INTERNATIONAL JOURNAL OF NATURAL AND APPLIED SCIENCES (IJNAS), VOL. 12, Special Edition (2019); P. 47 – 60, 2 TABLES, 6 FIGS.

Efficacy of the algorithm(s) in analytical software packages on DNA sequence data analysis

F. U. Ogban[1*], U. O. Udensi[2], G. A. Inyang[1], F Osang[3]

ABSTRACT

Most recently, there has been upsurges of DNA and Protein sequence data deposited in Genbank 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. This work intends to check the efficacies of the underlying algorithms in Molecular Evolutionary Genetic Analysis (Mega 7) and Phylogenetic Analysis Using Parsimony (PAUP 4) version 4 with respect to Maximum Likelihood, Parsimony and Neighbour-Joining methods. The underlying benchmarks shall be the Accuracy, Running time, Execution time and the response time. Understandably, these bioinformatics software are produced or written with algorithms that will enable the software to resolve the intended goal(s).The conundrum here is the fact that these software developers or programmers import algorithm of these software based on what they want to achieve. As such, student researchers using this software may not understand the intent of the programmer, implying that the result from such analysis might be less informative. Undoubtedly, every researcher is poised to achieving a certain goal and if the analytical tools used are not specific to achieving the desired goal(s), the results obtained thereof may cause wrong conclusion to be drawn.

INTRODUCTION

Most recently, there has been an upsurge of DNA and protein sequence data deposited in genbank or databases that are subjected to analysis, which is made possible through the utilization of bioinformatics tools. However, the accuracy and informativeness of these sequences often analyzed depend on the suitability of the bioinformatics Software employed during the analysis. Understandably, this bioinformatics software are produced or 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 algorithms of these software based on what they want to achieve. As such student researchers 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.

This is notwithstanding, the complication and differences arising from results got from different systems should resolve a specific parameter, such as phylogenetic tree using the same method. The results and runtime given the same stream of DNA sequence suggest to a large extent the correctness of the algorithm(s) thus the software package used. The implication is that algorithmic models contribute largely to the performance of the software packages used.

This paper is certainly of immense importance to the field of bioinformatics and computational biology. Due to the fact that the efficiency of the algorithms used in decision making remains very critical regarding DNA/RNA sequence data analysis, its findings will reveal how appropriate these algorithms and the corresponding packages are for sequence data analysis. In this paper, we will be limited to investigating the effectiveness of the algorithm(s) in analytical software packages as regards DNA sequence data analysis.

The scope will be based on analyzing sequence data from pigeon pea (Cajanus Cajan (L) Millop) using two analytical Software packages; Molecular Evolutionary Genetic Analysis (MEGA 7) version 7, and Phylogenetic Analysis Using Parsimony (PAUP4) version 4. In each of this, investigation will be on Maximum Likelihood, Parsimony and Neighbour-Joining methods as relates to phylogenetic reconstruction to test for species relatedness and divergence.

*Corresponding author**.** Email: felixogban@unical.edu.ng
[1]Department of Computer Science, University of Calabar, Calabar, Nigeria
[2]Department genetics and biotech, University of Calabar, Nigeria.
[3]Department of computer science, national open University of Nigeria

Aim and Objectives

The aim of this paper is to investigate the suitability of the algorithm(s) in analytical software packages on DNA/RNA sequence data analysis.

The Specific objectives are;

To examine the efficiency of two analytical software packages – MEGA 7 and PAUP4

To run DNA sequence data analysis with these analytical tools with special consideration to Phylogeny using Maximum Likelihood, Parsimony and Neighbour-Joining methods.

To propose a more optimal, generic Sequence Analysis and Alignment/merger Algorithm (SeAAA).

Related Work

Next-generation sequencing techniques are demonstrating promise in transforming research in life sciences (Schuster, 2007).

These techniques support many applications including metagenomics (Qin *et al.*, 2010), Ogban et al (2016) detection of Single Nucleotide Polymorphisms (SNPs) (Van Tassel*et al*., 2008) and genomic structural variants (Alkan*et al.*, 2009; Medvedev *et al*., 2009) in a population, DNA methylation studies (Taylor *et al.*, 2007), analysis of mRNA expression (Sultan *et al.*, 2008), cancer genomics (Guffanti*et al.*, 2009) and personalized medicine (Auffray*et al*., 2009). Some applications (e.g. metagenomics) require *de novo* sequencing of a sample (Miller *et al.*, 2010), while many others (e.g. variant detection, cancer genomics) require re-sequencing. For all of these software applications, the vast amount of data produced by sequencing poses many computational challenges (Hedges et al, 2015), Ogban et al (2016).

In this paper, a review of two (2) analytical software as regards DNA sequence data with respect to Maximum Likelihood, Parsimony and Neighbour- Joining Methods will be looked into.

i. **M**olecular **E**volutionary **G**enetic **A**nalysis (MEGA) Version 7
ii. **P**hylogenetic **A**nalysis **U**sing **P**arsimony Version 4 (PAUP 4)

(I) Molecular Evolutionary Genetics Analysis (Mega) Version 7

The Molecular Evolutionary Genetics Analysis (MEGA) version 7 software is the latest version of MEGA which contains many sophisticated methods and tools for phylogenomics and phylomedicine. In this major upgrade, 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.

The Molecular Evolutionary Genetics Analysis (MEGA) version 7 software aims to serve both of these purposes in inferring evolutionary relationships of homologous sequences, exploring basic statistical properties of genes and estimating neutral and selective evolutionary divergence among sequences. This has brought series of modifications of the software from MEGA to MEGA 7 all in a quest to achieving her goal of making available a wide variety of statistical and computational methods for comparative sequence analysis in a user-friendly environment.

