Relative Greenhouse Gases Emissions of MSA Algorithms: A comparison of T-Coffee and MAFFT

Dr Neeta Maitre
Assistant Professor
Department of Computer Engineering,
Cummins College of Engineering for Women , Pune, India

Abstract : Multiple sequence alignment is an important step in gene prediction. The sequences involved in the alignment process are assumed to be derived from a single ancestral sequence i.e. homologous. These sequences are subjected to sequence algorithms for the homology search. The results obtained by sequence alignment can be used for locating genes by identifying exon positions or can be utilized to interpret evolutionary origin. The time complexity of the algorithms governs the runtime, which in turn drives the greenhouse gas (GHG) emissions. GHG emissions are critical from a global sustainability point of view. In this work, a method to compare the relative standing of algorithms from GHG emissions point of view is developed. T-Coffee and MAFFT algorithms are used as examples. The peculiar feature of these algorithms is that both are progressive in nature and can cover all sized sequences used for alignment. MAFFT is observed to have relatively lower GHG emissions as compared to T-Coffee.

***IndexTerms* - Sustainability, Multiple sequence alignment algorithms, MAFFT, T-Coffee, GHG (Greenhouse Gases)**

I. INTRODUCTION

The field of bioinformatics is very vibrant. The major work in this field is related to sequences and their alignment. Multiple sequence alignments (MSA) are an essential and widely used computational procedure for biological sequence analysis in molecular biology, computational biology, and bioinformatics[6]. Multiple Sequence Alignment (MSA) methods refer to a series of algorithmic solutions for the alignment of evolutionarily related sequences while taking into account evolutionary events such as mutations, insertions, deletions, and re-arrangements under certain conditions[1]. The method of aligning three or more biological sequences having similar length is called “Multiple Sequence alignment”. The sequence alignment methods are divided into majorly two types: Pairwise alignment and progressive alignment. Both these techniques can be applied to the biological sequences namely DNA, RNA or protein, pairwise alignment has a limitation of aligning only two sequences at a time whereas progressive can align multiple sequences at a time. The pairwise alignment can be categorized under local as well as global alignment while MSA is acting under global category. In pairwise alignment of sequences using the global method uses the Neeleman-Wunch algorithm and the local alignment technique is governed by Smith -Waterman algorithm. The MSA usually employs progressive alignment method where the iterative implementation of pairwise algorithms is done.

MAFFT and T-Coffee are progressive in nature. The progressive algorithms undergo following steps:

- Calculation of number of possible pairs
- Pairwise alignment and distance calculation
- Guide tree generation and again the alignment

The basis for both , T- Coffee algorithm and MAFFT are based on the usage of progressive algorithm of alignment. The progressive alignment as shown in figure 1.

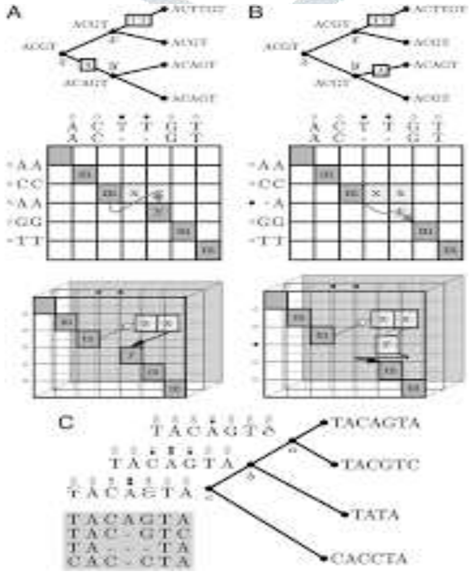


Figure 1: Process of Progressive alignment (Source:pnas.org)

Sustainability perspective of algorithms has now emerged as the concept of green algorithms. Green Algorithms (www.green-algorithms.org), which enables a user to estimate and report the carbon footprint of their computation. The Green Algorithms tool easily integrates with computational processes as it requires minimal information and does not interfere with existing code, while also accounting for a broad range of CPUs, GPUs, cloud computing, local servers and desktop computers[2].

This paper focuses on relevant parameters which can be considered under green algorithms and tries to compute them for selected progressive multiple sequence algorithms in bioinformatics namely T- Coffee and MAFFT.

