

A survey for the methods of detection and classification of genetic mutations

Rana Hikmet Saloom¹, Hussein K. Khafaji²

¹Department of Computer Science, Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers and Informatics, Baghdad, Iraq

²Department of Computer Communications Engineering, Al-Rafidain University College, Baghdad, Iraq

Article Info

Article history:

Received Jun 22, 2022

Revised Sep 1, 2022

Accepted Sep 16, 2022

Keywords:

Alignment

Bioinformatics

Classification

Deoxyribose nucleic acid

sequences

Mutation

ABSTRACT

Research into pathogenic mutations is vital and useful for understanding illness progression, prognosis, and gene-disease connections. Furthermore, pathogenic mutations might have negative implications, such as the development of illnesses or medical problems. In this article, we critically review published studies that are concerned with the detection of genetic mutations and their types using genomic data. Using the reporting items for systematic reviews and meta-analysis criteria, a complete search was conducted on IEEE, Scopus, the Web of Science, Google Scholar, and Elsevier. In our review, we included 73 papers out of a total of more than 150 that were initially discovered. The most common data types used to detect and predict models are deoxyribose nucleic acid (DNA) and protein sequencing data. The examined models have a good level of accuracy. We devised a methodology for developing a detective and predictive model that takes into account all stages of mutation identification and classification.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Rana Hikmet Tobia

Department of Computer Science, Informatics Institute for Postgraduate Studies

Iraqi Commission for Computers and Informatics

Baghdad, Iraq

Email: Phd202020565@iips.icci.edu.iq

1. INTRODUCTION

The massive data generated by genome sequencing, such as deoxyribose nucleic acid (DNA), Ribonucleic acid (RNA), and protein sequences, has resulted in an explosive and diverse growth in biological data, necessitating the use of intensive computing and big data analysis techniques to store, organize, analyze, and integrate it [1]. The data requires a complicated mathematical analysis to understand biological data using the multidisciplinary study of computer science and information technology known as bioinformatics or computational biology [2]. Bioinformatics has made a significant contribution to medical research in recent years. Bioinformatics data, such as genomic sequence analysis, offers a high potential for detecting any mutation in the human genome, paving the path for their correction utilizing cutting-edge biological technology [3]. Computational genomics today does a massive database search of genetic data. The challenge is obtaining optimal alignments by extracting information from various databases. The alignment score is a crucial statistic for determining how similar two DNA sequences are [4].

Sequence alignment is required to detect the comparable nature of biological molecules (e.g., proteins, DNA, or RNA), which is a key challenge in bioinformatics. It aids in the formation of many biological molecular interactions. The sequence alignment may be divided into two broad categories. By matching two biological sequences, pair-wise sequence alignment (PSA) is used to discover the greatest extent of similarity, allowing one to assess the degree of similarity and the potential for homology. Multiple sequence alignment (MSA) is a sort of

sequence alignment that involves aligning three or more biological sequences to determine an evolutionary relationship between the query sequences [5]. Since the 1970s, specialists have been extensively studying mutation testing, and significant breakthroughs have been made in terms of ideas, theory, technology, and empirical data. The existing literature studies have summarized the most significant realizations, but we don't know how mutation testing is used. Our objective is to identify and characterize the most common mutation testing applications, as well as to assess the level of replicability of empirical mutation testing investigations.

2. BIOINFORMATICS

Bioinformatics is an interdisciplinary topic that mixes programming and science to develop approaches and algorithms for analyzing and comprehending biological data. It arose from the necessity for rules for huge data collected by atomic physics, and it now plays an essential role in the disciplines of genetics and bioinformatics [6]. Bioinformatics combines biology, computer science, and information technology into one field with three sub-disciplines. These include analysis of large data sets, interpretation and analysis of nucleotides, amino acids, and protein structures, and development of tools that aid in effective access to various types of information [7]. Bioinformatics involves the analysis of huge amounts of genetic data. In order to analyze the genome, sequence data, and anticipate the structure and function of biological macromolecules that are utilized in the construction of the genome, the multidisciplinary area of bioinformatics collaborates with the biological sciences, statistics, and information technology [8]. Finding suitable methods to determine the similarities between DNA sequences is one of the most difficult tasks. Computers are designed specifically for this purpose and made available to a large number of academics all over the world [9].

3. BIOLOGICAL SEQUENCES ANALYSIS

In bioinformatics and contemporary biology, biological sequence analysis compares, aligns, indexes, and analyzes biological sequences. The fundamental level of biological information is DNA, RNA, and protein sequences. Below we will discuss the most important aspects of the biological sequence in detail to reach the central dogma of living organisms.

3.1. Deoxyribose nucleic acid (DNA)

DNA is composed of two strands that are twisted together to form a double helix, as well as four nitrogenous bases (Adenine, Thymine, Cytosine, and Guanine), a sugar molecule called deoxyribose, and a phosphate molecule [10]. DNA sequencing is an essential technique in biomedical research because it provides access to crucial and hugely encoded information in the human genome. It's possible to store a tremendous amount of information in one gram of DNA, approximately 215 million terabytes per gram. This is why it's critical to create more powerful DNA sequence analysis tools and algorithms [11], [12].

In DNA sequencing analysis, the following biological words are used [13]:

- a) Codons: These are triplicate groupings of three nucleotides.
- b) Genes: DNA is made up of a collection of genes that work together as a unit. A single gene can be encoded sequences known as Exons or non-coding sequences known as Introns, and it can contain a specific process, a set of data, or protein-coding.
- c) Chromosome: Genes, nucleotide sequences, and other regulatory components are wrapped around protein molecules to form giant DNA structures called chromosomes.
- d) Genome: A genome is a collection of creatures that contain DNA, chromosomes, genes, and nucleotides.

3.2. Ribonucleic acid (RNA)

Ribonucleic acid (RNA) is a nucleotide polymer that looks a lot like DNA. However, unlike DNA, RNA is normally found as a single strand and hence can establish base-pair connections with itself. messenger RNA (mRNA) is the most well-known kind of RNA, which conveys a copy of the genetic material from the nucleolus for translation into proteins. Non-coding RNA (ncRNA) strands, which do not code for protein machinery but become functional cellular machinery themselves, are also essential RNA examples. Transfer RNA (tRNA), ribosomal RNA (rRNA), and human telomerase RNA are examples of this kind of RNA (hTR). Adenine, Cytosine, Guanine, and Uracil are the four base nucleotides that make up RNA molecules. Furthermore, RNA is a key component of telomerase, a protein involved in DNA replication maintenance [14].

3.3. Protein

Proteins are one of the most significant counterparts in our bodies since they govern cell functions. Proteins, in particular, play a role in a variety of interactions, and we may control the processes in humans' bodies by permitting or disallowing certain activities [15]. Proteins are the cell's workhorses. The role of a protein and the biological function in which it is engaged is critical knowledge for extracting knowledge

about identifying disease-specific proteins, discovering novel medications, developing high-productivity, pest-resistant crops, and so on [16]. The biomedical and pharmaceutical industries both benefit from a better understanding of protein activity at the molecular level [17]. Protein annotation, for example, makes it easier to create new tools for disease prevention, diagnosis, and therapy [18]. Researchers can develop tests to describe a protein's function (for example, an assay to evaluate the execution of a specific biological function and see if the protein plays a role in such executions). Knowing the diversity and breadth of the protein universe will be beneficial, and because of recent breakthroughs in sequencing technology, the number of genomic sequences gathered is expanding at an exponential rate [19].

4. THE CENTRAL DOGMA

The process is analogous to DNA transcription, splicing, and RNA translation in living organisms, and it is the central tenet of molecular biology. The following are concise descriptions of the transcription, splicing, and translation processes:

- a) Transcription and splicing: A DNA sequence that makes up a gene is read from the promoter (beginning position) to the end. According to particular tags, non-coding sections (Intron) are deleted, and the remaining coding areas (Extron) are reconnected and capped. The sequence is then transcribed into a single-stranded mRNA sequence (messenger RNA). mRNA is transported from the nucleus to the cytoplasm.
- b) Translation: As the protein is created, the mRNA sequence is translated into an amino acid sequence. During translation, the ribosome reads the segment starting from certain three bases, then reads three bases (a codon) from the mRNA at a time and translates them into one amino acid; there are also specific finishing three bases to signify the end of the translation. Figure 1 illustrates the process of transcription and translation.

In essence, introns are cut out and exons are maintained throughout the transcription and splicing phases to generate mRNA, which will do the translation task. Codons are translated into amino acids according to the genetic code table throughout the translation process [20]. Figure 2 illustrates the genetic code table.

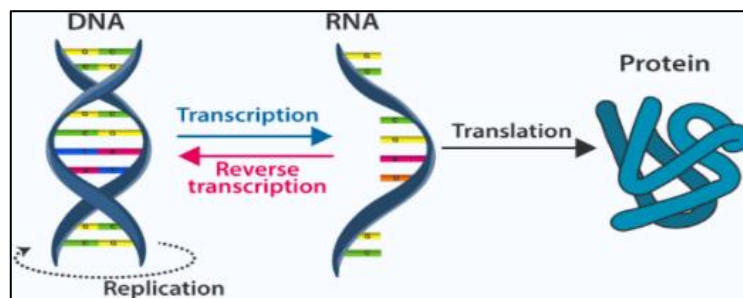


Figure 1. Transcription and translation

		Second letter					
		U	C	A	G		
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA } Stop UAG } Stop	UGU } Cys UGC } UGA } Stop UGG } Trp	U C A G	
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G	
	A	AUU } AUC } Ile AUA } AUG } Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G	
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G	

Figure 2. Genetic code table

5. MUTATIONS

Variations aren't created by accident. All of the differences we see in today's population are the product of previous mutations in chromosomes that are handed down to the next generation. A genetic mutation is an alteration in a person's DNA code that might harm them [21]. One alteration in one of the nucleotide bases in the DNA sequence causes the most prevalent form of genetic variation [22]. A pathogenic mutation has a detrimental impact on genes and their functioning, resulting in a disease state or genetic disorder [23]. Sickle cell anemia, cystic fibrosis, and local hemochromatosis are all diseases caused by genetic mutations [24].

There are two primary classifications for genetic mutations: Genetic mutations inherited from parents that are present throughout a person's life in every cell in his body and are referred to as germline mutations since they occur in the parents' germ cells [25]. The other type of genetic mutation is an acquired mutation, which occurs in certain cells over a person's lifetime and is present in particular cells [26]. This form of mutation happens when there are errors in DNA transcription during cell division, or when there is a component in the environment or radiation [27]. Missense, insertion, duplication, deletion, nonsense, and frameshifting are some of the other forms of mutations. These include the significant impacts of genetic mutations, which are shown in extremely lethal illnesses, such as cancer [28].

6. DNA SEQUENCES ALIGNMENT

DNA is the genetic substance that allows genetic information to be passed along from generation to generation. Organisms split from their predecessors throughout time as their DNA changes due to evolution. As a result, one of the most significant requirements in biological research is the examination of live organisms' DNA sequences. The method of establishing the precise sequence of nucleotides in a nucleic acid molecule is known as DNA sequence alignment [29], and it is frequently used in bioinformatics to determine the molecular sequence of an unknown DNA sequence preserved by natural selection in evolution [30]. Depending on the application, there are a variety of approaches to align DNA sequences, such as sequencing with near-universal alignment in NGS and comparing DNA sequences using local, global, and multiple sequence alignment [31]. Because DNA is sorted from least symmetrical to comparable sequences according to the degree of alignment, the alignment of two DNA sequences is mathematically dependent on the degree of alignment computed using various methodologies. As a result, we may deduce that the high degree of similarity between the two sequences indicates that they are connected to evolution [32]. Figure 3 illustrates DNA sequence alignment.