Historically, MEGA 7 has included likelihood methods for estimating evolutionary distances between sequence pairs as well as distance-based and Maximum Parsimony methods for inferring phylogenic trees. Ogban et al (2016). With MEGA 7, facilities for selecting the best-fit model of DNA and protein substitution, estimating the extent of rate variation among sites, testing molecular clocks among species and paralogous genes, reconstructing nucleotides and amino acids in the ancestral sequences, and inferring phylogenic trees have been added. These additions will provide an integrated solution for analysis of molecular sequences using a variety of statistical methods.

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

Software packages on DNA sequence data analysis

Establishing Variation in Distances (Pairwise)

Estimating the number of nucleotide or amino acid substitutions needed to compute evolutionary distances is one of the most important subjects in molecular evolutionary genetics and comparative genomics. Evolutionary distances are required for reconstructing phylogenic trees, assessing sequence diversity within and between groups of sequences, and estimating times of species divergence, among other things. Ogban et al (2016) MEGA 7 contains many statistical methods for estimating the evolutionary distance (actual number of substitutions per site) between sequences based on the observed number of differences. The methods included correct for multiple substitutions by taking into account the transition/transversion bias, unequal base frequencies, varying substitution rates among sites, and heterogeneous substitution patterns among lineages. Researchers can choose any of these options from a simple dialogue box (Figure 1). MEGA 7 divides distances into three groups – nucleotide, synonymous–non synonymous and amino acid – based on the properties of the sequence data and the type of substitutions being considered.

Nucleotide distances estimate the number of nucleotide substitutions per site between DNA sequences. Analytical formulas for estimating these distances under many substitution models are included in MEGA 7 (Table 1). Under some models, numbers of transition and transversion substitutions per site also can be estimated separately.

Table 1. Nucleotide substitution models for phylogenetic analysis

Substitution Model	Transition/ transversion bias	Base frequency bias	Rate variation among sites	Heterogeneous patterns among lineages
Jukes–Cantor[28]			Yes	
Kimura[24]	Yes		Yes	
Tamura[26]	Yes	Yes (G+C)	Yes	Yes
Tamura–Nei[25]	Yes	Yes	Yes	Yes

In Phylogenetic analysis, appropriate substitution models are considered (Table 1). The most common models are:

JUKES AND CANTOR[69] MODEL **JC69 model** (Jukes and Cantor, 1969): The Jukes-Cantor model is the simplest model that proposes a correction of the number of observed substitutions. Assumes that the probability of mutating a nucleotide for another is independent of the position of the said nucleotide and the nucleotide itself:

The probability of changing A by C, G or T is identical to alpha/3. In the same way for C, G and T.

KIMURA[80] MODEL

K80 model (Kimura, 1980): Kimura proposed a refinement of the Jukes-Cantor model which takes into account the greater probability of observing transitions than observing transversions and therefore, depends on two parameters:

The probability of observing a transition, alpha.

The probability of observing a transversion, beta.

TAMURA[92] MODEL

T92 model (Tamura, 1992): T92 is a simple mathematical method developed to estimate the number of nucleotide substitutions per site between two DNA sequences, by extending Kimura's (1980) two-parameter method to the case where a G+C-content bias exists. This method will be useful when there are strong transition-transversion and G+C-content biases.

TAMURA AND NEI[93] MODEL

TN93 model (Tamura and Nei 1993): The TN93 model distinguishes between the two different types of transition, *i.e.*, (A <-> G) is allowed to have a different rate to (C<->T). Transversions are all assumed to occur at the same rate, but that rate is allowed to be different from both of the rates for transitions.

Sequence Diversity (DNA and Protein Sequences)

Sequence alignment is usually the first step in comparative sequence analysis. It is the process of identifying homologous nucleotide (or amino acid) positions among a set of sequences. Building these alignments involves many steps such as acquiring sequences from databanks, performing computational sequence alignments, and manual fine tuning of the initial alignment.

Data acquisition in MEGA 7 routinely involves obtaining gene sequences from databanks using a web browser. Homologous sequences usually are searched in the Basic Local Alignment Search Tool *(*BLAST*)* procedure by using either a gene name (or other attributes such as the GenBank accession numbers) or a query

sequence. In both cases, a set of sequences is found and displayed on the computer screen. From this set, researchers may select all or some of the sequences based on specific criteria, for example, taxonomic sampling of chosen species and/or sequence matching score. Usually at this point, investigators begin the mundane, frustrating task of cutting and pasting the sequences from the web-browsers or saving them to files and then processing them for sequence alignment. To streamline this process, MEGA 7 now includes an integrated web-browsing facility (Figure 1). Researchers can use it in the same way as the commercial web browsers, such as Internet Explorer, Opera mini or fire fox. Because the MEGA 7 web-browsing facility is a wrapper around the full-function HTML browser in the Microsoft Windows operating system, it works even if a commercial browser is not installed on the computer. The MEGA 7 browser therefore can be used as a general-purpose web browser.

In the MEGA 7 web browser, once investigators have generated the list of desired sequences, they click on 'ADD TO ALIGNMENT' where upon MEGA 7 parses the sequences automatically and sends them to the Alignment Explorer (AE)(Figure 1).This web exploration and data retrieval system will help investigators in their everyday activities without the need to reinvent protocols and allows them to use the novel and modified data searching capabilities provided by GeneBank and other servers without requiring a MEGA 7upgrade.

