IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 11, 2019, accepted April 13, 2019, date of publication May 10, 2019, date of current version June 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916154

Current Trends and Ongoing Progress in the Computational Alignment of Biological Sequences

MUHAMMAD ISHAQ[1], (Fellow, IEEE), ASFANDYAR KHAN[1], (Member, IEEE), MAJID KHAN[2], AND MUHAMMAD IMRAN[1]

[1]Department of Computer and IT, The University of Agriculture, Peshawar 25000, Pakistan
[2]Department of Statistics, Math and Computer Science, The University of Agriculture, Peshawar 25000, Pakistan

Corresponding author: Muhammad Ishaq (ishaqafridipk@gmail.com)

ABSTRACT The computational techniques for nucleic acid and protein sequence comparison reduce the extensive burden of molecular biologists. The sequence alignment is one of the main research areas in bioinformatics, and comparative genomics and proteomics lead us to important discoveries in various fields of bioinformatics. Researchers develop and use different heuristics and evolutionary algorithms for optimal DNA and protein sequence alignment. There are different categories of improved computational sequence matching methods. In this paper, the goal is to cover almost all computational approaches toward sequence alignment. Different aspects and issues related to the optimal alignment of biological sequences will be analyzed. The sequence comparisons through mathematical and computational techniques have manifold benefits and importance in bioinformatics. Researchers recently explore proposing novel computational techniques for simultaneous matching of multiple sequences or multiple sequence alignment (MSA). Pairwise alignment, or the alignment of two sequences, is the basic building block of all alignment methods. The goal is to design optimal and relevant algorithms with less computational complexity and more efficiency.

INDEX TERMS Dynamic programming optimization, multiple sequence alignment, evolutionary computations, categories of sequence alignment, recent trends in sequence alignment, local and global alignment, pairwise alignment, structure based alignment.

I. INTRODUCTION

Nucleic acids are responsible for determining the structure and function of the living kingdom. The main constituent organic compound or component of all living organisms is protein. Genes are encoded by a specific pattern of DNA nucleotides; there is a specific nucleotide pattern for each amino acid which is also called as codons. The DNA codon sequence is then transcribed to messenger RNA (mRNA) and is transferred to ribosomal RNA (rRNA). The ribosome is comprised of rRNA that synthesizes proteins according to codon pattern of the mRNA with the help of tRNA (transfer RNA). Bioinformatics comprised of computer science tools and applications to efficiently handle important biological data. It can be used for protein data sorting, manipulation, and arrangement of nucleic acids.

Nucleic acids are made up of nucleotides, and a protein molecule is a poly peptide chain or combination of amino acids. Nucleotides in a specific order are referred to as a sequence. Sequence comparison reveals important information, such as relatedness between organisms. Sequence matching has historically been carried out in biochemical laboratories by molecular biologists. However, such kind of alignment is slow and practically impossible for enormous amount of data. Computational algorithms exploit biological comparisons and estimate the relatedness between sequences through alignment score.

The term biological sequence applies to nucleic acid and protein molecules with nucleotides and amino acids, respectively, in a properly ordered format, as shown in the figure three. Matching or comparison of nucleic acid and protein sequences point-by-point or cell-by-cell to find possible similarities and relationships among the sequences is called as sequence alignment. DNA sequences are comprised of

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang.

nucleotides and protein sequence is made up of the polypeptide chain of amino acids. If we take two proteins or DNA sequences as a row of symbols and each nucleotide and amino acid is represented by specific special symbols like thymine by 'T' and Leucine amino acid by 'L'. Comparing two rows of sequences, nucleotide by nucleotide or amino acid by amino acid, is a process called sequence alignment, as shown in the figure three. This is referred to as word or string based matching, as the residue nucleotides are represented by their respective symbols. Actual computational alignment of two DNA sequences in MATLAB workspace is shown in the figures 4.

In this diagrammatic illustration matches are represented by a vertical line from one nucleotide in one sequence to another nucleotide in the second sequence. Mismatch is represented by no line and gap in each sequence is represented by a dash in a sequence. The combine alignment score is the combination of all matches minus mismatch scores. Gap penalty is also applied. Usually gap penalty is set to zero for simplicity in any novel algorithm.

Sequence alignment gives us the measure of relatedness between nucleotides and amino acid sequences. By determining the relatedness between two sequences we are able to find out structural, functional and evolutionary relationships [1]. All computational alignment approaches are statistical probabilistic calculation of matches, mismatches and gaps. Some of them are mathematical assumptions and require pre-requisite information. The correct estimation of match, mismatch and gap determine the quality of computational algorithms.

II. RELATED WORK

Sum-of-pairs (SP) or Column score (CS) represent the quality of alignment. SPdist is a novel approach and is used to measure the sequence distance between mismatched residues in the query alignment. This technique gives better results than SP especially in terms of divergent reference alignments [2].

Through sequence comparison we find similarity. Similarity is the quantitative measure between two sequences. Choose two sequences, Select an algorithm that generates a score, Allow gaps (insertions, deletions). Match, mismatch and gap got specific assigned score in each algorithm. The combine score reflects the degree of similarity of two sequences. There are two kinds of sequence alignment tasks, e.g. pairwise sequence alignment and the more computationally complex task of multiple sequence alignment (MSA).The alignment of two sequences at a time is called as pairwise alignment and the matching of three or more sequences at a time is termed as multiple sequence alignment.

There is another kind of sequence alignment called as triplet alignment. Three sequences are aligned at a time in triplet comparison. Figure 1 show us the types and methods of sequence alignment. All novel computational techniques for sequence matching have some sort of relationship with these methods. Calculation of lower and upper bounds for optimal alignment of binary sequences can be carried out through various mathematical methods [3].

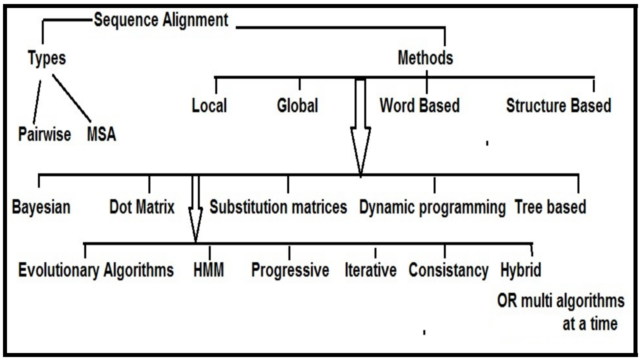
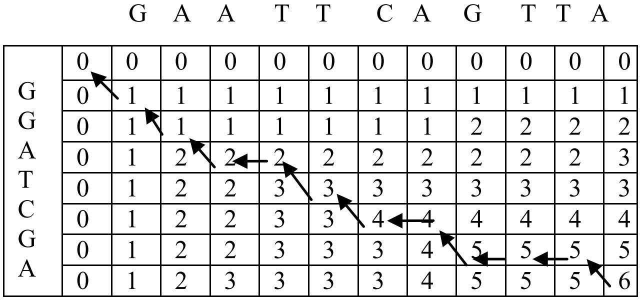


FIGURE 1. Different kinds of Sequence alignments methods and types.

Several methods are used for sequence alignment and analysis. Alignment free method is based on mapping symbolic sequences describing DNA, RNA and proteins onto vector spaces, in which many of the analysis can be performed more efficiently. In this method the rational is to represent sequences as numerical real valued vectors. The available tools like filtering techniques, normalization, dissimilarity estimation and clustering are applied to this domain. Probability, statistics and linear algebra are then at hand to provide a strong and extensive theoretical and computational background. It has the ability of effortlessly dealing with whole genomes, thus allowing the analysis of complete sequence information [4]. Traditional analysis of MSA is through single strand of multiple alignments. Directed acyclic graph (DAG) is used for a set of sampled alignment. Each node is an alignment column and each path through this DAG is a valid alignment. This approach provides a natural space distribution of MSA's to make the existing algorithm for alignment more efficient [58].

SNAPR algorithm is applied for efficient and accurate RNA-Seq alignment and analysis. This is an easy algorithm to convert raw RNA sequences to interpretable results. A hash table technique is used to utilize the processing and memory of high power machines. FASTQ and BAM file format are used as an input and the output is a sorted BAM file. The algorithm can read individual read counts and identify exogenous RNA species and gene fusion. SNAPR is used for future sequencing with longer reads [6]. Recently we experience many novel techniques for whole genome comparison. Some researchers propose compression models for MSA blocks [62]. One of such method is based on a mixture of finite-context models that address the problem of DNA bases and gap symbols together. This method further explores the correlation between sequences [7].

