STM JOURNALS

Research & Reviews: A Journal of
Bioinformatics

ISSN: 2393-8722
Volume 11, Issue 3, 2024
September—December
DOI (Journal): 10.37591/RRJoBI

https://journals.stmjournals.com/rrjobi

Review RRJoBI

Identifying Origin of Replication (ORI) Sites in Genomic Sequence Using Python-Based Programming and Motif Analysis in Bioinformatics

Kashif Raza Siddique*

Abstract

ORI sites serve a critical function in DNA replication serving as the beginning point of the process. Identifying the spots appropriately means a lot and is important for the biologists working in the lab. Detecting the ORI is not only vital for the detection of replication sites but is also important in numerous biological processes. In this study, we offer a unique approach employing Python-based motif analysis to discover the ORI sites within the supplied genomic sequence. It also highlights the problems of finding the ORI site and what are the alternatives. The sequence of the nucleotides or the genome sequence of the organisms that are utilized in this study as a reference for the detection of ORI has been supplied in the reference section. In this study, we discuss the method discovering the ORI of different organisms, we also share open-source Python programs and the genomic sequence that we utilized while completing this research. This research contributes to advance the field of bioinformatics by establishing a user-friendly framework for ORI site discovery, simplifying key exploration into DNA replication mechanisms and genome dynamics. In this paper, we have also spoken about the motif and the features. Also defined the circadian cycle and gene expressions and talked about the evening elements. Using skew arrays will increase the results and it will undoubtedly aid in detecting the ORI of complicated organisms like E. coli.

Keywords: ORI sites, DNA replication, Python-based motif analysis, genome sequence, bioinformatics

INTRODUCTION

The ORI site is where the replication begins, and it is conducted by molecular copy processes called DNA polymerase. Finding the ORI location is a crucial job not just for how cells multiply but many biomedical problems. The first human patient to benefit from gene therapy was a four-year-old girl with an immune deficiency illness in 1990 [1]. The kid had to live in a sterile environment since she was so prone to infections. Using a viral vector with an artificial gene that codes for a therapeutic protein, gene therapy aims to deliberately infect a patient lacking a critical gene. The viral vector multiplies once it is inside the cell, eventually producing many molecules of the therapeutic protein that ultimately improve the patient's condition. Finding the location of the ORI in the vector's genome and making sure any genetic modification they conduct, does not interfere with the vector's ability to reproduce inside the cell are essential for researchers [2].

***Author for Correspondence**
Kashif Raza Siddique
E-mail: kashifbt21@yahoo.com

[1]Student, Department of Biotechnology, Goel Institute of Technology and Management, Lucknow, Uttar Pradesh, India.

Received Date: March 10, 2024
Accepted Date: March 15, 2024
Published Date: November 11, 2024

Citation: Kashif Raza Siddique. Identifying Origin of Replication (ORI) Sites in Genomic Sequence Using Python-Based Programming and Motif Analysis in Bioinformatics. Research & Reviews: A Journal of Bioinformatics. 2024; 11(3): 22–33p.

Hidden Message

In the beginning, the researcher identified the ORI of the known bacterial genome. Studies

indicate that the bacterial genome region responsible for expressing ORI typically comprises several hundred nucleotides [3]. So, the first issue emerges in mind is, how does the bacterial cell know to begin replication within this little region within the much bigger chromosome, which contains of millions of nucleotides? A "Hidden message" instructing the cell to start replicating here should be present in the ORI region. It has come to our attention that DNA A, a protein that attaches to a specific region of the ORI called a DNA A box, is responsible for initiating replication. Let's suppose that the DNA A box is a message within the DNA sequence asking DNA A to "Bind here".

A reason was added to hunt for the common words on the ORI because of the many biological processes, certain nucleotide sequences emerge remarkably regularly in limited parts of the genome. This is often because certain proteins can only attach to DNA if a specific string of nucleotides is present, and if there are more occurrences of the string, then it is more probable that binding will successfully occur (It is possible that mutations will disrupt the binding process). So, let's find k-Mers present the full genomic sequence of Vibrio cholerae. To do so, the frequency map was constructed that generates the value. so an empty dictionary was generated (Freq= {}).

