Metrics and visualisation for crime analysis and genomics

Laros, J.F.J.

Chapter 10

Visualising Genomes in 3D using Rauzy Projections

We propose a novel visualisation method for DNA and other long sequences over a small alphabet, which is based on the construction of the family of Rauzy fractals for infinite words. We use this technique to find repeating structures of widely varying length in the input string as well as the identification of coding segments. Other properties of the input can also come to light using this technique.

10.1 Introduction

Projections of high dimensional structures onto a low dimensional surfaces (e.g. 2D, 3D) are commonly used to make structures in the data insightful.

Recognising patterns in long sequences is difficult for humans. The longer the sequence, the harder it gets. Furthermore, if the alphabet used for this sequence is small, the task gets even harder. If there are (small) deviations allowed in the patterns to be recognised, it will be nearly impossible without the aid of some sort.

In this chapter, we introduce a visualisation technique for DNA sequences using a projection onto a surface to investigate patterns in the DNA.

In Section 10.2 we explain the underlying idea of making fractals out of infinite words, which we adapt for our purposes in Sections 10.3. In Section 10.4 we show the visualisation results. Finally, related work is discussed in Section 10.5 and we conclude in Section 10.6.

10.2 Background

In this section, we describe the approach of Rauzy [59] to construct a fractal from an infinite word.

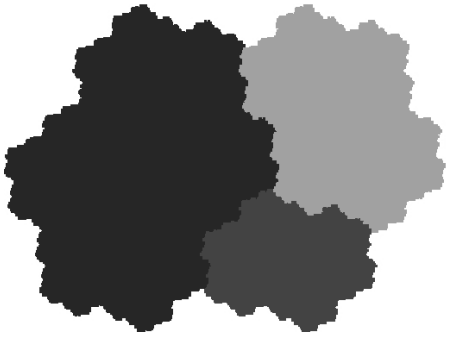

Figure 10.1: Standard Rauzy fractal

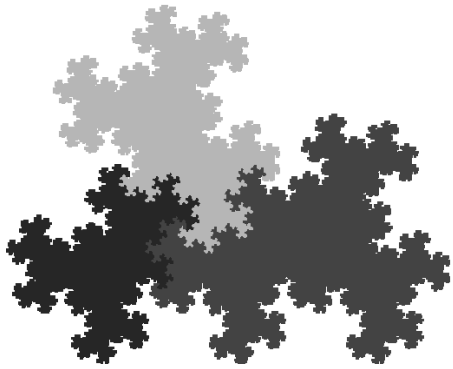

Figure 10.2: A "Rauzy" fractal using a different substitution

Given a finite alphabet Σ, we denote the set of all finite strings over this alphabet as Σ^*. In this section we take $\Sigma = \{0, 1, 2\}$.

Rauzy investigated the so-called tribonacci substitution:

$$\sigma : \begin{cases} 0 & \to & 01 \\ 1 & \to & 02 \\ 2 & \to & 0 \end{cases}$$

This substitution induces a *homomorphism*, again denoted by σ, from Σ^* to Σ^*. It is uniquely extended to Σ^* by requiring $\sigma(u \cdot v) = \sigma(u) \cdot \sigma(v)$ for all $u, v \in \Sigma^*$, where $\cdot$ denotes concatenation of strings.

Since 0 is a prefix of $\sigma(0)$, the following holds: $\sigma^n(0)$ is a prefix of $\sigma^{n+1}(0)(n = 1, 2, \ldots)$. Also $|\sigma^n(0)| \geq n \to \infty$ when $n \to \infty$. Therefore $(\sigma^n(0))_{n\in\mathbb{N}}$ defines a unique infinite word that has $\sigma^n(0)$ as finite prefix for each $n \in \mathbb{N}$. This word is invariant under the substitution (where we simultaneously substitute each letter). We call this word an accumulation point or fixed point of this substitution. For the given substitution σ, we get the word $01020100102010102010 0 \ldots$ as fixed point.

Rauzy used this infinite word to make a plot in 3-dimensional space, starting in $(0, 0, 0)$ and doing one step in the x-direction whenever a 0 occurs, a step in the y-direction when a 1 occurs and a step in the z-direction when a 2 occurs. All steps are of equal length. This results in a so-called *broken halfline* which "approximates" a halfline ℓ starting in the origin. The existence of this halfline is discussed in [59].

