Use of Recurrence Plots to Find Mutations in Deoxyribonucleic Acid Sequences

D. E. Rivas-Cisneros
Universidad Autónoma de Nuevo León
Pedro de Alba S/N, Niños Héroes, Ciudad Universitaria
San Nicolás de los Garza, N. L, México

This paper examines the use of recurrence plots to find changes in the deoxyribonucleic acid (DNA) sequence. The DNA sequence is entered into the recurrence plot algorithm in codons and not consecutively, in order to make it easier to find the nucleotide change. The results show that the codon arrangement makes it easy to identify two types of mutations, insertion and deletion, by means of the recurrence plot. It is also shown that the recurrence plot of a short codon sequence has a homogeneous structure. Similarly, a comparison is made with the digital signal processing methodology. Some limitations are the number of codons that can be entered into the algorithm because computationally it becomes very slow and matrix issues, which are mentioned in the discussion section.

Keywords: recurrence plot; DNA sequence; genetic mutations; digital signal processing

1. Introduction

Deoxyribonucleic acid (DNA) consists of four different nucleotides, which are adenine (A), guanine (G), thymine (T) and cytosine (C). These four bases, in different combination sequences, are the ones to define amino acid sequences of proteins. The synthesis of these specific proteins is the basis for all biological functions. A mutation can be defined as an alteration of the DNA sequence. A mutation may involve the substitution, deletion or insertion of nucleotides in the DNA sequence. When a mutation alters a gene, it can modify or even eliminate the usual function encoded in the protein and cause the appearance of an altered phenotype.

Sequence alignment method is an important term in DNA sequence analysis. This method sorts DNA, ribonucleic acid (RNA) or protein sequences to identify regions of similarity. There are two types of methods for finding similarities in DNA sequence analysis, which are [1, 2] alignment-based similarity analysis and alignment-free similarity analysis. The difference between these two methods is that the second has a lower computational complexity.

https://doi.org/10.25088/ComplexSystems.32.1.89

Using DNA-based tests, scientists and doctors can directly examine a patient's DNA for mutations associated with disease. There are various DNA tests to detect mutations, such as gene tests based on restriction enzyme analysis, PCR-based methods, genetic tests using DNA chips and genomic screening.

Recurrence plots (RPs) are a nonlinear analysis tool that shows the existence of recurrence or discontinuous patterns in time series [3]. The objective of this tool is to find the repetition of a pattern within the state space of a system or nonstationary data. It is also useful for finding differences and similarities between two or more systems.

The use of graphs is not new in DNA sequence analysis. Dot plots, which are a type of recurrence plot (RP), are used for DNA sequence analysis. According to the literature, the pioneers in the use of dot plots were Gibbs and Mcintyre [4], who used this type of method to analyze different DNA sequences of living beings. From here, several tools have been developed that use the dot plot method for the analysis of DNA sequences, such as [5–7]. Recently, RPs have been used for the analysis of protein sequences and DNA sequences [3, 8], as well as in the analysis of EEG and ECG signals [9–11]. Similarly, there are different methods that have been used for DNA sequence analysis, such as digital signal processing or long-range correlation [12–16]. As well, there are other methods of great relevance for DNA sequence analysis [17–19] that are based on prediction and entropy methods. The RP is a method that is also based on entropy.

A comparison is made with digital signal processing methodology. The digital signal processing method detects that a change in DNA sequence has occurred, but does not visualize the position where the mutation occurred. The RP method, which is proposed in this paper, detects and visualizes the position where changes in the DNA sequence occur. Another advantage of the RP method is that it does not require converting the DNA sequence into a numerical sequence.

The present paper aims to show that RPs can be a useful tool for finding DNA sequence changes. This paper is limited only to the simulation and identification of DNA mutations by RPs. It is proposed that the DNA sequence be entered in the form of codons into the algorithm [11], which is freely available on the web, to make it easier and less computationally time consuming to identify changes in the sequence.

The paper begins with a brief introduction about RPs and genetic concepts used in the present paper. A brief review of DNA sequence analysis by digital signal processing follows. Subsequently, the results, future work and conclusions are given.

