Research Article

Shannon Information and Power Law Analysis of the Chromosome Code

J. A. Tenreiro Machado

Department of Electrical Engineering, Institute of Engineering of Polytechnic of Porto,
Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal

Correspondence should be addressed to J. A. Tenreiro Machado, jtm@isep.ipp.pt

Received 8 June 2012; Revised 17 August 2012; Accepted 21 August 2012

1. Introduction

During the last years the genome sequencing project produced a large volume of data that is presently available for computational processing [1–14]. Researchers have been tackling the information content of the deoxyribonucleic acid (DNA), but interesting questions remain still open [15–21].

This paper addresses the information flow along each DNA strand. For this purpose several statistics are developed, and the relative frequencies of distinct types of symbol associations are evaluated. The concepts of character, word, word delimiter, and phrase are defined, and the information content of each chromosome message is quantified. Power law (PL) relationships emerge in the information locus. PL distributions, often known as heavy tail distributions, Pareto laws, Zipf laws, or others, have been largely reported in the modeling of distinct real phenomena [22–31]. It was recognized [11, 32–34] that DNA has an information structure that reveals long range behavior, somehow in the line of thought of systems with dynamics described by the tools of Fractional Calculus (FC) [35–37]. It is...
known the existence of a strong relationship between FC and PL; nevertheless, up to the present state of knowledge, no formal demonstration supported that observation based on empirical and experimental measurements. Therefore, it is not a surprise that both FC and PL descriptions emerge when analyzing DNA with distinct mathematical tools. In the present study PL descriptions are applied for condensing the charts characterizing the chromosomes of twenty-three species.

Having these ideas in mind, this paper is organized as follows. Section 2 presents the DNA sequence decoding concepts, the mathematical tools and formulates the algorithm that computes the information for each chromosome and species. Section 3 analyzes the DNA information dynamical content of 463 chromosomes corresponding to a set of twenty-three species. Finally, Section 4 outlines the main conclusions.

2. Preliminary Notes on the DNA Information

In the DNA double helix there are four distinct nitrogenous bases, namely, thymine, cytosine, adenine, and guanine, denoted by the symbols \{T, C, A, G\}. Each type of base on one strand connects with only one type of base on the other strand, forming the base pairing A – T and G – C. Besides the four symbols \{T, C, A, G\}, the available chromosome data includes a fifth symbol “N” which is believed to have no practical meaning for the DNA decoding.

For processing the DNA information a possible technique is to convert the symbols into a numerical value. In previous papers was adopted the direct symbol translation $1 + i0, C = -1+i0, T = 0+i, G = 0-i, N = 0+i0$, where $i = \sqrt{-1}$. We can move along the DNA strip, one symbol (base) at a time. The resulting values form a “signal” $x(t)$ where “t” can be interpreted as a pseudotime. The signal can be treated by the Fourier transform $F\{x(t)\} = \int_{-\infty}^{+\infty} x(t)e^{-i\omega t} dt$, where $\omega$ represents the angular frequency.

Figure 1 shows one example with the amplitude of the Fourier transform for chromosome 1 of the human being. The frequency interval $10^{-1} \leq \omega \leq 10^{0}$ is adopted and a PL approximation is superimposed revealing a strong correlation.
This technique has, however, one drawback which is the initial assignment of numerical values to the DNA symbols. Therefore, it is important to design an alternative method of analysis avoiding that problem, but, on the other hand, capable of revealing fractional order phenomena. Bearing this strategy in mind, in this paper is adopted an approach based on the histograms of symbol alignment, information theory, and PL approximations.

