A Reliable NJ Technique of Different Alignments

Mina Basheer*, Sawsan Kamal**

* computer science department, AL-Nahrain University, Iraq,minacom_84@yahoo.com
**computer science department, college of sciences, AL-Nahrain University, Iraq, skt@sc.nahrainuniv.edu.iq

Abstract— The rapid evolution of DNA sequencing technologies and the high amounts of molecular data used for analayzing phylogenetic tree take a great consumption in the computational performance as well as the memory demands. The evolutionary tree design is required for determining the relationship between the DNA sequences based on the differences within their genetic characteristics in addition to the extent of divergence between species. The major objective of phylogeny reconstruction is to describe evolutionary relationships in terms of relative recency of common ancestry. Although the pairwise sequence alignment (PSA); which defines a Local (LA) and Global alignment (GA), guarantees an optimal alignment for DNA sequences, it lacks in speed when used in huge sequences. Also, the multiple sequence alignment (MSA) guarantees a fast result, but lacks in the accuracy. Therefore, in this work both methods were utilized to determine the best alignment based on the shortest distance for each DNA sequence with a fast and a high accuracy. Where, the obtained result represnts an input to the utilized distance-based Neighbor-Joining (NJ) algorithm; which is the most common and fastest method for reconstructing a phylogenetic tree from the data sequences.

Index Terms— DNA, GA, LA, MSA, NJ, phylogenetic tree, PSA, sequence alignment.

1 Introduction

NOWADAYS, bioinformatics is a predominant expression of scientific research. A major measure of implementation has begun in this field. In reality, human genes consist of blocks of amino acids or proteins. Therefore, the main challenge is to analyze these sequences and infer information [1].

Bioinformatics could be expressed as the statistical, computing and mathematical methods pursue to solve biological problems implementing DNA and amino acid sequences and other relevant information. Bioinformatics might be named as an umbrella of several fields such as computer science, statistics, biology, chemistry, mathematics, etc. [1]. Phylogenetic or evolutional tree is a data structure demonstrating evolutionary relationships between biological species, or other entities. These relationships are named as the phylogeny of the species according to the differences or similarities in their genetic or physical characteristics [2].

The main objective of the study of molecular evolution relationships is to correct the reconstruction of the evolution history based on the convergence and the divergence of sequences between organisms; which means finding the correct tree chart with the lengths of the correct branches [3]. A phylogeny is the evolutionary history of a set of sequences, the primary objective of phylogeny reconstruction is describing the evolutionary relationships in terms of relative recency of common ancestry. These relationships can be represented as a branching diagram, or tree, with branches joined by nodes and leaded to terminals in the tips1of the tree. A grouping that contains a common ancestor and all the descendants (living and extinct) of that ancestor is called a clade. Fig. 1 shows a phylogeny of ARP gene. A monophyletic group is formed by Osrs2, the sister terminal to Zmrs2. According to the monophyletic sister group, PsPHAN and AtAS1 are paraphyletic containing AmPHAN, NtPHAN and LePHAN. 'Clade' and 'monophyletic group' can be used interchangeably, so as the 'grade' and 'paraphyletic group'. Nodes, terminals, branches, a clade, and a grade are indicated in [4].

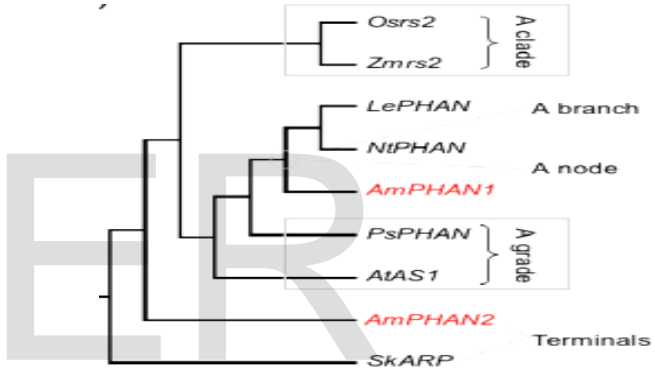

Fig. 1. Some phylogenetic expression [4].

From sequence analysis, phylogenetic empowers scientists to utilize different methods to infer evolutionary relationships. The phylogenetic analysis techniques generally generate phylogenetic trees that provide relationship between any set of species or their evolutionary history. From phylogenetic analysis, the most accurate tree describing the evolution of a sequence can be obtained by a DNA, RNA or amino acid sequences.

Beginning in the late nineteenth century, the analystes utilized the morphological features (e.g. state of quills, number of toes, and so on.) to rebuild the evolutionary histories. The sequencing innovation became less expensive and more accessible in the era between 1970's and 1980's. Also, the DNA and protein sequences became the main data source for analyzing and evolution the existing species. One of the fundamental reasons of the change from morphological to molecular characters was the accessibility of institutionalized computational strategies for reconstructing and analyzing molecular information [5].

Rebuilding the phylogenetic trees is useful in determining the relationship between DNAs sequences based on the differences within their genetic characteristics by implementing NJ.

2 Related Work

In [6], an algorithm has been introduced by Olson et al.; where a tree consisting of n taxa was constructed using a stepwise addition algorithm. An LA method was implemented to choose the best tree from each step which then gives a search space for likely trees. For n taxa a GA was done; where it scales so good with nearly linear speedups. Although, the method in [7] were considered to be obsolete, another version of this method was presented in [8].

