Applications of Fractal Geometry in DNA and Genome Studies

Dawid Baraskiewicz

$5^{th}$of June 2023

Abstract

In this paper I will cover some methods involving fractal geometry that are used in the analysis of DNA sequences. This paper will have a particular focus on the identification and study of VNTRs (Variable Number of Tandem Repeats), which are repeated sequences of nucleotides within a DNA sequence. The ability to identify VNTRs and their location within a DNA sequence leads to many applications, which are covered in the conclusion. This paper will also be concerned with the identification of other properties such as long-range power law correlations, patches and coding/non-coding regions. The Indicator Matrix method, the DNA walk and Detrended Fluctuation Analysis are explored, with the Indicator Matrix Method chosen as the most effective.

1 Introduction

A DNA sequence is made up of genes and each one of these genes contains the necessary information for the creation of different proteins. However, between these genes lie vast regions of non-coding DNA known as intergenic sequences. It has also been discovered that genes themselves include regions not used for coding. These regions are removed during the formation of mRNA [1] and are called introns [2]. The mRNA is then used for the assembling of proteins. The coding sequences in genes which are used during the formation of mRNA are called exons [3].

Genes are made up of building blocks called nucleotides of which there are four different varieties, two of which are referred to as purines and the other two are referred to as pyrimidines. These nucleotides can be arranged in different ways to produce three-letter sequences known as codons [4]. Codons contain the information that specifies which amino acid is needed at which position in a protein.

The two purine nucleotides are adenine and thymine, and the two pyrimidine nucleotides are cytosine and guanine. Adenine and thymine have complimentary shapes and in the case of DNA sequences they come in pairs that are held together by two hydrogen bonds. Thymine is a very specialised molecule and is essentially only found in DNA, and it is the only nucleotide to be excluded from RNA where it is replaced by uracil [5]. However, its purine counterpart adenine is slightly more diverse and other than being a component of DNA it also plays a role in protein synthesis and upon undergoing chemical reactions it has some useful derivatives including adenosine triphosphate, otherwise known as ATP, a common carrier of energy in living organisms. As expected, the remaining two nucleotides, guanine and cytosine, also pair up in DNA sequences and they are paired by three hydrogen bonds. Much like the previous pair the purine is the more diverse of the two molecules. Cytosine's only uses are in DNA and RNA, but guanine also appears in several products in the form of crystalline guanine where it adds an iridescent effect to the product.

The number of individual nucleotides found in a DNA sequence is staggeringly large; the longest genome discovered to date is that of the marbled lungfish, with a total of 130 billion nucleotides found in every strand of their DNA [6] which dwarfs the measly 3 billion found in the human genome. However, it is worth noting that a larger genome does not necessarily correspond to a more complex organism; it could also correspond to a larger portion of intergenic sequences in the DNA.

Due to the size of DNA sequences, it is to be expected that there will be repetition amongst the nucleotides. This repetition is known as a Variable Number of Tandem Repeats (VNTR) [7]. VNTRs can be split into two categories: minisatellites and microsatellites. Minisatellites are repeated sequences of nucleotides that consist of 10-100 nucleotides. Microsatellites are repeated sequences of nucleotides that consist of fewer than 10 nucleotides. A large portion of VNTRs is found in the non-coding regions of DNA which does not seem particularly useful at first but it could grant scientists the ability to more easily distinguish between coding and non-coding regions of DNA and hence isolate the coding regions for further study. Further applications, of which there are many will be outlined in the conclusion. DNA sequences also have regions known as patches which are areas that have a high density of a single nucleotide, these patches can create illegitimate long-range correlations that are induced by the patchy nature of the DNA sequence rather than any underlying unifications between the nucleotides.

It is very important for scientists to be able to identify VNTRs with as much ease as possible. This can be achieved by introducing different visual methods for the representation of DNA sequences, the most simple of which is writing down the nucleotides, represented by their initials, in order. For example, A, A, G, T, G, G, C, T etc. However, this method is very simple and not particularly efficient at representing large amounts of nucleotides or at finding VNTRs. Some more complex methods can make it significantly easier to spot VNTRs.

In this paper I will outline the following three methods used for visual representation of DNA sequences. The indicator matrix, the DNA walk and Detrended Fluctuation Analysis. A brief explanation of each of these will conclude the introduction.

