears simple at first glance but becomes more complex upon close inspection. In addition to essential and nonessential genes, there are many quasi-essential genes, which are not absolutely critical for viability but are nevertheless required for robust growth. [50]

What if this "robust growth" depends on compliance with DNA's intrinsic rules as well? Because nobody was aware of that possibility, the authors did not even consider it.[10]

1.4 The Palimpsest

Finally, are there more such rules? If this were a Gödelian universe, there would be infinitely more. Because the nature of the whole universe is not yet understood, the most we can assert so far is that, *for us*, the DNA sequences look like palimpsests. Etymologically, this word means "scraped clean and used again." However, over time, it has assumed the connotation that we employ today: "something having diverse layers." From the simplest interpretation of "layers", in which DNA is a sequence of mononucleotides, and at the next level, a sequence of dinucleotides and so on, up to the more sophisticated connotation in which the same DNA segment has multiple overlapping biological functions. If this view proves correct, then it is reasonable to speculate the existence of many unknown rules that drive and give birth to this amazing palimpsest harmony. Putting it differently, we believe that the DNA world is Gödelian.

1.5 The Alphabet

Deoxyribonucleic acid (DNA) was discovered by Johannes Friedrich Miescher (1844–1895) in 1871, more than 146 years ago [70]; however, until late 1950s, knowledge about this macromolecule and its functions in living organisms was scarce. It was known, for example, that DNA was composed of the nucleotides adenine (A), guanine (G), thymine (T), and cytosine (C); however, for many decades, it was believed that DNA was merely an unvarying and consecutive repetition of these four monomers, $(ATCG)_n$, where n indicates the number of

[10] Unfortunately, this evidence could not be assessed neither for nor against our hypothesis, because Dr. Venter's group did not make their synthetic sequences publicly available.

repetitions, which was also completely unknown. This theory was the so-called tetranucleotide hypothesis, in which DNA was considered to be one more polymer among many others. In Chargaff's words:

> The discovery of DNA by Miescher was followed soon after by the description of RNA in the laboratory of Hoppe-Seyler in Tübingen. Then began the long road – in this case nearly eighty years – which every biologically important, complicated chemical substance must travel: first its structure, then its function.

In fact, it took more 73 years for DNA's main function to be discovered by Oswald T. Avery (1877–1955) in 1944, when he published a memorable article [4], which, for the first time, revealed an association between deoxyribonucleic acid and the transmission of hereditary characteristics. He wrote:

> The evidence presented supports the belief that nucleic acid of the deoxyribose type is the fundamental unit of the transforming principle of Pneumococcus Type III.

1.6 The Birth of the Grammar of Biology

One of the people who quickly realized the potential of Avery's discovery was Chargaff. In an article commemorating the centenary of the discovery of DNA in 1971, he made the following remark about Avery's work:

> It certainly made an impression on a few, not on many, but probably on nobody a more profound one than on me. For I saw before me in dark contours the beginning of a grammar of biology. Just as Cardinal Newman in the title of a celebrated book, *The Grammar of Assent*, spoke of the grammar of belief, I use this word as a description of the main elements and principles of a science. Avery gave us the first text of a new language, or rather he showed us where to look for it. I resolved to search for this text. [13]

He was serious; he decided to abandon all that he was working on to start his own DNA research from scratch. The early years were not easy; Chargaff needed to develop more accurate methods for the chemical characterization of nucleic acids. After many years of bench work, he published his first significant results in 1950 [10, 11], which included Chargaff's first parity rule (CFPR). This rule states that in double-stranded DNA, the frequency of adenine is equal to the frequency of thymine; that is,

$$\mathbb{F}(A) = \mathbb{F}(T)$$

and the frequency of cytosine is equal to the frequency of guanine; that is,

$$\mathbb{F}(C) = \mathbb{F}(G)$$

where $\mathbb{F}$ indicates the frequency operator.

When summed up in these two equations, CFPR seems like a theoretical result without major practical consequences; however, it was actually a magnificent discovery and became one of the clues that, in addition to the X-ray images produced by Rosalind Franklin (1920–1958), resulted in the three-dimensional model of the double-helix structure of DNA [97, 98]. In Chargaff's own words:

> I believe that the double-stranded model of DNA came about as a consequence of our conversation; but such things are only susceptible of a later judgment:
>
> > *Quando Iudex est venturus.*
> > *Cuncta stricte discussurus!*
>
> When, in 1953, Watson and Crick published their first note on the double helix, they did not acknowledge my help and cited only a short paper of ours which had appeared in 1952 shortly before theirs, but not, as would have been natural, my 1950 or 1951 reviews.

The immediate consequence of the double-helix model is that DNA is formed by two complementary strands. Chargaff soon wondered what the individual properties of these strands were. He overcame the technical difficulties of his time, and in 1968, after eighteen years of study, he estimated the proportions of each nucleotide in single-stranded DNA [55]. He was surprised to find that a weaker version of CFPR, in which the equal sign $=$ gives way to the approximately equal sign $\approx$, was also valid in this case. This is known as Chargaff's second parity rule, which states that, in *single-stranded* DNA, the frequency of adenine is approximately equal to that of thymine; that is,

$$\mathbb{F}(A) \approx \mathbb{F}(T) \tag{1.1}$$

and the frequency of cytosine is approximately equal to that of guanine; that is,

$$\mathbb{F}(C) \approx \mathbb{F}(G) \tag{1.2}$$

The double-helix model fully explains CFPR; yet, there is still no such explanation for Eqs. (1.1) and (1.2).

With his studies, Chargaff definitively refuted the tetranucleotide hypothesis. He showed that the DNA sequence is more complex than initially assumed. By discovering the parity rules, he opened new horizons for DNA research.

Chapter 2
Modeling Human Nucleotide Frequencies

Abstract Fibonacci sequence is recurrent in nature. In this chapter, we model the human nucleotide frequencies through an optimization problem in which both the golden ratio and Chargaff's second parity rule play major roles.

> Mathematics is the science of patterns, and those patterns can be found anywhere you care to look for them, in the physical universe, in the living world, or even in our own minds. (Keith Devlin)

It was around 2005 that our interest in CSPR started to develop, after we, while studying human genome assembly, empirically rediscovered it. Due our lack of background in biochemistry, the result was so unexpected that we thought we'd made some mistake in the program we'd used to calculate nucleotide frequencies. Why should Eqs. (1.1) and (1.2) be valid in a single-stranded DNA sequence?[1] After double checking everything, we concluded that the result was sound, and immediately started to search in the literature to determine whether someone else had already reported it.

Until then, Chargaff was for us, an illustrious unknown.[2] After finding Chargaff's scientific papers, we were stunned to learn that he had made the same discovery in 1968. From our perspective, an attempt to explain that phenomenon was out of the question, and other studies [2, 34] had already proposed some explanations through the use of *plausible reasoning*. We've carefully chosen these two last words, and we'd like to use them as Georges Pólya (1887–1985) did [76]:

> We secure our mathematical knowledge by *demonstrative reasoning*, but we support our conjectures by *plausible reasoning*. A mathematical proof is demonstrative reasoning, but the inductive evidence of the physicist, the circumstantial evidence of the lawyer, the documentary evidence of the historian, and the statistical evidence of the economist belong to plausible reasoning. The difference between the two kinds of reasoning is great and manifold. Demonstrative reasoning is safe, beyond controversy, and final. Plausible reasoning is hazardous, controversial, and provisional.

Therefore, since we did not dare to explain it, we wondered whether the phenomenon could at least be mathematically modeled instead. Those are very

[1] To the best of our knowledge, no one knows the answer for sure yet.

[2] Unfortunately, Chargaff is not well known, even among scientists within life sciences.

different things. A mathematical model ultimately does not explain anything. If it is a good model, then it helps to predict *some* of the phenomenon's behaviors. What convinced us that nucleotide frequencies could be mathematically modeled? Morris Kline (1908–1992) revealed our major influence. He wrote:

> Greeks fashioned a conception of the universe which has dominated all subsequent Western thought. They affirmed that nature is rationally and indeed mathematically designed [56].

One could say that this conviction is one of the major principles of Western Science, for there would be no point in studying the natural world if the universe were ultimately chaotic. Science presupposes *Cosmos*[3] rather than *Chaos*. This idea was preserved and developed by the Judeo-Christian civilization as Dr. Rodney Stark, Distinguished Professor of Social Sciences at Baylor University, affirms:

> It is ideas that explain why science arose only in the west. Only Westerners thought that science was possible, that the universe functioned according to rational rules that could be discovered. We owe this belief partly to the ancient Greeks and partly to the unique Judeo-Christian conception of God as a rational creator. [87]

However, before presenting our mathematical model [102] which accurately predicts the nucleotide frequencies of the human chromosomes, we would like to summarize the major genomic breakthroughs in the last four decades. Although biologists may not need such an introduction, mathematicians and other researchers will certainly benefit from it. Because we do not intend to present an exhaustive history of the progression of genome research, we've decided to mention only three leading hallmarks: the technology used to sequence DNA, the Human Genome Project (HGP), and the ENCODE project. Each of them deserves its own book.

2.1 Sequencing DNA

The HGP would have been impossible without the work of Frederick Sanger (1918–2013). In 1977, he made a major contribution to the development of a DNA sequencing method, which is now referred to as "the Sanger method." For the first time, humanity was able to "read" a whole genome sequence.

As a matter of fact, the first organism to be sequenced had a relatively small genome. Still, the $\Phi X174$ bacteriophage's[4] 5386 nucleotides (or base pairs) represented a huge challenge in those days [81]. As Sanger wrote:

> The first DNA to be completely sequenced by the copying procedure was from bacteriophage $\Phi X174$ – a single-stranded circular DNA, 5,386 nucleotides long which codes for ten genes. The most unexpected finding from this work was the presence of "overlapping" genes. [80]

[3] From ancient Greek: order.

[4] A bacteriophage is a virus that attacks bacteria.

For this breakthrough Fred Sanger, as he is popularly known, received his *second* Nobel prize in Chemistry in 1980.

In its first version, which was limited to single-stranded DNA, the Sanger method demanded a lot of bench work (i.e., repetitive and meticulous manual tasks). As a result, only small genomes were serious candidate for sequencing. It is important to note that, since its very beginning, genomic analysis has required computer programs[5] either to perform certain critical tasks (such as compiling, editing and translating the sequence) or to search for a specific short sequence [49]. This intersection between Biology, Mathematics, Computer Science, and other fields would ultimately give rise to the field of Bioinformatics.

It did not take long for the first automated DNA sequencer to appear. It was developed in 1986. The genomic era had officially begun. For another 9 years, only viral and organelle genomes were sequenced. But in 1995, Dr. Craig Venter's group again made a significant contribution when they published the complete genomes of two bacteria: *Haemophilus influenzae* and *Mycoplasma genitalium*. At that point, the sequencing of cellular genomes became feasible. The next challenge was to read the human genome.

2.1.1 The Legacy of the Human Genome Project

The next major hallmark is certainly the sequencing of the human genome. Beginning in 1990, the The International Human Genome Sequencing Consortium[6] took more than a decade to publish the first draft and initial analysis of the human genome [51, 94]. However, it was not until 2003, when approximately 99% of the human genome's gene-containing regions were covered and close to US$ 2.7 billion dollars had been spent that the HGP was officially declared finished.

It is difficult to put the importance of the HGP into words. From a scientific standpoint, the HGP answered a lot of questions and raised many new ones. For instance, no one knew for certain the number of human genes. In 1964, without knowing of the existence of introns or large intergenic regions, Dr. Friedrich Vogel (1925–2006) estimated that the human genome had approximately 6.7 million genes [95]. Dr. Vogel himself considered this figure "disturbingly high." Over the years, the number continued to decrease. In 1990, an estimate of 100,000 genes was proposed by a joint National Institute of Health (NIH) and Department of Energy (DOE) report.[7] It reached its lowest value after the HGP was completed: 22,333 [75]. This amount is fewer than the number of grape genes.[8]

[5] Developed by Roger Staden and still in use today [86].

[6] https://www.genome.gov/10001772/all-about-the-human-genome-project-hgp.

[7] http://web.ornl.gov/sci/techresources/Human_Genome/project/5yrplan/index.shtml.

[8] Grape has about 30,343 genes.

2.1.2 The ENCODE Project

The HGP's estimate could put an end to the question of how many human genes there are. Nevertheless, with a reference human genome at hand, more sophisticated studies have resulted in new discoveries which, in turn, have brought the research back to an even more basic question: what is a *gene* in the first place? This question no longer has a universally accepted answer. In the past, "gene" used to refer to a genomic region that was transcribed and then translated into proteins. We used to believe that, in the case of humans, no more than 3% of the genome had been transcribed. In 2003, however, the National Human Genome Research Institute (NHGRI) launched a public consortium known as the "**ENC**yclopedia **O**f **D**NA **E**lements," or ENCODE. In 2012, the ENCODE published its most significant result, which was the finding that more than 80% of the human genome has been transcribed and seems to participate in at least one biochemical RNA- and/or chromatin-associated event [92]. In other words, more than 80% of the human genome seems to have some biochemical "function."[9]

Should we limit the definition of a gene to regions which produce proteins, or should we enlarge its scope to encompass all regions with some kind of biological function? Ironically, depending on the answer, Dr. Vogel's first guess would not be that wrong after all. In any case, one of the most important biological concepts should be settled sooner than later [40].

2.1.3 Pos-HGP: Next-Generation Sequencing Technologies

More importantly, the HGP was also remarkable in that it trained professionals and even created new scientific specialists. In 1998, Dr. Francis Collins and colleagues made the following statement [20]:

> The HGP has created the need for new kinds of scientific specialists who can be creative at the interface of biology and other disciplines, such as computer science, engineering, mathematics, physics, chemistry, and the social sciences. As the popularity of genomic research increases, the demand for these specialists greatly exceeds the supply. In the past, the genome project has benefited immensely from the talents of nonbiological scientists, and their participation in the future is likely to be even more crucial. There is an urgent need to train more scientists in interdisciplinary areas that can contribute to genomics. Programs must be developed that will encourage training of both biological and nonbiological scientists for careers in genomics. Especially critical is the shortage of individuals trained in bioinformatics.

It is important to note how much of what was said is still valid today, with the aggravating factor that the amount of genomic data is much higher. The advent of

[9] As every new discovery, the ENCODE project conclusions have received several critical reviews [27, 42].

Next-Generation Sequencing (NGS) technologies [69] made it possible to sequence a whole human genome for less than US$ 1000.00. Consequently, the number of organisms whose genomes have been sequenced has increased almost exponentially in recent years. The public sequence databases measure their storage capability in petabytes,[10] and will soon measure it in zettabytes.[11] In 1995, Dr. Michael S. Waterman[12] had already noticed this trend:

> The crudest measure of progress, the size of nucleic acid databases, has an exponential growth rate. Consequently, a new subject or, if that is too grand, a new area of expertise is being created, combining the biological and information sciences. Finding relevant facts and hypotheses in huge databases is becoming essential to biology. [96]

2.1.4 Multidisciplinary Approach

This colossal amount of data has imposed multidisciplinarity on the scientific community. Indeed, this worldwide enterprise has borne fruit: new genes have been discovered, and a whole new RNA-world [30] has been brought to light. An entirely new branch of life was even discovered [53]. Still, the DNA sequence has a dimension that has been almost completely neglected: its intrinsic properties.

For instance, no one knows for sure why there are so many repetitive sequences in some organisms' genomes. Almost half of the mammalian genome is composed of repetitive sequences [6]; in some plants, the proportion is even larger. Repetitive sequences, or simply "repeats," can be subdivided into those that are tandemly arrayed and those that are interspersed. Examples of the former class are microsatellites, minisatellites, and telomeres; examples of the latter class include transposable elements, or transposons. It is true that the influence of transposons present in the human germ line on gene expression can be envisaged by the fact that roughly one quarter of all analyzed human promoter regions harbor sequences derived from these elements [54]. However, most of the repetitive elements lack any recognizable biological function, and they seem to be part of what is referred to as "junk"[13] DNA.

More interesting, however, is the fact that, despite the lack of a correlation between the number of genes and complexity, there is an approximately linear correlation between genome size and the total number of DNA repetitive elements among the eukaryotes that have their genomes completed, though the contribution is more significant in larger genomes [43]. The mere existence of repeats is a riddle; their function, even more so. Could it be the case that repetitive sequences are linked to some unknown intrinsic property of DNA?

[10] 1 petabytes (PB) = 10^{15} bytes.

[11] 1 zettabyte = 10^{21} bytes.

[12] Dr. Waterman is the co-author of one of the most important algorithms in Computational Biology: the Smith-Waterman local alignment algorithm [85].

[13] The ENCODE project results call in question the concept of "junk" DNA.

These are just a few of the unanswered questions begging to be tackled. There are many more. By now, there is clearly no field of science that can study them alone. A multidisciplinary approach is essential. Biology needs Mathematics and vice-versa in order to achieve a better understanding of the language of DNA. Both sciences need Computer Science, Bioinformatics, Statistics, and other fields to contribute. It is ironic that, in order to decipher the language of DNA, scientists must first understand each other (which is no easy task). Over time, each field of science has developed its own dialect and symbolic codes. For this reason, we dare to propose the adoption of Mathematics as an Esperanto-like common language, for it provides the sharpest and most precise definitions. The next section will show how Mathematics has helped us to model human nucleotide frequencies.

2.2 Mathematical Modeling

Mathematical modeling has several different connotations, but some natural phenomena can be modeled as *Optimization Problems*. No one knows why, but it seems that nature is always optimizing itself.[14] Any optimization problem has three interconnected components: an objective function, variables, and constraints. We hold that there is no clearer or more concise presentation of those components than the one offered by Jorge Nocedal and Stephen Wright:

> Nature optimizes. Physical systems tend to a state of minimum energy. The molecules in an isolated chemical system react with each other until the total potential energy of their electrons is minimized. Rays of light follow paths that minimize their travel time.
>
> Optimization is an important tool in decision science and in the analysis of physical systems. To use it, we must first identify an objective, or a quantitative measure of the performance of the system under study. This objective could be profit, time, potential energy, or any quantity or combination of quantities that can be represented by a single number. The objective depends on certain characteristics of the system, referred to as *variables* or unknowns. Our goal is to find values of the variables that optimize the objective. Often the variables are restricted, or *constrained*, in some way. For instance, quantities such as electron density in a molecule and the interest rate on a loan cannot be negative.
>
> The process of identifying objective, variables, and constraint for a given problem is known as *modeling*. Construction of an appropriate model is the first step – sometimes the most important step – in the optimization process. If the model is too simplistic, it will not give useful insights into the practical problem, but if it is too complex, it may become too difficult to solve. [73]

In our case, all three components were present, but, lamentably, were not easily recognizable. This is the reason why mathematical modeling very often looks like *art*.

[14] “To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluos causes.” (Sir Isaac Newton) [72].

Umberto Eco's (1932–2016) book *On Beauty* [29] has a chapter entitled "From Abstract Forms to the depths of Material" in which he affirms that Michelangelo (1475–1564) used to instruct his pupils to "seek [their] statues among the stones." An even more poetic way to communicate the same sentiment is the supposedly direct quote from Michelangelo, in which he is reported to have said:

> I saw the angel in the marble and carved until I set him free.

This quote clearly reflects the similarity between the work of mathematicians and the work of artists, and it also helps us to understand why mathematicians usually have trouble explaining their ideas. Michelangelo likely experienced great difficulty in convincing his contemporaries that inside the raw stone was an angel. We face a similar challenge here. We must persuade others that human nucleotide frequencies can actually be modeled as an optimization problem.

Fortunately, Nocedal and Wright gave us excellent advice of how to proceed. First, we will show what led us to believe that nucleotide frequencies could be mathematically modeled. Then, we will present the three aforementioned components.

2.2.1 *The Line*

Arguably, the simplest way to begin any genome analysis is to perform some simple statistical measurements, including nucleotide frequencies and averages. In 2005, we tested the Chargaff's second parity rule for each one of the 24 human chromosomes ($22 + X + Y$), and it was definitively valid.

Moreover, by the very definition of the frequency, it is known that

$$\mathbb{F}(A) + \mathbb{F}(T) + \mathbb{F}(C) + \mathbb{F}(G) = 1.$$

The same equation can be rewritten with CSPR in the following way:

$$\underbrace{\mathbb{F}(A) + \mathbb{F}(T)}_{\mathbb{F}(A)\approx\mathbb{F}(T)} + \underbrace{\mathbb{F}(C) + \mathbb{F}(G)}_{\mathbb{F}(C)\approx\mathbb{F}(G)} = 1$$

Now, it is easy to derive the following equations:

$$\mathbb{F}(A) + \mathbb{F}(C) \approx \frac{1}{2} \tag{2.1}$$

or, equivalently,

$$\mathbb{F}(T) + \mathbb{F}(G) \approx \frac{1}{2} \tag{2.2}$$

or any other possible combination.[15] Note that the equal sign, $=$, has been replaced by the approximately equal sign, $\approx$.