In [9], a method has been introduced by Guindon and Gascuel that describes the phylogeny reconstruction. Based on the ML method, the simultaneous adjustment of tree topology and length of branches is the real-work of this method. Where it modifies the tree, at each iteration, to enhance the likelihood begining from an initial tree that is constructed by a fast distance based algorithm. Although, this method is topologically precise, it produces a randomness in the search; because of the intensive topological re-alignments.

A platform in [10] was developed by Dereeperet al. that perform multiple sequence arrangement, with the phylogenetic reconstruction and graphical demonstration of the inferred tree. This technique was done for the expert and non-expert persons that works in three modes. It gives a ready to use pipelining chain programs with activating all the parameters by default for the non-experienced user; in which it utilizes MUSCLE [11] for multiple sequence arrangement, Phyml [9] for tree building, and Gblocks [12] for auto- alignment. While, the option of editing the setting upon the need is done in the advance mode. The capability of running and testing the large methods in an efficient way is done in the 'A la Carte' mode; while keeping the overall interface similar to the advance mode. However, it works on a dedicated server with i/p & o/p limitations for some programs.

Our program is written in Python language and is independent of architecture and operating system. Hence it can be implemented in multiple environments. Also, the system has a simple implementation that works well for everyone with low cost as it does not require a dedicated hardware.

3 Methodology Of The Proposed System

The algorithm of this system consists of several steps as shown in Fig. 2. The proposed system creates three difference matrices of different alignments such as MSA and PSA. The NJ was implemented for computing the length of branches trees that will be used to build the requested phylogenetic trees for adenoviruses DNA sequences.

The Reconstruction the phylogenetic relationship is a hierarchal process that includes five levels:

1. The selection of DNA sequences.
2. The Sequence alignment.
3. The calculation of Distance matrix.
4. NJ Method of tree building.
5. The assessment of the resulting phylogeny.

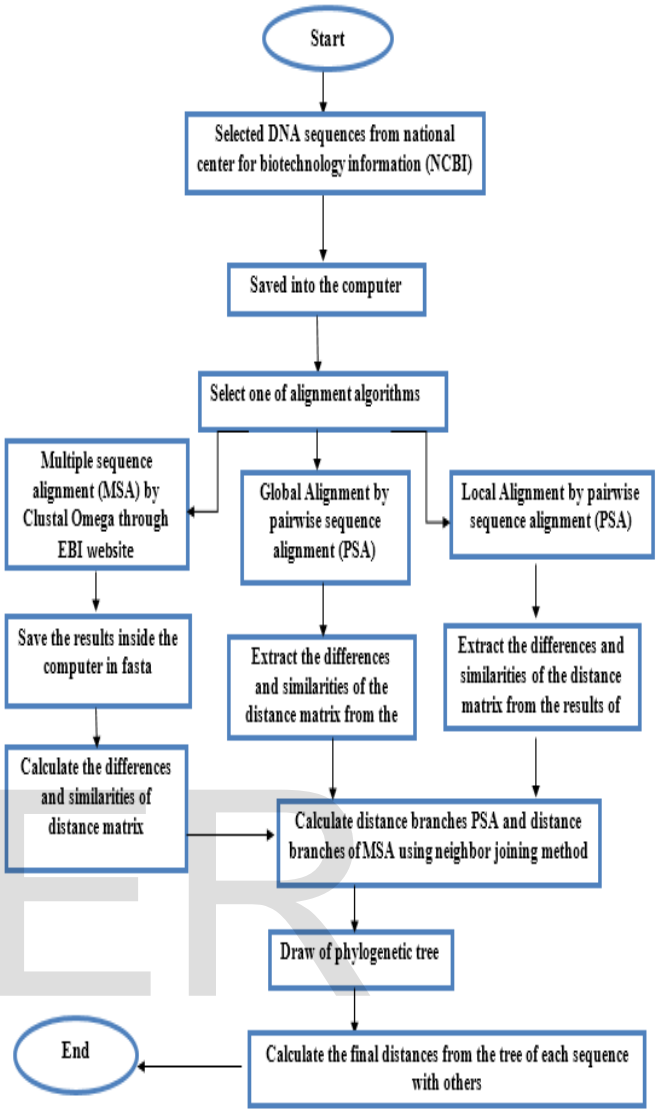

Fig. 2. General flowchart of system

3.1 The Selection of DNA Sequences

According to an advice from a bio science expert, 13 successive Adenoviruses AD (HAdv) were selected for one gene of Hexon worldwide type (Japan, Germany, Russia, etc) to determine the type of relationship (for human) between viruses and discover the origin of the virus.

3.2 The Sequence Alignment

Different types of sequence alignment have been applied such as the PSA (Local and GA) and MSA. The process of sequence alignment represents a real problem due to different reasons such as:

- The different in the length between two sequences.
- There may be only a relative small region between the matched sequences.
- The regions of variable lengths may have been inserted /deleted from the common ancestral sequence.

3.3 Distance Matrix Calculation

In this work, hamming distance has been implemented to compute the distance matrix of the difference between the corresponding DNA sequences; which represents the real evolutionary distance under estimation, while the possibility that multiple unseen mutations might have taken place at separated sites was not considered into account. The main benefit of such distance-based approaches is their computational speed.