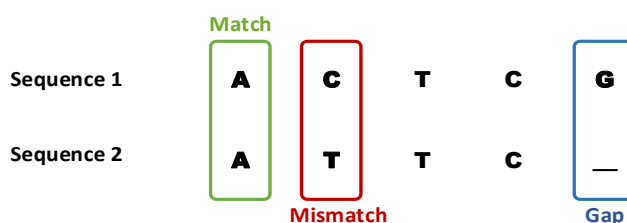


Figure 3. DNA sequences alignment

7. DNA MUTATION CLASSIFICATION

Even though a change in the nuclear acid sequence, or mutation, includes several categories, each with its own set of subcategories, a model to categorize mutations in the DNA sequence is required. At this point, the mutations detected in the DNA sequence are categorized using several classification algorithms based on their influence on the protein structure and function [33]. Based on their consequences, mutations can be divided into the following categories [34], [35]:

- a) Synonymous mutation: This sort of mutation happens when one of the nucleotides in the code that creates the same amino acid changes, as a result of several genetic codes referring to the same amino acid.
 - Missense mutation: A missense mutation occurs when one of the nucleotides changes, resulting in the creation of a completely different amino acid from the original amino acid.
 - Nonsense mutation: This form of mutation causes premature coding to be suspended due to a change in the nucleotide sequence in the DNA. This sort of alteration results in the production of a shortened protein, which is frequently non-functional.
 - Silent Mutation: Mutations in DNA that do not influence the organism's phenotype.
- b) Frameshift mutation: This form of mutation arises when several nucleotides in a DNA sequence are deleted or inserted.

The classification's goal is to create a model from the training dataset that can predict classes for fresh, unknown samples. Because of the non-numerical nature of biological sequence components, various sequence lengths, and other factors, classifying biological sequences is a tough task. Biological sequence categorization is the prediction of the kind of DNA sequence based on structural or functional similarities, followed by the prediction of the DNA sequence's function and relationships with other sequences to identify genes within DNA molecules [36], [37]. The procedure of picking characteristics is the most challenging challenge, as the features of the DNA sequence are difficult to collect, and the technique of representing it in general poses high-dimensional challenges [37], [38]. Figure 4 illustrates the types of mutations.

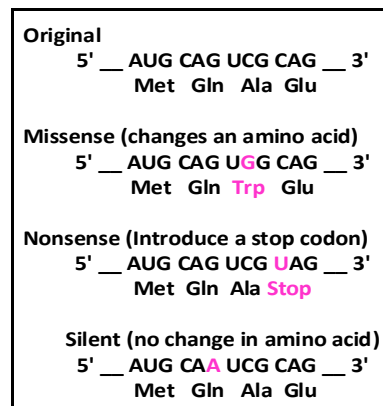


Figure 4. Type of mutations

8. LITERATURE SURVEY

In this section, we will discuss prior research on understanding the DNA sequence and how to detect genetic mutations that can arise within it, as well as research on categorizing genetic mutations inside the DNA sequence. In 2013, Yost *et al.* [38] developed a technique for identifying somatic variations using a sequencing analysis pipeline. To compare the methodology to previous mutation detection methods, two datasets are employed for experimentation: a collection of 80 tumor-normal spiked-in (TNS) pairs obtained from 38 different normal germline DNA samples and a combination of 8 normal DNA (MIX). They compared Mutoscope to other mutation callers using sequencing data generated from a mix of 8 normal DNA samples with genotypes that are known (MIX sample), resulting in somatic mutations with varied allelic proportions. The Mutoscope has been designed to maximize detection of the mutation and improve the accuracy of somatic mutation identification at low allelic fractions. In 2013, Cibulskis *et al.* [39] introduced MuTect, an approach that uses a Bayesian classifier to discover somatic mutations with very low allele fractions with only a few supporting reads and good specificity. Figure 5 illustrates the detection of a somatic point mutation using MuTect.

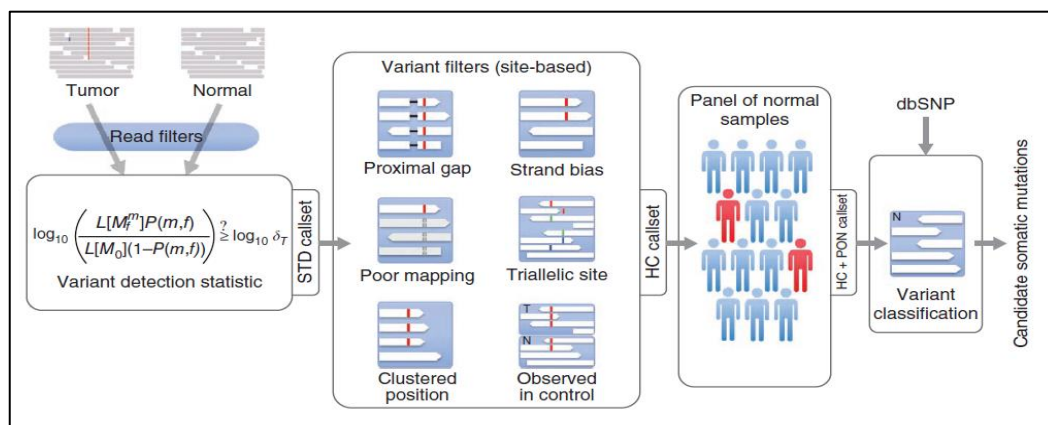


Figure 5. Detection of a somatic point mutation using MuTect [39]

In 2014, Saini and Dewan [40] proposed a unique wavelet-based graphical technique to discover single base alterations in the H5N1 influenza A virus. To pinpoint the areas of base alterations, the study examines the graphical plots of wavelet-transformed Hemagglutinin (HA) and Neuraminidase (NA) nucleotide sequences. The signal data, which may be utilized as a foundation for medication creation and the development of novel diagnoses, are a large number of previously sequenced genomic data that were collected from all over the world during epidemics. This technique helps in giving reliable and quicker findings for these data.

In 2015, Garai and Chowdhury [41] offered a novel sequence alignment approach that relies on a genetic algorithm (GA) for selecting the best alignment score between two sequences, which might be DNA or protein sequences. The suggested genetic-based technique, dubbed cascaded pairwise alignment with genetic algorithm (CPAGA), reduces the search space needed by dividing a big space into smaller subspaces. Before beginning the alignment operation, the sequence pair is decomposed into numerous parts. Even for longer sequences, such reduction improves the search process' capacity to find the global or near-global optimal solution. Several DNA and protein sequence pairings were used to test the approach. Also, they compared the CPAGA's alignment score to that of a few well-known and useful alignment approaches. Figure 6 illustrates initial sequence encoding in a population and pairwise aligning. A set of non-parametric statistical methodologies was used to evaluate CPAGA method performance, the results showed that CPAGA outperformed the other alignment processes.

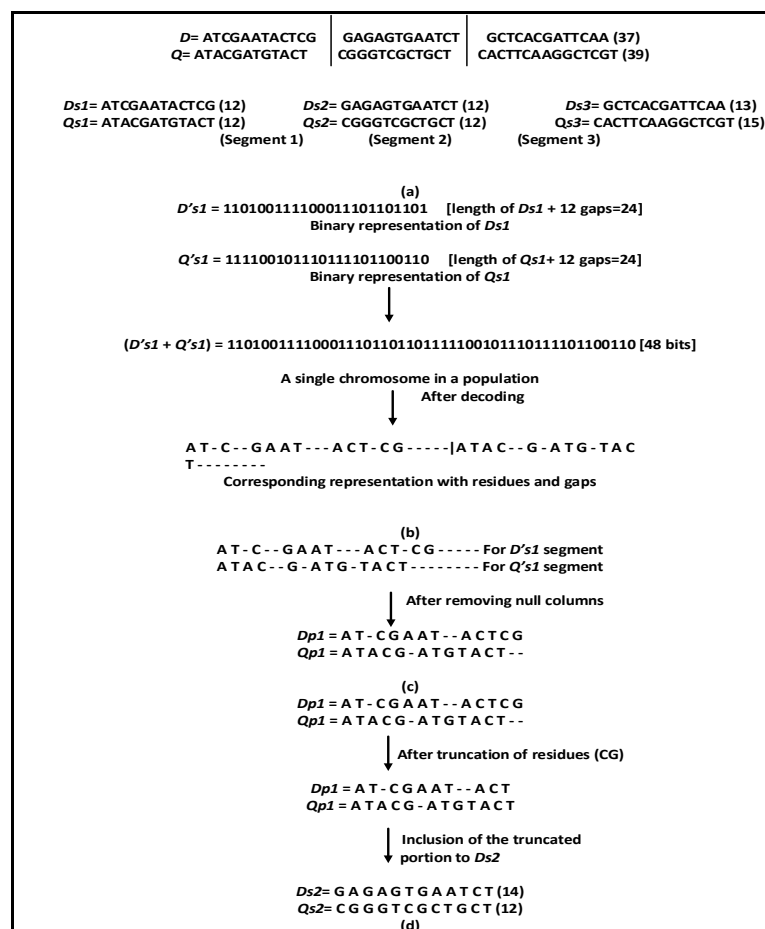


Figure 6. Pairwise alignment and initial sequence decoding in a population

In 2015, Lee *et al.* [42] proposed BulkAligner, an approach to a graph-based in-memory trinity distributed system. To get graph-form data, they trim the reference sequence k times (where k = the length of 134 sequence fragments) and divide each trimmed sequence into k -mer sequence fragments. To build a reference graph for each slave, they convert the reference sequence data into a graph representational data format similar to the de-Bruijn graph, as shown in Figure 7. BulkAligner has a throughput of at least 1.8 and up to 57 times faster than existing Hadoop techniques while maintaining the same or greater quality.

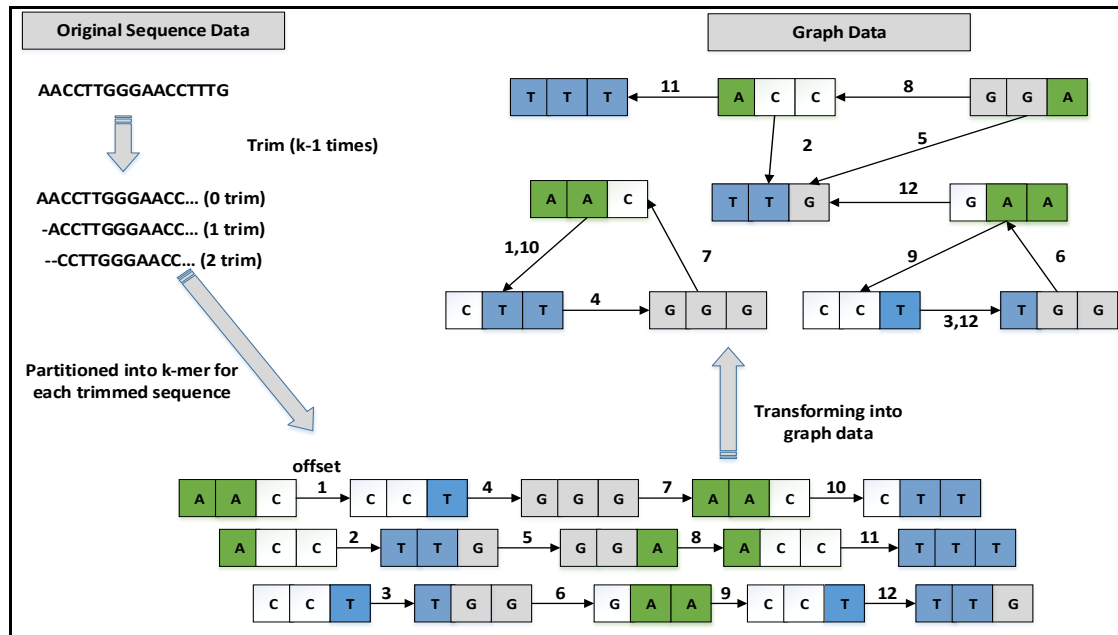


Figure 7. Converting the original sequence data to a graph data

In 2016, Abdullah *et al.* [43] proposed a systematic examination of how single nucleotide alterations in genetic codons might impact the amino acid residue produced. As the result of the substitution of basemutations producing alterations at the level of translation, a probability matrix was created to depict probable modifications and paths likely to be taken. They found that a certain group of amino acids is more likely to result from SNSs than a random sequence of amino acid substitutions.

