Vol. 19 (Supplementry Issue 3A), September 2025, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2025.3s.5

Identification of Conserved Regions in HIV-1 DNA Mutational Patterns Using Multiple Sequence Alignment

Mr Rajesh Jangade[1*], Dr Yazdani Hasan[2], Dr Suryakant Yadav[3], Mr Satyakam Pugla[1], Dr Mukesh Chourasia[1], Dr Jitender Kumar[1], and Dr Anil Barnwal[1]

[1]Center of Computational Biology and Bioinformatics, Amity University, Noida, UP, 201303
[2]Engineering and Technology, Noida International University, Greater Noida, UP, 201310
[3]Department of Computer Science, IIMT Group Of Colleges,. Greater Noida, 201310
*Correspending author: rkjangade@amity.edu

Abstract

HIV (Human Immunodeficiency Virus) targets and weakens the immune system. If left untreated, it can progress to AIDS (acquired immunodeficiency syndrome), a condition marked by a high viral load and increased infectiousness. Without treatment, individuals with AIDS typically survive about three years. While there is no cure for HIV, it can be managed effectively with proper medical care. This study focuses on identifying conserved regions within the genetic and protein sequences of HIV-1 from large-scale data. By leveraging multiple sequence alignment and machine learning techniques, conserved regions in the mutational patterns of HIV-1 are identified. These conserved regions hold significant potential for the pharmaceutical industry, particularly in development of vaccine.

Keywords: HIV-I, Mutational Pattern, Conserved region, Big Data, Multiple Sequence Alignment, Phylogenetic Tree

Introduction

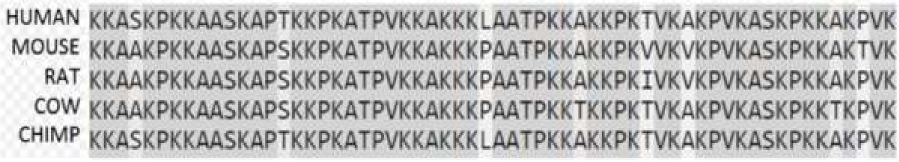
HIV infection in humans is believed to have originated from a type of chimpanzee in Central Africa. The chimpanzee version of the virus, known as simian immunodeficiency virus (SIV), likely crossed into humans during hunting activities when people met the infected blood of these animals [1][2][3]. Research suggests that HIV may have crossed from chimpanzees to humans as early as the late 1800s. Over time, the virus gradually spread across Africa and eventually reached other parts of the world. In the United States, evidence indicates that HIV has been present since at least the mid-to-late 1970s [1][2]. This study addresses two critical aspects relevant to the pharmaceutical field. First, it focuses on identifying conserved regions within HIV gene and protein sequences using multiple sequence alignment. Second, it examines the phylogenetic relationships of various HIV genes and proteins to provide deeper insights into their evolutionary patterns [4][5].

Understanding the conserved regions in various DNAs

Conserved regions are segments of RNA, DNA, or amino acid sequences that remain similar or identical across generations, either within the same species or among different species [6][7]. These sequences exhibit minimal or, in some cases, no changes in their composition over generations as the below (Figure 1) [8][9].

Challenges in India

In 1992, the Government of India established the National AIDS Control Organization (NACO) to lead the fight against AIDS. NACO's primary objectives include curbing the spread of HIV infection to mitigate the impact of AIDS, reducing morbidity, and lowering mortality rates associated with the disease [8][9]. The first phase focused on raising awareness about AIDS, establishing a surveillance system, and ensuring access to



```
HUMAN KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK
  RAT KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKAKPVK
  COW KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
CHIMP KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
```

Fig. 1: Example of Conserved region find in DNA of cross species

safe blood. In 1999, the second phase was introduced with two key objectives [9]. The primary objectives of the second phase were to reduce the spread of HIV infection and enhance India's capacity to respond to the HIV/AIDS crisis. Despite these efforts, 2005 marked the peak of the AIDS pandemic in India. In 2007, the third phase was launched with the goal of halting and reversing the epidemic [10][11][12]. Data from NACO indicates a significant decline in the number of new HIV cases in India since the peak of the AIDS pandemic in 2005. However, the rate of decrease has slowed in recent years [11][12][13]. The Indian HIV programme has expanded, evolved and implemented various new initiatives overs the years. Still there are some challenges faced by the NACO concluded as following[13][14]:

(i). Stigma and Discrimination
(ii) Access to testing services by people
(iii) Funding

Collection of data

Data for this study were collected from the following sources:

(i) DNA sequences of HIV were downloaded from the National Center for Biotechnology Information (NCBI) website.

