IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

A Comprehensive Review on Protein Sequence Analysis Techniques

P. Haritha[1], P. Shanmugavadivu[2*], S. Dhamodharan[3]

[1,2,3]Dept. of Computer Science and Applications, Gandhigram Rural Institute (Deemed to be University), Gandhigram, Dindigul, Tamil Nadu, India.

[*]*Corresponding Author: psvadivu67@gmail.com , Tel.: +91-94437-36780*

Available online at: www.ijcseonline.org

Accepted: 19/Jul/2018, Published: 31/Jul/2018

Abstract—Sequence analysis aims to explorethe pattern of DNA, RNA and Protein sequences. This analysis involves sequence alignment and extracts the patterns in sequences for the purpose of classification. Sequence analysis can be performed in several ways such as comparison of sequences, computation of similarity measure and mutation detection, based the degree of conserved sequences. It is useful to perform evolutionary analysis, prediction of gene/protein structures and reconstruction of DNA sequences. Sequence alignment plays a vital role in finding similarities among multiple sequences. This article gives an account on the familiar protein sequences alignment techniques being reported. Sequence alignments are of three types: Pair-wise Alignment, Multiple Sequence Alignment and Structural Alignment. The scope of this survey limits to the first two types of alignment. Pair-wise alignment handles a maximum of two sequences at a time, whereas multiple sequence alignment handles more than two sequences. This paper gives a synoptic view on the mechanisms, merits and demerits of select pairwise alignments and multiple sequence alignment techniques to find similarities among sequences.

Keywords—Protein Sequence, Protein Alignment, Local Alignment, Global Alignment, Multiple Sequence Alignment.

I. INTRODUCTION

The availability of voluminous database on protein sequences and the computational tools for protein sequence analysis has widened the frontiers of research on protein and its associated domains.Protein sequencedeals with identification and analysis of protein structure and polypeptic chains from the protein databases [1]. The sequence similarity measures aim to reveal the homology (i.e. similarity) among the given sequences and the query sequence in order to quantify the local and global sequences The inferences drawn from protein sequence analysis finds application in protein sequence classification, expression analysis, protein prediction, protein evolution as well as in Genetic Engineering and Bioinformatics. Proteins are classified based on the composition of amino acids. This article summarizes the principle and performance of select sequence alignment techniques designed for the computation of protein sequencessimilarities. In this research article, section I describe the basics of sequence analysis. Section II defines the common terminologies associated with genetics. Section III and Section IV explains on sequence analysis and sequence alignment techniques, respectively. In section V, the conclusions of this study are given.

II. BASIC TERMINOLOGIES

In this section, the common terminologies associated with DNA (Deoxyribonucleic Acid), RNA (Ribonucleic Acid) Codons and Proteins are described.

A. DNA

DNA is a molecular unit that depicts the genetic codes of an organism and its biological/genetic information. Each DNA molecule appears as a composition of two strands, twisted in anti-parallel fashion, in the form of double helix. The helical strands are made up of monomeric nucleotide units, known as polynucleotides. Each unit of nucleotide contains any one of the four nucleobases - *Adenine* (*A*), *Guanine (G)*, *Thymine* (*T*) and *Cytosine* (*C*), along with deoxyribose and a phosphate. The double helical structure is formed by *A* pairing with *T*and *C*pairing with *G*. This arrangement assures the replication of identical copies of DNA leading to the creation of daughter cells from the parent's cells, during cell division [2]B. RNARNA unit plays a vital role in processing the gene expressions. This single stranded molecule consists of four

base pairs namely*A*, *G*, *C*, and *U*(*Uracil*) [3]. The process of generating RNA from the copies of DNAis referred as transcription. The most common types of RNA are: mRNA (Messenger RNA), tRNA (Transfer RNA), rRNA (Ribosomal RNA). These RNAs are primarily engaged in protein synthesis. The functional characteristicsof RNA is categorized into two broad classes: informational andoperational. mRNA belongs to informational category whereas the other two types belongs to operational category. mRNA plays major part in proteinsencoding. In RNA, nucleotides*A*pairs with *U*, *C*with *G*and *G* with *U* [4].