II. EXAMPLE ALGORITHMS

There are various tools available in bioinformatics which can align the protein or nucleic acid sequences. These sequences follow algorithms either pairwise or iterative ways of alignment. The use of MSA can be for identification of conserved regions and for phylogenetic analysis. The time complexity of these algorithms is based on the amount of time required for the run of the algorithm as a function of the length of the genetic sequence. The algorithms discussed here are MAFFT and T-coffee. The time complexity of these algorithms is tabulated as in Table1.

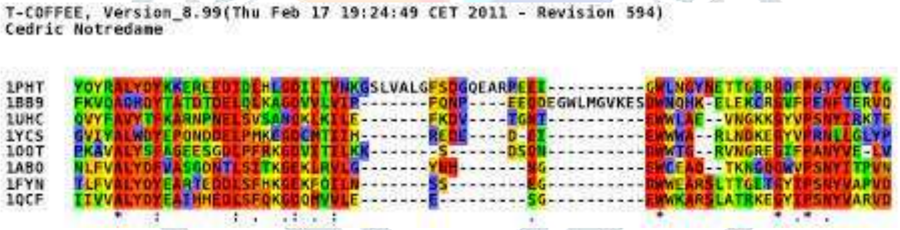
Table 1 Time complexity table

Multiple Alignment using Fast Fourier Transform(MAFFT)	The time complexity of the progressive method implemented in MAFFT is basically $O(N^2L) + O(NL^2)$, where L is the sequence length and N is the number of sequences.[3]
T-Coffee	The complexity of the whole procedure is $O(N^2L^2) + O(N^3L) + O(N^3) + O(NL^2)$ where N is the number of sequences and L is the average sequence length.[5]

2.1 T-Coffee

T-Coffee stands for Tree based consistency objective function for alignment evaluation.

It is majorly suitable for small alignments. It uses local as well as global alignment methodology to achieve multiple alignment. Advanced features of T-Coffee help to evaluate the quality of alignments and identification of motifs to some extent. This tool can take input sequences from a file or can be entered directly into the interface. After alignment, the out or the result summary is generated. The result summary consists of input sequences, tool output which contains a log file created, alignment in PHYLIP format, alignment in HTML format, alignment in Clustal format, alignment in MSF format, guide tree and building information for phylogram.



```
T-COFFEE, Version_8.99(Thu Feb 17 19:24:49 CET 2011 - Revision 594)
Cedric Notredame

1PHT    YOYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEI-----------GWLNGYNETTGERGDFPGTYVEYIG
1BB9    FKVQAQHDYTATDTDELQLKAGDVVLVIP-------FQNP----EEQDEGWLMGVKESDWNQHK-ELEKCRGVFPENFTERVQ
1UHC    QVYFAVYTFKARNPNELSVSANQKLKILE-------FKDV----TGNT-----------GWWLAE--VNGKKGYVPSNYIRKTE
1YCS    GVIYALWDYEPQNDDELPMKEGDCMTIIH-------REDE----D-EI-----------EWWWA--RLNDKEGYVPRNLLGLYP
1OOT    PKAVALYSFAGEESGDLPFRKGDVITILKK------S------DSQN-----------DWWTG--RVNGREGIFPANYVE-LV
1ABO    NLFVALYDFVASGDNTLSITKGEKLRVLG-------YNH--------NG-----------EWCEAQ--TKNGQGWVPSNYITPVN
1FYN    TLFVALYDYEARTEDDLSFHKGEKFQILN-------SS--------EG-----------DWWEARSLTTGETGYIPSNYVAPVD
1QCF    TIVVALYDYEAIHHEDLSFQKGDQMVVLE-------E---------SG-----------EWWKARSLATRKEGYIPSNYVARVD
            *  :          : .  .:. :              .              *          * .*. .
```

Figure2: Reformatted output of T-Coffee (Source-tcoffee.org)

2.2 MAFFT

MAFFT stands for Multiple Alignment using Fast Fourier Transform. The initial version of this algorithm was introduced in 2002. This is mainly used for medium to large sequence alignments. It uses fast fourier transform and hence reduces the computational time as compared to other tools. It has two methods of execution. One is progressive method and the other is the iterative refinement method. The former delivers comparable accuracy with reduced computation time and the latter gives almost 100 times faster than T-Coffee without loss in accuracy. Like T-Coffee, it also takes input sequences directly or from the file. The output shows N terminal alignment MAFFT is similar to T-Coffee but it is in the middle. It's C terminal is entirely different from that of T-Coffee.