Comparison of complete nucleic acid or protein sequences is called as global alignment, and partial sequence alignment is termed as local alignment. Local alignment is useful to find out sub optimal matching locus between two sequences [8]. Traditionally a variety of computational techniques are available for sequence alignment. For global alignment Needleman-wunsch [9] algorithm is used. Smith-waterman [10] algorithm is used in cases of local alignment.



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	3	4	5	5	5	5
A	0	1	2	3	3	3	3	4	5	5	5	6

FIGURE 2. Scoring of residue in a matrix (two dimensional array b) and trace backing to obtain an alignment of GAATTCAGTTA and GGATCGA.

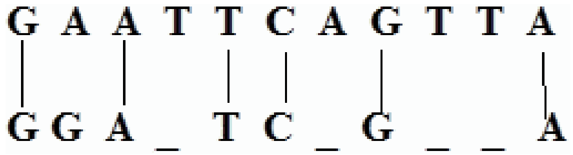


FIGURE 3. Diagrammatic illustration of Pairwise alignment through dynamic programming.

In both cases the longest common subsequence is obtained through dynamic programming.

III. ALIGNING THROUGH EXACT METHOD AND LATEST INPROVEMENTS

Dynamic programming is the exact method for computational sequence comparison. The application of dynamic programming requires certain conditions.

The sub-problems or matching of each cell residue are interrelated. Interrelations of sub-problems also mean that all pairwise alignments are linked through a tree in the process of progressive multiple sequence alignment.

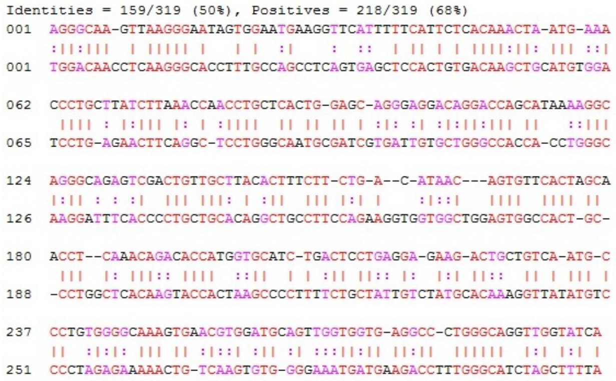
Optimal structure can be characterized as explained above for multiple matrices or multiple sequence alignment. The extent of similarity and non-relatedness can be defined. The characterize structure can be defined recursively like adding more sequences (repeat the same process) to the already aligned profile during MSA.

Solution for base cases has a termination criterion and if the construction of optimal solution is possible then dynamic programming is a better choice.

Dynamic programming consist of three steps, first initialization take place, then scoring of matrix and then alignment of two sequences through trace back mechanism. The matrix for sequence alignment through dynamic programming is shown in the figure 2 and the resultant pairwise alignment is shown in the figure 3.

IV. MERITS AND ISSUES IN ALIGNMENT

The function of protein is predicted by computational approaches. If the crystal structure of one protein is available then the 3D model of its unknown homologous in any database of proteins is retrieved through computational matching tools. Proteins binding regions can be identified through careful interpretation of multiple sequence alignment [1]. Comparison of two sequences point by point or cell by cell to discover relatedness or to find out common ancestry and functional similarity. Some techniques gave good results but there is always a chance of further improvement. Some methods predict the structure and function of proteins and the relationship between proteins and gene is traced out [1].



```
Identities = 159/319 (50%), Positives = 218/319 (68%)
001  AGGGCAA-GTTAAGGGAATAGTGGAATGAAGGTTCATTTTCATTCTCACAAACTA-ATG-AAA
     :||:|||  | ||||| |   | |  :|     :  ::  | || | | ||||:||: ||| ::|
001  TGGACAACCTCAAGGGCACCTTTGCCAGCCTCAGTGAGCTCCACTGTGACAAGCTGCATGTGGA

062  CCCTGCTTATCTTAAACCAACCTGCTCACTG-GAGC-AGGGAGGACAGGACCAGCATAAAAGGC
      ||||  :  |:||| |: | :||||  || || || | :|: :| :|:||:||| ||    ::|||
065  TCCTG-AGAACTTCAGGC-TCCTGGGCAATGCGATCGTGATTGTGCTGGGCCACCA-CCTGGGC

124  AGGGCAGAGTCGACTGTTGCTTACACTTTCTT-CTG-A--C-ATAAC---AGTGTTCACTAGCA
     |:|| : : :|  ||| |||: | :||  ||| |:| |     :|::|   ||||  |||| ||
126  AAGGATTTCACCCCTGCTGCACAGGCTGCCTTCCAGAAGGTGGTGGCTGGAGTGGCCACT-GC-

180  ACCT--CAAACAGACACCATGGTGCATC-TGACTCCTGAGGA-GAAG-ACTGCTGTCA-ATG-C
      |||  |: |||:: ||||  ::||  | | :|| ||:: |:  |:| ||::  || | ||| |
188  -CCTGGCTCACAAGTACCACTAAGCCCCTTTTCTGCTATTGTCTATGCACAAAGGTTATATGTC

237  CCTGTGGGGCAAAGTGAACGTGGATGCAGTTGGTGGTG-AGGCC-CTGGGCAGGTTGGTATCA
     ||  :|:|: ||| || :|::| :|| :| :::||:|| ||:||  ||||||  |:| |:| |
251  CCCTAGAGAAAAACTG-TCAAGTGTG-GGGAAATGATGAAGACCTTTGGGCATCTAGCTTTA
```

FIGURE 4. Actual alignment of two sequences.

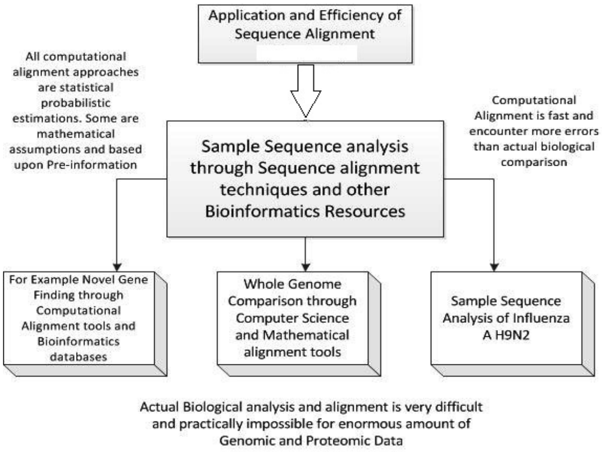

Application and Efficiency of Sequence Alignment
All computational alignment approaches are statistical probabilistic estimations. Some are mathematical assumptions and based upon Pre-information
Computational Alignment is fast and encounter more errors than actual biological comparison
Sample Sequence analysis through Sequence alignment techniques and other Bioinformatics Resources
For Example Novel Gene Finding through Computational Alignment tools and Bioinformatics databases
Whole Genome Comparison through Computer Science and Mathematical alignment tools
Sample Sequence Analysis of Influenza A H9N2
Actual Biological analysis and alignment is very difficult and practically impossible for enormous amount of Genomic and Proteomic Data

FIGURE 5. Merits and issues of computational sequence alignment tools and techniques.

The issues and problems related with sequence alignment are shown in the figure 5. Proper sequence alignment is necessary to reveal the structure and function of organism genome. Through proper alignment we can find relationship, linkage and interaction between nucleic acid and proteins. The phylogenetic origin of species is traced out. Sequence alignment is also helpful in forensic and medical sciences. Phylogenetic means the evolutionary origin or the history of organism lineage as they change through different times. The phylogenetic tree can be obtained from filtered and non-filtered MSA. Recent research investigation reveal that the tree generated from filtered MSA is worse than the tree obtained from unfiltered MSA [11]. The identification of non-coding DNA or RNA regions in the complete genome alignment is very important. Secondary structure of proteins and nucleotides is more stable and conserved. So the identification of non-coding regions through stable and conserved secondary structure is more reliable [12]. Outlier detection in MSA is another interesting area of research. In a given multiple sequence alignment an outlier sequence create many problems. OD-seq detects outlier through the examination of average distance of individual sequence from other sequences. The sensitivity and specificity of OD-seq is very high [13]. DECIPHER resolve the accuracy of alignment in cases of large number of sequences. DECIPHER R package is available from Bioconductor repository [14]. Figure 5 summarized the application and efficiency of Sequence Alignment.

V. RECENT TRENDS IN SEQUENCE MATCHING

In recent year we observe a lot of novel approaches toward sequence comparison. Cynthia Vinzant defines new lower bound for optimal alignments of binary sequences [3]. Alignment free comparison of genome sequences by numerical characterization is another novel method. Alignment free comparison of sequences was performed by computing the distances between vectors of the corresponding numerical characterization, which define the evolutionary relationship [53].