C. Codon

The sequence of nucleotide triplets are comprised of three mRNA nucleotides, depicting the genetic code, known as codons. These sequences follow the predefined base rules. These genetic codes are debased into 18 amino acids consisting of two, three, four or six codons [5].

D. Protein

The process of rephrasing the RNA sequenceswith respect to amino acids is referred as translation. The DNA is transformed into RNA, which in turn is synthesized into protein sequences. However, the protein sequence once generated cannot be traced back to DNA. These sequences are made up of amino acids. The process of protein synthesis
[6] involves three phases: Initiation; Locating the *AUG* initiator codon; and Elongation, which results in the formation of polypeptides (i.e. group of amino acids).

E. Types of Sequence Conservation

Conservation of protein sequence is divided into three classes: conserved; semi-conserved and non-conserved proteins. A detailed description on conservation is reported in [7]. For a given three distinct protein sequences for evolutionary analysis, the proteins are arranged in three separate rows. Each column of those protein sequences are compared for similarity. The sequence is labelled as conserved if the pattern of amino acids across the given three rows are the same. Among the three rows, if two rows alone exhibit similarity, then the label is semi-conserved and non-conserved otherwise.

III. SEQUENCE ANALYSIS

The abundance of DNA sequence data has prompted the exponential growth of research and development of computer-based analytical tools [8].The wide spectrum of tools and techniques reported in the literature, readily help the researchers to analyze the nucleotides and protein sequences. A few prominent techniques are outlined in this article.

A. Sequence Alignment

Sequence Alignment plays a major role in finding the similarity among sequences of proteins, DNA and RNA and thereby analyze the evolutionary relationships among those
units. The inferences ideally assist to explore the relationship among the collection of related proteins. The amino acids in the given set of protein sequences are compared. Conventionally, the amino acids are arranged in a linear fashion. The alignment tool looks for the identical amino acids in each column and otherwise blank is inserted [9].The applications of alignments include sequence assembly, sequence annotation, structural and functional prediction of genes/proteins, phylogeny and evolutionary analysis [10]. The primary types of alignments are: Pairwise Alignment; Multiple Sequence Alignment; and Structural Alignment. The pairwise sequence alignment considers only two sequences, whereas multiple sequence alignment relates more than two sequences to find similarities and conserved sequences.

B.Pairwise Alignment

Pairwise Alignment [11] is used to find the similarities between two sequences. This sequence alignment is able to reveal the exact matches of similar sequences. The types of pairwise alignment namely, Dot Plot Matrix Method and Dynamic Programming are discussed in the following section.

C.Dot Plot Matrix Method

Dot plot matrix method [12] compares two sequences, in order to find exact matches. One sequence is arranged at the top of the matrix, while the other sequence is arranged downwards at the left side of the matrix. The process of comparison starts at the first nucleotide notation in the first row and first character for every similar pattern being found, a dot is placed. Dot plot appears in diagonal, when the sequences are closely related. This approach is very easy to understand, analyze, interpret and implement. However, it requires extra memory space to store the duplicated exact matches is across the diagonal and the certain areas in the plot is considered as empty or noise.

D.Dynamic Programming

Dynamic Programming is suitable for solving larger problems by dividing it into several subproblems. According to sequence alignment, dynamic programming is implemented locally and globally[13]. It is assured to ascertain optimal alignment based on the scoring function.