2. Recurrence Plot

The RP method is an effective tool for the analysis of nonlinear signal, stationary or nonstationary data. This method is a two-dimensional representation created from a one-dimensional time series. The RP is created by means of an $N \times N$ matrix, where N is the number of data samples; this matrix is called R. The R matrix consists of zeros or ones, according to [4, 9]:

$$R_{ij} = \begin{cases} 1 : x_i \approx x_j \\ \qquad\qquad i, j = 1, \ldots, N \\ 0 : x_i \neq x_j \end{cases} \tag{1}$$

x_i and x_j are two samples of the signal at time points i and j. The concept of recurrence can be expressed as $\{x : \|x_i - x_j\| \leq \varepsilon\}$. If the distance between x_i and x_j is less than or equal to the threshold ε, it is said that there is a recurrence. The threshold value is determined empirically. According to equation (1), the RP is constructed by replacing 1 with dots and 0 with blanks. The RP has a black main diagonal line, called the line of identity (LOI). There are different typologies of RPs; here is a list of some of them [4]:

- Homogeneity: The process is stationary.
- White banding disruptions: The process is nonstationary.
- Periodic patterns: The process has characteristic cyclicities; the distance between periodic patterns is the period.
- Single isolated points: The states of the process are rare, do not persist for any time or have heavy fluctuations.
- Diagonal lines parallel to the LOI: The process is deterministic in the periodic sense (long diagonals) or chaotic sense (short diagonals).

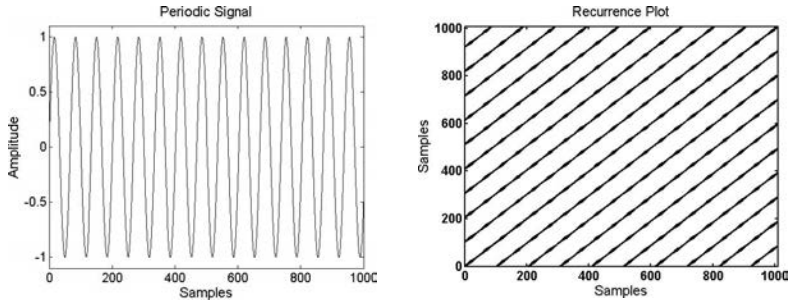
Figures 1 and 2 show a pair of time series and a string with their respective RPs. Figure 1(a) shows a periodic time series and Figure 1(b) its recurrence plot. The RP of Figure 1(b) shows long diagonal lines parallel to the LOI; this indicates that the process is periodic and deterministic. Figure 2(a) shows a time series of a white noise and Figure 2(b) its RP. The RP of Figure 2(b) shows a homogeneity structure indicating a stationary process.

Figure 3 represents the RP of the following tongue twister:

I WISH TO WISH THE WISH YOU WISH TO WISH, BUT IF YOU WISH THE WISH THE WITCH WISHES, I WOULD NOT WISH THE WISH YOU WISH TO WISH.

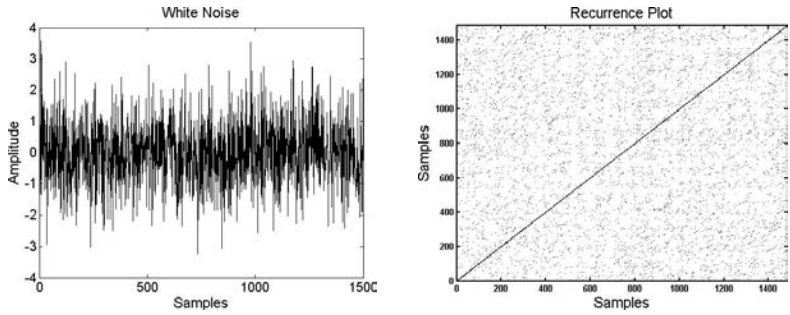
Which is input to the algorithm as a string. The dots inside the circle represent the word that recurs most often in the tongue twister; the isolated dots are the words that do not have much recurrence.

https://doi.org/10.25088/ComplexSystems.32.1.89



(a) (b)

Figure 1. RP of a periodic signal. (a) Periodic signal. (b) RP of Figure 2(a).



