Constructing of Phylogenetic Tree for Cortex based on DNA sequences

Dr.A.Rama Mohan Reddy[1], Chukka Santhaiah[2]

[1]*Professor, C.S.E Department, S.V.University , Tirupati , A.P, INDIA.*
[2]*Research Scholar, C.S.E Department, S.V.University , Tirupati , A.P, INDIA.*

Abstract: *The Phylogenetic expansion of genus is frequently described as the result of arbitrary transformations and predictable choice, which provides the awareness of phylogenetic trees. The large number of elevated superiority sequence data available has wrecked the reconstruction of phylogenetic trees as of sequence data achievable. In this work Cortex gene sequences are taken from NCBI as it's with accession number. The ClustalW loom is an important and trendy substitute for create phylogenetic trees based on sequence data, which can be constructed. The aim of work is to find the closeness of sequences and accelerate tree rebuilding by construction use of preceding information. Here put forward a new heuristic for incorporating a phylogenetic tree for DNA datasets. In this work we implement the proposed ClustalW for multiple alignments. The experiments performed show, which incorporating prior knowledge leads to a significant speed up, if large amounts are included.*
***Keyword*s:** *- ClustalW, Cortex, DNA, Phylogenetic tree.*

I. Introduction

The Cortex is the brain outer layer of neural tissue in humans and other mammals. It is divided into two cortices along the sagittal plane. The left and right cerebral hemispheres divided by the medial longitudinal fissure. The cerebral cortex plays a key role in memory, attention, perceptual awareness, thought, language, and consciousness.

The phylogentically most recent part of the cerebral cortex, the neo cortex is differentiated into six layers, the most ancient part of the cerebral cortex, the hippocampus, has at most three cellular layers.

Phylogenetics is the cram of development affairs linking organisms. Expertise has permit astonishing evolution in phylogenetics. One portion of this in the biology itself. The innovation of DNA and the capability for biologists to sequence DNA has, to say the slightest revolutionized the field. Computers have assisted extremely as well. To the side from their relieve in sequencing assignments, computers smooth the progress of a broad array of phylogenetic problem-solving activities.

The contribution of computers (or perhaps more appropriately, computer science) is text processing. Since DNA sequences come as a sequence of characters (from the alphabet A, C, T, G), there are several computer science algorithms that can be used as processing tools. The field of phylogenetics has applications to molecular biology, genetics, evolution, epidemiology, ecology, conservation biology, and forensics to name a few. Phylogenies are the chronological and evolutionary relationships among organisms. Researchers can employ this data to better understand how viruses spread or to study common biological processes between different species of life [1].

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences such as protein, DNA, or RNA. Typically it is implied that the set of equences share an evolutionary relationship, which means they are all descendents from a common ancestor. These regions may correspond to functional, structural, or evolutionary relationships between the sequences. Alignments can reflect a degree of evolutionary change between sequences that are descendants from a common ancestor. There is a relationship between phylogenies and sequence alignments [2].

There are various techniques used to create phylogenetic trees and most of them rely on aligned genetic sequences to perform this task. Probably the most popular genetic sequence alignment algorithm is ClustalW [2]. Although successful in its domain, ClustalW is very sensitive to highly divergent sequences. Therefore the purpose of this project is to modify the ClustalW sequence alignment algorithm so that it can be used to construct a more accurate tree when highly divergent sequences are present. Sequence alignment is a way of arranging sequences of DNA, RNA, or proteins in order to distinguish regions of similarity. ClustalW is a popular program used for multiple sequence alignment and for preparing phylogenetic trees. Its portability amongst various computing platforms is the main reason for its widespread use. Due to its popularity and the availability of source code, ClustalW was used for this project. The progressive alignment algorithm used by ClustalW to perform a multiple sequence alignment.