[15] $\mathbb{F}(A) + \mathbb{F}(G) \approx \frac{1}{2}$ or $\mathbb{F}(T) + \mathbb{F}(C) \approx \frac{1}{2}$.

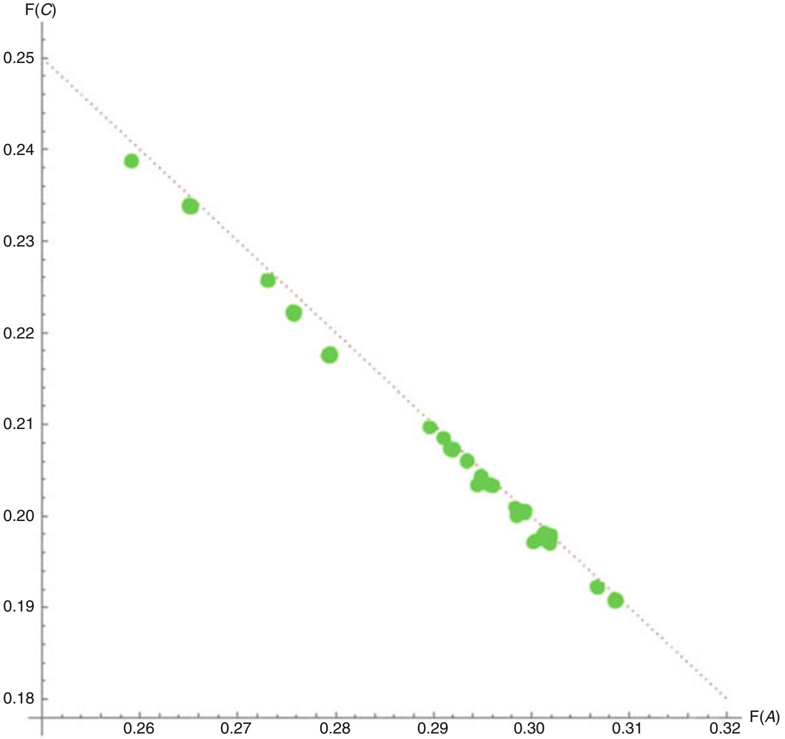


Fig. 2.1 The *red dotted line* $\mathbb{F}(A) + \mathbb{F}(C) \approx \frac{1}{2}$, and, in *green*, the observed points $(\mathbb{F}(A), \mathbb{F}(C))$ for each human chromosome (NCBI Build GRCh38.p2)

Equation (2.1) represents a line. Therefore, after plotting the points $(\mathbb{F}(A), \mathbb{F}(C))$ for each human chromosome, we discovered that they were not evenly distributed over the line:

$$\mathbb{F}(A) + \mathbb{F}(C) = \frac{1}{2}$$

An even spread would have been expected if they were randomly spread, but they seemed to be concentrated around certain points.

Figure 2.1 offers a better look into this claim. The red dotted line represents the line $\mathbb{F}(A) + \mathbb{F}(C) = \frac{1}{2}$, and, the green dots depict the Adenine and Cytosine frequencies, $(\mathbb{F}(A), \mathbb{F}(C))$, of the 24 human chromosomes.[16]

Some people may disregard Fig. 2.1 as uninteresting, but not mathematicians. Although the pattern in the green point distribution is not immediately recognizable, they are not evenly distributed, either. This dilemma could only be solved if we were able to build a mathematical model that could predict the observed frequency values.

[16] NCBI Build GRCh38.p2.

2.2.2 The Premises

In 2006, we proposed such a model [102].[17] It was simple and assumed only two reasonable premises. Moreover, it used the Fibonacci sequence and the Golden Ratio, which made it even more mathematically appealing.

The two basic premises are:

- the human nucleotide frequencies tend to *limit values* when the number of base pairs is sufficiently large.
- Chargaff's second parity rule is valid.

These two premises are deeply intertwined. The first premise requires that the nucleotide frequencies approach limit values instead of continuing to vary as the sequence grows. In fact, this assumption is always true for any finite sequence. However, our theoretical model did not require the sequence to be finite.

It should go without saying that the supposed "limit values" should be in agreement with CSPR. The problem is that short DNA sequences may not comply with CSPR. This begs the question: how short can a DNA sequence be and still satisfy CSPR?

We have performed several computational experiments in order to answer this question. One of the assays consisted of randomly selecting DNA segments of different lengths from actual genomic sequences, and then determining whether CSPR was valid. The results were not conclusive; thus, we have not yet reached a general and definitive answer. But based on our preliminary results, we at least can affirm that tens of thousands of base pairs were enough for the majority of DNA fragments to comply with CSPR. Therefore, for practical purposes, "sufficiently large" corresponds to a few tens of thousands. In the case of humans, the smallest chromosome, namely chromosome Y, has approximately 57 million base pairs, an amount which definitely satisfies this assumption.

2.2.2.1 The Golden Ratio

The next piece of information needed is the Golden Ratio. There are a lot of ways to introduce it, but we needed a way would simultaneously reveal the relationship with the Fibonacci numbers [31] and which would present a pure geometrical definition. Precisely for this reason, we will provide our own attempt to introduce it rather than directing the reader to another publication.

[17] This work took 2 years to be published.

In mathematics, one of the most famous integer sequences is, without a doubt, the sequence

$$\{1, 1, 2, 3, 5, 8, 13, \ldots\}$$

The reader should note that, beginning with 2, every Fibonacci number is the sum of the two numbers before it. Thus, $2 = 1 + 1$, $3 = 2 + 1$, $5 = 3 + 2$ and so on. Mathematically, we can state it using the following recurrence formula:

$$F(n+2) = F(n+1) + F(n), \tag{2.3}$$

together with the initial conditions $F(1) = 1$ and $F(2) = 1$.

In the West, the Fibonacci sequence was first described by Leonardo of Pisa (1170–1250), also known as Fibonacci, in his book *Liber Abaci*. The Fibonacci sequence appears in nature in different contexts: sea shell shapes, flower petals and seeds, just to name a few.

It is related to the *Golden Ratio*, ϕ, by the following limit:

$$\phi = \lim_{n\to\infty} \frac{F(n+1)}{F(n)}. \tag{2.4}$$

The Golden Ratio is associated with *Beauty* and *Perfection*, and for this reason, it is commonly found in art,[18] in music,[19] and nature, as in the case of sunflower heads (see Fig. 2.2).

The golden ratio is as old as Euclid (ca. 300 BC). He called it by a different name, the "extreme and mean ratio," and defined it as follows:

> A straight line is said to have been cut in extreme and mean ratio when, as the whole line is to the greater segment, so is the greater to the lesser.

First, it is important to note that ϕ is an irrational number. Thus, for practical purposes, musicians, painters, and artists in general use an approximate value, which, for the sake of argument, may be $\phi \approx 1.618$. In order to determine how it is applied, let us imagine that one painter should choose two quantities—length and width—for his painting. Nothing prevents him from choosing the former to be twice as long as the latter. Yet, and though no one can explain why, the artist could impress a deeper compositional effect on the observer, aesthetically speaking [68], if he chose them in the golden ratio.

Translating Euclid to modern terms, we have:

Remark 2.1 Given $a, b \in \mathbb{R}$ and $a > b > 0$, we say that a and b are in the Golden Ratio if

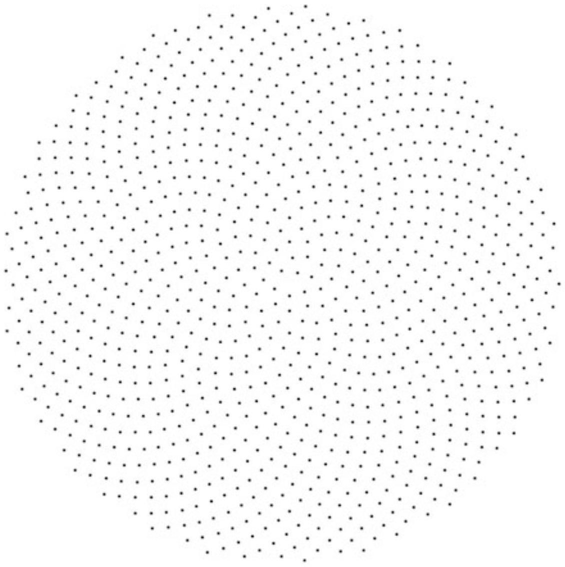
$$\frac{a+b}{a} = \frac{a}{b}$$

Keep Remark 2.1 in mind. We shall use it shortly.

[18] Leonardo da Vinci.

[19] Notably in Bartók and Debussy.

Fig. 2.2 The simulated sunflower seed pattern contains many spirals whose point coordinates are given by $\left(\sqrt{n}\cos(2\pi n\phi), \sqrt{n}\sin(2\pi n\phi)\right)$, where $n = 1, \ldots, 1000$



2.2.2.2 Chargaff's Second Parity Rule

> Scientific induction is actually the resultant of a parallelogram of rational and irrational forces. That is why in many respects Science is not a science, it is an art. (Erwin Chargaff)

Are nucleotide frequencies in the golden ratio? This is not simply a rhetorical question. Nevertheless, before the question can be considered seriously, it must be noted that the golden ratio presupposes two quantities rather than four (see Remark 2.1); thus, we must reduce the number of frequencies before trying to use the golden ratio in our model.

This obstacle is easily overcome because, in practice, CSPR reduces the independent frequencies to two, and although Eqs. (1.1) and (1.2) are the canonical form of CSPR, there are several different ways to mathematically state it. The soul of our model is here: combining CSPR and the golden ratio through mathematical formulas. This combination was achieved through two independent steps.

First Step

Remark 2.2 The division of the frequency of one nucleotide by the sum of the frequencies of the remaining nucleotides is in the proportion of three Fibonacci numbers.

Remark 2.2 is far from being CSPR. However, if we choose the three Fibonacci numbers appropriately and take only two quotients, then we get an alternative mathematical representation of CSPR.

Second Step

It is enough to select the following three Fibonacci[20] numbers and their quotients below:

$$\{F(n), F(n+1), F(n+k)\}\,, \tag{2.5}$$

It is important to note that, depending on k, the set may contain only two Fibonacci numbers. Namely, if $k = 0$ or $k = 1$, then we get the degenerate set $\{F(n), F(n+1)\}$.

Definition 2.1 We mathematically represent CSPR as:

$$\frac{\mathbb{F}(x_n)}{\mathbb{F}(y_n) + \mathbb{F}(z_n) + \mathbb{F}(w_n)} \propto \frac{F(n)}{F(n+k)}, \tag{2.6}$$

$$\frac{\mathbb{F}(y_n)}{\mathbb{F}(x_n) + \mathbb{F}(z_n) + \mathbb{F}(w_n)} \propto \frac{F(n+1)}{F(n+k)}, \tag{2.7}$$

$$\frac{\mathbb{F}(z_n)}{\mathbb{F}(x_n) + \mathbb{F}(y_n) + \mathbb{F}(w_n)} \propto \frac{F(n)}{F(n+k)}, \tag{2.8}$$

$$\frac{\mathbb{F}(w_n)}{\mathbb{F}(x_n) + \mathbb{F}(y_n) + \mathbb{F}(z_n)} \propto \frac{F(n+1)}{F(n+k)}, \tag{2.9}$$

where $k = 0, 1, 2, 3, .., N$, and $\mathbb{F}(x_n), \mathbb{F}(y_n), \mathbb{F}(z_n)$, $\mathbb{F}(w_n)$ represent the nucleotide frequencies, without any a priori association to actual nucleotides, when the number of base pairs is n, i.e.,

$$\mathbb{F}(x_n) = \frac{\hat{x}}{n}$$

where $\hat{x}$ stands for the number of nucleotide 'x'.

We acknowledge that it is not straightforward to see that Eqs. (2.6)–(2.9) are an alternative way to state CSPR. One possible attempt to grasp how the formulas above encapsulate CSPR is to notice that Eqs. (2.6) and (2.8) are proportional to the same quotient

$$\frac{F(n)}{F(n+k)}$$

[20] A particular case (i.e., "the division of the frequency of one nucleotide by the sum of the frequencies of the remaining nucleotides is in the proportion of three *consecutive* Fibonacci numbers") was originally proposed by Dr. Jean-Claude Perez in 1991 [74].

and that, similarly, Eqs. (2.7) and (2.9) are proportional to

$$\frac{F(n+1)}{F(n+k)}$$

Doubt may persist because the left sides of the equations are not as expected: they are not single nucleotide frequencies, but the quotient of nucleotide frequencies. In next section, we will show how to solve this apparent issue.

2.2.2.3 Limit Values

Now, let us impose our second assumption: that nucleotide frequencies tend to limit values when n is sufficiently large. Mathematically, it can be written as

$$\mathbb{F}(x) = \lim_{n\to\infty} \mathbb{F}(x_n) \tag{2.10}$$

$$\mathbb{F}(y) = \lim_{n\to\infty} \mathbb{F}(y_n) \tag{2.11}$$

$$\mathbb{F}(z) = \lim_{n\to\infty} \mathbb{F}(z_n) \tag{2.12}$$

$$\mathbb{F}(w) = \lim_{n\to\infty} \mathbb{F}(w_n) \tag{2.13}$$

It is also necessary to understand what happens with the quotients

$$\lim_{n\to\infty} \frac{F(n)}{F(n+k)}$$

and

$$\lim_{n\to\infty} \frac{F(n+1)}{F(n+k)}$$

Using Eq. (2.3) recursively, it is easy to get the following formula:

$$F(n+k) = F(k)F(n+1) + F(k-1)F(n). \tag{2.14}$$

We are particularly interested in the cases where n, the number of bases, is large, and where the quotient of the Fibonacci numbers tends toward a limit.

Mathematically, this case can be obtained using a few equations. Dividing (2.14) by $F(n+k)$, we get

$$1 = F(k)\frac{F(n+1)}{F(n+k)} + F(k-1)\frac{F(n)}{F(n+k)}. \tag{2.15}$$

When we take the limit as $n \to \infty$, then

$$1 = F(k) \lim_{n\to\infty} \frac{F(n+1)}{F(n+k)} + F(k-1) \lim_{n\to\infty} \frac{F(n)}{F(n+k)}, \tag{2.16}$$

However, we know that

$$\phi^{1-k} = \lim_{n\to\infty} \frac{F(n+1)}{F(n+k)} \tag{2.17}$$

and

$$\phi^{-k} = \lim_{n\to\infty} \frac{F(n)}{F(n+k)} \tag{2.18}$$

Thus, Eq. (2.16) can be written as

$$1 = F(k)\phi^{1-k} + F(k-1)\phi^{-k} \tag{2.19}$$

Finally, Eqs. (2.6)–(2.9) can be rewritten as

$$\frac{\mathbb{F}(x)}{\mathbb{F}(y) + \mathbb{F}(z) + \mathbb{F}(w)} = \phi^{1-k}, \tag{2.20}$$

$$\frac{\mathbb{F}(y)}{\mathbb{F}(x) + \mathbb{F}(z) + \mathbb{F}(w)} = \phi^{-k}, \tag{2.21}$$

$$\frac{\mathbb{F}(z)}{\mathbb{F}(x) + \mathbb{F}(y) + \mathbb{F}(w)} = \phi^{1-k}, \tag{2.22}$$

$$\frac{\mathbb{F}(w)}{\mathbb{F}(x) + \mathbb{F}(y) + \mathbb{F}(z)} = \phi^{-k}, \tag{2.23}$$

Remembering that $\mathbb{F}(x)$, $\mathbb{F}(y)$, $\mathbb{F}(z)$, and $\mathbb{F}(w)$ are frequencies, we have

$$\mathbb{F}(x) + \mathbb{F}(y) + \mathbb{F}(z) + \mathbb{F}(w) = 1. \tag{2.24}$$

Using Eq. (2.24), Eqs. (2.20)–(2.23) may be rewritten as

$$\frac{\mathbb{F}(x)}{1 - \mathbb{F}(x)} = \phi^{1-k}, \tag{2.25}$$

$$\frac{\mathbb{F}(y)}{1 - \mathbb{F}(y)} = \phi^{-k}, \tag{2.26}$$

$$\frac{\mathbb{F}(z)}{1 - \mathbb{F}(z)} = \phi^{1-k}, \tag{2.27}$$

$$\frac{\mathbb{F}(w)}{1 - \mathbb{F}(w)} = \phi^{-k}, \tag{2.28}$$

Equations (2.25) and (2.27) imply that

$$\mathbb{F}(x) = \mathbb{F}(z) \tag{2.29}$$

and, analogously, Eqs. (2.26) and (2.28) imply that

$$\mathbb{F}(y) = \mathbb{F}(w) \tag{2.30}$$

Equations (2.29) and (2.30) are CSPR, as we sought to demonstrate.

An immediate consequence of Eqs. (2.29), (2.30) and (2.24) is

$$\mathbb{F}(x) + \mathbb{F}(y) = \frac{1}{2} \tag{2.31}$$

2.2.3 Optimization Problem

Now, all of the elements necessary for revealing the optimization problem are in place. First, we have three equations in two variables[21] $\mathbb{F}(x)$ and $\mathbb{F}(y)$; namely:

$$\frac{\mathbb{F}(x)}{1-\mathbb{F}(x)} = \phi^{1-k} \tag{2.32}$$

$$\frac{\mathbb{F}(y)}{1-\mathbb{F}(y)} = \phi^{-k} \tag{2.33}$$

$$\mathbb{F}(x) + \mathbb{F}(y) = \frac{1}{2} \tag{2.34}$$

which can be rewritten as

$$\mathbb{F}(x) = \frac{\phi^{1-k}}{1+\phi^{1-k}} \tag{2.35}$$

$$\mathbb{F}(y) = \frac{\phi^{-k}}{1+\phi^{-k}} \tag{2.36}$$

$$\mathbb{F}(x) + \mathbb{F}(y) = \frac{1}{2} \tag{2.37}$$

The equations above represent a linear system. Using Eqs. (2.19) and (2.31), it is not difficult to show that the linear system is inconsistent, regardless of k.[22] In other

[21] Note that instead of denoting the variables by x and y as usual, we decided to keep the notation $\mathbb{F}(x)$ and $\mathbb{F}(y)$ just to remember that the variables represent frequencies. Please, do not interpret these notations as functions.

[22] In fact, only when $k \to \infty$ is the system consistent, but for practical purposes we are considering cases in which k is finite.

words, there is no value of k for which $\mathbb{F}(x)$ and $\mathbb{F}(y)$ satisfy all three equations at the same time. Therefore, the best we can do is to try to find an *approximative* solution through an optimization problem.

Note that Eq. (2.31) must be satisfied because $\mathbb{F}(x)$ and $\mathbb{F}(y)$ are frequencies and, by definition, Eq. (2.24) must hold. Therefore, we should try to minimize the difference

$$\left(\mathbb{F}(x) - \frac{\phi^{1-k}}{1+\phi^{1-k}}\right)$$

and the difference

$$\left(\mathbb{F}(y) - \frac{\phi^{-k}}{1+\phi^{-k}}\right)$$

under the condition that

$$\mathbb{F}(x) + \mathbb{F}(y) = \frac{1}{2}.$$

This is a classic optimization problem, and can be mathematically stated as

$$\min_{\mathbb{F}(x)+\mathbb{F}(y)=\frac{1}{2}} f_k(\mathbb{F}(x), \mathbb{F}(y)), \tag{2.38}$$

where

$$f_k(\mathbb{F}(x), \mathbb{F}(y)) = \left(\mathbb{F}(x) - \frac{\phi^{1-k}}{1+\phi^{1-k}}\right)^2 + \left(\mathbb{F}(y) - \frac{\phi^{-k}}{1+\phi^{-k}}\right)^2 \tag{2.39}$$

Given k, this minimization problem is sufficiently easy to solve, because its objective function is quadratic and the Jacobian of the constraint is full rank; therefore, the solution exists and is unique [73].

In Table 2.1, we list the solutions to the first eight values of k. It is not difficult to show that $(\mathbb{F}(x), \mathbb{F}(y)) \to (0.25, 0.25)$ as $k \to \infty$.

2.2.4 Experiment Follow-Up

More than a decade ago,[23] we performed an experiment using the available human genome data[24] to assess our mathematical model. The results were encouraging, as the observed data deviated less than 0.005 from the predicted values. However,

[23] The preprint of our manuscript has been publicly available at ArXiv since November of 2006. Cf.: http://arxiv.org/pdf/q-bio/0611041.pdf.