These approaches are optimally appropriate for the initial exploration of evolutionary relationships between sequences within a dataset. The percentage of each pairwise sequences was calculated by dividing the distance between each pairwise sequences by the total length of the sequences after the alignment process and the overall result was multiplied by 100 as shown below:

$$P = \{dist\ (s_{ij}) / L\ (s_{ij})\} \times 100 \quad (1)$$

Where P defines the percentage value and sij is the two sequences, dist (sij) represents a pairwise distance of sequences and L defines the Length of pairwise sequences. Fig. 3 presents the algorithm of distance matrix calculation:

Input: FASTA file of Alignment:

- MSA
- Local sequence alignment
- Global sequence alignment

Output: Distance matrix:

Begin: Read the file as lines:
Step 1: identify matches, mismatches, and gaps:
Step 2: define m=matrix:
Step 3: gap removal in the beginning and end of the aligned sequences:
Step 4: compute length of DNA sequences after gap removal:
Step 3: For loop:
Diff=0 % number of mismatches # **Hamming distance calculation**
Match=0
If base1~=base2
Diff=Diff+1
Else
Match=Match+1
Step 4: Divide Diff and Match by length of sequence (after gap removal) and obtain the percentage of Diff and Match:
Step 5: Store the distance matrix m:
Step 6: Save m into the computer as a text file.
End.

Fig. 3. Algorithm of the distance matrix calculation

3.4 NJ Method of Tree Building

After calculating the difference distance matrix, the conventional and NJ methods were used to compute the branches of the phylogenetic tree. It calculates the Q matrix for selecting the smallest (difference) value to be the reference that specifies the smallest value within the distance matrix of two DNA sequences. Fig. 4 demonstrates the steps of NJ calculation. The Q matrix and the distance between the new node with the sequences i and j have been computed as shown in the equations below:

$$Q\ (i,j) = (n-2) \times d(i,j) - \sum^{n}_{k=1} d(i,k) - \sum^{n}_{k=1} d(j,k) \quad (2)$$

Where, d(k ,n) defines the distance between sequences k & n.

$$dist_k = dist/2 + diff\_R / (2 \times (n-2)) \quad (3)$$

$$dist_n = dist - dist_k \quad (4)$$

Where, $dist_k$ defines the distance from node k to the new node, and $dist_n$ defines the distance from node n to the new node.

$$d(x,s) = 0.5 \times [d(y,s) + d(z,s) - d(y,z)] \quad (5)$$

Where, x defines the new node, s represents the node that is needed to compute the distance to and (y and z) are the joined pair-members

Input: Text files of Distance matrix:
Output: distance of branches of phylogenetic tree:

Begin: read the fasta file of DNA sequences
Step1: create a dictionary function for save all nodes and their branches
Step 2: compute matrix length:
Step 3: Sum the length of every sequence in the matrix:
Step 4: compute number of sequences (n)
Step 5: if n = 2 stop operation and save the result in the dictionary.
Step 6: Q matrix of $\hat{d}$ calculation based on Q matrix equation:
Step 7: Specify the criteria for choosing the smallest value of distance in DNA sequences:
Step 8: Calculate the pair of sequences i and $j (i \neq j)$ for which $Q(i,j)$ has its lowest value. These sequences are connected to a newly built node, which is connected to the central node.
Step 9: Compute the distance for every sequence in the pair to the new node.
Step10: distance Computation from each of the sequence outside of this pair to the new node.
Step12: After the new node creation, remove the old sequences related to the new created node
Step 13: Repeat the algorithm, replacing the pair of connected neighbors with the new node and implemented the distances computed in the previous step.
End.

Fig. 4. NJ calculation

3.5 The Assessment of the Resulting Phylogeny

This sections measures the accuracy by comparing the results of the acquired distances, resulted from the NJ (output of nj), to the real results of the distance matrix after the alignment process (input of nj). The Comparison between the true and estimated distances is done as illustrated in Fig. 5. For example, the distance matrices of LA differences tree for HAdv gene were obtained as follows:

S0, S10 =2.47 as input, 2.171 as output
S11, S12 = 16.34 in input and output
S2, S3 = 0.38 as input, 0.34

While the obtained results of GA differences tree for HAdv gene were as follows:

S11, S10 =5.77 as input, 5.99, 5.89 as output
S5, S8 =2.23 as input and output
S1, S7 = 8.25 as input, 8.92 as output

According to the above results, it is clear that the estimate distances were completely match to the real distances. Therefore, this method has proved that it has the capability to find the real tree. The steps of comparison process between the true and estimated distance matrix is shown in the Fig. below.

Input: true and estimated distance matrix

Output: the accuracy of estimated distance matrix

Step1: Compute r and $\hat{r}$ r and $\hat{r}$ denote the total distances of sequence (A) to all other sequences of true and estimated distance matrix respectively.

Step2: Create matrix for new distances according to the computed distances $\hat{d}$

Step3: Compare the estimated matrix with the true distance matrix to evaluate the NJ

End.

Fig. 5. The comparison steps between the True and Estimated values

4 Results & Discussion

The algorithm of this work consist of a hybrid technique that includes the fast and conventional NJ to build a phylogenetic tree of DNA sequence corresponding to viruses (adenovirus type Hexon) affect human. Three types of sequence alignments were used to create the phylogenetic tree. These alignments are: MSA and PSA (Local and GA). The goal of using these alignments is to select the alignment that gives a better similarity between the sequences and to compute the differences of matrix for building the phylogenetic tree using the NJ method. The analysis and discussion of each step of this works are demonstrated in the following sections.