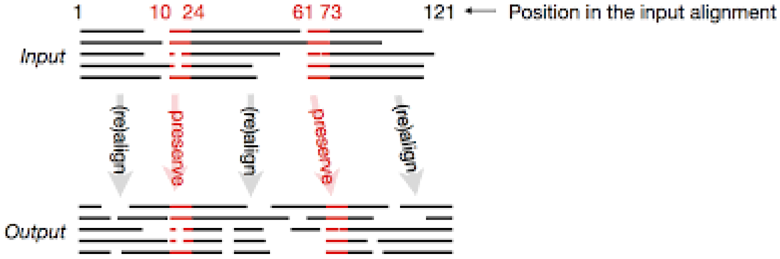


Figure3: Output of MAFFT(Source-mafft.cbrc.jp)

III. COMPARISON OF ALGORITHMS

Given a set of biological sequence (RNA, Proteins, DNA) the purpose of a Multiple Sequence Alignment method is to align the sequences in a way that will either reflect their evolutionary,functional or structural relationship [4].A wide range of computational algorithms have been applied to the MSA problem, including slow, yet accurate, methods like dynamic programming and faster but less accurate heuristic or probabilistic methods.[5]

The diagrammatic representation in T-Coffee and MAFFT is shown in figure 4 and 5. Both the methods show effective use of progressive alignment.

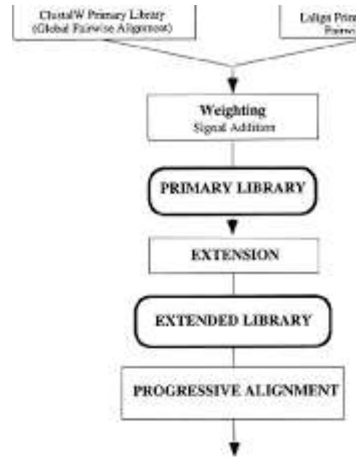

ClustalW Primary Library (Global Pairwise Alignment)
Lalign Primary Library (Local Pairwise Alignment)
Weighting
Signal Addition
PRIMARY LIBRARY
EXTENSION
EXTENDED LIBRARY
PROGRESSIVE ALIGNMENT

Figure 4: T-Coffee algorithm(Source-researchgate.net)

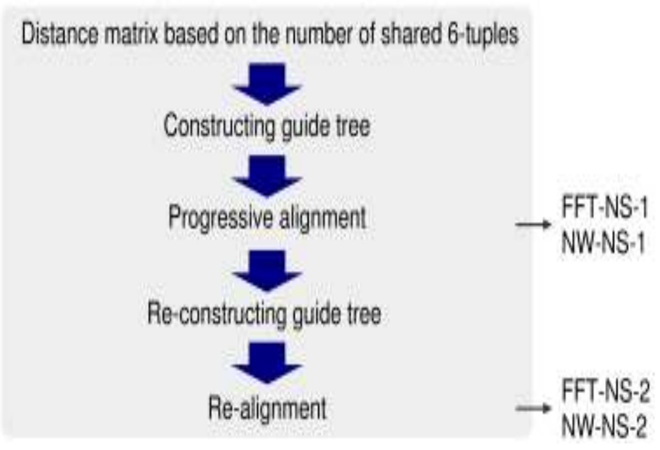


Figure 5: MAFFT algorithm(Source-mafft.cbrc.jp)

The basic governing parameters in both the cases are:

- Length of sequence (L)
- Number of sequences (N)

These parameters are subject to change based upon the input sequence/s and the maximum values are fixed for both the algorithms. For both these algorithms, the upper limit of number of sequences is 500 and the maximum size of sequence is restricted to 1MB size.

To compare the performance of these algorithms, the variable sequences and sequence lengths are subjected to these algorithms. The estimated run time computed as shown in Table 2.

