Plasmodium falciparum has rare correlation properties

Kushal Shah
*School of Computational and Integrative Science, Jawaharlal Nehru University, New Delhi - 110067, India.**

A plot of the correlation function of a given DNA sequence has certain characteristic features common to almost all organisms. One common feature is that the correlation values at distances that are multiples of three is higher than correlation values at other distances. Because of this such a correlation plot can be divided into two or three curves with different scalings. *P. falciparum* has a rare correlation property which is probably unique. I have analyzed genomes of many bacteria, fungi and protozoa and found that *P. falciparum* is the only organism whose DNA sequence correlation plot can be divided into four curves with different scalings. This property is neither shared by other species of the *Plasmodium* genus nor by other AT rich genomes. This could be a hint that the DNA sequence of *P. falciparum* has undergone certain rare mutational events.

I. INTRODUCTION

The DNA sequence contains all the information essential for the survival and growth of any organism [1]. Since the beginning of life on earth, these sequences have been subject to mutations which can be spontaneous [2] or induced [3]. Though spontaneous mutations would have produced a random DNA sequence, induced mutations lead to certain non-random features. More non-random features are incorporated in the DNA sequence by the process of natural selection (fitness) since any random sequence of nucleotides will not be able to successfully survive and replicate in given environmental conditions. The non-random nature of DNA sequences has also been supported by statistical analysis which has shown that these sequences have long-range correlations [4, 5].

Correlation studies of DNA sequences have not only contributed to a better statistical understanding of these naturally occuring sequences but also to the development of some biologically important insights and tools. In [6], it was shown that the correlation pattern of various organisms carry a characteristic signature which can be used to construct phylogenetic trees of various eukaryotes. This is an important contribution since unlike the other phylogenetic methods which are based on sequence alignment, a correlation based method is both much more simple and independent of the length of the DNA sequence. In another paper [7], jumps in correlation value along the DNA sequence were shown to be a better way to identify the origin of replication in various organisms. It was shown that a correlation based method is not only better than the conventional methods for identifying bacterial origins of replication, but it can also predict the origin of replication in certain higher organisms.

The above methods utilize the fact that the correlation pattern of each organism has certain characteristic features. However, the correlation pattern of the DNA sequences of almost all organisms also share some common features. One such interesting feature is that the value of correlation among nucleotides that are separated by a distance that is multiple of three is usually higher than the values of correlation at other distances [8]. Thus, a plot of correlation values with respect to nucleotide distance can usually be divided into two or three curves with different scalings (see Sec. III). This fact has also been linked to the presence of base 3 peridicities in genomes [9, 10]. Though this is true for almost all organisms, I have found a special case (*Plasmodium falciparum*) where the correlation plot corresponding to the Adenine (and Thymine) nucleotide can be divided into four curves with different scaling properties. This feature seems to be unique to *P. falciparum* and is most probably not shared by any other organism (not even other species of the *Plasmodium* genus).

In Sec. II, I have described the correlation method used. An analysis of the correlation plots of various organisms (*P. falciparum*, *P. knowlesi* and *D. discoideum*) is contained in Sec. III. At the end, Sec. IV contains the discussions and conclusions.

II. METHODS

The most commonly used way of calculating the autocorrelation of a DNA sequence is to convert it into a binary sequence. Since a DNA sequence is made up of four nucleotides (A,T,G,C), we can generate four binary sequences out of any given DNA sequence and each of these binary sequences can have different correlation properties. However, as it turns out in almost all cases, the correlation properties of A are very close to the correlation properties of T (similarly for G and C). Thus, in this paper, we will consider only the binary sequences associated with A and G. To generate the binary sequence corresponding to G, we assign a value $+1$ to every occurrence of G and -1 to all other positions (similarly for A). The lengths of both these sequences are identical and same as the length of the original DNA sequence.