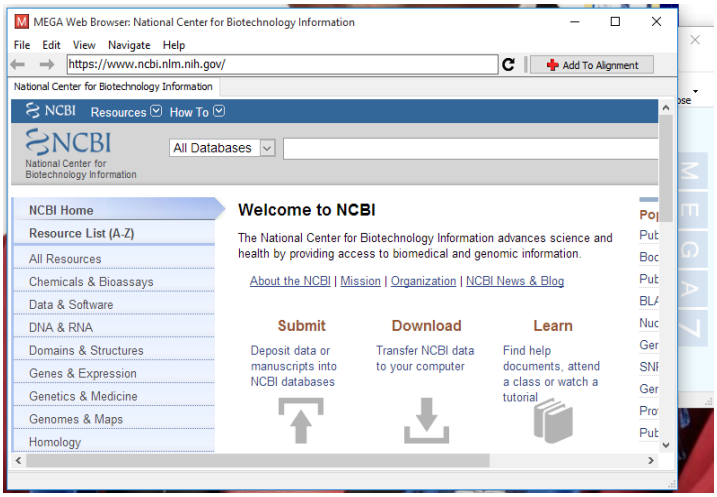


Fig. 1. The Alignment Explorer in MEGA 7 for creating, viewing and editing DNA.

Alignment Explorer (AE) Tool is a versatile tool for building DNA and protein sequence alignments.

It requires:

a. An extensive graphical user interface with facilities to edit sequence data such as the manual insertion of gaps and reverse complementation of DNA;
b. A computational capability for automated multiple sequence alignment;
c. Services for aligning coding sequences intuitively on DNA as well as protein
sequence levels; and
d. Facilities for the easy importation and exportation of the sequence data.

The Alignment Explorer (AE) offers two views of the data: DNA and (translated) amino acid sequences. In the grid, each row represents a single sequence and each column represents a site. Identity across all sequences is indicated by a '*' character in the top row of each column.

For automated sequence alignment, the Alignment Explorer (AE) includes a native implementation of CLUSTALW and is the most widely used multiple sequence alignment system for DNA and protein data. Sequence alignments can be edited manually and other operations on individual sites, columns and blocks can be performed with just a few mouse clicks. The AE provides unlimited undo capabilities and allows the user to construct alignments intuitively. Users can mark a rectangle (rows and columns of the source sequence) for alignment, invoke the integrated multi-sequence alignment module (CLUSTALW), specify appropriate alignment parameter values and initiate the sequence alignment. On the completion of the alignment, the Alignment Explorer (AE) automatically inserts the aligned sequences back into the source rectangle by expanding or contracting it appropriately. This allows for aligning of different regions of the sequence independently. For example, protein-coding nucleotide regions can be aligned separately from non-coding regions. For the protein coding regions, users can translate the selected sequences (or chosen rectangle) into protein sequences by a single mouse click, align the translated protein sequences using CLUSTALW, and then flip back to DNA sequences. The Alignment Explorer (AE) automatically adjusts the source nucleotide sequences as per the amino acid sequence alignment. Translated protein sequences can be further aligned manually even before the user comes back to the DNA sequences,

thus replacing a multi-step error-prone manual process by a simple and intuitive procedure.

Phylogenic Trees in Mega

Phylogenic trees infer the evolutionary relationships of species and patterns of gene duplications in multigene families.

They are also important for elucidating the patterns and processes of molecular evolution through studies of adaptive and neutral evolutionary changes. MEGA contains both distance-based and maximum parsimony (MP) methods for phylogenic reconstruction. It includes the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), the Neighbour-Joining (NJ) 50 method, and the Minimum Evolution (ME) method for inferring phylogenic trees using distance matrices. UPGMA is an agglomerative algorithm in which the tree is inferred, assuming constancy of the rate of evolution for all lineages. It should be used only if this assumption is satisfied. MEGA 7 contains a non-parametric test of the molecular clock to compare the rate of evolution in two sequences, given an outgroup sequence. The power of this test is similar to the Muse–Weir maximum likelihood ratio test.

In summary, for Phylogenetic trees, methods such as, Maximum Likelihood, Parsimony, and Neighbour-Joining methods are used in MEGA 7 (Figure 2).

Phylogenetic Sequence Analysis Using Mega 7

To carryout Phylogenetic analysis in MEGA 7, the following steps are required;

1. Save file as .FAS instead of .txt

2. Go to Align -> Retrieved from sequence -> upload the .FAS file

3. Under Alignment tab select your method of alignment i.e. CLUSTALW or

MUSCLE and perform the MSA accordingly.

4. Go to Data -> save session -> .MAS file

5. Close the save session

6. Go to File -> open a file/session -> select the saved .MAS file and then select

Analyze the MAS file

7. Two icons will appear in the main window i.e. with TA and close data sign

8. Go to Phylogeny -> and select the algorithm according to your choice

9. A pop up window will appear, requesting if continuation with the same data file is required i.e. .MAS file -> click yes

10. Finally, window with substitution model will appear -> click compute and tree

will be displayed.