Table 2 Estimated runtime for variable length (L) and number of sequences (N)

Algorithm	N	L	Estimated Runtime
MAFFT	1.00E+01	1.00E+05	1.00E+11
T-Coffee	1.00E+01	1.00E+05	1.10E+12
MAFFT	1.00E+05	1.00E+05	2.00E+15
T-Coffee	1.00E+05	1.00E+05	2.00E+20
MAFFT	1.00E+05	1.00E+02	1.00E+12
T-Coffee	1.00E+05	1.00E+02	1.01E+17

The observations are as follows:

- At constant L, MAFFT's estimated runtime is less than 10% of T-COFFEE.
- At constant N, MAFFT's estimated runtime is less than 5th order of magnitude of T-Coffee.

This computation is important to get its impact on Greenhouse gas (GHG) emission. The GHG emission is calculated using the formula [1]

$$\boldsymbol{E = t \times (n_c \times P_c \times u_c + n_m \times P_m) \times PUE \times 0.001}$$

$$E = t \times (n_c \times P_c \times u_c + n_m \times P_m) \times PUE \times 0.001$$

Here,

t = running time (hours)
n_c= number of cores
P_c= power draw of computing core

u_c= core usage factor
n_m= size of memory available (GB)
P_m= power draw from memory (Watt)
PUE = Power Usage Effectiveness

The elaboration of each term involved is as follows[1]:

- Running time: It depends upon the time complexity of the algorithm
- Power draw of computing core: This metric is provided by the manufacturer of the GPU or CPU. It is the thermal design power (TDP) and is measured in Watts. The normalization of TDP values is done in case of multiple cores.
- Power draw from memory: Background consumption majorly contributed to the power draw from memory. Total memory allocated mainly affects this factor more than the size of the database.
- Power draw from storage: Workload significantly hampers power draw from storage. The storage addressed here is SSD or HDD. Independent of the task in hand, it can be looked upon as a permanent record of data.
- Power Usage Effectiveness: It is the ratio of total power drawn by the facility and power used by IT equipment. It is also called the efficiency coefficient of a data center.

$$PUE = P_{tota\ l} / P_{compute}$$

A data centre PUE of 1.0 represents an ideal situation where all power supplied to the building is utilised by computing equipment.[1]

Under the standard conditions and the same working environment for both the algorithms, the energy consumption is directly dependent on "t"i.e running time of algorithms. According to EIA, the average carbon intensity of electricity production process is 417.3 kg/MWh. Hence, the GHG emissions would directly be influenced by the runtime. As a result, the relationships between the estimated GHG emissions for these two algorithms would be similar to the observed relationship between their estimated runtimes. It may be concluded that, for the given test cases from a sustainability standpoint, MAFFT performs better than T-COFFEE.

IV. Concluding remarks

Application of the factors corresponding to GHG is important in computational intensive tasks. Bioinformatics algorithms are dealing with a large number of sequences and also with the large sized sequences. Processing these sequences to get useful insights related to genetic predictions, mutations and other related tasks are computationally intensive. After analyzing the two MSA algorithms, namely T-coffee and MAFFT, used frequently in bioinformatics, the results show that the overall T-coffee algorithm takes more time than MAFFT. The conclusion for this tested scenario is that run time for an algorithm directly influences the energy required for the computation.

IV. Acknowledgement

I thank Dr Neeraj Hanumante (Post Doctoral Fellow at IIT-B) for providing a part of estimated runtime during this work.

References

[1]Maria Chatzou, Cedrik Magis et al, "Multiple Sequence Alignment Modeling: Methods and Applications", Briefings in Bioinformatics · November 2015
[2] Loïc Lannelongue, Jason Grealey, Michael Inouye, "Green Algorithms: Quantifying the carbon footprint of computation"
[3]Kazutaka Katoh and Hiroyuki Toh, "Recent developments in the MAFFT multiple sequence alignment program",BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 4. 286-298 doi:10.1093/bib/bbn013
[4] Jurate Daugelaite,Aisling O' Driscoll,and Roy D. Sleator, "Review Article An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics",Hindawi Publishing Corporation ISRN Biomathematics, Volume 2013
[5]https://en.wikipedia.org/wiki/Clustal
[6]https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/Clustal+Omega+Help+and+Documentation