Development and analysis of a parallel method for detecting chromosomal translocations and inversions in DNA sequences[⋆]

Lesia Mochurad [a,⋆,†], Renata Vladyka [a,†]

[a] *Lviv Polytechnic National University, 12 Bandera street, Lviv, 79013, Ukraine*

Abstract

In the current context of the growing volume of genetic research and the need for the fast and accurate detection of chromosomal abnormalities, developing efficient algorithms that provide high performance, scalability, and accuracy in comparing DNA sequences is a critical task. In this study, the aim was to create and analyze a parallel algorithm for detecting chromosomal translocations and inversions in DNA sequences. For this purpose, we implemented an algorithm capable of comparing DNA sequences to detect genetic abnormalities, including translocations and inversions. The use of parallel computing made it possible to significantly improve the efficiency of the analysis, reducing the algorithm's execution time and increasing scalability. Particular attention was paid to how the algorithm uses multithreading and how it efficiently distributes the load between threads. To detect inversions, an algorithm was developed that compares two DNA sequences and detects possible changes in nucleotide sequences. To find translocations, the Needleman-Wunsch algorithm was applied, which uses parallel computing to optimally align genetic fragments. The results of the algorithms have shown high efficiency in detecting chromosomal mutations, which is important for genetic research and can be used for medical diagnosis.

1. Introduction

This paper is devoted to the development of an algorithm for preventing chromosomal translocations and inversions. Chromosomal translocations and inversions are complex structural changes in genetic materials that occur as a result of disruptions in the organization of chromosomes. Translocations involve moving a part of a chromosome to another, causing genetic sequences to be interrupted and reconnected, while inversions occur when a chromosome fragment rotates around its axis, changing the order of genes and other genetic elements. These structural variations can occur as a result of various errors during cell division. Incorrect joining or rotation of chromosome parts can lead to important changes in genetic information, which is often associated with the development of various genetic diseases and developmental disorders [1].

The main reasons why it is important to avoid chromosomal translocations and inversions:

1. A translocation between chromosomes 21 and 14 can lead to "familial" Down syndrome, where one of the parents may be a carrier of the translocation. This, in turn, can affect the likelihood of having a child with Down syndrome.

2. Gametes carrying defective chromosomes with inversion often develop into non-viable or miscarried organisms in the early stages of embryogenesis.

3. Robertson translocations can cause differences in the number of chromosomes between closely related species, which affects their fertility and genetic stability.

4. If gametes with an inverted chromosome develop into organisms, 50% of the gametes of these organisms may not be viable. However, the other 50% may be viable and maintain the mutation in the population.

Thus, the diagnosis and prediction of chromosomal inversions and translocations are critical in genetic research and medical practice. Detection of these changes contributes to early diagnosis, which allows timely treatment and monitoring of patients. Predicting the risks of developing certain diseases based on the identified mutations also helps patients to realize their genetic risks and opportunities for prevention. To solve this problem, the Needleman-Wunsch algorithm is used [2]. However, one of the main problems with this algorithm is its high computational complexity; this algorithm has quadratic complexity. The latter makes it difficult to process large DNA sequences. The development of a parallel algorithm can reduce the execution time by simultaneously processing parts of the data [3], [4], [5]. Ensuring efficient parallelization requires balanced load distribution among threads to prevent idle time, which may involve optimizing data structures and algorithms. Proper synchronization is also essential to avoid conflicts when accessing shared resources, especially in algorithms with interdependent results. Maintaining result accuracy during parallel processing may require extra validation. Lastly, scalability is a key factor. It should be able to scale across different hardware platforms, from personal computers to supercomputer clusters, to achieve optimal results in different environments [6], [7].

Solving all the problems described above is an important step to improve the speed and accuracy of chromosomal abnormalities diagnosis, which, in turn, will increase the efficiency of genetic research and clinical practice.

