A SEQUENCE ANALYSIS AND ALIGNMENT SOFTWARE (SAAS) FOR OPTIMAL PERFORMANCE IN DNA SEQUENCE ANALYSIS

G A INYANG[1], F U OGBAN[2], U O UDENSI[3], E U OYOITA[1], E A EDIM[2]

[1]Department of Computer Science, University of Cross River State Nigeria.
[2]Department of Computer Science, University of Calabar, Calabar Nigeria.
[3]Department of Genetics and Biotechnology, University of Calabar, Calabar Nigeria.

E-mail: [1]gain140270@gmail.com, [2]fogban@gmail.com, [3]ugoji2012@gmail.com,
[1]emmanueloyoita@crutech.edu.ng, [2]edimemma@gmail.com

ABSTRACT

In recent time, there have been upsurges of DNA and Protein sequence data deposited in Genebank or databases that are subjected to analysis, which is made possible through the utilization of bioinformatics tools. However, the accuracy and informativeness of the sequences often analyzed depend on the suitability of the bioinformatics Software employed during the analysis. The thinking however, is that for phylogenetic reconstruction software for instance, using a particular method, Maximum Likelihood, for example, should display exact and similar output with same clustering/class pattern and bootstrap values. It therefore becomes urgent and imperative to delve into developing sequence analytical software whose command and algorithmic composition is the same, irrespective of the suite placed. This will aid interpretation of results, the researcher notwithstanding. This is the pivot on which this paper anchors. This paper will evaluate the extent and pattern of clustering of MEGA 7 and PAUP 4 using Maximum Likelihood, Parsimony and Neighbour-Joining Methods based on analyzing sequence data from pigeon pea (*Cajanus cajan* (L) Millsp) using the two analytical Software Packages (MEGA 7 and PAUP 4) with a comparative novel algorithm called Sequence Analysis and Alignment Software (SAAS), as it relates to phylogenetic reconstruction to test for species relatedness and divergence. The results from SAAS showed reduced variance of resolution into specific clusters as the number of the reads of the sequence decreases, and optimal performance in terms of runtime and accuracy of resolution into clusters.

Keywords: *Sequence Analysis and Alignment Software (SAAS), Pigeon Pea Plant, Phylogeny, Maximum Likelihood, Parsimony and Neighbour-Joining, Species Relatedness and Divergence Accuracy and Informativeness.*

1. INTRODUCTION

Most recently, there have been upsurges of DNA and protein sequence data deposited in gene bank or databases that are subjected to analysis, which is made possible through the utilization of bioinformatics tools. However, the accuracy and informativeness of these sequence often analyzed depend on the suitability of the bioinformatics Software employed during the analysis. Understandably, these bioinformatics software are written with algorithms that will enable the software to resolve the intended goal(s).

The puzzle here is the fact that these software developers or programmers write algorithm of these software based on what they want to achieve, which might not be the priority of the researcher. The implication is that, researcher using these software(s) may not understand the intent of the programmer, implying that the results from such analysis might be either disconcerted or wrongly interpreted.

Although various algorithms have been formulated and software packages developed for sequence data analysis, the developers of these algorithms and software packages optimize these analytical tools with respect to various considerations [1]. The worrisome perspective to this end is the fact that these software(s) even when used to resolve phylogeny of species adopting the same method gives different, conflicting and confusing outputs. It is obvious that the algorithmic commands differ as well as the defaults used by each programmer [2].

Sequence analysis is carried out for a variety of purposes, e.g. to find common subsequences (low information) or surprising subsequences (high information), to find repetition, signals, motifs, pattern or structure (low information), all against a background of chance matches, natural variation, evolution and mutation [3]. It has been argued that information content, as realized in the compression or message length criterion, captures an important aspect of these intuitions. Only by having a good statistical model of sequences is it possible to quantify 'common' and 'surprising' outcomes. Even if this is done subject to a number of simplifying assumptions then at least those assumptions, i.e. the design and parameters of a model, are explicit and are open to challenge and objective improvement [3]. Including the cost of the model itself prevents over-fitting and allows simple and complex models to be compared fairly. [3] Suggested that models based on finite-state machines have many practical advantages: their complexity is easy to quantify, the resulting inference algorithms are feasible, i.e. have reasonable complexity, and one can envisage a systematic search through at least the smaller finite-state machines.

Major algorithms used in gene sequence clustering can be divided into two categories according to the result format: hierarchical clustering algorithms and partitional clustering algorithms Hierarchical approaches may yield fairly good results, but they require the similarity of all pairs of sequences and quickly arrive at a bottleneck in terms of computational time and memory usage for large-scale data sets [4]. Hierarchical clustering is widely used for detecting clusters in genomic data. It generates a set of partitions forming a cluster hierarchy. According to linkage criteria, there are three hierarchical clustering methods including single-linkage clustering (SL), complete-linkage clustering (CL) and average-linkage clustering (AL)

The thinking however, is that for phylogenetic reconstruction software for instance, using a particular method, Maximum Likelihood, for example, should display exact and similar output with same clustering/class pattern and bootstrap values. It therefore becomes urgent and imperative to delve into developing sequence analytical software whose command and algorithmic composition is the same, irrespective of the suite placed. This will aid interpretation of results, the researcher notwithstanding. This is the pivot on which this paper anchors.

1.1 Aim and Objectives of Study

The aim of this research is to develop a Sequence Analysis and Alignment Software (SAAS) with optimal performance in terms of resolution power into classes / clusters of the genes in DNA sequence data, which will be same in result output despite the analytical software package.

The specific objectives are;

1. To evaluate the extent and pattern of clustering of MEGA 7 and PAUP 4 using Maximum Likelihood, Parsimony and Neighbour-Joining Methods.
2. To ascertain the reliability of the phylogenetic trees nodes using bootstrap estimation as benchmark.
3. To compute profile execution time in nanoseconds and accuracy of short reads for optimal performance in terms of resolution power into clusters.
4. To develop a sequence analysis and alignment algorithm (SeAAA) that will produce the same phylogenetic tree pattern in-terms of clustering of gene, the analytical package lodged notwithstanding.

