IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

Identification of Conserved Domains and Motifs among different types of Polymerases

Kavitha K.R. [1], Jyothsna B.S.[2], Sreenivasulu M.V.[3], Santhosh Kumar[4], Arjun M.A.[5], Naveen Kumar M.V[6], Ashok L.V[7], Ayaz Pasha[8]

[1]Registrar, Department of Botany and PG Studies, Nrupathunga University, Bangalore, Karnataka, India.

[2]Associate Professor, Department of Botany and PG Studies, Nrupathunga University, Bangalore, Karnataka, India.

[3]Associate Professor, Department of Botany and PG Studies, Nrupathunga University, Bangalore, Karnataka, India.

[4-8]M.Sc. Botany, Department of Botany and PG Studies, Nrupathunga University, Bangalore, Karnataka, India.

Abstract: Polymerases are enzymes that catalyze the synthesis of DNA or RNA polymers whose sequence is complementary to the original template. Domains are basic structural, functional and evolutionary components of proteins or polypeptide chains that fold upon themselves. We were interested in finding conserved domain regions in these different polymerases. We did multiple sequence alignments of whole genes. Domains were also identified using SMART database and were used for alignments. Motifs YMDD, GR, SP, DVE, LT are highly conserved motifs, mostly in DNA dependent RNA and RNA dependent DNA polymerases. We also saw a high degree of conservation in the domain sequences derived from SMART databases of Cyanobacterial organisms. We hypothesize that this conservation might be because these genes might have been involved in lateral or horizontal gene transfer.

Keywords: DNA polymerase, Reverse transcriptase, Conserved domains, horizontal gene transfer

1. INTRODUCTION

Polymerases are enzymes that catalyze the synthesis of DNA or RNA polymers whose sequence is complementary to the original template. Polymerisation is a process in which the nucleotides are strung together based on a template by an enzyme polymerase. Nucleotides are combined chemically to form long chains. Polymerase enzymes contain a polymerase domain that is used for nucleotide polymerisation, and also they usually will have an exonuclease domain for nuclease activity used for proofreading or DNA damage repair.
The following are the different types of polymerases.

- DNA dependent DNA polymerase
- DNA dependent RNA polymerase
- RNA dependent DNA polymerase
- RNA dependent RNA polymerase

DNA-dependent DNA polymerases are responsible for directing the synthesis of new DNA from deoxyribonucleotide triphosphates (dNTPs) opposite an existing DNA template, which contains the genetic information critical to an organism's survival.

1.1. DNA dependent RNA polymerase

RNA polymerase, is an enzyme that synthesizes RNA from a DNA template.
It catalyses the synthesis of RNA in vitro in the presence of a DNA template and the 4 ribonucleoside-triphosphates.

1.2. RNA dependent DNA polymerase (Reverse transcriptase)

Combining reverse transcription of RNA into DNA and amplification of specific DNA targets using polymerase chain reaction
It is primarily used to measure the amount of a specific RNA.

1.3. REVERSE TRANSCRIPTASE(RT)

A reverse transcriptase (RT) is an enzyme used to generate complementary DNA (cDNA) from an RNA template, a process termed reverse transcription. In retroviruses and retrotransposons, this cDNA can then integrate into the host genome, from which new RNA copies can be made via host-cell transcription.
Reverse transcriptases have been identified in many organisms, including viruses, bacteria, animals, and plants. In these organisms, the general role of reverse transcriptase is to convert RNA sequences to cDNA sequences that are capable of

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

inserting into different areas of the genome.

1.4. RNA dependent RNA polymerase

The RNA-dependent RNA polymerase (RdRp) or RNA replicase is an enzyme that catalyzes the replication of RNA from an RNA template.

RdRp is an essential protein encoded in the genomes of all RNA-containing viruses with no DNA stage.

1.5. FUNCTIONS:

In DNA: DNA repair is the common function

In RNA: RNA synthesis and transcription is the commonly found function

These are the common motifs of both RNA and DNA polymerase

The DNA polymerases are enzymes that replicate DNA based on a template. They create DNA molecules by assembling deoxyribonucleotides. Though the main function of the DNA polymerase is to replicate DNA they also participate in DNA Repair. (Table S1)

2. DOMAIN

Domains are basic structure, functional and evolutionary components of proteins.

polypeptide chain that folds upon itself.

In Bacteria there are 3 types of DNA polymerase.

They are of

- Polymerase I
- Polymerase II
- Polymerase III

In the synthesis process of DNA, Polymerase III and in the repair process polymerase I and II is required.

The finger domain functions to bind nucleoside triphosphate along with the template base.

Thumb domains play the role of processivity, translocation and making the position of DNA.