[24] The Human Genome sequence was downloaded from the NCBI site, and was Build 35.1.

Table 2.1 Solutions of the optimization problem for different values of k

k	$\mathbb{F}(x)$	$\mathbb{F}(x) \cong$	$\mathbb{F}(y)$	$\mathbb{F}(y) \cong$
0	$\frac{3+\sqrt{5}}{8+4\sqrt{5}}$	0.3090	$\frac{1+\sqrt{5}}{8+4\sqrt{5}}$	0.1909
1	$\frac{3+\sqrt{5}}{8+4\sqrt{5}}$	0.3090	$\frac{1+\sqrt{5}}{8+4\sqrt{5}}$	0.1909
2	$\frac{127+57\sqrt{5}}{420+188\sqrt{5}}$	0.3027	$\frac{83+37\sqrt{5}}{420+188\sqrt{5}}$	0.1972
3	$\frac{161+72\sqrt{5}}{550+246\sqrt{5}}$	0.2927	$\frac{114+51\sqrt{5}}{550+246\sqrt{5}}$	0.2072
4	$\frac{881+392\sqrt{5}}{3126+1398\sqrt{5}}$	0.2818	$\frac{682+305\sqrt{5}}{3126+1398\sqrt{5}}$	0.2181
5	$\frac{20583+9205\sqrt{5}}{75588+33804\sqrt{5}}$	0.2723	$\frac{17211+7697\sqrt{5}}{75588+33804\sqrt{5}}$	0.2276
6	$\frac{15908+7070\sqrt{5}}{59665+26683\sqrt{5}}$	0.2649	$\frac{3(9349+4181\sqrt{5})}{119330+53366\sqrt{5}}$	0.2350
7	$\frac{100793+45076\sqrt{5}}{388045+173539\sqrt{5}}$	0.2597	$\frac{186459+83387\sqrt{5}}{776090+347078\sqrt{5}}$	0.2402

we pointed out that, although the human genome project was declared finished in 2003, the sequence released included many gaps[25] and possibly misassembled regions [79] that could negatively interfere with our results. Thus, from the very beginning, we were aware that our results could change over time as more accurate assembly releases became public. We had no idea how complex the human genome actually was. For instance, a recent study sequenced 10,545 human genomes and found that, on average, each genome carries 0.7 Mbps of sequence that is not found in the reference genome [91]. If the HGP personnel knew that, they would never have decided to sequence several individuals (both male and female) to produce the reference genome. The premise was that human beings shared almost all genetic information. Unfortunately, the genome sequence varies more than initially assumed [25].[26]

Of course, this issue had an impact on our work. For instance, Fig. 2 of our former article depicted the solutions to the optimization problem as the center of red circles in which $r = 0.005$, and all of the nucleotide frequencies fell within one of the red circles. However, there was an intriguing empty red circle, meaning there was one solution to the optimization problem which had no nucleotide frequencies in its vicinity. That raised doubts about the correctness of our model. Would that empty red circle persist if more accurate sequences were available? Fortunately,

[25] The human genome is still incomplete due difficulties in cloning and assembling certain regions [32].

[26] Though, there are many genomic variations within species, the reference sequence is still a useful concept. However, novel genome projects try to use a single individual that is as homozygous as possible.

Table 2.2 Nucleotide frequencies for all human chromosomes

Chromosome	$\mathbb{F}(A)$	$\mathbb{F}(C)$	$\mathbb{F}(T)$	$\mathbb{F}(G)$	k
Chromosome 1	0.290997	0.208495	0.291759	0.208749	3
Chromosome 2	0.298448	0.200867	0.299264	0.201421	3
Chromosome 3	0.301302	0.198046	0.302045	0.198608	2
Chromosome 4	0.308621	0.190970	0.308943	0.191466	1
Chromosome 5	0.301767	0.197125	0.303167	0.197941	2
Chromosome 6	0.301891	0.197837	0.302052	0.198221	2
Chromosome 7	0.296019	0.203302	0.297002	0.203678	3
Chromosome 8	0.299337	0.200526	0.299104	0.201032	2
Chromosome 9	0.293428	0.206086	0.293823	0.206663	3
Chromosome 10	0.291726	0.207413	0.29286	0.208001	3
Chromosome 11	0.292019	0.207409	0.292575	0.207997	3
Chromosome 12	0.295700	0.203496	0.296632	0.204172	3
Chromosome 13	0.306656	0.192261	0.307841	0.193242	1
Chromosome 14	0.294518	0.203419	0.297141	0.204923	3
Chromosome 15	0.289570	0.209728	0.290094	0.210608	3
Chromosome 16	0.275750	0.222140	0.278399	0.223711	?
Chromosome 17	0.273038	0.225804	0.273812	0.227346	5
Chromosome 18	0.300301	0.197209	0.301946	0.200544	2
Chromosome 19	0.259099	0.238790	0.261514	0.240597	7
Chromosome 20	0.279422	0.217625	0.282534	0.220419	4
Chromosome 21	0.294851	0.204173	0.295770	0.205206	3
Chromosome 22	0.265135	0.233926	0.264830	0.236109	6
Chromosome X	0.301851	0.197059	0.302897	0.198194	1
Chromosome Y	0.298531	0.200121	0.301212	0.200136	2

new human genome sequences are released every so often, and we have been able to see how would our mathematical model would perform on this supposedly more complete and accurate data. The answer to this question is in Table 2.2.

As expected, there was some fluctuation in nucleotide frequencies, but they remained clustered around the predicted values. In Fig. 2.3, we have reproduced the same representation style of our former work: the solutions of the optimization problem for different values of k are depicted as red crosses, and the points $(\mathbb{F}(A), \mathbb{F}(C))$ are depicted in green for each one of the human chromosome frequencies. The dotted circles have their centers in the solutions of the optimization problem, and they have the same radius, which is equal to 0.005.

There are, nevertheless, two significant differences relative to our original experiment: (1) there is no empty circle anymore, but (2) there is one green dot (chr16) that does not belong to any circle. While the former difference is positive, the latter is worrisome. It is important to note that human chromosome 16 has a unique feature. Among all autosomal human chromosomes, chromosome 16 features one

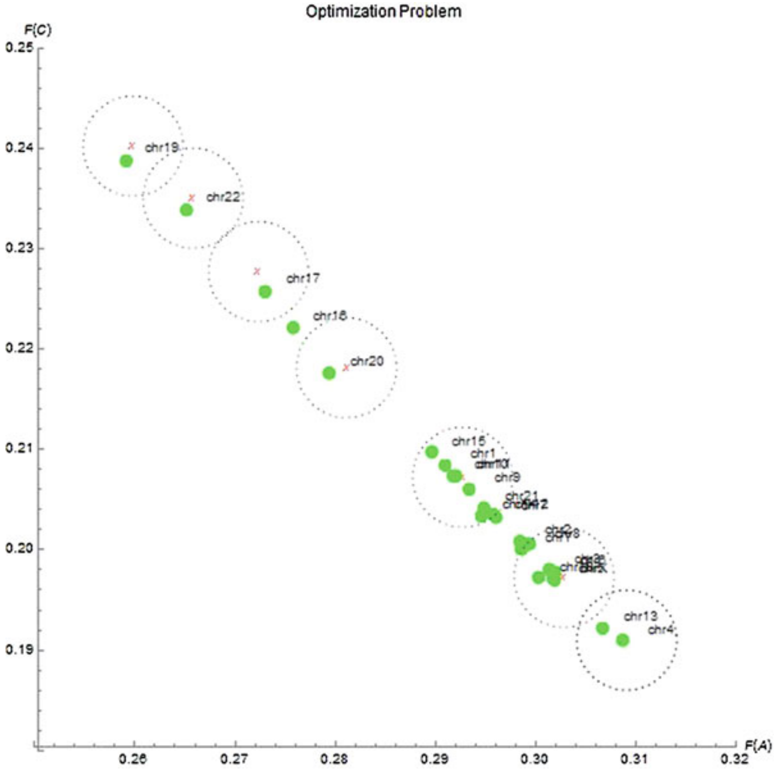


Fig. 2.3 In the figure, the solutions to the optimization problem are depicted as *red crosses*. Every *dotted circle* has the same radius ($r = 0.005$). The *green dots* show the nucleotide frequencies for each one of the human chromosomes. Unfortunately, due to their proximity, some chromosome labels are illegible. Two major differences from our former study: there is no empty circle and the chromosome 16 frequencies do not belong to any circle

of the highest levels of segmentally duplicated sequences [66]. The genomic average duplication percentage of the human chromosomes is approximately 5.3%, while that of chromosome 16 is 9.89%. This information is important because intrachromosomal duplications make sequencing and assembling issues even more complex. Chromosome 16 frequencies seem to be equidistant from two circles. Is it reasonable to wonder if chromosome 16 frequencies will fall within a circle in the future? Only more accurate human genome sequences will answer this question.

Chapter 3
Expanding the Grammar of Biology

Abstract The Symmetry Principle used to be the only generalization known for Chargaff's second parity rule. In this chapter, we present the conceptual theoretical framework used to discover four new higher order parity rules.

> Generalization is passing from the consideration of a given set of objects to that of a larger set, containing the given one. (Pólya)

> In the third book, we give an example of this in the explication of the System of the World; for by the propositions mathematically demonstrated in the former books, we in the third derive from the celestial phenomena the forces of gravity with which bodies tend to the sun and the several planets. Then from these forces, by other propositions which are also mathematical, we deduce the motions of the planets, the comets, the moon and the sea. (Sir Isaac Newton)

The link between human nucleotide frequencies and Fibonacci sequence was our first mathematical model of a DNA property. Despite its beauty, our model was limited in scope because it applied exclusively to human chromosomes and was restricted to the frequency of mononucleotides. In other words, it was an ad hoc model. Consequently, it fell far short our original aspirations.

We desired to make a significant contribution to the Chargaff's grammar of Biology. It could happen either by discovering new intrinsic rules of DNA or by building a conceptual theoretical framework that would enable further advancements. Surely those were extremely audacious goals, but science needs audacity. One cannot reach unknown places without taking unexplored paths.

The historical record of achievement was not encouraging, for all known intrinsic rules of DNA were essentially empirical findings. None of them were derived from abstract demonstrative reasoning. Over the years, we've often heard that "Biology is not an exact science" to which we've replied "yet, Biology is a science, and should therefore follow logical principles." Instead of becoming dismayed, we remained determined to reach our objectives. This was not stubbornness, but the firm conviction that Morris Kline was right when he wrote:

> Mathematics then discloses by reasoning secrets which nature may never have intended to reveal. The determination of the pattern of motion of celestial bodies, the discovery and control of radio waves, the understanding of molecular, atomic, and nuclear structures, and the creation of artificial satellites are a few basically mathematical achievements.

> Mathematical formulation of physical data and mathematical methods of deriving new conclusions are today the substratum in all investigation of nature [56].

It is true that both CSPR and the SP were susceptible to being disclosed by direct empirical inspection, but if there were more such rules, they would be harder to discover by observation alone; otherwise, they would have already been revealed, given the huge amount of data available. Perhaps, in the future, Artificial Intelligence[1] (AI) will do the job; at this point, however, we had to rely on good old-fashioned human intellectual powers to build a *mathematical formulation* that would embed the known rules and reveal new ones.

There is a popular maxim that affirms: "the right question is the first step toward solving any problem." This first step demands a deeper understanding of the problem. Only those who really grasp the subtle nuances of problems can ask the right questions. However, understanding is a necessary but not a sufficient condition. One may fully comprehend a given problem and may die without ever solving it. This is a sad truth. Even with no guarantee of success, we rummaged through the literature for leads that could help us to pose, with some luck, the right question.

3.1 The Right Question

The first hint was found as we reviewed the studies by Donald R. Forsdyke. Our attention was drawn to the *Bell and Forsdyke conjecture* [34], according to which CSPR was, in fact, a particular case of a higher-order rule. This principle was first stated as an alternative question by Forsdyke in 1995:

> Did evolutionary forces select for the Chargaff ratios in single DNA strands, with equality of complementary oligonucleotide frequencies being an automatic consequence? Alternatively, did evolutionary forces select for equality of complementary oligonucleotide frequencies, with Chargaff ratios being an automatic consequence? [33].

In 2004, Bell and Forsdyke concluded the following:

> If evolutionary forces cause a sequence to be "written" with parity at the level of single bases, then parity at the oligonucleotide level does not necessarily follow.[2] Since biological sequences show parity at the oligonucleotide level, this suggests that they were initially "written" at that level [34].

[1] The question "Can machines think?" was proposed by Alan Mathison Turing (1912–1954) in 1950. He wrote a manuscript [93] in which he devised a test called "the imitation game," and replaced the original question with "Are there imaginable digital computers which would do well in the imitation game?" On June 23, 2012, a chatbot named Eugene Goostman was the first to pass the Turing Test [83].

[2] Forsdyke proved this statement with a simple example (short DNA sequence) in which there was parity at the mononucleotide level, but no parity at the oligonucleotide level. Although his reasoning was essentially right, we would like to warn readers that long random DNA sequences with parity at the level of mononucleotides *necessarily* present parity at the oligonucleotide level. We will clarify this issue in Sect. 3.3.5.

Bell and Forsdyke knew that, in 1993, Vinayakumar V. Prabhu directly calculated oligonucleotide frequencies using genomic sequences available at the time, and discovered a generalization of CSPR, which, as mentioned previously, is known as the *Symmetry Principle*[3] (SP) [78]. This principle states that, for any given oligonucleotide, its frequency is approximately equal to its complementary reverse oligonucleotide. Mathematically, given any oligonucleotide w, the SP can be stated as:

$$\mathbb{F}(w) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(w))), \tag{3.1}$$

where $\mathscr{R}$ and $\mathscr{C}$ are the reverse and complement operators, respectively.[4] When applied to a single nucleotide, it is trivial to confirm that the SP is a generalization of CSPR.

In their own words:

> Nature " writes" with parity primarily at the oligonucleotide level. By default, there is parity at the mononucleotide level, as has been independently suggested by Baisnée et al (2002)[5] using a different approach.

There is therefore no doubt that they considered CSPR to be a consequence of the SP. This is now a time for the *imponderable*, for we had reasons to believe that there was another generalization of CSPR that had been missed completely. Just a few days before, we'd read an article written by Chargaff in 1979 [15], in which he described the relationship between nucleotide frequencies in double-stranded DNA using four identities, which, for the purposes of their adaptation to single-stranded DNA, can be rewritten as follows:

$$\mathbb{F}(A) + \mathbb{F}(G) \approx \mathbb{F}(T) + \mathbb{F}(C) \tag{3.2}$$

$$\mathbb{F}(A) \approx \mathbb{F}(T) \tag{3.3}$$

$$\mathbb{F}(C) \approx \mathbb{F}(G) \tag{3.4}$$

$$\mathbb{F}(A) + \mathbb{F}(C) \approx \mathbb{F}(T) + \mathbb{F}(G) \tag{3.5}$$

Equations (3.3) and (3.4) represent CSPR in its canonical form. It is easy to see that, if Eqs. (3.3) and (3.4) are true, then Eqs. (3.2) and (3.5) follow, and vice versa[6] In other words, we could also outline CSPR using identities (3.2) and (3.5) instead.

[3] Formal definition and discussion will be presented in Sect. 3.3.5.

[4] Formal definition in Sect. 3.2.

[5] This is article [5] in our reference list.

[6] In Chargaff's own words:

> The regularities of the composition of deoxyribonucleic-acids—some friendly people latter called them the 'Chargaff rules'—are as follows: (a) the sum of the purines (adenine and guanine) equals that of the pyrimidines (cytosine and thymine); (b) the molar ratio of adenine to thymine equals 1; (c) the molar ratio of guanine to cytosine equals 1. And, as a direct consequence of these relationships, (d) the number of 6-amino groups (adenine and cytosine) is the same as that of 6-keto groups (guanine and thymine) [14].

Therefore, there is no doubt that the SP can be considered a generalization of CSPR in relation to identities (3.3) and (3.4); however, the following question still remained:

> What would the generalization of CSPR be in relation to identities (3.2) and (3.5)?

To address this question, only Mathematics could provide the necessary conceptual and theoretical framework, for there were some critical questions that required clear concepts and an adequate theory to be properly answered. For example, Eqs. (3.2) and (3.5) consider the sum of two nucleotide frequencies on each side of the equation; in the case of more than one nucleotide,

- Which and how many oligonucleotides would be on each side of the equation?
- How many equations would there be?
- Is there one equation for dinucleotides and one for trinucleotides, and so on?

The first question is tricky, because it implicitly demands that the oligonucleotides be sorted into non-empty and disjoint subsets, which we will call partitions.[7] The number of different partitions of the oligonucleotide set is huge.

Just to give a glimpse into the complexity of the problem, consider the simplest case in which the oligonucleotides have only two nucleotides. The total number of dinucleotides is 16. How many partitions are there in a set of 16 elements? As it is well known, the number of partitions of a set depends on its cardinality, and it is given by

$$\mathscr{B}(n) = \sum_{k=0}^{n} \binom{n}{k} \mathscr{B}(k) \tag{3.6}$$

where n is the set cardinality and $\mathscr{B}(0) = \mathscr{B}(1) = 1$.

[7] Definition 3.7 in Sect. 3.3.2.

The recurrence formula given by Eq. (3.6) is called the *Bell number*. Thus, $\mathscr{B}(16) = 10{,}480{,}142{,}147$ which is a huge number considering that we have chosen the simplest case. Thus, it is not feasible to transverse the sample space even in the simplest case; therefore, there should be a "natural" set partition that could help us to answer the aforementioned questions.

3.2 Mathematical Definitions

Aristotle (384–322 BC) taught that

> The so-called Pythagoreans applied themselves to Mathematics, and were the first to develop this science; and through studying it they came to believe that its principles are the principles of everything [3].

What a bold statement! Why should the mathematical principles be the principles of everything? Were the Pythagoreans right? As the natural sciences developed, evidence supporting the Pythagoreans' claim grew. Today, there is no natural science in which Mathematics plays no role. Even in Biology, there are several examples of mathematical modeling [19]. Following this reasoning, can the DNA sequence be mathematically modeled? The short answer is yes, and in more than one way.

For our purposes, the DNA sequence may be modeled as a long string of four letters: A, C, G, and T. This small alphabet can generate words of different sizes, and produce a huge dictionary of all possible words. It is important to note that none of these words are likely to be associated with any biological functions. Our approach is theoretical in essence. The goal is only to study the properties of a subclass of these words (i.e., the subclass of words with the same number of letters). It is important to start by formally defining these concepts.

Definition 3.1 The DNA alphabet is composed of four nucleotides: Adenine, Thymine, Cytosine and Guanine. These nucleotides are presented by an alphabet made up of the letters A, T, C, and G, respectively. Therefore, the alphabet set, denoted by $\mathscr{A}$, is defined as $\mathscr{A} = \{A, C, G, T\}$.

Now, we need to define the subclass of words with the same number of letters. We will call this subclass the k-word dictionary.

Definition 3.2 Given that $\mathscr{A} = \{A, C, G, T\}$, the k-word dictionary, denoted by $\mathscr{W}^k$, is defined as the set of all words with exactly k number of letters (or nucleotides). The cardinality of $\mathscr{W}^k$, denoted by $|\mathscr{W}^k|$, is given by $|\mathscr{W}^k| = 4^k$.

Example 3.1 If $k = 2$, then $|\mathscr{W}^2| = 4^2 = 16$, namely

$$\mathscr{W}^2 = \{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT\}.$$

Example 3.2 If $k = 3$, then $|\mathscr{W}^3| = 4^3 = 64$, namely

$$\mathscr{W}^3 = \left\{ \begin{array}{c} AAA, AAC, AAG, AAT, ACA, ACC, ACG, ACT, \\ AGA, AGC, AGG, AGT, ATA, ATC, ATG, ATT, \\ CAA, CAC, CAG, CAT, CCA, CCC, CCG, CCT, \\ CGA, CGC, CGG, CGT, CTA, CTC, CTG, CTT, \\ GAA, GAC, GAG, GAT, GCA, GCC, GCG, GCT, \\ GGA, GGC, GGG, GGT, GTA, GTC, GTG, GTT, \\ TAA, TAC, TAG, TAT, TCA, TCC, TCG, TCT, \\ TGA, TGC, TGG, TGT, TTA, TTC, TTG, TTT \end{array} \right\} \tag{3.7}$$

The k-words have been referred to as k-mers in the literature [57]; however, we decided to keep the former nomenclature in order to emphasize the *grammar* analogy.