4.1 Experimental Tools

The codes of the proposed system are implemented by the programming language "Python" under the environment of windows7 ultimate 32-bit operating system inside a laptop computer with: Intel ® Core (TM) i5-4210U CPU @1.70 GHz 2.40 GHz and RAM14GB.

Python has the ability to handle specialized libraries such as BioPython that provide special and accurate functions in the field of bioinformatics. Biopython features comprise parsers for several Bioinformatics file formats such as (BLAST, Clustalw, FASTA, Genbank, etc.). In this work, the Python has been applied and made in a simplified form for bioinformatics by using high-quality, reusable scripts and modules.

4.2 The Results of Sequence Alignment

In the bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify the regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that the identical or similar characters are aligned in successive columns. Sequence alignments are also used for non-biological sequences, such as calculating the edit distance cost between strings in a natural language or in financial data.

4.2.1 MSA

MSA is the progressive algorithm. Clustal omega alignment is used in the process of MSA and the results are stored in a text file (FASTA) to be processed and used subsequently in the proposed system.

Table 1 shows the obtained results after applying the MSA for thirteen different DNA sequences and computing the difference matrices of sequences. MSA is convenient for aligning different length sequences of distance matrix value ranging from 61% up to 99.5%.

TABLE 1
THE MSA DISTANCE MATRIX PERCENTAGE (%) IN 30 SEQUENCES' DIFFERENCES OF (HADV) GENE

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S0	0.00	19.40	20.18	20.70	20.57	21.48	21.09	21.22	20.83	20.83	20.96	20.83	20.83
S1		0.00	3.15	2.84	2.77	3.18	2.80	2.73	2.90	2.63	2.88	2.56	2.84
S2			0.00	2.47	2.47	3.89	2.50	2.37	2.25	2.26	2.13	2.19	2.09
S3				0.00	0.17	1.85	1.54	1.41	0.97	1.23	0.91	1.23	0.86
S4					0.00	1.81	1.51	1.37	0.97	1.20	0.91	1.20	0.86
S5						0.00	0.89	0.74	1.09	0.67	0.91	0.59	0.86
S6							0.00	0.34	0.52	0.38	0.43	0.31	0.38
S7								0.00	0.19	0.24	0.16	0.17	0.11
S8									0.00	0.13	0.13	0.06	0.06
S9										0.00	0.11	0.07	0.05
S10											0.00	0.05	0.05
S11												0.00	0.00
S12													0.00

The alignment process is fast as it processes several sequences not only for pairwise sequences but for all sequences in one single time. The result of this alignment which is obtained by the clustal omega offers a sequence arrangement that has gaps at the beginning and end of each sequence pairs. So, in this work, these gaps were removed using python code to compute the distance matrix of differences without taking these gaps into account. Table 1 computes the distance matrix of difference.

4.2.2 LA

In this section, the calculation of two sequences by Smith-Waterman algorithm was implemented to obtain the best possible LA taking into account the calculated scoring system. It includes a gap scoring method and a substitution matrix. The Score considers the substitution, match, and mismatch. The alignment process is demonstrated as shown below.

1. Specifying the similarity threshold between pairwise sequences. In this work, five differences were chosen as a criteria for eliminating the dissimilar sequences.
2. The values of match, mismatch, and gaps were 2,-1,-4,-5.
3. Choosing the best similarity between pairwise sequences.

Table 2 and 3 shows the results of the computed difference,

and pairwise length matrices respectively, after applying the LA for thirteen different DNA sequences.

TABLE 2
THE LA DISTANCE MATRIX PERCENTAGE (%)

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S0	0.00	2.20	2.27	2.51	2.37	2.10	2.26	2.89	2.14	2.47	2.47	3.15	17.47
S1		0.00	0.07	0.31	0.17	0.00	0.06	0.59	0.05	1.24	1.20	2.46	17.77
S2			0.00	0.38	0.24	0.05	0.13	0.67	0.11	1.24	1.20	2.53	17.77
S3				0.00	0.34	0.34	0.52	0.89	0.43	1.55	1.51	2.71	17.77
S4					0.00	0.11	0.19	0.74	1.16	1.41	1.38	2.64	18.16
S5						0.00	0.06	0.86	0.05	0.86	0.86	2.70	19.15
S6							0.00	1.10	0.13	0.97	0.97	2.72	21.58
S7								0.00	0.91	1.86	1.82	3.08	18.28
S8									0.00	0.91	0.91	2.73	19.64
S9										0.00	0.17	2.74	17.65
S10											0.00	2.67	17.52
S11												0.00	16.34
Seq12													0.00

TABLE 3
LENGTHS BETWEEN THE SEQUENCES AFTER LA

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S0	2910	2910	2910	2910	2910	1858	1547	2697	1872	2910	2910	2886	784
S1	2910	2907	2907	2907	2907	1855	1544	2697	1869	2907	2907	2883	782
S2	2910	2907	2907	2907	2907	1855	1544	2697	1869	2907	2907	2883	782
S3	2910	2907	2907	2907	2907	1855	1544	2694	1869	2908	2907	2883	782
S4	2910	2907	2907	2907	2907	1855	1544	2694	1869	2907	2907	2883	782
S5	1858	1855	1855	1855	1855	1855	1855	1544	1857	1855	1855	1855	726
S6	1547	1544	1544	1544	1544	1544	1544	1546	1544	1544	1544	1544	644
S7	2697	2694	2694	2694	2694	1857	1546	2692	1871	2695	2695	2694	755
S8	1872	1869	1869	1869	1869	1842	1544	1871	1869	1869	1869	1869	713
S9	2910	2907	2907	2908	2907	1855	1544	2695	1869	2907	2907	2883	782
S10	2910	2907	2907	2907	2907	1855	1544	2695	1869	2907	2907	2883	782
S11	2886	2883	2883	2883	2883	1855	1544	2694	1869	2883	2882	2879	765
S12	784	782	782	782	782	726	644	755	713	782	782	765	753