1.2 Justification of the Work

It is obvious that the algorithmic commands differ as well as the defaults used by each programmer. The thinking however, is that for phylogenetic reconstruction software, using a particular method, Maximum Likelihood, for example, should display exact and similar output with same clustering/class pattern and bootstrap values.

It therefore becomes urgent and imperative to delve into developing sequence analytical software whose command and algorithmic composition is the same, the suite placed, notwithstanding; to aid interpretation of results by any researcher. This is the pivot on which this present research anchors. Additionally, the defaults used in this different software do not include drop down that show the orientation for this display or the direction of the parse tree such that every user knows the direction of clustering for uniformity of interpretation. Putting all together, these helps in developing a more robust software or package that is to be same, the suite where they are placed notwithstanding.

1.3 Limitation of the Work

The study is limited to evaluating the extent and pattern of clustering of MEGA 7 and PAUP 4 using Maximum likelihood, Parsimony and Neighbour-joining Methods. Having established that several other similar software produces different results for same methods, the scope is based on analyzing sequence data from pigeon pea (*Cajanus cajan* (L)Millsp) using two analytical Software Packages (MEGA 7 and PAUP 4) as a test frame to develop a comparative novel algorithm called Sequence Analysis and Alignment Algorithm (SeAAA). The results of Maximum likelihood, Parsimony and Neighbour-joining methods from MEGA 7 and PAUP 4 are expected to give an insight to the different presentations from these packages.

In each of this, investigation was carried out on Maximum Likelihood, Parsimony and Neighbour-joining methods as it relates to phylogenetic reconstruction to test for species relatedness and divergence. Human genome was not used.

2. LITERATURE REVIEW

Deoxyribonucleic acid (DNA), is a polymer made from four basic nucleotides; adenine (A), cytosine (C), guanine (G), and thymine (T). It is the main carrier of the genetic information or the genetic material. The DNA sequencing is the process in which this information is extracted by converting physical DNA molecules into a signal that describes the exact order and type of the constituent nucleotides.

Recently, DNA sequencing has revolutionized molecular biology, biomedicine and life sciences in general, [5]. The field of bioinformatics has made it possible to use computational and statistical methods analyzing sequence databases. These tools allow software developers and researchers to select methods and algorithms best suited to understand the function, structure, evolution, and adaptation of genes and species, [6]. The flood of data acquired from the increasing number of publicly available databases has led to new demands for bio-informatics software. With the growing amount of information resulting from high throughput experiments, new questions arise that often focus on the comparison of genes, genomes, and their expression profiles. Inferring new knowledge by combining different kinds of post-genomics data obviously necessitates the development of new approaches that allow the integration of variable data sources into a flexible framework, [7]; [1].

In the context of an explosion of new data management challenges, the field of bio-informatics is rapidly becoming the most critical step in realizing the full potential of genomics and proteomics. Population geneticists, ecologists, epidemiologists, and clinical researchers, have routinely used these statistical and computational models to evaluate information collected.

2.1 DNA Structure

DNA is made up of molecules called nucleotides. Each nucleotide contains a phosphate group, a sugar group and a nitrogen base. The deoxyribonucleic acid (DNA) molecule is the carrier of the genetic information in the human cells. It represents a single format onto which a broad range of biological phenomena can be projected for high-throughput data collection. In fact, all the information it contains are passed from organisms to their offspring during the process of reproduction. Most specifically, each DNA molecule consists of the four nucleotides; Adenine (A), Cytosine (C), Guanine (G), and Thymine (T), with backbones made of sugars and phosphate groups joined by ester bonds (Table 1). It is organized as a double-helix and the two strands of the DNA molecule are complementary to each other. The base-pairing is fixed: A is always complementary to T, and G is always complementary to C. This double-helix model was proposed in 1953 by James Watson, an American scientist, and Francis Crick, a British researcher, and it has not been changed much since then. This discovery was made by studying the X-ray diffraction patterns, and this allowed to the two scientists to build models and to figure out the double-helix structure of DNA, a structure that enables it to carry biological information from one generation to the following generation figure1.

The complete genome is composed of chromosomes, which are a set of different DNA molecules. Eukaryotes, i.e. organisms that have cells with a structure made by membranes, store the chromosomes in the nuclei of their cells. For example, in humans, the genome has a total of more than 3 billion of nucleotides that are organized in 23chromosomes, each of which appears in two copies in each cell. An important functional unit in a chromosome is the gene, which is described by the so called central dogma of molecular biology.

Table 1: Nucleosides and Nucleotides

Base	Nucleoside(base + sugar)	Nucleotide(base+sugar+ phosphate)
Adenine	Deoxyadenosine (sugar = deoxyribose)	Deoxyadenylic acid OR deoxyadenosine monophosphate
Guanine	Deoxyguanosine (sugar = deoxyribose)	Deoxyguanylic acid OR deoxyguanosine monophosphate
Cytosine	Deoxycytidine (sugar = deoxyribose)	Deoxycytidylic acid OR deoxycytidine monophosphate
Thymine	Deoxythymidine (sugar = deoxyribose)	Deoxythymidylic acid OR deoxythymidine monophosphate
Uracil (in RNA)	Uridine (in RNA)(sugar=ribose)	Uridylic acid OR uridine monophosphate