In the field of genetic research, there are many important aspects that relate to the study of genetic mutations, their diagnosis, and their impact on human health. Various articles offer valuable insights that emphasize the importance of such mutations in medicine, evolution, and biology. For example, paper [8] discusses the work of the Human Gene Mutation Database (HGMD), which collects and analyzes information about human genetic mutations. The authors describe the functions of the HGMD, its use in clinical diagnostics and research, as well as plans to expand the database, which include the integration of GTEx project data and the introduction of automated tools for mutation prediction. The advantages of the article are a detailed overview of the HGMD functions and an emphasis on its role in clinical practice, but the disadvantages are the complexity of automatic identification of new mutations and the lack of examples of real-world use of the HGMD in clinical cases.

In [9], the authors analyze different types of genetic mutations and their impact on human health, including the occurrence of diseases due to mutations in genes or chromosomes. The advantages of this article are a clear description of the causes and consequences of genetic diseases, as well as a detailed overview of the types of mutations. However, there are also disadvantages, such as general statements that are not always supported by specific data, as well as a lack of details about research methods. Nevertheless, the article is useful for familiarizing oneself with genetic mutations and their impact on the human body.

Paper [10] investigates the use of biodosimetry to assess radiation doses in US military personnel who participated in nuclear tests after World War II. The main goal of this study is to assess chromosomal aberrations, such as inversions and translocations, as a method of retrospective biodosimetry. The results show that inversions can be an effective way to establish radiation doses received decades ago and that their combination with translocations improves the accuracy of the estimate. The study also indicates the influence of age and smoking status on the frequency of aberrations, which is higher in older individuals and smokers.

The PETI method is discussed in [11]. It effectively induces precise DNA recombinations in human cells. PETI successfully generates recombination and inversion mutations in endogenous genomes and can correct disease-related inversions and translocations, making it promising for disease modeling and therapeutic approaches. This method is characterized by high accuracy and flexibility compared to other genome editing methods such as Prime-Del and twinPE, but it is still limited in its use for episomal mutations and requires further research to confirm its results.

The authors of [12] investigate the effectiveness of the third generation of sequencing analysis for detecting breakpoints in unbalanced chromosomal translocations. The use of Nanopore long reads allows for the detection of microdeletions, microinsertions, and other structural changes that complement translocations. This approach provides accurate genetic information necessary for diagnosis and treatment. However, the high cost and possible experimental failures point to the need for further research and optimization of methods.

The paper [13] discusses the role of chromosomal rearrangements in the genetic differentiation and evolution of populations, in particular, using the example of the spiny frog with a chromosomal translocation polymorphism. The authors use whole-chromosome staining (WCP) and genetic analysis to confirm the common origin of the translocations. They found that translocated chromosomes have a higher level of genetic differentiation due to recombination suppression, which contributes to genetic diversity and population differentiation. The article also discusses the mechanisms of recombination suppression, such as the accumulation of repetitive sequences and the capture of adapted alleles.

This review is rounded off by article [14], which analyzes the importance of chromosomal inversions in the plant kingdom and their role in plant evolution. The authors note that inversions are common in many groups of plants and are often associated with locally advantageous traits that promote adaptation and speciation. The article also discusses methods for detecting inversions, such as karyotyping, genetic mapping, and high-fidelity sequencing. The authors call for further research to better understand the origin, evolutionary role, and molecular mechanisms of inversions in plants.

Thus, the analysis of scientific studies shows that existing approaches to diagnosing and predicting chromosomal inversions and translocations have their advantages, but also significant disadvantages. For example, the HGMD database, although providing useful information for clinical practice, faces problems with the automatic identification of new mutations. The use of biodosimetry to estimate radiation doses, in particular by analyzing chromosomal aberrations, has proven effective, but is dependent on factors such as age and smoking status. The PETI method demonstrates high accuracy in genetic editing, but requires additional research to confirm its widespread use. Chromosomal rearrangements confirm their role in the genetic differentiation of populations, but require a deeper study of the mechanisms of recombination suppression. The study of inversions in plants also emphasizes the need for further research into the evolutionary mechanisms of these changes. This confirms the importance of continuing research in this area. In particular, the need to develop a parallel algorithm for the diagnosis and prediction of chromosomal inversions and translocations is driven by the need for faster and more accurate detection of genetic abnormalities, which is critical for effective treatment and management of diseases.

