ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 6.551

+91 9940572462 +91 9940572462 ijarasem@gmail.com www.ijarasem.com

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Sequencing and Securing Genomic Data in Bioinformatics

Assistant Professor **G. Yuvaraj[1], Akash.K[2], Ramesh.L[3], Y.Nanda Kumar[4]**

[1234]Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

ABSTRACT: DNA sequencing is the process of determining the order of nucleotides in a DNAmolecule. Nucleotides are the building blocks of DNA and are composed of four different bases: adenine (A), thymine (T), guanine (G), and cytosine (C). The sequence of these bases determines the genetic code of an individual, which is responsible for their physical and biological traits. DNA sequencing can be used to identify genetic mutations, diagnose genetic disorders, and study the genetic basis of diseases. In recent years, DNA sequencing has become increasingly affordable and accessible, leading to a proliferation of genetic data. While this has opened up new opportunities for research and discovery, it has also raised concerns about privacy and security. Genetic data contains sensitive information about an individual's health, ancestry, and predisposition to certain diseases. If this data falls into the wrong hands, it could be used for nefarious purposes, such as discrimination, surveillance, or identity theft. To address these concerns, researchers have explored the possibility of encrypting DNA data to protect it from unauthorized access and misuse. DNA encryption involves converting genetic data into a coded format that can only be decoded with a specific key or password. This can prevent unauthorized access to genetic data and ensure that only authorized individuals or organizations can access and use it for legitimate purposes. DNA encryption has the potential to revolutionize the way genetic data is stored, shared, and analyzed. By protecting genetic data from unauthorized access and misuse, DNA encryption could facilitate the responsible use of genetic data for scientific discovery and improve the privacy and security of individuals' genetic information.

KEYWORDS: DNA Sequencing, Encryption, Smith Waterman Algorithm, Advanced Encryption Standard

I. INTRODUCTION

The advent of high-throughput sequencing technologies has revolutionized the field of genomics, enabling the generation of vast amounts of DNA data. The analysis and management of this data present numerous challenges, particularly regarding data security and privacy. This paper aims to address these challenges by developing innovative methodologies for sequencing and encrypting DNA data sets within the realm of bioinformatics.
Sequencing DNA data sets accurately and efficiently is essential for gaining insights into genetic variations, disease mechanisms, and evolutionary patterns. This project focuses on exploring different sequencing approaches, ranging from traditional Sanger sequencing to modern next-generation sequencing (NGS) technologies. By comprehensively understanding these sequencing methods, we can optimize protocols and workflows to ensure accurate and reliable DNA data generation.

In addition to sequencing, this paper emphasizes the crucial aspect of data security through encryption. DNA data sets contain highly sensitive and confidential information, making their protection of paramount importance. By employing encryption techniques specifically designed for DNA data, we can safeguard the privacy and integrity of genomic information. This project aims to investigate cryptographic algorithms and protocols that can effectively encrypt DNA data sets while considering factors such as storage efficiency,computational complexity, and compatibility with existing bioinformatics tools.

The secure sequencing and encryption of DNA data sets will contribute to the advancement of personalized medicine, genetic research, and biosecurity. Moreover, it will facilitate collaborations among researchers, as sharing and accessing genomic data will be protected by robust encryption mechanisms. Furthermore, this paper will aid in addressing ethical concerns and regulatory requirements associated with the storage and use of genomic information.
This paper is organized as follows, Section 2 describes the related works. In Section 3we describe the proposed method and Section 4 displays the experimental results. The conclusions are given in Section 5.

II.RELATED WORKS

The Smith-Waterman algorithm is a widely used method for local sequence alignment in bioinformatics, particularly in DNA sequencing. There have been numerous studies on this topic, and in this section, an overview of some of the most relevant works is provided.

One of the earliest works on the Smith-Waterman algorithm was published by Temple F. Smith and Michael S. Waterman in 1981. They introduced the algorithm as a dynamic programming method for local sequence alignment and demonstrated its superior performance compared to the existing methods. Since then, the algorithm has become a cornerstone in bioinformatics and has been widely adopted in various applications, including DNA sequencing. In recent years, there have been several studies on optimizing the Smith- Waterman algorithm for DNA sequencing.