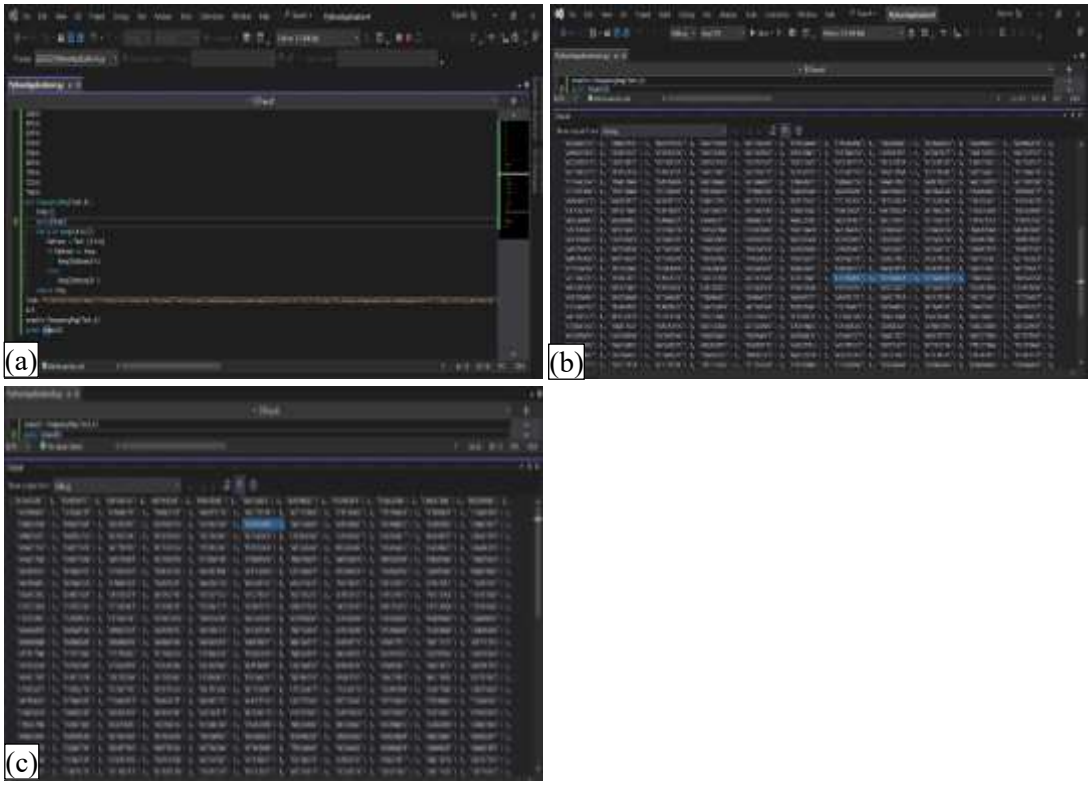

Figure 1. Python Code 1 (a), (b) and (c)

After applying the above code, the most repeated common 9-Mers in the sequence of the Vibrio cholerae will be found. So here are the 4 sequences of 9-Mers most repeated of 3 times.

"ATGATCAAG" "CTTGATCAT" "TCTTGATCA" "CTCTTGATC"

Now among the above 4 sequences, something interesting was witnessed that "ATGATCAAG" is a compliment or reverse compliment sequence to "CTTGATCAT". Now according to this point "ATGATCAAG" & "CTTGATCAT" appear 6 times either in compliment or reverse compliment form in the length of the ORI site of 540 nucleotides. This leads to the hypothesis that "ATGATCAAG" & "CTTGATCAT" indeed represent DNA A boxes that start replication. Now the "CTTGATCAT" presence and the repetition of this sequence is detected in the whole genome of the Vibrio cholerae.

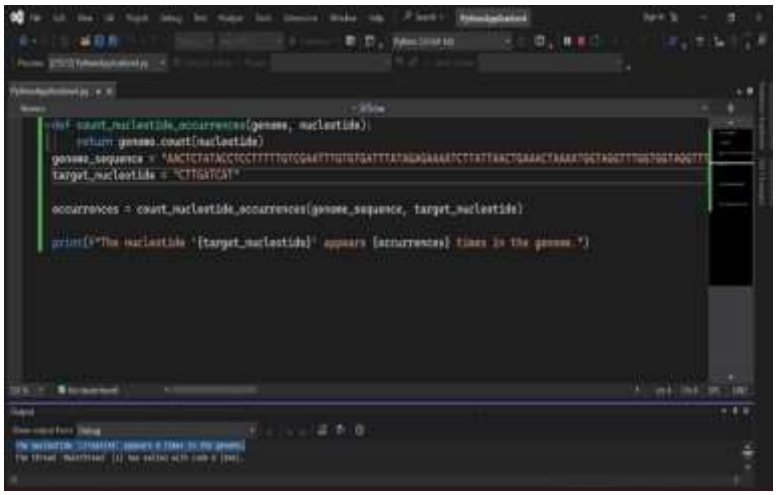

Figure 2. Python Code 2.

Now as the result says the "CTTGATCAT" has appeared 16 times in the full sequence. While the 3 is in the ORI site as we have seen in the last output of python code 1 shown in Figure 1. which is the smallest location where they are found the nearest. It establishes the presence of the ORI site and presents evidence that "ATGATCAAG" & "CTTGATCAT" may represent a DNA A box. But it is knownthat all the bacterial genomes don't have the same ORI sites. But does that even occur in the other bacterial site?

Now to verify that ORI of *Thermotoga petrophila* is tested which thrives in very hot surroundings. Its name is derived from the discovery in the water beneath oil reserves where the temperature surpasses 80ºC. The crucial fact is that the ORI of *Thermotoga petrophila* has 0 "CTTGATCAT". Output of python code2 is shown in Figure 2. So, it proved that different bacteria have different DNA A boxes to the DNA A protein. Also, using the above codes in this case shows that it has six 9-Mers with 3 times appearance [4].

"AACCTACCA" "AAACCTACC" "ACCTACCAC" "CCTACCACC" "GGTAGGTTT" "TGGTAGGTT"

To get the most frequent of them, a software called ORI-FINDER was utilized. This software chooses "CCTACCACC" along with its reverse complement "GGTAGGTTT" as a working hypothesis for the DNA A box in the Thermotoga petrophila which appeared 5 times (both). Now it is evident that if there's a new genome sequences the (500-3) clumps like "ATGATCAAG" & "CTTGATCAT" or "CCTACCACC" & "GGTAGGTTT" are not going to help. Now the difficulty is when this technique is built and used on the *E. coli* genome, hundreds of different 9-Mers were detected, creating (500-3) clumps on the genome. It is uncertain which of these 9-Mers represents the DNA A box in the ORI region. To tackle this problem, following process was followed.