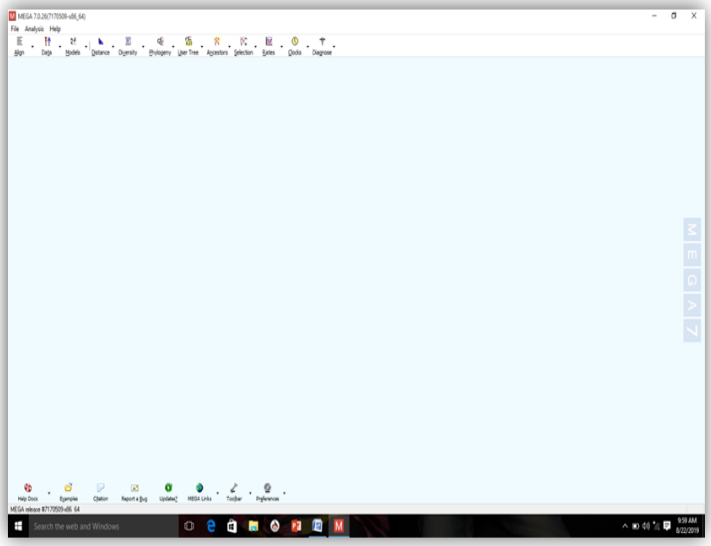


Fig. 2. The standalone interface for MEGA 7

Algorithms in Mega 7

(a) Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is a way to find maximum-likelihood estimates for model parameters when your data is incomplete, has missing data points, or has unobserved (hidden) latent variables. It is an iterative way to approximate the maximum likelihood function. While maximum likelihood estimation can find the “best fit” model for a set of data, it does not work particularly well for incomplete data sets. It works by choosing random values for the missing data points, and using those guesses to estimate a second set of data. The new values are used to create a better guess for the first set, and the process continues until the algorithm converges on a fixed point.

(b) Limitations of Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is very slow, even on the fastest computer. It works best when there is only a small percentage of missing data and the dimensionality of the data is not too big. The higher the dimensionality, the slower the

Expectation-step; for data with larger dimensionality, the-step may runs extremely slow as the procedure approaches a local maximum.

(ii) Phylogenetic Analysis Using Parsimony Version 4 (Paup 4)

Phylogenetic Analysis Using Parsimony Version 4 (PAUP4) is a program for phylogenetic analysis using parsimony, maximum likelihood, and distance methods (Figure 3). 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 is the rich array of options for dealing with phylogenetic trees 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. Pairwise distances can be summarized in a table or used to construct UPGMA and neighbor joining trees. In addition, PAUP4 can use the minimum evolution and least-squares functions to evaluate trees under the distance criterion.

Phylogenetic Sequence Analysis with Paup 4

In carrying out phylogenetic analysis with PAUP 4, the following steps are required:

1. Run PAUP4.0 and open the Nexus File, "smalldata_ptme3.nxs".
2. Execute your Nexus file in PAUP4.0 and conduct a neighbour-joining search (Choose "Distance", and then "Neighbor Joining/UPGMA" in the "Analysis" menu; refer to pp. 76-92 in PTME2). Root your NJ tree using midpoint rooting. Save the NJ tree to a file (note: this saves the new file).
3. Conduct a maximum Likelihood search.
4. Under "Print Trees", preview one of your Parsimony trees. Capture this image using either "Save as pict" or with the Grab utility so that you can print it out later and hand it in with your homework.
5. Compute and print out a 50% majority-rule consensus of your MPRs from smalldata_ptme3.nxs.
6. Carry out a bootstrap analysis of your "smalldata_ptme3.nxs" datablock. Ideally you should do 1000 bootstrap reps (approx. 10 minutes), but 500 should suffice. (Note: how different are the bootstrap values if you do 100 reps? 500 reps? 1000 reps?)
7. Print out a bootstrap consensus tree.

Algorithm Used in Phylogenetic Analysis Using Parsimony Version 4 (Paup 4)

A PAUP4 tree search has two components: a method for systematically generating trees that can be constructed from the data and an optimality criterion for evaluating these trees.

Three optimality criteria are available: maximum parsimony, distance, and maximum likelihood. According to the parsimony criterion, the optimal tree is the one that requires the least amount of evolutionary change to explain the data. (Such a tree is also referred to as the most parsimonious tree.) To use distance as a criterion, the program calculates a distance matrix from the aligned sequences and uses the matrix values to compute the sum of the branch lengths for each tree according to the minimum evolution algorithm. The distance criterion regards the optimal tree to be the one with the minimum sum of branch lengths. Maximum likelihood is a statistics-based method. Given a model for evolutionary change, a data set, and one or more trees, this method calculates the likelihood of the data set resulted from each tree. The tree with the highest likelihood is considered to be the optimal tree. For any of these criteria, it is possible that more than one tree will tie for the optimal tree. In that event, all of the optimal trees are reported at the conclusion of the search.

PAUP4 provides three methods to generate trees: an exhaustive search, a branch-and-bound search, and a heuristic search. Each of these three methods creates a candidate tree by adding branches to a partial tree in a stepwise-addition process until a complete tree has been constructed. The difference in the methods is in how the branches are added and what happens after a complete tree has been constructed. Here is a greatly simplified description of each method: *** Exhaustive search**. This method uses a systematic process of adding branches to partial trees to construct *all* possible trees that can describe the data, and computes a score for each complete tree. It is therefore guaranteed to find the optimal tree(s), but for large data sets, the time required may put a considerable dent in your computing budget and may even exceed your run time (Although, the number of sequences affects program run times.) In addition to finding trees, this search produces a frequency distribution of the scores of all possible trees.

*** Branch-and-bound search**. This method is also guaranteed to find the optimal tree(s). Unlike the exhaustive search, the branch-and-bound algorithm constructs trees with some "intelligence" rather than by brute force. It uses the same stepwise-addition process as the exhaustive search, but it computes the score of the partial tree each time it adds a branch. If the score of the partial tree is worse than that of the best complete tree found so far, the algorithm abandons this pathway and backtracks to a previous partial tree to use as the next starting point for adding branches. When the search is short-circuited in this way, a branch-and-bound search is faster than an exhaustive search.