The indicator matrix is a binary matrix concerned with individual nucleotides. It is most adept at finding patches, as well as finding VNTRs, although this becomes increasingly more difficult at larger scales.

The DNA walk is a one-dimensional random walk, not concerned with the individual nucleotides but rather in their classification as either a purine or a pyrimidine. Unlike the indicator matrix the DNA walk is less tailored towards studying VNTRs within a DNA sequence, the main application of the DNA walk is to provide a way of quantitatively measuring the correlations between nucleotides over long distances along a DNA sequence [8].

Detrended Fluctuation Analysis concerns itself with neither individual nucleotides nor their classification as a purine or pyrimidine. Detrended Fluctuation analysis was created to detect long-range correlations within a patchy DNA sequence (or any other system that exhibits patchiness), whilst avoiding the spurious detection of apparent long-range correlations caused by the DNA sequence's patchiness [9].

2 Definitions

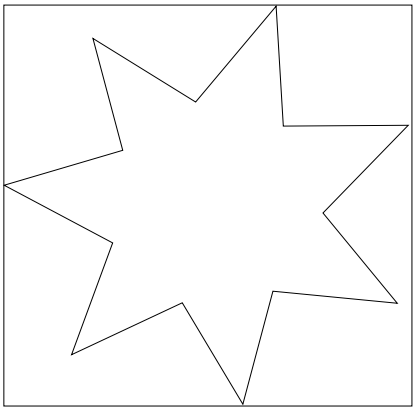
2.1 Fractals

As of yet, the world's leading mathematicians are unable to agree upon a defined definition of a fractal. However, the founding father of fractal geometry, Benoit Mandelbrot, has a definition that he abides by. 'A fractal is a set for which the Hausdorff-Besicovitch dimension strictly exceeds the topological dimension'[10].

2.2 Topological Dimension

The topological dimension, denoted dim(X) is the dimension we come across daily and a person without a background in mathematics will be familiar with it. It is defined by the number of unique coordinates needed to represent a point in space. To represent a point on a line one coordinate is required and hence it has a topological dimension of one. To represent a point in a cube three coordinates are required and hence it has a topological dimension of three. It is worth noting that calculating the topological dimension of a space, is not always trivial, but topological spaces of this nature will not be mentioned in this paper.

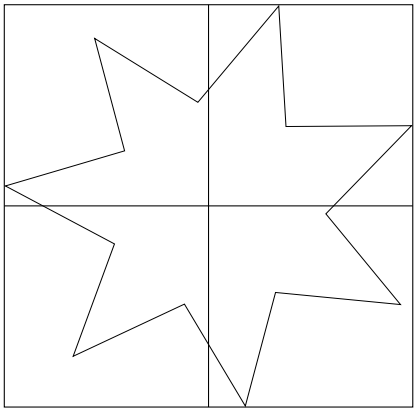
2.3 Fractal Dimension



[h]

Figure 1: An example where $\varepsilon = 1$

The Hausdorff-Besicovitch dimension, otherwise known as the fractal dimension, is a measure of roughness and while it is calculated computationally it is still important to understand how it is calculated. To find an object's fractal dimension we first encompass the object in the smallest possible square such that the objects outermost points are touching the sides of the square. We define the length ε of the sides of the square as equal to 1.



[h]

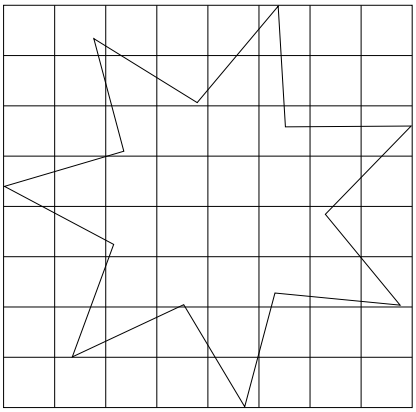
Figure 2: An example where $\varepsilon = \frac{1}{2}$

We define the number of boxes needed to encompass the object as $N(\varepsilon)$ and in the case of Figure 1 shown above we can see that $N(\varepsilon) = 1$ since there is only one box encompassing the object. We then halve the length of each of the sides of the squares so that $\varepsilon = \frac{1}{2}$ and it is simple to see that for this value of epsilon $N(\varepsilon) = 4$ as can be observed in Figure 2.