Each domain contributes the enzymatic activity that is DNA synthesis and deoxyribose phosphate lyase. During the repair of simple base lesions, these domains are termed Polymerase and lyase respectively.

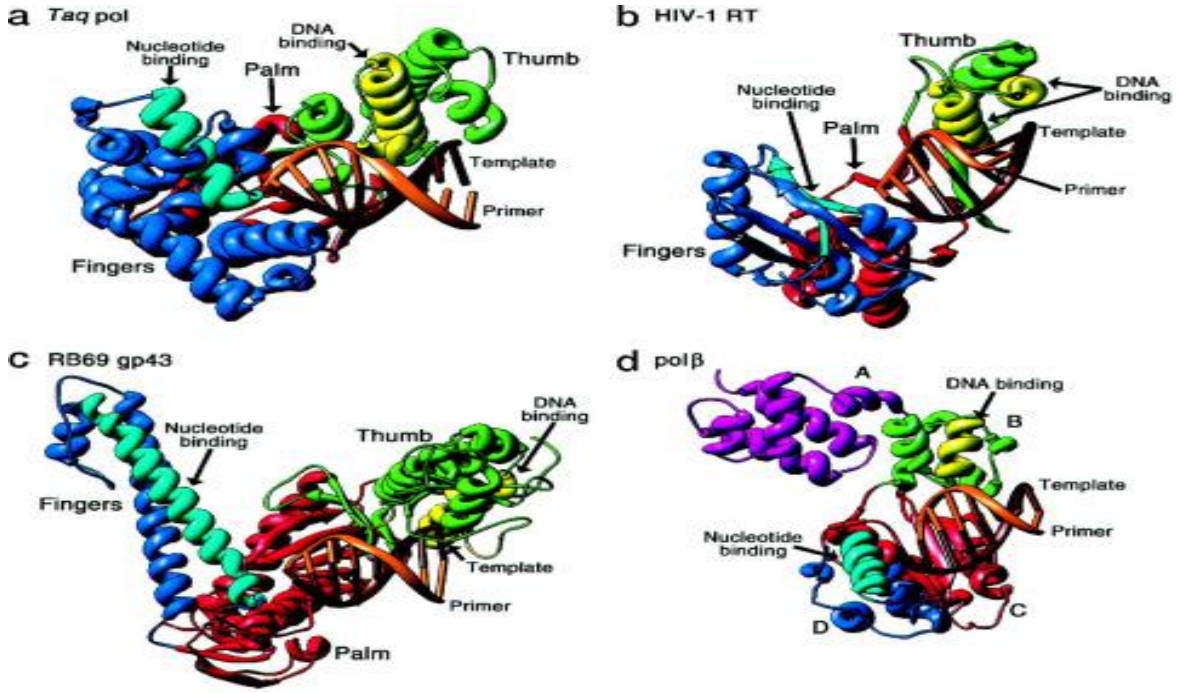

Fig 2: comparison of primer-template DNA bound to four DNA polymerases.

The following are different types found in Polymerase.

- Endonuclease domain
- Exonuclease domain

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

Polymerization and exonuclease domain

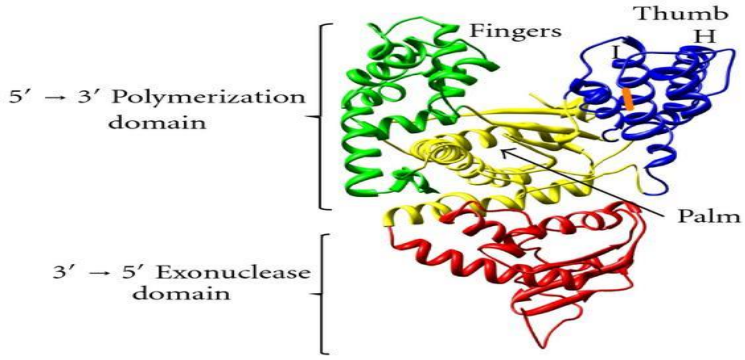

Fig 3: polymerization and endonuclease domains-

Structure of Klentaq1 (a) Structure of Klentaq1 in the absence of DNA (PDB #1KTQ) [7]. The overall structure of polymerases resembles that of a human right hand forming palm (yellow), fingers (green), and thumb (blue) domains. The palm domain houses the active site residues responsible for the 5 to 3 polymerase activity, and the 3 to 5 exonuclease domain (red) allows excision of misincorporated bases. The H1H2 Loop is disordered and missing from the structure (ends connected with an orange line).

The polymerase consists of a multidomain architecture not only the 5'-3' polymerization activity required for DNA replication but also 3'-5'and 5'-3' exonuclease activities.

The 3'-5' Exonuclease is achieved by removing incorrectly incorporated nucleotides from the growing DNA primer strand resulting in another opportunity to incorporate the correct base before continuing on with synthesis.