This study focuses over twenty-three species yielding a space of 463 chromosomes. Therefore, denoting by $N_j$ the number of chromosomes of species $j = 1, \ldots, 23$, we consider the \{Species, Tag, $N_j$\} given by \{Mosquito (Anopheles gambiae), Ag, 5\}_1, \{Honeybee, (Apis mellifera), Am, 16\}_2, \{Caenorhabditis briggsae, Cb, 6\}_3, \{Caenorhabditis elegans, Ce, 6\}_4, \{Chimpanzee, Ch, 25\}_5, \{Dog, Dg, 39\}_6, \{Drosophila simulans, Ds, 6\}_7, \{Drosophila yakuba, Dy, 10\}_8, \{Horse, Eq, 32\}_9, \{Chicken, Ga, 31\}_10, \{Human, Ho, 24\}_11, \{Medaka, Me, 24\}_12, \{Mouse, Mm, 21\}_13, \{Opposum, Op, 9\}_14, \{Orangutan, Or, 24\}_15, \{Cow, Ox, 30\}_16, \{Pig, Po, 19\}_17, \{Rat, Rn, 21\}_18, \{Yeast (Saccharomyces cerevisiae), Sc, 16\}_19, \{Stickleback, St, 21\}_20, \{Zebra Finch, Tg, 32\}_21, \{Tetraodon, Tn, 21\}_22 and \{Zebrafish, Zf, 25\}_23.

The DNA information decoding is addressed in this paper, and we start by defining the underlying concepts. The fundamental unit is the “symbol” that, in our case, consists in one of the four possibilities \{T, C, A, G\}, while “N” is simply disregarded. Each “character” is represented by an $n$-tuple association ($n = 1, 2, \ldots$) of the 4 symbols, resulting in a total of $4^n$ possible symbols per character. For example, with $n = 2$ we get a maximum of $4^2$ characters represented by the 16 two-symbol sequences \{TT, TC, TA, TG, CT, CC, CA, CG, AT, AC, AA, AG, GT, GC, GA, GG\}. The sequences are obtained when moving sequentially along the DNA. The characters may have different significance and are divided into two classes, namely, characters with relevant information, to be denoted in the sequel as “word characters,” and delimiters denoted as “spaces.” Therefore, joining consecutive “word characters” yields a “word,” that ends in the presence of one or more consecutive “spaces” (i.e., multiple spaces are considered as a single space). When the complete association of consecutive words is fulfilled, we obtain a “message.”

Figure 2 depicts a simple example of a message with 21 symbols and 3 words. The message \{ACTACGTTGCTTACGTTAACCA\} is processed according to the proposed scheme for $n = 2$ and considering the 2 sequences \{TT, AA\} as spaces, and the 14 sequences \{TC, TA, TG, CT, CC, CA, CG, AT, AC, AG, GT, GC, GA, GG\} as characters. Therefore, the resulting words are \{AC TA CG\}, \{GG GT TC AG\} and \{CC\}.

We verify that we may have words with different lengths and that it is considered as a single space any repetition of spaces. The message finishes when the end of the DNA strand is attained, and, therefore, it is not considered the case of multiple messages for each chromosome.
After defining the concepts for symbol, character (with the categories of word character and space), and message, we need to establish the numerical value to be adopted by \( n \) and the method for measuring the information. In what concerns \( n \) no \textit{a priori} optimal value is considered. Therefore, in the experiments is analyzed the influence when going from \( n = 1 \) up to \( n = 12 \), or, correspondingly, when going from \( 4^1 \) up to \( 4^{12} \) symbols per character. This evaluation is performed for one chromosome. Based on this first assessment, given the huge computational load required by high values of \( n \), the set of twenty-three species, totaling 463 chromosomes, is analyzed for \( n \in \{1, \ldots, 8\} \). In what concerns the information measurement it is adopted the Shannon information \([38-49]\) \( I_i = -\ln(p_i) \) where \( I_i \) represents the quantity of information of event \( i \) that has a probability \( p_i \). In this topic we can refer to [50] calculating also the Shannon information for short DNA words of differing lengths, where the authors find that genomes share universal statistical properties. It is also worth mentioning that other entropies, such as the Rényi, Tsallis, and Ubriaco definitions \([51, 52]\) were tested. Nevertheless, experiments with these expressions and distinct numerical values of the parameters did not reveal any significant conceptual difference. Therefore, for simplicity in the sequel it is adopted merely the Shannon definition.