In 2016, Anitha and Pushpa [44] suggested using waikato environment for knowledge analysis (Weka) platform to perform sequence alignment and extract important information from human chromosomes. A variety of filter techniques and clustering methods were used to combine chromosomes of animals with comparable associations. DNA clustering may be used to determine superfamily, family, and subfamily connections in DNA sequences at a cheap computational cost. This approach determines how closely the two sequences match, and which parts don't match. It is advantageous in the vast majority of scenarios when a lower computational load on the CPU is desired and reduced system memory utilization is desired.

In 2016, Yang *et al.* [45] improved the Smith-Waterman algorithm by introducing two-character alignment and improving the scoring function and matrix to increase specificity and minimize unpredictability. Changing the total memory of the score matrix to two-line storage dramatically decreases the complexity of the software space. Researchers have improved the rating matrix for HIV-1 infected individuals with a known transmission chain. They emphasize the smith-waterman technique for local alignment of DNA pairs. This algorithm considers the substitution of "similar" characters, which might increase the specificity and minimize the unpredictability.

In 2016, Rani and Ramyachitra [46] used a hybrid algorithm consisting of an artificial bee colony (GA-ABC) and a bacterial foraging optimization algorithm. The BALiBASE 3.0 benchmark dataset was used to compare the proposed technique against previous approaches. A multi-bacterial foraging optimization technique (MO-BFO) was compared to popular MSA methods such as Clustal Omega, Kalign, MUSCLE, MAFFT, and the hybrid genetic algorithm with artificial bee colony (GA-ABC) and ant colony optimization (ACO). Using GA-ABC, conserved blocks could not be retrieved. The alignment then is done with BFO, and the preserved blocks were retrieved.

In 2016, Kaghed *et al.* [47] applied heuristic approaches such as GA to find approximate solutions to multiple sequence alignment issues. The designed genetic algorithm makes use of a new assessment procedure, high mutation probability, new crossover operators, and a robust termination condition. The proposed genetic algorithm (GA) is capable of discovering excellent multiple sequence alignment with a low computing complexity. Multiple sequence alignment employing a genetic algorithm is used in our daily lives for a variety of biological purposes. Figure 8 depicts the created genetic algorithm that was proposed in this study.

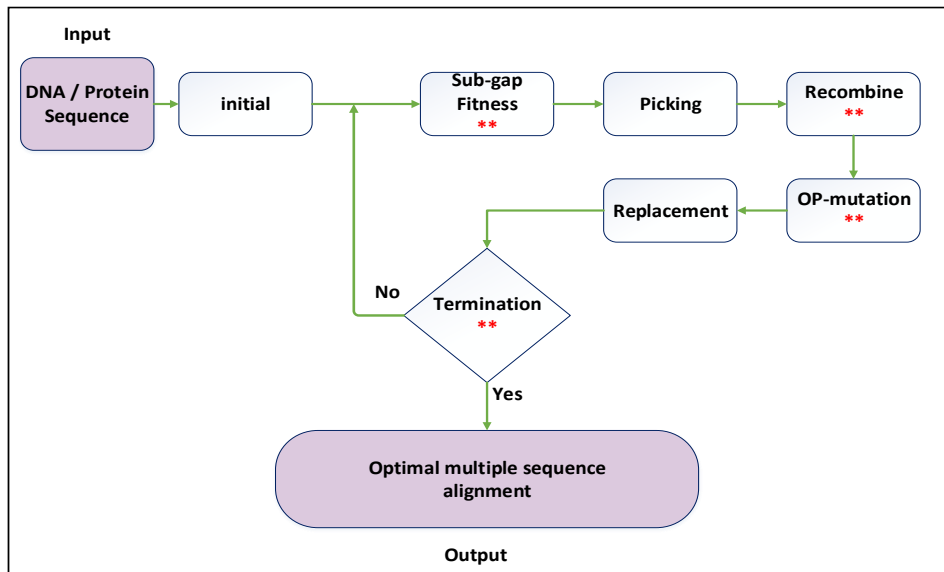


Figure 8. Block diagram of the suggested system in [47]

A memory-efficient pair-wise DNA sequence alignment approach was revealed by Ray *et al.* [48] in 2016. A unique notion of a pointing matrix is used to ensure the faster and more precise discovery of the best alignment. Using a dynamic programming paradigm, the suggested method discovers the best global alignment. The technique requires substantially less space for long DNA sequences than the Needleman-Wunsch approach. The entire procedure was tested on random100 faux samples, which produced DNA sequence pairings in ten different scenarios. The method took slightly longer to develop (9%-11%), but it took 34%-42% less time to find the best alignment than another method. Figure 9 illustrates the suggested alignment method's process block diagram.



Figure 9. Block diagram of the suggested alignment method's process

In 2016, Ismaeel and Mikhail [49] offered a concept for achieving a data mining, DM, technique using neural networks and large datasets. To address the disadvantages of prior methodologies, it also provides friendly forecasts, flexibility, and successfully categorized malignancies. This suggested strategy uses two approaches: firstly, bioinformatics algorithms such as BLAST, CLUSTALW, and others to determine whether or not there are malicious mutations; The other is DM with a neural network, which was chosen from 12 of the 53 TP53 gene database variables. The suggested data mining approach allows for a wide range of diagnostic and prediction options. It also classifies tumors based on alterations in the P53 sequence of the tumor protein. With the greatest performance, the backpropagation neural network algorithm utilizes the mean square error. Table 1 shows a comparison of two different methods with an effective data mining technique.

In 2017, Yang *et al.* [50] used colored petri net (CPN) for modeling and identifying the mutation type categories. The suggested methodology compares the bases of DNA strands to identify the mutation position and rate and then compares amino acid codons throughout the polypeptide chain to detect the mutation type. The model is useful for determining if the alternatives have an impact on the structure and function of proteins. The suggested model's usefulness and accuracy are demonstrated using biological examples and it is useful for determining if the alterations influence the function and structure of proteins. Figure 10 illustrates DNA mutation type categorization using the CPN model.

Table 1. Comparison of two different methods with an effective data mining technique [51], [52]

Features	The proposed method	Amin <i>et al.</i> [51]	Ismaeel and Yousif [52]
The aim	Specific cancer prediction, diagnosis, and categorization	Heart disease prediction and diagnosis	Position of mutation categorization, prediction, and diagnosis
General technique	Using two methods, including introducing a new data column to the database.	X	√
Tp53 genome included	√	X	√
DNA and Protein Analysis	√	X	√
Check for sequence homology	√	X	√
The database utilized	TP53 mutation database at UMD	Zurvey by the American Heart Association	TP53 mutation database at UMD
utilized method	Tools for bioinformatics and data mining (Back propagation algorithm) trainlm	Data mining strategies (neural network and genetic algorithm) trainlm	Fast Back propagation algorithm, bioinformatics equipment
Update weight functions			Quick back Propagation Network
Topology of ANN	(11-100-1)	(12-10-2)	(283-141-1)
Performances	0.1E10-13	0.034683	0.000006
Software	MATLAB R2015a	MATLAB R2012a	Neuro Intelligence Alyuda
In support of	Biomedical engineers, scientists, and bioinformatics physicians	Clinicians	Scientists

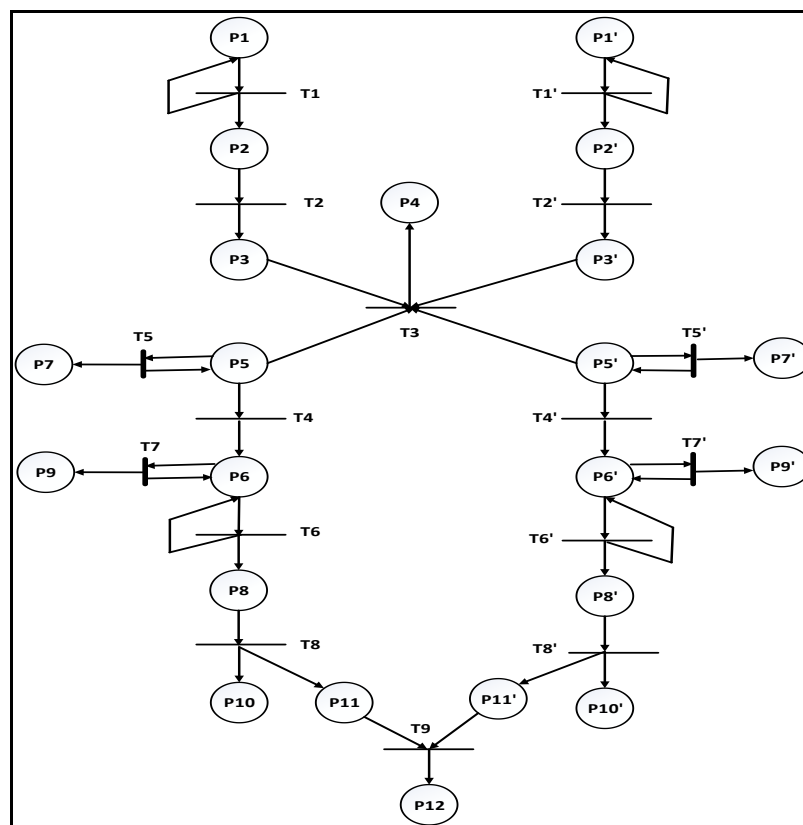


Figure 10. DNA mutation type categorization using the CPN model

In 2017, Kaghed *et al.* [53] compared the implementation of dynamic programming and the performance of the genetic algorithm (GA) implementation. Sequences of deoxyribonucleic acid (DNA) and protein sequences have been utilized. The results showed that the use of a genetic algorithm is preferable in comparison with the dynamic programming approach. The execution time and storage capacity are increased by increasing the lengths of the utilized sequences OR/AND expanding the number of the used sequences. Table 2 summarizes the observations of two sequence alignment techniques based on the two approaches.

Table 2. The summary of dynamic programming and genetic algorithm observations

Technique	Pros	Cons
Dynamic Programming	It is mathematically guaranteed to deliver the best alignment for a given collection of scoring functions.	However, because this solution needs time proportional to the product of the sequence lengths, it has exponential temporal complexity. Due to the enormous number of computing steps, it gets sluggish. As alignment sequences become larger, so does the memory need.
Genetic Algorithm	MSA's accuracy is improved with GA. These can be done several times or till convergence occurs. It is possible to use it to provide approximate results to the MSA issue. Utilizing a minimal quantity of computing resources	Because of the tradeoff between speed and precision, GA outcomes are occasionally poor (either a low-quality solution or a fast convergence rate).