Polymerization domain is also called endonuclease- the DNA double helix 5'-3' strand, Exonuclease domain is 3'-5' strand

2.1. Polymerization domain

Polymerization domain in different polymerases. For several DNA-dependent DNA polymerases, it has been shown that their synthetic and degradative activities are organized in two separated modules. The functional coordination required between them to accomplish successfully the replication process is provided by important contacts with the substrate contributed by residues coming from both modules. These domains are connected by a central "linker" region adjacent to the "EGG/A" motif, the putative limit of the polymerization domain.

3. Sequence alignment

Sequence alignment is a method of arranging the biological sequence to identify the regions of similarities

Pairwise Sequence Alignment (PSA) is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

In Pairwise alignment tools, emboss needle from EMBL-EBI aligns sequences end to end and is called global alignment.

3.1. Cyanobacteria

Cyanobacteria are also called blue-green algae. Cyanobacteria are photosynthetic and aquatic; cyanobacteria are often called "blue-green algae". Cyanobacteria, any of a large, heterogeneous group of prokaryotic, principally photosynthetic organisms. Cyanobacteria resemble eukaryotic algae.

They are able to fix atmospheric nitrogen fixation, decompose organic waste and residues, detoxify heavy metals, pesticides and growth of pathogenic microorganisms in soil and water.

They produce some bioactive compounds such as vitamins, hormones, and enzymes which contribute to plant growth

Cyanobacteria are not dependent on a fixed source of carbon and, as such, are widely distributed throughout aquatic environments. These include freshwater and marine environments and in some soils.

3.2. MSA

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 rearrangements under certain conditions. These methods can be applied to DNA, RNA or protein sequences.

Pairwise sequence alignment of virus Murmansk pox virus and Simian endogenous retrovirus have done and its name and accession number and results are given below.[Table 2.1.4(b)]

Multiple sequence alignments are essential in computational analysis of protein sequences and structures, with applications in structure modelling, functional site prediction, phylogenetic analysis and sequence database searching.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

Constructing accurate multiple alignments for divergent protein sequences remains a difficult computational task, and alignment speed becomes an issue for large sequence datasets.

3.3. INTERPRO

Interpro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to new protein sequences in order to functionally characterize them. The contents of Interpro consist of diagnostic signatures and the proteins that they significantly match. The signatures consist of models (simple types, such as regular expressions or more complex ones, such as Hidden Markov models) which describe protein families, domains or sites. Models are built from the amino acid sequences of known families or domains and they are subsequently used to search unknown sequences (such as those arising from novel genome sequencing) in order to classify them. Each of the member databases of InterPro contributes towards a different niche, from very high-level, structure-based classifications (SUPERFAMILY and CATH-Gene3D) through to quite specific sub-family classifications (PRINTS and PANTHER).

Interpro intention is to provide a one-stop shop for protein classification, where all the signatures produced by the different member databases are placed into entries within the InterPro database. Signatures that represent equivalent domains, sites or families are put into the same entry and entries can also be related to one another.

3.4. SMART

Simple Modular Architecture Research Tool (SMART) is a biological database that is used in the identification and analysis of protein domains within protein sequences. SMART uses profile-hidden Markov models built from multiple sequence alignments to detect protein domains in protein sequences.

3.5. LATERAL GENE AND HORIZONTAL GENE TRANSFER:

Horizontal gene transfer (HGT) or lateral gene transfer (LGT) as a general mechanism leads to biodiversity and biological innovations in nature. Lateral (or horizontal) gene transfer (LGT) refers to the transmission of genes between individuals without a direct vertical inheritance from parents to their offspring

Horizontal gene transfer is made possible in large part by the existence of mobile genetic elements, such as plasmids (extrachromosomal genetic material), transposons (“jumping genes”), and bacteria-infecting viruses.
In nature, transformation, transduction, and conjugation are the principal mechanisms of HGT.

4. METHODS

4.1. PSA (Pairwise sequence alignment)

The following sequences were retrieved for pairwise sequence alignment using an emboss needle of Polymerase partial [homosapien] and jaagsiekte sheep retrovirus. The accession numbers are AAF88167 and AAK38686. (Table S2)
Pairwise sequence alignment of virus Murmansk pox virus and Simian endogenous retrovirus. The accession numbers are YP_009468319 and QNI40181. (Table S3)