In practice, the increase in speed is dependent on the data and on the optimality criterion. When the optimality criterion is parsimony, a branch-and-bound search is usually faster than an exhaustive search. When the optimality criterion is not parsimony, branch and bound offers little or no improvement over an exhaustive search. There is no way of predicting if the branch-and-bound algorithm will speed up a search. For some data sets, the branch-and-bound search reverts to an exhaustive search.
*** Heuristic search**. This method is not guaranteed to find the optimal tree(s). However, it is the fastest type of search and is the only realistic option for large data sets. An initial complete tree is constructed, either by using one of several stepwise-addition methods or by creating a neighbour-joining tree (see "Tree Reconstruction Using Neighbour-Joining" below). Next, branches and/or sub-trees of this initial tree are swapped to grossly rearrange the tree to see if this improves the score. Several branch-swapping schemes are available. Because the heuristic method provides no guarantees, you should repeat the search using different options for the stepwise-addition and branch-swapping steps to be confident about the results.

Limitation of Algorithm used in Phylogenetic Analysis using Parsimony Version 4 (Paup 4)

Depending on the number of sequences being analyzed, the length, the degree of similarity among sequences, the type of search, and the optimality criterion, PAUP4 tree searches can take from less than a second to days or weeks of computer time. In some cases, the analysis may never come to a conclusion. The number of possible trees grows enormously with the number of sequences, so that four sequences have only three possible trees, seven sequences have 945 possible trees, 10 sequences have over 2 million possible trees, and 11 sequences have over 34 million possible trees. Because of this, the exhaustive and branch-and-bound methods for tree searches ("alltrees" and "bandb" and bootstrap analyses using these methods) should not be attempted for more than 10 or 11 sequences, unless the sequences are very similar or very short, or unless steps are taken that will constrain the search in some way (for example, using a very low upper bound setting for the branch-and-bound search). Searches that use the maximum likelihood criterion can be very slow because of the amount of computation involved. Searches that use the distance criterion and the parsimony criterion are much faster than searches using maximum likelihood (Scheet, and Stephens, 2006). For some data sets, parsimony is faster than distance; for other data sets, the reverse is true. The neighbor-joining algorithm is fastest of all, since it reconstructs a single tree from a star phylogeny rather than creating and evaluating large numbers of trees.

Run Times for Sample Data

The Run Time in PAUP4 for sample data requirements can rise in response to increases in data size and changes in program parameters. This is because of the large amounts of time and computer resources that a tree search can consume. It is not a good idea to run one of the search methods on a large set of newly *aligned sequences.*

METHODOLOGY

For this paper, a reference DNA sequence data (genome) is already available for the species Pigeon Pea Plant (Udensi et.al, 2011).

The first step was to compare the reference sequence data (genome), then alignment, and run analysis by mapping the reads of the data. After the alignment and analysis of the reference sequences, all adenines, cytosines, guanines, and thymines represented by As, Cs, Gs and Ts will be merge to produce an A or T or C or G at the first position rather than have a consensus nucleotide. After the merging, the model parameters were set and each of the methods (Maximum Likelihood, Parsimony and Neighbour-Joining) is run to produce the phylogenetic trees. Figure. 2.

Algorithmic Review

An algorithm is a step-by-step procedure that utilizes a finite number of instructions for automated reasoning and the calculation of a function. In other words, it is an unambiguous specification of how to solve a class of problems. Algorithms can perform calculations, data processing and automated reasoning task. (Wakeley, 2009) So far, many algorithms have been developed to overcome these challenges of sequence data analysis and these have been made available to the scientific community as software packages (Li and Homer, 2010). However, these results in most cases gives different results of the same data in different given software of the same module. Ogban et al (2019)

Results from Existing Analytical Software Tools

For this paper, two (2) analytical software as regards DNA sequence data are considered with respect to Maximum Likelihood, Parsimony and Neighbour- Joining Methods; **M**olecular **E**volutionary **G**enetic **A**nalysis (MEGA) Version 7 and **P**hylogenetic **A**nalysis **U**sing **P**arsimony Version 4 (PAUP 4)
Running the analysis and alignment, the reference DNA sequence data (genome) for species of Pigeon Pea Plant (Udensi et al; 2011) is used. The following phylogenetic trees in Figure 3 represent the results from running the analysis of this sequence data using these two analytical tools.

Software packages on DNA sequence data analysis

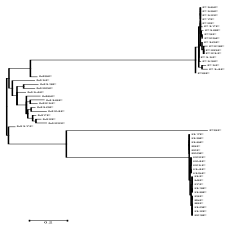
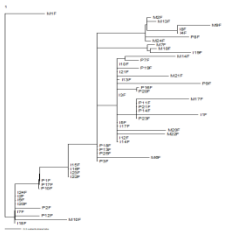
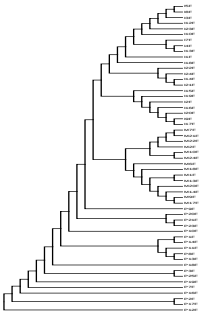
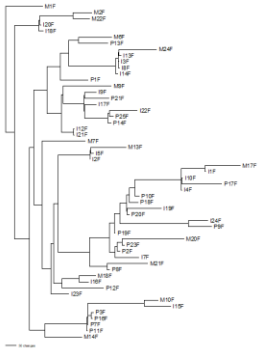
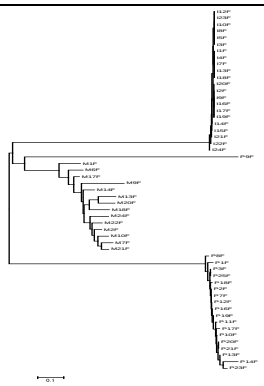
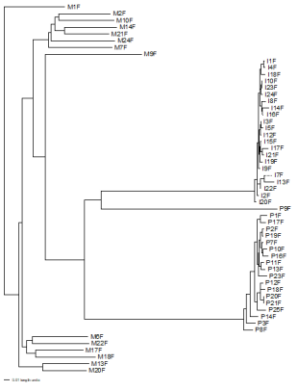
Molecular Evolutionary Genetics Analysis (Mega) Version 7.	**Phylogenetic Analysis Using Parsimony Version 4 (Paup 4)**
 MEGA Phylogenetic Tree of Pigeon Pea Plant based on the three genes (MATK, ITS and PETB) using Maximum likelihood Method. (a)	 Paup4 Phylogenetic Tree of Pigeon Pea Plant based on the three genes (MATK, ITS and PETB) using Maximum likelihood Method. (b)
 MEGA Phylogenetic Tree of pigeon Pea Plant based on three (MATK, ITS and PETB) genes using Parsimony Method. (c)	 PAUP 4 Phylogenetic Tree of Pigeon Pea Plant based on three genes (MATK, ITS and PETB) using Parsimony Method. (d)
 MEGA Phylogenetic Tree of Pigeon Pea Plant based on three genes (MATK, ITS and PETB) using Neighbor-Joining Method. (e)	 PAUP 4 Phylogenetic Tree of Pigeon Pea Plant based on three genes genes (MATK, ITS and PETB) using the Neighbor-Joining Method. (f)