IV. SEQUENCE ALIGNMENT METHODS

A. Local Alignment

The fundamental task of bioinformatics is to perform alignment on DNA as well as on protein sequences. The further analysis on evolutionary relationship helps to determine whether the protein sequences are inherited from common ancestor. The evolution of sequences can make changes among the existing sequences by insertions and deletions of amino acids. Hence, the sequence alignment is able to build the true evolutionary relationship by matching conserved regions and by detecting mismatches; gaps are substituted in the regions where mutations (alterations of nucleotides or proteins) occur. Conventionally, alignment is not performed with the entire sequences. Instead, local similarity is considered. In this way, Smith Waterman [14] algorithm is used to find the optimal local alignments of the sequences.It involves only parts of the sequences. The mechanism of *Smith Waterman algorithm*is: Initialize the matrix with zeros in 0^{th} row and 0^{th} column.The scoring matrix is computed. The technique fixes scores for matches, mismatches and gap penalty for amino acids in protein sequences. To illustrate, the score is assigned as3 for every match found; 2 for mismatch; and 1 for gap penalty.Finally, the scores are summed up to obtain the similarity score. From the lowest right corner of the matrix, the optimal path is traced, starting from highest score to zero score. The idea of Smith-Waterman is to put 0s in the equations, in order to trim the prefixes and suffixes of the alignment with negative score. This approach results in optimal local alignment.

I.FASTA

FASTA [15] representing FAST-All (i.e. Fast comparison of proteins or nucleotides) was proposed in 1995 which is an improved version of FASTP (Fast Protein) developed in 1985. *FASTA* can perform DNA searches and evaluate the statistical significance. The types of *FASTA* program are TFASTAX, TFASTAY, FASTAX and FASTAY. TFASTA and TFASTAY manipulate queries across DNA libraries, whereas FASTAX and FASTAY manipulate protein database. FASTA is designed on a heuristic algorithm that searches for k-tuple sub-words. Initially, *FASTA* identifies the regions which are identical. In second step, PAM-250 matrix is used for rescore the best regions, identified in the previous step. The high scoring diagonals are joined along with gaps.Finally, it generates score of optimal alignment, using *Smith-Waterman algorithm*. The output of *FASTA* program are in four parts, the first part consists of information about database which compares the sequences and exposes the similarities using the alignment technique of *FASTA;* the second part represents histogram that displays the distribution of scores; the third lists the matched sequences and statistical information and the fourth displays the alignments.

II.BLAST

BLAST stands for Basic Local Alignment Search Tool. Itcompares two different proteins or nucleotides. The variants of BLAST are:megaBLAST, BLASTN (BLAST Nucleotide), BLASTP (BLAST Protein), BLASTX, TBLASTN, PSI-BLAST, RPSBLAST and DELTA-BLAST. megaBLASTsearches for nucleotide-nucleotide similarity sequences. BLASTN finds nucleotide-nucleotide distant sequences. BLASTP is used to compare protein-protein sequences using BLASTX and TBLASTN as the basis to perform the searches. BLASTX performs searches on protein database for a translated nucleotide query. TBLASTX performs searches on nucleotide database for a translated protein query. PSI-BLAST initially performs BLASTP to

collect information to determine Position-Specific-Scoring Matrix (PSSM) and then it uses PSSM to search protein sequences in a database. RPSBLAST searches a protein query quickly across a database using PSSM. DELTA-BLAST works similar to RPSBLAST but faster than RPS BLAST.

BLAST performs local alignment in three phases: Setup, Preliminary Search and Traceback [16]. The setup phase produces a set of words and their fixed length according to the query given to search. The preliminary phase, detects the matches corresponding to the words and score is created. In Traceback, gapped extensions compute insertions and deletions. BLAST[17] works faster than dynamic programming methods such as Waterman, Wunsch. BLAST can perform alignments with two sequences whereas BLAST2 results multiple local alignment among two sequences.

III.Fast Dynamic Programming

Smith Waterman algorithm takes quadratic time and space to perform alignments, whilst FASTA and BLAST reduce the time complexity of sequence alignment. This method uses N-gram encoding and reduced amino acids. In [18], the amino acids are

reduced and clustering is performed to group the similar amino acid as one group. N-gram of non-overlapping words is used and they are encoded based on reduced amino acids. Then, Smith Waterman algorithm is applied on reduced amino acids to find similarities. The N-gram Smith Waterman algorithm consists of two phases: transformation phase and similarity calculation phase. The former transforms the sequences into integer, based on group code. The latter one takes those integer values as input and calculates scoring matrix to find maximum similarity score. Thus, it reduces the length of protein sequences while comparing the sequences.