(ii) Protein sequences of various HIV-1 mutants were obtained from the HIV Database and Analysis Unit at the NIH: National Institute of Allergy and Infectious Diseases.

(iii) Additional relevant data were gathered from the National AIDS Control Organization (NACO) website.

Multiple sequence alignment to find conserved regions in multiple DNZs

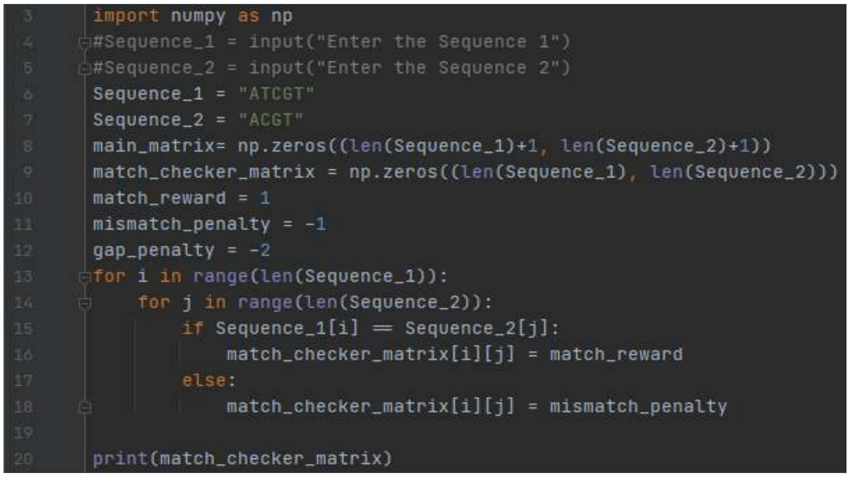
As mentioned in the introduction, conserved regions within multiple HIV DNA sequences can be identified using Multiple Sequence Alignment (MSA) methods. Sequence alignment is employed to detect regions of similarity between two or more DNA sequences [15][16]. These similarities typically indicate functional equivalence and an evolutionary relationship between two DNA sequences. So, what exactly is multiple sequence alignment (MSA)? MSA involves aligning more than two sequences to assess the similarities among them. There are several methods available to perform MSA on multiple DNA sequences, such as ClustalW, Clustal Omega, Muscle, MAFFT, and T-Coffee. In this study, Clustal Omega was used to perform multiple sequence alignment on the HIV virus sequences collected from the NCBI website and the HIV database [17][18].

Python code for sequence alignment

This section presents a Python code for sequence alignment using the Needleman-Wunsch method. The code takes two sequences as input and outputs their aligned form. To maximize similarity in the sequence alignment, all necessary logic and parameters are implemented, such as match, mismatch, and gap penalties. First, a scoring matrix is generated, and then the traceback method is applied to obtain the final alignment [19].

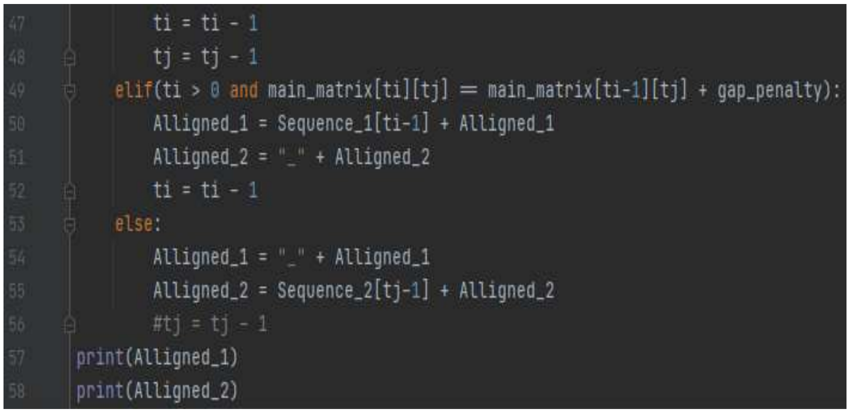
Use of mafft for multiple sequence alignment

One human HIV DNA sequence file, titled HIV1_ALL_2018_genome_DNA, was downloaded from the NIH website [19]. This large file, approximately 53 MB in size, contains 4,005 DNA sequences and is used for multiple sequence alignment with MAFFT. It represents a significant dataset of HIV DNA