As pointed out in the very definition, the cardinality of $\mathscr{W}^k$, i.e., the number of words within $\mathscr{W}^k$, is given by $|\mathscr{W}^k| = 4^k$. Note that, even for small values of k, the cardinality of $\mathscr{W}^k$ may be huge.

For instance, when $k = 20$, we have $|\mathscr{W}^{20}| = 4^{20} = 2^{40} = 1{,}099{,}511{,}627{,}776$. There are more than 1 trillion 20-words. As a comparison, let us assume that a given human genome assembly has exactly 3.2 billion base pairs. This would imply that, in the worst-case scenario,[8] it would have $3.2 \times 10^9 - (k - 1)$ k-words. Surely, the human genome would never use all the words available in $\mathscr{W}^{20}$, because it is three orders of magnitude smaller.

Thus, what is the smallest value of k at which the human genome uses all possible words? The answer is $k = 10$, or $|\mathscr{W}^{10}| = 4^{10} = 1{,}048{,}576$. The whole human genome can be written with only 1,048,576 10-words, which gives us an idea of how many repetitive sequences exist in the human genome: on average, every human 10-word is used about 3051 times.

If we consider the human genome release known as GRCh38.p7 the occurrence distribution is far from uniform (Fig. 3.1). The most frequent 10-word is *AAAAAAAAAA* with 3,130,363 occurrences, or almost 0.1% of the genome, while there are five 10-words which appear only twice:

$$TCGCGACGTA$$

$$CGACGATCGA$$

$$TCGACGTACG$$

$$TATTCGCGCG$$

$$CGTAACGCGC$$

[8] One in which all k-words are different. We know that this not true, because there are a lot of repetitive sequences within human genome.

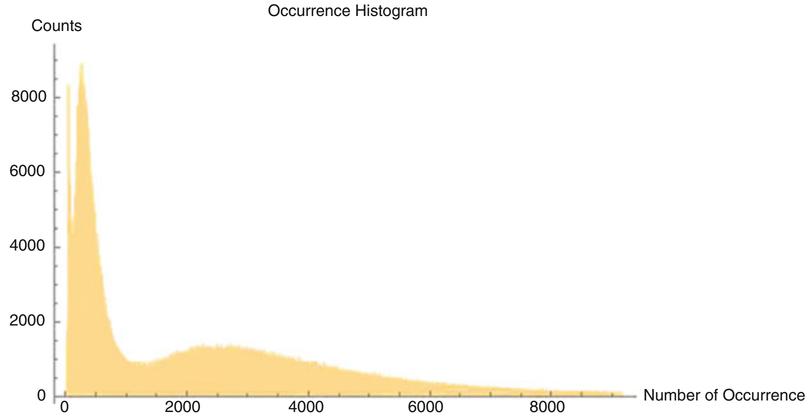


Fig. 3.1 The occurrence histogram of 10-words of the human genome release GRCh38.p7

Thus, every genome assembly should have its own k value for which the whole dictionary is used. Let us define it.

Definition 3.3 Given any genome assembly $\mathbb{G}$, let k^* denote the number of **distinct** k-words in $\mathbb{G}$. Thus, we define the function, $\mathscr{K}(k)$, as

$$\mathscr{K}(k) = \frac{k^*}{|\mathscr{W}^k|}.$$

Thus, *Completeness*, $\mathbb{K}$, is defined as the largest k such that $\mathscr{K}(k) = 1$.

One simple example should be helpful for understanding Definition 3.3.

Example 3.3 Let us assume that $\mathbb{G} = ATGATCTGTCACGAGCTA$ and $k = 2$. Therefore, there are 17 2-words in $\mathbb{G}$. However, there are only 12 distinct 2-words. Therefore,

$$\mathscr{K}(2) = \frac{12}{16} = 0.75.$$

Furthermore, $\mathbb{K} = 1$, because when $k = 1$, we have $\mathscr{K}(1) = 1$.

The most important factor, however, is that $\mathbb{K}$ is defined for a specific genome assembly. Therefore, we cannot talk about the $\mathbb{K}$ of the human genome without specifying its release. Today, there are thousands of human individuals whose genome assembly is publicly available. Although we expect that each one will have the same $\mathbb{K}$ value individually, if we could merge all of the assemblies, the resulting $\mathbb{K}$ value would probably be greater than 10 (let us say 11) due to the genomic differences between individuals. Although we do not dare to define *Completeness* for a species, we think that $\mathbb{K} + 1$ would be a good guess.

Figure 3.2 below shows the $\mathbb{K}$ distribution across the genome assemblies of the three kingdoms of life.

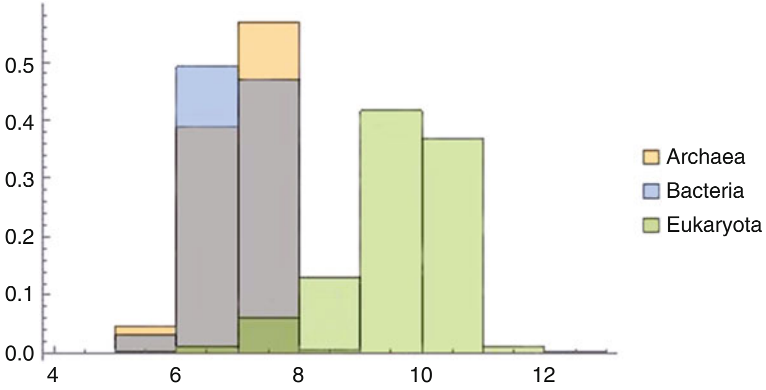


Fig. 3.2 Completeness distribution

Remark 3.1 Most organisms belonging to Archaea and Bacteria have a completeness value of $6 \leq \mathbb{K} \leq 7$. Most species in the Eukaryota kingdom have a completeness value of $9 \leq \mathbb{K} \leq 10$. Thus, as expected, eukaryotic genome assemblies have, in general, $\mathbb{K}$ values greater than those of the two other kingdoms. The smallest eukaryotic value is $\mathbb{K} = 5$, and it belongs to *Entamoeba histolytica* assembly JCVI-ESG2-1.0. The only organism with $\mathbb{K} = 12$ is *Ixodes scapularis*,[9] popularly known as the deer tick. Its assembly is *Ixodes scapularis* assembly JCVI-ISG-i3-1, It uses all 16,777,216 12-words. Biology never ceases to amaze us: it is startling that such a small parasite whose genome assembly is about 2.1 Gbps in size [44] uses more distinct words than a 3.2 Gbps human genome assembly.

3.3 Operators over $\mathscr{W}^k$

Now that we have defined the basic concepts, it is time to go a little further. Recalling Courant once more,"what matters" is the *relationship*. In this regard and from a strictly mathematical perspective, we could define a large number of operators over $\mathscr{W}^k$. However, considering our problem, there is only one that is frequently used in *Molecular Biology* and *Bioinformatics*: the complement-reverse operator. Its corresponding operation may be broken down into two independent operators: the reverse operator, $\mathscr{R}$, and the complement operator, $\mathscr{C}$. Although their definitions are fairly obvious, we provide them below for the sake of thoroughness.

Definition 3.4 Given any word $w \in \mathscr{W}^k$, the reverse of w, denoted by $\mathscr{R}(w)$ is also a word in $\mathscr{W}^k$ whose letters are w letters in the reverse order. Let $w = a_1a_2 \ldots a_{k-1}a_k$, then $\mathscr{R}(w) = a_ka_{k-1} \ldots a_2a_1$, where $a_i \in \mathscr{A}, i = 1, \ldots, k$.

[9] In North America, it is the main vector of Lyme disease.

Example 3.4 Let $w = ATGT$. Thus, $\mathscr{R}(w) = TGTA$.

Example 3.5 Let $w = AAAAAAAAAAAA$. Thus, $\mathscr{R}(w) = AAAAAAAAAAAA = w$.

Example 3.5 shows that $\mathscr{R}$ operator has *fixed-words.*[10]

Definition 3.5 Given any word $w \in \mathscr{W}^k$, the complement of w, denoted by $\mathscr{C}(w)$ is also a word in $\mathscr{W}^k$ whose letters are the complement of w. Let $w = a_1a_2 \ldots a_k$, then $\mathscr{C}(w) = \mathscr{C}(a_1)\mathscr{C}(a_2)\ldots\mathscr{C}(a_k)$ where $\mathscr{C}(A) = T$, $\mathscr{C}(T) = A$, $\mathscr{C}(C) = G$, $\mathscr{C}(G) = C$, and $a_i \in \mathscr{A}, i = 1, \ldots, k$.

Definition 3.5 introduces the biological concept of *base-pairing*. Because the molecular structure of DNA is a double helix, every nucleotide has a corresponding pair. Thus, A pairs with T, and C pairs with G. The explanation is ultimately physical and chemical in nature. Through biochemical assays and deep reasoning, Chargaff was the first person to discovery this property of DNA. He described this magical moment as follows:

> One late afternoon, while sitting at the desk in my narrow tube of an office on the fifth floor of the medical school, I asked myself: "What would happen if I assume that DNA contains equal quantities of purines and pyrimidines?" I took all the data we had on the molar proportions of adenine and guanine and of cytosine and thymine and corrected each set of give a total of 50 percent: there emerged—like Botticelli's Venus on the shell, though not quite as flawless—the regularities that I then used to call **the complementarity relationships** and that are now known as **base-pairing** [14].

Example 3.6 Let $w = ATGT$. Thus, $\mathscr{C}(w) = TACA$.

Example 3.7 Let $w = ATCGGCTAAA$, Thus, $\mathscr{C}(w) = TAGCCGATTT$.

3.3.1 *The Complement-Reverse Operator*

The complement-reverse operator is defined as the composite operator

$$\mathscr{C} \circ \mathscr{R} = \mathscr{C}(\mathscr{R}(w))$$

or

$$\mathscr{R} \circ \mathscr{C} = \mathscr{R}(\mathscr{C}(w))$$

over a k-word w.

It follows from their very definitions that

$$\mathscr{C}(\mathscr{R}(w)) = \mathscr{R}(\mathscr{C}(w)),$$

so we will use the two interchangeably.

[10] Definition 3.10 in Sect. 3.3.2.

In many Molecular Biology writings, it is customary to find the adjectives "reverse" or "complement" to refer to the complement-reverse of a DNA sequence. The context is usually sufficient to avoid any ambiguity.

Returning to the properties of the operators, it is important to note that both operators are *Involutions*.

Definition 3.6 An operator is an involution when its inverse is the operator itself. Mathematically,

$$\mathscr{C}(\mathscr{C}(w)) = w$$

and

$$\mathscr{R}(\mathscr{R}(w)) = w.$$

Example 3.8 Let $w = ATC$. Thus, $\mathscr{C}(\mathscr{C}(ATC)) = ATC$, and $\mathscr{R}(\mathscr{R}(ATC)) = ATC$.

3.3.2 Induced k-Word Set Partition

One of the goals that led us to start this study was the desire to discover how to sort the oligonucleotides into non-empty disjoint subsets in order to sum them up. Obviously, this task of assigning an oligonucleotide to a particular group cannot be arbitrary; otherwise, the number of possibilities would be prohibitive (*Bell number*). Thus, we need to single out a criterion to separate the oligonucleotides into classes.

Perhaps one example could help us to grasp the idea. How could we sort the set containing all human beings? There are several possibilities. Without loss of generality, we could put all the individuals that share the same birthday (considering only month and day) into the same subset. Thus, there would be 366 such subsets.[11] Moreover, since human beings are born only once, the subsets are disjoint, and because there are approximately seven billion human beings, it is almost certain that the subsets will be non-empty as well. Now, consider the following problem: given two human beings, say X and Y, how could we tell whether they belonged to the same subset? In this scenario, it would be enough to answer the question: *"Does X have the same birthday as Y?"* If the answer is yes, then they belong to the same subset; otherwise, they belong to different subsets. The key point here is to establish a relationship between X and Y, which, in our simple example, was "has the same birthday as."

The above example is simple, but it illustrates our true challenge very well: to uncover the relationship between k-words in order to sort them out into disjoint subsets. The decomposition of a set into non-empty disjoint subsets has a special name in Mathematics: *partition*.

[11] Don't forget February 29.

Definition 3.7 A partition of a set $\mathscr{X}$ is its decomposition into disjoint and non-empty subsets $\mathscr{X}_i$, where $i = 1, \ldots, n$ and $\mathscr{X} = \bigcup_{i=1}^{n} \mathscr{X}_i$

Let us consider some examples.

Example 3.9 Let $\mathscr{X} = \left\{ \begin{array}{c} AA, AC, AG, AT, \\ CA, CC, CG, CT, \\ GA, GC, GG, GT, \\ TA, TC, TG, TT \end{array} \right\}$. One possible set partition of $\mathscr{X}$, denoted by $\mathscr{P}$, is

$$\mathscr{P} = \left\{ \begin{array}{c} \{AA, AC, AG, AT\}, \\ \{CA, CC, CG, CT\}, \\ \{GA, GC, GG, GT\}, \\ \{TA, TC, TG, TT\} \end{array} \right\}$$

The partition $\mathscr{P}$ divides the set $\mathscr{X}$ into four disjoint (the intersection is the empty set) and non-empty subsets. However, there are several other partitions of $\mathscr{X}$. Just to illustrate another valid alternative,

$$\mathscr{P}' = \left\{ \begin{array}{c} \{AA, CC, TT, GG\}, \{AC, CA\}, \\ \{GA, AG\}, \{TA, GC, CG, AT\}, \\ \{TC, CT\}, \{TG, GT\} \end{array} \right\}$$

As Example 3.9 shows, a single set may have several distinct partitions. As we said earlier, the number of partitions of a set $\mathscr{X}$ depends on the number of elements in $\mathscr{X}$, and it is given by the *Bell number*, denoted by $\mathscr{B}$.

Example 3.10 When $k = 3$, the well-known *codon table* is actually one of the $\mathscr{B}(64) \approx 10^{66}$ partitions[12] of the 64 trinucleotides. In this particular case, each subset of trinucleotides either codes for an amino acid or for the stop signal.

What relationship between the k-words may induce a set partition? There are several such relationships. However, it just so happens that the $\mathscr{R}$, $\mathscr{C}$ and $\mathscr{C} \circ \mathscr{R}$ operators may actually be used to establish such relationship between k-words.

[12] As a comparison, the number of atoms in the “observable” universe is estimated to be about 10^{80}.

3.3.3 First Insight[13]

Definition 3.8 We say that two words $w_i, w_j \in \mathscr{W}^k$ belong to the same subset or *Equivalence Class* (EC), if, and only if, one of them is obtained from the other through the operators $\mathscr{R}$, $\mathscr{C}$, or any of their composite operators.

The term *Equivalence Class* is actually very appropriate. It means that the elements of that subset are *equivalent* to each other with regards to a given relationship. In this case, the equivalence relationship is given by the operators $\mathscr{R}$, $\mathscr{C}$, as well as by any of their composite operators. It is similar to the case of our example relation of "has the same birthday as," but stated mathematically.

However, in our example, the "has the same birthday as" relation only indicates whether two elements belong to the same EC. Now, consider a different problem. Suppose that you know John was born on July 29. With that information alone, you would not be able to find the other members of this EC. It would be great if given any element from an EC, we could use it to find all of the other elements. It just so happens that the equivalence operators defined in Definition 3.8 have this marvelous additional property. They not only say whether two elements belong to the same EC, but also that all of the elements of an EC can be found if any of its elements are known. In Mathematics, we say that the ECs are *closed* under those operators. Let us define a closed set.

Definition 3.9 A set is *closed* under an operator if that operation returns an element of the set when evaluated on elements that belong to the set.

Our oligonucleotide ECs are closed under the operators $\mathscr{C}$, $\mathscr{R}$, and, without loss of generality, $\mathscr{C} \circ \mathscr{R}$. We can explore this property.

Let us give some examples.

Example 3.11 If $k = 1$, then the 1-word set partition induced by $\mathscr{R}$, $\mathscr{C}$ and $\mathscr{R} \circ \mathscr{C}$ is

$$\mathscr{P}^1 = \{\{A, T\}, \{C, G\}\}.$$

This partition has only two ECs. Note that A and T belong to same EC because $\mathscr{C}(A) = T$ and $\mathscr{C}(T) = A$. The same can be said about the class $\{C, G\}$. The $\mathscr{R}$ operator plays no rule when $k = 1$.

[13] This word in English has several different connotations. Here, it has a special meaning coined by Dr. Bernard Joseph Francis Lonergan (1904–1984) in his philosophical masterpiece "Insight: A Study of Human Understanding [65]," in which term means "act of understanding."

Example 3.12 If $k = 2$, then the 2-word induced set partition is

$$\mathscr{P}^2 = \left\{ \begin{array}{c} \{AA, TT\}, \{AT, TA\}, \\ \{CC, GG\}, \{CG, GC\}, \\ \{AC, CA, TG, GT\}, \\ \{AG, GA, CT, TC\} \end{array} \right\}$$

In this example, there are six subsets, and it is clear that all three operators are needed. For instance, $\mathscr{R}(AC) = CA$, $\mathscr{C}(AC) = TG$, $\mathscr{R}(\mathscr{C}(AC)) = GT$ and, of course, $\mathscr{C}(\mathscr{C}(AC)) = AC$. We have chosen the 2-word AC to generate the other words, but this choice is *arbitrary* in the sense that any word in this same EC would do as well.

Another important piece of information is that some equivalence classes have only two elements instead of four. This happens because, for the first time, fixed-words appear, the definition of which is as follows:

Definition 3.10 Given any operator $\mathscr{F}$ over $\mathscr{W}^k$, a fixed-word of the operator $\mathscr{F}$ is defined as the word $w \in \mathscr{W}^k$, where $\mathscr{F}(w) = w$.

As we stated earlier, the $\mathscr{R}$ operator has fixed words, as does the $\mathscr{C} \circ \mathscr{R}$ operator when k is even. For example,

$$\mathscr{C}(\mathscr{R}(AT)) = AT, \ \mathscr{C}(\mathscr{R}(TA)) = TA, \ \mathscr{C}(\mathscr{R}(GC)) = GC$$

and

$$\mathscr{C}(\mathscr{R}(CG)) = CG.$$

Therefore, $\{AT, TA, GC, CG\}$ are fixed words of the operator $\mathscr{C} \circ \mathscr{R}$, and

$$\mathscr{R}(AA) = AA, \ \mathscr{R}(TT) = TT, \ \mathscr{R}(CC) = CC$$

and

$$\mathscr{R}(GG) = GG.$$

Thus, $\{AA, TT, CC, GG\}$ are fixed-words of the operator $\mathscr{R}$.

Remark 3.2 The operator $\mathscr{C}$ has no fixed-words.

Remark 3.2 is an idiosyncrasy of operator $\mathscr{C}$ that will prove important in the future.

Example 3.13 If $k = 3$, then the 3-word set partition is

$$\mathscr{P}^3 = \left\{ \begin{array}{c} \{AAA, TTT\}, \{AAT, ATT, TAA, TTA\}, \\ \{TTG, CAA, AAC, GTT\}, \{CTT, AAG, GAA, TTC\}, \\ \{ATA, TAT\}, \{ATC, GAT, CTA, TAG\}, \\ \{ATG, CAT, GTA, TAC\}, \{ACA, TGT\}, \\ \{TGA, TCA, AGT, ACT\}, \{CCA, TGG, ACC, GGT\}, \\ \{GCA, TGC, ACG, CGT\}, \{TCT, AGA\}, \\ \{GCT, AGC, TCG, CGA\}, \{AGG, CCT, GGA, TCC\}, \\ \{CAC, GTG\}, \{CAG, CTG, GAC, GTC\}, \\ \{CTC, GAG\}, \{CCC, GGG\}, \\ \{GCC, GGC, CCG, CGG\}, \{GCG, CGC\} \end{array} \right\}$$

This partition has 20 subsets.[14]

In the general case, the cardinality of $\mathscr{P}^k$ is given by

$$|\mathscr{P}^k| = 2^{k-1} + 4^{k-1}.$$

Thus, $|\mathscr{P}^1| = 2$, $|\mathscr{P}^2| = 6$, $|\mathscr{P}^3| = 20$, and so on. This number increases exponentially. For instance, $|\mathscr{P}^{10}| = 262{,}656$.

3.3.4 Generating Set

Now it is time to explore the fact that our ECs are closed under the equivalence operators. In Example 3.12, we mentioned that one single k-word is sufficient to generate all other k-words in the same EC. In other words, only one k-word is needed to represent its class. We will call the collection of these representative k-words a *Generating Set*.