The main contribution of this paper:

- A new parallel method for detecting chromosomal translocations and inversions using multithreading and parallel computing (OpenMP) is proposed. This method ensures optimal load distribution between threads, reducing processing time and increasing scalability.
- The Needleman-Wunsch alignment method has been parallelized to compare sequences, which has increased the accuracy of translocation detection. This approach allows for more accurate identification of chromosomal abnormalities, outperforming traditional sequential alignment methods.
- When using eight threads, the proposed method achieves speedups of up to 3.8 for translocations and 3.4 for inversions. This confirms its suitability for processing large amounts of data and the possibility of further optimization on multi-core or GPUs.

2. Methodic and materials

2.1. General scheme of the method for detecting chromosomal mutations

The task is to develop a method for detecting mutations in DNA sequences. There are two sets of DNA for inversion and four for translocation: half of them are taken as a reference, i.e. without mutations, and the rest with mutations. This method should be able to detect two types of mutations: inversions and translocations.

Let X and Y be DNA sequences of length n and m, respectively, $X=(x_1, x_2, ..., x_n)$ and $Y=(y_1, y_2, ..., y_m)$, where x_i and y_i take the values $\{A, C, T, G\}$. Sequence Y is the same sequence X, but at a certain interval from i to $i + k$ the sequence is inverted, i.e. $Y[i:i+k]$! $X[i:i+k]$ and $Y[i:i+k] = \text{reverse}(X[i:i+k])$. To search for inversions, the algorithm comparing two DNA sequences

is parallelized. This algorithm detects mismatches in nucleotide sequences and determines the gaps where nucleotides in DNA molecules do not match. After that, the found gaps are checked for inversions. The result of this check is the determination of the gaps where inversions are detected.

To search for translocations, we need to parallelize the Needleman-Wunsch algorithm, which consists in constructing an optimal alignment, using the optimal alignments of the initial fragments of the original sequences obtained in the previous steps. For two sequences X and Y with elements x_I (0<i<n) and y_j (0<j<m), we construct the matrix F. The element of this matrix contains the weight (score) of the best alignment of elements x_1, x_2, ..., x_i and y_1, y_2, ..., y_j of the sequences X and Y, respectively. We build the matrix F recursively. We assign the starting point zero weight $F(0,0)=0$. Then we fill in the matrix in ascending order of both indices, i.e. from the upper left corner to the lower right. If we have already defined $F(i-1,j-1)$, $F(i-1, j)$, $F(i, j-1)$, then we can define $F(i, j)$, as

$$F(i, j) = max \left\{ \begin{array}{l} F(i - 1, j - 1) + S(x_i, y_j) \\ F(i - 1, j) + \delta \\ F(i, j - 1) + \delta \end{array} \right. ,$$

where:

- $F(i, j)$ – the value in the alignment matrix in the position (i, j);
- $S(x_i, y_j)$ – the value of the penalty (or reward) for matching (match/mismatch) characters x_i and y_j;
- δ – the value of the gap penalty, which can be negative or positive.

This formula recursively calculates the values in each cell of the alignment matrix $F(i, j)$, considering the three possible ways to reach the current cell: diagonally (matching), vertically (insertion/deletion in the first sequence), and horizontally (insertion/deletion in the second sequence). Figure 1 shows a general flowchart of the chromosomal mutation detection algorithm
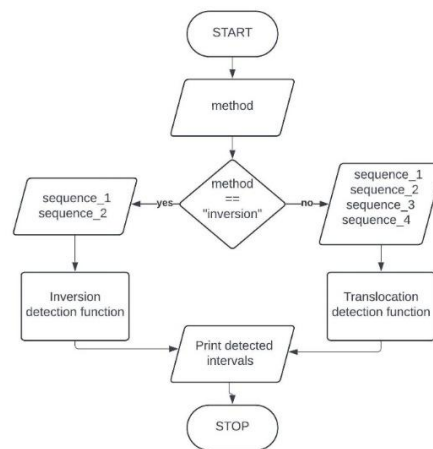
Figure 1: Flowchart of the chromosomal mutation detection algorithm