Table.1 Comparison of Local Alignment Methods

Author(s)/ Method/ Methodology	Advantages	Disadvantages
R. Mott,2005 Smith–Waterman It includes three steps such as initialization of matrix, Computation of Score and Traceback. Here, a part of the sequence is only involved [14].	Finds optimal local alignment. Exhibits best performance in terms of accuracy.	Time consuming Computationally complex
E.S. Donkor, et. al., 2014. FASTA PAM250 matrix is used and rescores the highest scoring regions [15]	Uses words to perform searches. Rescore assures optimal alignment. Fast and suitable for sequence alignment.	It is less sensitive to work with highly divergent sequences. N-grams (words) tend to overlap, while searching for similarities.
T. Madden, 2013. BLAST Works in three phases: Setup;	Performs faster than FASTA	Uses reasonable amount of resources.
Preliminary and Tracebacks. [16], [17].		Doesn't guarantees on quality
N.A.A. Rashid, et al.,, 2006. Fast Dynamic Programming N-gram words and Reduced amino acids are used[18].	Similar amino acids are grouped to ensure minimize space. Non-overlapping N-gram words are usedfor similarities search.	Requires more space and time.

B. Global Alignment

Global alignment works on the entire sequence of a given DNA or protein sequence.*Needleman Wunsch* Algorithm [19] is used for sequences alignment in most of the bioinformatics studies. It performs global alignment and determines the optimum matches between any two similar sequences. The computational process of this method is [20]: The scoring matrix is initialized with the sequences at the first row and column. A gap is appended at the$(0,0)^{th}$ cell and initialized as 0. The score is computed and each cell is filled with the maximum value among three scoring alternatives. Traceback starts from the last cell of the matrix at the bottom right. It moves backward through three different moves such as diagonally, left or upwards. The direction is decided based on yield of maximum value among three scoring alternatives. If the direction is diagonal, then the letters are aligned as such. For the left direction, a gap is introduced in the sequence located at the left side of the matrix. For upward direction, a gap is introduced in the sequence located at the top of the matrix.

I. AVID

The *AVID alignment* method [21]is developed to overcome the defects of existing global alignment methods. It is fast, reliable and sensitive in finding homologous regions and it eliminates false positive problem occurred in local alignment programs. It can align thousands of sequence pairs. The performance of this method is estimated with both *local and global alignment* methods. Itis developed to align larger genomic sequence. It includes components such as repeat masking, finding matches using suffix trees and anchor selection. Repeat masking classifies the matches into two groups such as overlapping repeats and non-overlapping repeats. Finding matches using suffix trees calculate the maximal match by concatenating two sequences into single string and places *N* between them. Then, if the matches cross the boundary of those two sequences, then is said to be maximal match. In anchor selection, the process of anchoring and aligning of sequences is performed after finding matches. This method is suitable for lengthy sequences. If it is applied in short sequences, the process will be executed faster but with lower alignment quality. After anchoring process, the final global alignment is performed.

II.LAGAN

LAGAN (Limited Area Global Alignment of Nucleotides)[22] uses three major steps to align genomic sequences. The steps are: i). generation of local alignments of two sequences, ii) constructing a rough global map and iii) computation of global alignment on a rough global map. In *AVID*, exact matching words are used to perform local alignment. *LAGAN* uses *CHAOS* algorithm that uses exact words to detect local alignments. *CHAOS*is applied in areas of anchors recursively to construct global alignment.

III.ACANA

Accurate Anchoring Alignment (ACANA) [23] performs both local and global alignment and it uses an anchoring technique. *ACANA*performs sequence analysis even with cross species. This method uses anchoring to perform alignment. Smith Waterman is recursively applied to identify near optimal local alignment as anchoring regions, instead of considering matched words as anchoring regions. It does not deal with sequence in versions. A new algorithm [24] is used for solving inversion problems.

