*Biological Forum – An International Journal* **16(11): 142-152(2024)**

**ISSN No. (Print): 0975-1130**
**ISSN No. (Online): 2249-3239**

# Gene-wise Analysis of Viral Genomes: Trends and Variations in GC Content with a Novel Graphical Tool for Gene-specific Calculation

***Sumeet Kumar Parida[1,2], Samarth Godara[2*], Shbana Begam[3], Shruti Godara[4] and Shambhavi Yadav[4]***
*[1]Centre for Post-Graduate Studies, OUAT, Bhubaneswar (Odisha) India.*
*[2]ICAR-Indian Agricultural Statistics Research Institute (New Delhi), India.*
*[3]ICAR-National Institute for Plant Biotechnology (New Delhi), India.*
*[4]ICFRE-Forest Research Institute, Dehradun (Uttarakhand) India.*

*(Corresponding author: Samarth Godara*)*

*(Received: 23 August 2024; Revised: 23 September 2024; Accepted: 21 October 2024; Published: 14 November 2024)*
*(Published by Research Trend)*

**ABSTRACT: Despite the known significance of GC content in shaping viral genome stability, evolution, and host adaptation, insightful analysis of GC content across viruses' genes remains underexplored. In this direction, the existing studies primarily focus on isolated viral families or specific host-virus interactions, leaving a gap in understanding the broader patterns and implications of GC content variability across viral genes. To address this gap, the present study introduces GCVirolens, a GUI-based software designed to analyze GC content variations across the genes of virome. It is a standalone, Windows-based bioinformatics tool developed in Python. The study systematically compares gene-specific GC content across various viruses. Twenty viral genomes were selected for analysis, with ten from plant species and ten from animal species. By evaluating the gene-specific GC content of these viral genomes, the research provides novel insights into factors that may influence viral adaptation to different hosts and environmental conditions. The results showed that the genes of animal viruses exhibited a wider range of GC content. In contrast, the genes of plant viruses displayed more consistent GC content. The broader GC content range observed in animal viruses may reflect their need to adapt to more complex and varied host immune systems. In contrast, the narrower GC content range in plant viruses might be tied to plant cells' relatively uniform metabolic conditions. The results contribute to a deeper understanding of virus genome dynamics and provide a foundation for future research on viral adaptation mechanisms, which could inform the development of targeted antiviral strategies.**

**Keywords:** GC content, Gene-wise analysis, Host adaptation, Virus genome, Viral adaptation

**Abbreviations:**
GC-Guanine and Cytosine
GFF-Generic Feature Format
GUI-Graphical User Interface
TMV- Tobacco Mosaic Virus
ZYMV-Zucchini Yellow Mosaic Virus
BMV- Brome Mosaic Virus
CaMV- Cauliflower Mosaic Virus
CMV- Cucumber Mosaic Virus
PPV- Plum Pox Virus
PV- Potato Virus
TSWV- Tomato Spotted Wilt Virus
TYLCV- Tomato Yellow Leaf Curl Virus.

## INTRODUCTION

The global scientific community is increasingly focused on viral genomics due to recent pandemics and outbreaks, highlighting the need for advanced tools to understand virus behaviour at a molecular level. The analysis of genomic sequences, particularly in virome, is crucial for understanding the biology, evolution, and pathogenicity of viruses (Moniruzzaman *et al.*, 2020). Key aspects that researchers typically focus on when studying a virus genome include genome structure and organization, gene content and functions, replication strategy, mutation rates and genetic diversity, adaptation to hosts, GC content and nucleotide composition, etc. In these viral genome studies, GC content analysis has not been explored in a comprehensive manner, despite its importance in understanding various biological factors (Wang *et al.*, 2023). As we know, GC content plays a crucial role in genome stability, gene expression regulation, mutation rates, and evolutionary adaptation. Studying GC content can provide insights into how viruses evolve, adapt to different hosts, and respond to environmental pressures, making it an essential aspect of genomic analysis. Therefore, a broader investigation into GC content variations is necessary to grasp its impact on viral behaviour and pathogenicity fully.

In this direction, the available tools like EMBOSS GeeCee (Rice *et al.*, 2000), and GC-Profile (Gao and Zhang 2006) for calculating GC content fail to provide gene-specific insights, especially in viral genomes where rapid mutations and diverse genetic structures demand more granular analysis. As viral genomes continue to be sequenced at an unprecedented rate, the need for specialized tools to analyze this data becomes increasingly pressing (Rice *et al.*, 2000). Existing tools for GC content analysis are either generalized for complete genomes or are not user-friendly for researchers focusing on viruses. These tools typically do not provide an easy mechanism to integrate annotations, such as those found in GFF files, with sequence data from FASTA files to yield gene-specific GC content (Cock *et al.*, 2009). Nowadays, several bioinformatics tools allow for GC content calculation across genomes. At the same time, these tools are usually designed with a broader scope, lacking the specificity and simplicity required for analyzing viromes at the gene level. Tools like EMBOSS GeeCee or online calculators provide fundamental GC content analysis but cannot seamlessly integrate genomic and annotation data (Bano & Khan 2023).

To address these limitations, we present ‘GCVirolens’, a user-friendly Python-based software tool designed explicitly for gene-wise GC content analysis of viral genomes. GCVirolens bridges the gap between genomic data and functional analysis by enabling the seamless integration of reference virus genome (fasta) and reference annotation Generic Feature Format (GFF) files. This allows researchers to perform detailed, gene-level GC content analysis with minimal effort. Our software adapted a graphical user interface (GUI), making it accessible to researchers and scholars alike. Unlike many tools concentrating on whole-genome GC content, GCVirolens allows users to calculate GC content for each gene, providing more detailed insights into viral genome organization. By supporting GFF file input, the software ensures that users can analyze specific genes based on their annotations, improving the accuracy and relevance of the analysis. GCVirolens offers the following features:

- A user-friendly GUI-based interface that eliminates the need for command-line expertise.
- One-click gene-wise GC content analysis with no requirement for parameter adjustments.
- Free access to a Windows-compatible executable version.
- Full availability of the software's source code, promoting reproducibility and facilitating modifications by other researchers.
- A fully standalone application requiring no pre-installed dependencies for use.

The methodology underlying GCVirolens involves parsing GFF annotation files to extract gene coordinates, which are then used to identify corresponding sequences in the FASTA file. The software calculates the GC content for each gene, using start and end positions defined in the GFF file. This ensures that the analysis is focused and accurate, reflecting the actual genetic structure of the virus.

In the second phase of the study, after developing the GCVirolens software, we used the software on a range of viral genomes (comprising ten animal viruses and the other ten from plant viruses). GCVirolens proved its capability by processing viral genomes of different sizes and complexities, performing the gene-wise analysis efficiently. The data for both plant and animal viruses showed significant variation in GC content across genes, demonstrating the value of gene-specific analysis. Furthermore, the results obtained from the proposed software were compared and evaluated in the article's discussion section.