The halving of ε continues and we want the squares encompassing the object to be as small as possible, so we let $\varepsilon \to 0$. This cannot feasibly be done by hand and so must be done by computer. Once this has been achieved the fractal dimension is defined as,

$$D_f = lim_{\varepsilon \to 0} \frac{logN(\varepsilon)}{log(\frac{1}{\varepsilon})} \tag{1}$$

In this modern age it is not a problem to find an object's fractal dimension computationally, however without the help of a computer it is an incredibly tedious task, and the nature of this dimension is one of the reasons why fractal geometry only started its development in recent years with the help of computers.



[h]

Figure 3: An example where $\varepsilon = \frac{1}{8}$. Unlike the other examples shown in Figures 1 and 2, one can see that not all of the boxes are being used to cover the shape.

3 Long Range Power Law Correlations in DNA

A power law can be defined as the following,

$$f(x) = x^{-n} \tag{2}$$

Fractals, and systems with fractal-like behaviours, have two distinct properties that show that they have a relation with power laws. These properties are self-organisation and self-similarity. Self-organisation is a process in which order spontaneously arises within an initially disordered system. A self-similar object is one that is similar or identical to a smaller part of itself. Data that has come from a system that is both self-similar and self-organised cannot be modelled by any conventional distributions. This is because any order exhibited from such a system is shown by correlations between different orders of magnitude, this rules out the possibility of correlations being exhibited by a conventional distribution on a set scale. Correlations of this type are best described by power laws because of their ability to move through different orders of magnitude [11]. We can also show that the fractal dimension obeys a power law by the following rearrangement

of equation (1),

$$D_f = \frac{logN(\varepsilon)}{log(\frac{1}{\varepsilon})} \tag{3}$$

$$logN(\varepsilon) = D_f \times log(\frac{1}{\varepsilon}) \tag{4}$$

$$logN(\varepsilon) = log(\varepsilon^{-D_f}) \tag{5}$$

$$N(\varepsilon) = \varepsilon^{-D_f} \tag{6}$$

There have been several studies published about the appearances of long-range power law correlations in DNA. Some have come to the conclusion that DNA sequences do display long-range power law correlations [12] whereas others have concluded the opposite [13]. After further study the following conclusion was reached. Non-coding regions of DNA sequences display long-range power law correlations whereas coding regions of DNA sequences do not display long-range power law correlations. The reason for this is unclear at present.

4 Indicator Matrix

We start by defining the following,

$$A = Adenine\ G = Guanine\ C = Cytosine\ T = Thymine \tag{7}$$

$$\zeta = (A, T, C, G) \tag{8}$$

From this we can define the sequence of nucleotides in DNA as the finite sequence,

$$S = \mathbf{N} \times \zeta \tag{9}$$

Such that individual members of S can be defined as,

$$S = x_h, x_h = (x_1, x_2, \cdots, x_n), n < \infty, x\ \epsilon\ \zeta \tag{10}$$

Where h is the position of x in the sequence and on the matrix.

The indicator function [14] is the map,

$$u = S \times S \rightarrow [0, 1] \tag{11}$$

Such that,

$$u(x_h, x_k) = 1\ if\ x_h = x_k\ (x_h, x_k \epsilon S) \tag{12}$$

$$u(x_h, x_k) = 0\ if\ x_h \neq x_k\ (x_h, x_k \epsilon S) \tag{13}$$

From the aforementioned equations we can see that a DNA sequence of length n can be represented as an $n \times n$ symmetrical matrix with binary values $(0, 1)$

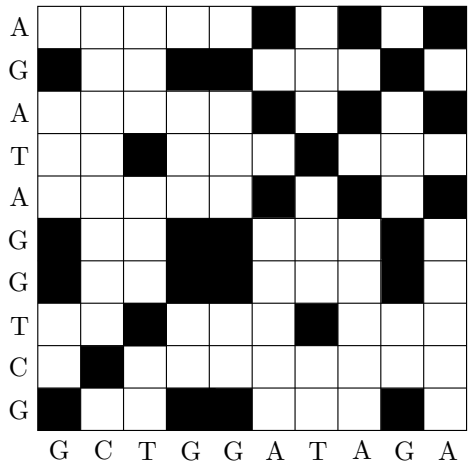


Figure 4: An example of an Indicator Matrix

A method used to further enhance the visual representation of the indicator matrix, is as follows. When $u(x_h, x_k) = 1$ replace the 1 with a black square when $u(x_h, x_k) = 0$ replace the 0 with a white square.