The Meselson and Stahl Experiment

In 1958, the most beautiful experiment in biology took place. It included three theories of conflicting DNA replication models: dispersive, conservative, and semiconservative. A DNA is read in the 5ˡ → 3ˡ direction [5]. Now the replication procedure is used to locate the ORI. The replication fork has two-strand leading and lagging strands. The mutation rate will be higher in lagging due of the single strand than the leading strand because of the double strand. While analyzing the nucleotide counts for *Thermotoga petrophila* it was noticed that the frequencies of "A" and "T" were equal but the "C" was more frequent in the forward (leading) strand. these differences are due to the "C" tends to deaminate or convert into a "T". Let's follow up on this concept and determine the number of nucleotides in the reverse and forward strands.

Table 1. Represents the number of nucleotides present in the reverse and the forward strands and their differences in *Thermotoga petrophila*.

	"C"	**"G"**	**"T"**	**"A"**
Entire strand.	427419	413241	491488	491363
Reverse strand.	219518	201634	243963	246641
Forward strand.	207901	211607	247525	244722
Difference.	+11617	–9973	–3562	+1919

Single-stranded DNA causes deamination rates to increase 100-fold, which lowers the amount of cytosine on the forward strand. These lead to producing mismatched pairs of "T" & "G". These mismatched pairs will farther into the "T" & "A" pair when the bond is healed in the following round of the replication, which accounts for the decrease in the "G" on the reverse strand. Now, this idea generated by deamination is applied to locate the ORI in a circular bacterial genome as "C" is more frequent in one half of the genome and less frequent in the other half [6].

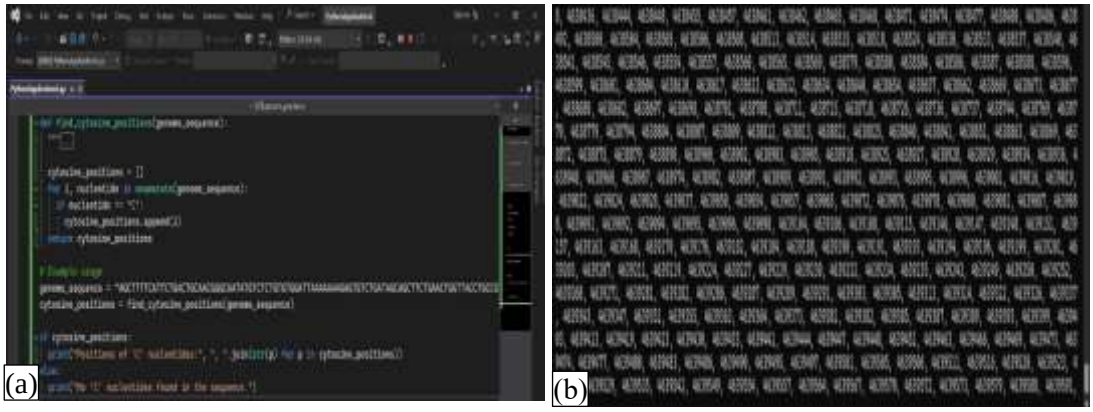

Figure 3. Python Code3 (a) and (b)

There is the output of 3 command pages, of a bacterial genome, such as *E. coli* which includes roughly 4.5 million nucleotides. There will be a need to conduct over ten trillion comparisons to produce a symbol array which will take many days on a fast home computer operating several millions of comparisons per second. Not only the occurrences of "C" but also "G" have the same occurrence disparities. Thus, the current objective is to maintain the overall discrepancies between the counts of "G" and "C." This difference indicates whether we are on the forward or reverse strand, depending on whether it increases or decreases shown in Figure 3 [7].

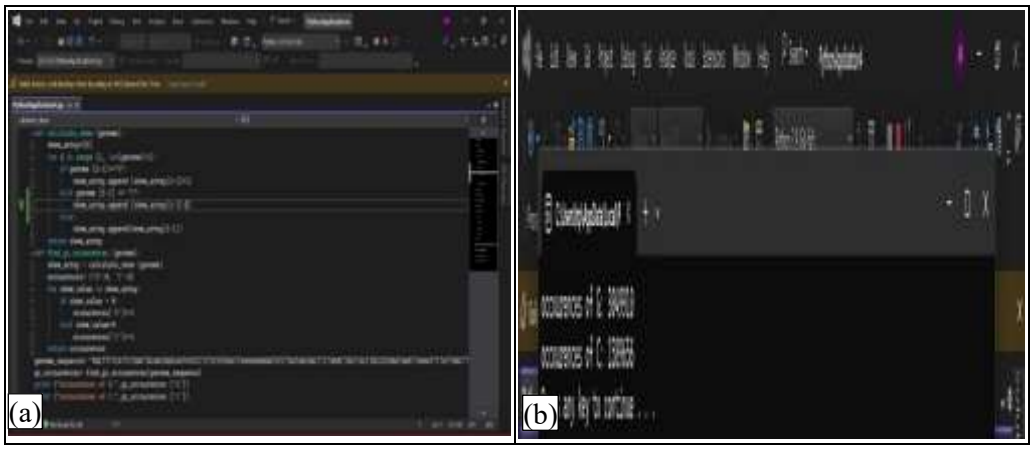

Figure 4. Python Code 4 (a) and (b).