II. Methodology

The methodology purpose is to get the nucleotide sequences by using DNA sequences that are related to same sequences. In this process first select the four DNA sequences entire this process. Here the Cortex is taken for this processing. To get the nucleotide sequence from amino acid sequence by performing different operations on selected protein. The methodology follows different operations.

a. Search the protein form database that is related to Cortex DNA sequences.
b. Give that protein in the NCBI database for find the sequence in the form of FASTA [3].
c. Then select the FASTA format of DNA from NCBI [3].
d. The output is gives graphic summary and description of gene. And it displays related genes.
e. Based on gene accession number to find the nucleotide sequences.
f. Then select the all sequences, copy the FASTA format of the sequences and paste in ms-word format for ClustalW operation [4].
g. Finally perform the ClustalW operation and it provides rooted and Unrooted phylogenetic tree [4].

In the above operations we get the nucleotide sequence, and also to create the phyologenetic tree based on nucleotide sequences. To get the phylogenetic tree with the help of DNA sequences. The output file contain rooted and unrooted phylogenetic tree. In phylogenetic tree the similar sequences are form as taxa. The remaining sequences are grouped as a whole tree of similar tree. The execution of the nucleotide process may take few seconds/minutes based on sequence length, processor and internet speed.

III. Results

ClustalW Input

General Setting Parameters:
Output Format: CLUSTAL
Pairwise Alignment: ⦿ FAST/APPROXIMATE   ◯ SLOW/ACCURATE

Enter your sequences (with labels) below (copy & paste): ⦿ PROTEIN ◯ DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

```
AAAACAAAGTCAAGAAAACAACTGGGTCGGATCATGCTAAAAGGAGATAATATTACTCTG
CTCCAAAGTG
TTTCCAACTAGCAGTGAATGGTGAAGTCTGTAGGATGTTGAGAAGACCCCTTGAGCGTGT
TTAAAGATGT
CTGCTGCAACCTGCATTTACTCAACT
```

Or give the file name containing your query
[Browse...] No file selected.

[Execute Multiple Alignment] [Reset]

>gi|40103413|gb|BG086892.2|BG086892 H3131H01-5 NIA Mouse 15K cDNA Clone Set Mus musculus cDNA clone H3131H01 5', mRNA sequence
GACCGTCCGAGCGGCGGGCGGCGGACGACGTTCGTCATTTGGTGCGGGAGGGAGCGCGAGAGCG
GCGGTCGAGCCTCCGGTGTCCCAAGCTAGAGGTAAGCATGGCAGAACAGGAACCCACTGCTGAGC
AGCTTGCGCAGATAGCTGCAGAGAATGAGGAAGATGAGCACTCTGTGAACCACAAGCCTCCAGCC
CAGAAGAGCATCCAGGAGATCCAGGAACTGGACAAGGACGATGAAAGCCTCCGAAAGTACTAGG
AGGCCCTGCTGGGCCGAGTAGCTGTCTCTGCAGACCCCAACGTCCCCAATGTCATTGTGACTCGAC
TGACCTTGGTATGCAGCACTGCCCCAGGCCCTCTGGAACTGGACCTGACAGGTGATCTGGAGAGCT
TCAAGAAACAGTCGTTTGTCTTGAAGGAAGGTGTGGAGTACCGGATAAAAATCTCTTTCCGGGTGA
ACAGAGAGATCGTGTCCGGCATGAAGTACATCCAACACACATACAGGAAAGGGGTCAAAATTGAC
AAGACTGACTACATGGTCGGGAGCTATGGGCCGAGGGCCGAG

Fig 1: The FASTA format of input Protein

ClustalW Result

```
WARNING: possibly wrong combination

---------------------------------
Selected type :    PROTEIN
Query sequence:    DNA
---------------------------------
```

[clustalw.aln][clustalw.dnd][readme]