In 2018, Reeta *et al.* [54] suggested a method for ranking the list of sick genes found in autistic individuals. This is accomplished through the use of the Naive Bayesian classification technique. The Naive Bayesian technique is an effective classifier since it is simple to construct because each structure has its own priority and this technique can be used to predict autism in its early stages and is very effective at predicting autism. Figure 11 depicts the architecture for autism prediction and the outcomes of this study.

In 2018, Ramakrishnan *et al.* [55] explained that RLALIGN and RLALIGN will be beneficial for realignment operations that require optimizing sections of a broader alignment. Unlike traditional algorithms, RLALIGN is unaware of the scoring scheme's nature, allowing for easy application to a wide range of issue types. Figure 12 depicts the A3C block diagram. RLALIGN is an RL-based approach for solving the MSA issue. It can be taught to align moderate-length sequences accurately and scale to longer sequences using multiple algorithms. The results are on par with, if not better than, those produced by well-known alignment algorithms.

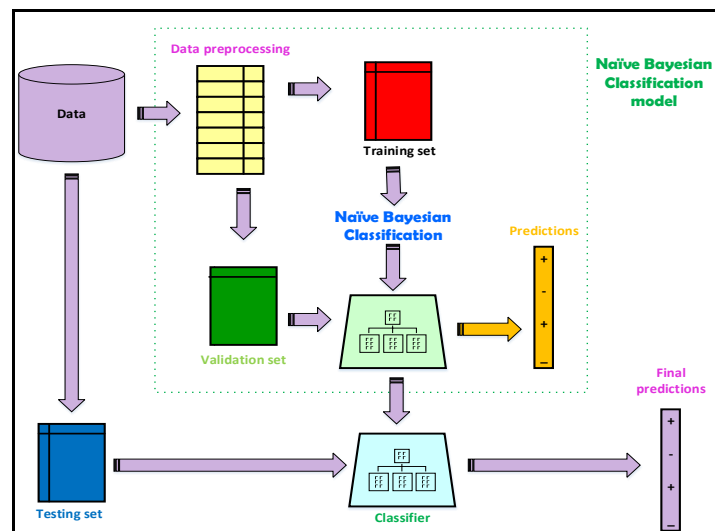


Figure 11. The architecture of autism prediction model

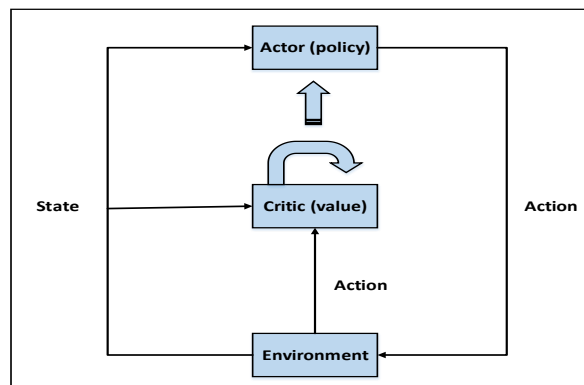


Figure 12. A3C block diagram

In 2018, Abedulridha and Al-Shamery [56] developed two approaches to tackle the pairwise sequence alignment problem. The first approach focuses on breaking DNA into parts with adaptive interleaving windows to separate the DNA tape into matched and non-matching regions. In the second approach, a multi-zone genetic algorithm (MZGA) is offered as an improved technique. A new crossover strategy based on cut-points and gap location are offered. The technique was tested on a real-world dataset of DNA with lengths ranging from 66 to 26.037 bases. The suggested approach met the DNA sequences' best alignment score. Figure 13 illustrates the block schematic of the proposed system, which is generalizable.

In 2018, Sun *et al.* [57] offered the pairwise alignment algorithm for very long sequences (PAAVLS) method. The algorithm is based on the Smith-Waterman algorithm's dynamic programming technique. It significantly decreases memory consumption and may be used to align very lengthy sequences. The algorithm finds six aligned sections between the two lncRNAs. The suggested PAAVLS method has a running time that is comparable to that of the classic Smith-Waterman algorithm.

In 2018, Wood *et al.* [58] created a machine-learning-based somatic mutation finding strategy called Cerebro. As shown in Figure 14, Cerebro identifies somatic mutations with high-confidence while limiting false positives using machine learning. A regular blood DNA sample was used to train the model and exome regions were collected and read twice using NGS techniques. This method revealed probable false-positive and false-negative changes in 74% of mutation calls in coupled tumor-normal genomic information from 1,368 the cancer genome atlas (TCGA) samples.

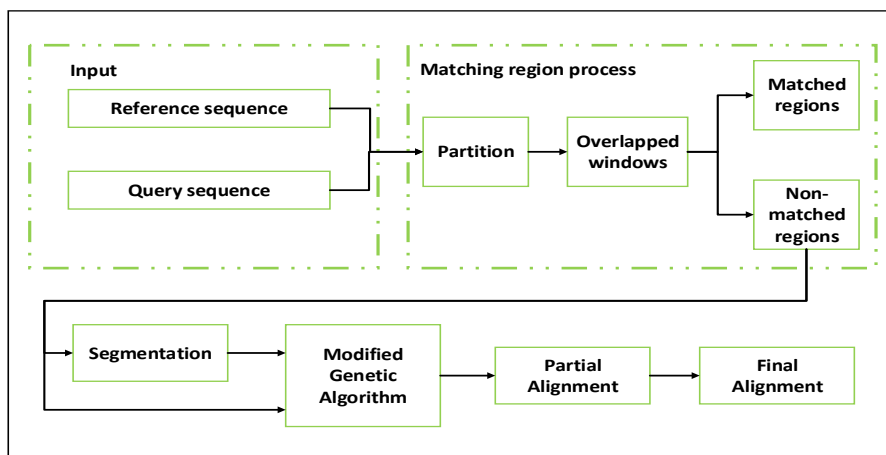


Figure 13. The block schematic of the proposed system

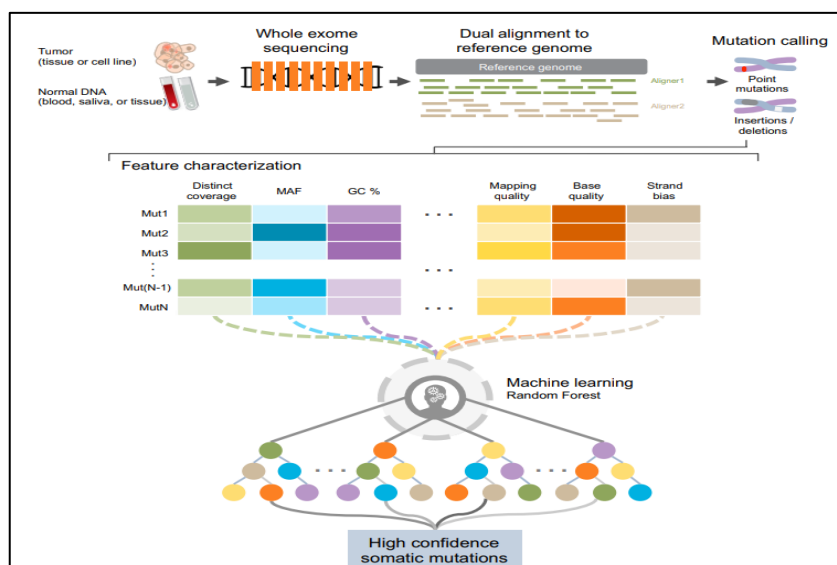


Figure 14. Cerebro overview for detecting somatic mutations [58]

In 2018, to align lengthy genomic sequences, Liao *et al.* [59] proposed an adaptively banded smith-waterman (ABSW) approach that is hardware compatible. In addition, dynamic overlapping, a heuristic approach for making overlap of a band of subsequences to increase accuracy, is presented. Also, suggest a hardware design of banded Smith-Waterman and traceback to allow ABSW hardware acceleration. A heuristic algorithm, dynamic overlapping, is proposed to improve the accuracy of alignment. Figure 15 depicts an overview of the proposed ABSW algorithm.

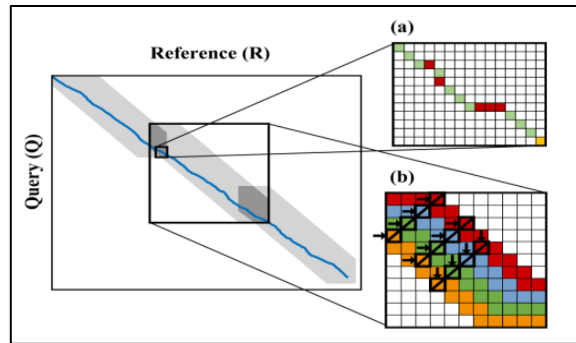


Figure 15. The proposed ABSW [58]

In 2019, Kar *et al.* [60] suggested an approach based on the DIT-FFT algorithm to locate the exonic region with the aid of integer value encoding for changing the DNA sequences. Period 3 components are read out from the output spectrum using digital filters, which also filter out undesired high-frequency noises from DNA sequences. To reduce background noise, introns, which are non-coding regions, are suppressed. On four genomic sequences, the proposed approach is evaluated with one or more exons. For the provided genomic data, the suggested approach is highly efficient, little processing time, and offers excellent accuracy.

Sudha and Vijaya [61] use supervised machine learning techniques to develop a model to identify syndromic ASD by identifying mutations that underpin these traits in 2019. Predictions were estimated using a tenfold cross-validation methodology. Figure 16 shows the proposed framework. A decision tree classifier outperformed previous learning algorithms in predicting autism-like behavior, with a 94% accuracy rate.

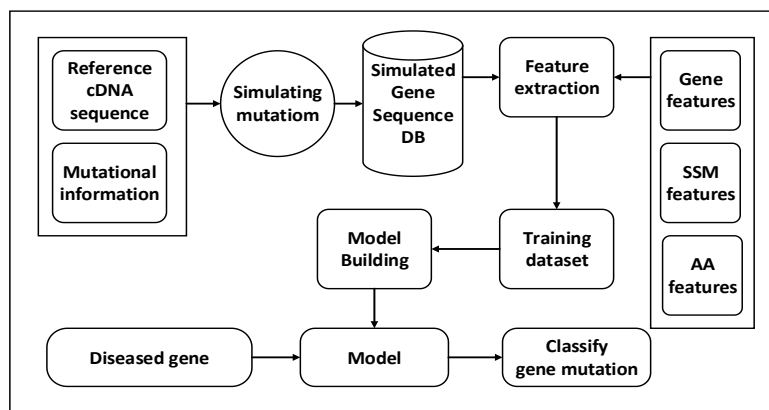


Figure 16. Proposed framework

In 2019, Irawan *et al.* [62] presented a method for identifying the mutation in a coronavirus' DNA sequence using the Needleman-Wunsch algorithm. This research performed an alignment and determined the locations of mutations in both sequences. The Needleman-Wunsch technique may be used to discover mutations in the DNA sequence. Tests were performed on 10 coronavirus DNA kinds of Infectious Bronchitis. The difference between the first and last types of viruses was 39%.