Now we can take the plane through the origin which is perpendicular to the line ℓ and project the broken halfline onto this plane. This results in the *Rauzy fractal*, depicted in Figure 10.1. The colour of the projected points depends on whether the next step from the original point on the broken halfline is in the x-, y-, or z-direction.

When a different substitution is used, fractals can be made in the same manner. We see such a fractal in Figure 10.2. The substitution in applied here is:

$$\sigma : \begin{cases} 0 & \to & 01 \\ 1 & \to & 2 \\ 2 & \to & 0 \end{cases}$$

10.3 Application to DNA

We can apply a similar technique on a DNA sequence. First we have to make a plot in 4-dimensional space by associating each nucleotide (`A`, `C`, `G` and `T`) with one of the spacial dimensions and analogous to the construction described above building a broken halfline, but now in four dimensions.

Then we can make a 2- or 3-dimensional projection by either choosing a plane or a hyperplane through the origin and by projecting the broken halfline upon that (hyper)plane.

One big difference with the Rauzy approach in three dimensions is that the choice of the line ℓ which approximates the broken halfline is not pre-determined. A DNA string does not have the nice mathematical properties that the tribonacci fixed point possesses, so there is no clear preference for ℓ. One choice could be the line going through the begin- and endpoint of the broken halfline. This is possible since a DNA string is finite. However, this will result in different choices for ℓ for different DNA strings. This might not be the best choice. Therefore, we shall look (only) at a solution that uses a fixed ℓ for every input.

If ℓ is chosen well, we expect to find the following in the projected image:

- A non-predictable walk for information rich parts of the DNA.
- A true random walk for random parts.
- Lines (or approximate lines) for repeating parts of the DNA.
- Large copies of substrings in the DNA, that can be easily visualised.

The term *non-predictable walk* should be taken loosely, because we know that coding DNA (and therefore information rich DNA) has a higher `GC`-content than other parts of the DNA. Therefore, the walk will tend to go into the `GC`-direction, and second, DNA is clearly not random.

Since we have so much freedom in the choice of ℓ, we can also look at the projection in the following way: In principle, we can choose four vectors in the plane and use these four vectors to make the projection. Associate the first vector with the nucleotide `A`, the second one with `C`, and so on. Now, to make an insightful projection, we want the four vectors to have the following properties:

- The vectors should be of comparable length.

- The four vectors should add up to 0.
- Every subset of three vectors should be independent.

By adhering to these properties, we arrive in the same point if we consecutively plot a combination of all four letters, this would thus indicate a perfect repeat of those four letters, or a repeat of all four letters in any order.

When using an interactive program like `gnuplot` to visualise the data, the angle between the vectors can easily be adjusted (especially when making 2-D projections) by stretching the image. We must however, make sure that the vectors are not chosen parallel to the axes, otherwise the stretching will only result in the alteration of the length of the vectors. This property of interactive programs, along with the ability to zoom in, makes the exploration of dense areas in our visualisation possible.

We made an interactive web application available via [44]. This demonstration has access to the first 1,000,000 base pairs of the Human chromosome 1 and can visualise any substring. For performance reasons, we added a *step size* to make the rendering of large substrings possible. The data is located on the webserver and it is kept small for practical purposes. Making more data available will have no effect on the visualisation application.

10.4 A number of DNA sequence visualisations

In this section, we shall make several projections, on both two and tree dimensional hyperplanes. In each case, we choose a set of vectors for the nucleotides `A`, `C`, `G` and `T`. We denote these vectors by $v_{\mathtt{A}}$, $v_{\mathtt{C}}$, $v_{\mathtt{G}}$ and $v_{\mathtt{T}}$ respectively.

10.4.1 Projections in three dimensions

For three dimensions, there is, apart from symmetry, a natural choice of the projection line ℓ and therefore the resulting vectors. This set is the one that defines a tetrahedron (centred at the origin), as shown in Table 10.1. In other words, the convex hull of the set of endpoints of these vectors forms a tetrahedron. All vectors in this set have the same length, the four of them sum up to 0 and each three-element subset is independent. Furthermore, the vectors are uniformly distributed, i.e., the angle between each pair of vectors is equal.

$$\begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{pmatrix}$$

Table 10.1: Vertices of a tetrahedron

We associate the first row with $v_{\mathtt{A}}$, the second row with $v_{\mathtt{C}}$ and so on. In the remainder of this section, we shall use these vectors for our projections.