In our case, for a \( n \)-tuple symbol encoding, the occurrence of the \( i \)th character within the \( 4^n \) set has probability \( p_{\text{char},n}^i \) leading to information \( -\ln(p_{\text{char},n}^i) \), and, therefore, the total information content of a word \( I_{\text{word},n} \) yields

\[
I_{\text{word},n} = -\sum_{i=1}^{m} \ln(p_{\text{char},n}^i),
\]

where \( m \) represents the total number of word characters including the first space. In fact, it was numerically evaluated the effect of including, or not, the space information but, due to its low importance, the final effect is negligible. Therefore, it is considered the inclusion of one space as the information for delimiting the word, while further consecutive repetitions of spaces are disregarded.

The message information is the sum of all word information:

\[
I_{\text{mes},n} = \sum_{i=1}^{r} I_{\text{word},n},
\]

where \( r \) denotes the total number of words included in the message (i.e., the chromosome).

The information measurement requires the knowledge of \( p_{\text{char},n}^i \). While we can expect an equilibrium of probabilities for \( n = 1 \), that may be not true for larger values of \( n \). Therefore, in the sequel it is adopted a numerical procedure that starts by reading the chromosome message based on the \( n \)-tuple character setup leading to the construction of one histogram per chromosome. In the set of \( 4^n \) bins are chosen, by inspection, those that are more frequent (and have smaller information content) for the role of spaces. In a second phase, the relative frequencies, which are adopted as approximants to the probabilities, and the information values (2.1) and (2.2) are calculated numerically while traveling along the DNA strand.

This strategy does not consider some \textit{a priori} optimal value of \( n \). Therefore, as mentioned previously, several distinct values of \( n \) will be studied before establishing any conclusions.
3. Capturing the DNA Information

We start by considering Human chromosome 12 (Ho12) and $n = \{1, \ldots, 12\}$. This chromosome is represented by a medium size file (130 Mbytes) and may be considered a good compromise between length and computational load.

Figure 3 depicts the histograms for $n = \{1, 2, 3, 4\}$ where, for simplifying the visualization, the characters are ordered by decreasing magnitude of relative frequency. For the histograms construction two counting methods were envisaged: (i) counting with disjoint set of $n$ symbols and (ii) counting the sets while sliding one symbol at a time. At first sight it seems that (i) is the most straightforward, but if we consider that we do not have reliable information for starting and synchronizing the counting, then method (ii) is more robust and, therefore, is adopted in the sequel.

Figure 4 shows the word information dynamics when travelling along the Ho12 strand for $n = \{1, 2, 3, 4\}$. We observe the existence of quantum information levels that somehow vanish when $n$ increases. This is due to finite number of quantifying levels of information that occur before a space terminates a word. The number of quantum levels increases with $n$ while...
the length of each word increases. Besides this interesting effect, we also note a considerable randomness and a uniform behavior along all length of the strand.

The total chromosome information, the number of words $N_w$, and the average word information $I_{av}$ versus $n$ are depicted in Figures 5(a) and 5(b). We verify a maximum of the total chromosome information for $n = 3$. For larger values of $n$ the information decreases slightly due to the effect of dropping out repeated consecutive spaces. Therefore, we can say that large values of $n$ seem to lead to a slightly better estimate of the total information content, while the cases of $n = 1$ or $n = 2$ lead to an inferior measurement process. We also observe that the number of words decreases with $n$ but its average information varies in the opposite way. Therefore, it is relevant to plot one variable against the other, with $n$ as parameter (Figure 5(c)). A PL trendline approximation demonstrates that the two quantities are inversely proportional. In fact, we get numerically $I_{av} = aN_w^b$ with $a = 2.07 \times 10^5$, $b = 2.07 \times 10^{-1}$.

Figure 4: Word information versus length for the Ho12 and $n = \{1, 2, 3, 4\}$.
For other values of $n$ the resulting histograms reveal identical characteristics, namely, two characters with a very large relative frequency (depicted at the left part of the histograms of Figure 3). Furthermore, experiments with other chromosomes lead to similar results. The two characters are simply a succession of symbols $A$ or $T$ and the corresponding $n$-tuples (i.e., $A \cdots A$ and $T \cdots T$) are adopted in the sequel as “spaces.”