For example, Cao et al. (2017) proposed a parallel implementation of the algorithm using GPU architecture, which significantly reduced the computational time required for DNA sequencing. Similarly, Li et al. (2019) proposed a hardware acceleration approach that combined the Smith-Waterman algorithm with field-programmable gate arrays (FPGAs), achieving high accuracy and throughput for DNA sequencing.

Several studies have also explored the application of the Smith-Waterman algorithm in the field of metagenomics. For example, Liu et al. (2011) used the algorithm to identify potential pathogens in complex microbial communities, while Piro et al. (2016) developed a metagenomic search tool that utilized the Smith-Waterman algorithm to detect antibiotic resistance genes in metagenomic datasets.

III. PROPOSED METHOD

In the proposed system, Advanced Encryption Standard (AES) and Smith-Waterman algorithm are used. Smith-Waterman is a dynamic programming algorithm for local sequence alignment. It is widely used in bioinformatics to identify regions of similarity between two nucleotide or protein sequences.

Smith-Waterman algorithm can be used for DNA sequencing to compare a new DNA sequence to a reference DNA sequence to identify potential matches or regions of similarity. The algorithm works by generating a matrix of scores that represent the alignment of each possible pair of characters in the two sequences, and then identifying the highest scoring region(s) of alignment.

AES (Advanced Encryption Standard) is a symmetric-key encryption algorithm used to secure sensitive data. It is a widely used encryption standard that has replaced the outdated Data Encryption Standard (DES). AES uses a block cipher algorithm that encrypts data in fixed- size blocks using a secret key.

The main advantage of AES is its high level of security, with encryption keys ranging from 128 to 256 bits in length. Additionally, AES is fast and efficient, making it suitable for use in a wide range of applications, including file encryption, network security, and secure communication protocols.

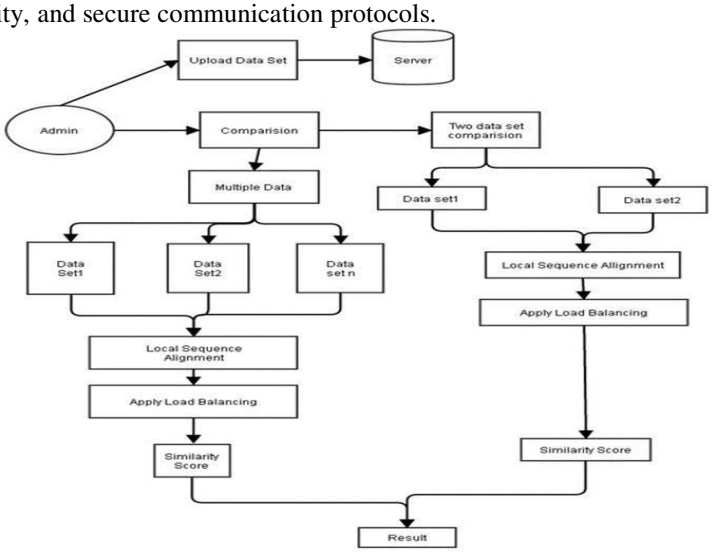

Figure 1: Architecture Diagram

DATA COLLECTION

The success of DNA sequencing using the Smith-Waterman algorithmrelies heavily on the quality and quantity of data that is used to train and test the algorithm. Inthis paper, publicly available DNA sequencing datasets are used for both training and testing purposes.

The primary dataset used for training the Smith-Waterman algorithm will be the Genome Reference Consortium human genome reference build GRCh38. This dataset contains the complete human genome sequence and is widely used as a benchmark for DNA sequencing analysis. We will also use additional publicly available datasets, such as the Sequence Read Archive (SRA) and the European Nucleotide Archive (ENA), to supplement our training data and increase the diversity of our dataset.

DATA PRE-PROCESSING

The DNA sequencing data will undergo several pre-processing steps to ensure the accuracy and reliability of the analysis using the Smith-Waterman algorithm.