Fig. 2. a,b,c,d,e,f: Phylogenetic Trees from Analyzed Sequence Data from Pigeon Pea Plant

Table 2. Showing the result of Analysis and Alignment of Pigeon Pea Plant (MAT K, PET B and ITS) gene using Maximum Likelihood, Parsimony and Neighbour-Joining Methods

Analytical Software Package	Analytical Method	Clustering Resolutions	
		Major Clusters	Sub-Clusters
Molecular Evolutionary Genetic Analysis (MEGA) 7	*Maximum Likelihood*(ML)	3	3
Phylogenetic Analysis Using Parsimony (PAUP) 4	*Maximum Likelihood(ML)*	1	16
Molecular Evolutionary Genetic Analysis (MEGA) 7	*Parsimony(Pars)*	2	11
Phylogenetic Analysis Using Parsimony (PAUP) 4	*Parsimony(Pars)*	2	16
Molecular Evolutionary Genetic Analysis (MEGA) 7	*Neighbour-Joining(NJ)*	2	2
Phylogenetic Analysis Using Parsimony (PAUP) 4	*Neighbour-Joining(NJ)*	2	3
Total Clusters number.		**12**	**51**

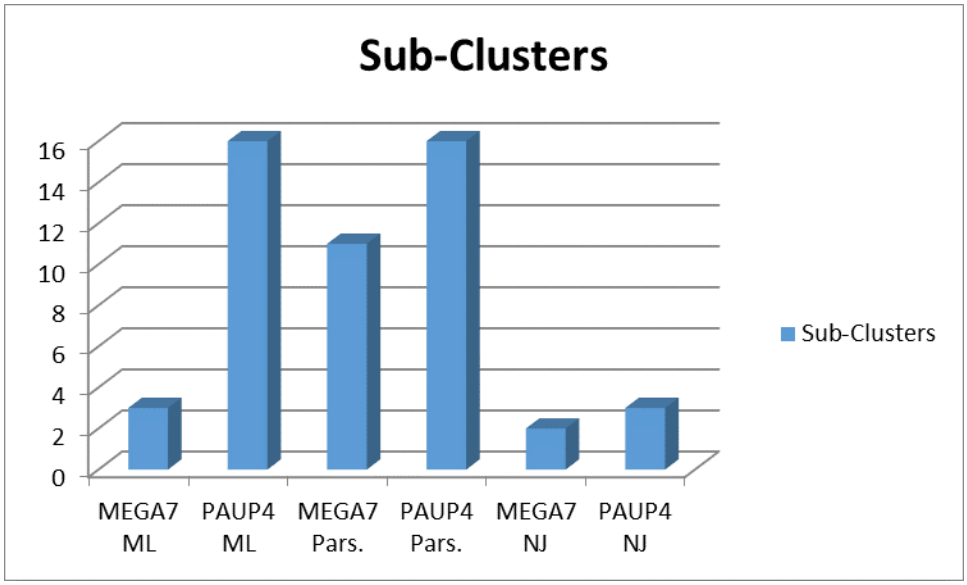


Fig. 3. Column plot of the Sub-clusters readings of the tools using ML, Pars and NJ.