The GCVirolens software contributes to the field of viral genomics by providing a specialized tool for gene-wise GC content analysis. It also facilitates more detailed analyses of viral genes, which can lead to new insights into viral evolution, function, and pathogenicity. By integrating genomic sequences (FASTA) with functional annotations (GFF), GCVirolens facilitates a deeper understanding of viral genome organization. By making gene-wise GC content analysis more accessible, GCVirolens can accelerate research and support the development of new therapeutic strategies. The implications of this research are broad, impacting fields such as virology, epidemiology, and bioinformatics. The ability to analyze gene-specific GC content in viral genomes will enhance our understanding of viral biology and could inform the design of vaccines, antiviral drugs, and diagnostic tests. GCVirolens can also be applied to a wide range of viral species, making it a valuable tool for both ongoing and future viral outbreaks.

The rest of the article is structured as follows: starting section presents a comprehensive literature review, highlighting previous research and relevant theoretical frameworks. Next, provides a brief description of the methodological approach undertaken by the software “GCVirolens”. Furthermore, offers details about the viruses that were undertaken, an overview of the results obtained, and a discussion of the gene-specific GC content output. Furthermore, a supplementary sheet is included with the article, elucidating the viruses’ diverse families, genome length, host organism, gene number, type, and accession number. At the last part discusses the obtained results and correlates the outputs with the findings of multiple existing studies. The section also sheds light on the implications rendered by the research, its limitations, and future scope followed by, conclusion and future scope.

## LITERATURE REVIEW

The relationship between genomic GC content and mutation rates has been a topic of significant interest in molecular genetics. High-GC regions in eukaryotic genomes, including *Saccharomyces cerevisiae*, often correlate with increased meiotic recombination rates (Eyre-Walker, 1993; Birdsell, 2002). This phenomenon suggests that GC-rich sequences may serve as hotspots for genetic exchange, potentially influencing evolutionary trajectories (Gerton *et al.*, 2000; Petes, 2001).

Additionally, the influence of transcription on mutation rates has been highlighted, with evidence suggesting that active transcription can elevate mutation frequencies, particularly in regions of high GC content (Kim *et al.*, 2007; Jinks-Robertson & Bhagwat 2014). The mechanisms underlying increased mutation rates in GC-rich regions have been explored in various contexts. For example, it has been suggested that the error-prone DNA polymerase ζ plays a critical role in generating mutations in these regions, particularly during DNA replication (Northam *et al.*, 2006).

Furthermore, studies have indicated that DNA polymerase slippage significantly contributes to deletions and duplications in GC-rich sequences (Tran *et al.*, 1995; Kokoska *et al.*, 2000). In addition to mutation rates, the impact of GC content on recombination has been extensively documented. High-GC regions are associated with elevated levels of meiotic and mitotic recombination (Blat & Kleckner 1999; St Charles & Petes 2013). The mechanisms driving this association may involve the preferential formation of double-strand breaks (DSBs) in GC-rich sequences, essential for initiating recombination events (Andersen & Sekelsky 2010; Symington *et al.*, 2014).

Wolfe *et al.* (1989) initially proposed that variations in synonymous substitution rates among genes correlate with GC content, suggesting that both mutation rates and GC content are influenced by the relative concentrations of nucleotide precursors during DNA replication in mammalian germ cells. This hypothesis, termed the mutationist hypothesis, posits that fluctuations in nucleotide precursor concentrations can lead to differences in mutation rates and GC content across genes (Wolfe *et al.*, 1989).

Subsequent studies have reinforced the idea that GC-rich isochores arise due to replication timing differences among DNA segments. Bernardi *et al.* (1985) characterized compositional isochores as long DNA segments with homogeneous GC content, noting that GC-rich isochores tend to replicate early in the cell cycle, while GC-poor isochores replicate later (Holmquist, 1987). However, Eyre-Walker (1992) challenged this view, presenting evidence that both GC-rich and GC-poor isochores can replicate at varying times during the somatic cell cycle, although these findings may not apply to germ cells. Recent models have attempted to describe the relationship between mutation rates and GC content quantitatively. The current study proposes a model that integrates the effects of misincorporation, correction efficiency, and the next-nucleotide effect to explain the inverted-V-shaped distribution of mutation rates relative to GC content. This model suggests that under normal physiological conditions, the equilibrium GC content in a sequence is approximately equal to the GC proportion in the nucleotide precursor pool (Gu and Li 1994).

Genomic GC content and optimal growth temperature (Topt) in prokaryotes have been extensively researched and debated. Previous studies have suggested that GC-rich genomes are more stable at elevated temperatures due to the additional hydrogen bond present in GC pairs compared to AT pairs (Bernardi *et al.*, 1986; Galtier & Lobry 1997). This stability is thought to confer a selective advantage in thermophilic environments, leading to a positive correlation between GC content and growth temperature. Early investigations into this relationship yielded mixed results. For instance, Hurst and Merchant (2001) found a significant positive correlation between Topt and the GC content of structural RNA genes, such as tRNAs and rRNAs, while failing to observe similar correlations for whole-genome GC content across a broad range of prokaryotic species. This discrepancy was attributed to the functional constraints imposed on protein-coding genes, which may obscure the thermal adaptation hypothesis (Hurst & Merchant 2001; Musto *et al.*, 2004).

Musto *et al.* (2004) further examined the correlation within closely related prokaryotic families, revealing a higher incidence of positive correlations than expected by chance. However, subsequent studies highlighted the influence of sample size and the presence of outlier species on the observed correlations (Marashi & Ghalanbor 2004; Basak *et al.*, 2005). These findings underscored the complexity of the relationship and the need for careful consideration of phylogenetic factors in analyses. In recent years, advancements in genomic sequencing and the availability of curated databases have facilitated more comprehensive analyses. Sato *et al.* (2020) compiled a large dataset of prokaryotic growth temperatures, enabling a re-evaluation of the correlation between GC content and Topt. The current study by Hu *et al.* (2022) builds upon this foundation, utilizing a significantly larger dataset comprising 681 bacteria and 155 archaea. Their phylogenetic comparative analyses reveal robust positive correlations between Topt and various measures of GC content in bacteria, while the correlation in archaea remains ambiguous, particularly when halophilic species are excluded. This research not only clarifies previous contradictory observations but also suggests that thermal adaptation may play a role in shaping genomic GC content in prokaryotes.

Russell *et al.* (2023) presents a novel probabilistic model that incorporates various sequence-level features, including local sequence context, length, and GC content, to predict trimming probabilities more accurately. Their findings indicate that GC content is predictive of sequence-breathing dynamics, suggesting that the structural properties of DNA play a significant role in the trimming process. This model not only refines the understanding of Artemis's function but also extends its applicability to other adaptive immune receptor loci, including TCRα, TCRγ, and IGH sequences. The identification of a preferentially trimmed sequence motif, independent of GC content, raises intriguing questions about the specificity of the trimming mechanism. While Artemis is generally regarded as a structure-specific nuclease, the presence of a sequence motif suggests that additional factors may influence trimming outcomes (Ma *et al.*, 2005; Chang and Lieber 2016). Further research is needed to explore the mechanistic basis of this motif and its implications for receptor diversity.