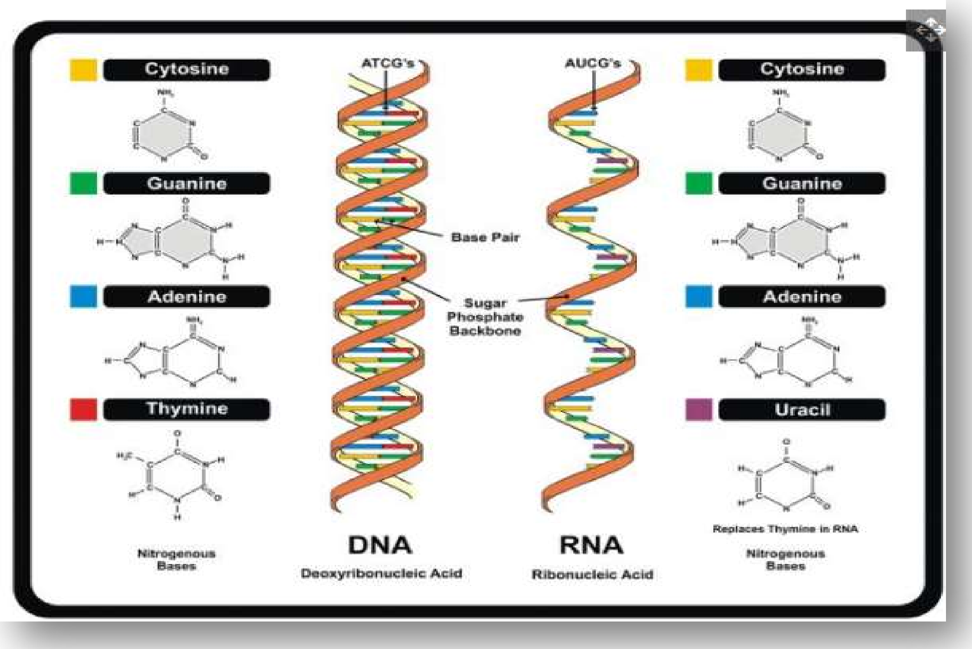


FIG. 1: DNA Structure:
Source :(www.livescience.com. December, 2019)

The central dogma in molecular biology says that portions of DNA, called gene regions, are transcribed into ribonucleic acid (RNA), which is then translated into proteins (figure 1). The RNA molecules differ from the DNA ones because they are often single stranded and the Thymine (T) nucleotide is replaced by Uracil (U).

2.2. DNA Sequencing.

DNA sequencing is a method used to determine the precise order of the four nucleotide bases – adenine, guanine, cytosine and thymine - that make up a strand of DNA. These bases provide the underlying genetic basis (the genotype) for telling a cell what to do, where to go and what kind of cell to become (the phenotype). Nucleotides are not the only determinants of phenotypes, but are essential to their formation. Each individual and organism has a specific nucleotide base sequence. The possible letters are A, C, G, and T, representing the four nucleotide bases of a DNA strand — **adenine**, **cytosine**, **guanine**, **thymine.**

DNA sequencing played a pivotal role in mapping out the human genome, and is an essential tool for many basic and applied research applications today. It has provided an important tool for determining thousands of nucleotide variations associated with specific genetic diseases.

2.3. DNA Analysis

Bioinformatics and the field of molecular biology have provided analytical software tools on how to analyze DNA sequence data. DNA analysis involves the following methods

i. Polymerase chain reaction (PCR)
ii. Gel electrophoresis.
iii. Polymerase chain reaction (PCR)
iv. DNA sequencing.

Obtaining a large number of identical DNA fragments is very important. From a computer science perspective, having the same string in 109 copies does not mean much since it does not increase the total amount of information. However, most experimental techniques (like gel electrophoresis, used for measuring DNA length) require many copies of the same DNA fragment. Since it is difficult to detect a single molecule or even a hundred molecules with modern instrumentation, amplifying the DNA to yield millions or billions of identical copies is often a prerequisite of further analysis.

2.4. Molecular Evolutionary Genetics Analysis (MEGA) Version 7 Software.

The Molecular Evolutionary Genetics Analysis (MEGA) version 7 software is one of the versions of MEGA which contains many sophisticated methods and tools for phylogenomics and phylomedicine. Usually, this software undergoes reversions, which is geared towards making it better in meeting the concerns of end-users, MEGA 7 has been optimized for use on 64-bit computing systems for analyzing larger datasets. This application is a desktop application designed for comparative analysis of homologous gene sequences either from multigene families or from different species with a special emphasis on inferring evolutionary relationships and patterns of DNA and protein evolution [8].

MEGA7 software infers evolutionary relationships of homologous sequences explore basic statistical properties of genes as well as estimate neutral and selective evolutionary divergence among sequences. This has brought series of modifications of the software from the previous version of MEGA to MEGA 7. All these are geared towards achieving the goal of making available a wide variety of statistical and computational methods for comparative sequence analysis in a user-friendly environment. [1]

In summary, MEGA 7 is an integrated work bench used by biologists for data mining from the web, aligning sequences, conducting phylogenetic analyses, testing evolutionary hypothesis and generating publication quality displays and descriptions [6].

2.5. Phylogenetic Analysis using Parsimony (PAUP 4) Version 4

Phylogenetic Analysis Using Parsimony (PAUP) version 4 is a program for phylogenetic analysis using parsimony, maximum likelihood, and distance methods. The program features an extensive selection of analysis options and model choices, and accommodates DNA, RNA, protein and general data types. Among the many strengths of the program are the rich arrays of options for dealing with phylogenetic tree including importing, combining, comparing, constraining, rooting and testing hypotheses.

PAUP4 provides a wide range of pairwise distant measures, from simple absolute differences to more complicated model-based corrected distances. In addition, PAUP4 can use the minimum evolution and least-squares functions to evaluate trees under the distance criterion.

2.6. Methods used for Phylogeny

The methods briefly discussed below are commonly used in producing phylogenetic trees in MEGA 7 and PAUP 4.

2.6.1. Maximum composite likelihood (ml) method

The maximum likelihood method uses standard statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees. The method requires a substitution model to assess the probability of particular mutations; roughly, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability. ML is an exhaustive method that searches every possible tree topology and considers every position in an alignment, not just informative sites. By employing a particular substitution model that has probability values of residue substitutions, ML calculates the total likelihood of ancestral sequences evolving to internal nodes and eventually to existing sequences. It sometimes also incorporates parameters that account for rate variations across sites.

It works by calculating the probability of a given evolutionary path for a particular extant sequence. The probability values are determined by a substitution model (either for nucleotides or amino acids), [9]. Maximum likelihood is thus well suited to the analysis of distantly related sequences, but it is believed to be computationally intractable to compute due to its NP-hardness

2.6.2. Parsimony method (PM)

Parsimony analysis is the second primary way to estimate phylogenetic trees from aligned sequences. It may be used to estimate "species" or "gene" phylogenies.