Let us denote the binary sequence corresponding to the nucleotide A by $\{a_i : i = 1,2,...,N\}$ and that corresponding to G by $\{g_i : i = 1,2,...,N\}$ with $a_i, g_i \in \{+1,-1\}$. Now, the autocorrelation functions for the sequences $\{a_i\}$ and $\{g_i\}$ are given by [11, 12]

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

*Electronic address: kkshah@mail.jnu.ac.in

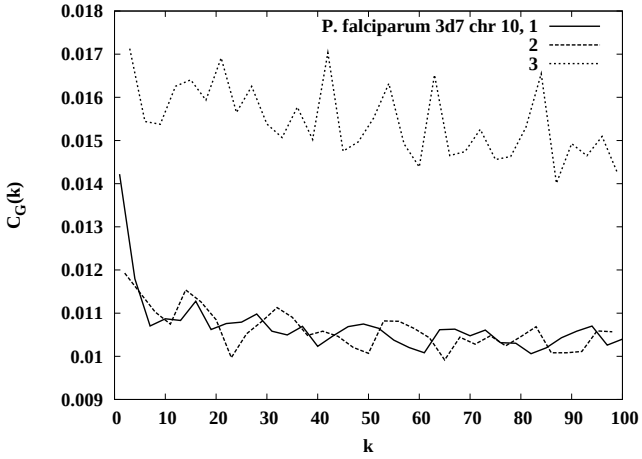

Figure 1: $C_G(k)$ vs. k for *P. falciparum* 3d7 chromosome 10. Curve 1 corresponds to $k = 1 \pmod 3$, curve 2 corresponds to $k = 2 \pmod 3$ and curve 3 corresponds to $k = 0 \pmod 3$. As can be seen, the values of $C_G(k)$ in curve 3 are higher than that for curve 1 and 2. This correlation pattern is similar to that observed in many other organisms.

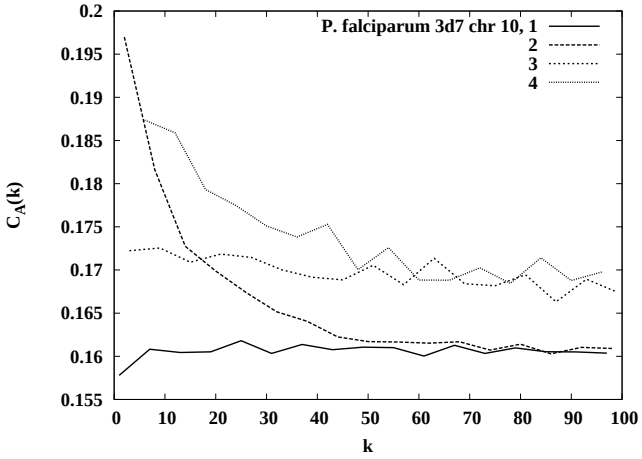

Figure 2: $C_A(k)$ vs. k for *P. falciparum* 3d7 chromosome 10. Curve 1 corresponds to $k = 1 \pmod 6$ and $k = 5 \pmod 6$, curve 2 corresponds to $k = 2 \pmod 6$ and $k = 4 \pmod 6$, curve 3 corresponds to $k = 3 \pmod 6$ and curve 4 corresponds to $k = 0 \pmod 6$. Among all the correlation plots we have analyzed, it is only the correlation plot for A (and T) of *P. falciparum* 3d7 that has four distinct scalings. This property makes this correlation pattern quite unique.

$$C_G(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} g_j g_{j+k} \tag{2}$$

where $k \in \{1,2,3,...,N-1\}$. Similarly, we can also define $C_T(k)$ and $C_C(k)$ corresponding to the nucleotide T and C respectively. One important feature of the auto-correlation function is that its value (between 0 and 1) is independent of the length of the sequence which makes it easier to compare DNA sequences of varying lengths. Lower value of $C(k)$ corresponds to lower correlation strength and vice-versa. The value of $C(k)$ for a typical random sequence will be zero and for a highly correlated sequence will approach unity.

All DNA sequences used in this paper were downloaded from NCBI [14] and PlasmoDB [15] websites.

III. RESULTS

The auto-correlation plot for A,T,G or C of almost any organism shows a distinct characteristic, namely that the magnitude of the auto-correlation for $k = 0 \pmod 3$ is higher than that for other values of k. This feature can also be linked to the presence of base 3 periodicity in genomes [8]. Though this feature is also present in the plots of both $C_G(k)$ and $C_A(k)$ for *P. falciparum*, the plot of $C_A(k)$ has some distinct features which are probably unique in nature.