4.2. MSA

Multiple sequence alignment using a tool Embos MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/) done for the following viruses is given in Table S4.
Multiple sequence alignment of Human immunodeficiency virus, Hepatitis B virus,
Woodchuck hepatitis virus and Simian endogenous virus. Their accession numbers are ACB36867, AGA95798, AAA69573 and AAC97565. (Table S5)
Multiple sequence alignment of Porcine mast adenovirus B, Red sea bream iridovirus, California sea lion adenovirus, Cyprinid herpesvirus 4, Varanid herpesvirus, Procavid Herpesvirus, Sea otter poxvirus, Strigid Herpesvirus 1 and their accession numbers are QNQ79208, BAK14264, YP_009032607, AIS36178, AAS17072, ABK41482, AGZ62590 and AWB09341. (Table S6)
Multiple sequence alignment of Mycobacterium tuberculosis, Calothrix. sp and Nostoc. sp. And their accession numbers are AYM47576, AEF33333 and AEF33338 (Table S7)
Multiple sequences of 30S ribosomal proteinS17(Aphanizomenon), MULTISPECIES:30S ribosomal protein S17(Calothrix)
And MULTISPECIES:30S ribosomal protein S17(Cyanobacteria). (Table S8)
Multiple sequence alignment of Bluetongue virus, Hantaan orthohantavirus, Tomato spotted wilt ortho tospovirus, Cactus virus X, Cucumber mosaic virus, Tobacco mosaic virus and Garlic latent virus. Their accession numbers are ADI49523, AAK01302, QEL52523, NP_148778, CAA25494, CAA46846 and CAA92815.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

4.3. INTERPRO: "The following sequences were used to obtain domain information in Interpro" **(https://www.ebi.ac.uk/interpro/)**
Using InterPro found the domains of Human immunodeficiency virus and hepatitis B. The accession numbers are ACB36867 and AGA95798 (Table S9)

4.4. SMART
Smart tool found the conserved sequences of Cyprinid herpesvirus 4, Procavid herpesvirus 1, Red sea bream iridovirus, Cea otter poxvirus & Varanid herpesvirus 1. And their accession numbers are AIS36178, ABK41482, BAK14264, AGZ62590 and AAS17072. (Table S10)
"These domain sequences were then aligned using a multiple sequence alignment tool MUSCLE".

5. RESULTS

5.1. PSA: Pairwise global sequence alignment of Partial polymerase (homo sapiens) and Jaagsiekte sheep using the tool Emboss NEEDLE and their results.

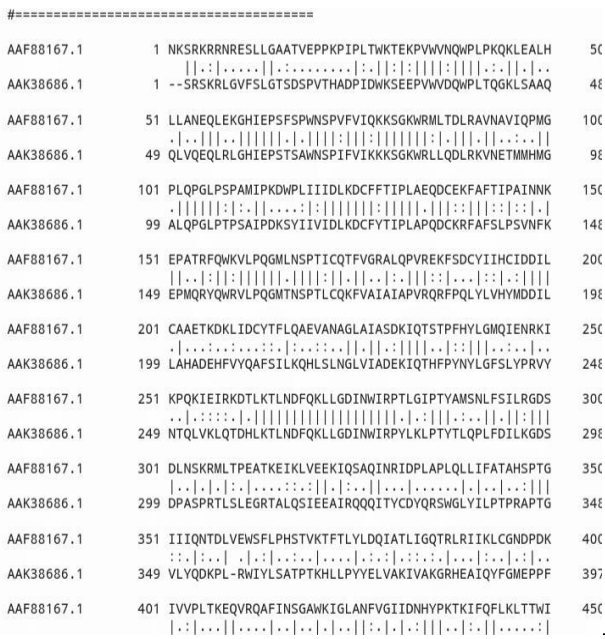

Fig 4: MSA results of RNA dependent DNA polymerase viruses.

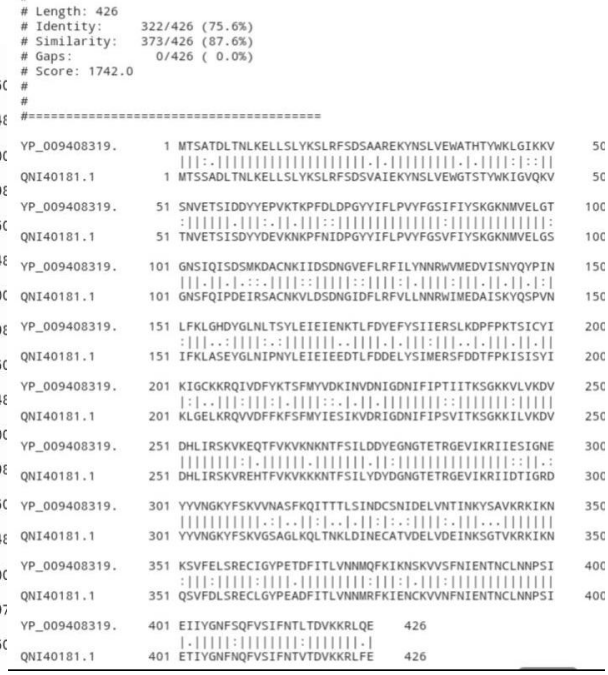

Fig 5: MSA results of RNA dependent DNA polymerase viruses

Pairwise alignment of the two protein sequences are RNA dependent DNA polymerases. the length showed an identity of **41%,** similarity of **55%** and gaps between them **10.5%**.