IV.FOGSAA

FOGSAA (Fast Alignment Global Sequence Alignment Algorithm) [25] shows same results of *Needleman-Wunsch* method with less amount of time. It gains time of 70%-90% for highly similar nucleotide sequences and 54%-70% for sequences of 30%-80% similarity. *FOGSAA* constitutes higher number of matches along with lower number of mismatches and gaps. It provides improved alignment than the heuristic alignment. It builds a tree which is used for finding optimal alignment. It starts building the tree with the given sequences. At an intermediate point, it is verified if some other branch is better than the current one. If found so, it is expanded further. This procedure is repeated until it gets optimal alignment. This method performs better, compared to the other methods namely *ACANA*, *AVID* and *CLUSTAL*. The major drawback of this technique is that it requires exponential time in the worst case.

V.Approximate Global Alignment

Generally, the global alignment methods/tools suffer from two major problems: high computational time; and poor quality of alignment for less time and space. As a remedial approach, approximate global alignment [26] proposes two dynamic programming methods namely, bounded and unbounded alignment to perform global alignment problems. In first method, the percentage of approximation is considered as input. On computing the matrix, the distance distribution of unaligned suffixes is calculated. The best Traceback passes through the entry which suffers at most *k* edit operations with a probability more than (100–*p*)%. This is said to be *p*% approximation for bounded alignment. The second method is for unbounded alignment, which is similar to bounded, but upper bound is calculated initially for unbounded alignment.

VI.GPCODON Alignment

GPCODON overcomes the traditional alignments, which are not able to handle larger sequences. The alignments are based on codons instead of nucleotides. This method [27] tends to increase accuracy of pairwise global alignment. It achieves better alignment with high score empirical codon substitution matrix which is used score computation. After calculating matrix, Traceback is performed to get optimal alignment. This method finds optimal alignment between two DNA sequences based on codon, instead of nucleotides. This method should be enhancedand it should be verified for larger datasets alignment.

VII.Parallel Approach

Parallel approaches mainly aim to speed up the computation of alignment. This approach [28] proposes a modification in *Needleman-Wunsch algorithm*, to achieve faster computation. This method accepts two inputs sequences to be aligned. The

sequences are arranged as rows and columns that initialize an array. The scoring scheme assigns *1* for match and *-1* for mismatch. Threads are used to fill the array using the principle of parallel processing. Finally, Tracebackis performed to obtain the best alignment.

Table.2 Comparison of Global Alignment Methods

Author(s)/ Method/ Methodology	Advantages	Disadvantages
S.R. Harris, et al., 2014 Needleman- Wunsch Involves three steps: Matrix Initialization; Computing Score and Traceback. It involves the entire sequence [19], [20].	Finds optimal global alignment	Consumes more time and space.
N. Bray et al., 2003 AVID Anchoring selection is used to find matches of lengthy sequences [21].	Aligns larger genomic regions. Assures faster computation	If the sequence is shorter, it is executed quickly. . Depicts lower quality in alignment
M. Brudno,et al., 2015 LAGAN CHAOS algorithm is used to detect local alignments and this algorithm is applied recursively in the area of anchors, to get global alignment [22]	Operates with in exact words to perform alignments.	It uses homologous sequences.
W.Huang, et al,, 2006 ACANA Anchoring technique is used to analyze sequences of cross species [23]	Aligns divergent sequences at both local and global levels	Score recalculation by dynamic programming makes the computation costlier.
A. Chakraborty, et al2013 FOGSAA Tree is constructed to find optimal alignment [25]	It is Faster and provides better results	Requires exponential time in worst case
T. Kahveci, et al, 2005 Approximate Global Alignment Bounded and Unbounded alignment are used[26]	Three times faster than other bounded alignment. Two Times faster than other unbounded alignment	Eliminating entries from the matrix makes it slower.
Z.A. Fareed, et al2016 GPCodon Codons are used for alignment [27]	Provides accurate results on pairwise global alignment	Execution time should be reduced
M. Sethi, et al2016 Parallel Approach Parallel approach is used for comparing two sequences[28]	Time taken for execution is less than serial implementation	Formation of threads demands additional computation time.

C. Multiple Sequence Alignment

Multiple Sequence Alignment is used to align more than two sequences. Such alignments show the conserved sequence regions among several sequences. This method constructs phylogentic tree that represents evolutionary relationships. The two types of methods are: Progressive and Iterative.