2.2 Parallelizing the detection of chromosomal inversions and translocations

This paper uses OpenMP technology to detect chromosomal inversions and translocations in the genome. OpenMP allows parallelization of the code, reducing computation time and increasing program efficiency. Testing on a different number of threads showed a significant speedup of the algorithm. By working with shared memory, OpenMP simplifies the exchange of data between threads, which helps to synchronize them.

The main work of the algorithm for finding chromosomal inversions can be divided into two steps:

1. Detection of chromosomal inversion gaps.
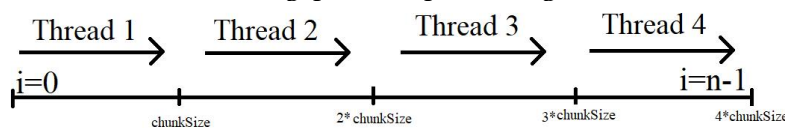
2. Analysis and verification of the results.

The detection of chromosomal inversion gaps is performed by comparing the reference with the target in order to detect mismatches. Accordingly, in a sequential algorithm, the search is performed from the beginning of the array to its end (see Figure 2).
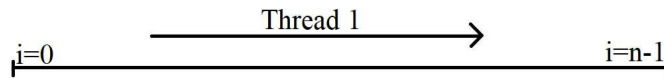
Figure 2: Visualization of searching for inversion gaps in a sequential algorithm

This process was parallelized. The array is divided into N intervals of equal length, where N is the number of threads. Accordingly, having a gap length of 2000 characters and $N = 4$, each thread will search for a gap of 500 characters. Each thread will store the found gaps in its own array and, upon completion of execution, add it to the general array for further processing.

To parallelize the algorithm, we used the *schedule(static, chunkSize)* directive, which statically distributes tasks between threads. *chunkSize* – the size of the block that will be processed by a particular thread. The use of the *static* parameter ensures that the intervals will be consistent. Visualization of the search for inversion gaps in the parallel algorithm is shown in Figure 3.


Figure 3: Visualization of the search for inversion gaps in the parallel algorithm

Another resource-intensive process is the analysis and verification of the found inversion gaps. To speed up this process, we perform an initial check to see if the detected nucleotide sequence in the target is an inversion of the corresponding nucleotides in the reference. If the condition is met, we save the gap in the format (start, end, inversion), where the start is the beginning, the end is the end of the gap, and inversion indicates that this nucleotide sequence is an inversion and no further processing is necessary. Otherwise, additional analysis is required.

Additional analysis includes merging gaps and checking the newly formed sequences for inversion. This process cannot be parallelized and is identical to that of the sequential algorithm [15], [16].

The sequential algorithm for finding translocations involves the use of the Needleman-Wunsch method, which consists of sequentially calculating the elements of a matrix based on previously found matrix values. The idea of parallelizing this algorithm is to calculate a specific row or column of the matrix in a separate thread (see Figure 4). With this implementation, the problem of a resource race arises. It is necessary to assume that all three elements that are used to calculate a particular *(i; j)* element of the matrix are calculated. For this logic, the matrix is used to identify whether the *(i; j)* element is calculated or not. Waiting for the necessary elements is realized by means of a spin lock, the essence of which is to place the thread in a loop that constantly checks the necessary condition for calculating the element.