In the parsimony approach, the goal is to identify that phylogeny that requires the fewest necessary changes to explain the differences among the observed sequences.

2.6.3. Neighbour-joining (NJ)

Neighbor-joining methods apply general cluster analysis techniques to sequence analysis using genetic distance as a clustering metric. The simple neighbor-joining method produces unrooted trees, but it does not assume a constant rate of evolution (i.e., a molecular clock) across lineages.

2.7. Bootstrapping

Bootstrap involves resampling with replacement from one's molecular data to create fictional datasets, called **bootstrap** replicates, of the same size. Specifically, the molecular data is typically organized as **a multiple sequence alignment** (MSA) of species ×n characters. It is a resampling method by independently sampling with replacement from an existing sample data with same sample size n, and performing inference among these resampled data. *towardsdatascience.com* (April, 2020).

To make a bootstrap replicate of the primates data, bases are sampled randomly from the sequences with replacement and concatenated to make new sequences. The same number of bases as the original multiple alignments is used in this *analysis*, and then gaps are removed to force a new pairwise alignment (figure 2).

It is possible to generate new data points by re-sampling from an **existing sample**, and make inference just based on these new data points.

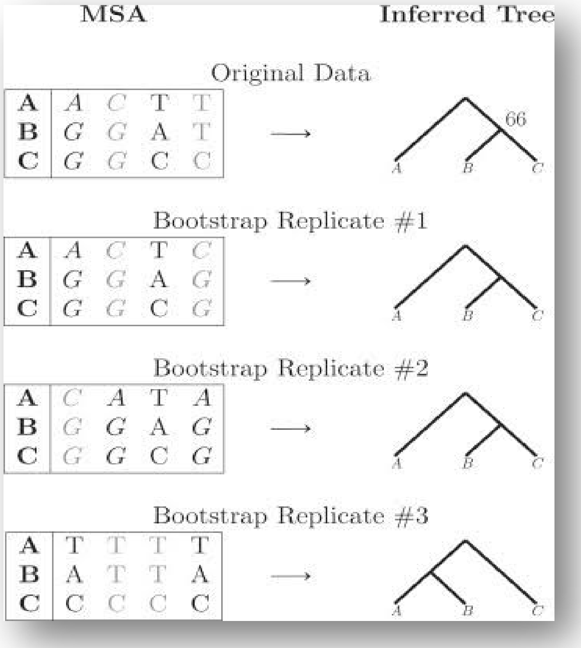


Fig. 2: Bootstrapping- To Test for the Reliability of the Phylogenetic Trees Nodes.
Source: An Overview/Sciencedirect.Com, 2020

3. DESIGN METHODOLOGY

The design methodology clearly involves various steps in analyzing and running alignment with the two analytical software(s), Molecular **E**volutionary **G**enetic **A**nalysis (MEGA) Version 7 and **P**hylogenetic **A**nalysis **U**sing **P**arsimony (PAUP) Version 4.

Notably here, the file formats are FASTA for MEGA and NEXUS for PAUP 4 respectively. The algorithms used in these software(s) are Maximum Expectation Algorithm for MEGA, while for PAUP 4, the algorithms are Exhaustive Search (ES), Branch and Bond(BBS), and Heuristic Search (HS) Algorithms. Both Software algorithms have analysis, simulation and alignment modules.

Running analysis, simulations and alignment with these modules using the methods of Maximum Likelihood (ML), Parsimony(PM) and Neighbour-Joining (NJ) of the reference DNA sequence data from species of pigeon pea, with the two analytical software packages; (MEGA) Version 7 and (PAUP) Version 4 revealed the results in fig. 3. In this, attempt was made to decipher the variations in phylogenetic reconstruction of gene sequences of *matK, petB* as well as *ITS* of pigeon pea accessions using these analytical software packages.

3.1 Results from MEGA 7 and PAUP 4 Analytical Software(s)

In running analysis and alignment with the two analytical software(s), the reference DNA sequence data (genome) for species of pigeon pea [10] was used. The following phylogenetic trees in fig. 3 represent the results obtained from running the analysis of this sequence data using these two analytical tools.

The phylogenetic trees of the three (3) genes (*matK*, *petB* and *its*) sequences were reconstructed using three methods- Maximum Likelihood, Parsimony and Neighbour-Joining in two software-MEGA 7 and PAUP 4. Results obtained fig. 4 revealed that clustering pattern was gene-dependent, which was clearly separated using Maximum Likelihood (ML) and Neighbour-Joining (NJ) in MEGA software. However, though using parsimony, the clusters were still based on specific genes sequences, but the clustering pattern was at variant with those of ML and NJ.

The three methods (ML, PM and NJ) on PAUP 4 showed the same clustering pattern, though still maintaining the gene-specific clustering. Critically, NJ and PM on PAUP 4 clustered the gene sequences approximately the same way from bottom to top.