Figure 1 shows the plot of $C_G(k)$ vs. k for *P. falciparum* 3d7 chromosome 10. Curve 1 corresponds to $k = 1 \pmod 3$, curve 2 corresponds to $k = 2 \pmod 3$ and curve 3 corresponds to $k = 0 \pmod 3$. As can be seen, the magnitude of $C_G(k)$ in curve 3 is higher than that for curve 1 and 2. This correlation pattern is same as that observed in many other organisms.

Figure 2 shows the plot of $C_A(k)$ vs. k for *P. falciparum* 3d7 chromosome 10. Curve 1 corresponds to $k = 1 \pmod 6$ and $k = 5 \pmod 6$, curve 2 corresponds to $k = 2 \pmod 6$ and $k = 4 \pmod 6$, curve 3 corresponds to $k = 3 \pmod 6$ and curve 4 corresponds to $k = 0 \pmod 6$. This clearly shows that the correlation for the nucleotide A of *P. falciparum* 3d7 chromosome 10 has four different scalings. This feature is shared by $C_A(k)$ and $C_T(k)$ of not only all *P. falciparum* 3d7 chromosome 10 but also all other chromosomes of this organism. It is an interesting finding that only the correlation patterns

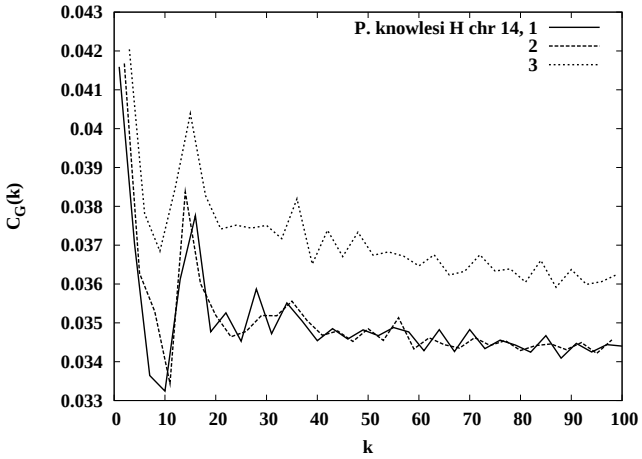

Figure 3: $C_G(k)$ vs. k for *P. knowlesi* strain H chromosome 14. Curve 1 corresponds to $k = 1 \pmod 3$, curve 2 corresponds to $k = 2 \pmod 3$ and curve 3 corresponds to $k = 0 \pmod 3$. As can be seen, the values of $C_G(k)$ in curve 3 are higher than that for curve 1 and 2 (at least for $k \leq 30$). This correlation pattern is similar to that observed in many other organisms.

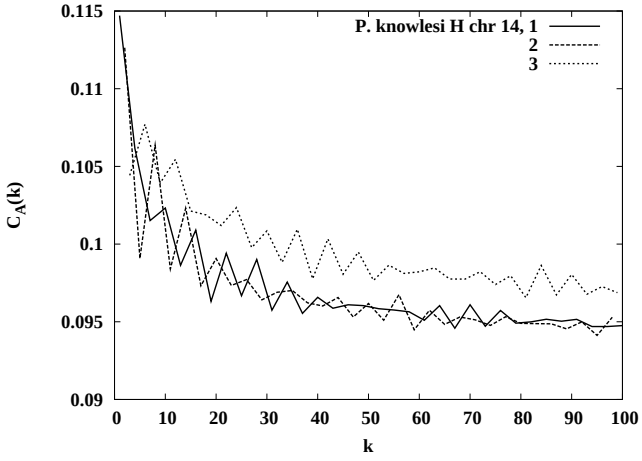

Figure 4: $C_A(k)$ vs. k for *P. knowlesi* strain H chromosome 14. Curve 1 corresponds to $k = 1 \pmod 3$, curve 2 corresponds to $k = 2 \pmod 3$ and curve 3 corresponds to $k = 0 \pmod 3$. As can be seen, the values of $C_G(k)$ in curve 3 are higher than that for curve 1 and 2 (at least for $k \leq 30$). This correlation pattern is similar to that observed in many other organisms.