To parallelize the algorithm, we used the *schedule(static, 1)* directive, which statically distributes tasks between threads. Using the static parameter ensures that the work for the threads is distributed evenly, and the second parameter ensures that each subsequent row or column is calculated under a different thread.
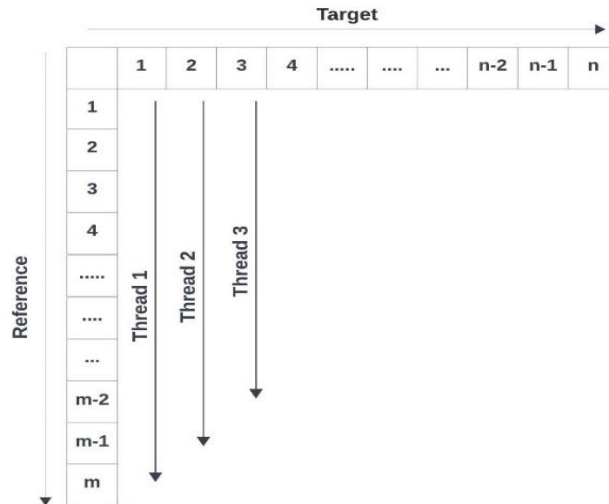
Figure 4: Visualization of the matrix calculation process

Thus, the main elements of the novelty of the proposed method include:

- Parallelization of the algorithm for detecting chromosomal inversions;
- Using OpenMP for static parallelization;
- Parallelization of the Needleman-Wunsch algorithm for translocation detection;
- Implementation of thread synchronization via spin lock;
- Development of a mechanism for checking inversions without additional processing in case of successful inversion detection.

3. Results

The data for testing were taken from the DNA sequence dataset [17]. This dataset is a dataset that contains DNA sequences and contains information about nucleotide sequences consisting of four possible bases: adenine (A), cytosine (C), guanine (G), and thymine (T).

First, we present the results of the parallel algorithm using OpenMP technology to find inversions and translocations (see Table 1 and Table 2, respectively). The program execution time will depend on the number of threads and the number of sequences in the file.

Table 1
Program execution time (milliseconds) for searching for inversions in parallel calculations

n	Number of threads			
	1	3	4	8
4021	8	7	8	10
8042	27	19	16	15
32161	97	61	45	34
192261	611	313	222	180

Table 2
Program execution time (seconds) for searching for translocations in parallel calculations

n	Number of threads			
	1	3	4	8
349	0.3	0.17	0.14	0.4
1045	1.8	1.2	0.9	0.8
4177	25	14.5	10	9
5801	57	31	21	15

Numerical experiments were performed on four different samples of nucleotide sequences. According to Table 1, it can be concluded that parallelization is not advisable for small values of n,

and the calculation speed improves with increasing data volume. Similarly, Table 2 shows that a small matrix dimension is not optimal for a large number of threads. Similarly to the algorithm for detecting inversions for translocations, the parallelization efficiency improves with increasing data size.

Further, based on the results obtained, we calculate the speedup $S_p = \dfrac{T_s(n)}{T_p(n)}$ and efficiency $E_p = \dfrac{S_p(n)}{p}$ of the proposed parallel method, where $T_s(n)$ – the execution time of the sequential algorithm, $T_p(n)$ – the execution time of the parallel algorithm, and p is the number of threads. The results are shown in Table 3 and Table 4

Table 3
Speedup and efficiency indicators of the parallel algorithm for detecting inversions

n	2		4		8	
	speedup	efficiency	speedup	efficiency	speedup	efficiency
4021	1.14	0.14	1.00	0.13	0.80	0.10
8042	1.42	0.18	1.69	0.21	1.80	0.23
32161	1.59	0.20	2.16	0.27	2.85	0.36
192261	1.95	0.24	2.75	0.34	3.39	0.42

Table 4
Speedup and efficiency indicators of a parallel algorithm for translocation detection

n	2		4		8	
	speedup	efficiency	speedup	efficiency	speedup	efficiency
349	1.76	0.22	2.14	0.27	0.75	0.09
1045	1.50	0.19	2.00	0.25	2.25	0.28
4177	1.72	0.22	2.50	0.31	2.78	0.35
5801	1.84	0.23	2.71	0.34	3.80	0.48