As the output reveals the occurrences of "G" & "C" are 3049910 & 1589656 shown in Figure 4. But as the notion offered for this code was that if occurrences of "G" are more than the occurrences of "C" also shown in the output signifies it is a forward strand. Now the next thing is to find where the maximum of the "G" & "C" occurred within less distance.

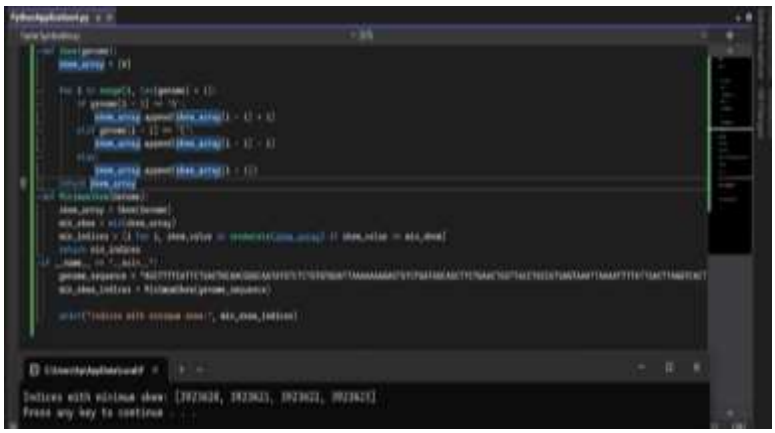

Figure 5. Python Code.5

Now as the result reveals the maximum number of instances of "G" & "C" are at 3923620, 3923621, 3923622, & 3923623 shown in Figure 5. All the output is near 4000000, so it is predicted that the ORI is close to it [8]. Now making the screw diagram below in Figure 6.

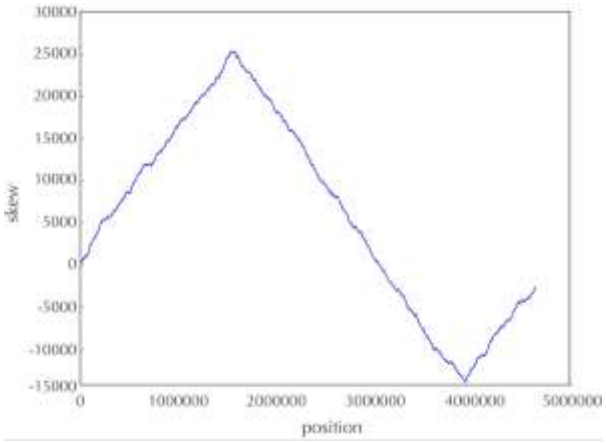

Figure 6. Screw diagram.

After solving it, the least skew value was obtained, roughly placing the origin at position 3923620 in *E. coli*. Keeping with this theme, let's search this area for a "HIDDEN MESSAGE" that would indicate a possible DNA A box. By resolving the common word problem in a 500-length window beginning at position 3923620, no 9-Mers (or their parts) that occur three or more times can be found [9]. So even if the position of ORI in *E. coli* has been located, it feels as if it is still not located the DNA A boxes. So, to examine the sequence of Vibrio cholerae and the most frequent of them while re-examining you will notice that there are 3 occurrences of "ATGATCAAG" and its reverse complement "CTTGATCAT", but the main thing is that it also contains additional occurrence of "ATGATCAAC" & "CATGATCAT" in only single nucleotide mismatch. But the 9-Mers makes more sense since DNA A can link not only to "perfect" DNA A boxes but to their small changes as well [10]. So, now our goal is to tweak our earlier approach for the frequent words problem to find DNA A boxes by recognizing frequent K-Mers, maybe with mismatches.

Approximate Pattern count (AAAAA, AACAA, GCATAAA, TTAAAGAG, 1) = 4

Because AAAA appears 4 times in this string with one mismatch observe that two of this overlap. In our final attempt to find DNA A boxes in the neighbourhood of the E. Because of skew fluctuations, even if the ORI is detected at position 3923620, it should not be assumed that it is located exactly at this location. Let's find the most common 9-Mers (within 1 mismatch) starting at location 3923620 of the *E. coli* within a window of length 500.

Figure 7. Output codes.

As the output of the code indicates the most frequent 9-Mers are "TGTGGATAA", "TTTGGATAA", "TTATACACA" & "TTATCCACA" shown in Figure 7, the experimentally confirmed DNA A box in E. It is "TTATCCACA" and is certainly the most frequent 9-Mers together with its reverse complement "TGTGGATAA" (with 1 mismatch). Moreover, "TTATCCACA" represents the most prevalent 9-Mers in this 500-nucleotide frame. It is not only one "GGATCCTGG", "GATCCCAGC", "GTTACCAC", "AGCTGGGAT" and "CTGGGATCA" (together with their reverse complement) appear 4 times. Although the function of any of these additional 9-Mers in the *E. coli* genome is unknown, it is known that hidden messages of many different kinds of frequently cluster inside genomes, and most of these messages are unrelated to replication. However, even supplying biologists with a tiny collection of 9-Mers as potential DNA A boxes is a huge aid since one of these 9-Mers is correct (even though bioinformaticians should cooperate with biologists to check their predictions) [11].