(a) (b)

Figure 2. RP of a white noise signal. (a) White noise signal. (b) RP of Figure 3(a).

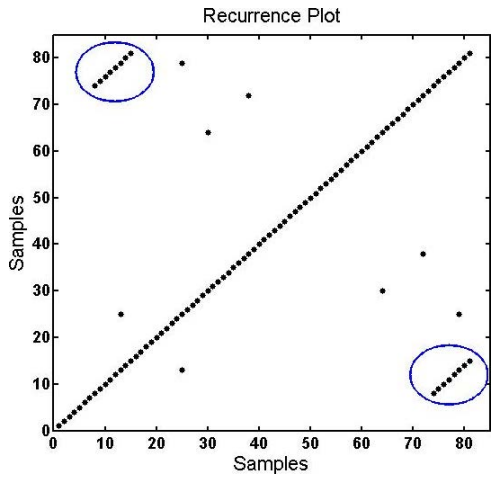


Figure 3. RP of the tongue twister.

2.1 Deoxyribonucleic Acid Sequence, Codons and Mutations

As mentioned earlier, the DNA sequence is made up of the following bases: A, G, T, C. A codon is a triplet of nucleotides that specifies a particular amino acid or a start or end signal in the genetic code [20]. The DNA sequence is usually written or expressed consecutively, and to decode or read the sequence, it is separated into codons to know the types of amino acids.

A mutation is an alteration of the DNA sequence. These alterations may involve the substitution, deletion or insertion of one or more nucleotides in the DNA sequence. Such mutations can occur spontaneously as a result of biological and chemical processes in the organism. They can also be due to external factors, such as chemicals and radiation.

Table 1 shows a small DNA sequence obtained from the website [21], expressed in codons. The codons sequence in Table 1 was used to obtain the results shown in Section 3.

CCT CCT CCC CCG CCT CGG CTC AGG AGT TGT TTT CCT CCC CCG TGG TCG
CCG GGC TGG CAA GGG GTC TGG GGC AAG GGC CAG GGA AGG ATG CGG CTG
CAA AGG TAA CGG TGT GCG CGG TGC CTC TCC CGC GGT CGT GAA AGT TTG
CTC TGG CGT GCG AGG CGT GCC CGC CGG GTG CTC CGC CGC GCC GCG CTG
CCC TGG CCC GAG GCT CCC CGG GGG CGG GCA GGG GCC GCG CCG GGC CAG
GGC AGG GCT GGG CAG GGG CCT CCC GGG CGG CCT TTA CTG CCT TGT CCC
CAT CAA AGG CCG CAG CTG CTC GAG TGG CGG CTG TAG CTT TGA AAC TGC
CAT AAA GAG CGA TGG AGG GCG GGG AGG CGG GGG ATG TTC AGA GCA GTG
CCC AGG AGG GTG GCC GGG AGA GGA GGG GCC CGA GGG GGC TGC TCG GGG
AGC GCG GCA GCA GCA GGG GCG TGA AGG GCA GGG CCC CGG GTC ACA GTC
CCA GGG GCC GAG TGC CCG CCC CCG TCC CCT CCG GTC CAG ATC CAG CCG
CGC CCC CGT CCC CTC CGG TCC AGA TCC AGC CGC CGT GCG

Table 1. DNA sequence obtained from [21], expressed in codons.

3. Review of Digital Signal Processing Methods for Deoxyribonucleic Acid Sequence Analysis

There are different digital signal processing (DSP) methods to detect periodicities or repetitive features of a DNA sequence, such as Fourier transform, Ramanujan–Fourier transform and periodic power spectrum (PPS) [12–14]. In the aforementioned methodologies, the alphabetic DNA sequence is converted into a numerical sequence and then undergoes DSP for power spectrum analysis in the sequence.