In 2019, Akshayaa *et al.* [63] proposed a model based on the one-shot method that has been developed for early identification and monitoring of DNA repair mutations. The goal of the project is to provide a one-shot mutation detection method for early detection of "DNA repairing point mutations" from

tissue biopsy pictures. The recommended tool's accuracy was found to be 70%, making it a promising model for predicting DNA repair mutations. Figure 17 shows the model structure.

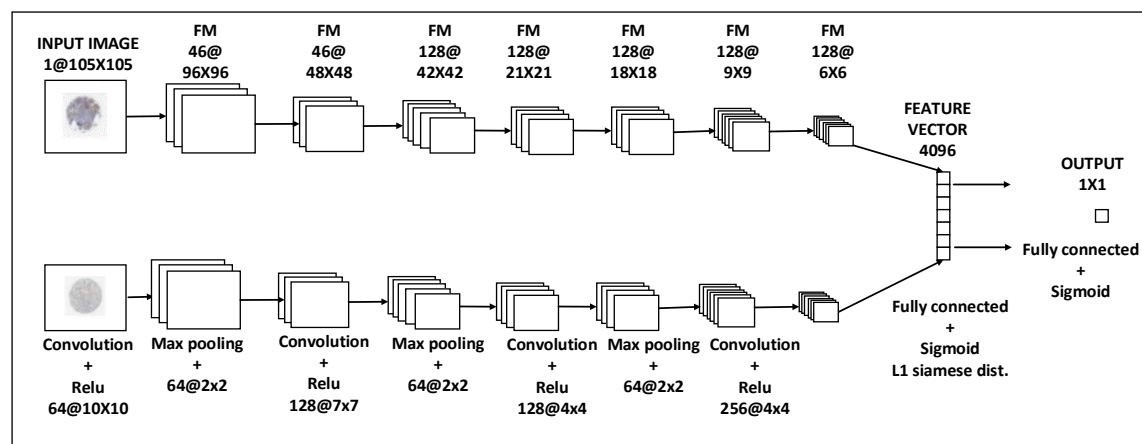


Figure 17. Proposed model structure

The computational hurdles in biological sequence processing are described by Gancheva and Georgiev [64] in 2019. This work investigates the effectiveness of sequence alignment using a concurrent multithreaded software version of the Needleman-Wunsch method. The building of the score matrix takes the greatest time and permits parallelization in the Needleman-Wunsch method. The existence of relationships between the matrix's members is a substantial barrier to parallelization. As a result, the procedure described below is applied:

- Count of threads they adopted the letter N.
- The matrix's initializations are carried out in parallel.
- The matrix is divided into N-row chunks.
- Compute the first N-1 elements of the first row, the first N-2 elements of a second row, and the first N-n elements of an nth row for each block.
- The next element here on the processed row is calculated by each thread.
- Calculate the remaining components of the row's ends.
- Steps should be repeated for every block of the matrix.
- trace steps backward.

Experimental estimates of execution time and speedup have been made to improve global sequence alignment as the number of cores rises, according to performance evaluations and scalability evaluations. In 2020, Rehmat *et al.* [65] introduced a new technique to understand the implicit properties that produce pathogenic variants. "NLP-SNPPred" consumes biomedical literature and generates vector representations. As shown in Figure 18, these representations are placed into machine learning models to determine which alterations are pathogenic and which are neutral as shown in Figure 18. Their findings indicate that NLP can accurately predict the functional significance of protein-coding changes with only a few additional biological parameters.

In 2020, Yilmaz [66] looked at the use of DNA sequences to predict mutation susceptibility. Using dna2vec, they first constructed word vectors for the human and mouse genomes. Their finding revealed that for a specified k-mer, the k-mers with the highest cosine similarity are those with the highest mutation count. Dna2vec and other embedding algorithms may be used to show mutation or variation properties of genomes using only the genome sequence. This might help researchers better understand the process or dynamics of mutations in the genome, as well as provide insights into how genes change over time.

In 2020, Delibas *et al.* [67] developed a top-k n-gram matches-based alignment-free DNA sequence alignment method of investigation. The method's phylogenetic relationships demonstrate that trees are almost identical to MEGA software's results. Their findings reveal that a small number of common sequence patterns may be used to describe DNA sequences. The capacity of the approach to exhibit effectiveness in sequences of various lengths can also be shown in Table 3. The MEGA7 program analysis time is based on the computation times for the ClustalW and Muscle alignment-based approaches.

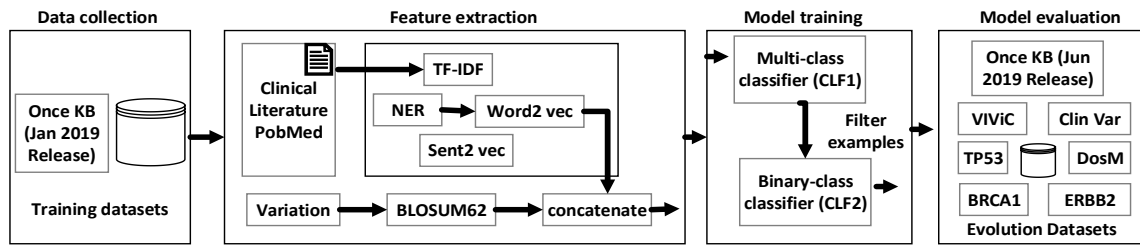


Figure 18. A brief description of the suggested technique in [65]

Table 3. The suggested approach and alignment-based methods comparison

Dataset	Computation times (sec)		
	Top-k n-gram	ClustalW	Muscle
NCBI genomic information for NADH dehydrogenase subunit 4 gene from 12 species	0.19	6.60	2.02
13 bacteria's 16S ribosomal DNA	0.15	17.70	5.85
NCBI database information on the mitochondrial genomes of 18 eutherian animals.	16.26	4528.05	2877.58

In 2020, Berman *et al.* [68] devised a unique machine learning framework that accurately predicts gene mutations by combining generative adversarial networks (GANs) with recurrent neural networks (RNNs). For this deep learning approach, influenza viral sequences were chosen as an excellent test case. MutaGAN generated sequences with a Levenshtein median distance of 2 amino acid residues from a given "parent" protein sequence. It was able to add at least one mutation from the worldwide influenza viral population to the majority of parent proteins. These findings indicate the MutaGAN framework's ability to help with pathogen forecasting. Figure 19 illustrates the architecture of MutaGAN Framework

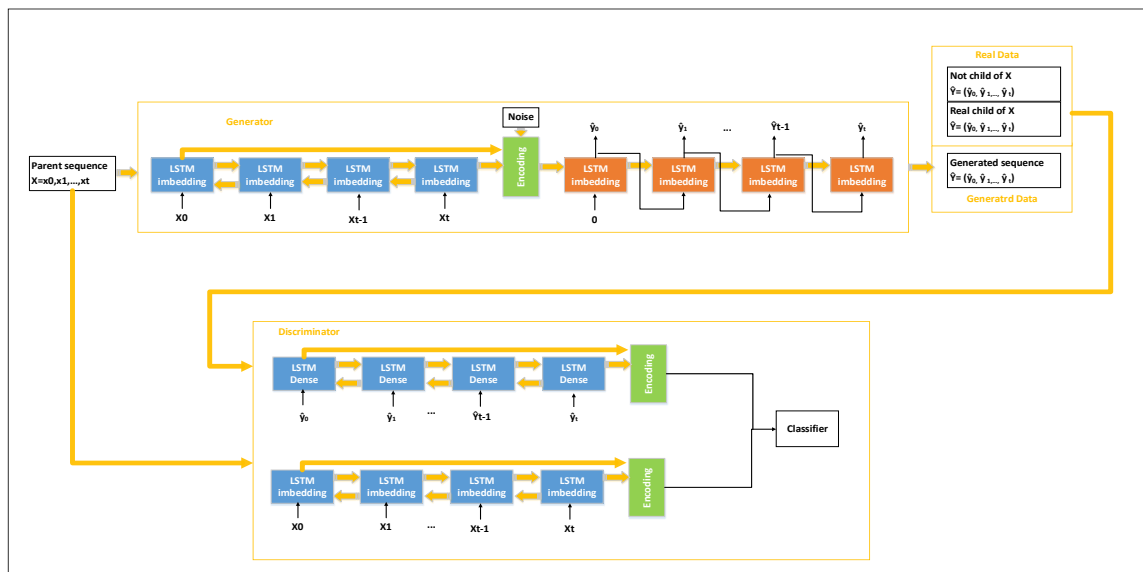


Figure 19. The architecture of MutaGAN framework

In 2020, Liao *et al.* [69] formed a convolutional neural network (CNN) platform using entire slide image data from the cancer genome atlas (TCGA) dataset and hepatocellular carcinoma (HCC) tissue microarrays. They trained the CNN and invented the probability of each slice in two ways. They found that the proposed technique is effective for diagnosing liver cancer. The information and deep learning architecture provided in this study are summarized in Figure 20.

In the year 2020, Kyal *et al.* [70] presented a performance-based method to the Needleman-Wunsch global sequence alignment technique. They were able to eliminate several bottlenecks using hardware-based design FPGA. The execution times of sequential CPU-based design, parallel GPU-based design with CUDA-C, and hardware-based design with FPGA are compared. Both techniques outperformed the sequential CPU-based solution significantly.

In 2021, Perera and Wannigeb [71] offered a technique focused on enter star and progressive approaches for MSA. A collection of DNA sequences of HIV-1 from injured individuals with a defined chain of transmission is used to conduct the evaluation. According to the findings, the new approach generates output with a higher sum of pairwise scores than center star techniques, and the resulting final alignment can yield a more accurate phylogeny than center star and progressive techniques.

In 2021, Das *et al.* [72] proposed the codon feature-based amino acid sequence analyser (CoFASA), which is a unique alignment-free approach for nucleotide sequence similarity analysis. They construct 20-dimensional characteristics for each coding DNA sequence and protein sequence. These characteristics are used to conduct phylogenetic analysis on the candidate sequences. Feature vectors were shown to be beneficial in differentiating comparable sequences obtained from the same protein or gene in experiments. In analyzing any huge sequence, the suggested approach is straightforward and computationally effective. Figure 21 depicts the overall phases of the suggested alignment-free sequence analysis method.

