Project report
Identification of replication origin in DNA sequence
ELL796

Abhishek Pathak

1 Introduction

Identification of the origin of replication in genomes is an important problem with varied applications. The point of origin of replication can be identified with experimental techniques, for example by removing certain genes/parts of the genome, and seeing its effect on the ability of the genome to replicate. Apart from wet-lab based approaches, there has also been work on computational prediction of the origin of replication or ORI. In this work, we look at three existing computational methods - cumulative GC skew, auto-correlation based measure and a new auto-correlation computed using fourth roots of unity. We study these methods on certain genomes. The genomes used in this study are bacterial genomes.

2 Datasets used

The bacterial genomes used in this study are:

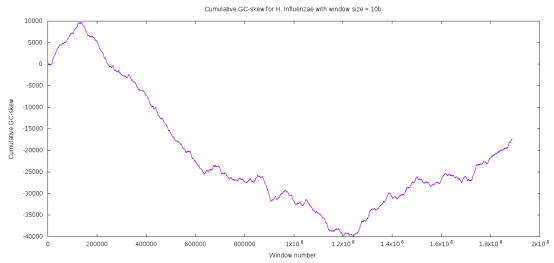
- *B. subtilis* (NC_000964)
- *M. jannaschii* (NC_000909)
- *N. tabacum* plastid (NC_001879)
- *E. coli* (NC_017626)
- *H. Influenzae* PittGG, complete genome (CP000672.1)

These bacterial genomes (and most bacterial genomes in general) are composed of a single 'chromosome' that is circular in shape. The origin of replication is on a point in this circular genome. From here, replication starts and extends in both directions till it reaches the point opposite to the ORI.

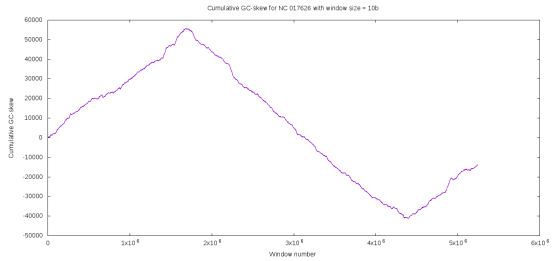
3 Replicating Cumulative GC-skew

Cumulative GC-skew is just that - the cumulative value of the GC-skew in a chromosome across consecutive windows of consideration. GC-skew for a window is defined as $\frac{n_G - n_C}{n_G + n_C}$, where n_G is the number of G's and n_C is the number of C's in the window.

Some figures of cumulative GC-skew for various genomes are as follows.



(a) Cumulative GC skew for *H. Influenzae*



(b) Cumulative GC skew for NC_017626

Figure 1: Cumulative GC-skew plots

As can be seen, the clear trend of increasing followed by decreasing cumulative skew is very prominent in these genomes.

We also tried cumulative GC skew on other genomes.

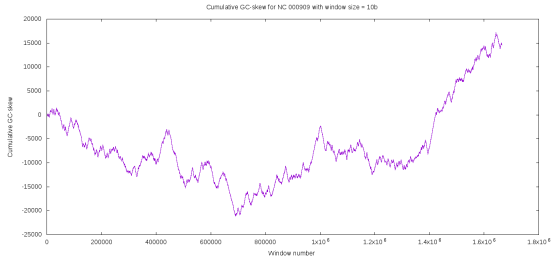


Figure 2: Cumulative GC skew for NC_000909

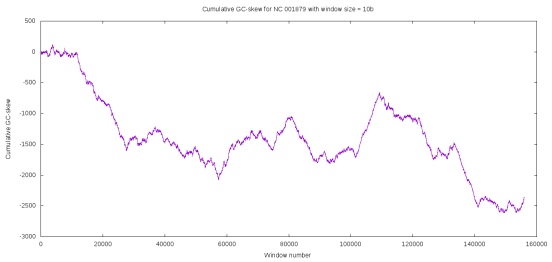


Figure 3: Cumulative GC skew for NC_001879

Clearly, here the method is not so effective - there is a trend of decreasing followed by increasing in NC_000909, but none which can help us infer the ORI in a straightforward manner.

In this, we implement and study cumulative GC-skew. It was observed that varying the window size over which skew is computed did not significantly alter the plots or the inference.