The PPS method can detect nucleotide changes in the DNA sequence. In [12] the comparison between a correct DNA sequence

and the same DNA sequence with a mutation is shown. Here is the definition of the PPS algorithm [12]:

$$\mathrm{PPS}(p) = \sum_{\alpha\in\{\mathrm{A,T,G,C}\}} \left(\sum_{q=0}^{p-1} f_{\alpha q}^2 + \sum_{k=0, j=0, k>j}^{p-1} S_p(k, j) f_{\alpha j} f_{\alpha k} \right) \tag{2}$$

where p is the periodicity of the DNA sequence and PPS, f_α is the congruent derivative vector of the sequence $\alpha \in \{\mathrm{A, T, G, C}\}$ and S_p is the spectrum transform matrix of size p.

Next, using the PPS method, the comparison of the DNA sequence (with periodicity 3) mentioned in the previous section against the same sequence but inducing a mutation is shown. Figure 4 shows the comparison between the correct DNA sequence with the same sequence but with a one-nucleotide substitution; a small change in the peaks is visualized. The peak value of the correct sequence is 1247 and of the sequence with nucleotide substitution is 1286. Figure 5 shows the comparison between the correct DNA sequence with the same sequence but with the deletion of one nucleotide; a considerable change in the peaks is visualized. The peak value of the sequence with the deletion is 896. Figure 6 shows the comparison between the correct DNA sequence with the same sequence but with the insertion of one nucleotide; a considerable change in the peaks is visualized. The peak value of the sequence with the insertion is 890. As can be seen in the graphs in Figures 4–6, the PPS method detects changes in DNA sequence with respect to the original.

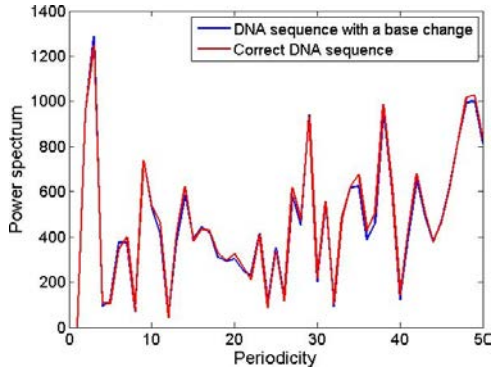


Figure 4. PPS of the correct DNA sequence with periodicity 3 against the same sequence but with a one-nucleotide substitution.

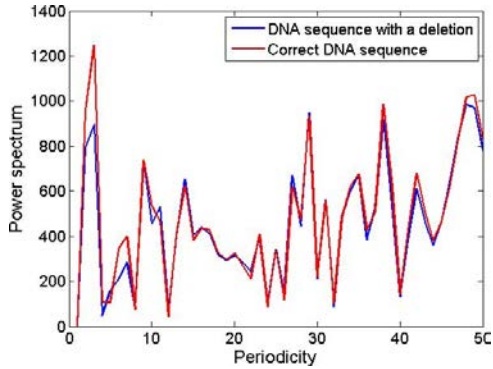


Figure 5. PPS of the correct DNA sequence with periodicity 3 against the same sequence but with a one-nucleotide deletion.

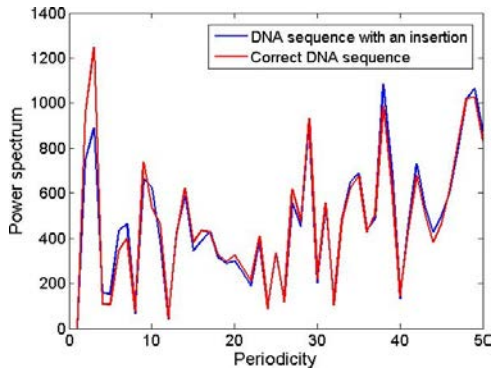


Figure 6. PPS of the correct DNA sequence with periodicity 3 against the same sequence but with a one-nucleotide insertion.

4. Results and Discussion

First, the RP of the codon sequence in Table 1 is shown. Subsequently, the RPs are presented when simulating a substitution, deletion and insertion into the codon sequence.