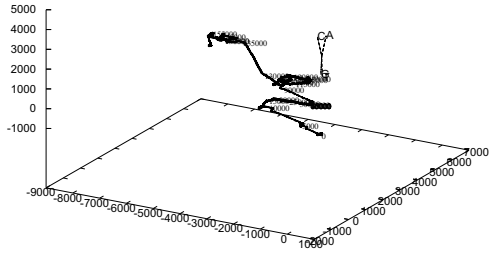

Figure 10.3: The first 160,000 nucleotides of the human Y-chromosome

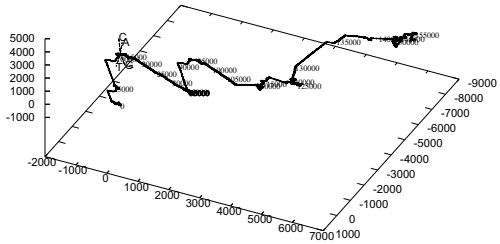

Figure 10.4: The first 160,000 nucleotides of the human Y-chromosome

As input for our three dimensional projection, we use the first 160,000 nucleotides of the human Y-chromosome [69] (build 18). This results in the picture shown in Figure 10.3. For clarity, we include the four vectors and the associated nucleotides in the visualisation.

In Figure 10.4, we see the exact same data and projection, but shown from a different angle.

This figure is a better representation of the data, more structures can be seen directly from this angle. We shall discuss the findings in detail in Section 10.4.2.

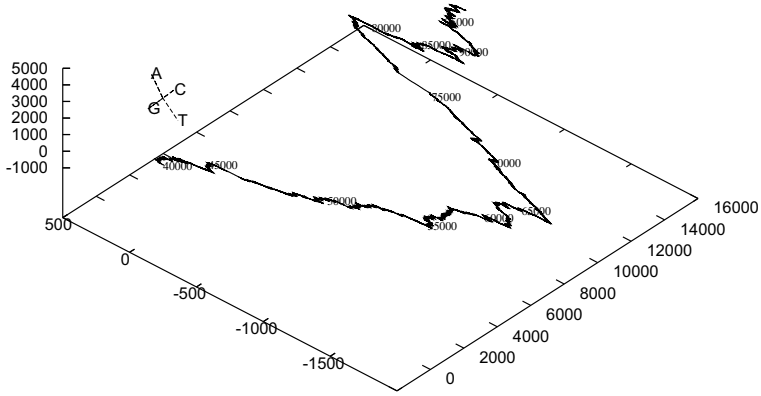

Figure 10.5: Offset 40,000–100,000 of the human chromosome 1

In Figure 10.5, we see a part of the first human chromosome, again, we shall discuss the findings in detail in Section 10.4.2.

10.4.2 Projections in two dimensions

Since, in two dimensions, there is no way to choose four vectors of equal length, of which all subsets are independent, we must choose the vectors in such a way that the lengths differ, to satisfy the independency constraint. Therefore we can choose e.g., the following vectors for `A`, `C`, `G` and `T` respectively: $(5, 7)$, $(-7, 6)$, $(8, -4)$ and $(-6, -9)$. By choosing these vectors, we have the property that every pair is independent, and the lengths are comparable. For practical purposes, we have chosen integer coordinates.

As input we use the first 160,000 nucleotides of the human Y-chromosome [69] (build 18). This results in the following projection:

