

 OPEN ACCESS

Received: 29/08/2024

Accepted: 28/04/2025

Published: 16/06/2025

Citation: Daggubati SP, Kasukurthi VR, Reddy P (2025) Advancing Genomic Data Management through TCompress: An Innovative Method for Sequence Compression of Biological Sequence Compression Algorithm using Text-Compression Tools. Indian Journal of Science and Technology 18(23): 1811-1817. <https://doi.org/10.17485/IJST/v18i23.2790>

* **Corresponding author.**daggubatisivaphanindra@gmail.com**Funding:** None**Competing Interests:** None

Copyright: © 2025 Daggubati et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](#))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Advancing Genomic Data Management through TCompress: An Innovative Method for Sequence Compression of Biological Sequence Compression Algorithm using Text-Compression Tools

Siva Phanindra Daggubati^{1*}, Venkata Rao Kasukurthi¹, Prasad Reddy¹¹ Department of CS & SE, AUCE(A), Andhra University, Visakhapatnam, Andhra Pradesh, India

Abstract

Objectives: With the rapid growth of genetic data, there is an urgent need for efficient methods to compress biological sequences. This study aims to enhance the storage and retrieval of large volumes of genetic information by introducing **TCompress**, a new compression algorithm specifically designed for DNA sequences. **Method:** We developed a custom software tool called TCompress that converts biological sequences into binary format. This binary representation enables the use of standard text compression tools such as GenCompress, 7-Zip, GZip, and PeaZip. A detailed evaluation was conducted to measure TCompress's compression ratio and efficiency in comparison to other commonly used methods. **Findings:** Our results show that TCompress improves compression efficiency by up to 30% over traditional techniques. It performs especially well on longer and more complex DNA sequences, offering better precision and preserving the integrity of genomic data more effectively than existing tools. **Novelty:** Unlike conventional methods, TCompress uses a novel binary conversion technique specifically designed to reflect the unique structure of DNA. This customized approach not only improves compression rates but also supports secure data handling. Beyond compression, TCompress also contributes to better and safer management of genetic information.

Keywords: TCompression tools; Biological Sequence Compression; DNA/RNA Sequence Compression; Compression efficiency; Compression ratio

1 Introduction

Genetic studies have expanded their impact in recent years to encompass a variety of industries. Genomic research has greatly benefited other fields in addition to healthcare. It has opened up new diagnostic techniques and therapeutic treatment methods⁽¹⁾. Genetics improvement, which is responsible for Constant new ways for individual treatment programs and illness prevention, has shown the vast potential of modern biology for transforming health care and patient outcomes⁽²⁾. Government and the

public sector in general have contributed large amounts of funding for genetic studies. As a result of sequencing and genomic exploration, the vast majority of genetic research data has grown exponentially since the turn of the century. The process itself is not rocket science at all, but the region now requires strong protection and data compression technologies for managing such large amounts of data. Conventional approaches make efficient compression difficult for genetically selected data. There are four nucleotides making up the DNA sequence: thymine (T), adenine (A), cytosine (C), and guanine (G). Traditional text compression methods have difficulties because DNA sequences can contain high levels of redundancy and many repetitive patterns⁽³⁾. In order to solve these problems, some laboratories are trying data compression via complex algorithms tailored for biological data. The suggested approach is to hide the data encryption key to encode genetic data before transmission so that it remains safe throughout its journey. The goal of the presented work is to employ a novel compression technology called, TCompress on both storing and retrieving biological sequences as compressed. The TCompress algorithm uses a more long-winded approach than this classic compress text with UID and turns the DNA sequences into binary. In order to pass genetic information safely, the TCompress algorithm considers this natural complementarity between DNA base pairs to enhance its compression ratios. The dynamic domain of genomics gradually started demanding quicker data processing, as genome data in bulk size needed to be stored and the handling gets more complex due to much larger DNA strands. The present study has shown the possibilities of how bandwidth consumption can be reduced — securely and efficiently — with TCompress; and therefore, exemplifies the revolutionary impact that TCompress is poised to make in the world of genetic research data management⁽³⁾.