4.1 Recurrence Plot of Codon Sequence

Figure 7 represents the RP of the codon sequence shown in Table 1. According to the RP in Figure 7, the typology is homogeneity. It is well known that a DNA sequence has stationary process characteristics [3, 22].

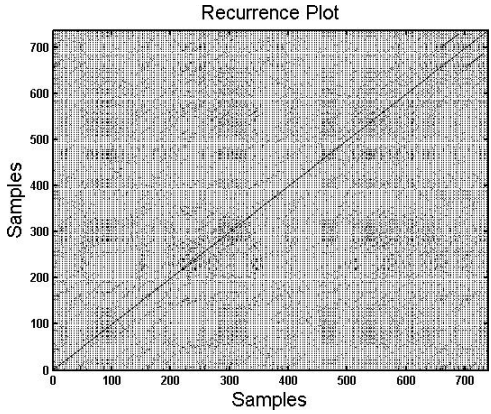


Figure 7. RP of the DNA sequence arranged in codons.

4.2 Substitution

To simulate the substitution, a single nucleotide change is made at codon 50 of the sequence in Table 1. The substitution generates a change of one nucleotide for another, so the number of nucleotides of the sequence does not change. To know the position where the mutation occurred, a matrix subtraction of the original sequence against the mutated sequence is made.

The lines in Figure 8 depict the position where the nucleotide substitution occurred. The position number is given by the position of the codon multiplied by three (since each codon has three nucleotides) plus the sum of spaces before the codon where the mutation occurred.

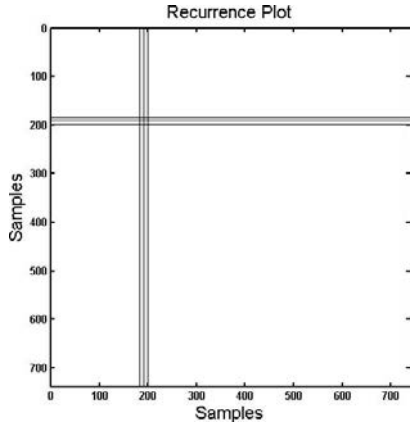


Figure 8. RP of the substitution.

4.3 Deletion and Insertion

To simulate the deletion, one nucleotide is deleted at codon 41 of the sequence in Table 1. When a nucleotide is eliminated from the codon, it is left with two nucleotides, so it is directly reflected in the RP, as shown in Figure 9, where the grayish-white stripe shows the position where the deletion occurred.

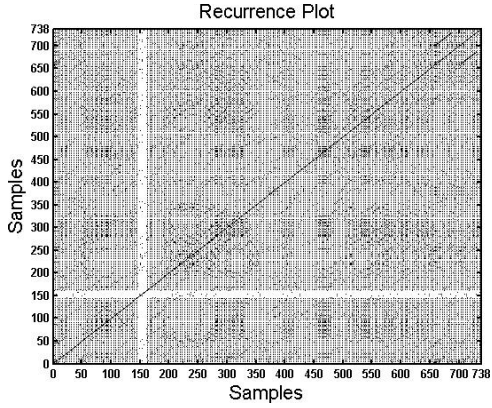


Figure 9. RP of the deletion.

To simulate the insertion, a single nucleotide is added at codon 70 of the sequence in Table 1. When a nucleotide is added to the codon, it is left with four nucleotides, so it is directly reflected in the RP, as shown in Figure 10, where the grayish-white stripe shows the position where the insertion occurred. In the case of deletion and insertion, the position number is given in the same way as explained for substitution.

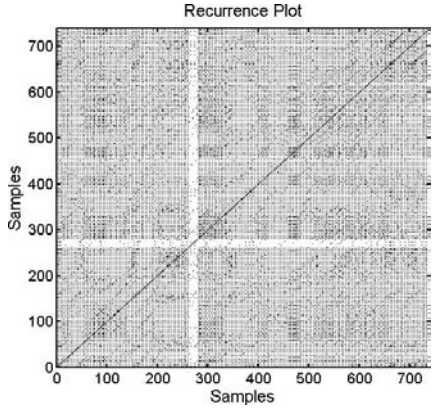


Figure 10. RP of the insertion.

https://doi.org/10.25088/ComplexSystems.32.1.89

4.4 Discussion

Figure 11 depicts the RP of a deletion and insertion at the codons 41 and 70, respectively. The deletion and insertion process can be obtained in a single RP, since the codon triplet is broken.