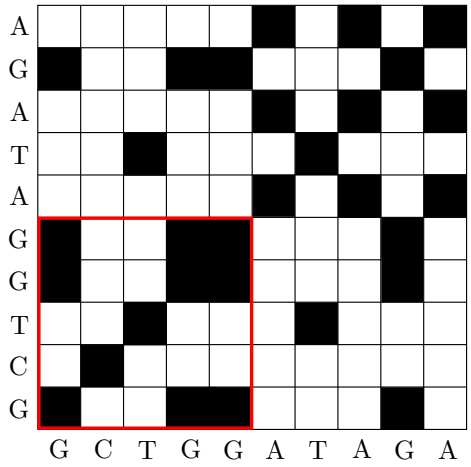


Figure 5: Highlighted in red is a 5×5 minor of a 10×10 matrix, where $p(n) = 11$.

This is a much more useful representation of the indicator matrix because the coloured squares are much easier to identify, especially from distance, than ones and zeros, hence this version of the indicator matrix allows for you to study larger portions of a DNA sequence.

DNA sequences can be incredibly long and contain anywhere from 10^6 to 10^9 individual nucleotides, so the patterns made from the indicator matrix are incredibly complex. However using an adaptation of the equation (1),

$$D_f = \frac{1}{N} \sum_{n=2}^{N} \frac{log\ p(n)}{log\ n} \tag{14}$$

where $p(n)$ is the average number of ones or black squares found in an $n \times n$ minor of the $N \times N$ matrix $u(x_h, x_k)$. We can computationally calculate the fractal dimension of the indicator matrix. An example is given in Figure 4.

5 Analysing Data from the Indicator Matrix

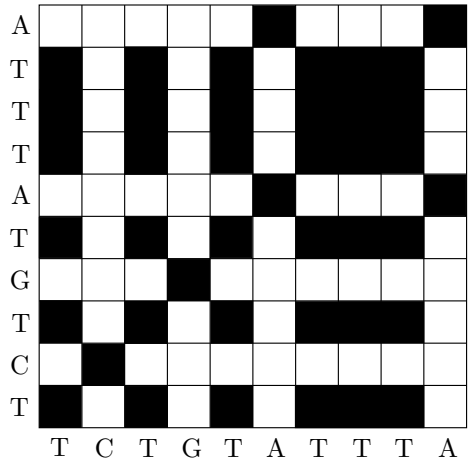


Figure 6: The indicator matrix of a portion of a DNA sequence with an excess of one nucleotide.

This section will outline the meanings of some of the visual characteristics of indicator matrices as well as look at what similarities between different indicator matrices might mean.

As can be seen in Figures 3, 4, 5, and 6, no matter what DNA sequence is used there will always be a constant chain of coloured squares that start at the origin and go along the diagonal $y = x$. This is because the same DNA sequence is going along both the x-axis and the y-axis, and so for any nucleotide a distance i along the x-axis there is a corresponding nucleotide a distance i up the y-axis, these

two points will meet at the coordinate (i, i), which lies on the aforementioned diagonal $y = x$. From here on out I will refer to this as the central diagonal.

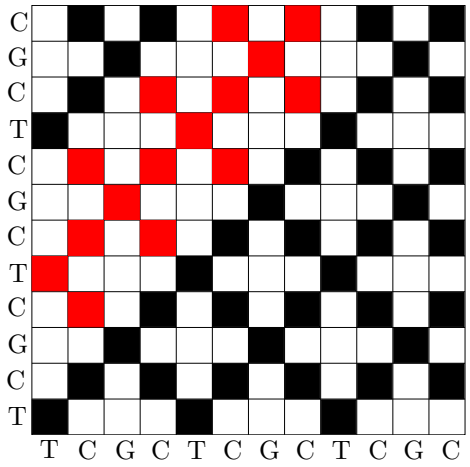


Figure 7: The indicator matrix of a DNA sequence that contains the VNTR T, C, G, C. One of the diagonal patterns produced by the VNTR is highlighted in red. It is worth noting there is an overlap in the diagonal patterns produced.

If we look at Figure 5 we can see that a large clump of coloured squares forms around the central diagonal when patches appear in a DNA sequence. Due to the size of DNA sequences and the patches they contain it would be much more fruitful to study these patches via larger indicator matrices than the one used in Figure 5. It is worth noting that the DNA sequence used in Figure 5 is just an example and would not be considered a patch since there aren't enough nucleotides.