Table 3 shows the way of aligning the sequences based on the similar length with obtaining the subsequences that lead to the best similarities. The implementation of LA is somehow slow; because its computation in finding the similar areas is more difficult. The LA depends on the pairwise technique that deals with one of the entire sequences as a reference sequence then compare it with other sequence, based on the specified threshold to remove the dissimilar region beyond it and align it with the rest of sequence. This algorithm shows an acceptable accuracy for different length of sequences. The accuracy value ranges between 78.42 and 99.93 as shown in the previous table.

4.2.3 GA

This algorithms has been applied to the HEXON gene of the adenovirus organism that is selected from different regions of the world

After applying the GA for thirteen different DNA sequences and computing the difference matrix, the results of difference matrix, and pairwise length matrix are shown in Tables 4 and 5 respectively. Fig. 6 shows part of the GA process for different length, note that the alignment process has been implemented between S2 of 2907 base, and S12 of 753 base; where S12 was extended by a gab to equalize it with S2.

```
s2 :  ATGGCTACCCCTTCGATGATGCCGCAGTGGTCTTACATGCACATCTCGGGCCAGGACGCCTCGGAGTACCTGAGC
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
s12 : ATGGCTACCCCCTCGGAG----T-A--------C-----C----T--------------GA-G--C--C--CC-G

Score match : 753

Missmatch =  74.097% match= 25.9% end length 2907
```

Fig. 6: part of GA

TABLE 4
LENGTH BETWEEN THE PAIRWISE OF SEQUENCES AND THE DIAGONAL OF THIS TABLE REPRESENTED OF BLUE COLOR THE REAL LENGTH OF EACH SEQUENCES

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S0	2910	2971	2973	2980	2976	2940	2931	2979	2941	2979	2979	2994	2910
S1		2907	2909	2916	2912	2907	2908	2920	2908	2943	2942	2973	2907
S2			2907	2918	2914	2907	2908	2921	2909	2943	2942	2975	2907
S3				2907	2917	2911	2911	2928	2912	2951	2951	2980	2908
S4					2907	2909	2910	2924	2910	2948	2947	2978	2909
S5						2909	2910	2924	2910	2948	2947	2978	2909
S6							1544	2700	1871	2918	2918	2905	1646
S7								2692	2706	2949	2948	2958	2701
S8									1869	2923	2923	2920	1939
S9										2907	2912	2981	2907
S10											2907	2979	2907
S11												2879	2884
S12													753

TABLE 5

THE PERCENTAGE OF DISTANCE MATRIX OF DIFFERENCES OF SEQUENCES IN GA OF THIRTEEN SEQUENCES OF (HADV) GENE

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S0	0.00	4.21	4.34	4.80	4.54	37.93	48.04	11.95	37.50	4.73	4.73	6.65	74.12
S1	4.21	0.00	0.14	0.62	0.34	36.19	46.94	8.25	35.76	2.45	2.38	5.38	74.1
S2	4.34	0.14	0.00	0.75	0.48	36.19	46.94	8.32	35.82	2.45	2.38	5.51	74.1
S3	4.80	0.62	0.75	0.00	0.69	36.41	47.10	8.78	36	2.98	2.98	5.84	74.14
S4	4.54	0.34	0.48	0.69	0.0	36.30	47.04	8.52	35.88	2.78	2.71	5.71	74.18
S5	37.93	36.19	36.19	36.41	36.30	0.00	16.86	31.90	2.23	37.03	37.03	37.77	64.38
S6	48.04	46.94	46.94	47.1	47.04	16.86	0.00	43.11	17.58	47.46	47.46	47.75	60.45
S7	11.95	8.25	8.32	8.78	8.52	31.9	17.58	0.00	31.45	10.14	10.07	11.66	72.45
S8	37.5	35.76	35.82	36	35.88	2.23	43.11	31.45	0.00	36.61	36.61	37.4	64.78
S9	4.73	2.45	2.45	2.98	2.78	37.03	47.46	10.14	36.61	0.00	0.34	5.9	74.1
S10	4.73	2.38	2.38	2.98	2.71	37.03	47.46	10.07	36.61	0.34	0.00	5.77	74.1
S11	6.65	5.38	5.51	5.84	5.71	37.77	47.75	11.66	5.90	37.4	5.77	0.00	74.06
S12	74.12	74.10	74.10	74.14	74.18	64.38	60.45	72.45	64.78	4.10	74.10	74.06	0.00

As shown in the table above, the percentage of sequence differences in accordance to sequence12 is very high; which represents a high divergence. As result, it leads to an issue when doing the alignment process. Therefore, the sequences end up with an alignment of very large gaps.

4.3 Phylogenetic Trees Construction using NJ

In this work, NJ method is implemented to construct a tree for all alignment methods of aforementioned sequences' differences.