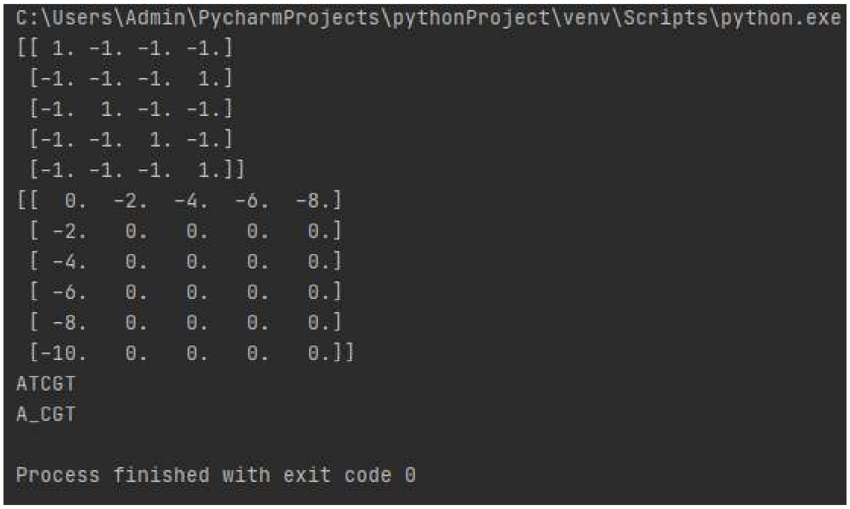


```
import numpy as np
#Sequence_1 = input("Enter the Sequence 1")
#Sequence_2 = input("Enter the Sequence 2")
Sequence_1 = "ATCGT"
Sequence_2 = "ACGT"
main_matrix= np.zeros((len(Sequence_1)+1, len(Sequence_2)+1))
match_checker_matrix = np.zeros((len(Sequence_1), len(Sequence_2)))
match_reward = 1
mismatch_penalty = -1
gap_penalty = -2
for i in range(len(Sequence_1)):
    for j in range(len(Sequence_2)):
        if Sequence_1[i] == Sequence_2[j]:
            match_checker_matrix[i][j] = match_reward
        else:
            match_checker_matrix[i][j] = mismatch_penalty

print(match_checker_matrix)
```

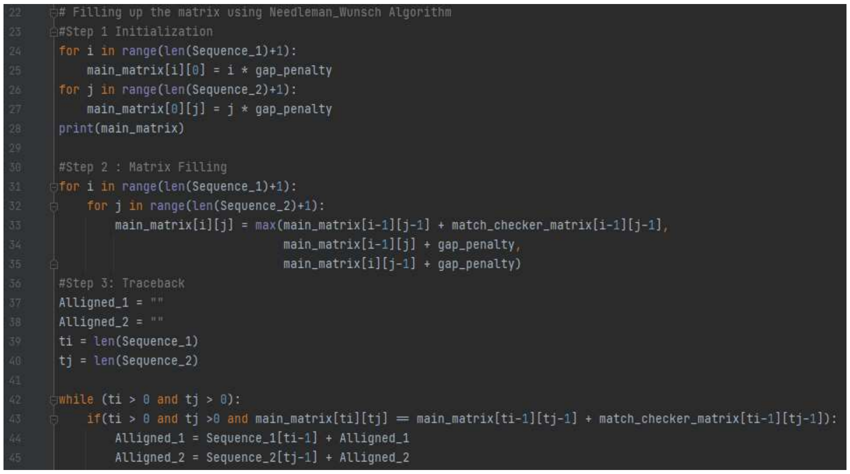


```
        ti = ti - 1
        tj = tj - 1
    elif(ti > 0 and main_matrix[ti][tj] == main_matrix[ti-1][tj] + gap_penalty):
        Alligned_1 = Sequence_1[ti-1] + Alligned_1
        Alligned_2 = "_" + Alligned_2
        ti = ti - 1
    else:
        Alligned_1 = "_" + Alligned_1
        Alligned_2 = Sequence_2[tj-1] + Alligned_2
        #tj = tj - 1
print(Alligned_1)
print(Alligned_2)
```



```
C:\Users\Admin\PycharmProjects\pythonProject\venv\Scripts\python.exe
[[ 1. -1. -1. -1.]
 [-1. -1. -1.  1.]
 [-1.  1. -1. -1.]
 [-1. -1.  1. -1.]
 [-1. -1. -1.  1.]]
[[  0.  -2.  -4.  -6.  -8.]
 [ -2.   0.   0.   0.   0.]
 [ -4.   0.   0.   0.   0.]
 [ -6.   0.   0.   0.   0.]
 [ -8.   0.   0.   0.   0.]
 [-10.   0.   0.   0.   0.]]
ATCGT
A_CGT

Process finished with exit code 0
```