In Figure 6 we can see that there is a microsatellite in the DNA sequence used. This is represented on the matrix via repeated diagonal patterns on the central diagonal and on the diagonals parallel to it.

One study collected data on the DNA of the influenza virus A H1N1 [15] in different regions of the world. Indicator matrices were made from the DNA of each of the viruses. The matrices displayed fractal-like properties including a non-integer fractal dimension and self-similarity.

The indicator matrixes for the DNA of each of the viruses could be put into groups in which every member of a group exhibited the same

visual characteristics. Interestingly when the fractal dimension for each of the matrices was calculated they could all be put into groups where each one shared the same fractal dimension up to $10^{-2}$and these groups were very similar to those formulated from the matrix's visual characteristics.

An assumption could be made here that if two indicator matrixes share similar fractal dimensions then they will also share similar visual characteristics. Using some of the observations made about Figures 3, 4, 5, and 6 we can also conclude that the sharing of visual characteristics between two matrices could include some of the following correlations between the two DNA sequences used for creating the matrices. Similar patches and VNTRs as well as similar placement of the patches and VNTRs within the DNA sequences.

A second, more recent study was carried out that looked at the indicator matrices of the RNA of the SARS-CoV2 coronavirus [16]. A total of 21 samples of RNA were taken and every sample had an indicator matrix generated. The fractal dimension of the indicator matrices was then calculated and each of the samples had a very similar if not identical value of $D_f = 1.63 \pm 0.03$. Upon analysing the 21 matrices some recurring patterns were noticed. This corresponds to the conclusion drawn from the previous study on the DNA of the influenza virus A H1N1 [17].

This conclusion is further supported when the data gathered for the SARS-CoV2 coronavirus is compared to SARS-CoV, SARS-like and MERS-CoV coronaviruses. When compared to the indicator matrices of SARS-CoV and MERS-CoV which have a similar fractal dimensions of $D_f = 1.60$ and $D_f = 1.63$ respectively it is observed that the three matrices share visual characteristics as expected. However upon being compared to the coloured indicator matrices generated from the RNA of SARS-like coronavirus which have a more significantly different fractal dimension of $D_f = 1.58 \pm 0.055$ there were no apparent similarities observed in the indicator matrices. Both of these observations support the conclusion that indicator matrices of similar fractal dimensions share visual characteristics.

6 DNA Walk

A DNA walk [18], shown in Figure 7, is defined by the two following rules. It moves up one vertical unit $u(i) = +1$ if there is a pyrimidine at a distance i along the DNA sequence. It moves down one vertical unit $u(i) = -1$ if there is a purine at a distance i along the DNA sequence.

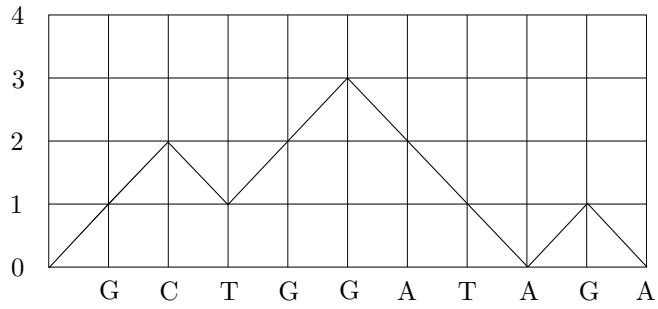


Figure 8: A DNA walk of net displacement 0

The net vertical displacement of a DNA walk is defined as,

$$y(l) = \sum_{i=1}^{l} u(i) \tag{15}$$

The question asked by researchers is whether these DNA walks display short range correlations or long-range correlations as those observed in fractal phenomena. This question can be answered with the help of Detrended Fluctuation Analysis which will be outlined in the next section.