THE CIRCADIAN CLOCK

The circadian clock is an internal clock that regulates the daily schedules of all living things, including plants, animals, and bacteria. This clock never stops ticking, as anyone who has ever felt the agony of jet lag will attest. Like any watch, rats and research volunteers in bunkers naturally maintain a nearly 24-hour cycle of activity and rest under complete darkness. But much like any timepiece, the circadian clock is not infallible, and this can result in a genetic condition known as Delayed Sleep Phase Syndrome (DSPS) [12]. There are various questions raised by the necessity of a molecular foundation for the circadian clock. How do individual cells in plants and mammals tell what time it is, let alone microbes? Is the "clock gene" present?

Can the genes also be identified that are involved in "breaking" the circadian rhythm to cause DSPS? When mutant flies with irregular circadian, rhythms were discovered in the early 1970s, Ron Konopka and Seymour Benzor were able to identify the gene responsible for the mutation. The first clue came from the identification of a clock gene akin to this in mammals, which took biologists two or more decades to discover. These genes, whose names include timeless clock and cycle, control hundreds of other genes' activities and exhibit a high degree of interspecies evolutionary conservation [13]. Thousands more circadian genes have been discovered to date. Since maintaining a plant's circadian clock is essential to its survival, about it is essential to know how many plants are affected by the sun's rise and set times. Indeed, biologists estimate that a hundred thousand plant genes are circadian, including the gene associated to photosynthesis, photorespiration, and flowering but what does it mean for a gene to be circadian?

GENE EXPRESSION

"DNA makes RNA makes PROTEIN" is the fundamental precept of molecular biology. The four nucleotides A, G, C, and T make up the RNA strands that are first translated from the DNA corresponding to a gene. The amino acid sequence of the protein is subsequently translated from the RNA transcript and has a function in the cell. The resulting strand of RNA is translated into an amino acid sequence in this way. Codons are non-overlapping 3-Mers that are formed during translation of the RNA strand. Next, genetic coding is used to translate each codon into 20 amino acids. An amino acid string that spans a 20-letter alphabet can be used to represent the final sequence. The Figure 8 below illustrates the amino acids that are encoded by each of the 64 RNA codons (several codons encode the same protein/amino acid).

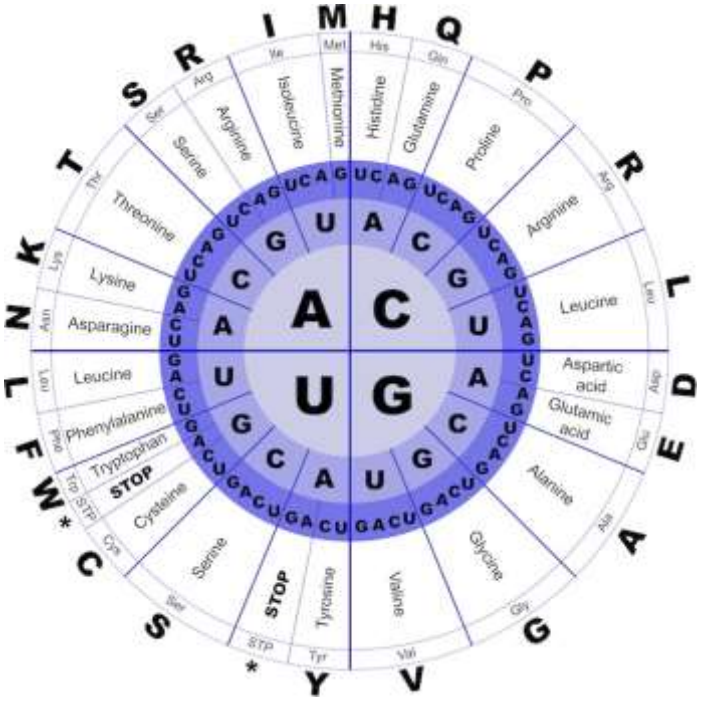

Figure 8. Illustrates the amino acids that are encoded by each of the 64 RNA codons

The key point is that the cells can transcribe different genes into RNA at different rates. This variance in the production of the gene transcripts, and gene expression, explains how a brain cell and a skin cell have the same DNA but perform vastly different functions. Variation in gene expression throughout the day also accounts for how the cell can keep track of time.

REGULATORY PROTEINS

It appears that each plant cell maintains its own schedule for day and night, and that the three plant genes LHY, CCA1, and TOC1 serve as the master timekeepers. These genes, along with the regulatory proteins they encode, are frequently subject to environmental stimuli (such sunshine or the availability of nutrients) that enable an organism to change the expression of certain genes [14]. Since the transcription factors or master regulatory proteins that these three proteins encode are responsible for turning on and off other genes, they can all affect the transcription of other genes. When a transcription factor binds to a particular short DNA sequence known as a transcription binding site or regulatory motif in the gene upstream region – a region that is 600–1000 nucleotides long and preceding to start of the gene. It would have been straightforward if the regulatory motifs were totally covered, but the reality is more complex as the regulatory motifs may change at some points [15]. However, how can these regulatory patterns be found in the absence of prior knowledge about their appearance? There is a need to create an algorithm for motif finding the challenge of discovering a "HIDDEN MESSAGE" shared by a group of strings.