The Fig. 7, 8 and 9 (with their related tables) demonstrate the connection between branches that creates a new node, where the red squares represent the group sisters. For example, the sequences (3 & 7), (9 & 10), and (3 & 4) were grouped as a sister group for MSA, LA and GA respectively. Table 9 gives the finals results after implementing NJ method to the sister group of each tree.

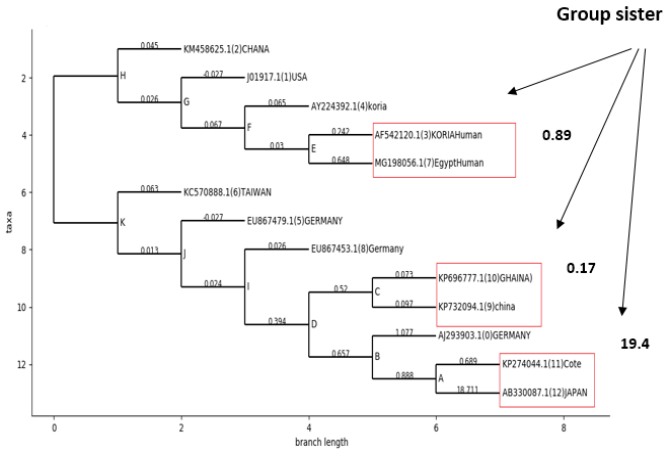

Fig. 7 NJ for construct tree after MSA (Differences)

TABLE 6

MSA FINAL RESULTS OF DIFFERENCES TREE FOR HADV GENE RELATED TO THE PREVIOUS FIG

Nodes	KC570888	EU867479.	EU867453.1(	AJ293903.1	KP274044.1(	AB330087.1	KP696777.1(	KP732094.1	KM4586	J01917.:	AY2243	AF5421	MG1980
	(6)	(5)	(8)	(0)	(11)	(12)	(10)	(9)	(2)	(1)	(4)	(3)	(7)
(9)china	1.111	1.008	1.037	2.351	2.851	20.873	0.17	0	1.093	1.047	1.206	1.413	1.819
(8)Germany	0.126	0.023	0	2.154	2.654	20.676	1.013	1.037	0.108	0.062	0.221	0.428	0.834
(7)EgyptHumon	0.831	0.757	0.831	2.036	3.436	21.458	1.795	1.819	0.816	0.768	0.743	0.89	0
(6)TAIWAN	0	0.049	0.126	2.228	2.728	20.75	1.087	1.111	0.108	0.062	0.221	0.428	0.834
(5)GERMANY	0.049	0	0.023	2.125	2.625	20.647	0.984	1.008	0.031	-0.015	0.144	0.351	0.757
(4)koria	0.221	0.144	0.221	2.323	2.166	20.188	1.182	1.206	0.203	0.105	0	0.337	0.743
(3)KORIAHu...	0.428	0.351	0.428	2.53	3.03	21.052	1.389	1.413	0.41	0.312	0.337	0	0.89
(2)CHANA	0.108	0.031	0.108	2.21	2.71	20.732	1.069	1.093	0	0.044	0.203	0.41	0.816
(12)JAPAN	20.75	20.647	20.676	20.676	19.4	0	20.849	20.873	20.732	20.686	20.188	21.052	21.458
(11)Cote	2.728	2.625	2.654	2.654	0	19.4	2.827	2.851	2.71	2.664	2.166	3.03	3.436
(10)GHAINA)	1.087	0.984	1.013	2.327	2.827	20.849	0	0.17	1.069	1.023	1.182	1.389	1.795
(1)USA	0.067	-0.015	0.067	2.164	2.664	20.686	1.023	1.047	0.044	0	0.105	0.317	0.718
(0)GERMANY	2.228	2.125	2.154	0	2.654	20.676	2.327	2.351	2.21	2.164	2.323	2.53	2.936