Select tree menu   ▼   Exec

```
CLUSTAL 2.1 Multiple Sequence Alignments

Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: gi|40103413|gb|BG086892.2|BG086892    563 aa
Sequence 2: gi|31570427|gb|AW555640.2|AW555640    600 aa
Sequence 3: gi|40110469|gb|BG075861.2|BG075861    550 aa
Sequence 4: gi|40073572|gb|BG085367.2|BG085367    376 aa
Start of Pairwise alignments
Aligning...


(Partial alignment)
Sequences (1:2) Aligned. Score: 21.492
(Partial alignment)
Sequences (1:3) Aligned. Score: 17.2727
(Partial alignment)
Sequences (1:4) Aligned. Score: 26.5957
(Partial alignment)
```

```
CLUSTAL 2.1 multiple sequence alignment


gi|40110469|gb|BG075861.2|BG07      -GAGAGGAAGTTGGATTTAATGACATGGAAGGAAGAGGGCCCAGCAACAA
gi|40073572|gb|BG085367.2|BG08      --------------------GTGACTT------------------------
gi|40103413|gb|BG086892.2|BG08      --GACCGTCCGAGCGGCGGGCGGCGGACGACGTTCGTCATTTGGTGCGGG
gi|31570427|gb|AW555640.2|AW55      CATTGTGTTGCTTTATTCAGTTTGCTCCAAGGGCAGAATCAATAGTAGGG
                                                                                    .

gi|40110469|gb|BG075861.2|BG07      GCCAGGCTTTTGCCTCCCCAGCCCTCCCTGGGCTTCCCAAGCTGCCTCAA
gi|40073572|gb|BG085367.2|BG08      ----------------------CCACCATGGCGTACCGCGGCCAGGGCCA
gi|40103413|gb|BG086892.2|BG08      AGGGAGCGCGAGAGCGGCGGTCGAGCCTCCGGTGTCCCAAGCTAGAGGTA
gi|31570427|gb|AW555640.2|AW55      ATCACAAACATTATTTTGAGTGGCTCACAAGCCTGATCAGTAAAACCTAA
                                              . *.    *      .  ..  . .         *

gi|40110469|gb|BG075861.2|BG07      TGGGATGTTATTGATTCGGGCACGTCACAATGTGTGGACCACCCTGGAAT
gi|40073572|gb|BG085367.2|BG08      AAAAGGTGCAGAAGGTGATGGTGCAGCCCAT----------CAACCTTATCT
gi|40103413|gb|BG086892.2|BG08      AGCATGGCAGAACAGGAACCCACTGCTGAG------CAGCTTGCGCAGAT
gi|31570427|gb|AW555640.2|AW55      AAAACAGGAAGTGTTACACAAGTTGCAAATAAAG-CCTCCTGGAGTTCAT
                                    :. .  * :.:   .  .   * *          *:  .    .*

gi|40110469|gb|BG075861.2|BG07      AGGGGCATCAAGATAGGAGGTAC--CCAGTTCAGAAGGTGAGTCCCAAAT
gi|40073572|gb|BG085367.2|BG08      TCAGATACTTGCAAAATAGATCT--CGAATTCAGGTG--------------
gi|40103413|gb|BG086892.2|BG08      AGCTGCAGAGAATGAGGAAGATG--AGCACTCTGTGAACCACAAGCCTCC
gi|31570427|gb|AW555640.2|AW55      TGGCATGATAGCAGTGTAAGGACGGTAAATGCAACAGGTGTTAACATCTC
                                    :     . .   . : :. *..      ..  *:.  .

gi|40110469|gb|BG075861.2|BG07      ACCCCCTTTCGTCTTTCTGCCCCACCACAGAGGGGTAGTTCCGCCCCTCT
gi|40073572|gb|BG085367.2|BG08      ---------TGGCTGTATGAACAAGTGAATATG-----------------
gi|40103413|gb|BG086892.2|BG08      AGCCCAGAAGAGCATCCAGGAGATCCAGGAACTGGACAAGGACGATGAAA
gi|31570427|gb|AW555640.2|AW55      ACACAGGACTTGGAACCTGCTCACTTCTGGACAATTCTGGGACCCTTGTT
                                            :  .:*    .      . *

gi|40110469|gb|BG075861.2|BG07      GCCTCTATAGCTGCCCCCACAGCCAGGCTCACAGGTGTGTCAAAGGCCCT
gi|40073572|gb|BG085367.2|BG08      ------------------------CGGATAGAAGGTTGTATTATTGGCTTT
gi|40103413|gb|BG086892.2|BG08      GCCTCCGAAAGTACTAGGAGGCCCTGCTGGGCCGAGTAGCTGTCTCTGCA
gi|31570427|gb|AW555640.2|AW55      GCAAACCCTGCAGATACCAATGCTGATTTTCTTTGATATGAAAATATGACA
                                         .                      :  .:          :

gi|40110469|gb|BG075861.2|BG07      GGCCCTAGCAGTTAGACTCCAGTTCTGTGAG--CCTGAGAAGGGGATGTCAT
gi|40073572|gb|BG085367.2|BG08      G---ATGAGTACATGAACCTCGTATTAGATG-ATGCAGAAGAGATTCATT
gi|40103413|gb|BG086892.2|BG08      GACCCCAACGTCCCCAATGTCATTGTGACTCGACTGACCTTGGTATGCAG
gi|31570427|gb|AW555640.2|AW55      TTACACAACTCTTCGTTGGGCCTTACTCTTGCTGCACTTGAGTGTTCCTA
```
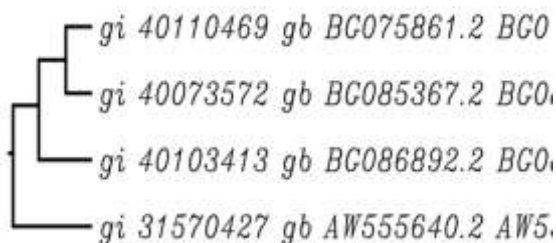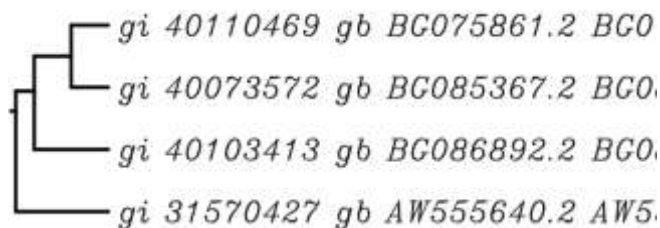
Fig 2: Without Branch length(UPGMA)
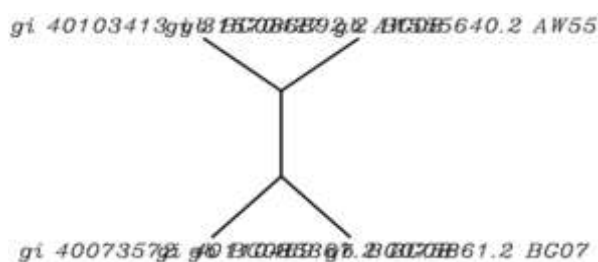