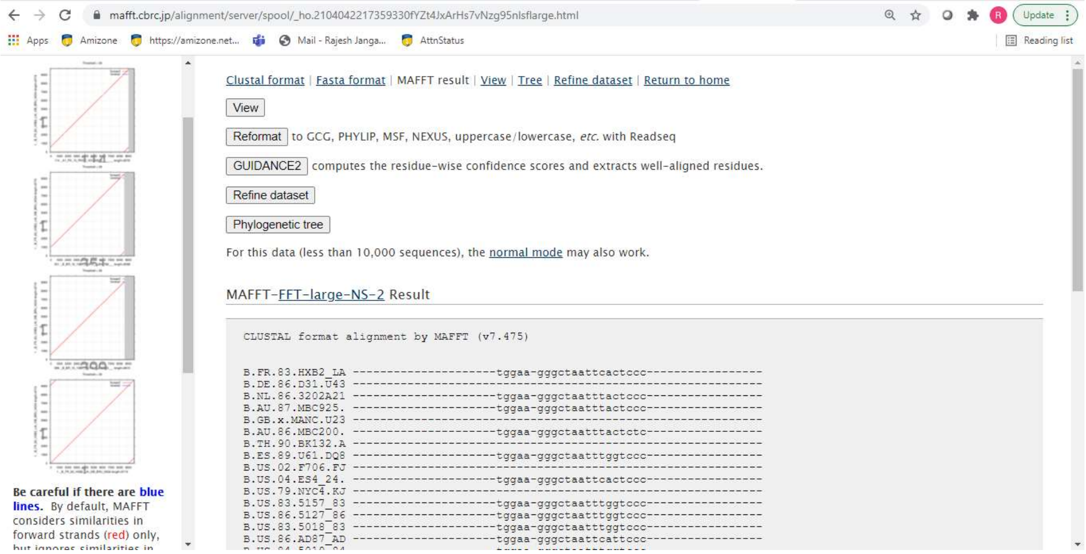


```
# Filling up the matrix using Needleman_Wunsch Algorithm
#Step 1 Initialization
for i in range(len(Sequence_1)+1):
    main_matrix[i][0] = i * gap_penalty
for j in range(len(Sequence_2)+1):
    main_matrix[0][j] = j * gap_penalty
print(main_matrix)

#Step 2 : Matrix Filling
for i in range(len(Sequence_1)+1):
    for j in range(len(Sequence_2)+1):
        main_matrix[i][j] = max(main_matrix[i-1][j-1] + match_checker_matrix[i-1][j-1],
                                main_matrix[i-1][j] + gap_penalty,
                                main_matrix[i][j-1] + gap_penalty)

#Step 3: Traceback
Alligned_1 = ""
Alligned_2 = ""
ti = len(Sequence_1)
tj = len(Sequence_2)

while (ti > 0 and tj > 0):
    if(ti > 0 and tj >0 and main_matrix[ti][tj] == main_matrix[ti-1][tj-1] + match_checker_matrix[ti-1][tj-1]):
        Alligned_1 = Sequence_1[ti-1] + Alligned_1
        Alligned_2 = Sequence_2[tj-1] + Alligned_2
```



mafft.cbrc.jp/alignment/server/spool/_ho.2104042217359330fYZt4JxArHs7vNzg95nlsflarge.html
Update
Apps
Amizone
https://amizone.net...
Mail - Rajesh Janga...
AttnStatus
Reading list

Clustal format | Fasta format | MAFFT result | View | Tree | Refine dataset | Return to home

View

Reformat to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, *etc.* with Readseq

GUIDANCE2 computes the residue-wise confidence scores and extracts well-aligned residues.

Refine dataset

Phylogenetic tree

For this data (less than 10,000 sequences), the normal mode may also work.