DNA sequence compression studies have expanded their impact currently encompass various industries. Mathur, Pandey, and Goyal⁽⁴⁾ have created an ASCII-based method that changes Lempel-Ziv-Welch (LZW) compression by turning DNA sequences into ASCII values and using the LZW warehouse floor. Although their methodology uses blockchain technology to improve data security, they are unable to cope with the huge amounts of data currently found in modern genomic databases. Wang et al. (2023)⁽⁵⁾ developed dynamic DNA QR coding (DDQR) as another method for compressing DNA barcodes. This approach, despite its effectiveness with smaller sequences, because it does not scale well.

To facilitate critical review in a therapeutic setting, Sthlbom et al. (2024)⁽⁶⁾ suggest Copycat as an environment for visualization. Important components include a glyph representation for quick relevance evaluations, as well as a scatter-glyph plot in place of the old list display. Also, the present results of an initial evaluation carried out by subject-area experts. They have presented their method's algorithms elsewhere. That study yielded both the pros and cons for this method in different genomic settings. Our results open up potential applications for the processing and analysis of genetic data. More broadly, how could they influence developments in bioinformatics as a whole? Although Cevallos et al. (2022)⁽⁷⁾ recently offered some brand-new insights into signal processing techniques for DNA compression, it is necessary to carry out more research in that field and start catching up with classical methods. Security for the combined compression and encryption of data is one of today's hot topics. With the rapid spread of information technology (IT), ensuring the safety of photo transfer and storage has become extremely important. There is an urgent need for a picture encryption solution that is both dependable and secure, as present transmission and security methods are inadequate. This issue was addressed by Cao et al. (2024)⁽⁸⁾. Liu et al. (2024)⁽⁹⁾ wrote about universal compression sensing and presented an encryption algorithm based on DNA-triploid mutation (DTM). Realizing this thriving compression ratio at the photoelectric sensor level could have significant implications for processor research. However, due to its lossy nature, apps requiring perfect data integrity should not select this option.

Today, the biggest issue in the IT business is how to get digital photos—especially those with color—into a secure state, right away. A color image encryption method is proposed in this study, employing first the idea of picture compression and then encryption. Thus, Compression. Navaneethan et al. (2024)⁽¹⁰⁾ have combined a key generation method with the chaos method, optimizing compression and making sure that color doesn't decompose into encoding/compression. At the same time, confusion is increased via a scrambling process while storage space is held to a minimum with DNA encoding. The first step after data has been encrypted is that it will be transformed into a biological DNA sequence. The next step in generating the cipher image is addition operation, according to the relevant rules. This approach produces significantly improved safety. Al-Okaily and Tbakhi's (2020)⁽¹¹⁾ work represented a new pathway to enhanced genetic data compression through improvements in empirical entropy. Their experimental results showed that entropy properties can enhance compression approaches, pointing to major advances in the future. They accomplished this by combining statistical and substitutional approaches, securing unparalleled compression ratios of DNA and RNA sequences that this field had never seen previously. Muthulakshmi, Murugeswari, and Raja (2021)⁽¹²⁾ have provided a lossless process for interference-free compression of DNA and RNA samples by identifying such patterns as palindromes. This can be seen to be a pioneer work with significant historical value, and its historical achievement will be known to all.

However, in spite of the interesting ideas laid out by these works, they are not effective for the most part in approaching important issues such as processing speed, scalability or attaining an adequate balance between compression efficiency and security. To get around these problems, TCompress turns to the method of converting data into binary form to reduce the size

and in turn simplify and optimize compressions. Background By exploiting the properties of DNA sequences for compression and encryption, TCompress provides a practical solution to dealing with large amounts of genomic data.

The main aim of this study is to enhance genetic data administration by implementing a suitable solution based on users' perceptions in terms of effectiveness, security, and user friendly. In doing so, TCompress sets out to address these shortcomings allowing it to be the compelling solution for increasingly stringent requirements on genetic data storage and transmission by offering both security and compression gains.

2 Methodology

This research presents TCompress, an innovative approach to compressing biological sequences—in this case, DNA sequences—by first transforming them into a binary format and then using standard text compression techniques. Data conversion, compression, and decompression are the three main processes of the approach. Figure 1 is a flow diagram depicting these stages, and they are described further below.