Fig 3: With branch length (UPGMA)

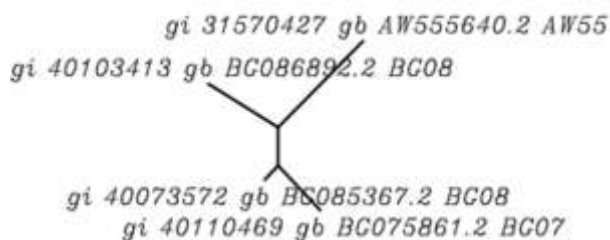Fig 4: Unrooted tree without branch length (NJ)

Fig 5: Unrooted tree with branch length(NJ)

IV. CONCLUSION

It has been shown to a great extent curiosity in phylogenetic systematic recently. Since the reason that it is a technique for biologists to rebuild the pattern of events that have led to the distribution and diversity of life. Due to the "Tree of Life" proposals, and studies the researchers have tried to find some new methods to combine large number of trees to construct phylogenies on hundreds, or even thousands of species. For understanding the language of genes and proteins we have to find a suitable model of how they have evolved in the course of evolution. Because of this we need to develop tree building methods which discover the process of evolution. These kinds of methods have gained importance with the advent of molecular biology.

References

[1].   Chukka Santhaiah, Dr.A.Ramamohanreddy ,IJERD *Volume 10, Issue 6 (June 2014), PP.26-32*
[2].   Thompson, Julie D., Desmond G. Higgins, and Toby J. Gibson. "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research 22 (1994): 4673-4680.
http://www.ncbi.nlm.nih.gov/
http://www.genome.jp/clustalw/