EVENING ELEMENTS

Using DNA arrays, Steve Kay (2000) determined which genes in the Arabidopsis Thaliana plant are active throughout different periods of the day [16]. After that, he separated the upstream regions of about 500 genes that showed circular behaviour and searched for patterns that frequently appeared in these regions. This upstream section's unexpectedly frequent word, "AAAATATCT," appears 46 times when concatenated into a single string. Kay named this the "evening element," or "AAAATATCT," and conducted a straightforward experiment to confirm that this is, in fact, the regulatory motif in Arabidopsis Thaliana that controls the production of circadian genes. When the evening element in the upstream region of a gene was altered, the gene's circadian activity was eliminated. In contrast, plants' evening element is comparatively conserved, making them more difficult or easy to locate motifs with lots of variations [17]. A "Frequent Words with Mismatches Problem" could be created by converting the Frequent Words Problem. However, as it does not fully reflect the biological problem of motif finding, concatenating all the strings into a single string is not satisfactory. A pattern that clusters or recurs regularly within a DNA string is called a DNA box. A regulatory motif, on the other hand, is a pattern that appears at least once at several different, dispersed sites throughout the genome [18].

MOTIFS

When the set of circadian genes in plants was determined by Steve Kay using a DNA array, individual instances of motifs would be stored in a computational problem formulation for motif searching based on the degree to which a motif resembles a "ideal motif," or a transcription factor binding site that binds to the transcription factor most effectively. Since the exact theme is not known, nine mer from each string were chosen and assigned a score to each k-mer based on how similar they are to one another shown in Figure 9.

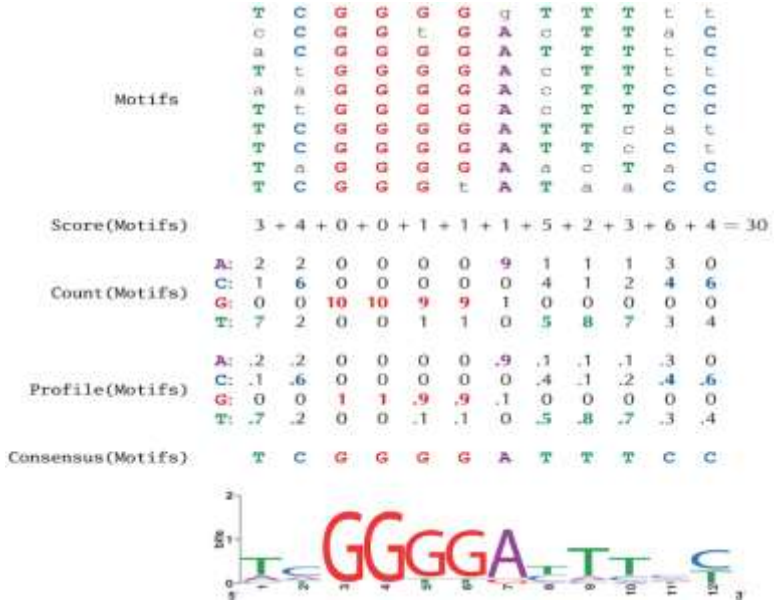

Figure 9. Motif.

MOTIF IN TUBERCULOSIS

The Mycobacterium tuberculosis bacterium (MTB) is the infectious agent that causes tuberculosis (TB), an infection that claims over a million lives annually [19]. Medication-resistant forms of tuberculosis are resurfacing, even though tuberculosis transmission has dramatically diminished. Since MTB can live in humans for decades without causing illness, it is a highly effective pathogen. Latent MTB infections, in which the germs lie dormant in the host's body and may or may not reactivate later, affect one-third of the world's population [20]. Suppressing TB outbreaks is challenging due to the high prevalence of latent infections. As a result, scientists are keen to learn what causes the illness to be latent and how MTB becomes active within a host.

How MTB survives during latency and why it can remain latent for so long are unknown. MTB may be able to stop producing most of its genes and remain dormant, much like bears hibernating in the winter, due to latent TB's resistance to medications. Bacteria that hibernate are known as sporulating because many of them produce metabolically inactive, protective spores that can withstand severe environments, enabling the germs to remain in the environment until conditions improve. Hypoxia, or low oxygen levels, is generally associated with latent forms of tuberculosis. With low-oxygen environments, biologists have discovered that MTB goes dormant, most likely with the hope that the host's lungs will eventually recover enough to allow the disease to spread. Finding the MTB genes necessary for the formation of the latent state in hypoxic environments is crucial since MTB demonstrates an amazing capacity to endure for years without oxygen. Finding a master regulator, or transcription factor, that "senses" a low oxygen environment and initiates a genetic programme that changes the expression of several genes, enabling MTB to adapt to hypoxia, is of interest to biologists. A transcription factor called dormancy survival regulator (DosR) controls many genes, the expression of which is markedly altered in hypoxic conditions. It was discovered by researchers in 2003. But DosR's transcription factor binding site was still unknown, and it was still unclear how it controlled these genes. Using a DNA array experiment, scientists were able to identify 25 genes whose expression levels differed significantly in hypoxic environments, which helped them to solve this mystery. Considering the 250 nucleotide upstream regions of these genes, our goal is to identify the "hidden message" that DosR employs to regulate the expression of these genes [21].