## METHODOLOGY

The GCVirolens software is developed as a simple, user-friendly tool for gene-specific GC content analysis of viral genomes. GC content is an important metric in genomics research, as it affects various biological functions, including genome stability and gene regulation. GCVirolens simplifies this complex task by providing a graphical interface requiring minimal computational expertise. It allows users to input viral genome sequences in FASTA format alongside GFF annotation files and outputs the GC content for each gene. Fig. 1 illustrates the steps undertaken by GCVirolens to extract the required insights. To ensure the reproducibility of this research, the complete code for each of the mentioned modules is available at: *https://github.com/ICAR-BIOINFORMATICS/GCVirolens.git*.
The explanation of each step is as follows:

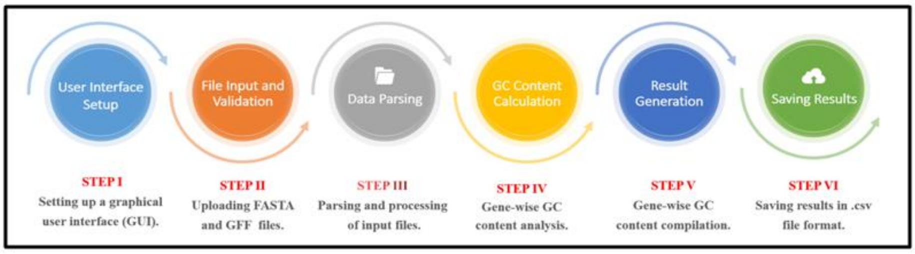

User Interface Setup
File Input and Validation
Data Parsing
GC Content Calculation
Result Generation
Saving Results
STEP I
Setting up a graphical user interface (GUI).
STEP II
Uploading FASTA and GFF files.
STEP III
Parsing and processing of input files.
STEP IV
Gene-wise GC content analysis.
STEP V
Gene-wise GC content compilation.
STEP VI
Saving results in .csv file format.

**Fig. 1.** Methodology used by GCVirolens software to extract gene-wise GC content.

**1. User Interface Setup:** The first step involves setting up a GUI using the ‘Tkinter’ package. The interface includes buttons for uploading files, executing the analysis, and saving results, making it user-friendly. Users are guided through the process with labels and status updates on the UI.

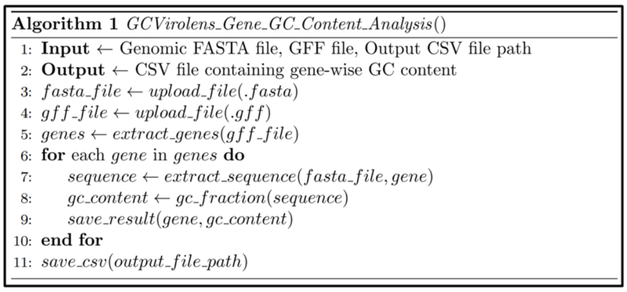
**2. File Input and Validation:** In this step, users upload the required FASTA (genome sequence) and GFF (gene annotation) files using the ‘filedialog’ module. The software checks if both files are properly uploaded, and error handling is employed to ensure that the analysis does not proceed without both files.

**3. Data Parsing:** Once the files are uploaded, the FASTA file is parsed using Biopython’s ‘SeqIO’ module, while the GFF file is processed to extract gene coordinates. The software identifies the gene locations in the genome, preparing the data for GC content analysis.

**4. GC Content Calculation:** Using the parsed data, the software extracts individual gene sequences based on their start and end positions from the GFF file. It then calculates the GC content for each gene using Biopython’s ‘gc_fraction’ function, providing a gene-wise analysis.

**5. Result Generation:** The results of the GC content analysis are compiled into a ‘DataFrame’ that includes gene IDs, sequence positions, and calculated GC content. This table is displayed within the software and prepared for saving.

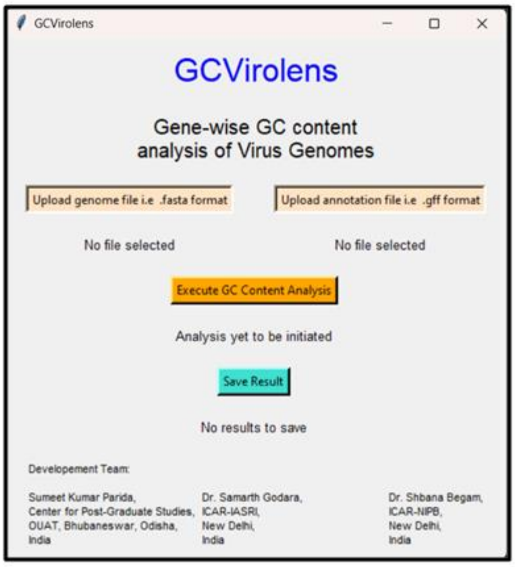
**6. Saving Results:** The final step allows users to save the generated results in CSV format using the ‘asksaveasfilename’ method. The file is stored locally, containing the full gene-wise GC content analysis, which can be used for further research.



```
Algorithm 1 $GCVirolens\_Gene\_GC\_Content\_Analysis()$
Input ← Genomic FASTA file, GFF file, Output CSV file path
Output ← CSV file containing gene-wise GC content
$fasta\_file \leftarrow upload\_file(.fasta)$
$gff\_file \leftarrow upload\_file(.gff)$
$genes \leftarrow extract\_genes(gff\_file)$
for each $gene$ in $genes$ do
    $sequence \leftarrow extract\_sequence(fasta\_file, gene)$
    $gc\_content \leftarrow gc\_fraction(sequence)$
    $save\_result(gene, gc\_content)$
end for
$save\_csv(output\_file\_path)$
```

**Algorithm 1**: Pseudocode of the primary modules of GCVirolens software.

For a broad understanding of the working of the primary modules of the software, Algorithm 1 gives the pseudocode of the software. The algorithm takes a genomic FASTA file (with DNA sequences), a GFF file (with gene annotations), and an output CSV file path as inputs. The function 'upload_file(.fasta)' and 'upload_file(.gff)' allow users to upload the respective files via a Tkinter GUI interface. The function 'extract_genes (gff_file)' parses the GFF file to extract gene regions, such as their start and end positions on the genome. For each gene, 'extract_sequence(fasta_file, gene)' fetches the corresponding DNA sequence from the uploaded FASTA file. The function 'gc_fraction(sequence)' computes the GC content (the proportion of guanine (G) and cytosine (C) bases) for the extracted gene sequence. Later, the GC content values are saved using 'save_result(gene, gc_content)' for each gene. Finally, the data is stored in a CSV file with 'save_csv(output_file_path)'. This workflow ensures that the gene sequences are extracted from the genome and their GC content is calculated, allowing researchers to analyze the GC content distribution in virus genomes.



GCVirolens

GCVirolens

Gene-wise GC content
analysis of Virus Genomes

Upload genome file i.e .fasta format
Upload annotation file i.e .gff format

No file selected

No file selected

Execute GC Content Analysis

Analysis yet to be initiated

No results to save

Developement Team:

Sumeet Kumar Parida,
Center for Post-Graduate Studies,
OUAT, Bhubaneswar, Odisha,
India

Dr. Samarth Godara,
ICAR-IASRI,
New Delhi,
India

Dr. Shbana Begam,
ICAR-NIPB,
New Delhi,
India

**Fig. 2.** GCVirolens Graphics User Interface.

The user interface of the GCVirolens software is designed for ease of use, as illustrated in Fig. 4. Users can upload genomic files in FASTA format and annotation files in GFF format for analysis. Once the files are selected, the software performs GC content analysis, providing real-time status updates. Upon completion, users can save the results in CSV format.