Definition 3.11 A set containing exactly one k-word of each equivalence class is called a *Generating Set* (GS).

From the very definition of a GS follows that *it is not unique*. For instance, there are 4,294,967,296 different generating sets for $\mathscr{P}^3$. Because there are no selection criteria, the choice of the GS elements is completely arbitrary from a mathematical point of view. It goes without saying that this factor does not exclude the existence of selection criteria of another nature.

Let us consider some examples.

[14] The same number of amino acids.

Example 3.14 $\mathscr{G}^2 = \{AA, AT, CC, CG, AC, AG\}$ is a GS of $\mathscr{P}^2$, as is $\mathscr{G}^{2'} = \{TT, TA, GG, GC, CA, GA\}$ or any set satisfying Definition 3.11.

Example 3.15 If $k = 3$,then

$$\mathscr{G}^3 = \left\{ \begin{array}{c} AAA, AAT, TTG, CTT, ATA, ATC, ATG, ACA, TGA, CCA, \\ GCA, TCT, GCT, AGG, CAC, CAG, CTC, CCC, GCC, GCG \end{array} \right\}$$

is a GS of $\mathscr{P}^3$.

One obvious application of the concept of GS is the representation of the $\mathscr{W}^k$. Without loss of generality, let us assume that $k = 3$; therefore, $|\mathscr{W}^3| = 64$. However, it is not necessary to store all 64 elements in order to represent $\mathscr{W}^3$. It is enough to store any one of its GSs, which each contain only 20 elements. Thus, we get a reduced representation of any $\mathscr{W}^k$.

Remembering that $|\mathscr{P}^k| = 2^{k-1} + 4^{k-1}$ and $|\mathscr{W}^k| = 4^k$, we can show that

$$\lim_{k\to\infty} \frac{|\mathscr{P}^k|}{|\mathscr{W}^k|} = 0.25$$

which means that our compression power tends to move toward one fourth as k increases.

3.3.5 The Riddle of the Symmetry Principle

As we anticipated in Chap. 3, Dr. Vinayakumar V. Prabhu [78] used computational and statistical analyses in 1993 to discover the following generalization of CSPR into k-words:

Definition 3.12 The Symmetry Principle (SP) is given by $\mathbb{F}(w) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(w)))$, where $w \in \mathscr{W}^k$.

The SP states that the frequency of a k-word w is approximately equal to its complement-reverse k-word frequency. Therefore, in a sufficiently long sequence, the frequencies of w and $\mathscr{C}(\mathscr{R}(w))$ should be approximately the same. No one knows why. It doesn't have to be, but it is.

Although the SP seems to be counterintuitive, it is, actually, expected by chance alone in random DNA sequences, *provided* that the mononucleotide frequencies follow CSPR.[15] Indeed, in these cases, the property is even stronger: a full equifrequency is observed in each equation [26]:

$$\boxed{\mathbb{F}(w) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(w))) \approx \mathbb{F}(\mathscr{C}(w)) \approx \mathbb{F}(\mathscr{R}(w))}$$

[15] We will refer to these sequences as *Semi-random*.

It is not difficult to grasp why. Let's assume that a random DNA sequence follows CSPR, such as $\mathbb{F}(A) \approx \mathbb{F}(T)$ and $\mathbb{F}(C) \approx \mathbb{F}(G)$. The expected frequency value of any k-word within this sequence is the product of its mononucleotide frequencies. For example,

$$\mathbb{F}(ATC) \approx \mathbb{F}(A)\mathbb{F}(T)\mathbb{F}(C).$$

However, given the fact that this sequence follows CSPR, we have

$$\mathbb{F}(ATC) \approx \mathbb{F}(T)\mathbb{F}(A)\mathbb{F}(G)$$

which can be rewritten as

$$\mathbb{F}(ATC) \approx \mathbb{F}(G)\mathbb{F}(A)\mathbb{F}(T) \approx \mathbb{F}(GAT).$$

Therefore, $\mathbb{F}(ATC) \approx \mathbb{F}(GAT)$.

Remark 3.3 The intriguing question about the SP is why the property is weaker in actual DNA sequences, or in other words, why the *symmetry* is "broken" for

$$\boxed{\mathbb{F}(w) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(w)))} \neq \boxed{\mathbb{F}(\mathscr{C}(w)) \approx \mathbb{F}(\mathscr{R}(w))} \qquad (3.8)$$

Equation (3.8) shows that, in any EQ with four elements, the word frequencies are grouped into two subsets for the following: whenever $\mathbb{F}(w) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(w)))$ holds, so does $\mathbb{F}(\mathscr{C}(w)) \approx \mathbb{F}(\mathscr{R}(w))$ and vice-versa as the following theorem shows:

Theorem 3.1 *Given $w \in \mathscr{W}^k$ and the operators $\mathscr{C}$ and $\mathscr{R}$, then*

$$\mathbb{F}(w) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(w))) \Leftrightarrow \mathbb{F}(\mathscr{C}(w)) \approx \mathbb{F}(\mathscr{R}(w)) \qquad (3.9)$$

Proof Theorem 3.1 proof is divided in two steps. First let us assume that

$$\mathbb{F}(w) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(w)))$$

is true. Then,

$$\mathbb{F}(\mathscr{C}(w)) \approx \mathbb{F}(\mathscr{R}(\mathscr{C}(\mathscr{C}(w)))).$$

However, the $\mathscr{C}$ operator is an involution; therefore,

$$\mathscr{C}(\mathscr{C}(w)) = w.$$

Thus, it follows that

$$\mathbb{F}(\mathscr{C}(w)) \approx \mathbb{F}(\mathscr{R}(w)).$$

Now, let us assume that

$$\mathbb{F}(\mathscr{C}(w)) \approx \mathbb{F}(\mathscr{R}(w))$$

is true. Thus, we can write

$$\mathbb{F}(\mathscr{C}(\mathscr{R}(w))) \approx \mathbb{F}(\mathscr{R}(\mathscr{R}(w))).$$

However, operator $\mathscr{R}$ is an involution, which implies that $\mathscr{R}(\mathscr{R}(w)) = w$. Thus,

$$\mathbb{F}(\mathscr{C}(\mathscr{R}(w))) \approx \mathbb{F}(w). \qquad \square$$

The mathematical symbol $\Leftrightarrow$ means “if and only if.” Thus Theorem 3.1 is quite strong, and it asserts that, in a given equivalence class, its word frequencies are grouped two by two, or as $\mathbb{F}(w) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(w)))$ and $\mathbb{F}(\mathscr{C}(w)) \approx \mathbb{F}(\mathscr{R}(w))$.

One important remark is that Theorem 3.1 brings forward the usually implicit relationship between $\mathbb{F}(\mathscr{C}(w))$ and $\mathbb{F}(\mathscr{R}(w))$. However, nothing is said about the relationship between the frequencies of those two groups. Their frequencies are either different, as in actual DNA sequences, or similar, as in random DNA sequences that follow CSPR.

3.3.6 Palindromic Sequences

As mentioned previously, for every $k \geq 2$, the operators $\mathscr{R}$ and $\mathscr{C} \circ \mathscr{R}$ (when k is even) have fixed-words. Another important remark is the following:

Remark 3.4 The $\mathscr{C} \circ \mathscr{R}$ operator fixed-words, also known as *palindromic sequences*, satisfy the SP by definition.

Remark 3.4 has important consequences. First, let us understand it through a simple example. Fixed-words are defined as those $w \in \mathscr{W}^k$ where $w = \mathscr{C}(\mathscr{R}(w))$; consequently

$$\mathbb{F}(w) = \mathbb{F}(\mathscr{C}(\mathscr{R}(w))).$$

Now, consider the 2-words *CG* and *GC* (or *AT* and *TA*). It is known that, in some organisms, $\mathbb{F}(CG) \ll \mathbb{F}(GC)$, and this fact does not violate the SP because *CG* is a fixed-word of the $\mathscr{C} \circ \mathscr{R}$ operator. Thus, its frequency should be equal to the CG itself (which is always true), regardless of the frequency of *GC*.[16]

Palindromic sequences play major roles in molecular biology. There is a class of enzymes called restriction enzymes that cut double-stranded DNA. Some of them recognizes palindromic sequences. For instance, the *EcoR1* restriction enzyme recognizes the palindromic sequence *GAATTC*, meaning that, wherever this palindrome occurs, EcoR1 cuts the DNA at that position. We cannot forget that the most powerful tool of Synthetic Biology is known by the acronym CRISPR which stands for "Clustered Regularly Interspaced Short *Palindromic* Repeats" [100].

From the insertion of genes into plasmid vectors during gene cloning to the latest genome editing tool, palindromic sequences have several important biotechnological applications. We do not intend to review all known functions of palindromic sequences, but we cannot fail to mention that the recognition/methylation by DNA [amino]-methyltransferases [103] uses DNA palindromes as well.

Remark 3.5 How convenient it is that such important sequences have no frequencies bonds at all.

3.4 New Parity Rules

In the introduction of his book *Principles of Mathematical Logic* [48], David Hilbert (1862–1943) stated that

> The great advances in mathematics since antiquity, for instance in algebra, have been dependent to a large extent upon success in finding a usable and efficient symbolism.

Every mathematician recognizes this truth, and pursues its application whenever possible. That is the reason why we were so meticulous in the previous sections. The "symbolisms" that we've introduced are ripe and can now reveal their consequences. The following information will show how.

[16] This example will aid in the understanding of the discussion in Remark 4.1.

There are 4^k k-words in $\mathscr{W}^k$; therefore, from the very definition of frequency, we can write

$$\sum_{i=1}^{4^k} \mathbb{F}(w_i) = 1. \tag{3.10}$$

Equation (3.10) gives no insight into the k-word frequencies relationship. Each word w_i, where $i = 1, .., 4^k$, seems to have its own independent frequency $\mathbb{F}(w_i)$. However, we already know that this is not true. In order to bring forward the relationship between frequencies, let us start by applying the definition of GS to Eq. (3.10).

We can rewrite it in the following way: let $\mathscr{S}^k = \{g_1, g_2, \ldots, g_t\}$ be a GS, where $t = |\mathscr{P}^k|$. Therefore, we have

$$\sum_{i=1}^{t} \mathbb{F}(g_i) + \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i))) + \mathbb{F}(\mathscr{C}(g_i)) + \mathbb{F}(\mathscr{R}(g_i)) = 1. \tag{3.11}$$

In order to avoid an unnecessarily complex notation, Eq. (3.11) is actually a simplified version, because some equivalence classes have only two elements rather than four. Of course, in those cases, there are only two terms $\mathbb{F}(g_i)$ and $\mathbb{F}(\mathscr{C}(g_i))$ to add together (either $g_i = \mathscr{C}(\mathscr{R}(g_i))$ or $g_i = \mathscr{R}(g_i)$).

Theorem 3.1 and the SP imply that Eq. (3.11) could be rewritten with fewer terms because only two terms, $\mathbb{F}(g_i)$ and $\mathbb{F}(\mathscr{C}(g_i))$ for $i = 1, \ldots, t$, are actually independent. This remark is better understood with an example: when $k = 3$, we have $\mathbb{F}(AAA) \approx \mathbb{F}(TTT)$, $\mathbb{F}(ACG) \approx \mathbb{F}(CGT)$, $\mathbb{F}(CGG) \approx \mathbb{F}(CCG)$ and so on. Ultimately, there are only 32 independent frequencies instead of 64 (total number of 3-words or codons), which means that only $\frac{1}{2}$ of the number of 3-word frequencies are actually "independent." However, we can not forget that some EQs have only two elements, which prevents us from going further and drawing false conclusions.

Yet there is a way to approach Eq. (3.11) that prevents any misrepresentation. First, let us organize the terms of Eq. (3.11) into matrix form. Each row will contain an EQ. Thus, the matrix will have four columns: $\mathbb{F}(w)$, $\mathbb{F}(\mathscr{C}(\mathscr{R}(w)))$, $\mathbb{F}(\mathscr{C}(w))$ and $\mathbb{F}(\mathscr{R}(w))$, in that order.

Definition 3.13 The **Math Table** (MT) is a matrix with $t = |\mathscr{P}^k|$ rows and four columns. Each row is associated with an EQ, and the variables $\mathbb{F}(g_i)$, $\mathbb{F}(\mathscr{C}(\mathscr{R}(g_i)))$, $\mathbb{F}(\mathscr{C}(g_i))$, and $\mathbb{F}(\mathscr{R}(g_i))$, in that order, correspond to its columns. Furthermore, because the operators $\mathscr{R}$ and $\mathscr{C} \circ \mathscr{R}$ have fixed-words, whenever a fixed-word appears, only the columns $\mathbb{F}(g_i)$ and $\mathbb{F}(\mathscr{C}(g_i))$ should be filled in.

Table 3.1 When $k = 2$, there are 6 EQs; therefore, the MT has 6 rows and 4 columns

Class	$\mathbb{F}(g)$	$\mathbb{F}(\mathscr{C}(\mathscr{R}(g)))$	$\mathbb{F}(\mathscr{C}(g))$	$\mathbb{F}(\mathscr{R}(g))$
1	AA	–	TT	–
2	AC	GT	TG	CA
3	AG	CT	TC	GA
4	AT	–	TA	–
5	CG	–	GC	–
6	CC	–	GG	–

Remark 3.6 The name *Math Table* (MT) is used only to emphasize its mathematical nature rather than its biological nature.

Remark 3.7 Let us take $k = 2$ and the 2-word CG. Thus, we have

$$CG = \mathscr{C}(\mathscr{R}(CG)) = CG$$

and

$$\mathscr{C}(CG) = \mathscr{R}(CG) = GC.$$

The 2-word CG is a fixed-word of the operator $\mathscr{C} \circ \mathscr{R}$; therefore, which columns of the MT should be filled in? Taking advantage that the operator $\mathscr{C}$ has no fixed-words, in order to avoid ambiguity, whenever fixed-words appear, *by definition*, only the columns corresponding to $\mathbb{F}(CG)$ and $\mathbb{F}(\mathscr{C}(CG))$ should be filled in (see Table 3.1).

Remark 3.8 The MT's first column is a GS.

Example 3.16 When $k = 2$, we have Table 3.1.

Example 3.17 When $k = 3$, we have Table 3.2.

3.4.1 Second Insight

By organizing the k-word frequencies of real genomic data in this way, two hidden identities which would have otherwise remained unnoticed were discovered. No matter which GS was chosen, when we summed up the k-word frequencies in each column, almost independently of k[17] we observed that

[17] When $k = 2$, as will be discussed in Remark 4.1, the identities are violated slightly.

$$\sum_{i=1}^{t} \mathbb{F}(g_i) \approx \sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(g_i)), \tag{3.12}$$

and

$$\sum_{i=1}^{t} \mathbb{F}(\mathscr{R}(g_i)) \approx \sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i))). \tag{3.13}$$

Remark 3.9 Equations (3.12) and (3.13) cannot be derived from the SP. Actually, considering the SP, one would expect that

$$\sum_{i=1}^{t} \mathbb{F}(g_i) \approx \sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i)))$$

Table 3.2 When $k = 3$, there 20 EQs; therefore, its MT has 20 rows and 4 columns

Class	$\mathbb{F}(g)$	$\mathbb{F}(\mathscr{C}(\mathscr{R}(g)))$	$\mathbb{F}(\mathscr{C}(g))$	$\mathbb{F}(\mathscr{R}(g))$
1	AAA	–	TTT	–
2	AAT	ATT	TTA	TAA
3	TTG	CAA	AAC	GTT
4	CTT	AAG	GAA	TTC
5	ATA	–	TAT	–
6	ATC	GAT	TAG	CTA
7	ATG	CAT	TAC	GTA
8	ACA	–	TGT	–
9	TGA	TCA	ACT	AGT
10	CCA	TGG	GGT	ACC
11	GCA	TGC	CGT	ACG
12	TCT	–	AGA	–
13	GCT	AGC	CGA	TCG
14	AGG	CCT	TCC	GGA
15	CAC	–	GTG	–
16	CAG	CTG	GTC	GAC
17	CTC	–	GAG	–
18	CCC	–	GGG	–
19	GCC	GGC	CGG	CCG
20	GCG	–	CGC	–

and, in conjunction with Theorem 3.1,

$$\sum_{i=1}^{t} \mathbb{F}(\mathscr{R}(g_i)) \approx \sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(g_i)).$$

However, we should remember that the operator $\mathscr{C} \circ \mathscr{R}$ has fixed-words (see Remark 3.4), which implies that, only if we summed up its fixed-word frequencies twice (on both sides of the equal sign) would those formulas be true.

Remark 3.10 What is intriguing in Eq. (3.12) is that, usually, $\mathbb{F}(g_i) \neq \mathbb{F}(\mathscr{C}(g_i))$, yet when $\mathbb{F}(g_i)$ and $\mathbb{F}(\mathscr{C}(g_i))$ are separately summed up for $i = 1, \ldots, t$, the resulting values are approximately equal, as asserted by Eq. (3.12). The same reasoning may be applied to Eq. (3.13).

Although it may not be obvious at first, Eqs. (3.12) and (3.13) are the desired generalizations of Eqs. (3.2) and (3.5). To demonstrate this, consider the simplest case of a single nucleotide, a 1-word. In this case, the alphabet is equal to the dictionary: $\{A, C, G, T\}$. Using equivalence operators, we have only two ECs: $\{A, T\}$ and $\{C, G\}$. Without loss of generality, we can choose the set $\{A, C\}$ as the GS. Thus, $t = 2$ and $g_1 = A$ and $g_2 = C$. Because the operator $\mathscr{R}$ plays no role in this case, Eqs. (3.12) and (3.13) would be equal to

$$\underbrace{\mathscr{F}(A) + \mathscr{F}(C) \approx \mathscr{F}(T) + \mathscr{F}(G)}_{\text{Eq. (3.5)}}$$

Similarly, if we choose GS as $\{A, G\}$, we would have

$$\underbrace{\mathscr{F}(A) + \mathscr{F}(G) \approx \mathscr{F}(T) + \mathscr{F}(C)}_{\text{Eq. (3.2)}}$$

Therefore, it is shown that we obtained the alternative generalization for CSPR. With this new formalization, we can return with more confidence to the questions raised previously and provide the following answers:

- Which oligonucleotides do we add on each side of the equation?

  In Eq. (3.12), we will add up the frequencies of the oligonucleotides of the GS on one side of the equation, and the frequencies of the complementary forms for the same elements of the GS on the other side; similarly, in Eq. (3.13), we will add up frequencies of the reverse forms on one side, and the frequencies of the reverse complementary forms of the GS elements on the other side.

(continued)

- How many oligonucleotides will we add up?

  It depends on the number of elements in GS. For words with a single nucleotide, we will have $t = 2$; for two nucleotides, $t = 6$; for three nucleotides, $t = 20$, and in general, for words with k nucleotides, we will have $t = 2^{k-1} + 4^{k-1}$.
- How many equations would there be? Would there be an equation for dinucleotides and one for trinucleotides, and so on?

  In its most general form, there would be only two equations, and they do not depend on the number of nucleotides in each word. In other words, the equations have the same form and are self-similar (a *fractal-like* property), regardless of the number of nucleotides.

3.4.2 *Third Insight*

Given that Eqs. (3.12) and (3.13) are actually generalizations of CSPR, could we generalize Eqs. (2.1) and (2.2) as well? The answer is yes. As in our work published in 2008 [102] and taking in account that, by definition, the sum of frequencies of all words, regardless of the number of nucleotides, is always equal to one, we have

$$\underbrace{\left(\sum_{i=1}^{t} \mathbb{F}(g_i) + \sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(g_i))\right)}_{\sum_{i=1}^{t} \mathbb{F}(g_i) \approx \sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(g_i))} + \underbrace{\left(\sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i))) + \sum_{i=1}^{t} \mathbb{F}(\mathscr{R}(g_i))\right)}_{\sum_{i=1}^{t} \mathbb{F}(\mathscr{R}(g_i)) \approx \sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i)))} = 1. \tag{3.14}$$

But based on Eqs. (3.12) and (3.13), we have

$$2\sum_{i=1}^{t} \mathbb{F}(g_i) + 2\sum_{i=1}^{t} \mathbb{F}(\mathscr{R}(g_i)) \approx 1. \tag{3.15}$$

and

$$2\sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(g_i)) + 2\sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i))) \approx 1. \tag{3.16}$$

This implies that

$$\sum_{i=1}^{t} \mathbb{F}(g_i) + \mathbb{F}(\mathscr{R}(g_i)) \approx \frac{1}{2} \tag{3.17}$$

and

$$\sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(g_i)) + \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i))) \approx \frac{1}{2} \tag{3.18}$$

Equations (3.17) and (3.18) are the true generalizations of Eqs. (2.1) and (2.2).