An important characterization of the graphs formed by DNA walks is the square root of the mean fluctuation function about the average displacement, denoted $F(l)$. It is defined as the square root of the difference between the average of the square and the square of the average,

$$F(l) = \sqrt{\overline{[\Delta y(l) - \overline{\Delta y(l)}]^2}} = \sqrt{\overline{[\Delta y(l)]^2} - \overline{\Delta y(l)}^2} \tag{16}$$

although it is more commonly written in the form $F^2(l)$as defined below,

$$F^2(l) = \overline{[\Delta y(l) - \overline{\Delta y(l)}]^2} = \overline{[\Delta y(l)]^2} - \overline{\Delta y(l)}^2 \tag{17}$$

$\Delta y(l)$ is defined as the difference in vertical displacement between some start point l_0 and some end point l,

$$\Delta y(l) = y(l_0 + l) - y(l) \tag{18}$$

The bars in equations (16) and (17) indicate an average over all possible positions l_0 in the DNA sequence.

7 Detrended Fluctuation Analysis

In order to distinguish between sequences with and without long-range power law correlations an appropriate scaling analysis of the correlation properties is required. The method described in this paper is called Detrended Fluctuation Analysis [19]. The method is as follows.

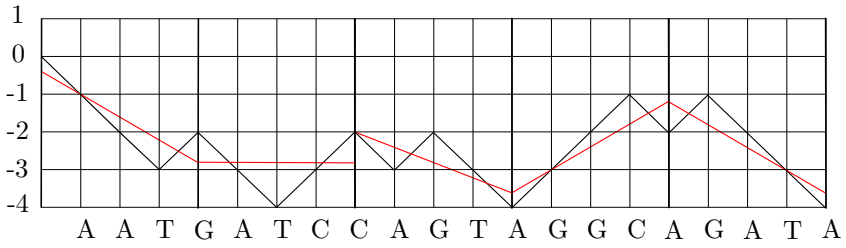


Figure 9: A DNA sequence split into 5 sections of length 4. The line of best fit is highlighted in red in each section.

First you must divide a DNA sequence of length N into $\frac{N}{l}$ non-overlapping boxes where each box contains l nucleotides. Define the 'local trend' of each of the boxes, as the line of best fit for the DNA walk in that box. The net displacement of the local trend of each box will be denoted as $x(n)$. The detrended walk, $y_l(n)$, is defined as the difference between the net displacement of the DNA walk of a box, calculated using equation (15), and the net displacement of the local trend of that same box,

$$y_l(n) = y(n) - x(n) \tag{19}$$

Then calculate the average variance about the detrended walk in each box by using the following equation.

$$F_d^2(l) = \frac{1}{N}\sum_{n=1}^{N} y_l^2(n) = \frac{1}{N}\sum_{n=1}^{N}(y(n) - x(n))^2 \tag{20}$$

If a nucleotide sequence only has short-range correlations, then the detrended DNA walk must have properties consistent with that of a random walk so, $F_d(l) \sim l^{\frac{1}{2}}$, if the sequence does contain long-range correlations, then, $F_d(l) \sim l^{\alpha}$, where α is the critical exponent [20] and $\alpha \neq \frac{1}{2}$. If $\alpha < \frac{1}{2}$ then the long-range power laws show an alteration of different nucleotides whereas when $\alpha > \frac{1}{2}$ the long-range power law correlations show a persistence of a singular nucleotide.

If we plot the function, $F_d(l) = l^\alpha$ on a double-logarithmic graph, $log\ F_d(l) = \alpha log\ l$, then we get a straight line. From this it is easy to compute the gradient of the double-logarithmic graph, find the value of α and establish whether a nucleotide sequence exhibits long-range power law correlations. If the sequence does you can also establish whether those correlations show an alteration of different nucleotides or the persistence of a singular nucleotide.

8 Conclusion

In the conclusion I will go over some of the more promising applications of the methods outlined in this paper.

Out of all of the methods outlined in this paper for identifying VNTRs, the indicator matrix demonstrates the most value. Detrended Fluctuation Analysis is a useful technique for finding long-range power law correlations, however, this is of no use in identifying VNTRs as they occur over a shorter scale. Hence, the comparison is only between the indicator matrix and the DNA walk. The indicator matrix is the victor in this comparison because one is able to fruitfully study much larger portions of a DNA sequence at a time with it when compared to the DNA walk. This is due to it being easier to distinguish the 2-dimensional patterns on the matrix than the 1-dimensional patterns of the DNA walk. This also means that any repeated patterns caused by VNTRs would be easier to identify via the indicator matrix than the DNA walk.