## EXPERIMENTS AND RESULTS

The experiments conducted in the study were run on a system with an Intel Core i5 processor, 8 GB RAM, and a 512 GB SSD, using Windows 11 OS. Python 3.10 was employed, with the following libraries: 'Tkinter' for the GUI, 'Biopython' for sequence parsing and GC content calculations, and 'pandas' for data management. We used GCVirolens to analyse the gene-wise GC content of ten plant and ten animal viral genomes, using corresponding FASTA and GFF files. Table 1, in the supplementary sheet, provides a comprehensive overview of various viruses undertaken in the study, detailing their host organisms, genome types, lengths, gene counts, and accession numbers. It includes plant viruses like Tobacco mosaic virus (TMV) and Cucumber mosaic virus, which infect species such as tobacco and cucumber, as well as animal viruses like Rabies lyssavirus and Nipah virus, affecting mammals like dogs, bats, and humans. Viral genomes range from single-stranded RNA (ssRNA) to double-stranded DNA (dsDNA), and the number of genes varies from 2 to 7, depending on the virus type.

Fig. 3 illustrates the genome length (in kbp) and the number of genes of the undertaken plant viruses. The green bars represent genome length, while the red bars depict the number of genes. The graph highlights that the Tomato spotted wilt virus has the longest genome at 16.6 kbp but only 5 genes. In contrast, the Cauliflower Mosaic Virus has 7 genes despite a genome of only 8 kbp. The Tobacco mosaic virus, and Tomato yellow leaf curl virus both have 6 genes, but their genome lengths vary from 2.8 to 6.4 kbp.

Fig. 4 compares the undertaken animal viruses based on their genome length (in kilobase pairs, kbp) and the number of genes they possess. Among the considered viruses, Covid has the longest genome length at 29.9 kbp and the highest gene count of 11. Mayaro and Rabies Lyssavirus have the same genome length around 12 kbp(11.4 and 11.9) and 4 and 5 genes, respectively. In contrast, the Lujo virus has the smallest genome (10.4 kbp) and only 4 genes. Other viruses like Hendra, Langya, and Nipah exhibit genome lengths around 18 kbp with 6 genes each.

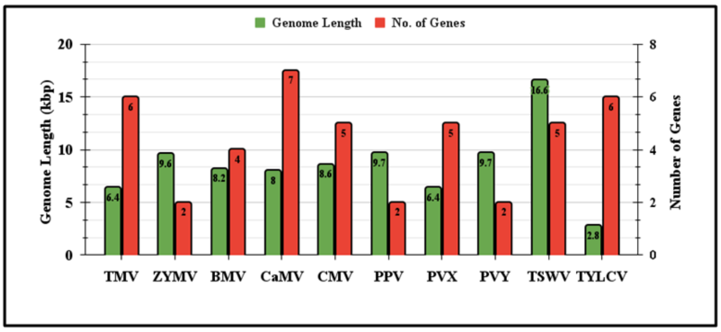


**Fig. 3**. Comparison of the undertaken Plant Viruses in terms of their Genome Length and Number of Genes: TMV, ZYMV, BMV, CaMV, CMV, PPV, PVX, PVY, TSWV, TYLCV.

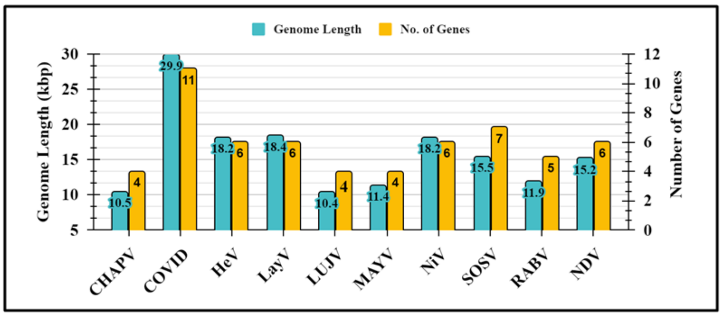
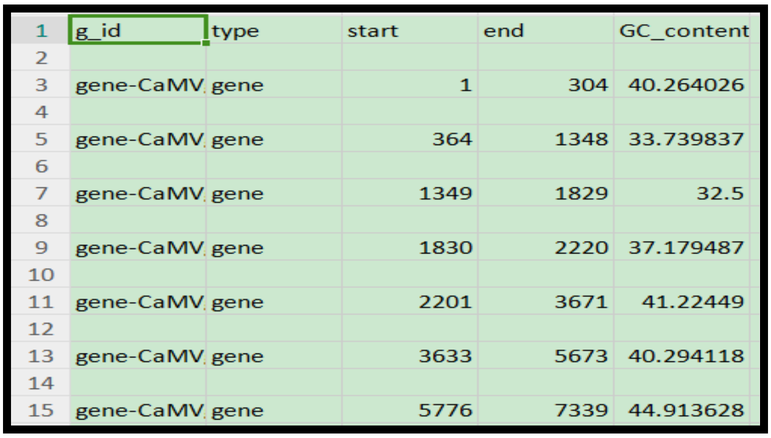


**Fig. 4.** Comparison of the considered Animal Viruses in terms of their Genome Length and Number of Genes: CHAPV, COVID, HeV, LayV, LUJV, MAYV, NiV, SOSV, RABV, NDV.