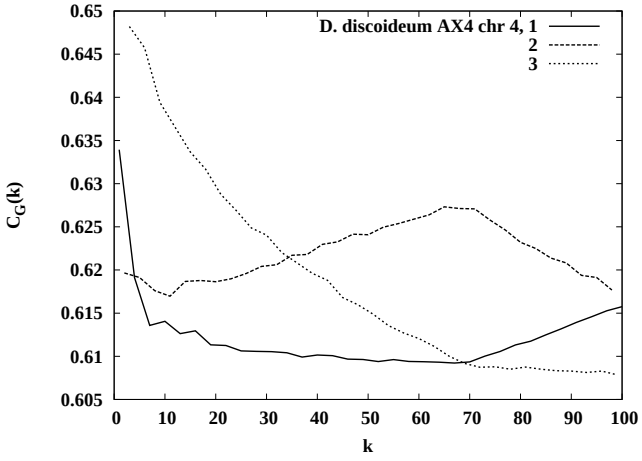

Figure 5: $C_G(k)$ vs. k for *D. discoideum* AX4 chromosome 4. Curve 1 corresponds to $k = 1 \pmod 3$, curve 2 corresponds to $k = 2 \pmod 3$ and curve 3 corresponds to $k = 0 \pmod 3$. As can be seen, the values of $C_G(k)$ in curve 3 are higher than that for curve 1 and 2 (at least for $k \leq 30$). This correlation pattern is similar to that observed in many other organisms.

corresponding to the Adenine and Thymine nucleotides of *P. falciparum* have this rare property. The correlation patterns corresponding to the Guanine and Cytosine residues show a behavior that is similar to that of many other organisms.

A question that arises is whether this same correlation pattern is observed in other species of *Plasmodium*. I have analyzed several chromosomes belonging to *P. vivax*, *P. berghei*, *P. chabaudi* and *P. knowlesi* but did not find separation into four scalings in any of them. Figures 3 and 4 show the plots for $C_G(k)$ and $C_A(k)$ in *P. knowlesi* strain H chromosome 14 and it can be clearly seen that this organism has correlation features similar to that found in many other organisms, namely that the correlation for $k = 0 \pmod 3$ is higher than that for other values of k. Similar results were obtained for chromosomes of other species belonging to the *Plasmodium* genus which have an unusually high AT content. In order to rule out the possibility that the rare correlation property of *P. falciparum* could be a result of its high AT content, I have analyzed the correlation plots of other organisms (not belonging to *Plasmodium* genus) with a high AT content and found that the correlation feature of *P. falciparum* (80.61% AT content) is indeed unique. Figures 5 and 6 show the plots for $C_G(k)$ and $C_A(k)$ of *D. discoideum* AX4 chromosome 4 (77.74% AT content) and it can be clearly seen that this organism also has the standard correlation features found in many other organisms.

Among all the correlation plots I have analyzed, it is only the correlation plot for A (and T) of *P. falciparum* 3d7 that has four distinct scalings. This property makes this correlation pattern quite unique. I have also analyzed the genomes of many other bacteria, fungi and protozoa but not found this feature in any of them. This hints at the possibility that some unique evolutionary events have happened in *P. falciparum*

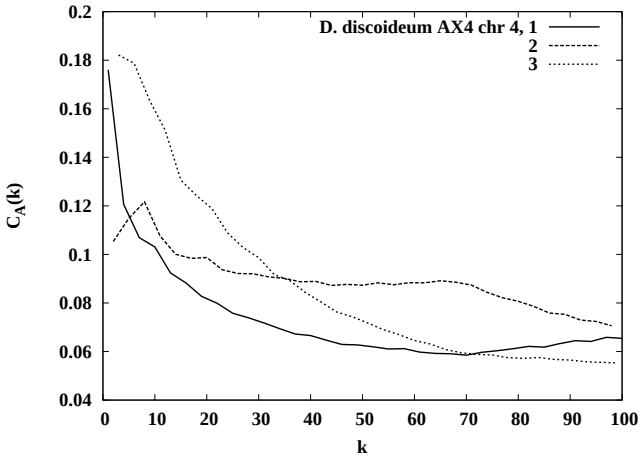

Figure 6: $C_A(k)$ vs. k for *D. discoideum* AX4 chromosome 4. Curve 1 corresponds to $k = 1 \pmod 3$, curve 2 corresponds to $k = 2 \pmod 3$ and curve 3 corresponds to $k = 0 \pmod 3$. As can be seen, the values of $C_G(k)$ in curve 3 are higher than that for curve 1 and 2 (at least for $k \leq 30$). This correlation pattern is similar to that observed in many other organisms.

which are not shared by any other organism.

