

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

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

²Unfortunately, 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's⁴ 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]

³From ancient Greek: order.

⁴A 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 programs⁵ 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 Consortium⁶ 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

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

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

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

⁸Grape 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 "**ENCyclopedia Of DNA Elements**," 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

⁹As 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. Waterman¹² 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 multidisciplinary 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?

¹⁰1 petabytes (PB) = 10^{15} bytes.

¹¹1 zettabyte = 10^{21} bytes.

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

¹³The 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 superfluous 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, Nosedal 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 $\mathbb{F}(T) + \mathbb{F}(C) \approx \frac{1}{2}$.

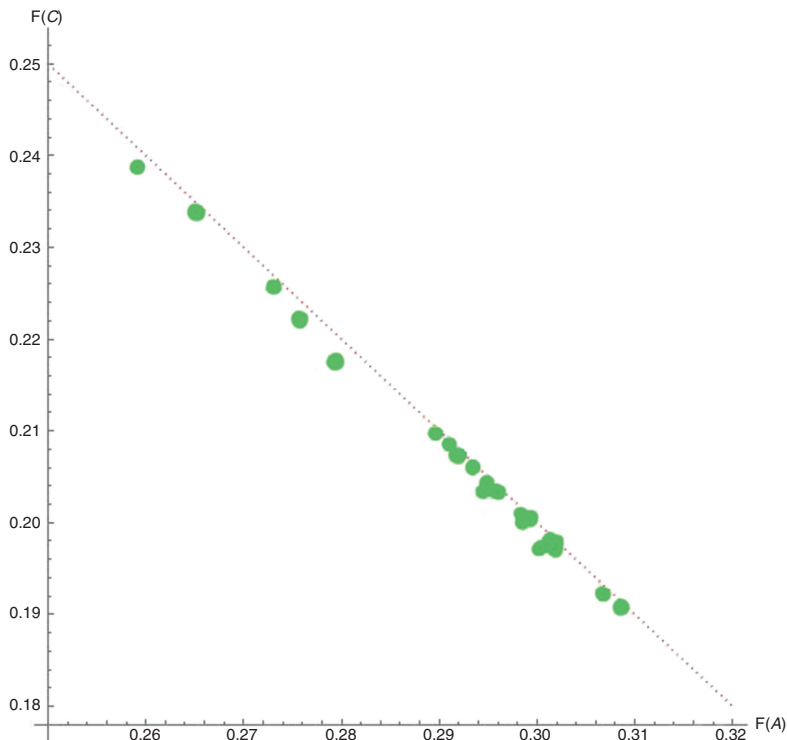


Fig. 2.1 The *red dotted line* $F(A) + F(C) \approx \frac{1}{2}$, and, in *green*, the observed points $(F(A), F(C))$ for each human chromosome (NCBI Build GRCh38.p2)

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

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

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 $F(A) + F(C) = \frac{1}{2}$, and, the green dots depict the Adenine and Cytosine frequencies, $(F(A), F(C))$, 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.

¹⁶NCBI 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.

¹⁷This work took 2 years to be published.

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

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

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:

$$F(n + 2) = F(n + 1) + F(n), \quad (2.3)$$

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:

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

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

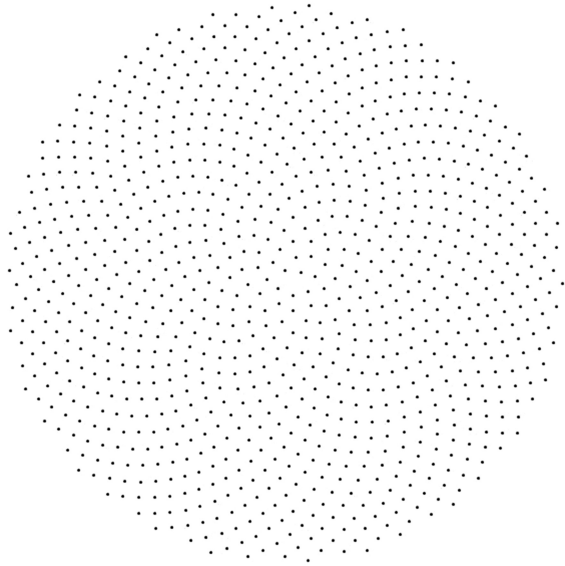
$$\frac{a + b}{a} = \frac{a}{b}$$

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

¹⁸Leonardo da Vinci.

¹⁹Notably in Bartók and Debussy.