A.Progressive

This method aligns sequences, which are closely related, and adds the remaining sequences, which are less related. This mechanism uses a series of pairwise alignment [29]. *Needleman Wunsch* pairwise alignment is used iteratively to perform multiple sequence alignment and to construct evolutionary tree [30].

I.Clustal Series

Clustal [31] program combines progressive alignment technique with memory-efficient dynamic programming. A series of pair-wise alignment follows branching order of guide tree which is used as a reference while constructing the progressive multiple sequence alignments. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method constructs guide tree. *ClustalV* is comprised of alignment of existing alignments and *Neighbour Joining method*is used to construct multiple sequence alignment. The subsequent multiple sequence alignment generates the tree. *ClustalW* improved alignment algorithm includes sequence weighting, position specific gap penalties and choice of weight to the matrix. Neighbour Joining method is used for the construction of dendogram instead of *UPGMA* method.

II.T-Coffee

The major problem with progressive alignment is the selection of initial sequence with which the process of alignment begins. This process may invite some errors. *T-Coffee* [32] tries to solve this problem. A mixture of global and local pairwise alignments is used to compute multiple alignments. The best fits of pairwise alignments are selected from the input library to construct multiple alignments. This makes the method to work faster and assures minimum errors.

B.Iterative

Iterative method is used to overcome the error incurred due to progressive alignment. This method can realign the previously aligned sequences, in order to determine high quality alignment score. The iteration stops when the score values converge.

I.DIALIGN

DIALIGN [33] synthesizes both global and local alignments. It performs local alignment while the parts of sequences are not related. Global alignment is performed if the entire length of the sequences is globally related. This method aligns region of similarities. It combines global and local alignment to perform multiple alignments.

II.SAGA

The problem of performing alignment without objective functions and errors in intermediate alignments is addressed in [34].It uses Genetic Algorithm and Sequence Alignment by Genetic Algorithm (*SAGA*) to perform globally multiple sequence alignment. This method sets an objective function to obtain best alignment. Initially, generation zero is generated, which considers alignment as their population. The choice of selecting next generation is based on fitness, which is measured by the Objective Function. These steps are iterated. The new pieces of alignments are obtained by mutations and they are synthesized by crossover. The iterations stops when there are no further improvements in alignments.

Table 3 Comparison of Multiple Sequence Alignments

Authors/ Method/ Methodology	Advantages	Disadvantages
J.D. Thampson, et al., 1994 *Clustal Series* Neighbour Joining method is used [31].	Used to create pylogenetic trees. Execution time is reduced. Neighbourhood joining method can handle sequences long in length	Errors made in initial alignment are not rectified
C. Notredame, et al.,, 2000 *T-Coffee*	It attempts to solve the	This works faster but still with

A mixture of local and global alignments are performed to get multiple alignment [32]	disadvantage of Clustal series	minimum errors.
B. Morgentern, 2004 *Dialign* Dialign combines both global and local alignments leaving un related sequences [33]	Mixture of global and local alignments are performed on related sequences	It is too slow to align lengthy sequences.
C. Notredame and D. G. Higgins, 1996 *SAGA* Sequence Alignment by Genetic Algorithm [34]	Optimization is performed based on the objective function.	Involves higher processing time.

V. CONCLUSION

This research article summarizes the techniques used for sequence analyses. The pairwise alignment is used to find the similarities between any two given sequences. Multiple Sequence Alignment is used to generate a tree that describes the evolutionary relationship among sequences. This survey helps to understand the merits and demerits of the reported methods. This analysis helps to explore the possible research gaps that would lead to the development of newer ides for better sequence alignment of proteins in term of accuracy and time complexity.

REFERENCES

[1] M. Lehman, *"Experiments with Algorithms for DNA Sequence alignment"*, Computer Science, Simpson College, Indianola, Iowa 50125, pp.1-14, 2005.

[2] D. Fenstermacher, *"Introduction to Bioinformatics"*, Journal of American Society for information Science and Technology, Vol.56, pp.440-446, 2005.