Noteworthy that there is no direct reference to the number of nucleotides in either formula. Therefore, they should be valid for every value of $k = 1, 2, 3, \ldots, K^*$, where K^* is an upper bound due the fact that actual DNA sequences are finite in length. Hence, we have the same formulas regardless of the value of k. It is tempting to think of a DNA sequence as a fractal-like palimpsest written layer over layer with the same alphabet. In spite of the level, the same invariant property summarized in Eqs. (3.17) and (3.18) is found.

3.4.3 *Alternative Path*

Before finishing this section, we would like to discuss a very interesting question that may have occurred to some readers: could we have used the SP to derive Eqs. (3.17) and (3.18)? The answer is yes. However, we decided to use Eqs. (3.12) and (3.13) instead because they prevent us from deriving two flawed frequency rules as well. In order to tackle this issue, let us rewrite Eq. (3.11) as follows:

$$\sum_{i=1}^{t} \underbrace{\mathbb{F}(g_i) + \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i)))}_{\mathbb{F}(g_i) \approx \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i)))} + \underbrace{\mathbb{F}(\mathscr{C}(g_i)) + \mathbb{F}(\mathscr{R}(g_i))}_{\mathbb{F}(\mathscr{C}(g_i)) \approx \mathbb{F}(\mathscr{R}(g_i))} = 1.$$

In addition to Eqs. (3.17) and (3.18), it is possible to produce these two additional formulas:

$$\sum_{i=1}^{t} \mathbb{F}(g_i) + \mathbb{F}(\mathscr{C}(g_i)) \approx \frac{1}{2}$$

and

$$\sum_{i=1}^{t} \mathbb{F}(\mathscr{C}(\mathscr{R}(g_i))) + \mathbb{F}(\mathscr{R}(g_i)) \approx \frac{1}{2}$$

However, these additional equations are flawed. The reason is essentially the same as that which is presented in Remark 3.9.

Chapter 4
"In God We Trust; All Others, Bring Data"

Abstract No matter how elegant and beautiful a mathematical model is, it must be confronted with nature's sovereign behavior in order for its true worth to be assessed. In this chapter, we use all publicly available genome reference sequences to evaluate the new parity rules.

> It would seem to me that man cannot live without mysteries. One could say the great biologists worked in the very light of darkness. (Erwin Chargaff)
>
> Models — in contrast to those who sat for Renoir — improve with age. (Erwin Chargaff)

As we anticipated and unlike typical findings in Biology, our results do not follow from direct empirical observation. In fact, we would never reach them using empirical methods alone. They were derived through an old-fashioned mathematical method: abstract *demonstrative reasoning*.

This is the last chapter, and so far, we have presented no empirical evidence of our last results. There are beautiful mathematical models that are useless from a practical point of view. For instance, before the existence of the *codon table*, George Gamow (1904–1968) proposed the "diamond code" for associating nucleotides with amino acids [35]. Gamow's model[1] was mathematically appealing, but plainly wrong. Nature does not care about how we think it should behave, even when our models are elegant and beautiful. Therefore, despite our results' undoubtedly aesthetic value, they must be confronted with nature's sovereign behavior in order for their true worth to be determined.

During our undergraduate studies, we read Pólya's book *How to Solve it?* [77]. The objective of the book was to teach students to solve mathematical and non-mathematical problems. Since then, we've been trying to follow his heuristics, which can be summarized as a four-step methodology:

- Understanding the problem
- Devising a plan
- Carrying out the plan
- Looking back

[1] Despite being wrong, Gamow's diamond code is still remembered for its beauty!

The "Looking back" step consists of examining the solution obtained and answering two questions: "Can you derive the results differently?" and "Can you check the result?"

4.1 "To Be or Not to Be"

As for the first question, the worst-case scenario would be if our results could be obtained by chance alone. It may not seem so at first, but this possibility is quite plausible. For starters, it is important to realize that Eqs. (3.17) and (3.18) are directly linked to the MT. Note that Eq. (3.17) is just the summation of its first and last columns, while Eq. (3.18) is the summation of its second and third columns. Whatever the value of k, this implies that Eq. (3.17) comprises half of the frequencies, as does Eq. (3.18). The problem is when half of the frequencies from any assay are randomly selected and summed up, the expected resulting value is 0.5. Thus, we need to perform extra experiments in order to avoid the possibility of our results being a random artifact.

Let us begin by confirming what we've just said. Consider any assay where there are N frequencies, f_k in which $k = 1, \ldots, N$, or

$$\sum_{k=1}^{N} f_k = 1.$$

If we randomly select half of those frequencies, what is their expected sum?

We have performed 1,000,000 random trials to address the above question. Figure 4.1 summarizes the result. The expected value, as anticipated, is $\mu_{random} = 0.5$, and the standard deviation (SD) is $\sigma_{random} \approx 0.036$.

Does this result mean that our equations are expected by chance? This question cannot be answered yet, because we do not know what the corresponding distribution is when actual data is used.

The k-word frequencies of real DNA sequences have some known patterns.[2] We could mention CSPR or the SP which, as we've seen, state that some frequencies occur at least twice.[3] Therefore, CSPR alone is enough to show that k-word frequencies are not completely random. So, in order to properly answer the question, we need to take our investigation a little further. We must investigate the μ_{real} and σ_{real} values when real data is used. Remember that our former assays were completely random: both the frequencies themselves and the selection of the half of them were random.

[2] Of course, our new findings should be considered one example of such regularities, but because we still do not know if they are true (in the sense that actual DNA sequences actually satisfy them), we're going to set them aside for a while.

[3] For the sake of argument, we are considering that $\mathbb{F}(w) = \mathbb{F}(\mathscr{C}(\mathscr{R}(w)))$.

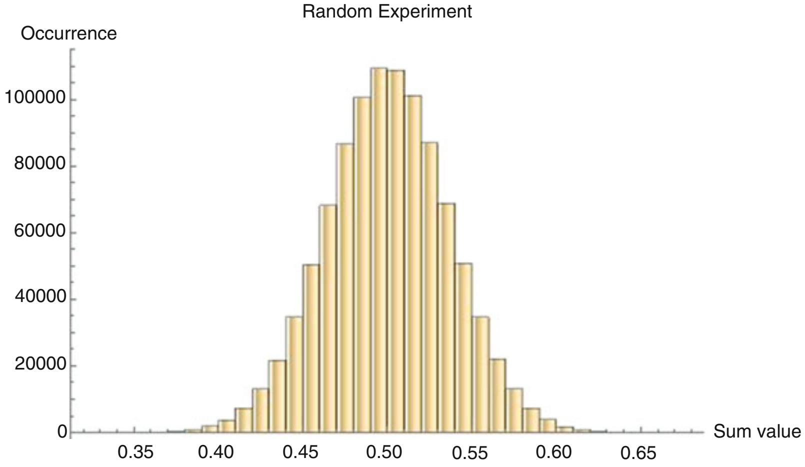


Fig. 4.1 Histogram of the sum of half of frequency values

Without loss of generality, let us consider $k = 3$ and the human genome.[4] This time, only the selection of half of the frequencies will be random. We have calculated the 64 corresponding frequencies for every 3-word within the human genome. We randomly selected 32 frequencies 1,000,000 times and summed them up. The result is depicted in Fig. 4.2.

Although the histogram curve shape and the $\mu_{real} = 0.5$ are identical to those of the previous experiment, the SD is actually different: $\sigma_{real} \approx 0.028$. Figure 4.3 may help to clarify this difference. As we expected, real DNA frequencies reduce the SD value, which means that the data dispersion is smaller. Consequently, our null hypothesis cannot yet be avoided. We need to design an assay that can definitively settle this matter.

So far, both experiments relied on random selection of frequencies, and one of our major contributions[5] is actually a very particular selection criterion: the Generating Set. As we've mentioned previously, the $\mathscr{R}$, $\mathscr{C}$ and $\mathscr{R} \circ \mathscr{C}$ operators are actually *Closed Equivalence Class Operators*, meaning they induce a natural k-word set partition into *Closed Equivalence Classes*. Once the EQs are established, however, the choice of the GS elements is arbitrary. This arbitrariness will help us to devise the experiment to give a definitive answer to the persistent question.

The new assay is essentially an adaptation of the last, in which the 3-word frequencies are grouped into EQs and the GS elements are randomly chosen 1,000,000 times.

[4] Actually, we have performed the experiment using more than 60,000 organisms, as will be shown latter.

[5] And perhaps our most important one.

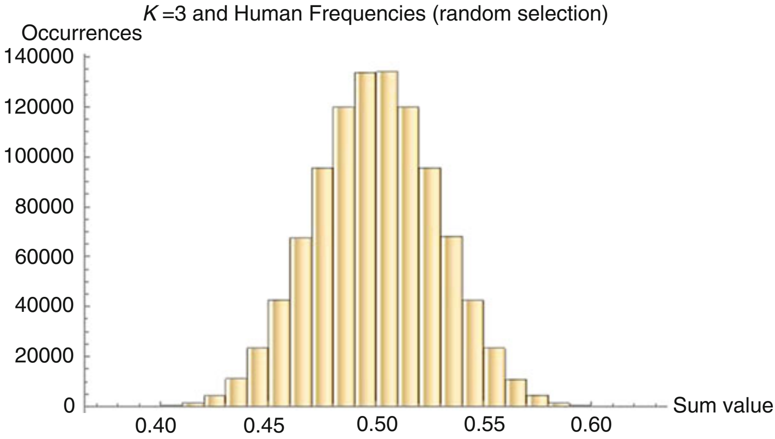


Fig. 4.2 Histogram of the sum of half of 32 3-word human frequency values

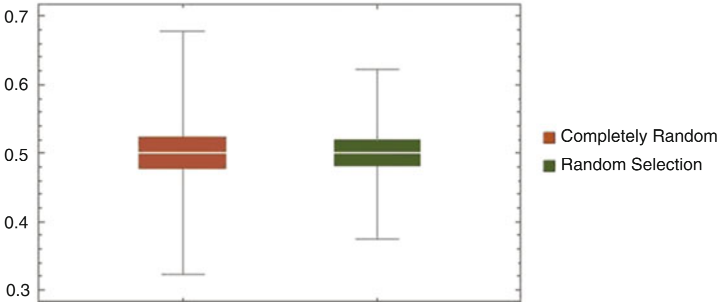


Fig. 4.3 This *box-and-whisker plot* depicts the two experiment results side by side

Once more, the histogram's curved shape and the $\mu_{GS} = 0.5$ are identical (Fig. 4.4); however, the SD is approximately two orders of magnitude smaller than the previous ones: $\sigma_{GS} = 0.000478502$.

Despite our efforts to produce a better visual representation, the box-and-whisker plot representing our last experiment is only slightly visible in Fig. 4.5 because

$$\sigma_{GS} \ll \boxed{\sigma_{real} < \sigma_{random}}$$

After this last experiment, it is possible to answer beyond any reasonable doubt that our results using the MT, regardless the GS element choice, are not expected by chance alone.

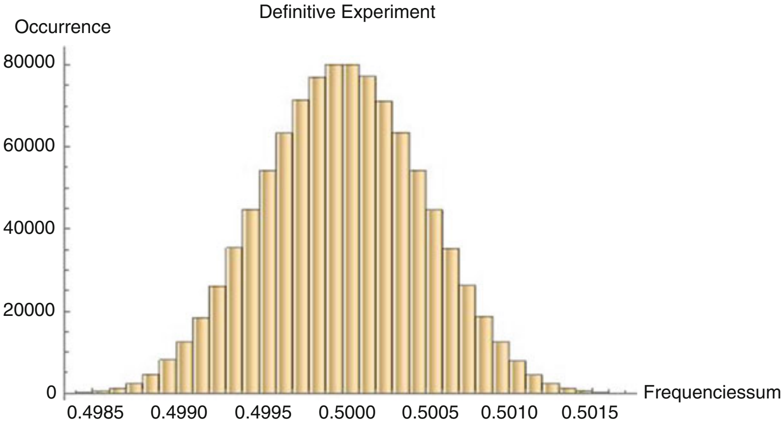


Fig. 4.4 Histogram of the last experiment

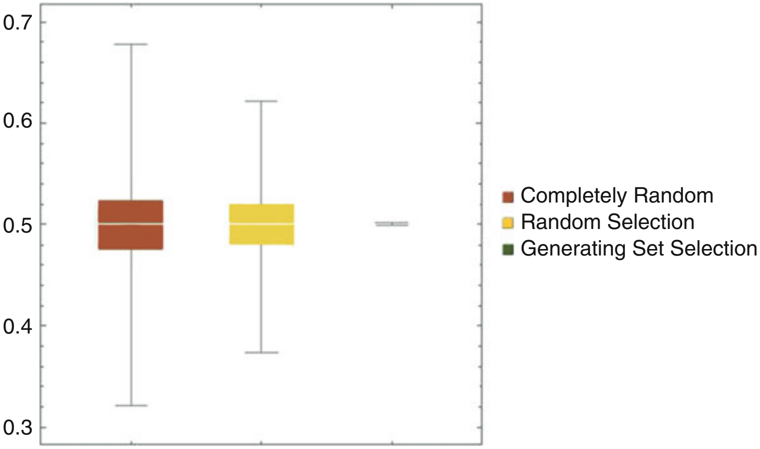


Fig. 4.5 This *Box-and-whisker plot* shows the results of all of the experiments

4.2 The Three Kingdoms of Life

Now it is time to address Polya's second question: "Can you check the results?" So far, all experiments performed have used human data and have been limited to $k = 3$. What about other organisms? And how do our equations behave with other k values? Both questions were answered in our former work [101]. Yet, because we had selected only 32 organisms, doubts remained regarding the appropriateness of extrapolating our conclusions to organisms left off that list.

It could be argued that, without a comprehensive assay, our results may be irremediably incomplete or even flawed. This objection relies on the fact that genomic sequences vary immensely from one organism to the next. For instance, it is known that bacteria have a huge range of GC-content[6] and few repetitive sequences. On other hand, there are whole taxonomic classes that do not vary much in GC-content, but which may have 90% of their sequences composed of repeats. Thus, in the presence of these known (and unknown) idiosyncrasies, it is reasonable to question whether our equations remain valid regardless of the variability of the genomic sequence.

Moreover, even though our objective was to create a list of “representative” organisms from the very beginning, we soon realized that such a list is always deficient: today, even the concept of *species* [84] is being questioned, which makes the task of selecting a representative species ultimately meaningless. Thus, we gave up the idea of a “representative” list and decided to download the whole set of available genomic *reference* sequences deposited at the National Center for Biotechnology Information (NCBI). Even this list often changes, for new genomic sequences are constantly added. In any case, while writing this book, we were able to download 60,541 reference genomic sequences from the NCBI.

The genomic sequences that were downloaded may be grouped into the three kingdoms of life [99]: Bacteria, Archaea and Eukaryota. Table 4.1 shows the sequence frequency by group. The Bacteria set is clearly overrepresented. In order to deal with this unbalanced dataset and for analytical purposes, we decided to analyze each group separately.

It is important to note that there are some species with more than one individual. If we remove the redundancy, there are 9558 supposedly different species. To better reflect the redundancy, among Eukaryota, there are only nine species with more than one individual; for instance, there are three representatives of *Cryptococcus neoformans*.[7] Within Archaea, there are 39 species with more than one representative. The *Methanosarcina mazei*[8] has 62 individuals. And finally, the Bacteria kingdom is by far the most redundant. There are 1408 species with more than one exemplar. *Streptococcus pneumoniae*,[9] for instance, has 6829 individuals.

Table 4.1 Sequence proportions from three Kingdom of Life

Kingdom	*No. of sequences*	%
Archaea	514	0.84
Bacteria	59,436	98.17
Eukaryota	591	0.97

[6] GC-content (GC-ratio or G+C ratio) is defined as

$$\frac{\mathbb{F}(C) + \mathbb{F}(G)}{\mathbb{F}(A) + \mathbb{F}(T) + \mathbb{F}(C) + \mathbb{F}(G)}.$$

[7] It is an opportunistic fungal pathogen that may cause meningitis.

[8] Probably the only anaerobic methanogen known to produce methane using all three metabolic pathways.

[9] Also known as *Pneumococcus*. It is recognized as a major cause of pneumonia.

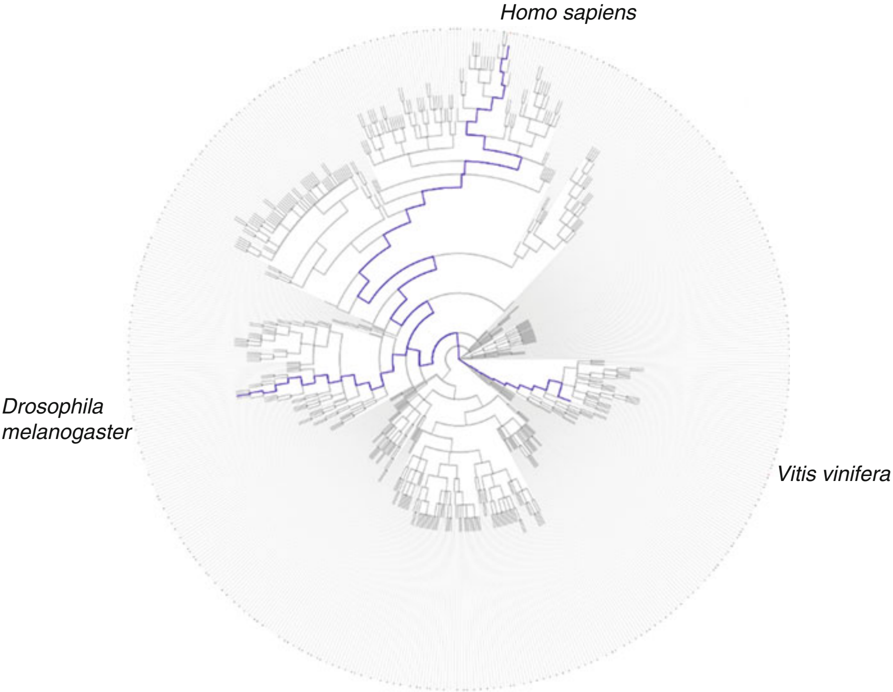


Fig. 4.6 Phylogenetic tree for Eukaryota. On the *left side*, we have *Drosophila melanogaster*, at the *top*, *Homo sapiens* and on the *right*, *Vitis vinifera*

Unfortunately, despite our best efforts, it was not feasible to show the complete phylogenetic tree of our data. Nevertheless, to provide an indication of how comprehensive our data was, we have drawn the Eukaryota tree (Fig. 4.6).

The assessment of 60,541 genomic sequences is no small task. However, we used the High Performance Computing (HPC) infrastructure of the *Laboratório Multiusuário de Bioinformática da Embrapa*, which has enough computer power to face this challenge. Thus, we calculated the MTs ranging from $k = 1, \ldots, 8$ for each of the genome sequences.

We've decided to present the results in a box-and-whisker plot (see Fig. 4.7), which encapsulates the main statistical information such as median and the first and third quartiles, as well as the maximum, the minimum, and the variation.

In Fig. 4.7, for each value of k, there are three box-and-whisker plots corresponding to the Archaea, Bacteria and Eukaryota kingdoms, respectively.

The values do not deviate significantly from 0.5, except when $k = 2$. The variability of the Eukaryota kingdom is much smaller than that of the other two groups. The Bacteria kingdom has the largest variability. This may be partially explained by the large number of organisms in this category, and by the fact that only 4512 (7.59%) of 59,436 sequences are actually complete genomes; the other 54,924 (92.41%) are incomplete drafts (scaffolds or contigs). Below, we will present some important remarks regarding Fig. 4.7.

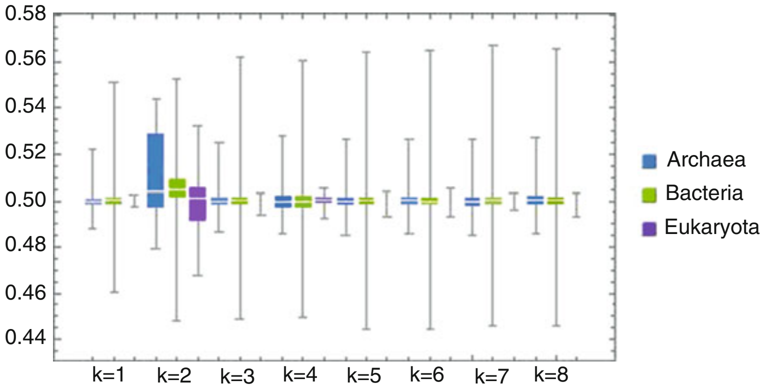


Fig. 4.7 *Box-and-whisker plot* representing the value of Eq. (3.17) for $k = 1, \dots, 8$

4.2.1 A Plausible Explanation for the k2-Effect

Remark 4.1 An interesting part of Fig. 4.7 is that, when $k = 2$, we get the largest variability in all groups. We will call this phenomenon the "k2-effect." Even the most rule-compliant kingdom, the Eukaryota, exhibits this awkward feature. To understand what is happening, let us look at the frequency distribution of each dinucleotide in order to figure out a plausible explanation.