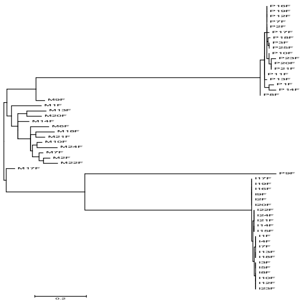
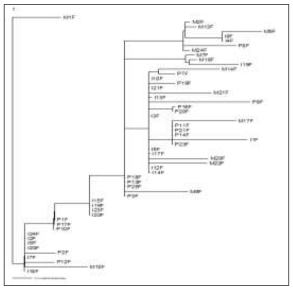
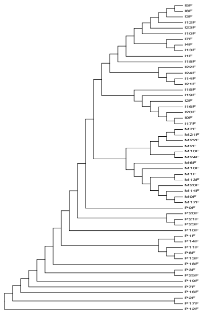
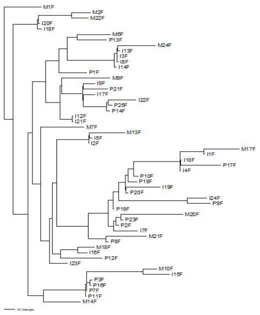
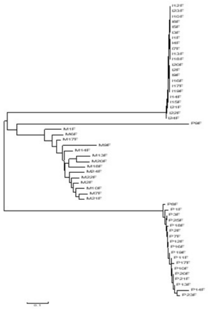
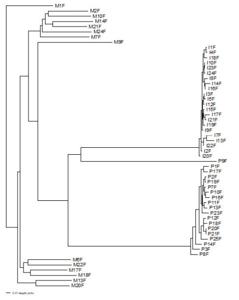
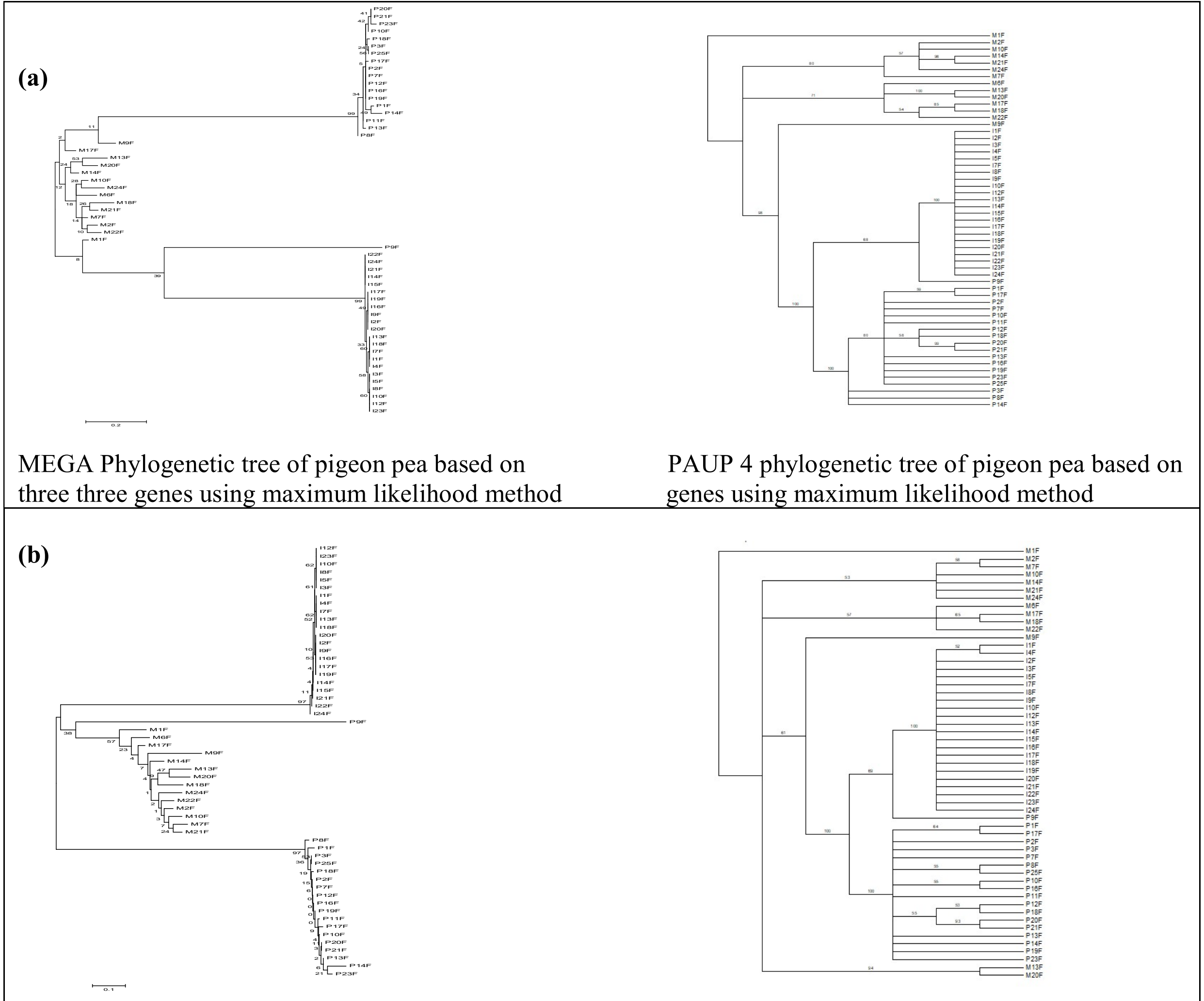
MOLECULAR EVOLUTIONARY GENETICS ANALYSIS (MEGA) VERSION 7.	PHYLOGENETIC ANALYSIS USING PARSIMONY VERSION 4 (PAUP 4)
(a)  **MEGA Phylogenetic Tree of pigeon pea based on the three genes (*matk, its and pepb*) using Maximum likelihood Method.**	 **Paup4 Phylogenetic Tree of pigeon pea based on the three genes (*matk, its and pepb*) using Maximum likelihood Method.**
(b)  **MEGA Phylogenetic Tree of pigeon pea based on three (*matk, its and pepb*) genes using Parsimony Method.**	 **PAUP 4 Phylogenetic Tree of pigeon pea based on three genes (*matk, its and pepb* using Parsimony Method.**
(c)  MEGA Phylogenetic Tree of pigeon pea based on three genes (*matk, its and pepb*) using Neighbor-Joining Method.	 PAUP 4 Phylogenetic Tree of pigeon pea based on three genes (*matk, its and pepb*) using the Neighbor-Joining Method.

FIG. 3: Phylogenetic Trees from Analyzed Sequence Data from Pigeon Pea Plant, Showing inconsistency in clustering of three genes (matk, its and pepb) using MEGA and PAUP 4 Analytical Software

3.2 Result from Bootstrapping

Using the bootstrapping method to test for the reliability of the phylogenetic trees nodes, and considering the three methods (Parsimony, neighbor-joining and maximum likelihood); these two software (MEGA 7 and PAUP 4) revealed varying reliability strengths. Comparing the bootstrap values (figure 4), ML showed better reliability, which is followed by the NJ method on the MEGA but the three methods on PAUP 4 revealed high bootstrap values between nodes of clusters in the phylogenetic trees.