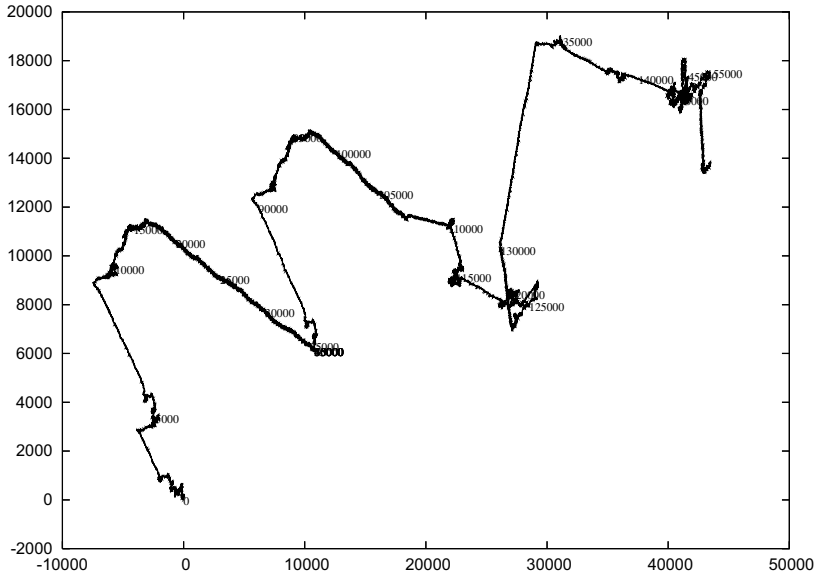

Figure 10.6: The first 160,000 nucleotides of the human Y-chromosome

The numbers in this plot denote the offset in the DNA, the plot starts in $(0, 0)$. We immediately see that two of the three assumptions from Section 10.3 can be verified for this input data. There are lines, which denote (approximate) repeats. For example, the line that starts somewhere near $(-3000, 4500)$ and ends in $(-7500, 9000)$ contains a large number of approximate copies of the string `CCCCGCTCCTCCCCTCGGGACCACCCCAGA`. In the region near $(23000, 9000)$, marked by the offset 115000, we see a part where the walk seems random. Moreover, we can see an extremely large substring from $(-2500, 3000)$ to $(5500, 8500)$ (approximately offset 5000 to 25000) that repeats itself from $(11000, 6000)$ to $(18500, 11500)$ (approximately offset 93000 to 107000).

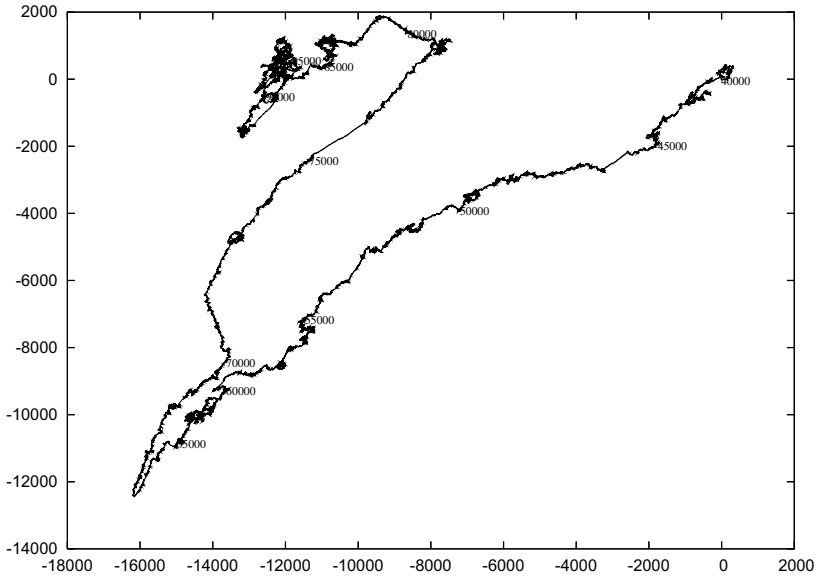

Figure 10.7: Offset 40,000–100,000 of the human chromosome 1

In Figure 10.7 we see a part of the human chromosome 1. No large repeats are noticeable in this part of the genome, but we can clearly see some short repeating sequences. The string `TTC` for example, is repeated a number of times from position $(-2600, -2200)$ to $(-3200, -2700)$. Another obvious short repeat `AAG`, can be seen from position $(-11200, -2200)$ to $(-9700, -1200)$, approximately.