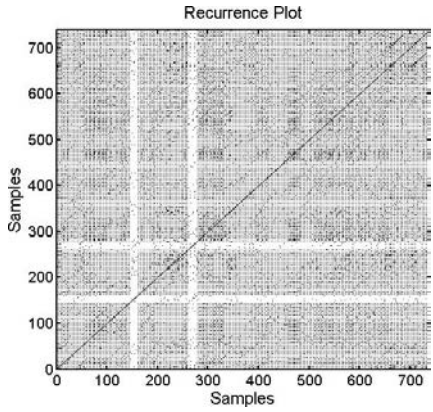


Figure 11. RP of deletion and insertion at the same time.

The substitution process occurs when there is a change of nucleotide for another, so the triplet of the codon is not altered. As mentioned earlier, a matrix subtraction is done to know where the substitution occurs by means of the RP. For matrix considerations, it is not possible to obtain the three types of mutation in a single RP. As future work, it is proposed to improve the RP algorithm to solve the matrix problem.

5. Conclusions and Future Work

According to the results, the methodology of recurrence plots (RPs) can be an additional tool to identify genetic mutations. In the present paper, the types of mutations are simulated within the deoxyribonucleic acid (DNA) sequence, which is arranged in codons for easier identification. A comparative analysis between digital signal processing and recurrence plot (RP) methodologies is performed. The result is favorable to the RP, since this method is able to detect and visualize where the mutations occur. Hence, an area of opportunity is to have a base DNA sequence within the algorithm and then introduce the sequence to be analyzed, and by means of the RP visualize where a substitution, deletion and insertion of nucleotides occur, according to the original sequence. This can be possible by fixing the matrix

problem, which we have done in this specific case, as cited in the discussion section.

References

[1] A. Zielezinski, S. Vinga, J. Almeida and W. M. Karlowski, "Alignment-Free Sequence Comparison: Benefits, Applications, and Tools," *Genome Biology*, **18**(1), 2017 186. doi:10.1186/s13059-017-1319-7.

[2] P. Waz and D. Bielinska-Waz, "Non-standard Similarity/Dissimilarity Analysis of DNA Sequences," *Genomics*, **104**(6, Part B), 2014 pp. 464–471. doi:10.1016/j.ygeno.2014.08.010.

[3] C. L. Webber, Jr. and N. Marwan, eds., *Recurrence Quantification Analysis: Theory and Best Practices*, Cham: Springer International Publishing, 2015. doi:10.1007/978-3-319-07155-8.

[4] A. J. Gibbs and G. A. Mcintyre, "The Diagram, a Method for Comparing Sequences: Its Use with Amino Acid and Nucleotide Sequences," *European Journal of Biochemistry*, **16**(1), 1970 pp. 1–11. doi:10.1111/j.1432-1033.1970.tb01046.x.

[5] K. M. Seibt, T. Schmidt and T. Heitkam, "FlexiDot: Highly Customizable, Ambiguity-Aware Dotplots for Visual Sequence Analyses," *Bioinformatics*, **34**(20), 2018 pp. 3575–3577. doi:10.1093/bioinformatics/bty395.

[6] J. Krumsiek, R. Arnold and T. Rattei, "Gepard: A Rapid and Sensitive Tool for Creating Dotplots on Genome Scale," *Bioinformatics*, **23**(8), 2007 pp. 1026–1028. doi:10.1093/bioinformatics/btm039.

[7] W. Chen, B. Liao, W. Zhu, H. Liu and Q. Zeng, "An Ant Colony Pairwise Alignment Based on the Dot Plots," *Journal of Computational Chemistry*, **30**(1), 2009 pp. 93–97. doi:10.1002/jcc.21039.