The principle of "one codon, one amino acid" was established during protein translation in 1961 by Sydney Brenner and Francis Crick [21]. They discovered that the amount of protein produced in a gene was significantly impacted by the removal of one or two nucleotides in succession. The protein was surprisingly little changed when three consecutive nucleotides were removed. Charles Yanofsky discovered in 1964 that a gene's product and its corresponding protein are collinear, i.e., the protein's first amino acid is coded for by the first codon, its second by the second codon, and so on [22]. For the following thirteen years, researchers thought that a protein was encoded as a long string of successive nucleotide triplets. But Phillip Sharp and Richard Roberts' independent discovery of split genes in 1977 demonstrated differently, posing the computational difficulty of locating genes based solely on genomic sequence [23]. Adenovirus single strand DNA is opposed by the viral protein known as hexon, which is encoded by sharp hybridised RNA. He predicted that there would be a one-to-one hybridization of RNA and DNA bases if the hexon gene were continuous. When Sharp looked at the RNA-DNA hybridization under an electron microscope, he was astonished to observe three loop structures instead of the continuous duplex section indicated by the contiguous gene notion (image below) [24]. This finding demonstrated that the four non-contiguous adenovirus genomic segments needed to construct the hexon mRNA. To create a split gene, these four sections – called exons – are divided by three fragments – called introns – that are shown by the loops in the illustration below in Figure 10.

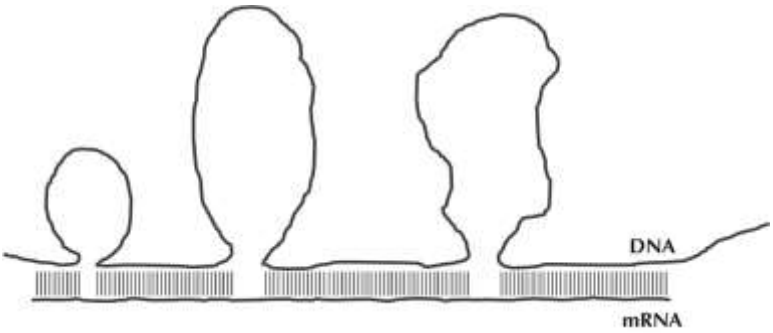

Figure 10. Introns that are shown by the loops.

An intriguing conundrum was raised by the finding of split genes: What is the fate of the introns? Stated differently, RNA produced from a split gene (also called precursor mRNA or pre-mRNA) should

be longer than RNA that acts as a template for protein synthesis (called messenger RNA or mRNA). A few biological processes include cutting the introns from pre-mRNA and concatenating the exons into a single mRNA string [24]. The spliceosome is a molecular device that facilitates the splicing process, which converts pre-mRNA into mRNA. Multiple new fields of study were initiated by the finding of split genes. There is ongoing dispute among biologists over the function of introns; while some are regarded as "junk DNA," others contain essential regulatory elements. Moreover, the way a gene is divided into exons can differ between species. A gene in the chicken genome, for instance, can have fewer exons than a gene of the same type in the human genome.

Certain locations in many biological motifs include two nucleotides that bind to transcription factors roughly equally well. In the yeast S. cerevisiae, for example, the sixteen nucleotides that comprise the CSRE transcription factor binding site are made up of five strongly conserved sites and eleven poorly conserved sites, all of which have the same frequency of two nucleotides [25].

The first thing to keep in mind when creating a motif scoring function is that each column of Profile (Motifs) represents a probability distribution or a set of nonnegative numbers that add up to 1. For example, the probability of 0.2, 0.6, 0.0, and 0.2 for A, C, G, and T, respectively, is represented by the second column in the image from the first step.

$$H(p1, \ldots, pN) = -\sum_{i=1}^{N} p_i . log_2 p_i$$

There are two ways of searching a motif either by greedy technique or randomized technique.

CONCLUSIONS

This study showcases the effectiveness of bioinformatics tools based on Python for identifying the Origin of Replication (ORI) site in the genomes of *V. cholerae*, *T. Petrophila*, and *E. coli*. By using a combination of computational techniques and data analysis, the ORI site was successfully located, which provided valuable insights into the replication initiation process. The methodologies utilized in this research pave the way for further exploration and understanding of genome replication in *V. cholerae*, *T. Petrophila*, and *E. coli*. This study underscores the importance of bioinformatics in unraveling complex biological processes and highlights the potential of Python as a powerful tool in genomic research.