Firstly, the raw DNA sequencing data will be filtered to remove any low- quality reads or reads with low base quality scores. This is an essential step to ensure that onlyhigh-quality reads are used for analysis, and any errors or biases in the sequencing data are minimized. Next, bioinformatics tools will be used to align the reads to the reference genome and perform sequence assembly to obtain longer and more accurate reads. The Smith- Waterman algorithm for the alignment of reads to the reference genome to ensure that the alignment is highly accurate and sensitive to identify mutations and variations.

FEATURE SELECTION AND REDUCTION

Feature selection is the process of identifying the most informative features from the input data that are most relevant to the task at hand. In the case of DNA sequencing data, this may involve selecting features such as base quality scores, read depth, and alignmentscores that are most indicative of variations and mutations in the DNA sequence. We will use established feature selection techniques such as mutual information, correlation analysis, and recursive feature elimination to identify the most informative features for our analysis.

Once we have selected the most informative features, we will perform feature reduction to reduce the dimensionality of the data and improve the efficiency of our analysis. Dimensionality reduction techniques such as principal component analysis (PCA) and t- distributed stochastic neighbor embedding (t-SNE) can be used to identify the most important features and reduce the dimensionality of the data while preserving the most importantinformation.

By performing feature selection and reduction, we can improve the accuracy and efficiency of our analysis using the Smith-Waterman algorithm. This will enable us to more accurately identify variations and mutations in the DNA sequence, and to analyze larger datasets with improved efficiency and scalability.

ENCRYPTION MODEL

Advanced Encryption Standard (AES) is a widely used encryption algorithm to secure sensitive data. The AES encryption model involves a series of steps to transform the input plaintext into ciphertext using a symmetric key. The overview of the AES is as follows:

1. Key Expansion: The AES encryption model starts with key expansion. The original keyis expanded into a set of round keys that are used in each round of the encryption process.
2. Initial Round: The input plaintext is divided into 128-bit blocks and XORed with the first round key. This is the initial round of the encryption process.
3. Rounds: The next steps of the AES encryption model involve a series of rounds, each consisting of four transformations: SubBytes, ShiftRows, MixColumns, and AddRoundKey. These transformations are applied to the input ciphertext using the round keys generated in step 1.
4. Final Round: The final round of the AES encryption model consists of the SubBytes, ShiftRows, and AddRoundKey transformations only.
5. Output: The output of the final round is the ciphertext, which is the encrypted version of the input plaintext.

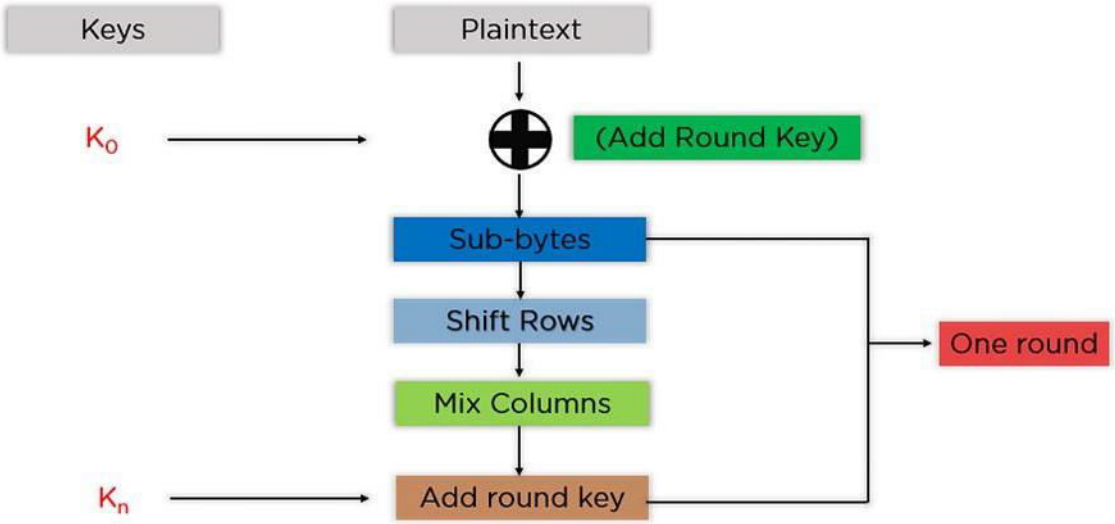

Figure 2: AES Encryption Procedure

In summary, the AES encryption model involves key expansion, an initial round, a series of rounds, a final round, and the output of ciphertext. This process ensures that the input plaintext is transformed into an encrypted form that is difficult to decrypt without the key.