Pairwise alignment of the two protein sequences is RNA dependent DNA polymerases. The length of the sequence is **426** of which **75.6%** Identity, **87.6%** are similarities and gap between them is **0.0%**.

5.2. MULTIPLE SEQUENCE ALIGNMENT (MSA):
In order to see the conservation of protein sequences with each class of the polymerases we did MSA
MSA: RNA dependent DNA polymerase

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

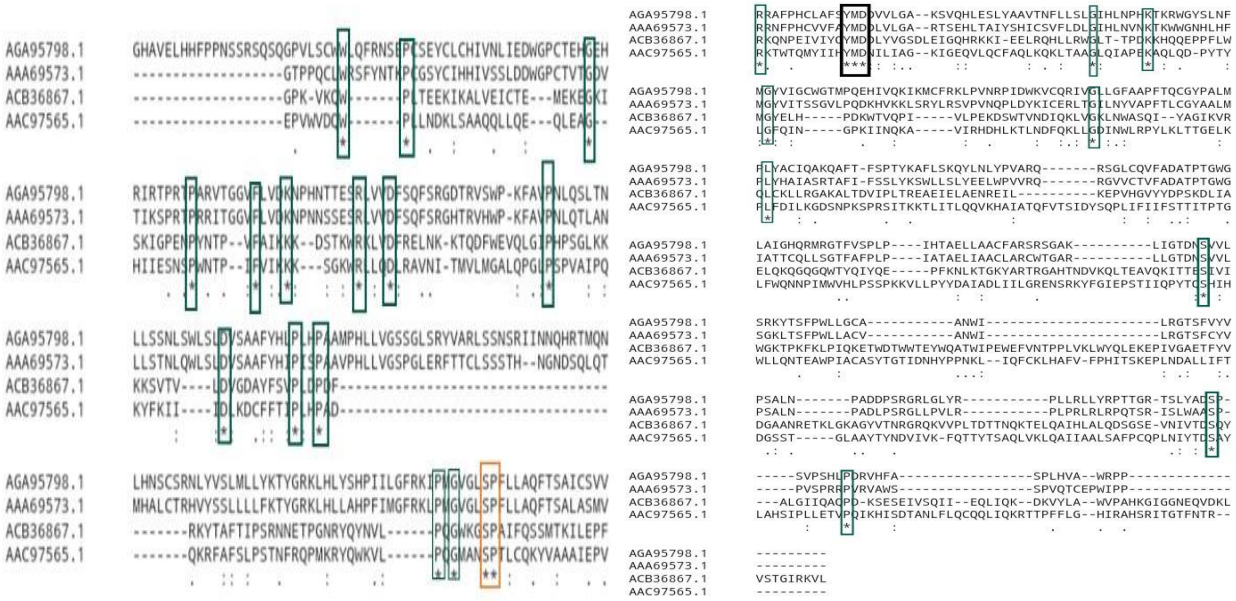

Fig 6 (a): MSA results of RNA dependent DNA polymerase. Fig 6(b): Results of RNA dependent DNA polymerase by multiple sequence alignment using tool MUSCLE.

The conserved regions are marked in green, black and orange boxes in alignment RNA dependent DNA polymerase. The "YMDD" is highly conserved in all the DNA polymerase.

5.3. DNA dependent DNA polymerase

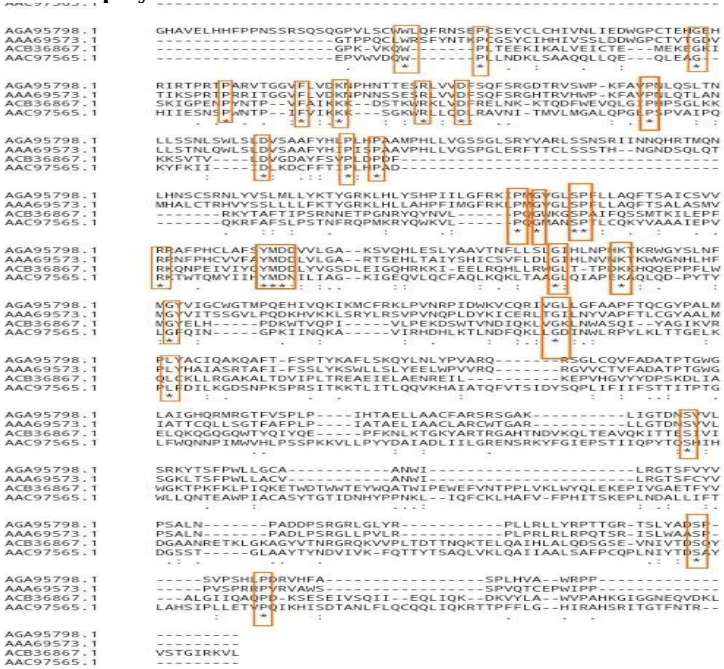

Fig 7: Results of DNA dependent DNA polymerase by multiple sequence alignment using tool MUSCLE.