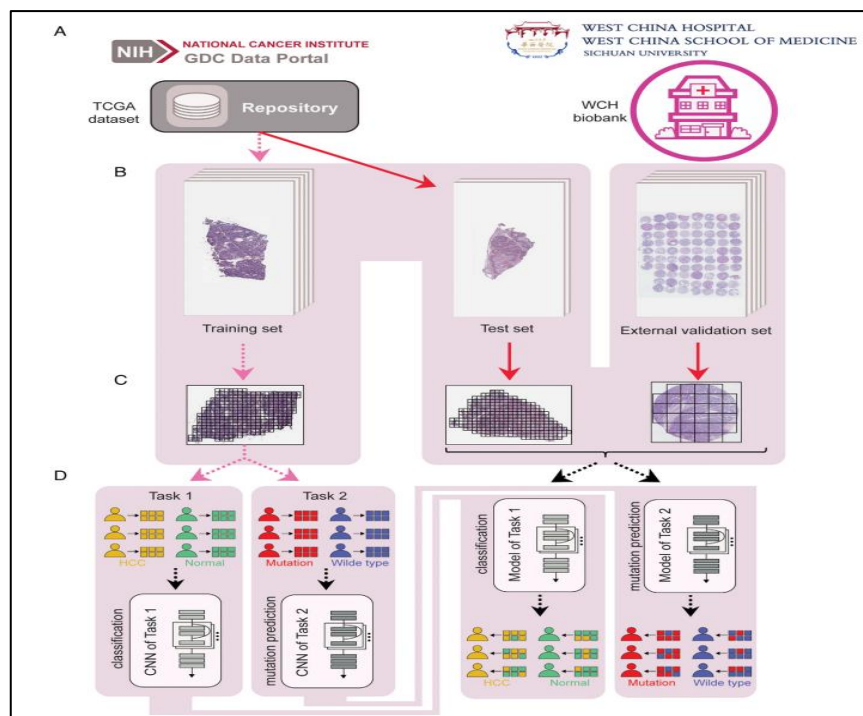


Figure 20. Summary of the data and the deep learning framework [69]

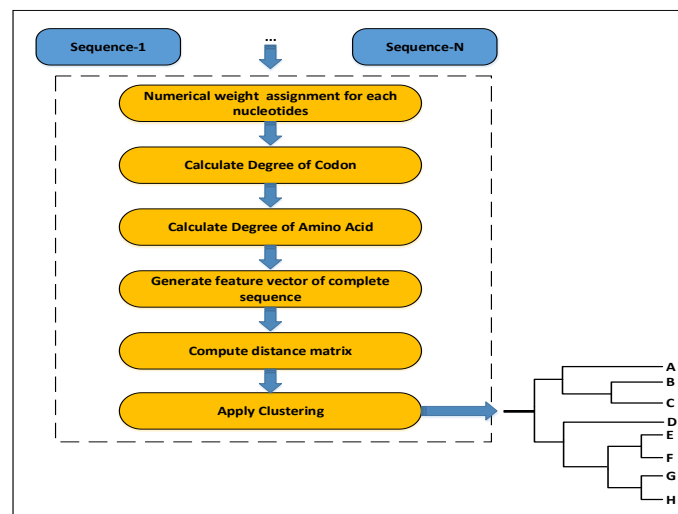


Figure 21. Overall phases of the suggested alignment-free sequence analysis method

In 2021, Rashed *et al.* [73] presented a work that shows an efficient software/hardware digital combination of two commonly used DNA sequence matching methods. The suggested method focuses on the parallelization of all of these core algorithms within specified constraints. It rely on the popular alignment parallelization techniques for DNA sequences under specified constraints. The suggested MATLAB approach for local and global alignment produces vastly better elapsed time and GCUPS than the current state-of-the-art technique. Figure 22 depicts a procedure of an SW or NW algorithm on an FPGA.

In 2021, He *et al.* [74] offered the correlation coefficient feature vector (CCFV), a unique alignment-free technique that specifies a correlation measure of the L-step delay of a nucleotide position from its placement in the original DNA sequence. The technique is used to analyze the evolution of basic human viruses such as SARS-CoV-2, Dengue virus, Hepatitis B virus, and human rhinovirus. The computational cost of multiple sequence alignment is avoided using this strategy, which leads to improved comparison speed.

In 2021, Zuo *et al.* [75] offered a mutation detection technique based on the position index of a feedback rapid learning neural network. Single nucleotide polymorphism (SNP) and InDel mutations, as well as structural mutations, may be analyzed using position correlation of sequences. The mutation sites found by position index will be more than the ones identified by Bcftools, Freebye, Vanscan2, and Gatk, according to experimental data. The feedback learning neural network index position model is shown in Figure 23.

In 2022, Prakash and Ganapathi [76] provided a cache-efficient parallel approach to accomplish the sequence alignment with gap penalty issues for shared-memory devices. In this work, an effective r-way divide-and-conquer technique is described. It has asymptotically greater parallelism and data localization. For some dynamic programming issues, a different matrix splitting method than just a 2-way split of a dynamic programming table across every aspect may assist in improving cache complexity.

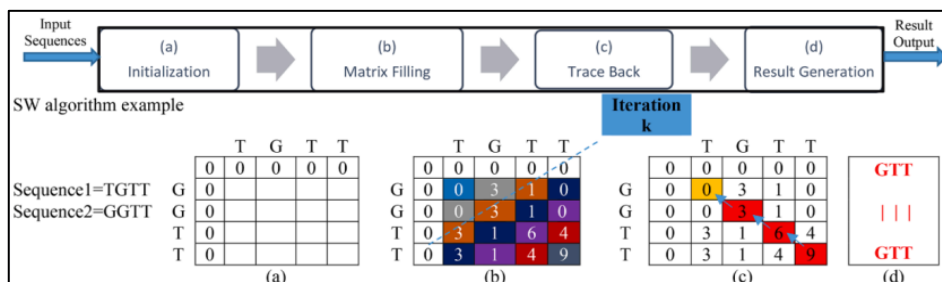


Figure 22. The procedure of an SW or NW algorithm on FPGA [73]

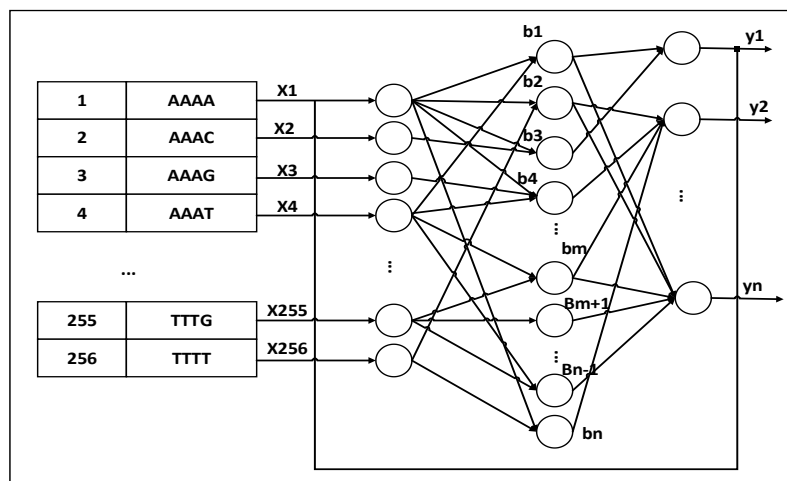


Figure 23. Feedback learning neural network index position model

9. SUMMARY

Table 4 summarizes the literature relevant to the detection and classification of genetic mutations. Following what has been seen, we outline the advantages and disadvantages discovered from the works discussed above, as shown in Table 5. The information in the two tables below can assist in understanding the approaches' advantages and disadvantages so that they can be modified to optimize their accuracy.

Table 4. Summary of the methods of the detection and classification of genetic mutations

Related Work No.	Methodology	Method Name	Sequence Type	Reference No.	Goal	Dataset
1	Needleman and Wunsch		DNA	46	DNA sequence alignment	NCBI
				65	Align DNA Sequences	
				57	identify a mutation in DNA sequences of Coronavirus	
2	Smith-Waterman	ABSW	DNA	59	Multithreaded Parallel Sequence Alignment	
				55	utilizing constant memory discovers alignment of a couple of arbitrary lengthy sequences	
				43	improves the Smith-Waterman algorithm	
3	Bayesian classifier	PAAVLS	RNA	53	Sequence alignment	
				38	with very low allele percentages, identify somatic mutations	
				50	The algorithm compares the similarity between both the individual's genome and the sick genome in the training set to predict autistic behavior.	
4	Genetic Algorithm	GARS	DNA and protein	45	Multiple sequence alignment	NCBI
				49	identify the best between the two methodologies (dynamic programming and Genetic Algorithm)	
				39	find maximum matches between two sequences	
5	GA&ABC&BFO	MO-BFO	DNA	52	sequence alignment	Cosmic dataset benchmark
				44	Multiple sequence alignment	
				70	mutation detection in the genome sequence	
6	Neural Network	MutaGAN	DNA	47	predict, mutation diagnoses, and classify cancer.	TP53
				62	forecast future biological population development and genetic mutations	
				54	Search for high-quality somatic mutations techniques for detecting mutations	
				64	prediction of somatic mutation	
7	Deep learning	Cerebro	DNA	56	create a model to detect syndromic ASD by categorizing the mutations that cause it	SFARI Gene2 database
				58	Prior detection of DNA damage and targeting the condition with a personalized approach	
				51	Multiple sequence alignment (MSA)	
8	Reinforcement Learning	RLALIGN	DNA	60	Identifying mutations	OncoKB
				NLP-SNPPred		
9	NLP		DNA	63	DNA sequence similarity	NCBI
				61	examined DNA sequence-based prediction of mutation susceptibility	
10	Center star graph-based in-memory distributed system		DNA	66	Multiple sequence alignment (MSA)	HIV database
11	Colored Petri Net (CPN), divide-and-conquer and not-in-place matrix transposition	BulkAligner	DNA	39	Sequence alignment	NGS
12			DNA	48	modeling and determining the classification of the type of the mutation	
13			DNA	71	sequence alignment	
14	alignment-free	CoFASA	DNA	67	similarity analysis of nucleotide sequences	Benchmark datasets
15	genomic signal processing	CCFV	DNA	69	Sequence alignment	SARS-CoV-2 datasets, a DENV dataset, and an HBV dataset
				74		
				75		

Table 5. Benefits and drawbacks of the methods in the review

Related Work No.	Methodology	Reference No.	Benefits	Drawbacks
1	Needleman and Wunsch	46 65 57 59	It discovers the ideal alignment solution for the sequences.	Alignment requires more time to complete, which reduces performance.
2	Smith-Waterman	55 43 53	It allows for both local and global matching and supports flexible penalties and affine gaps. Parallelizing diagonal elements is possible. Shared memory access is possible. Easy to implement and evaluation of the conditional probability is simple.	mathematically challenging large memory consumption to store the interim findings, hence, the computational complexity of traceback.
3	Bayesian classifier	38 50	Very quick - no iterations necessary because the probabilities may be calculated immediately. Therefore, this method is helpful in situations when training speed is crucial.	Zero probability issue: We may have zero class probabilities if we come across terms in the testing data for a given class that are absent from the training data.
4	Genetic Algorithm	45 49 39 52		The main drawback of this strategy is the "local minimum" issue, which results from the algorithm's greedy nature. Because more sequences are added to the alignment, this means that any errors made in any intermediate alignments cannot be fixed. Furthermore, there isn't an objective function that can be applied to determine if one alignment is better than another or whether the optimal alignment under the circumstances has been identified.
5	GA&ABC&BFO	44	This method's main benefit is its ability to align collections of sequences with known tertiary structures quickly, simply, and with good sensitivity.	
6	Neural Network	70 47 62 54 64 56	high precision and easy implementation.	difficulty in obtaining the ideal neural network parameters. The neural network will be put into use, and the encoding technique will be optimized.
7	Deep learning	58	With an one shot technique, early diagnosis, and detection of DNA repair mutations	This model has the greatest accuracy of 70% for recognizing the particular mutation between DNA repair mutations. Although this method produced acceptable findings for small sequences, it is computationally prohibitive.
8	Reinforcement Learning	51	The method produced acceptable findings for small sequences	The importance of NLP approaches in predicting the structures and functionalities of the genetic sequence data is increasing. Finding the best NLP strategies to complete a given assignment is, regrettably, never a simple endeavor.
9	NLP	60 63 61	Acceptable classification execution time of the genetic mutation	Every sequence must be compared to each other to determine how similar they are, and only then can the center star alignment be completed, which takes a lot of time and resources.
10	Center star	66	the approach yields more high accuracy than progressive alignment techniques.	Despite the loss in a disk and network I/O resulting from the input/output format's high size.
11	graph-based in-memory distributed system	39	showed a strong performance in sequence alignment that allowed polymorphisms for longer readings.	
12	Colored Petri Net (CPN)	48	A practical modeling technique to analyze and convey molecular processes intuitively	Petri nets are difficult to scale. Therefore, efforts to replicate biological systems using regular Petri nets have mostly been limited to very tiny models up.
13	divide-and-conquer and not-in-place matrix transposition	71 67	faster than other algorithms. The technique effectively uses the cache memory since it takes up little space and handles basic subproblems there instead of contacting the slower main memory.	The algorithm gives a sub-optimal alignment
14	alignment-free	69	reduces the computational burden of multiple sequence alignment, speeding up the comparison of sequences as a result.	Alignment strategies are already mature, while only a few alignment-free strategies have seriously questioned the reliability and validity of alignment-based procedures. Alignment-free methods are still in their infancy and have a great deal of room for improvement.
15	genomic signal processing	74 75	Analyzing genetic sequence simpler, quicker, and more accurate.	