In Figs. 4.8, 4.9, 4.10, the dinucleotide frequencies were sorted in such a way that each dinucleotide and its reverse-complement[10] were placed next to each other. All figures show that, when $k = 2$, there are two pairs of $\mathscr{C} \circ \mathscr{R}$ operator fixed-words: $\{AT, TA\}$ and $\{CG, GC\}$, which, on *average*, exhibit the following pattern:

$$\mathbb{F}(CG) \ll \mathbb{F}(GC) \tag{4.1}$$

and

$$\mathbb{F}(AT) \gg \mathbb{F}(TA). \tag{4.2}$$

Inequalities (4.1) and (4.2) are not new. Back in 1962, Swartz and colleagues [90] observed that $\mathbb{F}(CG)$ was much lower than its expected value. In their own words:

> The most striking example is the frequency of the CpG sequence, which are only approximately one-third of the value calculated from the base composition, whereas those of the isomeric GpC sequence are close to the calculated values.

[10] Of course, the exceptions were the $\mathscr{C} \circ \mathscr{R}$ operator fixed words: $\{AT, TA, CG, GC\}$.

Fig. 4.8 Archaea dinucleotide frequencies

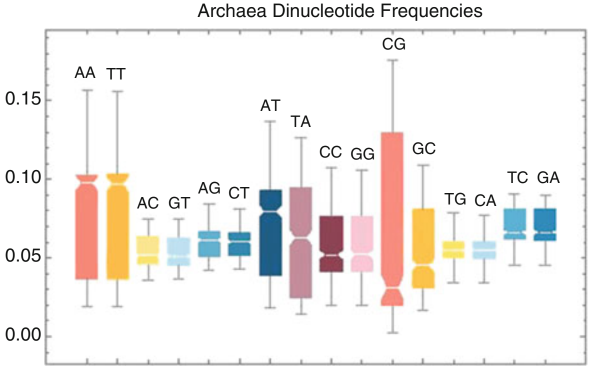


Fig. 4.9 Bacteria dinucleotide frequencies

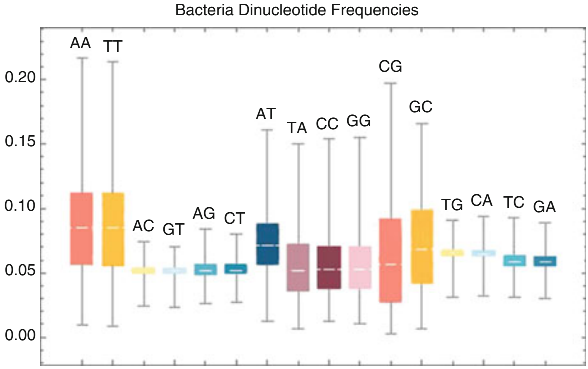
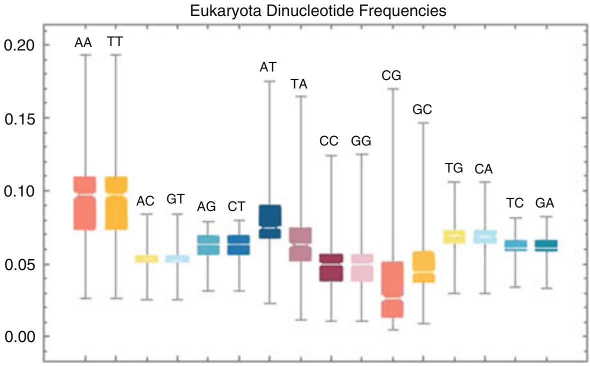


Fig. 4.10 Eukaryota dinucleotide frequencies



The biological community uses CpG to refer to the CG dinucleotide. Analogously, GpC refers to the GC dinucleotide and so on. Because several cited authors use that notation, we have maintained its meaning.

We cannot help but note that Swartz and colleagues failed to realize that their data implied that TpA, to a lesser extent, was also below to its expected value.

The first sound biological explanation to this phenomenon appeared in 1980, when Sir Adrian Peter Bird associated the deficiency of CpG with DNA methylation [7]:

> Might the function of DNA methylation be to increase the mutation rate? It is difficult to argue conclusively for or against this possibility. One point against it, however, is that CpG would become progressively rarer in the DNA, thereby canceling out the importance of its mutability. An alternative possibility is that the function of DNA methylation lies in some other direction, whose advantages outweigh the attendant disadvantage of high mutability. What that function might be is not yet clear.

Prophetic words—since then, DNA methylation has proven to exhibit many important biological functions [52], including its unexpected role in "epigenetic[11] memory" [8]:

> The heritability of methylation states and the secondary nature of the decision to invite or exclude methylation supports the idea that DNA methylation is adapted for a specific cellular memory function in development.

While DNA methylation can be a plausible explanation for CpG deficiency in animals and plants, the reason for TpA depletion is completely unknown [28].

Regardless of their biological causes, we conjecture that the $CpG \times GpC$ and $ApT \times TpA$ frequency differences may negatively impact the validity of Eq. (3.17). Thus, in order to double check this possibility, let us fully expand Eq. (3.17) for $k = 2$, considering, without loss of generality, the Generating Set $GS = \{AA, AC, AG, AT, CC, CG\}$:

$$\begin{aligned} &\mathbb{F}(AA) + \mathbb{F}(AC) + \mathbb{F}(AG) + \mathbb{F}(AT) + \mathbb{F}(CC) + \mathbb{F}(CG) \approx \\ &\quad \mathbb{F}(TT) + \mathbb{F}(TG) + \mathbb{F}(TC) + \mathbb{F}(TA) + \mathbb{F}(GG) + \mathbb{F}(GC) \end{aligned} \tag{4.3}$$

Equation (4.3) is the desired expansion. In order to simplify the notation, let us denote it as:

$$\Delta_1 = \mathbb{F}(AA) + \mathbb{F}(AC) + \mathbb{F}(AG) + \mathbb{F}(AT) + \mathbb{F}(CC) + \mathbb{F}(CG) \tag{4.4}$$

and

$$\Delta_2 = \mathbb{F}(TT) + \mathbb{F}(TG) + \mathbb{F}(TC) + \mathbb{F}(TA) + \mathbb{F}(GG) + \mathbb{F}(GC) \tag{4.5}$$

Thus, Eq. (4.3) implies that

$$\Delta = \frac{\Delta_1}{\Delta_2} \approx 1 \tag{4.6}$$

Now, to verify whether Eqs. (4.1) and (4.2) have any impact on Eq. (4.6), let us define three additional quotients:

$$\delta_{\{AT,CG\}} = \frac{\Delta_1 - \mathbb{F}(CG) - \mathbb{F}(AT)}{\Delta_2 - \mathbb{F}(GC) - \mathbb{F}(TA)} \tag{4.7}$$

[11] May be defined as a "change in gene expression without base sequence alteration" [82].

$$\delta_{AT} = \frac{\Delta_1 - \mathbb{F}(AT)}{\Delta_2 - \mathbb{F}(TA)} \tag{4.8}$$

$$\delta_{CG} = \frac{\Delta_1 - \mathbb{F}(CG)}{\Delta_2 - \mathbb{F}(GC)} \tag{4.9}$$

Note that, in order to obtain $\delta_{\{AT,CG\}}$, it is necessary to subtract from the numerator of Δ the frequencies $\mathbb{F}(CG)$ and $\mathbb{F}(AT)$. Likewise, we subtracted from the denominator of Δ the frequencies $\mathbb{F}(GC)$ and $\mathbb{F}(TA)$. In other words, $\delta_{\{AT,CG\}}$ is Δ without the "problematic" operator $\mathscr{C} \circ \mathscr{R}$ fixed-words. The same reasoning applies to δ_{AT} and δ_{CG}.

Figures 4.11, 4.12, and 4.13 below show paired histograms for Δ (left side) and $\delta_{\{AT,CG\}}$, δ_{AT} and δ_{CG} (right side), respectively.

Looking at Figs. 4.11, 4.12 and 4.13, it is clear that δ_{CG} average is closer to 1 and that its standard deviation is the smallest. Based on this observation, we are led to conclude that

$$\mathbb{F}(CG) \ll \mathbb{F}(GC)$$

is the inequality that carries the most responsibility for the k2-effect.

4.2.1.1 There is No k3-Effect

Remark 4.2 If we had a *"k3-effect"* instead, we would immediately associate it with Richard Gratham's (1922–2009) genome hypothesis:

> The genetic code is used differently by different kinds of species. Each type of genome has a particular coding strategy, that is, choices among degenerate bases are consistently similar for all genes therein. [41]

However, there is no k3-effect at all. This is no trivial result. While in the case of Eukaryota, this absence could be explained by the small number of protein coding regions relative to the genome as whole; the same reasoning does not apply to the Bacteria and Archaea kingdoms. Thus, we expected to observe a k3-effect in Bacteria and Archaea, but we don't. Therefore, we are led to conclude that codon bias is not enough to break our rules and produce a k3-effect, even in very protein-rich genomes.

4.2.1.2 What About Non-compliant Sequences?

Remark 4.3 Non-compliant genome assemblies are as important as the compliant ones. Our scientific knowledge is always incomplete. Scientific advancements usually depend on the deep inspection of outliers. For instance, Albert Einstein (1879–1955) received his Nobel prize in Physics not for his outstanding Theories of Relativity (Special and General), but for his studies of a classical non-compliant

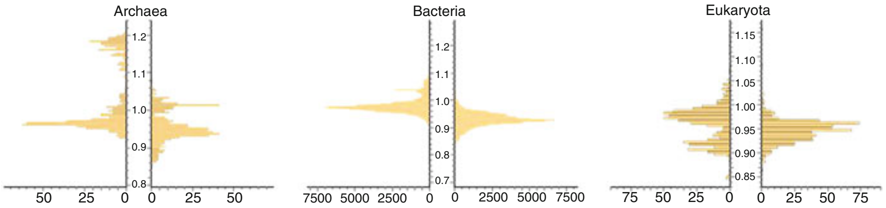


Fig. 4.11 Three paired histograms showing Δ on *left side* and $\delta_{\{AT,CG\}}$ on *right side* for Archaea, Bacteria, and Eukaryota, respectively. The Δ average and standard deviation values are: $\{\mu_{Arc} = 1.02876, \sigma_{Arc} = 0.09079\}$, $\{\mu_{Bac} = 0.983073, \sigma_{Bac} = 0.032939\}$ and $\{\mu_{Euk} = 0.961499, \sigma_{Euk} = 0.0389073\}$. The $\delta_{\{AT,CG\}}$ average and standard deviation values are: $\{\mu_{Arc} = 0.958223, \sigma_{Arc} = 0.04326\}$, $\{\mu_{Bac} = 0.92904, \sigma_{Bac} = 0.0286364\}$, and $\{\mu_{Euk} = 0.946848, \sigma_{Euk} = 0.0286364\}$

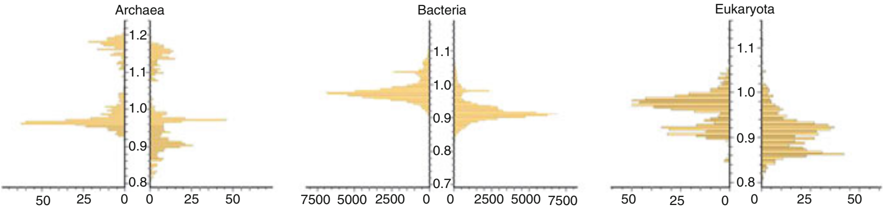


Fig. 4.12 Three paired histograms showing Δ on *left side* and δ_{AT} on *right side* for Archaea, Bacteria, and Eukaryota, respectively. The Δ average and standard deviation values are: $\{\mu_{Arc} = 1.02876, \sigma_{Arc} = 0.09079\}$, $\{\mu_{Bac} = 0.983073, \sigma_{Bac} = 0.032939\}$ and $\{\mu_{Euk} = 0.961499, \sigma_{Euk} = 0.0389073\}$. The δ_{AT} average and standard deviation values are: $\{\mu_{Arc} = 0.990651, \sigma_{Arc} = 0.101922\}$, $\{\mu_{Bac} = 0.923481, \sigma_{Bac} = 0.0366809\}$, and $\{\mu_{Euk} = 0.908992, \sigma_{Euk} = 0.0431361\}$

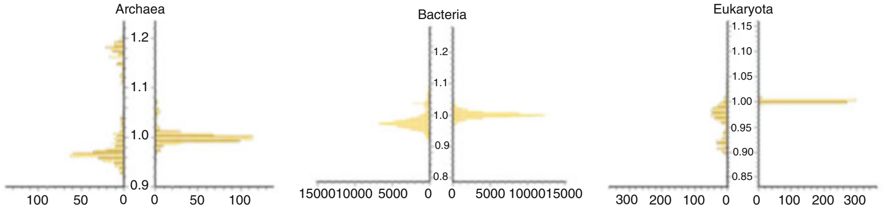


Fig. 4.13 Three paired histograms showing Δ on *left side* and δ_{CG} on *right side* for Archaea, Bacteria, and Eukaryota, respectively. The Δ average and standard deviation values are: $\{\mu_{Arc} = 1.02876, \sigma_{Arc} = 0.09079\}$, $\{\mu_{Bac} = 0.983073, \sigma_{Bac} = 0.032939\}$ and $\{\mu_{Euk} = 0.961499, \sigma_{Euk} = 0.0389073\}$. The δ_{CG} average and standard deviation values are: $\{\mu_{Arc} = 1.00145, \sigma_{Arc} = 0.015637\}$, $\{\mu_{Bac} = 1.00034, \sigma_{Bac} = 0.0134426\}$, and $\{\mu_{Euk} = 0.999988, \sigma_{Euk} = 0.00211276\}$

electromagnetic theory phenomenon known as the Photoelectric Effect. Both compliant and non-compliant phenomena should be studied, for they reveal different aspects of the same problem; however, the non-compliant ones force us into deep thought which may lead to new insights. We call the study of biological outliers *Freak Biology*.

Table 4.2 Non-compliant genome assemblies

Kingdom	Sequences	%
Archaea	28	5.44
Bacteria	2083	3.82
Eukaryota	2	0.33

In order to identify the non-compliant genome assemblies, we had to define what it meant to be non-compliant. We decided to use the term "non-compliant" to describe the genome assemblies that deviate from the expected value (0.5) by more than one standard deviation. See Table 4.2.

The only two non-compliant eukaryote genome assemblies are those of *Dipodomys ordii*[12] (Assembly accession: GCF_000151885.1) and *Eimeria tenella*[13] (Assembly accession: GCF_000499545.1). Of course, we must investigate each of these non-compliant genome assemblies. We hope that their study may lead to new discoveries.

4.2.1.3 Completeness and Rule Compliance

Remark 4.4 What happens when the genome assembly does not use the entire k-word within the $\mathscr{W}^k$? Do our equations still hold? Figure 4.7 suggests that they do. Note that the value of Eq. (3.17) does not deviate from 0.5 even when $k = 8$. While most of the Eukaryota kingdom exhibits Completeness $\mathbb{K} \geq 8$ (see Fig. 3.2), the same is not true for the Archaea and Bacteria kingdoms. Because a significant portion of the Archaea and Bacteria kingdoms exhibit Completeness $\mathbb{K} < 8$, we would expect, at least for those kingdoms, either a k7-effect or a k8-effect. However, there is no k7 or k8-effect at all. This is a very interesting observation.

In order to get a glimpse into the relationship between Completeness and our equations' validity, we ran the experiment for human genome assembly GRCh38.p7 for $k = 1$ to $k = 18$. The result is reported in Table 4.3.

According to our definition, the completeness of this assembly is $\mathbb{K} = 10$; however, Table 4.3 shows that only when $k = 16$ does the value of Eq. (3.17) start to deviate significantly from 0.5. At the same time, approximately one quarter of the 15-words are actually present (see Fig. 4.14).

Note that, when $k = 15$, even with almost half of the k-words absent, our rules still hold. This is no trivial result, either. Furthermore, looking at Table 4.3, we realize that, with almost one quarter of the k-words, the value of Eq. (3.17) does not deviate substantially from 0.5, either. Based on these observations, we conjecture that the organisms' k-words are highly interdependent.

[12] Ord's kangaroo rat. It is native to western North America.

[13] This is an intracellular protozoan parasite. It causes avian coccidiosis in chickens.

Table 4.3 Relationship between $\mathscr{K}$ and $\sum_{i=1}^{t} \mathbb{F}(g_i) + \mathbb{F}(\mathscr{R}(g_i))$ value

k	$\mathscr{K}$	Equation (3.17) value	$\lvert\mathscr{W}^k\rvert$
1	1	0.499138	4
2	1	0.489046	16
3	1	0.499310	64
4	1	0.500199	256
5	1	0.499396	1024
6	1	0.498957	4096
7	1	0.499152	16,384
8	1	0.498378	65,536
9	1	0.499167	262,144
10	1	0.498525	1,048,576
11	0.999776	0.498808	4,194,304
12	0.990219	0.499034	16,777,216
13	0.929167	0.499520	67,108,864
14	0.756606	0.502097	268,435,456
15	0.510436	0.517065	1,073,741,824
16	0.264477	0.577907	4,294,967,296
17	0.098486	0.686536	17,179,868,184
18	0.007994	0.780472	68,719,476,736

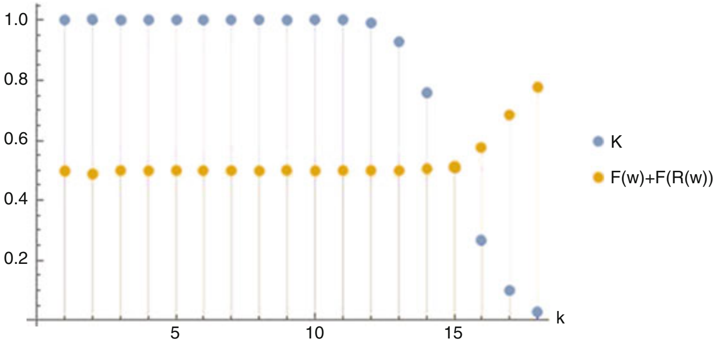


Fig. 4.14 The values of $\mathscr{K}$ and Eq. (3.17) at the same figure

There is no perfect theory, but a good one should pass critical tests until it eventually fails and gives way to a new theory.[14] In addition to the trials described in

[14] According to the Modern Evolutionary Synthesis, biologists used to believe that there was no relationship between the direction in which mutations occur and the direction that would lead to enhanced fitness (Random genetic variation); however, new data suggest that some phenotypic variants are more likely than others (non-random phenotypic variation). Similarly, it was taken for granted that acquired characters were not inherited (Genetic inheritance), yet new evidence imply that acquired characters can play evolutionary roles by biasing phenotypic variants subject to selection (Inclusive inheritance) [61].

this chapter, we've performed many other experiments to assess our findings. So far, they've passed all tests. For instance, from the very beginning, we've conjectured that our conceptual theoretical framework could be used to generalize Szybalski's transcription rule as well. In fact, we've been working on this project, and our preliminary results are definitely promising. Another important line of research is the study of those few genome sequences that did not follow our new rules. What can we learn from them? Did they fail for some idiosyncratic biological reason? Did their genome assemblies have major mistakes? Whatever the case, we are sure that we will learn a lot.

4.3 The Synergy Between Mathematics and Biology

Before finishing this chapter, we would like to reinforce the special role that Mathematics must play in Biology. In 2004, Dr. Joel E. Cohen published an article [19] in which he described the impact of the discovery of the microscope on the seventeenth century. Arguably, the best word to describe this event is "revolution." Suddenly, biologists realized that there were living beings that were invisible to the naked eye. A few people suspected their existence, but without solid evidence, these beings belonged to the realm of fairy tales. The microscope was fundamental for Biology to see the reality that our natural vision was unable to reach. Dr. Cohen also affirmed that "Mathematics is Biology's next microscope, only better." From our point of view, his analogy is almost perfect. Indeed, Mathematics may help us to see otherwise "invisible" realities.