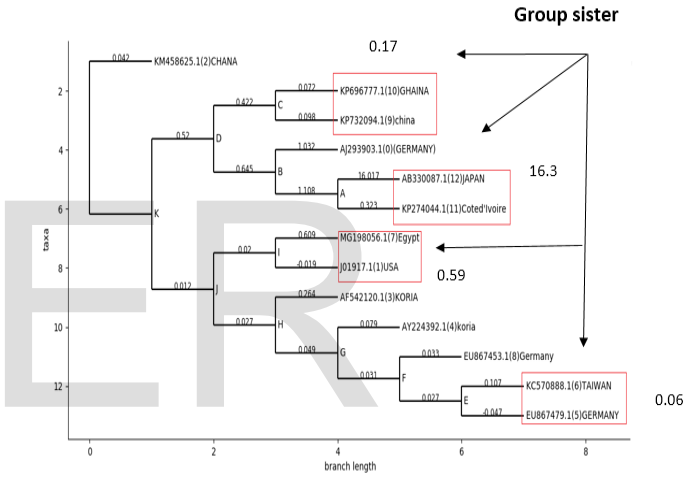

Fig. 8 NJ for construct tree after LA (differences) tree

TABLE 7
FINAL RESULTS OF DIFFERENCES TREE (LOCAL) FOR HADV GENE RELATED TO FIG. 8

Nodes	AJ29390	AB33008	KP27404	KP6967	KP7320	AF5421	AY2243	EU8674	KC5708	EU8674	MG1980	J01917.:	KM45862
	(0)	(12)	(11)	(10)	(9)	(3)	(4)	(8)	(6)	(5)	(7)	(1)	(2)
(0)(GERMANY)	0	18.157	2.463	2.171	2.197	2.5	2.364	2.349	2.45	2.296	2.838	2.21	2.239
(12)JAPAN	18.157	0	16.34	18.264	18.29	18.593	18.457	18.442	18.543	18.389	18.931	18.303	18.332
(10)GHAINA	2.171	18.264	2.57	0	0.17	1.317	1.181	1.166	1.267	1.113	1.655	1.027	1.056
(9)china	2.197	18.29	2.596	0.17	0	1.343	1.207	1.192	1.293	1.139	1.681	1.053	1.082
(1)USA	2.21	18.303	2.609	1.027	1.053	0.292	0.156	0.141	0.242	0.088	0.59	0	0.055
(2)CHANA	2.239	18.332	2.638	1.056	1.082	0.345	0.209	0.194	0.361	0.374	0.683	0.055	0
(5)GERMANY	2.296	18.389	2.695	1.113	1.139	0.324	0.09	0.013	0.06	0	0.716	0.088	0.374
(8)Germany	2.349	18.442	2.748	1.166	1.192	0.377	0.143	0	0.167	0.013	0.769	0.141	0.194
(4)koria	2.364	18.457	2.763	1.181	1.207	0.392	0	0.143	0.244	0.09	0.784	0.156	0.209
(6)TAIWAN	2.45	18.543	2.849	1.267	1.293	0.478	0.244	0.167	0	0.06	0.87	0.242	0.361
(11)CotedIvoire	2.463	16.34	0	2.57	2.596	2.899	2.763	2.748	2.849	2.695	3.237	2.609	2.638
(3)KORIA	2.5	18.593	2.899	1.317	1.343	0	0.392	0.377	0.478	0.324	0.92	0.292	0.345
(7)Egypt	2.838	18.931	3.237	1.655	1.681	0.92	0.784	0.769	0.87	0.716	0	0.59	0.683

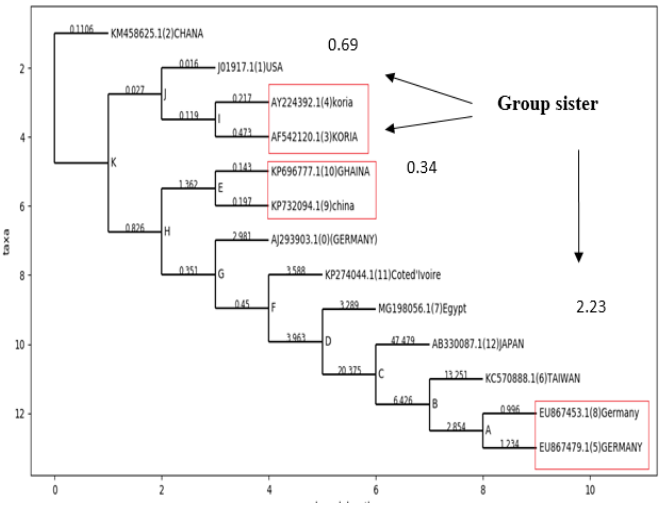

Fig. 9 NJ for construct tree after GA (differences) (tree)

TABLE 8
THE FINAL RESULTS OF GA SIMILARITY TREE FOR HADV GENE RELATED TO THE PREVIOUS FIG.

Nodes	KP696777.1	KP732094.	AJ293903	KP274044	MG19805	AB33008	KC570888.1(	EU867453.1	EU867479	J01917.1(1)L	AY224392	AF54212(	KM458625.1(2)CHA
	(10)	(9)	(0)	(11)	(7)	(12)	(6)	(8)	(5)	(1)	(4)	(3)	(2)
(11)Cot...	5.894	5.948	7.019	0	10.84	75.405	47.603	38.202	38.44	5.258	5.578	5.834	5.3256
(7)Egypt	9.558	9.612	10.683	10.84	0	71.143	43.341	33.94	34.178	8.922	9.242	9.498	9.2886
(8)Germ...	36.92	36.974	38.045	38.202	33.94	57.755	17.101	0	2.23	36.284	36.604	36.86	36.3516
(5)GER...	37.158	37.212	38.283	38.44	34.178	57.993	17.339	2.23	0	36.522	36.842	37.098	36.5896
(6)TAIW...	46.321	46.375	47.446	47.603	43.341	67.156	0	17.101	17.339	45.685	46.005	46.261	45.7526
(1)USA	2.374	2.428	4.201	5.258	8.922	73.487	45.685	36.284	36.522	0	0.352	0.608	0.1536
(2)CHANA	2.4416	2.4956	4.2686	5.3256	9.2886	73.5546	45.7526	36.3516	36.5896	0.1536	0.4736	0.7296	0
(4)koria	2.694	2.748	4.521	5.578	9.242	73.807	46.005	36.604	36.842	0.352	0	0.69	0.4736
(3)KORIA	2.95	3.004	4.777	5.834	9.498	74.063	46.261	36.86	37.098	0.608	0.69	0	0.7296
(10)GHA...	0	0.34	4.837	5.894	9.558	74.123	46.321	36.92	37.158	2.374	2.694	2.95	2.4416
(9)china	0.34	0	4.891	5.948	9.612	74.177	46.375	36.974	37.212	2.428	2.748	3.004	2.4956
(0)(GER...	4.837	4.891	0	7.019	10.683	75.248	47.446	38.045	38.283	4.201	4.521	4.777	4.2686
(12)JAPAN	74.123	74.177	75.248	75.405	71.143	0	67.156	57.755	57.993	73.487	73.807	74.063	73.5546