Studying VNTRs can yield some very useful results. Tumour cells have reduced control over their own replication, (this includes the replication of their DNA), because of this, large amounts of VNTRs may be lost or gained during each round of mitosis that occurs. Therefore VNTRs (specifically microsatellites) analysed in primary tissue have been routinely used in cancer diagnosis to assess tumour progression [21].

The study of VNTRs is also beneficial in genetic linkage analysis. Microsatellites can be used as genetic markers when scanning a genome in search of a gene responsible for a certain phenotype. This use of microsatelites has sucesfully led scientists to the discovery of the genes responsible for type 2 diabetes and prostate cancer [22][23].

Once the presence of a power law in a DNA sequence has been established by the use of Detrended Fluctuation Analysis and the DNA walk, and by borrowing methods from the modern theory of critical

phenomena [24] we are able to quantify them with the critical exponent, α. Quantification of this kind of scaling behaviour for apparently unrelated systems allows us to recognise similarities between different systems, leading to underlying unifications that might otherwise have gone unnoticed [25].

It is important to remember the existence of the long-range power law correlations and the underlying unifications they suggest. As regardless of the cause of the correlations their presence in non-coding regions of DNA sequences and their absence in coding regions of DNA sequences must be accounted for by future explanations of global properties in gene organisation and evolution [26].

9 References

[1] National Human Genome Research Institute: Messenger RNA (mRNA). Authors: Shurjo K. Sen

[2] National Human Genome Research Institute: Intron. Authors: Paul P. Liu

[3] National Human Genome Research Institute: Exon.

[4] National Human Genome Research Institute: Codon.

[5] National Human Genome Research Institute: Uracil. Authors: Lawrence Brody

[6] CABI Compendium: Protopterus aethiopicus aethiopicus (marbled lungfish). Authors: Tsungai Zengeya

[7] National Human Genome Research Institute:

[8] A universal rule for the distribution of sizes. Authors: Salingaros N.A, West B.J.

[9] Mosaic organization of DNA nucleotides. Authors: C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger

[10] The Fractal Geometry of Nature. Authors: Benoit Mandelbrot.

[11] Self-similarity and Power Laws. Authors: Tiina Komulainen. p. 3

[12] Power Law Correlations in DNA Sequences. Authors: S.V Buldyrev.

[13] Long-Range Correlation and Partial $\frac{1}{f^{\alpha}}$ Spectrum in a Noncoding DNA Sequence. Authors: W.Li, K.Kanenko.

[14] Fractals and Hidden Symmetries in DNA. Authors: Carlo Cattani.

[15] Fractals and Hidden Symmetries in DNA. Authors: Carlo Cattani.

[16] Fractal signatures of SARS-CoV2 coronavirus, the indicator matrix, the fractal dimension and the 2D directional wavelet transform: A comparative study with SARS-CoV, MERS-CoV and SARS-like coronavirus. Authors: Sid-Ali Ouadfuel.

[17] Fractals and Hidden Symmetries in DNA. Authors: Carlo Cattani.

[18] Fractal landscape analysis of DNA walks. Authors: C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, F. Sciortino, M. Simons, H.E. Stanley, S. Havlin. p. 1

[19] Mosaic organization of DNA nucleotides. Authors: C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger

[20] Long-range power-law correlations in condensed matter physics and biophysics. Authors: H.E.Stanley, S.V.Buldyrev, A.L.Goldberger, S.Havlin, C.-K.Peng, M.Simons. p. 1

[21] Molecular Biomarkers and classification models in the evaluation of the prognosis of colorectal cancer. Authors: Sideris M, Papagrigoriadis S.

[22] Telomeres-Structure, Function, and Regulation. Authors: Weisi Lu, Yi Zhang, Dan Liu, Zhou Songyang, Ma Wan.

[23] Genetic linkage analysis in the age of whole-genome sequencing. Authors: Jurg Ott, Jing Wang, Suzanne M. Leal.

[24] Power Law Correlations in DNA Sequences. Authors: S.V Buldyrev.

[25] Fractals in Biology and Medicine. Authors: Shlomo Havlin, Sergey V Buldyrev, Ary L Goldberger, Rosario N Mantegna, S. M Ossadnik, Chungkang Peng, Michael P Simons. p. 13

[26] Fractals in Biology and Medicine. Authors: Shlomo Havlin, Sergey V Buldyrev, Ary L Goldberger, Rosario N Mantegna, S. M Ossadnik, Chungkang Peng, Michael P Simons. p. 19