For instance, all the properties of the Higgs boson particle[15] were mathematically predicted [18] before mankind had the technology to measure them.[16] There are several such examples in science. However, instead of tiring the reader with other cases, we would instead like to emphasize that Mathematics helps us to see "abstract" entities that nevertheless obey strict logical rules. The relationship between actual physical objects and those abstract entities is an endless discussion in Philosophy. They may *not* be related to each other, but there is a lot empirical evidence showing that they are. Thus, Mathematics may serve as a lens to helps us to see "invisible" patterns in biological data.

Dr. Cohen listed five biological challenges that could be addressed by mathematicians. His first item reads as follows:

> Understand cells, their diversity within and between organisms, and their interactions with biotic and abiotic environments. The complex networks of gene interactions, proteins, and signaling between the cell and other cells and the abiotic environment is probably incomprehensible without some mathematical structure perhaps yet to be invented.

[15] Also known as the *God Particle*.

[16] In other words, Higgs boson properties were not discovered through direct observation.

We fully agree with him. We would just like to add that the same reasoning applies to the intrinsic properties of DNA. The rules presented in this book would not have been discovered without the mathematical structure that we invented. We hope that the empirical evidence presented in this chapter is enough to convince the reader that DNA sequences actually have some underlying mathematical structure.

This book is a good example of the idea that Mathematics is Biology's next Microscope, only better. We hope that our fellow mathematicians will build upon our work, and, in order to tackle even more complex intrinsic properties of DNA, develop new mathematical tools that show that *"Biology is Mathematics' next Physics, only better."*

Postscript

> And the LORD God said, Behold, the man is become as one of us, to know good and evil: and now, lest he put forth his hand, and take also of the tree of life, and eat, and live for ever (Genesis 3, 22)

Will there be a future? This is another ancient and unanswered question. We believe that all preceding generations were confronted with it to an extent, and all cultures have likely wondered whether humanity could last much longer. For instance, in the early twentieth century, José Ortega y Gasset (1883–1955) wrote:

> No one knows toward what center human things are going to gravitate in the near future, and hence the life of the world has become scandalously provisional [39].

Yet, humanity still thrives, and will likely continue to do so for while.

The previous generation witnessed the split of the nucleus of the atom. The sound produced by that fission killed thousands. For the first time in 4.280 billion years, a single living being was able to destroy the whole world. We take for granted that no rational being would dare to do it, but someone has already informed us that reason appeared relatively late over the course of human evolution, and it is just a thin layer over our beastly nature. Nevertheless, we are still here, wondering and wandering around. Only statistics can explain why the tragedy has not yet occurred: few individuals possess access to nuclear weapons.

Nowadays, as Chargaff said, we are meddling with another nucleus: the cell's. Once again, another nucleus has given us the power to eliminate all life on earth. Life is indeed very fragile. The difference, this time, is that the planet would remain intact. It would continue to orbit the sun for ages. No living beings at all. Only the lifeless minerals and the ashes of what used to be alive. No one to regret the past errors or to speculate about the future.

Atoms and cells are invisible to the naked eye. It is amazing that such minuscule entities hide within themselves such enormous destructive power. Yet this is an one-sided and unfair picture. We cannot forget to mention that nuclear energy delivers electric power to whole cities, and that, in the future, it may be the fuel that will be used to take us to other planets. Analogously, Synthetic Biology may be applied to cure several human diseases, and, if employed in agriculture or animal breeding, may save mankind from starvation due to a food shortage caused by climate change. All of this means that both atom and cell nuclei may be adopted to benefit the biosphere as well. Who decides?

The only animal able to choose is the *Homo sapiens*. The Latin root of the word "sapiens" means *wise*.[1] Wisdom and Knowledge are very different things. While there is no doubt that our world is full of knowledge, the same cannot be said of wisdom. And it is *wisdom* that we desperately need more of. The last few centuries have shown that knowledge without wisdom produces a paradoxical effect: instead of expanding human horizons, it reduces them to a single question which was best expressed by Albert Camus (1913–1960) in the opening line of his book *The Myth of Sisyphus*:

> There is only one really serious philosophical question, and that is suicide.

More *wisdom* than knowledge is necessary to refute the "Devil's doctrine" mentioned by Chargaff. Our ancestors knew that some deeds cannot be undone, and unless we are absolutely sure what we are doing, we should not trifle with matters of life and death. Why are we lying to ourselves and ignoring this common sense advise? Chargaff was right about the author of the maxim "what can be done must be done," for it certainly came surreptitiously from the *very center of Hell* where Dante Alighieri (1265–1321) put Lucifer.

We hold that this fragile thing called *life* lies between two *massive events*. The first was discovered by the Catholic priest Georges Lamaître (1894–1966), and it was the origin of our Universe. We do not know yet the nature of the second. It may be another tragic collision with a celestial body, or the "Big Freeze" that some physicists predict based on the latest cosmological evidence. But we can also force the ultimate extinction of life to come earlier either as a deafening *mushroom cloud* or as a silent *gene drive* that will sweep away all living organisms from the face of our planet. Nothing can be done to prevent the laws of Physics from following their natural course, but *homo Sapiens* can avoid the last two possibilities. On a geological time scale, our species is still in the early embryonic stage. It would be a pity to discover which cataclysm will put an end to all life, before realizing "what life is."

> Lord, forgive them, for they know what they do!
> (Karl Kraus [59])

Acknowledgements We would like to thank the Embrapa Multiuser Bioinformatics Laboratory (Laboratório Multiusuário de Bioinformática da Embrapa) for providing the high-performance computational infrastructure. This work was partially funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)—grant 304549/2011-8.

[1] Whoever chose the Latin word *sapiens* instead of *cogitans* was aware of the difference between them.

References

1. Akbari, O.S., Bellen, H.J., Bier, E., Bullock, S.L., Burt, A., Church, G.M., Cook, K.R., Duchek, P., Edwards, O.R., Esvelt, K.M., Gantz, V.M., Golic, K.G., Gratz, S.J., Harrison, M.M., Hayes, K.R., James, A.A., Kaufman, T.C., Knoblich, J., Malik, H.S., Matthews, K.A., O'Connor-Giles, K.M., Parks, A.L., Perrimon, N., Port, F., Russell, S., Ueda, R., Wildonger, J.: Safeguarding gene drive experiments in the laboratory - multiple stringent confinement strategies should be used whenever possible. Science **349**(6251), 927–929 (2015)
2. Albrecht-Buehler, G.: Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc. Natl. Acad. Sci. U. S. A. **103**(47), 17828–17833 (2006)
3. Aristotle: Metaphysics, Book I, 985b
4. Avery, O.T., MacLeod, C.M., McCarty, M.: Studies on the chemical nature of the substance inducing transformation of pneumococal types. J. Exp. Med. **79**, 137 (1944)
5. Baisnée, P.-F., Hampson, S., Baldi, P.: Why are complementary DNA strands symmetric? Bioinformatics **18**, 1021–1033 (2002)
6. Bannert, N., Kurth, R.: Retroelements and the human genome: new perspectives on an old relation. Proc. Natl. Acad. Sci. U. S. A. **101**, 14572–14579 (2004)
7. Bird, A.P.: DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**(7), 1499–1504 (1980)
8. Bird, A.P.: DNA methylation patterns and epigenetic memory. Genes Dev. **16**, 6–21 (2002)
9. Breitling, R., Takano, E., Gardner, T.S.: Judging synthetic biology risks. Science **347**(6218), 107 (2015)
10. Chargaff, E.: Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experimentia **6**, 201 (1950)
11. Chargaff, E.: Structure and function of nucleic acids as cell constituents. Fed. Proc. **10**, 654–659 (1951)
12. Chargaff, E.: A quick climb up mount olympus. Science New Ser. **159**(3822), 1448–1449 (1968)
13. Chargaff, E.: Preface to a grammar of biology: a hundred years of nucleic acid research. Science **172**, 637–642 (1971)
14. Chargaff, E.: Heraclitean Fire: Sketches from a Life Before Nature. The Rockefeller University Press, New York (1978)
15. Chargaff, E.: How genetics got a chemical education. Ann. N. Y. Acad. Sci. **325**, 345–360 (1979)
16. Chernoff, Y.O.: Mutation processes at the protein level: is Lamarck back? Mutat. Res. Rev. Mutat. Res. **488**(1), 39–64 (2001)

17. Chesterton, G.K.: Orthodoxy. Garden City, New York (1959)
18. CMS Collaboration: Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Phys. Lett. B **716**, 30–61 (2012)
19. Cohen, J.E.: Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. Plos ONE **2**(12), e439 (2004)
20. Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., and the members of the DOE and NIH planning groups: New goals for the U.S. Human Genome Project: 1998–2003. Science **282**(5389), 682–689 (1998)
21. Courant, R., Robbins, H.: What is Mathematics? An Elementary Approach to Ideas and Methods. Oxford University Press, New York (1996)
22. Devlin, K.: Mathematics: The Science of Patterns - The Search for Order in Life, Mind, and the Universe. Scientific American Library/Henry Holt and Company, New York (1997)
23. Dobzhansky, T.: Nothing in biology makes sense except in the light of evolution. Am. Biol. Teach. **35**(3), 125–129 (1973)
24. Dodd, M.S., Papineau, D., Grenne, T., Slack, J.F., Rittner, M., Pirajno, F., O'Neil, J., Little, C.T.S.: Evidence for early life in Earth's oldest hydrothermal vent precipitates. Nature **543**, 60–64 (2017)
25. Dolgin, E.: The genome finishers. Nature **462**, 17 (2009)
26. Dong, Q., Cuticchia, A.J.: Compositional symmetries in complete genomes. Bioinformatics **17**(6), 557–559 (2001)
27. Doolittle, W.F.: Is junk DNA bunk? a critique of ENCODE. Proc. Natl. Acad. Sci. U. S. A. **110**(14), 5294–5300 (2013)
28. Duret, L., Galtier, N.: The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. Mol. Biol. Evol. **17**(11), 1620–1625 (2000)
29. Eco, U.: On Beauty: A History of a Western Idea. Seeker & Warburg, London (2004)
30. Eddy, S.R.: Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. **2**(12), 919–929 (2001)
31. Fibonacci, L., Singler, L.E. (Translator): Fibonacci's Liber Abaci. Springer, New York (2002)
32. Flintoft, L.: Filling gaps in the human genome. Nat. Rev. Genet. **14**, 676 (2013)
33. Forsdyke, D.R.: Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J. Mol. Evol. **41**, 573–581 (1995)
34. Forsdyke, D.R., Bell, S.J.: Purine-loading, stem-loops, and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. Appl. Bioinformatics **3**, 3–8 (2004)
35. Gamow, G.: Possible relation between deoxyribonucleic acid and protein structures. Nature **173**, 318 (1954)
36. Gantz, V.M., Bier, E.: The mutagenic chain reaction: a method for converting heterozygous to homozygous mutations. Science **348**(6233), 442–444 (2015)
37. Gantz, V.M., Jasinskiene, N., Tatarenkova, O., Fazekas, A., Macias, V.M., Bier, E., James, A.A.: Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. Proc. Natl. Acad. Sci. U. S. A. **112**, E6736–E6743 (2015)
38. Gardner, M.: The Colossal Book of Mathematics. W. W. Norton & Company, New York (2001)
39. Gasset, J.O.: The Revolt of the Masses. W. W. Norton & Company, New York (1993)
40. Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., Snyder, M.: What is a gene, post-ENCODE? history and update definition. Genome Res. **17**, 669–681 (2007)
41. Grantham, R.: Workings of the genetic code. Trends Biochem. Sci. **5**(12), 327–331 (1980)
42. Graur, D., Zheng, Y., Price, N., Azevedo, R.B.R., Zufall, R.A., Elhaik, E.: On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol. Evol. **5**(3), 578–590 (2013)

43. Gregory, T.R.: The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. Ann. Bot. **95**, 133–146 (2005)
44. Gulia-Nuss, M., Nuss, A.B., et al.: Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat. Commun. **7**, 10507 (2016)
45. Hardy, G.H.: A review – the psychology of invention in the mathematical field. Math. Gaz. **30**, 111–115 (1946)
46. Hardy, G.H.: A Mathematician's Apology. Cambridge University Press, Cambridge (2012)
47. Hargittai, I., Hargittai, M.: Candid Science: Conversations with Famous Chemists. Imperial College Press, London (2000)
48. Hilbert, D., Ackermann, W.: Principle of Mathematical Logic. AMS Chelsea Publishing, Providence (1999)
49. Hutchison III, C.A.: DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. **35**(18), 6227–6237 (2007)
50. Hutchison III, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., Pelletier, J.F., Qi, Z.-Q., Richter, R.A., Strychalski, E.A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K.S., Smith, H.O., Glass, J.I., Merryman, C., Gibson, D.G., Venter, J.C.: Design and Synthesis of a minimal bacterial genome. Science **351**(6280), 1414 (2016)
51. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature **409**, 860–921 (2001)
52. Jones, P.A., Takai, B.: The role of DNA methylation in mammalian epigenetics. Science **293**(5532), 1068–1070 (2001)
53. Jones, M.D., Forn, I., Gadelha, C.M., Egan, M.J., Massana, R., Richards, T.A.: Discovery of novel intermediate forms redefines the fungal tree of life. Nature **474**, 200–203 (2011)
54. Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E.V.: Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. **19**, 68–72 (2003)
55. Karkas, J.D., Rudner, R., Chargaff, E.: Separation of B. subtilis DNA into complementary strands, II. Template functions and Composition as determined by transcription with RNA polymerase. Proc. Natl. Acad. Sci. U. S. A. **60**, 915–920 (1968)
56. Kline, M.: Mathematics for the Nonmathematician. Dover, New York (1967)
57. Kong, S.-G., Fan, W.-L., Chen, H.-D., Hsu, Z.-T., Zhou, N., et al.: Inverse symmetry in complete genomes and whole-genome inverse duplication. PLoS ONE **4**(11), e7553 (2009). doi:10.1371/journal.pone.0007553
58. Koonin, E.V., Wolf, Y.I.: Is evolution Darwinian or/and Lamarckian? Biol. Direct **4**, 42 (2009)
59. Kraus, K.: Half-Truths & One-and-a-Half Truths, Karl Kraus Selected Aphorisms. The University of Chicago Press, Chicago (1990)
60. Krauss, L.M.: A Universe from Nothing: Why There Is Something Rather Than Nothing. Atria Books, Miami (2013)
61. Laland, K.N., Uller, T., Feldman, M.W., Sterelny, K., Muller, G.B., Moczek, A., Zablonka, E., Odling-Smee, J.: The extended evolutionary synthesis: its structure, assumptions and predictions. Proc. R. Soc. B **282**, 20151019 (2015)
62. Landrain, T., Meyer, M., Perez, A.M., Sussan, R.: Do-it-yourself biology: challenges and promises for an open science and technology movement. Syst. Synth. Biol. **7**, 115–126 (2013)
63. Lartigue, C., Glass, J., Alperovich, N., Pieper, R., Parmar, P.P., Hutchison 3rd, C.A., Smith, H.O., Venter, J.C.: Genome transplantation in bacteria: changing one species to another. Science **317**(5838), 632–638 (2007)
64. Ledford, H.: Garage biotech: life hackers. Nature **467**, 650–652 (2010)
65. Lonergan, B.J.F.: Insight: a study of human understanding. Philosophical Library, New York (1965)
66. Martin, J., et al.: The sequence and analysis of duplication-rich human chromosome 16. Nature **432**, 988–994 (2004)
67. Matthew, P.: On Naval Timber and Arboriculture; with Critical Notes on Authors Who Have Recently Treated the Subject of Planting. Adam Black, London (1831)

68. McManus, I.C.: The aesthetics of simple figures. Br. J. Psychol. **71**, 505–525 (1980)
69. Metzker, M.L.: Application of next-generation sequencing technologies - the next generation. Nat. Rev. Genet. **11**(1), 31–46 (2010)
70. Miescher, F.: Ueber die chemische Zusammensetzung der Eiterzellen. Medicinisch-chemische Unters. **4**, 441–460 (1871)
71. Mitchell, D., Bridge, R.: A test of Chargaff's second rule. Biochem. Biophys. Res. Commun. **340**, 90–94 (2006)
72. Newton, I.: The Principia - Mathematical Principles of Natural Philosophy. Snowball Publishing, Dallas (2010)
73. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research. Springer, New York (2000)
74. Perez, J.-C.: Chaos, DNA and neuro-computers: a golden link. Specul. Sci. Technol. **14**, 336–346 (1991)
75. Pertea, M., Salzberg, S.L.: Between a chicken and a grape: estimating the number of human genes. Genome Biol. **11**(5), 206 (2010)
76. Polya, G.: Mathematics and Plausible Reasoning. Princeton University Press, New Jersey (1954)
77. Polya, G.: How to Solve It: A New Aspect of Mathematical Method. Princeton University Press, New Jersey (1988)
78. Prabhu, V.V.: Symmetry observation in long nucleotide sequences. Nucleic Acids Res. **21**, 2797–2800 (1993)
79. Salzberg, S.L., Yorke, J.A.: Beware of mis-assembled genomes. Bioinformatics **21**, 4320–4321 (2005)
80. Sanger, F.: Determination of Nucleotide Sequences in DNA. Nobel Lecture, 8 December 1980
81. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchinson, C.A., Slocombe, P.M., Smith, M.: Nucleotide sequence of bacteriophage $\Phi X174$ DNA. Nature **265**(5596), 687–695 (1977)
82. Sano, H.: DNA methylation and Lamarckian inheritance. Proc. Jpn. Acad. **78**, Ser. B, 293–298 (2002)
83. Shah, H., Warwick, K., Vallverdú, J., Wu, D.: Can machines talk? comparison of Eliza with modern dialogue systems. Comput. Hum. Behav. **58**, 278–295 (2016)
84. Shapiro, B.J., Leducq, J.-B., Mallet, J.: What is speciation? PLoS Genet. **12**(3), e1005860 (2016)
85. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. **147**, 195–197 (1981)
86. Staden, R., Beal, K.F., Bonfield, J.K.: The Staden package. Methods Mol. Biol. **132**, 115–130 (2000)
87. Stark, R.: How the West Won: The Neglected Story of the Triumph of Modernity. ISI Books, Delaware (2015)
88. Stove, D.C.: Against the Idols of the Age. Transaction Publishers, New Brunswick (1999)
89. Stove, D.C.: Scientific Irrationalism: Origins of a Postmodern Cult. Transaction Publishers, New Brunswick (2011)
90. Swartz, M.N., Trautner, T.A., Kornberg, A.: Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J. Biol. Chem. **237**, 1961–1967 (1962)
91. Telenti, A., Pierce, L.C.T., Biggs,W.H., di Iulio, J., Wong, E.H.M., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C., Brewerton, S.C., Bulsara, N., Garner, C., Metzker, G., Sandoval, E., Perkins, B.A., Och, F.J., Turpaz, Y., Venter, J.C.: Deep sequencing of 10,000 human genomes. Proc. Natl. Acad. Sci. U. S. A. **113**(42), 11901–11906 (2016)
92. The ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature **489**, 57–74 (2012)
93. Turing, A.M.: Computing machinery and intelligence. Mind **49**, 433–460 (1950)
94. Venter, J.C., et al.: The sequence of the human genome. Science **291**(5507), 1304–1351 (2001)

95. Vogel, F.: A preliminary estimate of the number of human genes. Nature **201**, 847 (1964)
96. Waterman, M.S.: Introduction to Computational Biology: Maps, Sequences and Genomes. Interdisciplinary Statistics. Chapman & Hall/CRC, Boca Raton (1997)
97. Watson, J.D.: The Double Helix: A Personal Account of the Discovery of the Structure of DNA. Atheneum, New York (1968)
98. Watson, J.D., Crick, F.H.C.: Molecular structure of nucleic acids. Nature **4356**, 737 (1953)
99. Woese, C., Fox, G.: Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. **74**(11), 5088–5090 (1977)
100. Wright, A.V., Nuñez, J.K., Doudna, J.A.: Biology and application of CRISPR systems: harnessing natures's toolbox for genome engineering. Cell **164**, 29–44 (2016)
101. Yamagishi, M.E.B., Herai, R.H.: Chargaff's grammar of biology: new fractal-like rules (2011). arXiv:1112.1528
102. Yamagishi, M.E.B., Shimabukuro, A.I.: Nucleotide frequencies in human genome and Fibonacci numbers. Bull. Math. Biol. **70**, 643–653 (2008)
103. Zinoviev, V.V., Yakishchik, S.I., Evdokimov, A.A., Malygin, E.G., Hattman, S.: Symmetry elements in DNA structure important for recognition/methylation by DNA [amino]-methyltransferases. Nucleic Acids Res. **32**(13), 3930–3934 (2004)

Index