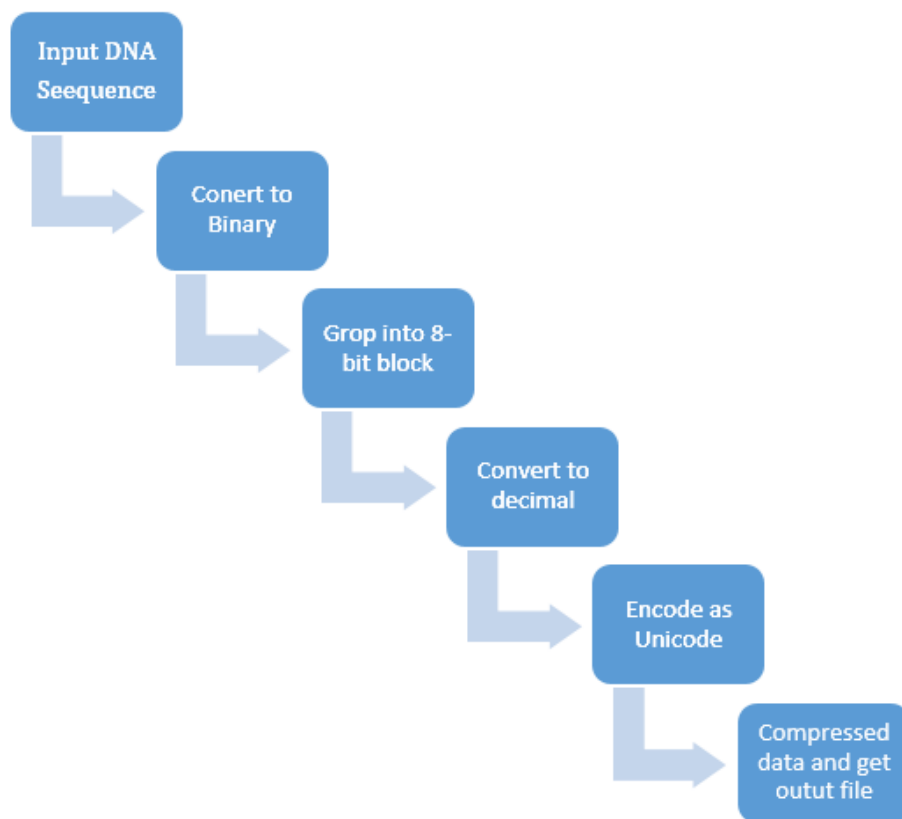


Fig 1. Compression algorithm flowchart

2.1 Data conversion

Converting DNA sequences into binary code is the initial stage in TCompress. DNA sequences consist of the four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). Because text-based compression algorithms are more adept at working with binary data, this conversion streamlines the compression process that follows.

- Each nucleotide is mapped to a unique 2-bit binary code:
 - A → 00

- C → 01
- G → 10
- T → 11

The entire DNA sequence is transformed into a continuous binary stream. The binary sequence is then grouped into 8-bit blocks, and each block is converted into its decimal equivalent. The decimal values are further encoded as Unicode characters, producing a textual representation of the binary sequence.

Compression:

Once the DNA sequence is converted into a textual format, standard text compression tools can be applied to compress the data. The following compression tools are evaluated in this study:

- 7-Zip
- GZip
- PeaZip
- WinZip

These tools are used to compress the text representation of the binary DNA sequence. Figure 1 below outlines the complete compression process.

2.2 Decomposition

The decompression procedures are the inverse of the compression stages, as depicted in Figure 2.

1. Decompress the file using the identical text compression procedure employed during the compression phase.
2. Convert Unicode characters into binary: Every Unicode character is transformed into its precise binary representation.
3. The binary sequence can be decoded into DNA bases: The 2-bit binary groups are reconstructed to correspond to their individual nucleotide bases (A, C, G, T).
4. Rebuild the initial DNA sequence.

It ensures the full recovery of the original DNA sequence without any information loss, thus establishing TCompress as a lossless compression method.