This is a powerful tool for genome comparison [15]. Alignment free method is also used for phylogenetic inference. D2 alignment free approach is usually used for the same purpose [16]. Protein MSA through permutation similarity is proposed to evaluate several algorithms. As the permutation similarity method only concerns with the relative order of different protein evolutionary distances, without taking into account the slight difference between the evolutionary distances, it can get more robust evaluation [17].

Analysis of MSA through correlation method finds out residue-positions whose occupation with amino acids change in a concerted manner. The position of specific residue is of vital importance in many areas of research. The residue position is important for protein function or stability [5].

The application of parallel processing in multiple sequence alignment is another catching area. There are a lot of parallel strategies for word based protein MSA. These methods can be extended to the structure based protein MSA. Waterman and Smith [10] is a famous algorithm for local alignments but at the expense of very high computing power and huge memory requirements. In one approach Smith-Waterman algorithm for local alignment run in a cluster of workstations using a distributed shared memory system [18]. This approach can be used for structure MSA. In cases of protein structure based multiple sequence alignment several searches and structure matching are involved which require a lot of time and processing speed. The use of a parallel computer cluster or Grid can reduce processing time and obtain an optimal result in less time. It may help us to quickly predict the function of proteins from its structure. A number of algebraic operation theories for linear and context-free grammars make possible to combine atomic and complex multidimensional grammars in cases of complex alignment problems [6].

According to one research method inside a structure based multiple sequence alignment; remote homologues proteins improve sequence alignment by extracting structural information from profiles of multiple structure alignment. A systematic search algorithm combined with a group of score functions based on sequence and structural information has been introduced in one procedure [4].

Structurally informed alignment has a manifold benefits and it can be useful in the phylogenetic analysis of biogenic amine receptors in vertebrates. A comprehensive high quality alignment is constructed to facilitate the biogenic amine receptors study [5].

Roy D. Sleater et al proposed MSA algorithm to be run in a parallelized fashion with the sequence data distributed over a computer cluster or server farm. The cloud computing technology improves the speed, quality and capability of MSA. They introduce next generation of cloud based MSA algorithm [19]. Some researchers evaluate the performance of parallel multiple sequence alignment on supercomputer like BlueGene/Q or JUQUEEN. A parallel I/O interface for simultaneous and independent access to single file collectively has been designed and verified [20]. David diaz and his co-researchers developed MC64-ClustalWP2 as a new implementation of Clustal W algorithm, integrating a novel parallelization strategy that significantly increases the performance when aligning long sequences in architecture with many cores. They analyze the software and hardware features in order to exploit and optimize the full potential of parallelism in many-core CPU systems.

To test the performance of their proposed algorithm they use hybrid computing system. MC64- Clustal WP2 has many fold benefits [21]. To improve the scalability of global sequence alignment an MPI based parallelization technique is proposed. In this method a parallel waveform algorithm based on a chunk size transformation to handle large datasets with message passing model exposes high speed up and scales linearly with the increasing number of processes [22]. Some researchers examine different multi-core machines by running a variety of MSA software [23]. In recent years we observe various kinds of novel techniques for parallel MSA like artificial bee colony optimization [24].

Structure based alignment is more comprehensive, conserved and informative. Conserved regions are much stable and predict correct function and structure of a given protein. The extent to which two structures align is to measure the root mean square deviation (RMSD) [25].

VI. MAOR CATEGORIES OF SEQUENCE ALIGNMENT

Computational alignment techniques are summarized as under.

A. PAIRWISE LOCAL AND GLOBAL SEQUENCE ALIGNMENT

Comparison of two sequences at a time is termed as pairwise sequence alignment. BLAST [26] and FASTA [27] are famous programs for pairwise alignment. BLAST [28] is used to compare a sequence with the entire online database, so it is an important discovery method of potential homologs. The inference of homology or common ancestry of organisms

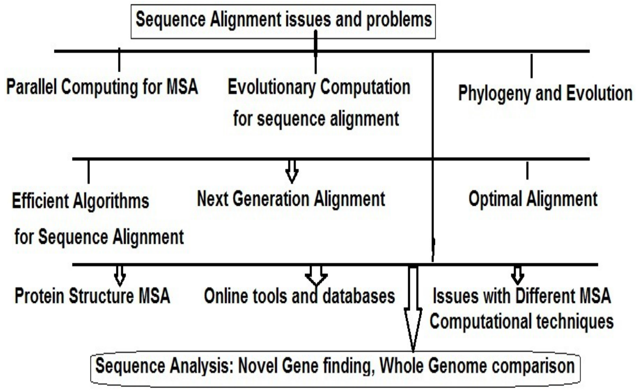

Sequence Alignment issues and problems
Parallel Computing for MSA
Evolutionary Computation for sequence alignment
Phylogeny and Evolution
Efficient Algorithms for Sequence Alignment
Next Generation Alignment
Optimal Alignment
Protein Structure MSA
Online tools and databases
Issues with Different MSA Computational techniques
Sequence Analysis: Novel Gene finding, Whole Genome comparison

FIGURE 6. Issues related with computational sequence matching.

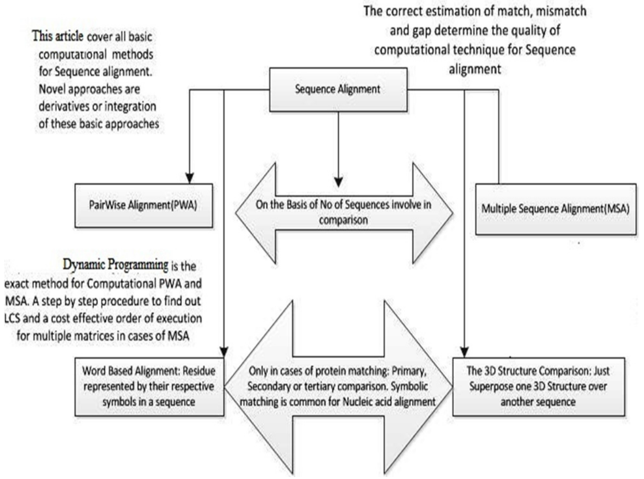

This article cover all basic computational methods for Sequence alignment. Novel approaches are derivatives or integration of these basic approaches
The correct estimation of match, mismatch and gap determine the quality of computational technique for Sequence alignment
Sequence Alignment
PairWise Alignment(PWA)
On the Basis of No of Sequences involve in comparison
Multiple Sequence Alignment(MSA)
Dynamic Programming is the exact method for Computational PWA and MSA. A step by step procedure to find out LCS and a cost effective order of execution for multiple matrices in cases of MSA
Word Based Alignment: Residue represented by their respective symbols in a sequence
Only in cases of protein matching: Primary, Secondary or tertiary comparison. Symbolic matching is common for Nucleic acid alignment
The 3D Structure Comparison: Just Superpose one 3D Structure over another sequence

FIGURE 7. Classification of computational sequence alignment techniques according to my understanding.

is satisfied when two sequences have sufficient similarity. Homology refers to the relationship between genes separated by the event of speciation (orthology) or the relationship between genes separated by the events of genetic duplication (paralogy). Generally protein sequences with 20-25% similarity are classified in the twilight zone. Some pairwise alignment methods are specialized for below twilight zone comparisons [29]. Figure 7 is the diagrammatic illustration of different categories of alignment techniques.

Coding for one-to-many multiple sequence alignment is another hot research topic nowadays. A code that take an input set of pairwise alignments and generates a one-to-many gapped multiple sequence alignment. CombAlign code is demonstrated by generating gapped multiple sequence alignment from structure based pairwise alignments. The multiple sequence or structure based alignment (MSSA) show individual residue-residue relationship, which enable the identification of similar and different regions between the alignment proteins [30].

Homology is the conclusion that we reach after measuring the similarity or percent identity between sequences. Homology cannot be measured in degrees and we can use the term partial homology in cases of partial similarity.

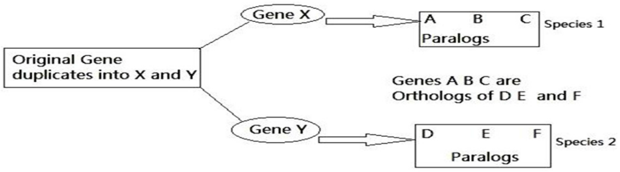

Gene X
A B C
Paralogs
Species 1
Original Gene duplicates into X and Y
Genes A B C are Orthologs of D E and F
Gene Y
D E F
Paralogs
Species 2

FIGURE 8. Diagram showing homology, orthology, and paralogy.

Homology implies the evolutionary relationship between sequences. Homology mean the percentage of similarity between sequences and homologs mean same or nearly same sequences in the same or different organisms. Homologs belong to the same species are called as paralogs and orthologues are those homologs that are found in different species of organisms. The figure 8 shows homology, orthology and paralogy.