[3] B. Alberts, A. Johnson, J. Lewis, *"From DNA to RNA"*, Molecular Biology", 4th Edition, 2002.

[4] J. Wu, J. Xiao, Z. Zhang, X. Wang, S. Hu, J. Yu, *"Ribogenomics: The Science and Knowledge of RNA"*, Genomics Proteomics Bioinformatics, Vol.22, pp. 57-63, 2014.

[5] R. Chen, H. Yan, K.N. Zhao, B. Martinac, G.B. LIU, *"Comprehensive Analysis of Prokaryotic Mechanosensation Genes: Their Characteristics in Codon Usage"*, DNA Sequence, pp. 1-10, 2006.

[6] G. Bertram, S. Innes, O. Minella, J. P. Richardson, L. Stansfield, *"Endless Possibilities: Translation Termination and Stop Codon Recognition"*, Microbiology, Vol.147, pp.255-269, 2001.

[7] T.J.J. Kelly, A.D. Lindeman, S.M. Bridges, *"Exploratory visual analysis of conserved domains on multiple sequence alignments"*, BMC Bioinformatics, Vol.10,pp. 1-18, 2009.

[8] B.H.A. Rehm, *"Bioinformatic Tool for DNA/ Protein Sequence Analysis, Functional Assignment of Genes and Protein Classification"*, Applied Microbiology Biotechnology, Vol.57, pp.579-592. 2001.

[9] C.B. Do and K. Katoh, *"Protein Multiple Sequence Alignment"*, Methods in Molecular Biology, Vol.484, pp. 379-413, 2009.

[10] S. Batzoglou, *"The Many Faces of Sequence Alignment",* Briefings in Bioinformatics, Vol.6, pp. 6-22, 2005.

[11] L. Mullan, *"Pairwise Sequence alignment"*, Briefings in Bioinformatics, Vol. 7, pp. 113-115, 2006.

[12] P. Manohar, S. Singh, *"Protein Sequence Alignment: A Review"*, World Applied Programming, Vol. 2, pp.141-145, 2012.

[13] Z.M. Zhou, Z.W. Chen, *"Dynamic Programming for Protein Sequence Alignment",* International Journal of Bio-Science and Technology, Vol.5, pp. 141-150, 2013.

[14] R. Mott, *"Smith-Waterman Algorithm"* University of Oxford, DOI: 10.1038/npg.els.0005263, pp. 1-5, 2005.

[15] E.S. Donkor, N.T.K.D. Dayie, T.K. Adiku, *"Bioinformatics With Basic Local Alignment Search Tool (BLAST) and Fast Alignment (FASTA)",* Journal of Bioinformatics and Sequence Analysis, Vol. 6, pp.1-6, 2014.

[16] T. Madden, *"The BLAST Sequence Analysis Tool",* National Center for Biotechnology Information, pp.1-10, 2013.

[17] T.A. Tatusova, T.L. Madden, *"BLAST2 Sequences, a New Tool For Comparing Protein And Nucleotide Sequences"*, FEMS Microbiology Letters, Vol. 174, pp. 1-6, 2006.