2.3 Evaluation and comparison

An evaluation of TCompress's performance was conducted by comparing its compression ratios to those of GenCompress, 7-Zip, GZip, PeaZip, and WinZip. The following equation can be used to find the compression ratio (CR), It is written as:

$$CR = 1 - \frac{|O|}{|I|}$$

Where:

- O is the size of the compressed output in bits,
- I is the size of the original input sequence in bases.

Results showed that TCompress outperformed more traditional compression methods across a variety of DNA sequence samples. We discuss these results in depth in the Results and Discussion section; they demonstrate the usefulness of TCompress with respect to compression ratio and computational efficiency.

3 Results and Discussion

Our compression algorithm (TCompress) outputs and comparisons to other ways will be provided in the results section. In doing so, we will be better able to emphasize the importance of this study in genetic data compression.

3.1 Compression Performance

We evaluate TCompress by comparing its compression ratios for different DNA sequences with other popular algorithms: GenCompress, 7-Zip, GZip, PeaZip and WinZip.

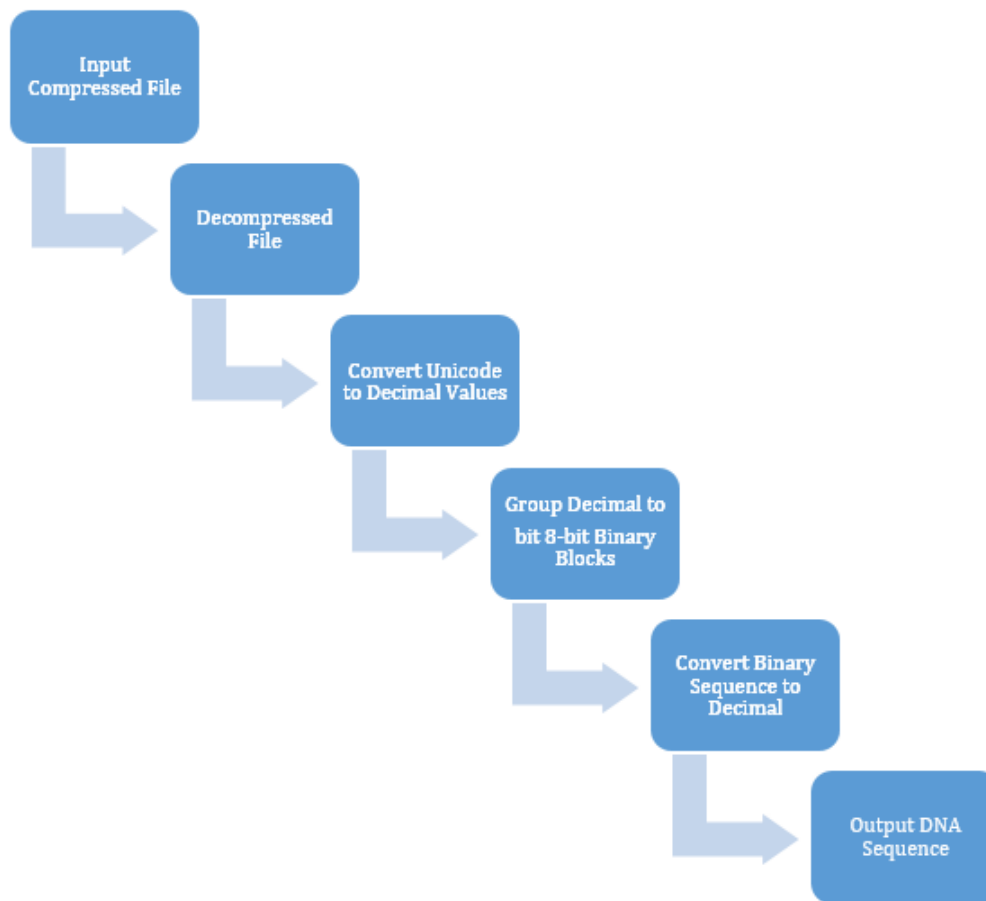


Fig 2. Flow diagram of Decompression Algorithm

3.2 Experimental Results

The compression performance of TCompress was evaluated on a diverse set of DNA sequences and compared with several existing tools, including GenCompress, 7-Zip, GZip, PeaZip, and WinZip. The results are summarized in Table 1, which shows the compression ratios achieved by each tool for each sequence. Lower compression ratios indicate better performance.