Similarity of two sequences is measured through empirical mathematical scoring matrices. There are two famous substitution matrices called as PAM and BLOSUM [31]. The evolutionary history of two sequences is revealed by homology through phylogenetic analysis [32]. Pairwise alignment is either local or global as discussed above. Phylogenetic analysis is carried out through various tools like MEGA [33] or MATLAB [34]. Characterization of errors in pairwise and multiple sequence alignment is a difficult task [23].

Partial sequence alignment or the comparison of most similar parts of two sequences is termed as local alignment. Local alignment finds optimally matching regions within two sequences. Smith-waterman [10] procedure is used for local alignment. Actually smith and waterman work on a complete mathematical analysis of RNA secondary structure [10]. RNA has a single helical structure.

Smith waterman or local alignment find and align most optimal or closely related fragments between two sequences. There are very few gaps in local alignment and long gaps are ignored.

Smith waterman algorithm is slow in processing and there are several parallel and high performance computing techniques to resolve this issue. The high computational power of NVIDIA-based general purpose graphic processing units (GPGPUs) can be accessed through PaSWAS alignment software. PaSWAS is a Smith waterman parallel implementation that runs on graphical hardware with improved performance. Score, number of gaps and mismatches can retrieve through this tool and the software reports multiple hits per alignment [2]. The test cases show the usability and versatility of this parallel alignment software utility [35].

Comparison of complete length-wise nucleic acid or protein sequences is called as global alignment. Global alignment extends from one end of a sequence to the other.

Global alignment gives us complete details of comparison between two sequences.

Recently we observe novel and modified global alignment techniques like normalized global alignment for protein sequences [36]. In order to reduce the length and composition dependence of global alignment scores, Z-score is computed with Monte-Carlo algorithm. This technique requires a great number of sequence alignments, leading to high computational cost. In this method a normalized global alignment score is introduced in order to correct the length dependencies of global alignment. This algorithm is based on fractional programming and defines a best ratio of alignment score and length [36]. Some algorithms for global alignment are a combination of many strategies. One algorithm proposes by Qi *et al.* [37] combine simple alignment algorithm with extension algorithm for largest common substring and graphical simple alignment tree (GSA). The GSA tree solves the problem of global alignment of two DNA sequences [37]. Progressive MSA need a guide tree to align sequences according to the topology of tree. The adaptive method of constructing guide tree is another quality approach and it mainly improves the accuracy of different progressive MSA tools [38]. Some researchers systematically explore the performance of different guide trees currently used for multiple sequence alignment [59]. In one effort the researchers claim that the pairwise distance based default guide tree performance is better than evolutionary guide tree in cases of structure derived reference alignment. The results of pairwise distance based default guide tree are still not optimal but even better than the average chained guide tree [39].

Through Needleman-wunsch method it is possible to determine whether significant homology exists between the proteins. This information is used to trace their possible evolutionary development. The largest number of amino acids of one protein that can be matched with those of a second protein allowing for all possible interruptions in either of the sequences. Comparisons are made from the smallest unit of significance, a pair of amino acids, one from each protein. All possible pairs are represented by a two-dimensional array, and all possible comparisons are represented by pathway through the array. A numerical value is assigned to each cell in the array and the maximum match is the largest number that would result from summing the cell values of every pathway [9].

B. MULTIPLE SEQUENCE ALIGNMENT

Comparison of more than two sequences at a time is called as multiple sequence alignment (MSA). In the process of multiple sequence alignment with any algorithm or method first pairwise alignment takes place. So pairwise alignment is the basic building block for multiple sequence alignment. Multiple sequence alignment is very useful in cases of evolutionary relationship between sequences. It is helpful to find out similarity and relationship between homolog sequences and the discovery of special motifs in a sequence.

For large number of sequences progressive alignment is a standard method. In this approach we experience the tradeoff between alignment accuracy and computational time. In one research finding the loss of information in the early steps result in an unstable final alignment. If the order of sequences is reversed in the input file then a totally new alignment is generated. This MSA technique is instable for large number of sequences. The researchers also determine the number of sequences for which the probability of instability is more noticeable [38].

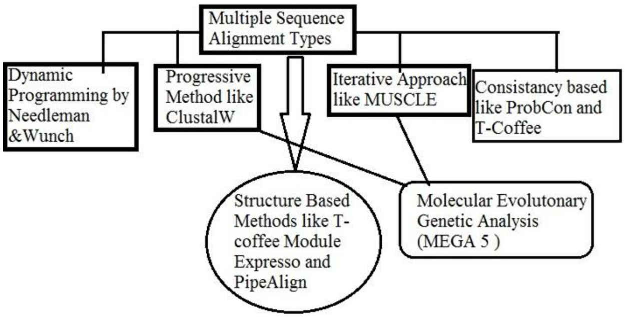

Multiple Sequence Alignment Types
Dynamic Programming by Needleman &Wunch
Progressive Method like ClustalW
Iterative Approach like MUSCLE
Consistancy based like ProbCon and T-Coffee
Structure Based Methods like T-coffee Module Expresso and PipeAlign
Molecular Evolutonary Genetic Analysis (MEGA 5)

FIGURE 9. Different kinds of multiple sequence alignment and related tools (personal perceptive).

FASMA [24] is a service to format and analyze sequences during the process of multiple sequence alignment. MSA reveal conserved secondary structures features and indels (insertion and deletion in evolutionary mutation). There are some useful web applications that extract indel regions and conserved blocks from protein multiple sequence alignment [40]. During meiosis or reduction division mutation take place and insertion and deletion of nucleotides and amino acids is common. Indels refer to the inserted or deleted sequence parts. Multiple sequence alignment is further divided into five basic types.

Multiple sequence alignment identifies conserved regions, patterns and domains. Conserved regions are required for structure and function and can carry out limited changes without affecting the structure and function of the given protein. Mutation is very frequent in non-conserved regions of proteins. Through Multiple sequences alignment Phylogenetic analysis is easy and it also generates position specific scoring matrices for subsequent searches. Large scale multiple sequence alignment with simultaneous phylogenetic inference is possible through parallel computing [41].

Protein alignment is more informative and accurate than nucleic acid alignment. Reasonable number of sequences is preferred. Selection of sequences with moderate similarity index gives clear results. It means a collection of not too similar and not too different sequences for multiple sequence alignment. Different kinds of multiple sequence alignment methods and related tools are shown in the figure 9.

C. WORD AND STRUCTURE BASED SEQUENCE ALIGNMENT

The term Word-based is used when the amino acids and nucleotides are represented in a sequence by their respective symbols, like nucleotide guanine by G and amino acid lysine by K. FASTA or Fast alignment [27] and BLAST [26] are word based methods. BLAST stand for Basic local alignment search tool. BLAST seeks high-scoring segment pairs (HSP). HSP are pairs of sequences that can be aligned with each other and got maximum aggregate score. The obtained score cannot be improved by extension and score must be above a certain threshold.

The alignment obtained through BLAST is either gapped or un-gapped. Direct approximate alignments through BLAST optimize a measure of local similarity. BLAST is simple and robust and it can be implemented in a number of ways and applied in a number of contexts. In addition to its flexibility and tractability to mathematical analysis, BLAST is faster than sequence comparison tools of comparable sensitivity [28].

BLAST tends to rely on amino acid distribution frequency and sometime result to false positives. BLAST for nucleotide query to retrieve nucleotide result is called as BLASTN. Another version of BLASTN called MegaBLAST is used to align very long and highly similar sequences and good for batch nucleotide searches. MegaBLAST is faster than BLASTN and is used for eukaryote organelles, whole chromosomes alignment or small organism genome searches.

BLAT or BLAST like alignment tool is similar in function to MegaBLAST. BLAT find out genome coordinates and determine gene structure of an unknown gene or sequence. BLAT also modifies markers of interests in the vicinity of a sequence. It display separate tracks of sequences and identify gene family members.

A comprehensive survey of web based multiple sequence alignment tools is conducted by Ken D.NguYen et al that cover all existing web based MSA techniques [55]. The author also amazed and identify discrepancies in these methods. A web based 'SeqAna' model is proposed that comprehensively cover all missing needs in this area [32].

In cases of local protein multiple sequence alignment the transitive consistency score (TCS) web server measure the local reliability and find out analogous residues positions. TCS scoring scheme for structural superposition and phylogenetic reconstruction is more accurate than other related methods [32].

VII. PROMINENT EFFORTS FOR SEQUENCE ALIGNMENT

A. DOT MATRIX SEQUENCE ALIGNMENT METHOD AND SUBSTITUTION MATRICES

In dot matrix [42] method a matrix store intermediate results and sequences are plotted on a graph. Each intersection point or dot on a graph represents a matching pair or similarity and differences between sequences.