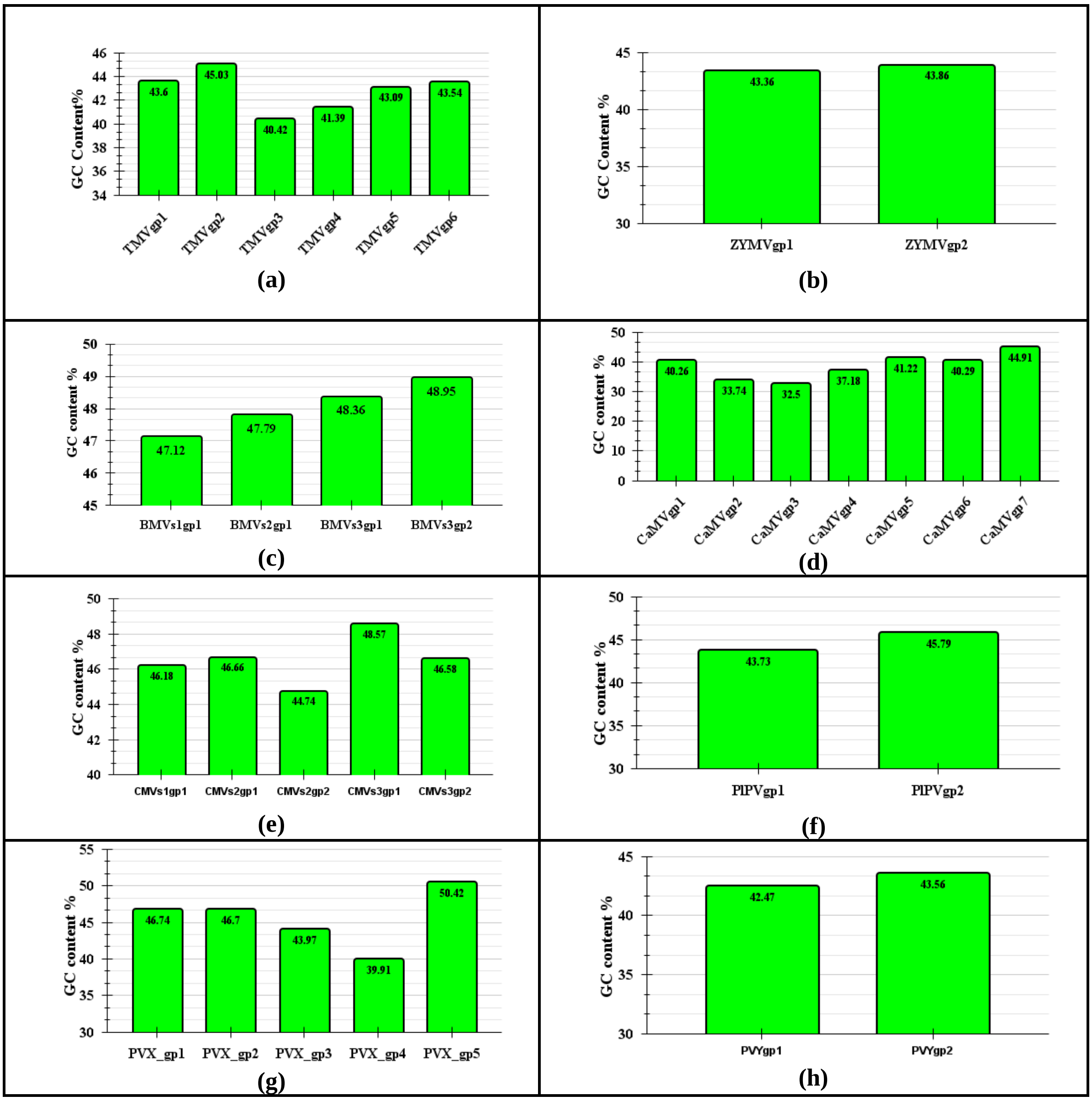
Sosuga and Newcastle viruses have similar genome sizes (15.5 kbp and 15.2 kbp, respectively) with 7 and 6 genes.

The analysis performed using the proposed software on 20 different viruses(both plant and animal) yielded a minimum gene GC content of 32.56% and a maximum of 54.13%. This gene-level breakdown is critical as GC content influences viral genomes' stability and gene expression. Across other viruses, similar patterns are expected where the GC content varies by gene, reflecting the virus's adaptation strategies in different host environments. An exemplar gene-wise GC content output of the Cauliflower Mosaic Virus genome is shown in Table 2.

**Table 2: Sample Output of the Software Showing Gene-wise GC Content of Cauliflower Mosaic Virus.**



| 1 | g_id | type | start | end | GC_content |
|---|---|---|---|---|---|
| 2 | | | | | |
| 3 | gene-CaMV | gene | 1 | 304 | 40.264026 |
| 4 | | | | | |
| 5 | gene-CaMV | gene | 364 | 1348 | 33.739837 |
| 6 | | | | | |
| 7 | gene-CaMV | gene | 1349 | 1829 | 32.5 |
| 8 | | | | | |
| 9 | gene-CaMV | gene | 1830 | 2220 | 37.179487 |
| 10 | | | | | |
| 11 | gene-CaMV | gene | 2201 | 3671 | 41.22449 |
| 12 | | | | | |
| 13 | gene-CaMV | gene | 3633 | 5673 | 40.294118 |
| 14 | | | | | |
| 15 | gene-CaMV | gene | 5776 | 7339 | 44.913628 |



**(a)**

**(b)**

**(c)**

**(d)**

**(e)**

**(f)**

**(g)**

**(h)**

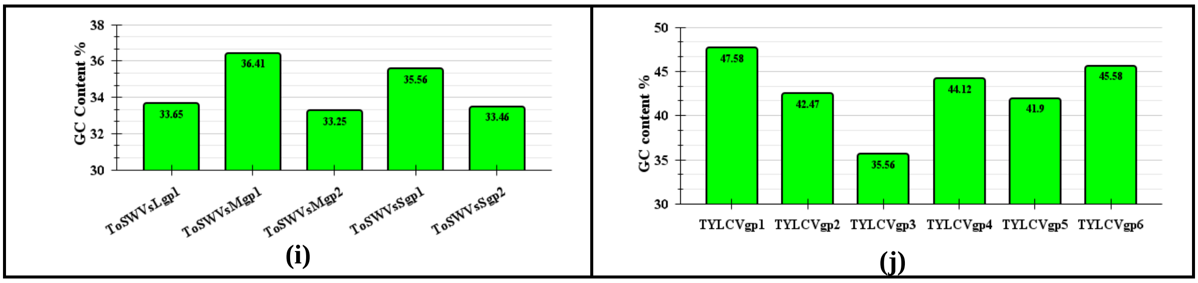


**Fig. 5**. Gene-wise GC Content of Various Plant Viruses [ (a)- Tobacco Mosaic Virus, (b)- Zucchini Yellow Mosaic Virus, (c)- Brome Mosaic Virus, (d)- Cauliflower Mosaic Virus, (e)- Cucumber Mosaic Virus, (f)- Plum Pox Virus, (g)- Potato Virus X, (h)- Potato Virus Y, (i)- Tomato Spotted Wilt Virus, (j)- Tomato Yellow Leaf Curl Virus]

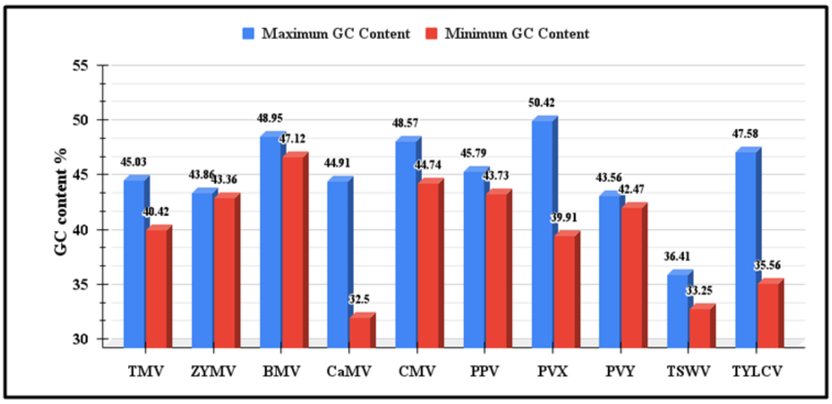
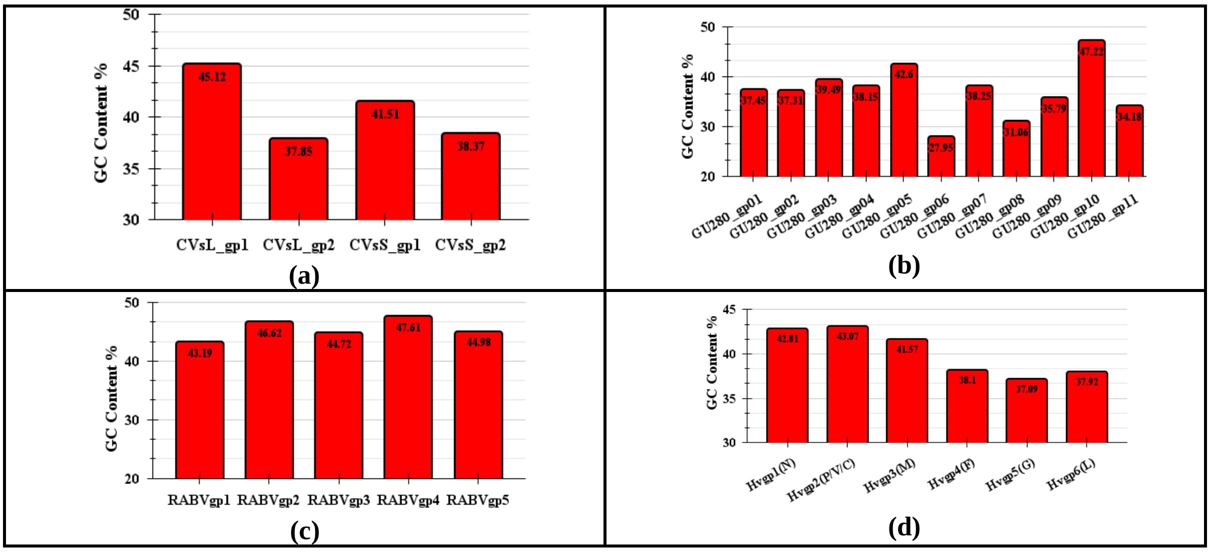