The conserved regions are marked in brown boxes in the alignment of DNA dependent DNA polymerase
The" YMDD motif is found to be highly conserved in all the DNA polymerase". The motif is found to be highly conserved in all the DNA polymerase".
YMDD is an important motif found in DNA dependent DNA polymerase as well as in RNA dependent RNA polymerase. Lamivudine resistance is a result of mutations in the YMDD motif in which rt203-206th codons (Y: tyrosine; M: methionine; D: aspartic acid; D: aspartic acid) of reverse transcriptase.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

5.4. DNA dependent RNA polymerase

Cyanobacteria.

The conserved regions are marked in boxes of brown and green color in the alignment of DNA dependent RNA polymerases.

“The DVE motif is found to be highly conserved” in DNA dependent RNA polymerase

The conserved regions are marked in color boxes of green, yellow, orange & brown color in the alignment of 30S ribosomal proteinS17 (Cyanobacteria).

"The HPKYGKI & ETRPLS motifs are found to be highly conserved"

"KERVG & HDEEN motifs are also highly conserved as shown in the above pics ".

Results of RNA dependent DNA polymerase by multiple sequence alignment using tool MUSCLE (3.8).

There is no conserved sequences are obtained in RNA dependent DNA polymerase

Fig 8: Results of DNA dependent RNA polymerase (cyanobacteria) by multiple sequence alignment using tool MUSCLE.

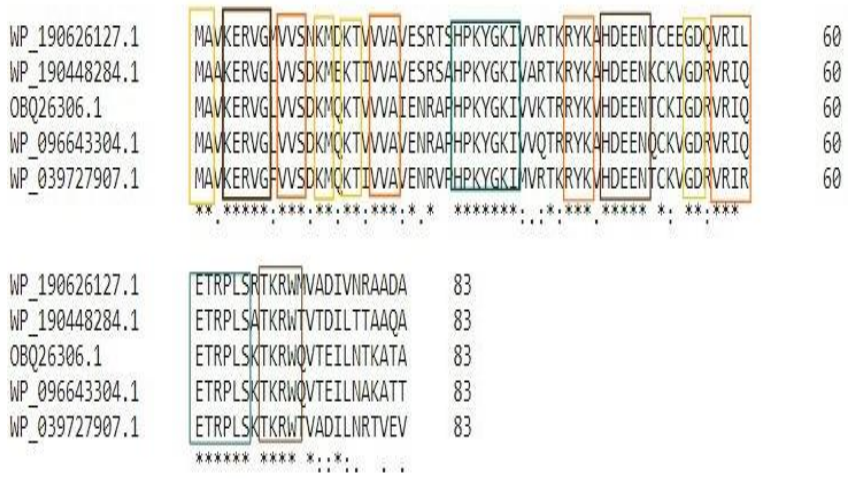

Fig 9: MSA results of 30S ribosomal proteinS17 Cyanobacteria

INTERPRO

Since the full-length sequences showed a lot of divergence between the organisms we obtained the domains and aligned domains specific sequences.

We obtained domains which are obtained by using the result of MSA between HIV and Hepatitis B.

The use of an RNA template to produce DNA, for integration into the host genome and exploitation of a host cell, is a strategy employed in the replication of retroid elements, such as the retroviruses and bacterial retrons. The enzyme catalyzing polymerisation is an RNA-directed DNA-polymerase, or reverse transcriptase (RT)