IV. DISCUSSION AND CONCLUSION

Though it is known that the correlation strength at $k = 0 \pmod 3$ is higher than that at other values of k, the reason for the presence of this property in almost all DNA sequences is not yet understood. What is known is that this property

is not due to differences in percentage occurrence of various nucleotides at the three positions [8]. In the case of *P. falciparum* also, the presence of four different scalings is not due to any inhomogeneity in the nucleotide percentage at the six positions along the DNA sequence. Uncovering the reason for the presence of such correlation structures in DNA sequences will surely lead to a better understanding of the underlying mutational dynamics and I hope that this reason will be found in the near future.

It has been shown in this paper that the correlation pattern of *P. falciparum* is quite rare and probably unique. This correlation pattern is different from even that of other *Plasmodium* species and other AT rich genomes. However, the correlation plot of *P. knowlesi* also has some properties different from that of *D. discoideum*. As can be seen in Figs. 5 and 6, the three distint curves of correlation for *D. discoideum* are quite smooth. However, as can be seen in Figs. 3 and 4, the various correlation curves of *P. knowlesi* are very noisy. This noisy behavior is also characteristic of the correlation patterns of species belonging to *Plasmodium*. Correlation plots of other organisms have a smooth behavior like that of *D. discoideum*. This could possibly hint at the presence of certain rare mutational pressures experienced by all the organisms belonging to *Plasmodium*.

Species belonging to *Plasmodium* are known to cause malaria, which is one of the most severe human infectous diseases [13]. Several major research groups around the world are working towards developing an effective drug to combat this deadly disease. Any success in this direction will crucially depend on the identification of certain special properties of the DNA sequence of *P. falciparum* and other *Plasmodium* species. In this paper, I have shown that the DNA sequence of *P. falciparum* has a very special property which as well might be unique to this organism. Since the correlation patterns in a DNA sequence are the result of numerous mutation and selection pressures, it is quite possible that the DNA sequence of *P. falciparum* has undergone certain uncommon evolutionary dynamics. A further investigation in this direction could help us in unravelling important phenomenon associated with the evolution of *P. falciparum* which, in return, might help us in developing effective drugs to combat malaria.

Acknowledgments

I would like to thank SCIS-JNU, the Center of Excellence, Department of Biotechnology (Government of India) for financial support.

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular biology of the cell* (Garland Science, New York, 2002).
[2] S. E. Luria and M. Delbruck, *Genetics* 28, 491 (1943).
[3] J. Cairns, J. Overbaugh and S. Miller, *Nature* 335, 142 (1988).
[4] W. Li, T. G. Marr and K. Kaneko, *Physica D* 75, 392 (1994).
[5] S. Hod and U. Keshet, *Physical Review E* 70, 015104(R) (2004).
[6] M. Dehnert, R. Plaumann, W. E. Helm and M. Hutt, *J. Comp. Biol.* 12, 545 (2005).
[7] K. Shah and A. Krishnamachari, BioSystems 107, 52 (2012).
[8] K. Shah and A. Krishnamachari, BioSystems doi: 10.1016/j.biosystems.2011.11.006 (in press).
[9] H. Herzel, O. Weiss and E. N. Trifonov, *Bioinformatics* 15, 187 (1999).
[10] E. N. Trifonov, *Physica A* 249, 511 (1998).
[11] K. G. Beauchamp and C. K. Yuen, *Digital methods for signal analysis* (Geroge Allen and Unwin, London, 1979).
[12] T. J. Cavicchi, *Digital Signal Processing* (John Wiley & Sons, New York, 2000).
[13] S. M. Rich, F. H. Leendertz, G. Xu et. al., Proc Natl Acad Sci USA 106, 14902 (2009)
[14] NCBI: ftp://ftp.ncbi.nlm.nih.gov/genomes
[15] PlasmoDB: http://plasmodb.org/plasmo