**Fig. 6**. GC Content of Various Plant Viruses and their Maximum & Minimum Value.

Among the viruses, the Tobacco Mosaic Virus shows moderate GC content variation, with a maximum of 45.03% and a minimum of 40.42%. The Zucchini yellow mosaic virus displays a meager range, with a high of 43.86% and a low of 43.36%, suggesting stability. Brome Mosaic Virus has relatively consistent GC content, ranging from 48.95% to 47.12%. In contrast, Cauliflower Mosaic Virus shows a wide variation from 44.91% to 32.5%, indicating potential structural differences within its genome. The Cucumber Mosaic Virus exhibits moderate variation with GC content ranging from 48.57% to 44.74%. Plum Pox Virus slightly ranges between 45.79% and 43.73%. Potato Virus X has the highest GC content at 50.42% and a significant drop to 39.91%, implying notable genomic variability (Fig. 5). Potato Virus Y shows minimal GC variation, ranging from 43.56% to 42.47%, while Tomato Spotted Wilt Virus has the lowest GC content overall, with a maximum of 36.41% and a minimum of 33.25%. Lastly, the Tomato Yellow Leaf Curl Virus exhibits a broad range, from 47.58% to 35.56%. Overall, the data indicate that some viruses, such as Cauliflower Mosaic Virus and Tomato Yellow Leaf Curl Virus, have substantial GC content variability, which may affect genome stability. In contrast, others like Brome Mosaic Virus and Potato Virus Y display more uniform GC content. While some viruses exhibit stable GC content across their genes, others display significant variability, indicating the potential for differences in their replication strategies, genomic evolution, and interaction with their plant hosts (Fig. 6).



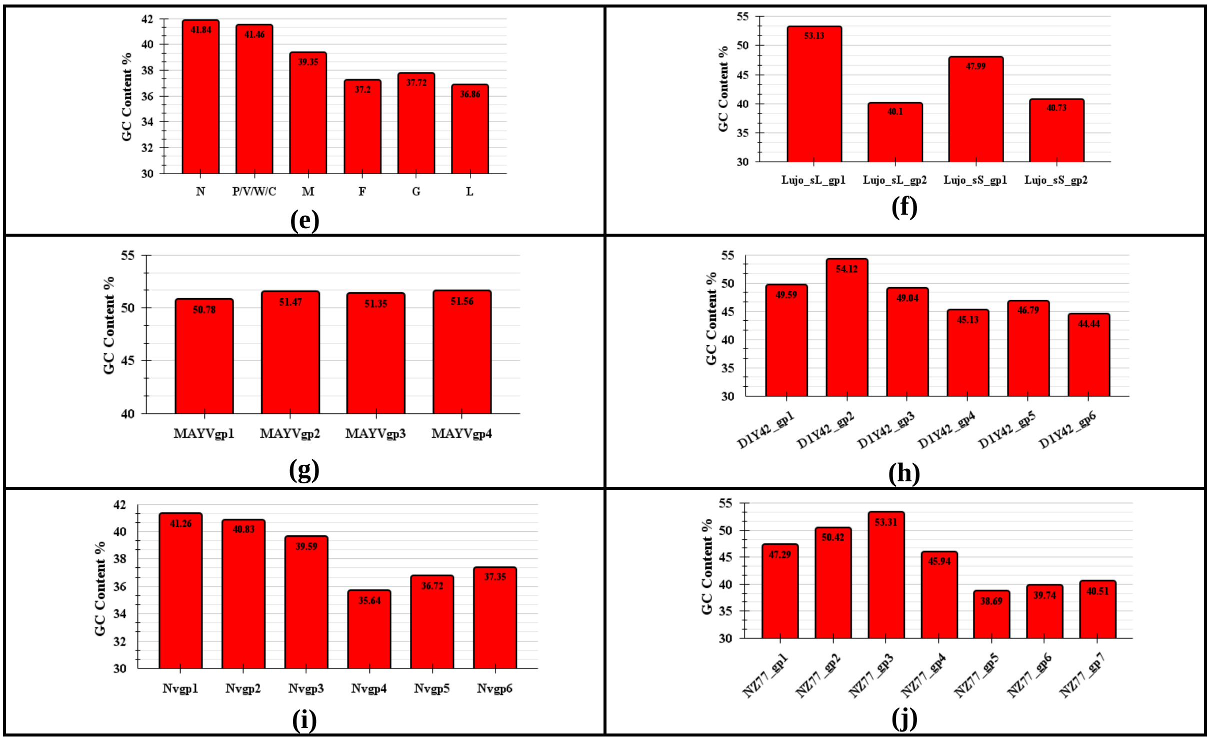


**Fig. 7.** Gene-wise GC Content of Various Animal Viruses [ (a)- Chapare Virus, (b)-severe acute respiratory syndrome coronavirus 2, (c) - Rabies Lyssavirus, (d)- Hendra Virus, (e)- Langya Virus, (f)- Lujo Virus, (g)- Mayaro Virus, (h)- Newcastle Disease Virus, (i)- Nipah Virus, (j)-Sosuga Virus]