MAFFT-FFT-large-NS-2 Result

```
CLUSTAL format alignment by MAFFT (v7.475)


B.FR.83.HXB2_LA ---------------------tggaa-gggctaattcactccc------------------
B.DE.86.D31.U43 -------------------------------------------------------------
B.NL.86.3202A21 ---------------------tggaa-gggctaatttactccc------------------
B.AU.87.MBC925. ---------------------tggaa-gggctaatttactccc------------------
B.GB.x.MANC.U23 -------------------------------------------------------------
B.AU.86.MBC200. ---------------------tggaa-gggctaatttactctc------------------
B.TH.90.BK132.A -------------------------------------------------------------
B.ES.89.U61.DQ8 ---------------------tggaa-gggctaatttggtccc------------------
B.US.02.F706.FJ -------------------------------------------------------------
B.US.04.ES4_24. ---------------------tggaa-gggctaattcactccc------------------
B.US.79.NYC4.KJ -------------------------------------------------------------
B.US.83.5157_83 ---------------------tggaa-gggctaatttggtccc------------------
B.US.86.5127_86 ---------------------tggaa-gggctaatttggtccc------------------
B.US.83.5018_83 ---------------------tggaa-gggctaatttggtccc------------------
B.US.86.AD87_AD ---------------------tggaa-gggctaattcattccc------------------
```

Be careful if there are blue lines. By default, MAFFT considers similarities in forward strands (red) only,

Fig. 2: Result of MSA using MAFFT

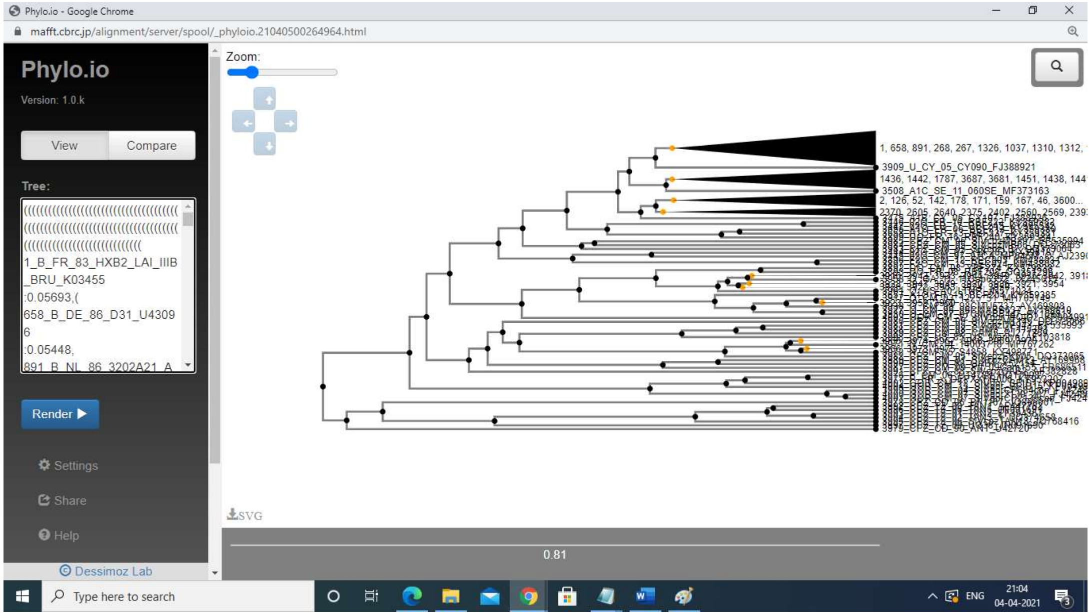
sequences, which is utilized here to identify conserved regions in the mutational patterns [19][20].

Although ClustalW and Clustal Omega are commonly used tools for multiple sequence alignment, their capacity is limited to aligning up to 4,000 sequences or a maximum file size of 4 MB. Since the HIV1_ALL_2018_genome_DNA file contains over 4,000 sequences and exceeds this size limit, MAFFT was chosen for alignment instead [21]. Above is the the result of multiple sequence alignment of *HIV1_ALL_2018_genome_DNA* using MAFFT (Figure 2).

The MAFFT also generates the phylogenetic tree Study of Phylogenetic Tree to find the during multiple sequence alignment (Figure 3)

Study of Phylogenetic Tree

After generating the phylogenetic tree, the next step is to identify conserved



((((((((((((((((((((((((((((((((((((
((((((((((((((((((((((((((((((((((((
(((((((((((((((((((((((((((
1_B_FR_83_HXB2_LAI_IIIB
_BRU_K03455
:0.05693,(
658_B_DE_86_D31_U4309
6
:0.05448,
891_B_NL_86_3202A21_A

Fig. 3: Phylogenic tree generated by MAFFT.

regions, often referred to as similarities between DNA sequences. These sequences and trees are deposited in relevant databases, such as NCBI's GenBank for sequences and TreeBase for trees, along with metadata indicating when and where the sequences were collected. If similarities between sequences are identified, efforts should be made to describe them in collaboration with taxonomists. Once described, these similarities represent the conserved regions in the mutational pattern of HIV-1 virus DNA.