10. CONCLUSION

We have analyzed several prior works of literature based on the identification of genetic mutations in genetic sequences and the categorization of mutation types in this study. These comparisons and summaries in terms of mutation prediction tasks, their types, data sources, and detection techniques, such as Smith-Waterman and Needleman-Wunsch algorithms, genetic algorithms, neural network architectures, genomic signal processing, and other listed techniques, may be of interest to researchers looking to improve their models. Such approaches offer a lot of promise for use, but due to poor interpretation and a lack of frequent validation, they are still a long way from being a reality. In short, these strategies can be improved by enhanced interpretation and a comprehensive model selection process.

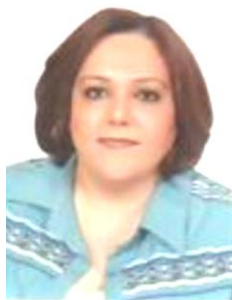
REFERENCES




- [1] B. A. Cheba, "Biotechnological applications of bioinformatics in the post genomic ERA," in 2019 *International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–6, doi: 10.1109/ICCISci.2019.8716439.
- [2] H. Ahmed, L. Alarabi, S. El-Sappagh, H. Soliman, and M. Elmogy, "Genetic variations analysis for complex brain disease diagnosis using machine learning techniques: opportunities and hurdles," *PeerJ Computer Science*, vol. 7, p. e697, 2021, doi: 10.7717/peerj-cs.697.
- [3] S. Kar, M. Ganguly, and S. Ghosal, "Prediction of coding region and mutations in human DNA by effective numerical coding and DSP technique," in 2021 *International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, pp. 180–185, doi: 10.1109/ICCCIS51004.2021.9397102.
- [4] K. Shruthi, A. Gupta, and D. Pavitra, "Sequence alignment using PSoC," in 2018 4th *International Conference for Convergence in Technology (I2CT)*, 2018, pp. 1–5, doi: 10.1109/I2CT42659.2018.9058082.
- [5] A. Sarkar and S. Banerjee, "FPGA implementation of DNA sequence alignment with traceback," in 2020 4th *International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 47–52, doi: 10.1109/ICECA49313.2020.9297554.
- [6] J. Parihar, P. Kansal, K. Singh, and H. Dhiman, "Assessment of bioinformatics and healthcare informatics," in 2019 *Amity International Conference on Artificial Intelligence (AICAI)*, 2019, pp. 465–467, doi: 10.1109/AICAI.2019.8701262.
- [7] R. Fando and M. Klavdieva, "Bioinformatics: past and present," in 2018 *International Conference on Engineering Technologies and Computer Science (EnT)*, 2018, pp. 34–36, doi: 10.1109/EnT.2018.00013.
- [8] A. B. Yousif, H. K. Al-Khafaji, and T. Abbas, "A survey of exact motif finding algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 2, p. 1109, Aug. 2022, doi: 10.11591/ijeecs.v27.i2.pp1109-1118.
- [9] W. J. da, S. Diniz, and F. Canduri, "Bioinformatics: an overview and its applications," *Genet Mol Res*, vol. 16, no. 1, pp. 10–4238, 2017, doi: 10.4238/gmr16019645.
- [10] P. Malathi, M. Manoj, R. Manoj, V. Raghavan, and R. E. Vinodhini, "Highly improved DNA based steganography," *Procedia Computer Science*, vol. 115, pp. 651–659, 2017, doi: 10.1016/j.procs.2017.09.151.
- [11] M. A. Islam, P. K. Datta, and H. Myler, "VLSI structures for DNA sequencing—a survey," *Bioengineering*, vol. 7, no. 2, p. 49, 2020, doi: 10.3390/bioengineering7020049.
- [12] S. Taluja, J. Bhupal, and S. R. Krishnan, "A survey paper on DNA-based data storage," in 2020 *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1–4, doi: 10.1109/ic-ETITE47903.2020.62.
- [13] B. B. Raj and V. C. Sharmila, "An survey on DNA based cryptography," in 2018 *International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*, 2018, pp. 1–3, doi: 10.1109/ICETIETR.2018.8529075.
- [14] B. Shabash and K. C. Wiese, "RNA visualization: relevance and the current state-of-the-art focusing on pseudoknots," *IEEE/ACM Transactions on computational Biology and Bioinformatics*, vol. 14, no. 3, pp. 696–712, 2016, doi: 10.1109/TCBB.2016.2522421.
- [15] G. Mirceva, I. Ivanoska, A. Naumoski, and A. Kulakov, "Feature selection for improved classification of protein structures," in 2019 42nd *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2019, pp. 1013–1018, doi: 10.23919/MIPRO.2019.8757005.
- [16] D. S. Kumar and P. K. Reddy, "Improved approach for protein function prediction by exploiting prominent proteins," in 2015 *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1–7, doi: 10.1109/DSAA.2015.7344865.
- [17] Z. Zhao and G. Rosen, "Visualizing and annotating protein sequences using a deep neural network," in 2020 54th *Asilomar Conference on Signals, Systems, and Computers*, 2020, pp. 506–510, doi: 10.1109/IEEECONF51394.2020.9443364.
- [18] A. S. Rifaioglu, T. Doğan, M. J. Martin, R. Cetin-Atalay, and V. Atalay, "DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks," *Sci Rep*, vol. 9, no. 1, pp. 1–16, 2019, doi: 10.1038/s41598-019-43708-3.
- [19] Z. Zhao, A. Cristian, and G. Rosen, "Keeping up with the genomes: efficient learning of our increasing knowledge of the tree of life," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–23, 2020, doi: 10.1186/s12859-020-03744-7.
- [20] C. S. Sreeja, M. Misbahuddin, and N. P. M. Hashim, "DNA for information security: A survey on DNA computing and a pseudo DNA method based on central dogma of molecular biology," in *International Conference on Computing and Communication Technologies*, 2014, pp. 1–6, doi: 10.1109/ICCT2.2014.7066757.
- [21] A. S. M. A. Mahmood, T.-J. Wu, R. Mazumder, and K. Vijay-Shanker, "DiMeX: a text mining system for mutation-disease association extraction," *PLoS One*, vol. 11, no. 4, p. e0152725, 2016, doi: 10.1371/journal.pone.0152725.
- [22] R. V. Kadumuri, and S. C. Janga, "Epitranscriptomic code and its alterations in human disease," *Trends Mol Med*, vol. 24, no. 10, pp. 886–903, 2018, doi: 10.1016/j.molmed.2018.07.010.
- [23] V. Kordopati *et al.*, "DES-mutation: system for exploring links of mutations and diseases," *Sci Rep*, vol. 8, no. 1, pp. 1–14, 2018, doi: 10.1038/s41598-018-31439-w.
- [24] H. Al-Mubaid, "Gene mutation analysis for functional annotations using graph heuristics," in 2019 *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2019, pp. 1–6, doi: 10.1109/CIBCB.2019.8791494.
- [25] Y. Guan *et al.*, "CRISPR/Cas9-mediated somatic correction of a novel coagulator factor IX gene mutation ameliorates hemophilia in mouse," *EMBO Mol Med*, vol. 8, no. 5, pp. 477–488, 2016, doi: 10.15252/emmm.201506039.
- [26] S. K. Viswanathan *et al.*, "Hypertrophic cardiomyopathy clinical phenotype is independent of gene mutation and mutation dosage," *PLoS One*, vol. 12, no. 11, p. e0187948, 2017, doi: 10.1371/journal.pone.0187948.
- [27] Z. Long, H. Li, Y. Du, M. Chen, J. Zhuang, and B. Han, "Gene mutation profile in patients with acquired pure red cell aplasia," *Annals of Hematology*, vol. 99, no. 8, pp. 1749–1754, 2020, doi: 10.1007/s00277-020-04154-8.