1. *V. cholerae* sequence: https://bioinformaticsalgorithms.com/data/realdatasets/Replication/Vibrio_cholerae.txt
2. *T. petrophila* sequence:http://bioinformaticsalgorithms.com/data/realdatasets/Replication/t_petrophila_oriC.txt
3. *E. coli* sequence: https://bioinformaticsalgorithms.com/data/realdatasets/Replication/E_coli.txt

REFERENCES

1. Jacob F, Brenner S, Cuzin F. On the regulation of DNA replication in bacteria. Cold Spring Harbor Symposia on Quantitative Biology. 1963;28:329–48. Available from: https://doi.org/10.1101/sqb.1963.028.01.048.
2. Messer W. The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication. FEMS Microbiology Reviews. 2002;26:355–74. Available from: https://doi.org/10.1111/j.1574-6976.2002.tb00620.x.
3. Harrison PW, Lower RP, Kim NK, Young JPW. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. Trends in Microbiology. 2010;18:141–8. Available from: https://doi.org/10.1016/j.tim.2009.12.010.
4. Gao F. Bacteria may have multiple replication origins. Frontiers in Microbiology. 2015;6:324. Available from: https://doi.org/10.3389/fmicb.2015.00324.
5. Zakrzewska-Czerwińska J, Jakimowicz D, Zawilak-Pawlik A, Messer W. Regulation of the

initiation of chromosomal replication in bacteria. FEMS Microbiology Reviews. 2007;31:378–87. Available from: https://doi.org/10.1111/j.1574-6976.2007.00070.x.
6. Leonard AC, Grimwade JE. The orisome: structure and function. Frontiers in Microbiology. 2015;6:545. Available from: https://doi.org/10.3389/fmicb.2015.00545.
7. Krause M, Rückert B, Lurz R, Messer W. Complexes at the replication origin of Bacillus subtilis with homologous and heterologous DnaA protein. Journal of Molecular Biology. 1997;274:365–80. Available from: https://doi.org/10.1006/jmbi.1997.1404.
8. Brilli M, et al. The diversity and evolution of cell cycle regulation in alpha-proteobacteria: a comparative genomic analysis. BMC Systems Biology. 2010;4:52. Available from: https://doi.org/10.1186/1752-0509-4-52.
9. Jaworski P, et al. Unique and universal features of epsilon proteobacteria origins of chromosome replication and DnaA-DnaA box interactions. Frontiers in Microbiology. 2016;7:1555. Available from: https://doi.org/10.3389/fmicb.2016.01555.
10. Richardson TT, Harran O, Murray H. The bacterial DnaA-trio replication origin element specifies single-stranded DNA initiator binding. Nature. 2016;534:412–6. Available from: https://doi.org/10.1038/na.
11. Ryan VT, Grimwade JE, Camara JE, Crooke E, Leonard AC. Escherichia coli prereplication complex assembly is regulated by the dynamic interplay among Fis, IHF, and DnaA. Molecular Microbiology. 2004;51:1347–59. Available from: https://doi.org/10.1046/j.1365-2958.2003.03906.x.
12. Bramhill D, Kornberg A. Duplex opening by DnaA protein at novel sequences in the initiation of replication at the origin of the E. coli chromosome. Cell. 1988;52:743–55. Available from: https://doi.org/10.1016/0092-8674(88)90412-6.
13. Kowalski D, Eddy MJ. The DNA unwinding element: a novel, cis-acting component that facilitates the opening of the Escherichia coli replication origin. EMBO Journal. 1989;8:4335–44.
14. Marczynski GT, Rolain T, Taylor JA. Redefining bacterial origins of replication as centralized information processors. Frontiers in Microbiology. 2015;6:610. Available from: https://doi.org/10.3389/fmicb.2015.00610.
15. Song C, Zhang S, Huang H. Choosing a suitable method for the identification of replication origins in microbial genomes. Frontiers in Microbiology. 2015;6:1049. Available from: https://doi.org/10.3389/fmicb.2015.01049
16. Song J, Ware A, Liu SL. Wavelet to predict bacterial ori and ter: a tendency towards a physical balance. BMC Genomics. 2003;4:17. Available from: https://doi.org/10.1186/1471-2164-4-17..
17. Gao F, Zhang CT. Ori-finder: A web-based system for finding oriCs in unannotated bacterial genomes. BMC Bioinformatics. 2008;9:79. Available from: https://doi.org/10.1186/1471-2105-9-79.
18. Kundal S, Lohiya R, Shah K. iCorr: Complex correlation method to detect the origin of replication in prokaryotic and eukaryotic genomes. arXiv. 2016. Available from: https://arxiv.org/abs/1606.02459
19. Maderankova D, Sedlar K, Vitek M, Skutkova H. The identification of replication origin in bacterial genomes by cumulated phase signal. In: 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). 2017. Available from: https://doi.org/10.1109/cibcb.2017.8058561.
20. Zhang G, Gao F. Quantitative analysis of the correlation between AT and GC biases among bacterial genomes. PLOS ONE. 2017;12:e0171408. Available from: https://doi.org/10.1371/journal.pone.0171408.
21. Lobry JA. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie. 1996;78:323–6. Available from: https://doi.org/10.1016/0300-9084(96)84764-x
22. Lobry JA. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie. 1996;78:323–6. Available from: https://doi.org/10.1016/0300-9084(96)84764-x

23. Luo H, Zhang CT, Gao F. Ori-under 2, an integrated tool to predict replication origins in the archaeal genomes. Frontiers in Microbiology. 2014;5:482. Available from: https://doi.org/10.3389/fmicb.2014.00482
24. Gao F, Zhang CT. DoriC: a database of oriC regions in bacterial genomes. Bioinformatics. 2007;23:1866–7. Available from: https://doi.org/10.1093/bioinformatics/btm255
25. Gao F, Luo H, Zhang CT. DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. Nucleic Acids Research. 2012;41:D90–D93. Available from: https://doi.org/10.1093/nar/gks990