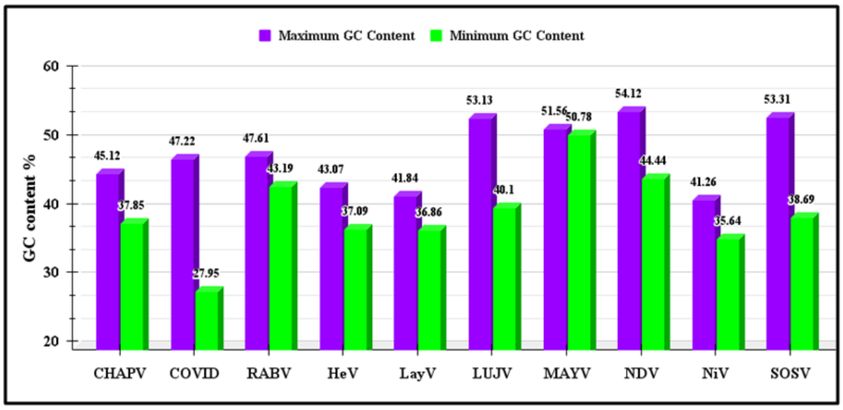


**Fig. 8**. GC Content of Various Animal Viruses and their Maximum & Minimum Value.

The Chapare Virus shows noticeable deviations, with GC content ranging between 37.85% and 45.12%. This variation may reflect differences in the gene's functions or structural stability within the virus. For the COVID-19 Virus, the graphs show a broader range of GC content, from 27.95% to 47.22%. This suggests a slightly more varied nucleotide composition across its genome, possibly due to different structural or functional roles of the genes within this virus. The Rabies Lyssavirus shows relatively stable GC content values across its genes, ranging from 43.19% to 47.61%. The narrow range indicates consistency in the nucleotide composition, which could correlate with the virus's genetic stability or conservation across its genes (Fig. 7).

In contrast, the Hendra virus exhibits moderate variation, with GC content ranging between 37.09% and 43.07% across different genes. This suggests that while there is some fluctuation in its genetic composition, the GC content remains within a specific range, possibly reflecting its adaptation to host environments. The Langya virus has a more consistent GC content, hovering around 36.86%-41.84%. This narrow range indicates little variation between genes, suggesting a highly stable and conserved genetic structure across the virus’s genome. For the Lujo virus, the GC content fluctuates more widely, from 40.10% to 53.13%. This higher GC content could point to a more stable RNA or DNA structure, possibly aiding in its resistance to mutation or degradation. The Mayaro virus has consistent GC content values close to 50-51% across its genes. The Newcastle disease virus shows greater diversity in its GC content, ranging from 44.44% to 54.12%. This variation indicates that different genes within the virus may have differing structural or functional properties that require varied nucleotide compositions. The GC content spans from 35.64% to 41.26% for the Nipah virus, reflecting moderate variability across genes. Lastly, the Sosuga virus displays a wide GC content range, from 38.69% to 53.31%. This suggests a higher degree of variability between its genes (Fig. 8).

## DISCUSSION

The present study focuses on developing and validating GCVirolens, a Python-based software for gene-wise GC content analysis of viral genomes. By integrating FASTA and GFF files, GCVirolens allows researchers to assess GC content at a more granular, gene-specific level, which is particularly crucial for analyzing rapidly mutating viral genomes. The tool was tested on 20 viral genomes (ten plant and ten animal viruses), demonstrating its capability to analyze diverse viral types. Results showed gene-specific GC content variations that provide insights into viral evolution, stability, and adaptation strategies in different hosts.

Interestingly, many of the obtained results correlate with the findings of multiple existing studies. The GC content findings reported by the proposed software for the Newcastle disease virus showed significant consistency with the findings of Qiu *et al.* (2011). According to their data, the GC content findings of 25 strains of NDV representing different genotypes (avg. of each genotype) for the genes named NP, P, M, F, HN, and L with their standard deviations being mentioned in a tabular form, which corresponds to most of the software findings (49.59, 54.12, 49.04, 45.12, 46.79, 44.44). Similarly Biswas *et al.* (2021) cited the works mentioning coronavirus, the largest genome among all known RNA viruses, with G + C contents varying from 32% to 43%, which goes to consolidate the GCVirolens software GC content findings, further confirmed by (Li *et al.*, 2020; Berkhout and van Hemert 2015). The complete genome of GD-SH-01 (a virulent wild Rabies Virus strain) is 11,923 nt in length, with a GC content of 45.42% (Luo *et al.*, 2012), similar to the output given by the proposed software. At the same time, Tobacco Mosaic Virus shows GC content values ranging from 40.42 to 45.03 with TMVgp1(43.59%), TMVgp2(45.03%), TMVgp3(40.42%), TMVgp4(41.38%), TMVgp5(43.08%), TMVgp6(43.54%) respectively. Comparing the software findings with that of the table provided in a study led by a group of scientists, namely Cheeran *et al.* (2023), further corroborates the software's accuracy. The table comprises of six genes of TMV and their nucleotide composition, mean GC content with standard deviation, etc. The gene-wise GC content analysis of viral genomes, as provided by our study, reveals significant insights into virus genome stability. This increased stability can protect the viral genome from environmental stress and contribute to its longevity within host cells. Higher GC content in certain areas can impact mutation rates, as transitions between GC pairs are less frequent, allowing viruses to maintain essential genes with fewer errors during replication. Thus, GC content can act as a regulator for mutation hotspots. By studying the GC content (the amount of guanine and cytosine in their genetic code), we can predict how stable a virus is and how likely it is to mutate. These insights can help design better antiviral strategies or breed virus-resistant plants. The developed tool, GCVirolens, can help scientists quickly find parts of the virus that are more likely to change or resist treatment, aiding in creating better solutions to stop the virus from spreading. The study focuses solely on GC content, offering valuable insights but not a comprehensive view of viral genomic complexity, particularly concerning protein functions and host-pathogen interactions. It overlooks the influence of epigenetic factors and relies heavily on high-quality FASTA and GFF files, which may be scarce for lesser-known viruses. Additionally, the analysis excludes RNA secondary structures and protein-coding efficiencies.

Future studies can expand the dataset to include a broader range of viruses to enhance statistical faculty and explore GC content variation across more viral families. Incorporating additional genomic parameters like codon usage, RNA folding patterns, and epigenetic modifications will provide a more comprehensive view of viral evolution and adaptation. Future research can further refine the tool’s capabilities, applying it to a wider range of viral species to explore the relationships between GC content and viral fitness and its role in host-virus interactions. Enhancing GCVirolens to include real-time viral mutation tracking could support public health measures during outbreaks. Finally, integrating machine learning algorithms to predict viral behaviour based on GC content could offer predictive insights for epidemiology and virology.

## CONCLUSION

Viral genome stability, gene expression, and host adaptability are influenced by GC content. Nonetheless, the existing tools need to be more precise for detailed gene-specific analysis across virus types. In this scenario, the proposed software, GCVirolens, offers user-friendly gene-specific GC content analysis for viral genomes. Its unique design focuses on systematic gene-wise analysis, a feature absents in current tools. Additionally, in the second phase of the study, the gene-wise GC content of 20 viral genomes, comprising 10 plant viruses and the other 10 representing animal viruses, was extracted and analyzed. Results from these experiments revealed significant variations in GC content across different viral species and between individual genes within the same genome. The proposed work's implications are substantial in virology. It enables researchers to understand gene-specific GC content, enhancing insights into viral evolution, genome stability, and host adaptation. In the future, authors intend to expand the study to analyze viral cross-species transmission and mutation trends. Comparing GC content across strains or species, and using machine learning to predict viral behaviour, would enhance its impact. In summary, GC Virolens offers a new, precise, and user-friendly method for analyzing viral genomes, filling a crucial gap in virology. It presents a foundation for advanced computational approaches in viral genome analysis, paving the way for future progress in virology research.