In substitution matrices each intersection or matching pair got a specific score [43]. Dot matrix analysis is dynamic but substitution matrices are static. There are two famous types of matrices, the percent accepted mutation (PAM) [31] and the blocks substitution matrix (BLOSUM) [44]. PAM is usually used for global alignment of closely related proteins and BLOSUM is used for local alignment of distantly related proteins. Substitution matrices give accurate measure of similarity between two sequences. Some researchers prefer BLOSUM over PAM and there is no single matrix that should be a complete enough for all sequence comparisons.

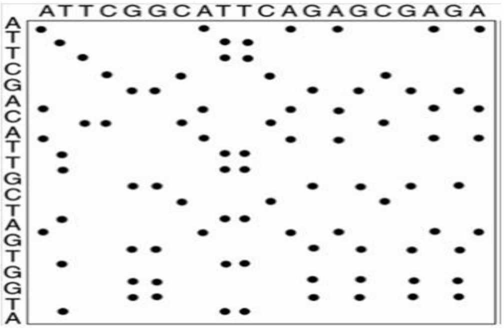


FIGURE 10. Dot matrix pairwise alignment.

Substitution matrices are actually pre-computed tables of numbers representing all possible transitions states for nucleotides and amino acids. Scoring matrices give us empirical weighting scheme representing physiochemical and biological characteristics of nucleotides and amino acids. This reveal side chain structure, function and chemistry. Protein substitution matrices affect amino acid classification [45].

There are two considerations in dealing with scoring matrices. The first is conservation; it means that what kind of residue substitution is possible that could not affect the function of protein. The second is the frequency of a residue amino acid or nucleotide within a sequence when deriving the scoring matrices for a given alignment.

Gap represent biological event, either insertion or deletion. If there is one gap per twenty residues then this is good. The Gap penalty can be deducted by the following equation 1.1 and Dot-matrix pairwise comparison is shown in the figure 10.

$$\begin{aligned} \text{Deduction for a gap} &= \mathrm{G} + \mathrm{Ln} \\ \text{where } \mathrm{G} &= \text{gap openingpenality} \\ \mathrm{L} &= \text{gap extensionpenality} \\ \mathrm{n} &= \text{length of the gap} \\ \text{and } \mathrm{G} &> \mathrm{L} \end{aligned}$$

Equation 1.1: Gap Penalty Deduction

B. BAYESIAN METHOD FOR SEQUENCE ALIGNMENT

This method is rarely used for pairwise alignment and is used to measure the evolutionary distance between DNA sequences. This method involves the probabilities of all aligned sequences in a profile, their gap scores and substitution matrix value to assess the probability of the next alignment.

No need to specify all parameters in Bayesian method. It describes the exact uncertainty and derives significant measure. This method assesses the probability of the alignment and there is no need of substitution matrix or gap scoring [46].

VIII. EVOLUTIONARY OR GENETIC ALGORITHMS FOR SEQUENCE ALIGNMENT

To trace out the evolutionary origin of sequences we have two steps process. First the comparison of multiple sequences

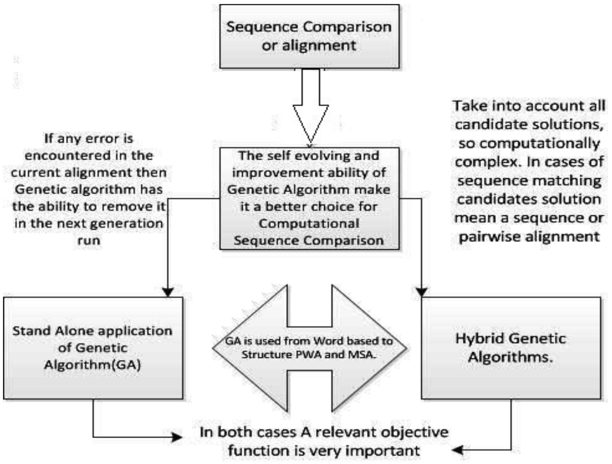

Sequence Comparison or alignment
The self evolving and improvement ability of Genetic Algorithm make it a better choice for Computational Sequence Comparison
If any error is encountered in the current alignment then Genetic algorithm has the ability to remove it in the next generation run
Take into account all candidate solutions, so computationally complex. In cases of sequence matching candidates solution mean a sequence or pairwise alignment
Stand Alone application of Genetic Algorithm(GA)
GA is used from Word based to Structure PWA and MSA.
Hybrid Genetic Algorithms.
In both cases A relevant objective function is very important

FIGURE 11. My personal understanding the types of genetic algorithm for sequence alignment.

and then to build the evolutionary or phylogenetic tree of all sequences in the given MSA. The inference of evolutionary parameters is drawn from maximum likelihood or Bayesian method. The changes in sequence is mediated by probabilistic substitution models. Some researcher's investigate the statistical properties of the above mentioned methods used in building the phylogenetic tree.

Simulation study of sequence divergence inference and phylogenetic tree were conducted. Such kind of analysis shows that nucleotide and amino acids are negatively affected by using inaccurate and overfitting guide tree. The effect is more pronounced for alignments involving more sequences. Amino acid results are more robust than nucleotide in cases of no inference strategy [47]. Such kind of sequence alignment algorithms based upon the concept of natural organic evolution and hence we name it evolutionary algorithms. The different categories of genetic algorithms are summarized in the figure 11.

First, initialize a population of sequences or alignments, then the selection of the fittest (optimal) candidate take place. Parents give birth to offspring's and variations occur in the population. Optimization of crossover, mutation and migration is different in each evolutionary algorithm. Child population is raised on the basis of candidate fitness. Evolutionary algorithms are applied in cases of multiple sequence alignment. The function of an evolutionary algorithm is to pick the next optimal pairwise alignment that has to be aligned with the already aligned profile.

During multiple sequence alignment three or more than three aligned sequences are sometime termed as a profile. Protein sequence profiles can be used to predict reliable aligned regions [48], [60]. Objective or fitness function is used to judge the fitness of solution. The fitness values determined in the objective function show us the evolutionary relationship and structural information of the aligned sequences. For better results more improvement and further refinement is required. Some famous sequence alignment algorithms that are based upon the evolutionary computational model are discussed below.

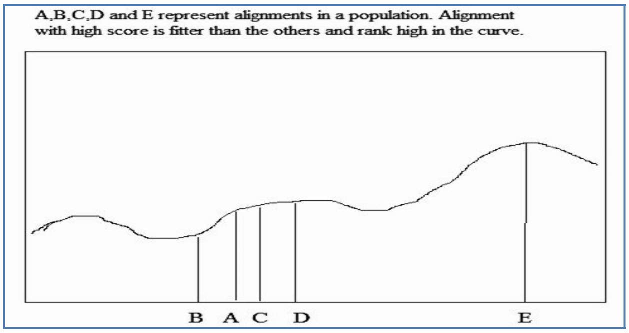


FIGURE 12. Graphical representation of hill climbing algorithm (according to personal perception).

A. HILL CLIMBING ALGORITHM FOR SEQUENCE ALIGNMENT

To search a fittest alignment inside a population we apply local search as in hill climbing algorithms. In this approach the current solution is jump to its neighbor solution, if the neighbor is fitter than the previous one then select it as a new solution and if it is worse than the previous then return to the previous solution. Selection through this algorithm is graphically shown by a curve like a hill, so this is called as hill climbing approach.). If A is the current solution then B is rejected because B is less fit than A, and we jump to solution C, which is fitter than A. The search is limited to neighborhood or only local search for fittest is possible through this approach. Solutions here mean a pairwise alignment of sequences. If there is no proper solution in the neighborhood then this algorithms stagnate at local maximum. Graphical representation of hill climbing algorithm is shown in the figure 12. Neighboring joining procedure is proposed by Hoksza and Svozil [56] in cases of 3D nucleic acid structural alignment. SETTER is a fast and accurate technique for RNA pairwise structure comparison. MutliSETTER is an extension of SETTER and is used in cases of multiple RNA structure alignment. SETTER decomposes RNA structure into non-overlapping structural parts. MutliSETTER combine the function of SETTER with ClustalW. ClustalW in this case is used for RNA structural superposition or alignment [48].

B. SIMULATED ANNEALING ALGORITHM FOR SEQUENCE ALIGNMENT

In this method a temperature parameter is initialized with population initialization. A non-local mutant can be accepted as a current solution, if the mutant is fitter than the current candidate. Simulated annealing resolve the issue of local solution stagnation. Migration of candidate solutions in a population is a continuous process and if the worse solution is temporarily moved then local search can find a fittest candidate solution in a population.

After variations in sequences and alignments, new uniform child population of fittest offspring is generated.