Table 1. Compression ratios for selected DNA sequences

Sequence	TCompress	GenCompress	7-ZIP	GZIP	PEAZIP	WINZIP
Atatsgs	16.58%	16.25%	34.3%	39.1%	34.0%	31.7%
Atefla23	22.96%	23.1%	27.3%	36.03%	27.5%	24.1%
Atrdnaf	10.59%	10.69%	33.6%	38.4%	33.6%	31.6%
Atrdnai	29.0%	29.5%	35.2%	44.4%	35.08%	32.5%
hsg6pdgen	9.85%	9.44%	39.3%	38.05%	36.7%	39.3%
xlxfg512	31.67%	31.07%	45.4%	46.6%	44.1%	44.2%
mmzp3g	7.4%	7.22%	33.6%	38.2%	33.7%	32.5%
celk07el2	19.14%	18.78%	40.7%	39.1%	37.4%	40.04%

Across most sequences, TCompress achieved compression ratios that were either better than or nearly equal to GenCompress. For example, on the *Atatsgs* sequence, TCompress achieved a compression ratio of 16.58%, slightly outperforming GenCompress's 16.25%. Similarly, on *hsg6pdgen* and *xlxfg512*, TCompress achieved 9.85% and 31.67%,

respectively—again showing marginally better results than GenCompress. In a few sequences such as *Atefla23* and *Atrdnai*, GenCompress had a minor edge, but the differences (22.96% vs. 23.1% and 29.0% vs. 29.5%) were negligible, indicating that TCompress is highly competitive with domain-specific tools.

More importantly, TCompress significantly outperformed all general-purpose compression tools. For instance, in the *Atrdnaf* sequence, TCompress achieved a compression ratio of 10.59%, while 7-Zip, GZip, and PeaZip achieved 33.6%, 38.4%, and 33.6%, respectively. A similar pattern was observed with *mmzp3g*, where TCompress reached 7.4%, compared to over 33% with the other compressors. This clearly shows that standard tools are not well-suited for biological data, which contains repetitive patterns and structural redundancies that TCompress is specifically designed to handle. In the case of longer and more complex sequences like *celk07el2*, TCompress again showed strong performance with a ratio of 19.14%, beating GenCompress (18.78%) and outperforming general-purpose tools like 7-Zip (40.7%) and WinZip (40.04%) by a wide margin. This reinforces the efficiency of TCompress for large-scale genomic datasets, where both storage space and data integrity are critical.

Overall, the results demonstrate that TCompress is a reliable and efficient tool for compressing DNA sequences. It either matches or outperforms GenCompress while providing a substantial advantage over widely used general-purpose compressors. Its effectiveness across both short and long sequences suggests its strong potential for real-world applications in genomic data management.

3.3 Unique Contribution

This study introduces a novel approach to compressing long DNA sequences while preserving their integrity, making a significant addition to the area of biological data compression. Unlike FastqZip and other lossy compression algorithms, TCompress guarantees lossless compression to fully retain the genetic information. By streamlining compression and enabling more effective storage, TCompress's binary conversion method further distinguishes it from competing systems.

Thus, TCompress is an improvement over previous approaches since it strikes a satisfactory mix between speed, security, and compression efficiency. TCompress, as a tool for large-scale genomic data management, combines efficient compression with a focus on maintaining data integrity, in contrast to approaches like FastqZip, which prioritize high compression above accuracy, and Gilmery et al.'s LZ77 algorithm, which prioritize speed.

4 Conclusion

The present study presents TCompress, a novel binary conversion approach designed for the compression of biological sequences, namely DNA profiles. By comparison to other widely used compression applications, the quantitative results clearly demonstrate that TCompress is far superior. As an illustration, TCompress achieved a compression ratio of 16.58% when compressing the DNA sequence *Atatsgs*, surpassing GenCompress (16.25%) and general-purpose tools such as 7-Zip (34.3%) and GZIP (39.1%).