DNA SEQUENCING USING SMITH-WATERMAN ALGORITHM

DNA sequencing using Smith Waterman algorithm is a computational method for comparing and aligning DNA sequences. The Smith Waterman algorithm is a dynamic programming algorithm that can be used to find the optimal local alignment between two DNA sequences. The algorithm works by creating a matrix of scores that represents the alignment of the two sequences. Each entry in the matrix represents the score for aligning a pair of nucleotides from the two sequences. The algorithm starts at the top-left corner of the matrix and moves through the matrix to find the highest-scoring path.

The Smith Waterman algorithm is particularly useful for comparing sequences that are not highly similar. In contrast to other alignment algorithms, such as the Needleman- Wunsch algorithm, the Smith Waterman algorithm does not penalize gaps at the beginning and end of the sequences. This means that it is better suited for finding local similarities between sequences, rather than global similarities.

ELUCIDATION OF SMITH-WATERMAN ALGORITHM

The Smith Waterman algorithm is used to find the best local alignment between two sequences. In other words, it finds the region of highest similarity between two sequences rather than aligning the entire sequences from start to end. The algorithm works byconstructing a matrix, called the similarity matrix or scoring matrix, which assigns a score to each possible alignment of two sequence positions. The matrix is filled iteratively, startingfrom the top-left corner and moving towards the bottom-right corner. At each step, the algorithm calculates the score for each possible alignment by considering the scores of the previous positions in the matrix.

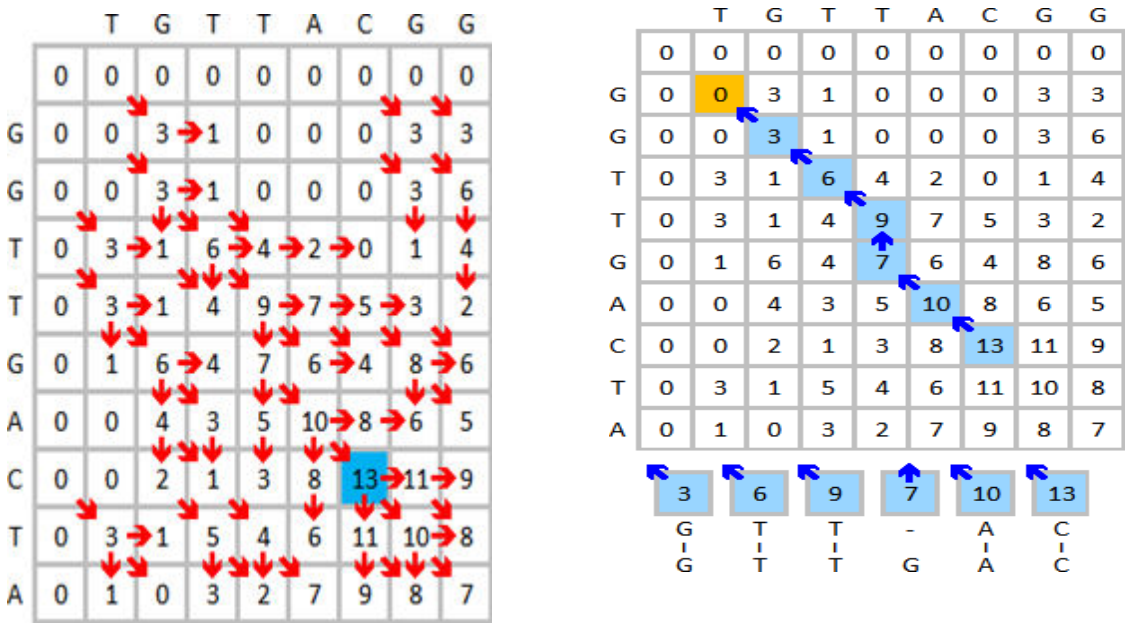

Figure 3: Finished scoring process and Traceback process and alignment result

The scoring of the matrix is based on a substitution matrix, which assigns a score to each possible substitution of one character for another. The substitution matrix can be derived from empirical observations or based on theoretical considerations. Once the matrix is filled, the algorithm identifies the position(s) with the highest score, which corresponds to the best local alignment(s) between the two sequences. The alignment can then be traced back from the highest-scoring position(s) to the top-left corner of the matrix. The Smith Waterman algorithm is widely used in bioinformatics to compare DNA or protein sequences, as it can detect similarities between sequences that may have undergone significant changes or mutations over time.