- [28] A. R. Lucena-Araujo *et al.*, "Combining gene mutation with gene expression analysis improves outcome prediction in acute promyelocytic leukemia," *Blood, The Journal of the American Society of Hematology*, vol. 134, no. 12, pp. 951–959, 2019, doi: 10.1155/2021/8689873.
- [29] Z. Elyazghi, L. el Yazouli, K. Sadki, and F. Radouani, "ABI base recall: Automatic correction and ends trimming of DNA sequences," *IEEE Transactions on NanoBioscience*, vol. 16, no. 8, pp. 682–686, 2017, doi: 10.1109/TNB.2017.2755004.
- [30] B. Chowdhury and G. Garai, "A review on multiple sequence alignment from the perspective of genetic algorithm," *Genomics*, vol. 109, no. 5–6, pp. 419–431, 2017, doi: 10.1016/j.ygeno.2017.06.007.
- [31] Y.-J. Song and D.-H. Cho, "Local alignment of DNA sequence based on deep reinforcement learning," *IEEE Open J Eng Med Biol*, vol. 2, pp. 170–178, 2021, doi: 10.1109/OJEMB.2021.3076156.
- [32] W. R. Pearson, "An introduction to sequence similarity ('homology') searching," *Curr Protoc Bioinformatics*, vol. 42, no. 1, pp. 1–3, 2013, doi: 10.1002/0471250953.bi0301s42.
- [33] Y. Zhu, C. S. Ong, and G. A. Huttley, "Machine learning techniques for classifying the mutagenic origins of point mutations," *Genetics*, vol. 215, no. 1, pp. 25–40, 2020, doi: 10.1534/genetics.120.303093.
- [34] P. D. Stenson *et al.*, "The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies," *Hum Genet*, vol. 136, no. 6, pp. 665–677, 2017, doi: 10.1007/s00439-017-1779-6.
- [35] S. A. O. Mohammed and A. A. Bourawy, "Modeling protein synthesis and DNA mutations using finite state machines," in *Proceedings of the 6th International Conference on Engineering & MIS 2020*, 2020, pp. 1–7, doi: 10.1145/3410352.3410773.
- [36] A. Yang, W. Zhang, J. Wang, K. Yang, Y. Han, and L. Zhang, "Review on the application of machine learning algorithms in the sequence data mining of DNA," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 1032, 2020, doi: 10.3389/fbioe.2020.01032.
- [37] G. lo Bosco and M. A. di Gangi, "Deep learning architectures for DNA sequence classification," in *International Workshop on Fuzzy Logic and Applications*, 2016, pp. 162–171, doi: 10.1007/978-3-319-52962-2_14.
- [38] S. E. Yost *et al.*, "Mutoscope: sensitive detection of somatic mutations from deep amplicon sequencing," *Bioinformatics*, vol. 29, no. 15, pp. 1908–1909, 2013, doi: 10.1093/bioinformatics/btt305.
- [39] K. Cibulskis *et al.*, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Nat Biotechnol*, vol. 31, no. 3, pp. 213–219, 2013, doi:10.1038/nbt.2514.
- [40] S. Saini and L. Dewan, "Graphical method to determine base change locations in genomic sequences of influenza a virus using wavelets," *WSEAS Trans Biol Biomed*, vol. 11, pp. 70–81, 2014.
- [41] G. Garai and B. Chowdhury, "A cascaded pairwise biomolecular sequence alignment technique using evolutionary algorithm," *Information Sciences*, vol. 297, pp. 118–139, 2015, doi: 10.1016/j.ins.2014.11.009.
- [42] J. Lee, Y. Yeu, H. Roh, Y. Yoon, and S. Park, "BulkAligner: A novel sequence alignment algorithm based on graph theory and Trinity," *Inf Sci (N Y)*, vol. 303, pp. 120–133, 2015, doi: 10.1016/j.ins.2015.01.011.
- [43] T. Abdullah, M. Faiza, P. Pant, M. R. Akhtar, and P. Pant, "An analysis of single nucleotide substitution in genetic codons-probabilities and outcomes," *Bioinformation*, vol. 12, no. 3, p. 98, 2016, doi: 10.6026/97320630012098.
- [44] V. Anitha and S. Pushpa, "DNA sequence alignment using clustering by WEKA," *International Journal On Advanced Computer Theory And Engineering (IJACTE)*, vol. 5, no. 1, 2016.
- [45] Z. Yang, R. Zhu, and L. Zhang, "The improvement and implementation on the algorithm for local alignment of pairs of DNA sequences," in 2016 *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2016, pp. 1316–1320, doi: 10.1109/IMCEC.2016.7867426.
- [46] R. R. Rani and D. Ramyachitra, "Multiple sequence alignment using multi-objective based bacterial foraging optimization algorithm," *Biosystems*, vol. 150, pp. 177–189, 2016, doi: 10.1016/j.biosystems.2016.10.005.
- [47] H. N. Kaghed, S. E. Al-Shamery, and F. E. K. Al-Khuzai, "Multiple sequence alignment based on developed genetic algorithm," *Indian Journal of Science and Technology*, vol. 9, no. 2, pp. 1–7, 2016, doi: 10.17485/ijst/2016/v9i2/84236.
- [48] S. S. Ray, A. Banerjee, A. Datta, and S. Ghosh, "A memory efficient DNA sequence alignment technique using pointing matrix," in 2016 *IEEE Region 10 Conference (TENCON)*, 2016, pp. 3559–3562, doi: 10.1109/TENCON.2016.7848720.
- [49] A. G. Ismael and D. Y. Mikhail, "Effective data mining technique for classification cancers via mutations in gene using neural network," *arXiv preprint arXiv:1608.02888*, 2016, doi: 10.1109/TENCON.2016.7848720.
- [50] J. Yang, J. Lian, H. Pu, and J. Gu, "Modeling and analysis of DNA mutation type based on colored Petri net," in 2017 *IEEE International Conference on Information and Automation (ICIA)*, 2017, pp. 864–869, doi: 10.1109/ICInfA.2017.8079024.
- [51] S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in 2013 *IEEE Conference on Information & Communication Technologies*, 2013, pp. 1227–1231, doi: 10.1109/CICT.2013.6558288.
- [52] A. G. Ismael and R. Z. Yousif, "Novel mining of cancer via mutation in tumor protein P53 using quick propagation network," *arXiv e-prints*, p. arXiv-1505, 2015.
- [53] N. H. Kaghed, E. S. Al, and F. E. K. Al-Khuzai, "Comparative study of genetic algorithm and dynamic programming of DNA multiple sequence alignment," *Journal of University of Babylon*, vol. 25, no. 2, pp. 403–414, 2017.
- [54] R. Reeta, G. Pavithra, V. Priyanka, and J. S. Raghul, "Predicting autism using naive Bayesian classification approach," in 2018 *International Conference on Communication and Signal Processing (ICCS)*, 2018, pp. 109–113, doi: 10.1109/ICCS.2018.8524491.
- [55] R. K. Ramakrishnan, J. Singh, and M. Blanchette, "Rlalign: a reinforcement learning approach for multiple sequence alignment," in 2018 *IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2018, pp. 61–66, doi: 10.1109/BIBE.2018.00019.
- [56] S. Q. Abedulridha and E. S. Al-Shamery, "Optimal pair DNA sequence alignment based on matching regions and multi-zone genetic algorithm," *International Journal of Engineering & Technology*, vol. 7, no. 4.19, pp. 751–756, 2018.
- [57] J. Sun, K. Chen, and Z. Hao, "Pairwise alignment for very long nucleic acid sequences," *Biochem Biophys Res Commun*, vol. 502, no. 3, pp. 313–317, 2018, doi: 10.1016/j.bbrc.2018.05.134.
- [58] D. E. Wood *et al.*, "A machine learning approach for somatic mutation discovery," *Sci Transl Med*, vol. 10, no. 457, p. eaar7939, 2018, doi: 10.1126/scitranslmed.aar7939.
- [59] Y.-L. Liao, Y.-C. Li, N.-C. Chen, and Y.-C. Lu, "Adaptively banded smith-waterman algorithm for long reads and its hardware accelerator," in 2018 *IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, 2018, pp. 1–9, doi: 10.1109/ASAP.2018.8445105.
- [60] S. Kar, M. Ganguly, and S. Das, "Using DIT-FFT algorithm for identification of protein coding region in eukaryotic gene," *Biomedical Engineering: Applications, Basis and Communications*, vol. 31, no. 01, p. 1950002, 2019, doi: 10.4015/S1016237219500029.
- [61] V. P. Sudha and M. S. Vijaya, "Machine learning-based model for identification of syndromic autism spectrum disorder," in *Integrated Intelligent Computing, Communication and Security*, Springer, 2019, pp. 141–148, doi:10.1007/978-981-10-8797-4_16.




- [62] M. I. Irawan, I. Mukhlash, A. Rizky, and A. R. Dewi, "Application of needleman-wunch algorithm to identify mutation in DNA sequences of Corona virus," in *Journal of Physics: Conference Series*, 2019, vol. 1218, no. 1, p. 012031, doi: 10.1088/1742-6596/1218/1/012031.
- [63] S. N. PV, S. Akshayaa, B. Vaisali, and K. N. PK, "DNA repair mutation detection using deep learning strategy—A pharmacogenomic perspective," in 2019 *Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2019, vol. 1, pp. 1–4, doi: 10.1109/i-PACT44901.2019.8960113.
- [64] V. Gancheva and I. Georgiev, "Multithreaded parallel sequence alignment based on needleman-wunsch algorithm," in 2019 *IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2019, pp. 165–169, doi: 10.1109/BIBE.2019.00037.
- [65] N. Rehmat, H. Farooq, S. Kumar, S. Hussain, and H. Naveed, "Predicting the pathogenicity of protein coding mutations using natural language processing," in 2020 *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 5842–5846, doi: 10.1109/EMBC44109.2020.9175781.
- [66] A. Yilmaz, "Assessment of mutation susceptibility in DNA sequences with word vectors," *Journal of Intelligent Systems: Theory and Applications*, vol. 3, no. 1, pp. 1–6, 2020, doi: 10.38016/jista.674910.
- [67] E. Delibaş, A. Arslan, A. Şeker, and B. Diri, "A novel alignment-free DNA sequence similarity analysis approach based on top-k-n-gram match-up," *Journal of Molecular Graphics and Modelling*, vol. 100, p. 107693, 2020, doi: 10.1016/j.jmgm.2020.107693.
- [68] D. S. Berman, C. Howser, T. Mehoke, and J. D. Evans, "MutaGAN: A Seq2seq GAN framework to predict mutations of evolving protein populations," *arXiv preprint arXiv:2008.11790*, 2020.
- [69] H. Liao *et al.*, "Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma," *Clin Transl Med*, vol. 10, no. 2, 2020, doi: 10.1002/ctm2.102.
- [70] C. Kyal, R. Kumar, and A. Zamal, "Performance-based analogising of needleman wunsch algorithm to align DNA sequences using GPU and FPGA," in 2020 *IEEE 17th India Council International Conference (INDICON)*, 2020, pp. 1–5, doi: 10.1109/INDICON49873.2020.9342078.
- [71] G. K. K. Perera and C. T. Wannige, "A hybrid algorithm for identifying partially conserved regions in multiple sequence alignment," *International Journal of Computers and Applications*, vol. 43, no. 10, pp. 979–986, 2021, doi: 10.1080/1206212X.2019.1628468.
- [72] J. K. Das, A. Sengupta, P. P. Choudhury, and S. Roy, "Mapping sequence to feature vector using numerical representation of codons targeted to amino acids for alignment-free sequence analysis," *Gene*, vol. 766, p. 145096, 2021, doi: 10.1016/j.gene.2020.145096.
- [73] A. E. E.-D. Rashed, M. Obaya, H. El, and D. Moustafa, "Accelerating DNA pairwise sequence alignment using FPGA and a customized convolutional neural network," *Computers & Electrical Engineering*, vol. 92, p. 107112, 2021, doi: 10.1016/j.compeleceng.2021.107112.
- [74] L. He, S. Sun, Q. Zhang, X. Bao, and P. K. Li, "Alignment-free sequence comparison for virus genomes based on location correlation coefficient," *Infection, Genetics and Evolution*, vol. 96, p. 105106, 2021, doi: 10.1016/j.meegid.2021.105106.
- [75] Z. Zuo *et al.*, "Gene position index mutation detection algorithm based on feedback fast learning neural network," *Computational Intelligence and Neuroscience*, 2021, doi: 10.1155/2021/1716182.
- [76] S. Prakash and P. Ganapathi, "An algorithm for the sequence alignment with gap penalty problem using multiway divide-and-conquer and matrix transposition," *Information Processing Letters*, vol. 173, p. 106166, 2022, doi: 10.1016/j.ipl.2021.106166.

BIOGRAPHIES OF AUTHORS



Rana Hikmet Saloom    is an assistant lecturer at the Ministry of Higher Education and Scientific Research-Iraq. She holds a B.Sc. degree from Al-Mansour University College and received her M.Sc. from the Iraqi Commission for Computers and Informatics-Information Institute for Postgraduates, Baghdad, Iraq, in 2000 and 2018 respectively. She is doing her Ph.D. research at the Iraqi Commission for Computers and Informatics-Institute of Informatics for Postgraduates, Baghdad, Iraq. Research interests include bioinformatics and data mining. She can be contacted by email: Phd202020565@iips.icci.edu.iq.



Hussein K. Khafaji    received the B.Sc., M.Sc., and Ph.D. degrees from the University of Technology, Baghdad, Iraq, in 1989, 1992, and 2002, respectively. He has been a professor of AI and data mining at Al-Rafidain University College since 2012. He is currently the Head of the Computer Communications Engineering Dept. at Al-Rafidain University College, Baghdad, Iraq. He is also a member of the advisory council of the Iraqi Commission for Computers and Informatics in the Ministry of Higher Education and Scientific Research in Iraq. He has published more than 70 refereed journal and conference papers in the fields of AI and data mining and its applications. He can be contacted by email: hussain.ketan.elc@ruc.edu.iq.