4 Replicating autocorrelation measure

The autocorrelation measure used uses the autocorrelation function $C(k)$ of a discrete sequence, $\{a_i : i = 1, 2, \ldots, N\}$ with $a_i \in \{+1, -1\}$, defined as

$$C(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} a_j a_{j+k} \tag{1}$$

The correlation measure C_G for the sequence is then defined as the average over all correlation values.

$$C_G = \frac{1}{N-1} \sum_{k=1}^{N-1} |C(k)| \tag{2}$$

Implementing this formulation and plotting the resulting graphs gives us the below results.

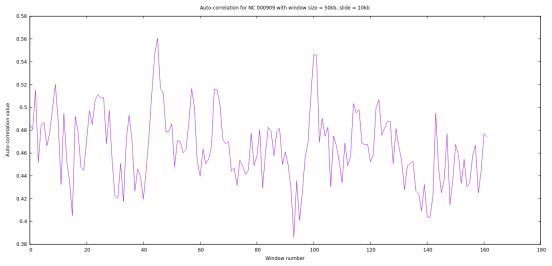


Figure 4: Auto-correlation for NC_000964

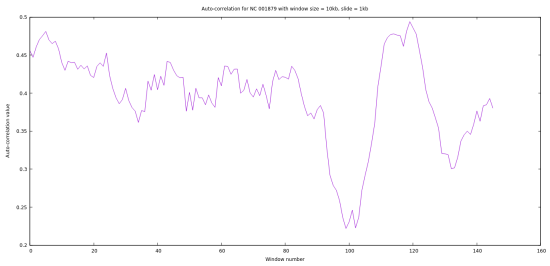


Figure 5: Auto-correlation for NC_001879

These plots are perfectly replicated for the cases as studied in [4].

5 Replicating iCorr measure

The iCorr measure is similar to the auto-correlation measure. This time, each nucleotide in the sequence is assigned a complex fourth root of unity to compute the measure.

iCorr also replicated perfectly - here are some results, which can be compared with the original paper [5].

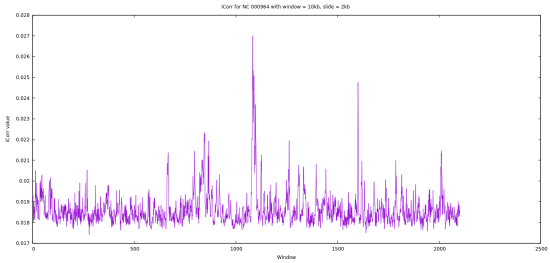


Figure 6: iCorrelation for NC_000964

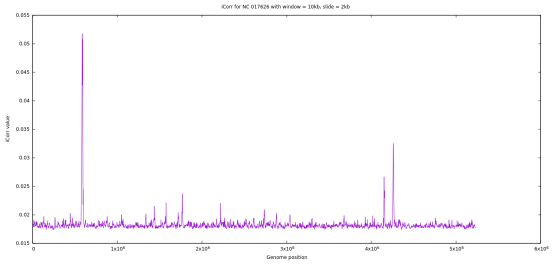


Figure 7: iCorrelation for NC_017626

6 Computational complexity

The complexity measure of implementing this algorithm for both autocorr and iCorr is given by $\mathcal{O}(\frac{NW^2}{s})$, where

- N = Number of elements in transformed sequence
- W = Window length
- s = Slide distance

7 Transforming sequences

To incorporate the extension of sequences from that of A, T, G and C nucleotides to that of 2-mers and 3-mers of nucleotides, the following steps were carried out.

- A mapping was created from all possible k-mers of nucleotides, to the roots of unity.
- For k-mers, number of possibilities $= 4^k$, so by taking $n = 4^k$, the mapping was created from n possible k-mers to the n roots of unity.

This sequence was then operated upon to compute autocorrelation measures. Since this measure involves roots of unity, this new method can be called ωCorr. Note that the sequences can be computed by taking the kmers to be overlapping or disjoint.

8 ωCorr on transformed sequences

Results for the ωCorr method are shown below.