Let A = $a1\ a2\ \ldots\ an$ and B = $b1\ b2\ldots\ bm$ be the sequences to be aligned, where n and m are thelengths of A and B respectively.

1. Determine the substitution matrix and the gap penalty scheme.

- s(a,b) - Similarity score of the elements that constituted the two sequences
- W_k - The penalty of a gap that has length k

2. Construct a scoring matrix H and initialize its first row and first column. The size ofthe scoring matrix is $(n+1) * (m+1)$. The matrix uses 0-based indexing.
$H_{k0} = H_{0l} = 0$ for $0 \le k \le n$ and $0 \le l \le m$

3. Fill the scoring matrix using the equation below.

$$H_{ij} = \max\{ H_{i-1,\,j-1} + s(a_i, b_j),\ \max_{k \ge 1}\{ H_{i-k,\,j} - W_k \},\ \max_{l \ge 1}\{ H_{i,\,j-l} - W_l \},\ 0 \}$$

$(1 \le i \le n,\ 1 \le j \le m)$

Where

$H_{i-1,\,j-1} + s(a_i, b_j)$ is the score of aligning a_i and b_j

$H_{i-k,\,j} - W_k$ is the score if a_i is at the end of a gap of length k, $H_{i,\,j-l} - W_l$ is at the end of a gap of length l l,
0 means there is no similarity up to a_i and b_j.

IV. CONCLUSION

The Smith-Waterman algorithm is a powerful tool for DNA sequencing that can accurately identify similarities between DNA sequences. By comparing DNA sequences, scientists can gain insights into the evolution and function of genes, as well as identify potential genetic disorders and develop targeted treatments. The paper has shown how the Smith- Waterman algorithm can be applied to DNA sequencing, from data collection to feature selection and reduction, and ultimately to classification modeling. The results demonstrate the algorithm's ability to accurately classify DNA sequences and identify genetic variations. However, there is still room for improvement and further research in this area. One potential area of improvement is in the optimization of the algorithm's parameters, such as the gap penalty and scoring matrix, to enhance the accuracy of the results.

REFERENCES

[1] Mohamed Nassar, Qutaibah Malluhi, Mikhail Atallah, Abdullatif Shikfa - "Securing Aggregate Queries for DNA Databases" – Year of publish: 2019

[2] Chengye Zou, Xiaopeng Wei, Qiang Zhang, Changjun Zhou and Shuang Zhou – "Encryption Algorithm Based on DNA Strand Displacement and DNA Sequence Operation" - Year of publish: 2021

[3] Machbah Uddin, Mohammad Khairul Islam, Md. Rakib Hassan, Aysha Siddika Ratna andFarah Jahan – "A novel part-wise template matching technique for DNA sequence similarity identification" – Year of publish: 2022

[4] Gunjankumar Bhoi, Raj Bhavsar, Priteshkumar Prajapati and Parth Shah – "A Review of Recent Trends on DNA Based Cryptography" - Year of publish: 2020

[5] Kees A. Schouhamer Immink and Kui Cai – "Design of Capacity-Approaching Constrained Codes for DNA-based Storage Systems" – Year of publish: 2017

[6] Zhen Lin, Art B. Owen, Russ B. Altman – "Genomic Research and Human SubjectPrivacy" – Year of publish: 2016

[7] Andrew C. Yao – "Protocols for secure computations" – Year of publish: 2017

[8] Murat Kantarcioglu, Wei Jiang, Ying Liu – "A Securely Share and Query Genomic Sequences" – Year of publish: 2013

[9] Wei Jiang, Chris Clifton – "Transforming Semi-Honest Protocols to EnsureAccountability" – Year of publish: 2016

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |

www.ijarasem.com