The candidates in a population are unique. Alignments are selected gradually base upon their similarity score. To prevent disruption caused by crossover and mutation functions, a profile or subpopulation based alignment is preferred. In simulated annealing suboptimal solutions evolve through variations to get an optimal solution. Profiles are then combined into an optimal multiple sequence alignment. All candidate solutions are accessed in a random rather than ordered manner [47]. So all potential solutions may not be visited during random search, which reduce computational complexity.

C. SAGA AND RAGA

SAGA stands for Sequence alignment genetic algorithm. The population is made up of complete multiple sequence alignment and the operators have direct access to the aligned sequences. Insertion or shifting of gaps takes place in a random or semi random manner. Populations of multiple alignments evolve by selection, combination and mutation.

The population consists of alignments and mutations shuffle the gaps using complex models. Each individual to be a multiple alignment and it is represented by two dimensional array in which each line represent an aligned sequence and each location or cell of array is a residue or a gap. The number of individuals in a population is constant with no duplicates. The design of proper operators reflects the true mechanism of molecular evolution. The Pseudocode of SAGA is as under [49].

```
Initialization: -
1. Create Go, an initial random population
   Selection: -
2. Evaluate the population of generation n (Gn)
3. If the population is stabilized then END
4. Select the individuals to replace
5. Evaluate the expected offspring
   Variation: -
6. Select the parents from Gn
7. Select an operator
8. Generate offspring
9. Keep or discard the new offspring in Gn+1
10. Go to step 6 until all Gn+1 is complete
11. n=n+1
12. Go to step 2
13. End
```

The goal of any GA is how to initialize and generate a diverse population in term of genotype, uniformity and scores. Manual collection of sequences from genomic and proteomic databases resolve this raised issue. Fitness is measured by scoring each alignment according to the chosen objective function. Better alignment got higher scores and therefore higher fitness.

The weakest half of the population is replaced by new offspring and the other half carried over unchanged to the next generation this is called as overlapping generation. Raw alignment score is converted into an expected offspring (EO).EO is used as a probable parent in the next population [49].

A variation operator is applied to the parent alignment to create offspring alignment. Twenty two different variations operator are available in SAGA. These variations are classified into a single-parent (mutation) or multiple-parent (crossover).The algorithm is itself terminated when there is no improvement for more than 100 generations. The design of proper variation operator is important. Crossovers generate new alignments by combining information contained in two existing alignments. Two parents alignment generate two offspring. The fittest offspring survive in the next generation.

The function of mutation operator is to insert/delete a gap. The alignment is divided into two groups. The gap insertion operator inserts a gap at a specific position in both groups. Groups are chosen by randomly splitting of estimated phylogenetic tree. During algorithm execution the probability are dynamically reassessed to reflect each operator performance. Efficiency and performance using different operators facilitate the use of dynamic scheduling method [47].

The island models propose several identical genetic algorithms that run parallel on separate processors [50]. After Every five generations the processors exchange some of their individuals between the evolving populations. The Island model is implemented in RAGA, The RNA version of SAGA. This distributed model is ten times faster than the non-parallel version [51]. Migration of best score individuals between processors is possible after specified number of generations. In SAGA the donor processor keep a copy of donated individuals and the migrating individual replace low-scoring alignment in the recipient genetic algorithm. Relevant and appropriate objective or fitness function is very important to the overall efficiency of any Genetic Algorithm (GA).

There are three kinds of objective functions used in SAGA, weighted sums of pairs, consistency based and taking nonlocal interaction into account. Consistency based is used in the COFFEE score and RAGA is based upon the nonlocal interaction. RAGA use island model in cases of parallel implementation. RAGA can be used to evaluate the alignment of two RNA sequences. The Sequence with a known secondary structure is called as the master and one that is homologous to the master but with unknown secondary structure is termed as the slave RNA sequence. In cases of weighted sum of pairs method [52] each aligned residue and gap is associated with a cost and the combine cost of multiple sequence alignment is the sum of individual pair cost and substitution cost. The computational cost of sum of pairs [33] can be given by

$$\mathrm{C} = \sum_{\mathrm{i}=1}^{\mathrm{n}-1} \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{W}_{\mathrm{i,j}}\, \mathrm{cost}\left(\mathrm{A}_{\mathrm{i}}, \mathrm{A}_{\mathrm{j}}\right)$$

Optimization of consistency based objective function uses the same technique as discussed above in the section of multiple sequence alignment. Consistency based objective function in cases of SAGA use an already aligned multiple sequence alignment as a guide for pairwise alignment in the process of multiple sequence alignment. SAGA, RAGA and PRAGA are available for free download. PRAGA is a parallel version of RAGA [57].

D. MESSY GENETIC ALGORITHM (mGA)

This algorithm is used for polypeptide structure prediction and work on building block hypothesis (BBH). Small pieces of solution (alignments) combine and recombine into larger pieces. Small pieces of alignment may be disrupted by crossover or mutation so this algorithm encodes the string position. String position in this case means the locus and it value or allele. And we achieve a true building block in this case. In this algorithm there are underspecified and over specified strings to exist in the population. Underspecified strings have no allele defined for locus while over specified have multiple alleles for the same locus. Start of open reading frame can also be declared as string position. Open reading frame have genes codons and are also called as exons or those areas of nucleic acid that encode proteins [56].

E. KENOBI ALGORITHM FOR SEQUENCE ALIGNMENT

The goal of this algorithm is to develop biologically useful alignment. This algorithm first aligns the most conserved portions of proteins, their cores, as represented by secondary structure elements (SSE) [54]. A genetic algorithm then optimizes the alignment according to an elastic similarity score. This evolutionary algorithm generates optimal alignment which is very near to manual biological alignment [47].

F. K2 ALGORITHM

Improved version of KENOBI algorithm with rapid vector based SSE. This algorithm also use genetic algorithm for calculating the statistical significance of resultant alignments. In K2 algorithm hybridization of fast vector-based SSE with slower but reliable genetic algorithm take place. This algorithm handles difficult problems. Vector based SSE mean to represent SSE for two proteins with vectors and then identify a set of equivalent vectors.

The purpose of vector based SSE is to reduce computational complexity of structure based alignment. Vector alignment can be computed very quickly and efficiently. This algorithm first finds optimal alignment within SSEs. Vector based alignment is an intelligent direct method for selecting initial population of alignments. The initial population is then refined in detail.

This algorithm work in three stages, first selection of best alignments of any protein SSE. A genetic algorithm manipulates these selected alignments to optimize the alignment of amino acid position with the given SSEs. Finally protein backbones are superposed based on the equivalencies determined in the first and second stages. K2 then searches for additional equivalent positions in the non-SSE regions. This hierarchical approach to alignment reflects the nature of protein structure.

IX. CONCLUSION

The correct estimation of match, mismatch and gap, determine the quality of a given alignment approach [2]. There are many issues related with optimal sequence matching. Dynamic programming optimization and its several improved variants for sequence comparison is an optimal choice. The computational sequence analysis and alignment is very important. Different kind of sequence alignment approaches and a lot of tools and technologies are in practice for this purpose. Parallel processors and specialized kind of hardware make easy the complex task of multiple sequence alignment [61].

Protein is the end product of nucleic acid gene codon translation. A comparison that take into account the three dimensional structure of protein is more important, due to more informative, stable and reliable arrangement of molecules. Online bioinformatics resources play a significant role in all kinds of alignments and analysis methods. The goal of next generation sequence alignment is the parallel processing of algorithms. The application of multiple algorithms together or hybrid approach also ensures optimality of matching techniques. Almost all computational approaches are statistical probabilistic estimations or mathematical optimizations.

Parallel genetic algorithms overcome the computational burden of multiple sequence alignment. The self-improving and evolving capable Genetic algorithms are applied for both word based and structure based MSA. Parallel computing and hybrid strategy reduces the time burden and improve efficiency. Right now we observe a race among researchers to develop and propose optimal alignment libraries or packages in different programming languages.

REFERENCES

[1] C. L. Pierri, G. Parisi, and V. Porcelli, "Computational approaches for protein function prediction: A combined strategy from multiple sequence alignment to molecular docking-based virtual screening," *Biochimica Biophysica Acta (BBA)-Proteins Proteomics*, vol. 1804, no. 9, pp. 1695–1712, Sep. 2010. doi: 10.1016/j.bbapap.2010.04.008.

[2] P. Bawono, A. van der Velde, S. Abeln, and J. Heringa, "Quantifying the displacement of mismatches in multiple sequence alignment benchmarks," *PLoS One*, vol. 10, no. 5, 2015, Art. no. e0127431. doi: 10.1371/journal.pone.0127431.

[3] C. Vinzant, "Lower bounds for optimal alignments of binary sequences," *Discrete Appl. Math.*, vol. 157, no. 15, pp. 3341–3346, Aug. 2009. doi: 10.1016/j.dam.2009.06.028.