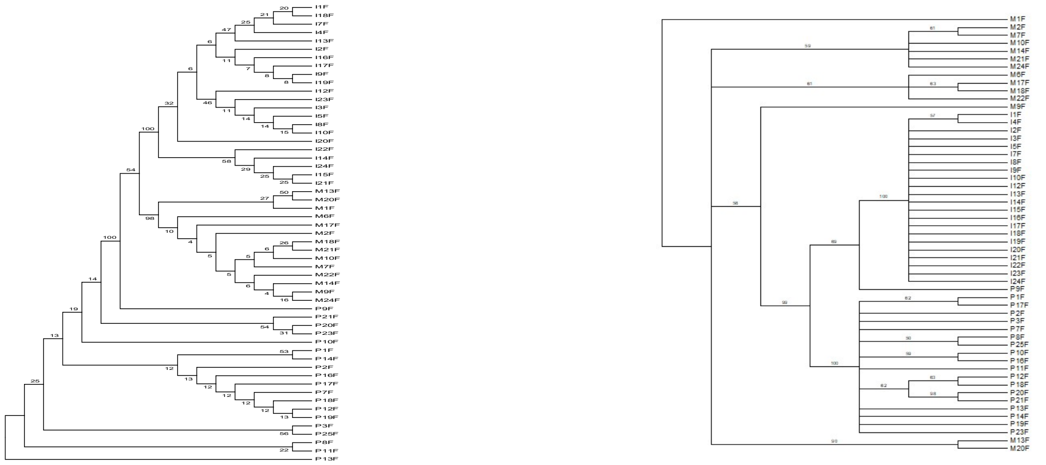


MEGA Phylogenetic tree of pigeon pea based on three three genes using maximum likelihood method

PAUP 4 phylogenetic tree of pigeon pea based on genes using maximum likelihood method

MEGA Phylogenetic tree of pigeon pea based on three genes using the Neighbor-Joining method

PAUP 4 Phylogenetic tree of pigeon pea based on three genes using Neighbor - Joining method



(c)

MEGA Phylogenetic tree of pigeon pea based on three genes using Parsimony method

PAUP 4 Phylogenetic tree of pigeon pea based on genes using Parsimony method

FIG.4: Bootstrapping method to test for the reliability of the phylogenetic trees nodes for the three (3) methods on the two (2) analytical software packages.

4. IMPLEMENTATION AND RESULTS

This paper evaluates the extent and pattern of clustering of MEGA 7 and PAUP 4 using Maximum Likelihood, Parsimony and Neighbour-Journing Methods based on analyzing sequence data from pigeon pea (*Cajanus cajan* (L)Millsp) using the two analytical Software Packages (MEGA 7 and PAUP 4) with a comparative novel algorithm called Sequence Analysis and Alignment Algorithm (SeAAA), as it relates to phylogenetic reconstruction to test for species relatedness and divergence.

In information theory, the distance between two strings of equal length is the number of positions at which the corresponding symbols are different. That is, the measurement of the minimum number of *substitutions* required to change one string into the other, or the minimum number of *errors* that could have transformed one string into the other. In a more general context, the distance is one of several string metrics for measuring the edit distance between two sequences [11].

The Sequence Analysis and Alignment Software (SAAS) basically involve reading a reference DNA in FASTA file format. It consists of analysis, simulations and alignment modules respectively. The methods include, SAAS Maximum Likelihood (SML), SAAS Parsimony (SPM) and SAAS Neighbour-Joining (SNJ) with the novel algorithm called, Sequence Analysis and Alignment Algorithm (SeAAA).

The Sequence Analysis and Alignment Software (SAAS) provide modules for analysis and alignment of sequence data. It will run on a Core i7 system for efficiency, speedy runtime performance and accuracy of cluster resolutions of genes.

The software, as comprehensive as it will be, is implemented in Java programing language with several utilities that can be used for the following;

- ➢ Reading a pre-existing reference DNA sequence data from one or more FASTA files.
- ➢ Generate a DNA sequence data based on input parameters (length, repeat count, repeat length, as well as repeat variability rate).
- ➢ Simulate reads in the genome based on input parameters of read length, coverage, and sequencing error rate.
- ➢ Apply Sequence Analysis and Alignment Algorithm (SeAAA) to the sequence data and the reads through a standardized interface.
- ➢ Parse the output of the alignment tool and calculate the number of reads that were correctly or incorrectly aligned.
- ➢ Display phylogenetic tree.

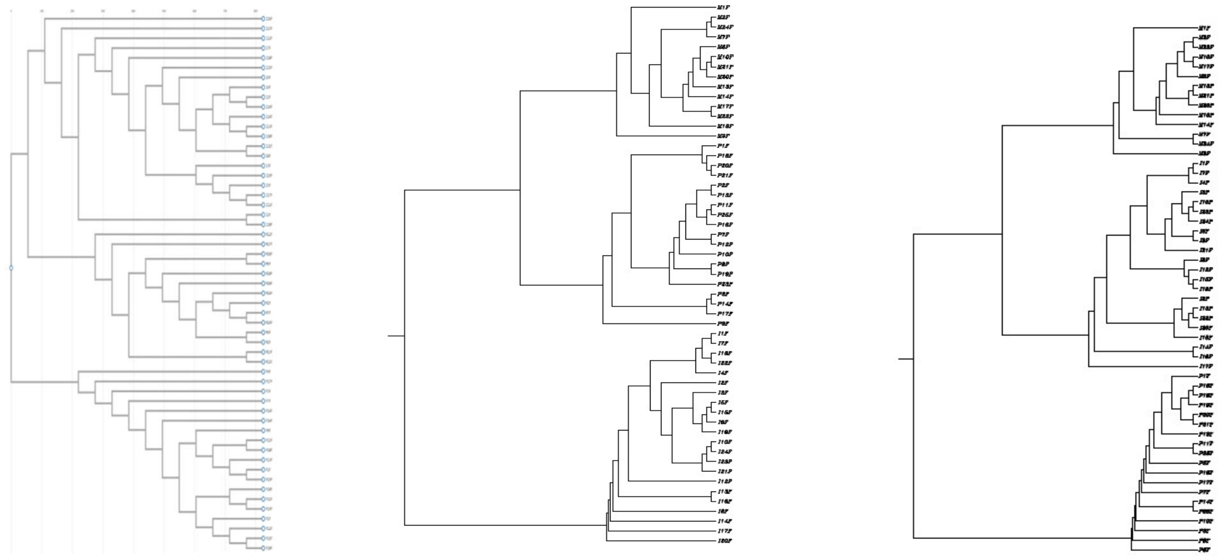
- Compute runtimes and measures of accuracy.