## FUTURE SCOPE

GCVirolens can be extended to analyze GC content across a wider range of organisms, including bacteria, plants, and higher eukaryotes, enabling comparative studies across diverse taxa. Future versions could include a web-based or cloud-hosted interface, allowing for broader accessibility and integration with online genomic databases.

**Acknowledgement**. The authors gratefully acknowledge the guidance, support and facilities provided by ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI) and Odisha University of Agriculture & Technology (OUAT), Odisha, which were instrumental in the development of the GCVirolens tool

**Conflicts of Interest**. None.

## REFERENCES

Andersen, S. L. & Sekelsky, J. (2010). Meiotic versus mitotic recombination: Two different routes for double-strand break repair: The different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes. *Bioessays*, *32*(12), 1058-1066.

Bano, H. & Khan, J. A. (2023). Identification of viruses naturally infecting patchouli (*Pogostemon cablin* (Blanco) Benth.) in India. *Biological Forum-An International Journal, 15*(6), 549-558).

Basak, S., Mandal, S. & Ghosh, T. C. (2005). Correlations between genomic GC levels and optimal growth temperatures: some comments. *Biochemical and biophysical research communications*, *327*(4), 969-970.

Berkhout, B. & van Hemert, F. (2015). On the biased nucleotide composition of the human coronavirus RNA genome. *Virus research, 202*, 41-47.

Bernardi, G., Mouchiroud, D., Gautier, C. & Bernardi, G. (1989). Compositional Patterns in Vertebrate Genomes: Conservation and Change in Evolution. *Evolutionary Tinkering in Gene Expression*, 133-142.

Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular biology and evolution*, *19*(7), 1181-1197.

Biswas, B., Nandi, R., Char, P., Bose, S. & Stergioulas, N. (2021). GW190814: on the properties of the secondary component of the binary. *Monthly Notices of the Royal Astronomical Society*, *505*(2), 1600-1606.

Blat, Y. & Kleckner, N. (1999). Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell*, *98*(2), 249-259.

Chang, H. H. & Lieber, M. R. (2016). Structure-Specific nuclease activities of Artemis and the Artemis: DNA-PKcs complex. *Nucleic acids research*, *44*(11), 4991-4997.

Cheeran, K., Suresh, K. P., Jacob, S. S., Gowda, C. S. S. & Gejendiran, N. (2023). Analysis of codon usage bias of six genes of replicase/coat protein of tobacco mosaic virus. *Indian J. Agric. Res*, *1*, 7.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A. & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422.

Eyre-Walker, A. (1992). Evidence that both G+ C rich and G+ C poor isochores are replicated early and late in the cell cycle. *Nucleic Acids Research*, *20*(7), 1497-1501.

Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *252*(1335), 237-243.

Galtier, N. & Lobry, J. R. (1997). Relationships between genomic G+ C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of molecular evolution*, *44*, 632-636.

Gao, F. & Zhang, C. T. (2006). GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic acids research*, *34*(suppl_2), W686-W691.

Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O. & Petes, T. D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, *97*(21), 11383-11390.

Gu, X. & Li, W. H. (1994). A model for the correlation of mutation rate with GC content and the origin of GC-rich isochores. *Journal of molecular evolution*, *38*, 468-475.

Holmquist, G. P. (1987). Role of replication time in the control of tissue-specific gene expression. *American journal of human genetics*, *40*(2), 151.

Hurst, L. D. & Merchant, A. R. (2001). High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1466), 493-497.

Jinks-Robertson, S. & Bhagwat, A. S. (2014). Transcription-associated mutagenesis. *Annual review of genetics*, *48*(1), 341-359.

Kim, N., Abdulovic, A. L., Gealy, R., Lippert, M. J. & Jinks-Robertson, S. (2007). Transcription-associated mutagenesis in yeast is directly proportional to the level of gene expression and influenced by the direction of DNA replication. *DNA repair*, *6*(9), 1285-1296.

Kokoska, R. J., Stefanovic, L., DeMai, J., & Petes, T. D. (2000). Increased rates of genomic deletions generated by mutations in the yeast gene encoding DNA polymerase δ or by decreases in the cellular levels of DNA polymerase δ. *Molecular and cellular biology*.

Li, Y., Yang, X., Wang, N., Wang, H., Yin, B., Yang, X. & Jiang, W. (2020). GC usage of SARS-CoV-2 genes might adapt to the environment of human lung expressed genes. *Molecular Genetics and Genomics*, *295*(6), 1537-1546.

Luo, Y., Zhang, Y., Liu, X., Yang, Y., Yang, X., Zhang, D. & Guo, X. (2012). Complete genome sequence of a highly virulent rabies virus isolated from a rabid pig in south China.

Marashi, S. A. & Ghalanbor, Z. (2004). Correlations between genomic GC levels and optimal growth temperatures are not 'robust'. *Biochemical and biophysical research communications*, *325*(2), 381-383.

Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature communications*, *11*(1), 1710.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valín, F. & Bernardi, G. (2004). Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS letters*, *573*(1-3), 73-77.

Northam, M. R., Garg, P., Baitin, D. M., Burgers, P. M. & Shcherbakova, P. V. (2006). A novel function of DNA polymerase ζ regulated by PCNA. *The EMBO journal*, *25*(18), 4316-4325.

Petes, T. D. (2001). Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics*, *2*(5), 360-369.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends in genetics*, *16*(6), 276-277.

Russell, M. L., Simon, N., Bradley, P. & Matsen IV, F. A. (2023). Statistical inference reveals the role of length, GC content, and local sequence in V (D) J nucleotide trimming. *Elife*, *12*, e85145.

Sato, Y., Okano, K., Kimura, H. & Honda, K. (2020). TEMPURA: database of growth temperatures of usual and RAre Prokaryotes. *Microbes and environments*, *35*(3), ME20074.

St. Charles, J. & Petes, T. D. (2013). High-resolution mapping of spontaneous mitotic recombination hotspots on the 1.1 Mb arm of yeast chromosome IV. *PLoS genetics*, *9*(4), e1003434.

Symington, L. S., Rothstein, R. & Lisby, M. (2014). Mechanisms and regulation of mitotic recombination in Saccharomyces cerevisiae. *Genetics*, *198*(3), 795-835.

Tran, H. T., Degtyareva, N. P., Koloteva, N. N., Sugino, A., Masumoto, H., Gordenin, D. A. & Resnick, M. A. (1995). Replication slippage between distant short repeats in Saccharomyces cerevisiae depends on the direction of replication and the RAD50 and RAD52 genes. *Molecular and Cellular Biology*, *15*(10), 5607-5617.

Wang, C., Pan, C., Yong, H., Wang, F., Bo, T., Zhao, Y. & Li, M. (2023). Emerging non-viral vectors for gene delivery. *Journal of Nanobiotechnology*, *21*(1), 272.

Wolfe, K. H., Sharp, P. M. & Li, W. H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, *337*(6204), 283-285.

**How to cite this article:** Sumeet Kumar Parida, Samarth Godara, Shbana Begam, Shruti Godara and Shambhavi Yadav (2024). Gene-wise Analysis of Viral Genomes: Trends and Variations in GC Content with a Novel Graphical Tool for Gene-specific Calculation. *Biological Forum – An International Journal, 16*(11): 142-152.