[8] I. Abdennaji, M. Zaied and J.-M. Girault, "Prediction of Protein Structural Class Based on Symmetrical Recurrence Quantification Analysis," *Computational Biology and Chemistry*, **92**, 2021 107450. doi:10.1016/j.compbiolchem.2021.107450.

[9] G. Baghdadi, M. Amiri, E. Falotico and C. Laschi, "Recurrence Quantification Analysis of EEG Signals for Tactile Roughness Discrimination," *International Journal of Machine Learning and Cybernetics*, **12**(4) 2021 pp. 1115–1136. doi:10.1007/s13042-020-01224-1.

[10] V. L. Hilarov, "Estimates of Heart Rate Variability by the Method of Recurrence Plots," *Technical Physics Letters*, **44**(12), 2018 pp. 1084–1086. doi:10.1134/S1063785018120246.

[11] Y. Chen and H. Yang, "Multiscale Recurrence Analysis of Long-Term Nonlinear and Nonstationary Time Series," *Chaos, Solitons & Fractals*, **45**(7), 2012 pp. 978–987. doi:10.1016/j.chaos.2012.03.013.

[12] C. Yin and J. Wang, "Periodic Power Spectrum with Applications in Detection of Latent Periodicities in DNA Sequences," *Journal of Mathematical Biology*, **73**(5), 2016 pp. 1053–1079. doi:10.1007/s00285-016-0982-8.

[13] S. Jeya Bharathi and P. Balamanikandan, "Spectral Analysis of Codons in the DNA Sequence of Fragile X Syndrome," *Journal of Medical Systems*, **43**(8), 2019 261. doi:10.1007/s10916-019-1408-0.

[14] N. Chakravarthy, A. Spanias, L. D. Iasemidis and K. Tsakalis, "Autoregressive Modeling and Feature Analysis of DNA Sequences," *EURASIP Journal on Applied Signal Processing*, **2004**(1), 2004 952689. doi:10.1155/S111086570430925X.

[15] H. Kargupta, "A Striking Property of Genetic Code-like Transformations," *Complex Systems*, **13**(1), 2001 pp. 1–32. complex-systems.com/pdf/13-1-2.pdf.

[16] R. Mansilla and G. Cocho, "Multiscaling in Expansion-Modification Systems: An Explanation for Long Range Correlation in DNA," *Complex Systems*, **12**(2), 2000 pp. 207–240. complex-systems.com/pdf/12-2-4.pdf.

[17] D. Pratas, M. Hosseini, J. M. Silva and A. J. Pinho, "A Reference-Free Lossless Compression Algorithm for DNA Sequences Using a Competitive Prediction of Two Classes of Weighted Models," *Entropy*, **21**(11), 2019 1074. doi:10.3390/e21111074.

[18] H. Zenil and P. Minary, "Training-Free Measures Based on Algorithmic Probability Identify High Nucleosome Occupancy in DNA Sequences," *Nucleic Acids Research*, **47**(20), 2019 e129. doi:10.1093/nar/gkz750.

[19] E. Rivals, M. Dauchet, J. P. Delahaye and O. Delgrange, "Compression and Genetic Sequence Analysis," *Biochimie*, **78**(5), 1996 pp. 315–322. doi:10.1016/0300-9084(96)84763-8.

[20] A. Som, S. Chattopadhyay, J. Chakrabarti and D. Bandyopadhyay, "Codon Distributions in DNA," *Physical Review E*, **63**(5), 2001 051908. doi:10.1103/PhysRevE.63.051908.

[21] "TCF20." National Library of Medicine. (Jan 24, 2023) www.ncbi.nlm.nih.gov/gene/6942.

[22] X. Gu and W.-H. Li, "Estimation of Evolutionary Distances under Stationary and Nonstationary Models of Nucleotide Substitution," *Proceedings of the National Academy of Sciences*, **95**(11), 1998 pp. 5899–5905. doi:10.1073/pnas.95.11.5899.