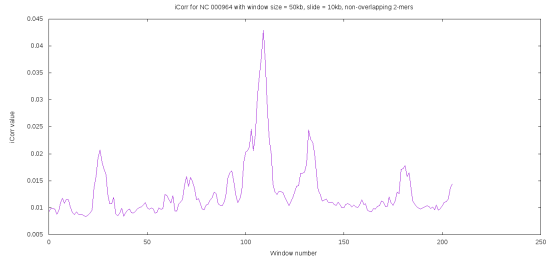


Figure 8: ωCorrelation for NC_000964 2-mers

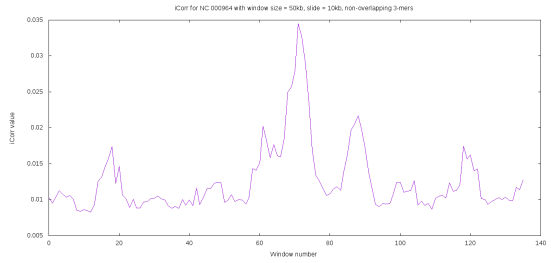


Figure 9: ωCorrelation for NC_000964 3-mers

Plots for 2-mers and 3-mers were quite similar here, as well as for some other genomes. This is surprising, as one would expect a different kind of behaviour of the curve for 2-mers

and 3-mers. This hints towards the possibility that for any two identical 2-mers in the genome, the 3-mers formed by taking the element next to these 2-mers are also likely to be the same.

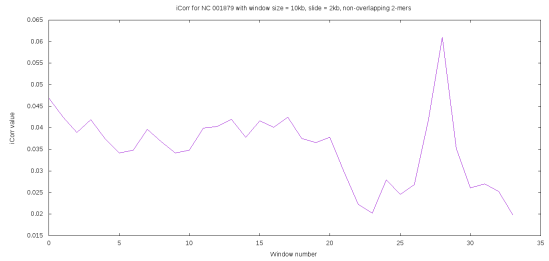


Figure 10: ωCorrelation for NC_001879

Again, this plot is seen to be strikingly similar to the corresponding auto-correlation plot for the same genome when only 1-mers (or single nucleotides) are considered at a time. This also points to the likelihood that two identical nucleotides in the sequence are followed by the same second nucleotide.

In general, it was observed that the sudden gradient changes and maxima/minima in the plots for 2-mers and 3-mers were very similar to those for the 1-mers/single nucleotide cases. In a few cases, the plot was different from the original. For instance, in the plot below, the three consecutive prominent maxima are worth investigating. The maxima which was present in the original iCorr plot, however, is present in 'reduced' form.

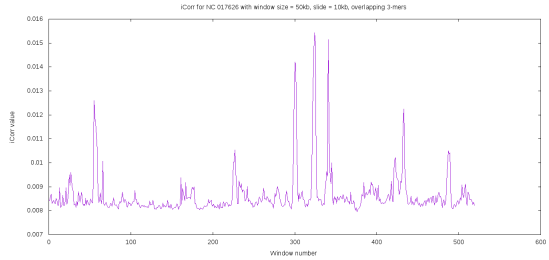


Figure 11: ωCorrelation for NC_000964

In many cases, the results were not very interpretable, or were similar to the iCorr and auto-correlation results.

9 Conclusion

We have studied many ways of computationally predicting ORI's in genomes, and devised a new scheme to apply the auto-correlation measures on the genome sequence via a transformation of the sequence. The results of the initial papers were successfully replicated and the computational complexity of calculating the auto-correlation measures was characterized. Finally, we transformed the sequences to sequences of 2-mers and 3-mers, and found that surprisingly often, the plots for these transformed sequences are quite similar to each other, and similar to the original auto-correlation measures as well.

References

[1] Yakovchuk, Peter and Protozanova, Ekaterina and Frank-Kamenetskii, Maxim D. *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.* Nucleic Acids Research, Volume 34, Issue 2, 1 January 2006, Pages 564 to 574, https://doi.org/10.1093/nar/gkj454

[2] Frederico LA, Kunkel TA, Shaw BR. *A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy.* Biochemistry, 1990 Mar 13;29(10):2532-7.

[3] Andrei Grigoriev. *Analyzing genomes with cumulative skew diagrams.* Nucleic Acids Research, 1998, Vol. 26, No. 10, Pages 2286-2290.

[4] Kushal Shah and Annangarachari Krishnamachari. *Nucleotide correlation based measure for identifying origin of replication in genomic sequences.* BioSystems 107 (2012) 5255.

[5] Shubham Kundal, Raunak Lohiya and Kushal Shah. *iCorr : Complex correlation method to detect origin of replication in prokaryotic and eukaryotic genomes.* arXiv:1701.00707 [q-bio.GN]