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

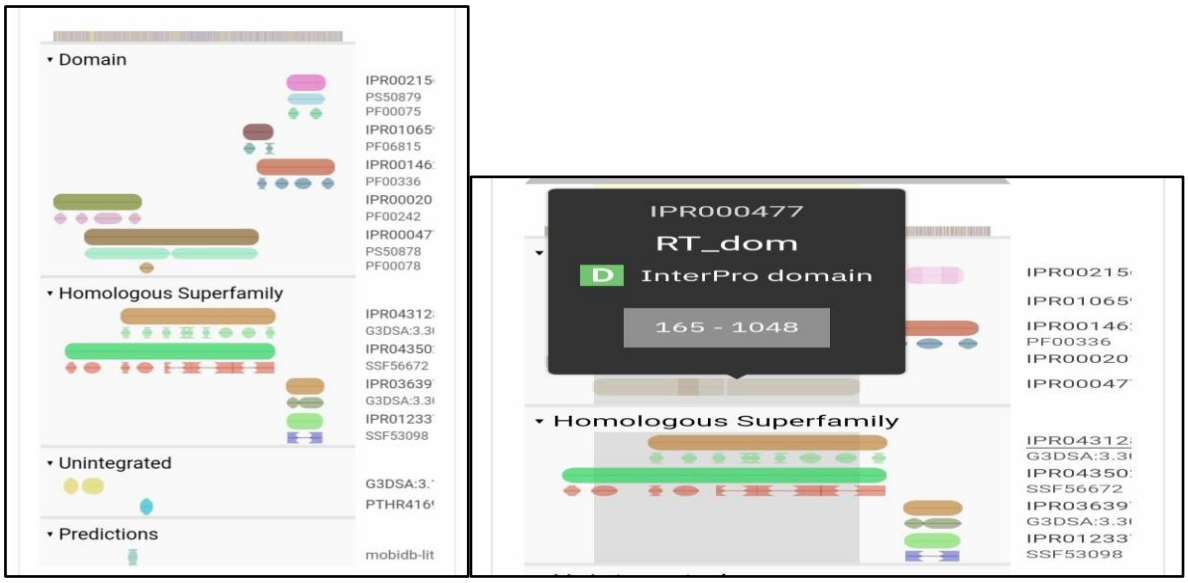

Fig 11: The domains or motifs that are obtained. Fig 11(a): Reverse transcriptase domain. by INTERPRO

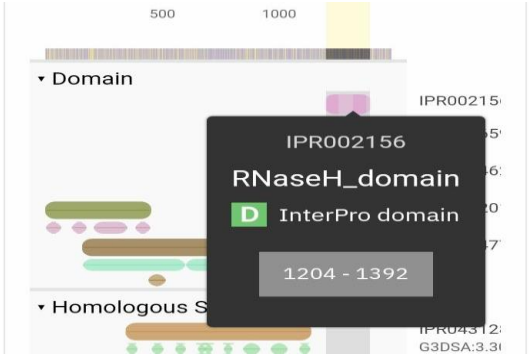

Fig 11(b): Ribonuclease H domain.

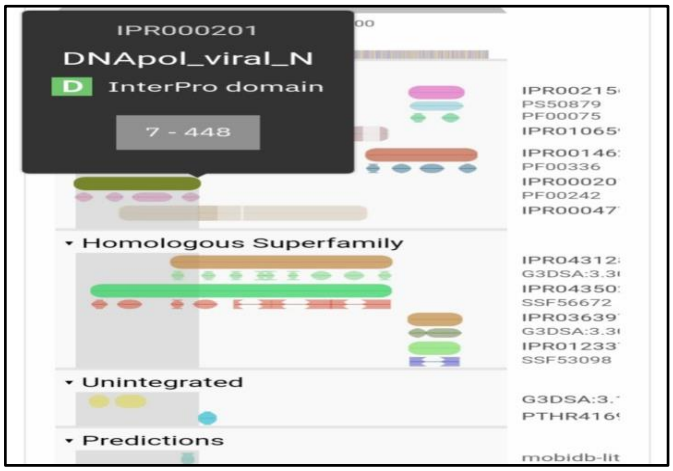

Fig 11(c): DNA polymerase N terminus virus.

The RNase H domain is responsible for the hydrolysis of the RNA portion of RNA x DNA hybrids, and this activity requires the presence of divalent cations (Mg2+ or Mn2+) that bind its active site. This domain is a part of a large family of homologous RNase H enzymes of which the RNase HI protein from Escherichia coli is the best characterised.

This domain is at the N terminus of hepadnavirus P proteins and covers the so-called terminal protein and the spacer region of the protein. This domain is always associated with IPR000477 and IPR001462.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

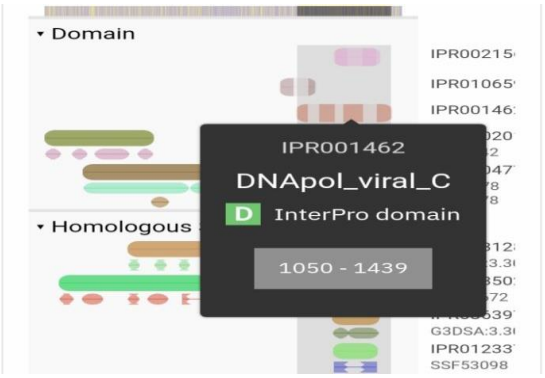

Fig 11(d): DNA polymerase C terminus virus.