Future Study

Currently, the global community, particularly the pharmaceutical industry, faces significant challenges in developing a permanent vaccine for the HIV-1/AIDS virus. Through phylogenetic studies, similarities or conserved regions between human DNA and HIV-1 DNA within the human body can be identified. By applying machine learning and artificial intelligence algorithms to this extensive collection of conserved regions and phylogenetic trees, one can pinpoint conserved regions across the entire genome, protein DNA sequences, and the full HIV-1 virus DNA sequence. These conserved regions could play a crucial role in the development of an HIV-1 vaccine.

Conflicts of Interest

The authors declare no conflicts of interest in this work.

References

1. Koonin EV, Altschul SF, Bork P, "BRCA1 protein productsFunctional motifs". Nat Genet, 13(3), pp.266-268, 1996,

2. Pagel P, Mewes HW, Frishman D, "Conservation of protein-pro-tein interactions - lessons from ascomycota". Trends Genet, 20(2), pp.72-76, 2004.

3. Jordan IK, Rogozin IB, Wolf YI, Koonin EV, "Essential genes aremore evolutionarily conserved than are nonessential genes in bacteria", Genome Res, 12(6), pp. 962-968, 2002.

4. Frishman D, Mewes HW, "Protein structural classes in five com-plete genomes", Nat Struct Biol, 4(8), pp. 626-628, 1997.

5. Gerstein M, “A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of proteinstructure”. J Mol Biol, 274(4), pp. 562-576, 1997.
6. Das R, Gerstein M, “The stability of thermophilic proteins: astudy based on comprehensive genome comparison”. FunctIntegr Genomics, 1(1), pp.76-88, 2000.
7. Thompson MJ, Eisenberg D. “Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability”. J Mol Biol, 290(2), pp. 595-604, 1999.
8. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S, “Under- standing the adaptation of Halobacterium species NRC-1 toits extreme environment through computational analysis of its genome sequence”. Genome Res, 11(10), pp. 1641-1650, 2001.
9. Gianese G, Bossa F, Pascarella S, “Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes. Proteins”, 47(2), pp. 236-249, 2002.
10. Narayan BeheraM. S. Jeevitesh, Justin Jose, Krishna Kant Alpana Dey, Javed Mazher, “Higher accuracy protein multiple sequence alignments by genetic algorithm“, Procedia Computer Science, Elsevier, Vol-108, pp. 1135-1144, 2017.
11. Lipman,D.J., et. Al, “A tool for multiple sequence alignment”, Proceedings of National Academy of Science; 86, pp. 4412-4415, 1989.
12. Thompson,J.D., et. Al, “CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”. Nucleic Acids Research., 22, pp. 4673-4680, 1994.
13. Edgar,R.C, “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. Nucleic Acids Research.,32, pp. 1792-1797, 2004.
14. Pei, Jimin, “Multiple protein sequence alignment, Current opinions in structural biology”, 1144 Narayan Behera et al. / Procedia Computer Science 108C (2017) 1135–1144 18, pp. 382:386, 2017.
15. Lloyd,S. and Snell,Q.O., “Accelerated large-scale multiple sequence alignment”. BMC Bioinformatics,; 12, pp. 466, 2011.
16. Collingridge, P.W. and Kelly, Steven, “MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments”. BMC Bioinformatics, 13, pp. 117, 2012.
17. Thompson, J.D., et. al. Balibase 3.0: Latest Developments of the Multiple Sequence alignment Benchmark. Proteins., 2005; 61, 127:136.
18. Holland, J.H. Adaptation in natural and artificial systems. Univ of Michigan press, 1975; Ann Arbor, MI.
19. Behera N. &NanjundiahV.trans-Gene Regulation in Adaptive Evolution: a Genetic Algorithm Model, Journal of Theoretical Biology 1997; 188, 153:162.
20. Behera N. &Nanjundiah V. Phenotypic plasticity can potentiate rapid evolutionary change. Journal of Theoretical Biology, 2004; 226, 177:184.
21. Zhang,C and Wong,A.K. A Genetic algorithm for multiple molecular sequence alignment. CABIOS, 1997; 13, 565:581.