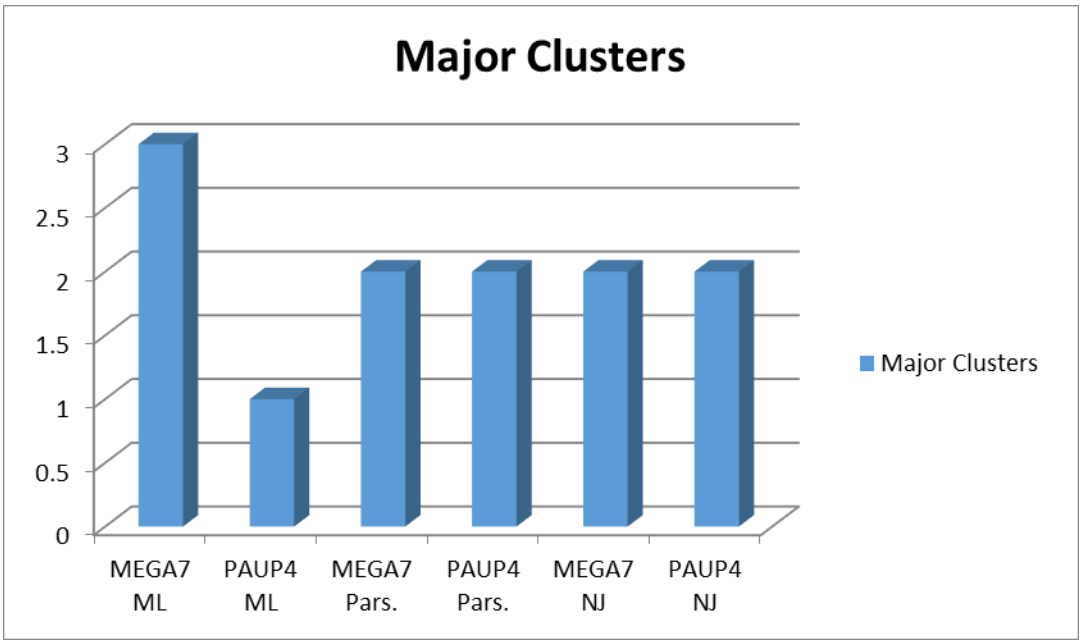


Fig. 4. Column plot of the Major clusters readings of the tools using ML, Pars and NJ.

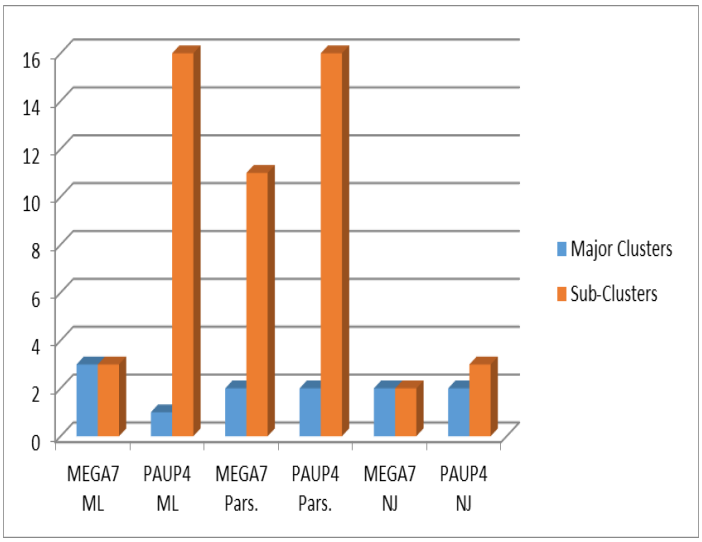


Fig. 5. Comparative Column plot of the Major and Sub- clusters readings of the tools using ML, Pars and NJ.

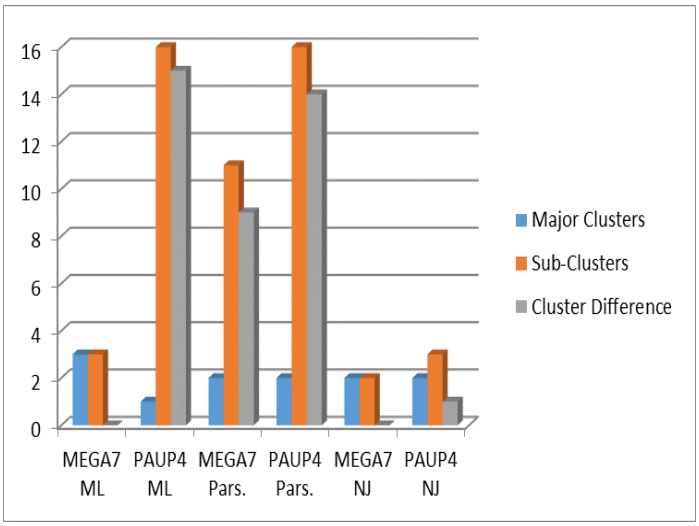


Fig. 6. Comparative Column plot of the Major, Sub and cluster difference readings of the tools using ML, Pars and NJ.

DISCUSSIONS

MEGA7.0 Using Maximum Likelihood Method

For the phylogenetic tree as shown in Figure. 4 using MEGA 7.0 software, there are three major clusters. Cluster one had two sub-

clusters with all ITS genes while only one pet-B (P9F) occupied the second sub-cluster. The second major clusters consist of only Maturase k gene (M17F). The third cluster consists of only Maturase k on one of its sub-clusters. However, one Maturase k gene (M9F) was found in the sub-cluster containing the pet-B gene as an out-group. See table 2, Fig. 3,4,5,6.

PAUP 4 Using Maximum Likelihood Method

In same Figure 2, it was revealed that the maximum likelihood tree from PAUP 4 generated a tree where there were no specific grouping of samples based on the genes. Fig. 3,4,5,6.

MEGA 7.0 Using Parsimony Method

The phylogenetic tree constructed by MEGA 7.0 using the parsimony method generated two major clusters. The first cluster had only one pet-B gene (P12F). All the remaining samples were divided into sub-groups under the second major cluster as shown in Figure 2, 3, 4, 5, 6

PAUP 4 Using Parsimony Method

The parsimony method of PAUP 4 also generated a tree with two major clusters. However, the samples were not grouped by the type of gene.

MEGA 7.0 Using Neighbour –Joining Method

The phylogenetics of the 3 genes in the pigeon pea constructed using MEGA 7.0 revealed two major clusters. The first cluster was made up of all the Pet-B genes. The second major cluster had two sub-clusters where all the Maturase k genes were grouped together with one pet-B gene as an out-group. The second sub-cluster had only the ITS genes.