[18] N.A.A. Rashid, R. Abdullah, A.Z.H.Talib, Z.Ali, *"Fast Dynamic Programming Based Sequence Alignment"*, IEEE Transactions, pp.1-7,2006.
[19] S.R. Harris, K. Chinyere, Okoro, *"New Approaches to Prokaryotic Systematics",* Methods in Microbiology, 2014.
[20] K. Benkrid, Y.Liu, A. Benkrid, *"A Highly Parametrized and Efficient FPGA-Based Skeleton for Pair-wise Biological Sequence Alignment"*, IEEE Transactions on very large scale integration (VLSI) Systems, Vol. 17,pp. 561-570, 2009.
[21] N. Bray, I. Dubchak, L. Pachter, *"AVID: A Global Alignment Program",* Genome Research, Vol. 13, pp. 97-102, 2003.
[22] M. Brudno, C. B. Do, G, M, Cooper, "*LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA"*, Genome Research, pp. 721-731, 2015
[23] W.Huang, D.M. Umbach, L.Li, *"Accurate Anchoring of Divergent Sequences",* Bioinformatics, Vol. 22, pp. 29-34, 2006.
[24] D.J.Cho, Y.S.Han, H.Kim, *"Alignment with Non-overlapping Inversions on Two Strings",* International Workshop on Algorithms and Computation, pp. 261-272, 2014.
[25] A. Chakraborty, S. Bandyopadhyay, *"FOGSSA: Fast Optimal Global Sequence Alignment Algorithm",* Scientific Reports, pp. 1-7, 2013.
[26] T. Kahveci, V. Ramaswamy, H. Tao, *"Approximate Global Alignment of Sequences",* Bioinformatics and Bioengineering, pp. 1-8, 2005.
[27] Z.A. Fareed, H.M.O. Mokhtar, A. Ahmed, *"GPcodon Alignment: A Global Pairwise Codon Based Sequence Alignment Approach",*Internation Journal of Database Management System, Vol. 8, pp. 1-12, 2016.
[28] M. Sethi, S. Singh, *"Parallel Approach for Global Alignment",*Internation Journal of Interdisciplinary Research, Vol. 2, pp. 315-217, 2016.
[29] J.D. Thampson, D.G. Higgins, T.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, Vol. 22, pp. 4673-4680, 1994.
[30] D.F. Feng, R.F. Doolittle, *"Progressive Sequence Alignment as a Prerequisite to Correct Pylogenetic Trees",* Journal of Molecular Evolution, Vol. 25, pp. 351-360, 1987.
[31] R.Chenna, H. Sugawara, T.Koike, *"Multiple Sequence Alignment with Clustal Series of Programs",* Nucleic Acid Research, Vol. 31, pp.3497-3500, 2003.
[32] C. Notredame, D. G. Higgins, J. Heringa, *"T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment",* Journal of Molecular Biology, Vol. 302, pp. 205-217, September 2000.
[33] B. Morgentern, *"DIALIGN: Multiple DNA and Protein Sequence Alignment AtBibiserv"*, Nucleic Acid Research, Vol.32, pp. 33-36, July 2004.
[34] C. Notredame, D. G. Higgins,*"SAGA: Sequence Alignment by Genetic Algorithm",* Nucleic Acid Research, Vol.24, pp. 1515-1524,April 1996.

AUTHORS PROFILE

Dr. Shanmugavadivu Pichai is currently the Professor of Computer Science and Applications, at Gandhigram Rural Institute (Deemed to be University), and is involved in Teaching, Research, and Extension. Dr.Pichai has 27+ years of Academic experience, and has guided/guiding Research Scholars, and funded-research projects of UGC, DST and ICMR for an outlay of Rs.150.6Lakh. Her research areas includeMedical Image Analysis, Healthcare Analytics. Parallel Computing, Digital Image Processing and Content-Based Image Retrieval. She has conducted a national conference, training programmes, and workshops, and has delivered 80+ lectures as Keynote Speaker, Chief Guest, and Guest Lecturer. She has edited three volumes of research publications and authored about 100+ research publications. Recipient of Indo-US 21st Century Knowledge Initiative Award 2015. She had been on an international academic assignment in Malaysia and USA. Dr.Pichai holds a Master's degree in Computer Applications (REC, Trichy), Ph.D. in Digital Image Restoration (GRI) and MBA (IGNOU).

P.Haritha pursed Bachelor of Sciencefrom Parvathy's Arts and Science college, Dindigul, India in year 2014, Master of Computer Science and Applications from Gandhigram Rural Institute, India in year 2017. She is currently pursuing M.Phil. in Department of Computer Science and Applications, GandhigramRural Institute since 2017. Her main research focuses on Data Analysis.

Mr. S. Dhamodharan pursed Bachelor of Science from Government Arts College, Melur, India in 2010 and Master of Computer Applications from Thanthai Periyar Government Institute of Technology, Vellore, India in year 2013. He is currently pursuing Ph.D and working as Project Fellow in Department of Computer Science and Applications, The Gandhigram Rural Institute (Deemed to be University), India since 2017. His main research work focuses on Digital Image Proessing.