Figure 6 shows the total information, that is, the information resulting from summing the information of all the chromosomes of each species $versus$ the corresponding number of chromosomes, for character encoding with $n = 8$. We observe a weak correlation between both variables.

Figure 7 shows the length of each chromosome $L_i^{crom}$ versus its information content $I_i^{crom,n}$, $i = 1, \ldots, 463$, estimated by the proposed method with $n = 8$. In this case we observe $b = -1.02$. For the rest of the chromosomes it was observed a similar type of behavior, but with different numerical values for the parameters.
a strong correlation between both variables, meaning that the implementation of the DNA code has a large similarity between all species. In fact, we can calculate a PL trendline over the 463 chromosomes yielding the relationship $I_{av}^{i,crom,8} = 0.79(L_{i,crom}^{8})^{1.03}$.

Bearing these ideas in mind it was decided to explore the PL behavior, that is, the relation $I_{av} = aN_{w}^{b}$, $a > 0$, $b < 0$, of the average word information $I_{av}$ versus the number of words $N_{w}$ (with $n$ as parameter) per chromosome. The extensive evaluation of the 463 chromosomes for $n = \{1, \ldots, 8\}$ leads to the locus $(a, b)$ of the PL trendline depicted in Figure 8. The point for chromosome DyYh is not included to allow a better visualization of the remaining set of points. Moreover, the individual chromosome labels are not included to make the plot more readable.

We verify that the map produces clear patterns, not only by grouping the chromosomes of each species but also by the relative positioning of the different species.
Nevertheless, the large number of points complicates the visualization. Therefore, it was decided to represent each species by a single point having for coordinates the geometric and arithmetic averages of parameters $a$ and $b$, respectively. Figure 9 depicts the resulting locus where is now easier to analyze the previously mentioned relations. The microchromosomes Ga32 and Tg16, which have a very small base pair counting, were not included in the calculations because they significantly disturb the results.

We verify the emergence of clusters that are in reasonable accordance with phylogenetics, going from the less “complex” species at left up to the most “complex” species at the right. The cluster of mammals is at the right and includes the subcluster of primates $\{\text{Ho, Ch, Or}\}$, with Ch closer to Hu than Or. In the rest of mammals it is interesting to see Po close to the primates and the position of the marsupial Op relatively distant from the placental mammals. In what concerns the rest of the points we notice Cb close to Ce and, in a middle position, the clusters of birds $\{\text{Ga, Tg}\}$, fishes $\{\text{Tn, St, Me, Zf}\}$, and insects $\{\text{Dy, Ds, Am, Ag}\}$.

In conclusion, the proposed information measure leads to an assertive and quantitative classification of chromosomes and species. Furthermore, it can be further explored for decoding in more detail other aspects of the DNA code in association with the FC tools.

4. Conclusions

Chromosomes have a code based on a four-symbol alphabet, and it can be analyzed with methods usually adopted in information processing. The information structure has resemblances to those occurring in systems characterized by fractional dynamics. Nevertheless, schemes based on assigning numerical values to the DNA symbols may deform the information, and alternative methods that avoid such problem need to be implemented. In this paper it was proposed a scheme based on the Shannon information theory. Bearing these ideas in
mind, the chromosomes were processed in the perspective of a PL relationship between the average information and the total number of words, for distinct values of character encoding. For condensing the information an averaging of the PL parameters was also adopted. The resulting locus revealed the emergence of clearly interpretable patterns in accordance with current knowledge in phylogenetics. The proposed methodology opens new directions of research for DNA information processing and supports the recent discoveries that fractional phenomena are present in this biological structure.

**Acknowledgments**


---

**Figure 9:** Locus of geometric average of $a$ versus the arithmetic average of $b$ for the twenty-three species.
Abstract and Applied Analysis


References


Submit your manuscripts at
http://www.hindawi.com