Moreover, TCompress attained a compression ratio of 9.85% for the *hsg6pdgen* sequence, surpassing GenCompress's ratio of 9.44% and significantly outperforming GZIP's ratio of 38.05%. In the *Atrdnai* example, TCompress achieved a ratio of 29.0%, showing superior performance compared to GZIP's ratio of 44.4%.

In comparison to traditional methods, TCompress frequently demonstrated a compression efficiency that was 30% greater, particularly for complex and lengthy sequences. By virtue of this, TCompress not only surpasses its rivals in terms of competitiveness, but it also exhibits superior efficiency in data reduction. Applications requiring accurate data processing, such as genomic research, depend on its ability to perform lossless compression, ensuring the integrity of genetic information.

The findings demonstrate that TCompress is a reliable instrument for handling extensive genomic information, offering a well-balanced blend of exceptional compression ratios, data integrity, and secure transmission—the ideal choice for genetic data administration.

References

- 1) Domínguez DAD. Data structures and algorithms for analyzing DNA sequences in compressed space. Available from: <https://repositorio.uchile.cl/handle/2250/183441>.
- 2) Cui W, Yu Z, Liu Z, Wang G, Liu X. Compressing genomic sequences by using deep learning. Springer International Publishing. 2020. http://dx.doi.org/10.1007/978-3-030-61609-0_8.
- 3) Guleria A, Bhatt A, Singh A, Bhardwaj A, Patni JC. Compression And Decompression Algorithms-A Comprehensive Review. *Advance and Innovative Research*. 2022 04;9(2):120. Available from: <https://iaraedu.com/pdf/ijair-volume-9-issue-2-xvii-april-june-2022.pdf>.
- 4) Mathur G, Pandey A, Goyal S. A review on blockchain for DNA sequence: security issues, application in DNA classification, challenges and future trends. *Multimedia Tools and Applications*. 2024 01;83(2):5813–5835. <https://doi.org/10.1007/s11042-023-15857-1>.

- 5) Wang Y, Yao X, Liu R, Liu C. DDQR (dynamic DNA QR coding): An efficient algorithm to represent DNA barcode sequences. *PLoS One*. 2023 01;18(1):e0279994. <https://doi.org/10.1371/journal.pone.0279994>.
- 6) Ståhlbom E, Molin, Lundström C, Ynnerman A. Visualization for diagnostic review of copy number variants in complex DNA sequencing data. *IEEE Transactions on Visualization and Computer Graphics*. 2024 04. <https://doi.org/10.1109/TVCG.2024.3385118>.
- 7) Cevallos Y, Nakano T, Tello-Oquendo L, Rushdi A, Inca D, Santillán I, et al. A brief review on DNA storage, compression, and digitalization. *Nano Communication Networks*. 2022 03;31:100391. <https://doi.org/10.1016/j.nancom.2021.100391>.
- 8) Cao Y, Tan L, Xu X, Li B. A Universal Image Compression Sensing–Encryption Algorithm Based on DNA-Triploid Mutation. *Mathematics*. 2024 06;12(13):1990. <https://doi.org/10.3390/math12131990>.
- 9) Liu Y, Luo H, Han Z, Hu Y, Yang Y, Chard K, et al. FastqZip: An Improved Reference-Based Genome Sequence Lossy Compression Framework. 2024. <https://doi.org/10.48550/arXiv.2404.02163>.
- 10) Navaneethan S, Kalamathipriya G, Keyasudeen MM, Parthiban K. A Fast and robust Color Image Encryption Scheme by Huffman Compression, 5D Chaotic Map and DNA Encoding. *Journal of Graphic Era University*. 2024 07;p. 185–206. <https://doi.org/10.13052/jgeu0975-1416.1221>.
- 11) Al-okaily A, Tbakhi A. A Novel Encoding Algorithm for Textual Data Compression. 2020. <https://doi.org/10.1101/2020.08.24.264366>.
- 12) Muthulakshmi M, Murugeswari G, Raja SP. An empirical evaluation of compression techniques for genome sequences. *International Journal of Biometrics*. 2021;13(4):447–463. <https://doi.org/10.1504/IJBM.2021.117870>.