[4] Z. Zhang, M. Lindstam, J. Unge, C. Peterson, and G. Lu, "Potential for dramatic improvement in sequence alignment against structures of remote homologous proteins by extracting structural information from multiple structure alignment," *J. Mol. Biol.*, vol. 332, no. 1, pp. 127–142, Sep. 2003. doi: 10.1016/S0022-2836(03)00858-1.

[5] S. J. Spielman, K. Kumar, and C. O. Wilke, "Comprehensive, structurally-informed alignment and phylogeny of vertebrate biogenic amine receptors," *PeerJ*, vol. 3, p. e773, Feb. 2015. doi: 10.7717/peerj.773.

[6] C. H. z. Siederdissen, I. L. Hofacker, and P. F. Stadler, "Product grammars for alignment and folding," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 3, pp. 507–519, May/Jun. 2015.

[7] S. R. Sathe and D. D. Shrimankar, "Parallelizing and analyzing the behavior of sequence alignment algorithm on a cluster of workstations for large datasets," *Int. J. Comput. Appl.*, vol. 74, no. 21, pp. 1–13, Jul. 2013.

[8] M. Zuker, "Suboptimal sequence alignment in molecular biology: Alignment with error analysis," *J. Mol. Biol.*, vol. 221, no. 2, pp. 403–420, Sep. 1991. doi: 10.1016/0022-2836(91)80062-Y.

[9] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970. doi: 10.1016/0022-2836(70)90057-4.

[10] M. S. Waterman and T. F. Smith, "RNA secondary structure: A complete mathematical analysis," *Math. Biosciences*, vol. 4, nos. 3–4, pp. 257–266, Dec. 1978. doi: 10.1016/0025-5564(78)90099-8.

[11] C. X. Chan, G. Bernard, O. Poirion, M. James Hogan, and A. Mark Ragan, "Inferring phylogenies of evolving sequences without multiple sequence alignment," *Sci. Rep.*, vol. 4, p. 6504, Sep. 2014. doi: 10.1038/srep06504.

[12] A. S. M. Hossain, B. P. Blackburne, A. Shah, and S. Whelan, "Evidence of statistical inconsistency of phylogenetic methods in the presence of multiple sequence alignment uncertainty," *Genome Biol. Evol.*, vol. 7, no. 8, pp. 2102–2116, Jul. 2015. doi: 10.1093/gbe/evv127.

[13] P. Jehl, F. Sievers, and D. G. Higgins, "OD-seq: Outlier detection in multiple sequence alignments," *BMC Bioinf.*, vol. 16, p. 269, Aug. 2015. doi: 10.1186/s12859-015-0702-1.

[14] E. S. Wright, "DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment," *BMC Bioinf.*, vol. 16, p. 322, Oct. 2015. doi: 10.1186/s12859-015-0749-z.

[15] G. Tan *et al.*, "Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference," *Systematic Biol.*, vol. 64, no. 5, pp. 778–791, 2015.

[16] G. Huang, H. Zhou, Y. Li, and L. Xu, "Alignment-free comparison of genome sequences by a new numerical characterization," *J. Theor. Biol.*, vol. 281, no. 1, pp. 107–112, Jul. 2011. doi: 10.1016/j.jtbi.2011.04.003.

[17] Z. Gong, F. Li, and L. Dong, "Performance assessment of protein multiple sequence alignment algorithms based on permutation similarity measurement," *Biochem. Biophys. Res. Commun.*, vol. 399, no. 4, pp. 470–474, Sep. 2010. doi: 10.1016/j.bbrc.2010.07.103.

[18] A. Boukerche, A. C. M. A. de Melo, M. Ayala-Rincón, and M. E. M. T. Walter, "Parallel strategies for the local biological sequence alignment in a cluster of workstations," *J. Parallel Distrib. Comput.*, vol. 67, no. 2, pp. 170–185, Feb. 2007. doi: 10.1016/j.jpdc.2006.11.001.

[19] J. Daugelaite, A. O'Driscoll, and R. D. Sleator, "An overview of multiple sequence alignments and cloud computing in bioinformatics," *ISRN Biomath.*, vol. 2013, Jun. 2013, Art. no. 615630. doi: 10.1155/2013/615630.

[20] P. Borovska, V. Gancheva, and S. Ko, "Scaling of parallel multiple sequence alignment on the supercomputer JUQUEEN," in *Proc. IEEE 7th Int. Conf. Intell. Data Acquisition Adv. Comput. Syst.*, Berlin, Germany, Sep. 2013, pp. 687–691.

[21] D. Díaz *et al.*, "MC64-ClustalWP2: A highly-parallel hybrid strategy to align multiple sequences in many-core architectures," *PLoS One*, vol. 9, no. 4, 2014, Art. no. e94044. doi: 10.1371/journal.pone.0094044.

[22] G. Parmentier, D. Trystram, and J. Zola, "Large scale multiple sequence alignment with simultaneous phylogeny inference," *J. Parallel Distrib. Comput.*, vol. 66, no. 12, pp. 1534–1545, Dec. 2006. doi: 10.1016/j.jpdc.2006.03.003.

[23] G. Landan and D. Graur, "Characterization of pairwise and multiple sequence alignment errors," *Gene*, vol. 441, nos. 1–2, pp. 141–147, Jul. 2009. doi: 10.1016/j.gene.2008.05.016.

[24] S. Costantini, G. Colonna, and A. M. Facchiano, "FASMA: A service to format and analyze sequences in multiple alignments," *Genomics, Proteomics Bioinf.*, vol. 5, nos. 3–4, pp. 253–255, 2007. doi: 10.1016/S1672-0229(08)60013-3.

[25] Y. Fu, Z. Z. Xu, Z. J. Lu, S. Zhao, and D. H. Mathews, "Discovery of novel ncRNA sequences in multiple genome alignments on the basis of conserved and stable secondary structures," *PLoS One*, vol. 10, no. 6, 2015, Art. no. e0130200. doi: 10.1371/journal.pone.0130200.

[26] T. Madden, "The BLAST sequence analysis tool," in *The NCBI Handbook [Internet]*, 2nd ed. Bethesda, MD, USA: National Center for Biotechnology Information (US), 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK153387/

[27] W. R. Pearson, "Rapid and sensitive sequence comparison with FASTP and FASTA," *Methods Enzymol*, vol. 183, pp. 63–98, 1990. doi: 10.1016/0076-6879(90)83007-V.

[28] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3pp. 403–410, Oct. 1990. [Online]. Available: http://www.blastalgorithm.com/. doi: 10.1016/S0022-2836(05)80360-2.

[29] J. D. Blake and F. E. Cohen, "Pairwise sequence alignment below the twilight zone," *J. Mol. Biol.*, vol. 307, no. 2, pp. 721–735, Mar. 2001. doi: 10.1006/jmbi.2001.4495.

[30] L. C. E. Zhou, "CombAlign: A code for generating a one-to-many sequence alignment from a set of pairwise structure-based sequence alignments," *Source Code Biol. Med.*, vol. 10, p. 9, Aug. 2015. doi: 10.1186/s13029-015-0039-1.

[31] D. W. Mount, "Comparison of the PAM and BLOSUM amino acid substitution matrices," *Cold Spring Harbor Protocols*, vol. 2008, 2008. doi: 10.1101/pdb.ip59.pdb.ip59.

[32] J.-M. Chang, P. Di Tommaso, V. Lefort, O. Gascuel, and C. Notredame, "TCS: A web server for multiple sequence alignment evaluation and phylogenetic reconstruction," *Nucleic Acids Res.*, vol. 43, pp. W3–W6, Jul. 2015. doi: 10.1093/nar/gkv310.

[33] V. K. Sohpal, A. Dey, and A. Singh, "MEGA biocentric software for sequence and phylogenetic analysis: A review," *Int. J. Bioinf. Res. Appl.*, vol. 6, no. 3, pp. 230–240, 2010.

[34] G. Amos, *MATLAB: An Introduction with Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2004.

[35] S. Warris, F. Yalcin, K. J. L. Jackson, and J. P. Nap, "Flexible, fast and accurate sequence alignment profiling on GPGPU with PaSWAS," *PLoS One*, vol. 10, no. 4, 2015, Art. no. e0122524. doi: 10.1371/journal.pone.0122524.

[36] G. Peris and A. Marzal, "Normalized global alignment for protein sequences," *J. Theor. Biol.*, vol. 291, pp. 22–28, Dec. 2011. doi: 10.1016/j.jtbi.2011.09.017.

[37] Z.-H. Qi, X.-Q. Qi, and C.-C. Liu, "New method for global alignment of 2 DNA sequences by the tree data structure," *J. Theor. Biol.*, vol. 263, no. 2, pp. 227–236, Mar. 2010. doi: 10.1016/j.jtbi.2009.12.012.