PAUP 4 Using Neighbour-Joining Method

Using the neighbour joining method, PAUP 4 generated a tree with two major clusters. All the ITS genes were grouped together as well as the pet-B gene. However, the Maturase k gene were divided into two sub-groups as shown in figure 2, 3, 4, 5, 6.

In summary, it can be seen that the phylogenetic trees produced vary in each of the methods from Molecular Evolutionary Genetic Analysis and Phylogenetic Analysis Using Parsimony showing disparity in the algorithms used in these analytical software package.

Proposed Sequence Analysis and Alignment Algorithm (SeAAA)

The proposed Sequence Analysis and Alignment/merger Algorithm (SeAAA) is to include sequence analysis and alignment/merger properties of sequence data (genome). The system will simulate short read sequencing of sequence data, runs them for a given set of configurations, align and merge the output of each simulation.

Implementation

The proposed Sequence Analysis and Alignment Algorithm (SeAAA) will provide modules for analysis and alignment/merger of sequence data for efficiency, speedy runtime performance and accuracy.

The algorithm, as comprehensive as it will be, will be implemented in Java with several utilities that can be used, including:

- Reading a pre-existing reference DNA sequence data from one or more FASTA files.
- Generate a reference 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 alignment/merger tools of the Sequence Analysis and Alignment Algorithm (SeAAA) to the sequence data and the reads through a standardized interface.
- Parse the output of the alignment/merger tool and calculate the number of reads that were correctly or incorrectly mapped.
- 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. This repeat variability rate is intended to introduce variability in the potential mappings of a read. Repeats

are quite common in real genomes (Cheung *et al.*, 2003), but of essence is the speed in performance runtime and accuracy.

Expected results from the proposed Sequence Analysis and Alignment Algorithm (SeAAA) should show significant differences in runtime performance and accuracy as the number of the reads of the sequence data decreases.

CONCLUSION

In considering and investigating the suitability of the algorithm(s) in existing analytical software packages in DNA sequence data analysis, a comparative analysis of their results and runtime performance in respect to alignment and authentic statistical basis, shows dependence on guesses, less accurate and slowness in runtime performance.

In this paper, we have explored two existing sequence data software analytical tools with focus on phylogeny using maximum likelihood, parsimony and neighbor-joining methods. Our result revealed inconsistency in the algorithms used in these analytical tools. As part of our future work, we will explore other sequence data software analytical tools in other areas of this research work or investigation.

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 the statistics to develop a unified analytical tool for sequence data analysis.

REFERENCE

Alkan C, Kidd JM, Marques-Bonet T, Aksay G (2009) Personalized copy number and segmental duplication maps Next-generation sequencing. *Nat. Genet.*, **41**:1061–1067.

Auffray C., Michael H. Sieweke, Frederic Geissmann (2009) Blood Monocytes: Development, Heterogeneity, and Relationship with Dendritic Cells Annual Review of Immunology
First published online as a Review in Advance on January 8, 2009 https://doi.org/10.1146/annurev.immunol.021908.132557

Cheung V. G, Jen K. Y, Weber T, Morley M, Devlin J., k.g.Ewens K G, Spielman A.S (2003) Genetics of Quantitative Variation in Human Gene Expression

Guffanti A, Iacono M, Pelucchi P, Kim N, Soldà G, (2009). A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics*, **10**:163–179

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.

Kimura,M. (1980) A simple method for estimating evolutionary rate ofbase substitutions through comparative studies of nucleotide sequences.J. Mol. Evol., 16:111–120.

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.

Li,H, Homer,N. (2010) A survey of sequence alignment algorithms for next generation sequencing. *Brief. Bioinformatics*, **11**:473–483.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. InMunro (ed.), Mammalian Protein Metabolism. Academic Press, NewYork, pp. 21–132

Miller,J.R, Koren S, Sutton G, (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**:315–327.

Ogban F.U, Udensi U.O, Inyang G. (2016) Diversity in Content and Analytical Algorithms of Biologist-Centric Software(s) for DNA and Sequence Data: Mega, DNASp, GenAlex and ARLEQUIN.

Qin,J,Li R, Raes J, Arumugam M (2010) A human gut microbial gene catalogue established by metagenomicsequencing. *Nature*, **464**:59–65.

Scheet, P. and Stephens,M. (2006) A fast and flexible statistical model for large scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*,**78**: 629–644.

Schuster,S.C. (2007) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**:16–18.

Sultan, M, Schulz MH, Richard H, (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* ,**321**:956–960.

Tamura K, Nei M, (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees Mol. Biol. Evol., 10 (93):512-526

Tamura, K (1992) Two parameters method for the Estimation of G+C content biasMol. Biol. Evol., 10 (93):12-26

Taylor,K.H, Kramer RS, Davis JW, Guo J, Duff DJ, Xu D, (2007) Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.*, **67**: 8511–8518.

Udensi O, Edu E A, Umana E J, Ikpeme E. U (2019). Estimation of Genetic Variability in Locally Grown Pulses (Cajans cajan (L.) Millsp and Vigna unguiculata (L.) Walp): A Panacea for Sourcing Superior Genotypes. Pakistan Journal of Biological Sciences 14(6):404-407.

Van Tassell, C.P., T.P.L. Smith, L.K. Matukumalli, J.F. Taylor, R.D. Schnabel, C. Taylor Lawley, C.D. Haudenschild, S.S. Moore, W.C. Warren, T.S. Sonstegard (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries Nat. Methods, 5 (2008), pp. 247-252

Wakeley, J. (2009). *Coalescent Theory. An Introduction*. Roberts and Company Publishers. Greenwood Village.