From the results, we can see that the proposed parallel method yields good speedup and efficiency with increasing number of threads and input data dimensionality, which indicates that the method is well scalable. It should be noted that all experiments were performed on a computer with eight logical processors. Therefore, they can be significantly improved with a more powerful computing system. The paper also investigates the accuracy achieved by the proposed approach for finding inversions and translocations. To do this, the program was run 1000 times with different lengths of DNA sets with pre-introduced mutations randomly, and the final result was calculated as an average. Table 5 presents the accuracy results of the proposed method on different data sizes, and Table 6 compares the accuracy of the Smith-Waterman algorithm [18] and the Needleman-Wunsch algorithm.

Table 5
Accuracy of the proposed parallel algorithms on different data sizes, %

Sequence length	100	200	500	1000
inversion	98.5	98.6	99.0	99.2
translocation	96.3	97.0	97.5	97.6

Table 6
Comparison of the accuracy of the Smith-Waterman and Needleman-Wunsch algorithms, %

Sequence length	Smith-Waterman algorithm	Needleman-Wunsch algorithm

100	96.7	96.3
200	97.3	97.0
500	97.4	97.5
1000	97.5	97.6

4. Conclusions and future research

In this paper, we investigated the effectiveness of parallel algorithms for detecting chromosomal mutations, such as inversions and translocations, using OpenMP technology. The algorithms were developed to compare DNA sequences and detect deviations in nucleotide sequences, which allowed for high accuracy in identifying inversions. For translocations, the parallel Needleman-Wunsch algorithm was used to ensure optimal alignment of DNA fragments. Numerical experiments demonstrated a significant performance improvement on different datasets: the best performance was 3.4 for inversions and 3.8 for translocations, achieved using eight threads. The results indicate an improvement in performance with increasing amounts of input data and the possibility of further optimization of the results obtained through the development of multi-core computing systems. To further improve the efficiency of the developed algorithm, we can consider its implementation on a graphics processing unit using CUDA technology, which will potentially provide even higher data processing speed.

5. Acknowledgements

6. Declaration on Generative AI

During the preparation of this paper, the authors utilized Grammarly to verify spelling and grammar accuracy.

7. References

[1]   S. Dias *et al.*, "Translocations and inversions: major chromosomal rearrangements during Vigna (Leguminosae) evolution," *Theor. Appl. Genet.*, vol. 137, no. 1, p. 29, Jan. 2024, doi: 10.1007/s00122-024-04546-8.
[2]   A. E. E.-D. Rashed, H. M. Amer, M. El-Seddek, and H. E.-D. Moustafa, "Sequence Alignment Using Machine Learning-Based Needleman–Wunsch Algorithm," *IEEE Access*, vol. 9, pp. 109522–109535, 2021, doi: 10.1109/ACCESS.2021.3100408.
[3]   L. Mochurad, A. Sydor, and O. Ratinskiy, "A fast parallelized DBSCAN algorithm based on OpenMp for detection of criminals on streaming services," *Front. Big Data*, vol. 6, p. 1292923, Oct. 2023, doi: 10.3389/fdata.2023.1292923.
[4]   L. Mochurad, "Implementation and analysis of a parallel kalman filter algorithm for lidar localization based on CUDA technology," *Front. Robot. AI*, vol. 11, p. 1341689, Feb. 2024, doi: 10.3389/frobt.2024.1341689.
[5]   V. Kovtun, T. Altameem, M. Al-Maitah, and W. Kempa, "Simple statistical tests selection based parallel computating method ensures the guaranteed global extremum identification," *J. King Saud Univ. - Sci.*, vol. 36, no. 5, p. 103165, May 2024, doi: 10.1016/j.jksus.2024.103165.

[6]   S. Saif, P. Das, S. Biswas, S. Khan, M. A. Haq, and V. Kovtun, "A secure data transmission framework for IoT enabled healthcare," *Heliyon*, vol. 10, no. 16, p. e36269, Aug. 2024, doi: 10.1016/j.heliyon.2024.e36269.