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



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 Fibonacci²⁰ numbers and their quotients below:

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

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:

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

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

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

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

where $k = 0, 1, 2, 3, \dots, N$, 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

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

²⁰A 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

$$\mathbb{F}(x) = \lim_{n \rightarrow \infty} \mathbb{F}(x_n) \quad (2.10)$$

$$\mathbb{F}(y) = \lim_{n \rightarrow \infty} \mathbb{F}(y_n) \quad (2.11)$$

$$\mathbb{F}(z) = \lim_{n \rightarrow \infty} \mathbb{F}(z_n) \quad (2.12)$$

$$\mathbb{F}(w) = \lim_{n \rightarrow \infty} \mathbb{F}(w_n) \quad (2.13)$$

It is also necessary to understand what happens with the quotients

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

and

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

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

$$F(n+k) = F(k)F(n+1) + F(k-1)F(n). \quad (2.14)$$

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

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

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

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

However, we know that

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

and

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

Thus, Eq. (2.16) can be written as

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

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

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

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

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

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

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

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

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

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

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

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

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

Equations (2.25) and (2.27) imply that

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

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

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

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

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

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 variables²¹ $\mathbb{F}(x)$ and $\mathbb{F}(y)$; namely:

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

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

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

which can be rewritten as

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

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

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

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

²¹Note 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.

²²In fact, only when $k \rightarrow \infty$ 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

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

where

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

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 $(\mathbb{F}(x), \mathbb{F}(y)) \rightarrow (0.25, 0.25)$ as $k \rightarrow \infty$.

2.2.4 Experiment Follow-Up

More than a decade ago,²³ we performed an experiment using the available human genome data²⁴ to assess our mathematical model. The results were encouraging, as the observed data deviated less than 0.005 from the predicted values. However,

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

²⁴The 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	$F(x)$	$F(x) \cong$	$F(y)$	$F(y) \cong$
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 gaps²⁵ 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 0.7Mbps 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,

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

²⁶Though, 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	$F(A)$	$F(C)$	$F(T)$	$F(G)$	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 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 $(F(A), F(C))$ 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

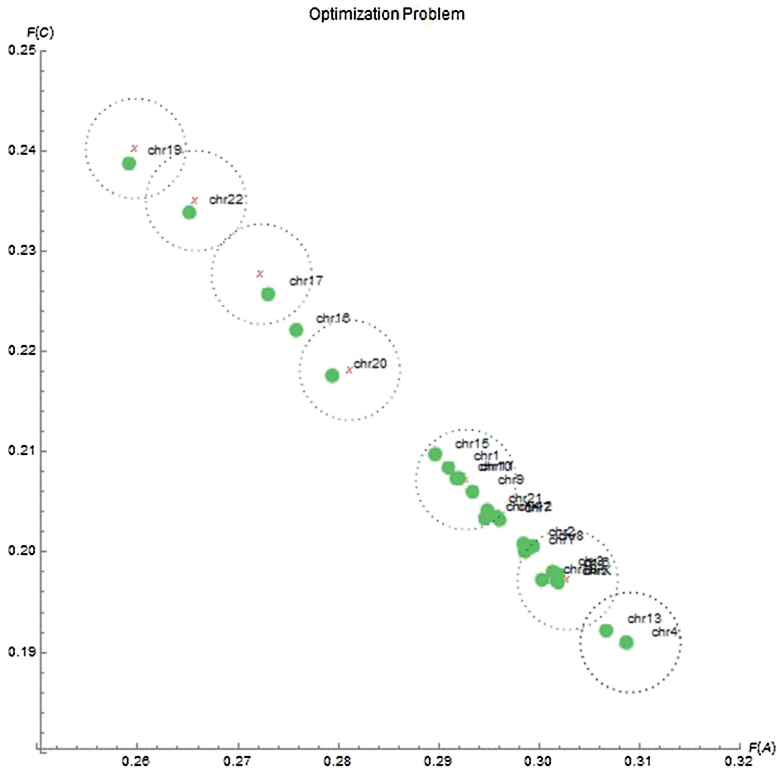


Fig. 2.3 In the figure, the solutions to the optimization problem are depicted as *red crosses*. Every *dotted circle* has the same radius ($r = 0.005$). The *green dots* show the nucleotide frequencies for each one of the human chromosomes. Unfortunately, due to their proximity, some chromosome labels 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.

Mathematical Grammar of Biology

Yamagishi, M.E.B.

2017, XII, 82 p. 19 illus., 17 illus. in color., Softcover

ISBN: 978-3-319-62688-8