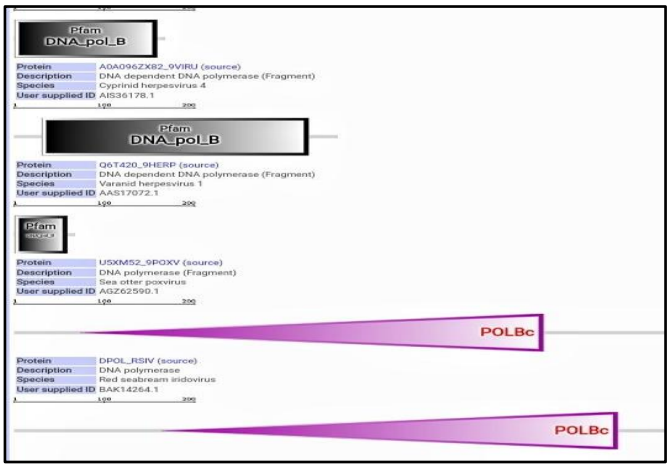

Fig 12: Smart batch retrieval results of MSA.

This domain is at the C terminus of hepatitis B-type viruses P proteins and represents a functional domain that controls the RNase H activities of the protein. The domain is always associated with IPR000201 and IPR000477.

SMART

Domains of DNA_ pol_ B, P fam, POLBc we got from DNA polymerase and the sequence which we used to multiple sequence alignment to find conserved domain regions. The sequences of DNA pol B domain were retrieved from the SMART database and only the domain sequences were aligned through multiple sequence alignment.

Fig 13: MSA results of DNA pol B & POLBc.

"GR motif is conserved and L motif is semi conserved" is obtained by MSA of Procavid herpesvirus, Cyprinid herpesvirus core, Varanid herpesvirus 1, Partial (sea otter pox virus) and Red sea iridovirus.

6. CONCLUSION

Here we conclude that we saw a high degree of conservation in the domain sequences derived from the databases of polymerases and in cyanobacterial organisms. Hypothetically, these conservations might be observed because of the motifs which are responsible for the lateral and horizontal gene transfer.

Finally, the cyanobacterial organism motifs are especially high conservation showing that, HPKYGKI, ETRPLS, KERVG and HDEEN.

Even though we got conserved regions in RNA dependent DNA polymerase and DNA dependent RNA polymerase, "YMDD" motifs are most conserved by Multiple sequence alignment and hence we assumed. There is a high degree of conservation in sequences of Cyanobacteria. So we assume that it is responsible for lateral or horizontal gene transformation in these polymerases.

7. AUTHORS NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the pa-per was free of plagiarism.

8. REFERENCES

[1] Delarue M, Poch O, Tordo N, Moras D, Argos P. An attempt to unify the structure of polymerases. Prot Engineering. 1990; 3:461–467. Pioneer work in the identification of conserved regions in various polymerase types and comparison with the E. coli polymerase I fold. [PubMed] [Google Scholar]

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9265

[2] Joyce CM. Choosing the right sugar: how polymerases select a nucleotide substrate. Proc Natl Acad Sci USA. 1997; 94:1619–1622. Comparison of residues in the active site of different types of DNA and RNA polymerases. Emphasis is put on the possible substrate discriminative mechanisms. [PMC free article] [PubMed] [Google Scholar]

[3] Falaschi A, Kornberg A (April 1966). "Biochemical studies of bacterial sporulation. II. Deoxy- ribonucleic acid polymerase in spores of Bacillus subtilis". The Journal of Biological Chemistry. 241 (7): 1478–82. doi:10.1016/S0021-9258(18)96736-0. PMID 4957767.

[4] Gouzy J., Corpet,F. and Kahn,D. (1999) Whole genome protein domain analysis using a new method for domain clustering. Comput. Chem., 23, 333–340. [PubMed] [Google Scholar]

[5] Joyce CM, Steitz TA. Function and structure relationships in DNA polymerases. Annu Rev Biochem. 1994; 63:777-822. doi: 10.1146/annurev.bi.63.070194.004021.PMID:7526780 doi:http://dx.doi.org/10.1146/annurev.bi.63.070194.004021

[6] Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. 55, 709–742 (2001).

[7] Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. Nature 405, 299–304 (2000).

[8] Letunic, Ivica; Doerks, Tobias; Bork, Peer (January 2015). "SMART: recent updates, new developments and status in 2015". Nucleic Acids Research. 43 (Database issue): D257–260. doi:10.1093/nar/gku949. ISSN 1362-4962. PMC 4384020. PMID 25300481.

[9] Bollum FJ (August 1960). "Calf thymus polymerase". The Journal of Biological Chemistry. 235 (8): 2399–403. doi:10.1016/S0021-9258(18)64634-4. PMID