[7]   V. V. Shkarupylo, I. V. Blinov, A. A. Chemeris, V. V. Dusheba, and J. A. J. Alsayaydeh, "On Applicability of Model Checking Technique in Power Systems and Electric Power Industry," in *Systems, Decision and Control in Energy III*, vol. 399, A. Zaporozhets, Ed., in Studies in Systems, Decision and Control, vol. 399. , Cham: Springer International Publishing, 2022, pp. 3–21. doi: 10.1007/978-3-030-87675-3_1.

[8]   P. D. Stenson *et al.*, "The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting," *Hum. Genet.*, vol. 139, no. 10, pp. 1197–1207, Oct. 2020, doi: 10.1007/s00439-020-02199-3.

[9]   S. Banoon, T. Salih, and A. Ghasemian, "Genetic Mutations and Major Human Disorders: A Review," *Egypt. J. Chem.*, vol. 0, no. 0, pp. 0–0, Oct. 2021, doi: 10.21608/ejchem.2021.98178.4575.

[10]   M. J. McKenna *et al.*, "Chromosome Translocations, Inversions and Telomere Length for Retrospective Biodosimetry on Exposed U.S. Atomic Veterans," *Radiat. Res.*, vol. 191, no. 4, p. 311, Feb. 2019, doi: 10.1667/RR15240.1.

[11]   J. Kweon, H.-Y. Hwang, H. Ryu, A.-H. Jang, D. Kim, and Y. Kim, "Targeted genomic translocations and inversions generated using a paired prime editing strategy," *Mol. Ther.*, vol. 31, no. 1, pp. 249–259, Jan. 2023, doi: 10.1016/j.ymthe.2022.09.008.

[12]   X. Zeng *et al.*, "Gene sequencing and result analysis of balanced translocation carriers by third-generation gene sequencing technology," *Sci. Rep.*, vol. 13, no. 1, p. 7004, Apr. 2023, doi: 10.1038/s41598-022-20356-8.

[13]   Y. Xia, X. Yuan, W. Luo, S. Yuan, and X. Zeng, "The Origin and Evolution of Chromosomal Reciprocal Translocation in Quasipaa boulengeri (Anura, Dicroglossidae)," *Front. Genet.*, vol. 10, p. 1364, Jan. 2020, doi: 10.3389/fgene.2019.01364.

[14]   K. Huang and L. H. Rieseberg, "Frequency, Origins, and Evolutionary Role of Chromosomal Inversions in Plants," *Front. Plant Sci.*, vol. 11, p. 296, Mar. 2020, doi: 10.3389/fpls.2020.00296.

[15]   N. F. Binti Abdul Rahim *et al.*, "Channel Congestion Control in VANET for Safety and Non-Safety Communication: A Review," in *2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, Kedah, Malaysia: IEEE, Dec. 2021, pp. 1–6. doi: 10.1109/ICRAIE52900.2021.9704017.

[16]   I. Fedorchenko, A. Oliinyk, J. A. J. Alsayaydeh, A. Kharchenko, A. Stepanenko, and V. Shkarupylo, "MODIFIED GENETIC ALGORITHM TO DETERMINE THE LOCATION OF THE DISTRIBUTION POWER SUPPLY NETWORKS IN THE CITY," Dec. 2020, doi: 10.5281/ZENODO.5163692.

[17]   Nagesh Singh Chauhan, "DNA sequence dataset." 2021. [Online]. Available: https://www.kaggle.com/datasets/nageshsingh/dna-sequence-dataset

[18]   Z. Xia *et al.*, "A Review of Parallel Implementations for the Smith–Waterman Algorithm," *Interdiscip. Sci. Comput. Life Sci.*, vol. 14, no. 1, pp. 1–14, Mar. 2022, doi: 10.1007/s12539-021-00473-0.