The ability to generate random reference genomes enables systematic studies of the effect of various factors on software performance. In particular, besides specifying the length of the reference genome, the adjustment of different repeat parameters—repeat count, repeat length and repeat variability rate (the probability of altering a base at each genome location during a repeat) will be achieved. The phylogenetic trees in figure 6, revealed after the various platforms for analyzing and aligning of the genes for variability and clustering in a potential mappings of a read. Repeats are quite common in real genomes, but of essence is the speed in performance runtime and accuracy.

4.1. Discussions on Results from Sequence Analysis and Alignment Software (SAAS)

The phylogenetic trees of the three (3) genes (*matK, pepB* and *its*) sequences were reconstructed using three methods- Maximum Likelihood, Parsimony and Neighbour-Joining with the Sequence Analysis and Alignment Algorithm (SeAAA). Critically, ML, NJ and PM clustered the gene sequences approximately the same way from bottom to top as the bootstrap values (fig. 4), revealing high reliability values between nodes of clusters in the phylogenetic trees.

The results from Sequence Analysis and Alignment Software (SAAS) (figure 5), therefore showed reduced variance of resolution into specific clusters as the number of the reads of the sequence decreases, and optimal performance in terms of runtime and accuracy of resolution into clusters.



SAAS Phylogenetic tree
Maximum Likelihood Method

SAAS Phylogenetic tree
Parsimony Method

SAAS Phylogenetic tree
Neighbour–Joining method

FIG. 5: Phylogenetic Trees of MatK, Pet and ITS genes from SAAS showing cluster resolutions of genes from SAAS ML, PM and NJ methods respectively.

TABLE 3:Runtime In Nanosecond Of Alignment Analysis With 1000 Reads

S/N	METHOD	MEGA 7 Time in nanosecond (ns)	PAUP 4 Time in nanosecond (ns)	SAAS Time in nanosecond (ns)
1	ML	932,000,000,000	48,000,000,000	33,000,000,000
2	PM	2,539,000,000,000	56,000,000,000	47,000,000,000
3	NJ	65,000,000,000	38,000,000,000	28,000,000,000

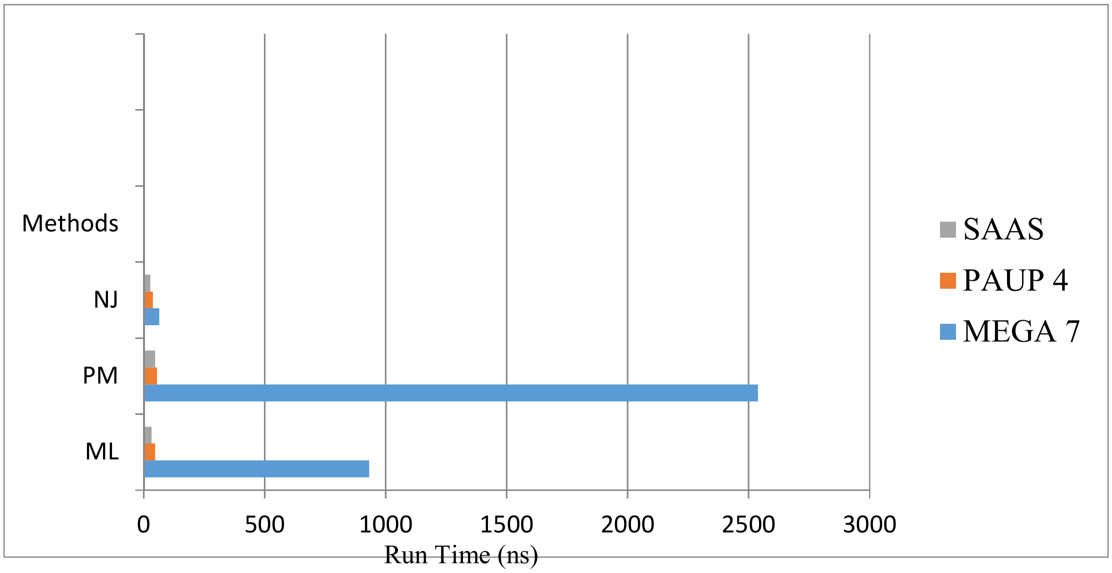


Fig. 6: Bar Chart Showing Run Time of the Methods in Nanoseconds.

5.1. Summary

In considering and investigating the optimal performance in terms of the resolution in clusters of the genes from Pigeon Pea Plant, a comparative analysis of two (2) existing sequence data software analytical packages was explored with focus on phylogeny using maximum likelihood, parsimony and neighbor-joining methods. Results in figure 4 revealed inconsistency in terms of clustering of genes from the algorithms used in these analytical software packages. These results in respect to alignment and authentic statistical basis, shows dependence on guesses and less accurate resolution power of clusters of the genes.

The revolution in sequence of data analysis due to the development of a number of new high-throughput sequencing technologies have fundamentally changed the way in which sequence data are analyzed, opening new perspectives and new directions that were not achievable or even thinkable. They have provided the opportunity for a global investigation of multiple genomes and transcriptomes. The large collection of computerized ("digital") nucleic acid sequences, protein sequences, or other polymer sequences stored on a computer known as Sequence Data [12] are fundamental data types in bioinformatics and computational biology [13].

Recent studies reveal also that BlastClust is less effective for clustering divergent sequences [14], and its performance strongly depends on the choice of optimal BLAST parameters including similarity threshold, percent identity, and alignment length [15]. CD-HIT-EST, on the other hand, does not provide hierarchical relationships between clusters of sequences. In many situations both CD-HIT-EST and BlastClust yield clusters with only one sequence [15]. All the traditional clustering methods based on sequence alignment encounter computational difficulties in dealing with large biological databases.