[38] K. Boyce, F. Sievers, and D. G. Higgins, "Instability in progressive multiple sequence alignment algorithms," *Algorithms Mol. Biol.*, vol. 10, no. 1, p. 26, 2015. doi: 10.1186/s13015-015-0057-1.

[39] Q. Zhan, Y. Ye, T.-W. Lam, S.-M. Yiu, Y. Wang, and H.-F. Ting, "Improving multiple sequence alignment by using better guide trees," in *Proc. 10th Int. Symp. Bioinf. Res. Appl. (ISBRA)*, Zhangjiajie, China. Jun. 2014, p. 383.

[40] P. Ajawatanawong, G. C. Atkinson, N. S. Watson-Haigh, B. Mackenzie, and S. L. Baldauf, "SeqFIRE: A web application for automated extraction of indel regions and conserved blocks from protein multiple sequence alignments," *Nucleic Acids Res.*, vol. 40, no. W1, pp. W340–W347, 2012. doi: 10.1093/nar/gks561.

[41] C. Sharma, P. Agrawal, and P. Gupta, "Article: Multiple sequence alignments with parallel computing," in *Proc. IJCA Int. Conf. Adv. Comput. Eng. Appl. ICACEA*, no. 5, Mar. 2014, pp. 16–21.

[42] D. W. Mount, *Dot Matrix Pairwise Sequence Comparison Bioinformatics: Sequence and Genome Analysis*, 2nd ed. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press, 2004, ch. 3. doi: 10.1101/pdb.top31.

[43] S. Henikoff and J. G. Henikoff, "Performance evaluation of amino acid substitution matrices," *Proteins*, vol. 17, no. 1, pp. 49–61, 1993.

[44] S. F. Altschul, "Amino acid substitution matrices from an information theoretic perspective," *J. Mol. Biol.*, vol. 219, no. 3, pp. 555–565, 1991.

[45] J. G. Esteve and F. Falceto, "Classification of amino acids induced by their associated matrices," *Biophys. Chem.*, vol. 115, nos. 2–3, pp. 177–180, Apr. 2005. doi: 10.1016/j.bpc.2004.12.023.

[46] B.-J. M. Webb, J. S. Liu, and C. E. Lawrence, "BALSA: Bayesian algorithm for local sequence alignment," *Nucleic Acids Res.*, vol. 30, no. 5, pp. 1268–1277, 2002.

[47] G. Fogel and D. Corne, *Evolutionary Computation in Bioinformatics*. San Francisco, CA, USA: Elsevier, 2003. pp. 41–43.

[48] M. T. Pervez *et al.*, "IVisTMSA: Interactive visual tools for multiple sequence alignments," *Evol. Bioinform. Online*, vol. 11, pp. 35–42, Mar. 2015. doi: 10.4137/EBO.S18980.

[49] C. Notredame and D. G. Higgins, "SAGA: Sequence alignment by genetic algorithm," *Nucleic Acids Res.*, vol. 24, no. 8, pp. 1515–1524, 1996.

[50] P. Borovska and V. Gancheva, "Massively parallel multiple sequence alignment on the supercomputer JUQUEEN," *Int. J. Comput.*, vol. 12, pp. 1–8, 2018.

[51] A. Y. Zomaya, *Grid Computing for Bioinformatics and Computational Biology*. Hoboken, NJ, USA: Wiley, 2007, chs. 2–5.

[52] J. Stoye, S. W. Perrey, and A. W. M. Dress, "Improving the divide-and-conquer approach to sum-of-pairs multiple sequence alignment," *Appl. Math. Lett.*, vol. 10, no. 2, pp. 67–73, Mar. 1997. doi: 10.1016/S0893-9659(97)00013-X.

[53] S. Vinga, "Editorial: Alignment-free methods in computational biology," *Briefings Bioinf.*, vol. 15, no. 3, pp. 341–342, 2014.

[54] S. Dietrich *et al.*, "Experimental assessment of the importance of amino acid positions identified by an entropy-based correlation analysis of multiple-sequence alignments," *Biochemistry*, vol. 51, no. 28, pp. 5633–5641, 2012. doi: 10.1021/bi300747r.

[55] K. D. Nguyen, "On the edge of web-based multiple sequence alignment services," *Tsinghua Sci. Technol.*, vol. 17, Aug. 6, pp. 629–637, Dec. 2012.

[56] D. Hoksza and D. Svozil, "Multiple 3D RNA structure superposition using neighbor joining," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 3, pp. 520–530, May/Jun. 2015.

[57] A. T. Magis, C. C. Funk, and D. And N. D. Price, "SNAPR: A bioinformatics pipeline for efficient and accurate RNA-seq alignment and analysis," *IEEE Life Sci. Lett.*, vol. 1, no. 2, pp. 22–25, Aug. 2015.

[58] J. L. Herman, Á. Novák, R. Lyngsø, A. Szabó, I. Miklós, and J. Hein, "Efficient representation of uncertainty in multiple sequence alignments using directed acyclic graphs," *BMC Bioinf.*, vol. 16, p. 108, Apr. 2015. doi: 10.1186/s12859-015-0516-1.

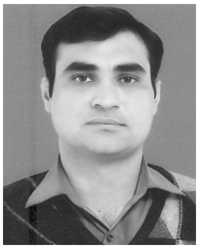
[59] F. Sievers, G. M. Hughes, and D. G. Higgins, "Systematic exploration of guide-tree topology effects for small protein alignments," *BMC Bioinf.*, vol. 15, no. 1, p. 338, 2014.

[60] M. L. Tress, D. Jones, and A. Valencia, "Predicting reliable regions in protein alignments from sequence profiles," *J. Mol. Biol.*, vol. 330, no. 4, pp. 705–718, Jul. 2003. doi: 10.1016/S0022-2836(03)00622-3.

[61] N. Sebastião, N. Roma, and P. Flores, "Hardware accelerator architecture for simultaneous short-read DNA sequences alignment with enhanced traceback phase," *Microprocessors Microsyst.*, vol. 36, no. 2, pp. 96–109, Mar. 2012. doi: 10.1016/j.micpro.2011.05.003.

[62] L. M. O. Matos, D. Pratas, and A. J. Pinho, "A compression model for DNA multiple sequence alignment blocks," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3189–3198, MAY 2013.

MUHAMMAD ISHAQ received the Ph.D. degree in computer science from Harbin Engineering University, China, in 2012. He became a member of several research societies in the relevant discipline. He is currently an Assistant Professor with The University of Agriculture, Peshawar. He is an active Researcher in the fields of bioinformatics, semantic sciences, and ontology engineering. Beside publication, projects, review services, and conference presentations, he also delivered plenary in National and International (IEEE ICCSNT 2011) and chaired many international conference sessions.



ASFANDYAR KHAN received the M.Sc. degree from the University of Peshawar, Pakistan, in 2003, and the Ph.D. degree from the Department of Computer and Information Sciences, University of Technology PETRONAS, Malaysia, in 2011. During his Ph.D. degree, he was a Graduate Assistant with the University of Technology PETRONAS. From 2011 to 2017, he was an Assistant Professor with the Department of Computer Science, University of Science and Technology, Bannu, Pakistan. He is currently a Faculty Member with the Department of Computer Science and IT, Institute of Business and Management Sciences (IBMS), The University of Agriculture, Peshawar, Pakistan. His research interests include wireless sensor networks, healthcare information security, cyber security, real-time systems, green computing, bioinformatics, and medical imaging systems. He is also a member of the IEEE Islamabad Section.

MAJID KHAN received the M.S. degree (Hons.) in computer science and the M.C.S. degree (Hons.) from The University of Agriculture, Peshawar (UAP). He is currently pursuing the Ph.D. degree with the University of Engineering and Technology, Peshawar.

He was a Lecturer with the Department of Statistics, Mathematics and Computer Science, UAP. He is a competent academic administrator having additional duties as a Senior Warden of hostel, a member of the Proctor Board, and the Assistant Director Vice Chancellor Secretariat. His research interests include mobile ad hoc networks (MANETs), cloud computing, wireless sensor networks, information security, and computational biology.

MUHAMMAD IMRAN received the Ph.D. degree (Hons.) in computer science (specifically in the area of cloud computing and data preservation) from the University of Vienna, Austria, in 2014. He is currently an Assistant Professor and a member of the Postgraduate Advisory Committee at the Institute of Business and Management Sciences, The University of Agriculture, Peshawar, Pakistan. He has published several research papers in international journals, has presented his work in many conferences of international repute, and has chaired many conferences. His research interests include cloud computing, data preservation, provenance, and service-oriented architectures.