TABLE 9
THE SISTER GROUP OF DIFFERENCES IN THREE TREES OF ALIGNMENT METHODS

No.	Sister group		Alignment type	Distance pairwise
1	S3: Korea	S7: Egypt	MSA	0.89
2	S9: china	S10: China	MSA	0.17
3	S11: coted'Ivoire	S12: japan	MSA	19.4
4	S9: China	S10: China	Local	0.17
5	S11: japan	S12: coted'Ivoire	Local	16.34
6	S7: Egypt	S1: USA	Local	0.59
7	S5: Germany	S6: Taiwan	Local	0.06
8	S3: Korea	S4: Korea	Global	0.69
9	S9: china	S10: china	Global	0.34
10	S5: Germany	S8: Germany	Global	2.23

4.4 System Comparison

In this work, the three types of alignments were implemented to compare the distances between each pair of sequence after implementing the NJ method. Refering to Table 9, the MSA and LA alignments present very close results of distances. While for GA, the results of distances have large differences in comparison to other alignment.

According to the above-mentioned table, multiple pair of sequences have been aligned to find the differnce of sequences for the three types of alignments. Where, the shortest distance given by MSA equals to 0.049; which defines the distance between seq6, and seq5. While the longest distance given by MSA equals to 20.75; which defines the distance between seq6, and seq12. In GA, the shortest distance is 0.153; which represents the distance between seq2, and seq1. While the longest distance is 67.156 that defines the distance between seq6, and seq12. In LA, the shrotest distance is 0.055; that defines the distance between seq2, and seq1, while the long distance (18.543) defines the distance between seq6 and seq12. It is clear that the MSA and LA provide a convenient alignment; because the distances between each pair of distances are small. While the GA provids unconveniant alignment as the distances between pair of sequences are large; due to the existing gaps that are not observed in local and MSA.

5 CONCLUSION

According to the experimental results, it is shown that the GA is not suitable for sequences of different lengths. Therefore, the short-length sequence is extended by adding some gaps to equialize the length of sequences. Although, the GA has proved to be is faster than LA in our experiments, LA is more suitable for sequences of different lengths, as it choose the most similar area between pairs within a sequence and eliminating the rest nucleotides of different lengths. However, its impact is ignoring the length of sequence with taking into account only the sub regions within a sequence that have a high similarity of bases.

The MSA is faster than both LA and GA since the alignment process is done only once for the whole sequences based on the best alike sequences. Also, it has not been involved when computing the execution time as the alignment process is already done using the Clustal Omega website, so that only the computation of distance matrix has been done in this work. The output result from the Clustal Omega website has been edited by removing the beginning and end-gaps to reduce the sequence-length; while maintaining the nucleotide positions with taking into consideration the shortest sequence for each pair to eliminate the dissimilarity. Finally in this work, the user has the capability to have the suitable alignment for creating the phylogenetic tree.

REFERENCES

[1] M. Khan and N. Rahman, "Bioinformatics : Analyzing DNA Sequence using BLAST", Department of Computer Science and Engineering, 2007 .

[2] R. K. Bíró, "Constructing Phylogenetic Trees," PhD Thesis, 2015.

[3] Z. M. Alkhafaji, and A. Abd-alhafed,"Bioinformatics", Baghdad Univ., 2012.

[4] J. Harrison, C., and J. A. Langdale.. "A Step by Step Guide to Phylogeny Reconstruction", Plant Journal, vol. 45, no. 4, pp. 561–72, 2006.

[5] T. Jakub., "Fast Algorithms for Large-Scale Phylogenetic Reconstruction,"http://uwspace.uwaterloo.ca/handle/10012/8011, 2013.

[6] G. Olson, H.Matsuda and R. Overbeek, "FastDNAml: A tool for construction of phylogenetic trees of dna sequences using maximum likelihood", Compu-tApplBiosci, vol 10, no. 1, pp. 41–48, 1994.

[7] D. Trystram and J. Zola, "Grid Computing", Bioinformatis and Computational Biology, John Wiley & Sons, Inc., Ho-boken, New Jersey, 2008

[8] CA. Stewart, D. Hart, DK. Berry, GJ. Olsen, and W.Fischer, "Parallel implementation and perfor-mance of fast DNAml a program for maximum likelihood phylogenetic inference", Procs of SC, vol. 32, no. 1, 2001.

[9] S. Guindon and Gascuel. O., "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum like-lihood", Syst. Biol, vol. 52, no. 5, pp. 696–704, 2003.

[10] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, JF. Dufayard, S. Guindon, V. Lefort, M. Lescot, JM Claver and O. Gascuel, 2007, "Phylogeny.fr: robust phylogenetic analysis for the non-specialist", Nucleic Acids Research, vol 36, no. 1, pp. 465–469, 2007.

[11] RC.Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity", BMC Bio-informatics, vol. 5, no. 113, 2004.

[12] J. Castresana, "Selection of conserved blocks for mul-tiple alignments for their use in phylogenetic alignments", Mol. Biol, vol. 17, no. 4, pp. 540–552.