[16] presented a new approach involving a new alignment-free distance measure based on k-tuples, DMk (Distance Measure based on k-tuples), and a modified bisecting K-means clustering algorithm, mBKM (modified Bisecting K-Means algorithm). mBKM aims to speed up the clustering process by using the alignment-free similarity measure, and is able to produce either a hierarchical clustering or a partition clustering result.

Phylogenetic measures of diversity, or phylo-diversity, have been developed as a way of deriving an objective, quantitative metric that reflects different intrinsic values of species or areas for several purposes – (of conservation prioritization [17]; [18]. These measures have in common that they are derived from some function of the branch lengths of a phylogeny, as distinct from taxonomic measures that depend on the number of nodes or species in the tree [18]. Metrics may use just a single edge from a phylogenetic tree—such as the tip length connecting a species to its nearest relative—or a path between tips in the phylogeny, or the sum of edges for a clade or sample of tips [19].

5.2 Conclusion

In consideration of the above, it became urgent and imperative to delve into developing a standardized Sequence Analysis and Alignment Software (SAAS) with a novel algorithm, whose composition and result obtained thereof is the same, despite the analytical software used.

The results from the Sequence Analysis and Alignment Algorithm (SeAAA) showed reduced variance of resolution into specific clusters, and optimal performance in terms of runtime and accuracy of resolution into clusters (figure 5).

5.3 Recommendation

The following divergence in results produced running analysis with two analytical software tools in sequence data relating to phylogenetic tree to an extent revealed inconsistency in the algorithms in these software tools. Therefore, the need for further study on other analytical tools in other areas of computational biology or bioinformatics will further justify the efficacy of the algorithms in the software analytical tools.

The field of computational biology or bioinformatics requires the collaboration of team players in biological sciences, computer science, software engineering and statistics to develop a *unified analytical tool* for sequence data analysis with which same method with a given parameter should produce the same or very similar result. Thereby, not giving room for different interpretations of one result from same method. This is often prevalence when a choice for pivot or seed for computation is arbitrary for different developed module.

REFERENCE

[1]. Ogban, F. U; Udensi, U. O; Inyang, G. A; Osang, F (2019b) Efficacy of the algorithm(s) in analytical software packages on DNA sequence data analysis. International Journal Of Natural And Applied Sciences (IJNAS) 12 (1), P. 47 – 60

[2]. Ogban, F U, Udensi U O, Inyang G A, Osang F, (2019a) Diversity in Content and Analytical Algorithms of Biologist-Centric Software(s) for DNA and Sequence Data: Mega, DNASp, GenAlex and ARLEQUIN. International Journal of Natural and Applied Sciences (IJNAS) Vol 12, Special Edition (2019: P. 61 – 73

[3]. Allison L, Stern L., Edgoose T., Dix T.I. (2000) Sequence complexity for biological sequence analysis. Computers and Chemistry 24 (2000) 43–55, 0097-8485/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved. PII: S0097-8485(99)00046-7

[4]. Hao X, Jiang R, Chen T (2011): Clustering 16 S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics. 2011, 27 (5): 611-618. 10.1093/bioinformatics/btq725.

[5]. Cao, D. Ganesamoorthy, A.G. and Elliott, D. (2016). Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION sequencing. *Giga Science*, 5, 1-12.

[6]. Kumar, S., Nei, M., Dudley, J., and Tamura, K. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, 9(4), 299-306. DOI: 10.1093/bib/bbn017

[7]. Goodman, R..P., Berry, R.M. and Turberfield, A. .J. (2004).The single-step synthesis of a DNA tetrahedron.*Chem. Commun*.,12, 1372–1373

[8]. Kumar, S, Stecher G, Peterson D, Tamura K. (2012). MEGA computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28:2685–2686.

[9]. Hedges, S. B, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. MolBiol E vol 32:835–845.

[10]. Udensi, O. U. (2020) Infering genetic diversity in pigeon pea *(Cajanus Cagan)* (L) Mil(LP) accessions from *pepB*, *metK* and *ITS* genes

[11]. Ogban, FU; Edim, EA; Essien, EE; Ofem, OA; Edim, EB (2021) Upshot of Percentage Consistency and the Sum-of-Pairs Score Factor on the Algorithmic Structures of Some DNA Analysis Tools; International Journal of Computer Science and Information Security (IJCSIS …Vol 19: 8 P. 32-43

[12]. Nakamura, Y, (2010) Translational repression by the ocyte-specific protein P100 in Xenopus. *Dev Biol* 344(1):272-83

[13]. Andrey D. P and Alla L. L (2016) Encyclopedia of Bioinformatics and Computational Biology Volume 1, 2019, Pages 38-6

[14]. Alam I, Cornell M, Soanes DM, Hedeler C, Wong HM, Rattray M, Hubbard SJ, Talbot NJ, Oliver SG, Paton NW: A Methodology for Comparative Functional Genomics. JIntegr Bioinform. 2007, 4 (3): 69-

[15]. Picardi E, Mignone F, Pesole G: EasyCluster: a fast and efficient gene-oriented clustering tool for large-scale transcriptome data. BMC Bioinformatics. 2009, 10 (Suppl 6): S10-10.1186/1471-2105-10-S6-S10.

[16]. Jiang H, Wong W.H. (2010) SeqMap: mapping massive amount of oligonucleotides to the
genome Bioinformatics.24:2395–2396.

[17]. Crozier, R. H. (1997). Preserving the information content of species: genetic diversity, phylogeny, and conservation worth. Annual Review of Ecology and Systematics, 28(1), 243– 268. https://doi.org/10.1146/annurev.ecolsys.28.1.243.

[18]. Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., … Mazel, F. (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biological Reviews, 92(2), 698– 715. https://doi.org/10.1111/brv.12252.

[19]. Kondratyeva, A., Grandcolas, P., & Pavoine, S. (2019). Reconciling the concepts and measures of diversity, rarity and originality in ecology and evolution. Biological Reviews, 94(4), 1317– 1337. https://doi.org/10.1111/brv.12504.