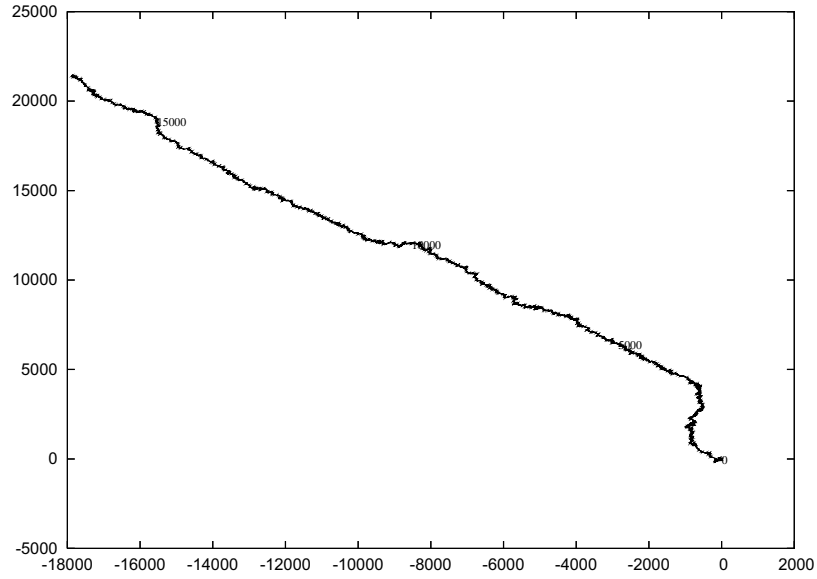

Figure 10.8: The complete mitochondrial DNA

In Figure 10.8, we see the DNA of mitochondria found in the human species.

The most noticeable is the drift in the C-direction. Further analysis show that this is due to the relatively low content of G-nucleotides in the mitochondrial DNA (13% versus 25–31% for the other nucleotides).

10.5 Related work

In the DNA-rainbow [3] project, plots of DNA were made by assigning a colour to a nucleotide; green for Adenine, red for Thymine, white for Guanine and blue for Cytosine. Undetermined positions were given a grey colour. Then the nucleotides were plotted to a standard bitmap with a width of 770 pixels. The resulting pictures, viewable with any picture viewer and/or editor, give a colourful impression of the DNA. Most of the pictures appear random, but in some parts, repeating parts of the genome can be seen in the form of diagonal lines. A nice property of this approach is that GC-rich areas (which are associated with encoding DNA), can be spotted right away.

An obvious shortcoming of this approach is the fixed width of the picture. Short repeating sequences which have a total length of under 770, will not stand out. Very long repeating sequences will not stand out either, since it will look like a random block that is repeated. However, for the detection of short repeating sequences (that are repeated for a large number of times), this approach seems very well suited.

In [62], several visualisation methods are investigated. One of them is essentially the same as in the previously described project, but here the width of the columns can be changed to detect (small) repeats of different length. Other visualisation methods include information hiding by using less colours, and usage of expert knowledge to emphasise some subsequences, like start and stop codons. Furthermore, a translation to amino acids can also be made with this technique. According to the authors, their proposed method is useful for sequences up to a length of 2,000 base pairs.

In [53], several visualisation techniques are reviewed: the “random walk” visualisation, as well as a fractal visualisation and a visualisation based on entropy-like parameters which are calculated within a sliding window. The “random walk” resembles the visualisation discussed in this chapter, with the exception that the directions that are associated with the nucleotides are fixed and the technique is limited to two dimensions.

10.6 Conclusions and further research

We have shown that all our hypothesis were confirmed; the simple repeats indeed show up as lines in our visualisation. What we did not expect were the large repeats (the thick lines mentioned in Section 10.4), although it is a rather nice result. Furthermore, we detected a large approximate repeat on the first part of the Y-chromosome. This repeat is already known by genetic experts, but it is

nice to have detected it with no excessive calculation, like the alignment of two large sequences, provides hope for further exploration of DNA in this way.

As recommendations for further research, we suggest using colours as an extra coding scheme. By doing this, we can see the direction of a line in our visualisation. For example, it is hard to distinguish between simple repeats `AAG` and `CTT` because they are almost each others inverse. The lines resulting from these repeats will have approximately the same slope (see Figure 10.7 for an example of this), although their contents is different. The only ways to distinguish them at this point is to measure the exact slope, or by looking at the offsets and thereby finding the orientation of the line. By using a colour coding scheme, the colour of the line will represent the orientation directly.

An other interesting extension would be to make the line ℓ along which we project the 4-D structure onto a surface a parameter that can be changed in real time. A potential user could try to find a projection, better suited for his or her purposes.