Sequence Alignment

7

Manoj Kumar Gupta, Gayatri Gouda, N. Rajesh, Ravindra Donde, S. Sabarinathan, Pallabi Pati, Sushil Kumar Rathore, Ramakrishna Vadde, and Lambodar Behera

Abstract

The sequence analysis is one of the most effective and commonly applied methods (explicity or implicitly) in biological research. Thus, in this chapter, author attempted to understand the basics of sequence analysis and how researchers implement various computational tools to achieve them. Information obtained revealed that alignment can be either global and local or pairwise sequence alignment and multiple sequence alignment. For performing these alignment, various algorithms like dynamic programming, heuristic algorithms, or probabilistic methods have been developed. Sequence analysis helps us to detect evolutionary relationship as well as scan motifs by taking into consideration of various events, such as mutations, insertions, deletions, and reordering under some circumstances. Thus, sequence alignment serves as an essential requirement for the most of the biological research ranging from genomics to proteomics. However, our perception of alignment biases remains primitive.

M. K. Gupta (✉) · G. Gouda · R. Donde · L. Behera
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

N. Rajesh · R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_7

Thus, there is an urgent requirement to explore the effect of alignment bias on broad comparative genomics accuracy. In the near future, information present in this chapter will be useful for retriving information biological sequence.

Keywords

Dynamic programming · BLAST · FASTA · Needleman–Wunsch algorithm · gap penalties · Smith-Waterman algorithm

Abbreviations

BLAST	Basic local alignment search tool
DP	Dynamic programming
HMM	Hidden Markov Model
MAFFT	Multiple sequence alignment based on Fast Fourier Transform
MSA	Multisequence alignments
MUSCLE	Multiple sequence comparison by log-expectation
PA	Progressive alignment
PSA	Pairwise sequences alignment
UPGMA	Unweighted pair group method with arithmetic mean

7.1 Introduction

More than 12 million organisms reside on the earth. This biodiversity is mainly due to distinct genomic and proteomic sequences contained in these organisms. These sequences store unique information that modulates various processes required for the survival of these organisms [1]. DNA sequence comparison is a unique approach to evaluate gene-level variations amongst these organisms and to study their differences and similarities [1]. What "similarities" are identified to rely on the alignment process's objectives. The easiest way for comparing two same-length sequences is to identify the number of matching characters. The attribute that calculates sequence similarity is known as the alignment value of two sequences. On the contrary, the degree of dissimilarity between sequences is known as the sequence distance. The amount of characters that do not align is known as the hamming distance. However, while estimating similarity, this approach does not take into consideration of normal biological activities like insertion or deletions.

The classic definition of sequence alignment includes estimating the so-called "edit distance," which normally equals the minimum number of insertions, substitution, and deletion that are necessary for transforming one sequence into another [2]. Earlier several algorithms, like Smith & Waterman and Needleman & Wunsch have been developed for computing "edit distance" [3, 4]. These algorithms were originally developed for protein-protein alignment and subsequently employed for

DNA sequence alignment. In the majority of the real-life scenarios, nevertheless, these algorithms seem inefficient for DNA alignment owing to their runtime as well as memory requirements [2].

To date, several kinds of alignment approaches, like prediction-based methods, pairwise sequences alignment (PSA), profile-based methods, multisequence alignments (MSA), and the structure-based methods have been proposed [5]. The most frequently used are PSA and MSA. In PSA, per sequence is aligned once a time. It is the easiest method of aligning and can be achieved with two strategies: local and global. The MSA approach could also be implemented using local or global strategies but is much more complex. During MSA, many protein sequences are organized into a rectangular array, and residues that are either homologous or identical are placed in one column. MSA is generally employed for detecting conserved regions in protein sequence and for designing protein's secondary and tertiary structures. Homology, as well as evolutionary relationships between sequences, may also be derived via MSA approaches because MSA has an underlying postulation, i.e., all matching sequences would share evolutionary homology [5]. Alignment results are also a requirement for many other downstream analyses, like drug design. Nevertheless, results generated by different methods can be quite diverse [6]. Thus, there is an urgent requirement for the development of systematic metrics that may provide explicit guidance on the strengths as well as shortcomings of the different sequence alignment algorithms. This, in turn, will help us to deduce a more significant relationship between sequences. Considering the above, in this chapter, the author attempted to provide an overview of sequence alignment with a summary of popular specific algorithms, methods, and approaches which underlie the most current method of sequence alignment.

7.2 Basic Terminology

A sequence alignment is a basic analysis in almost every biological study (implicit or explicit). The main objective of sequence alignment is to detect the homologous sites in sequences [7]. Homology is a qualitative argument and identifies shared ancestral relations between sequences. Two distinct types of homology exist, i.e., paralogs (shared ancestry due to a duplication event) and ortholog (shared ancestry due to a speciation event) [8]. *"By definition, orthologs are genes that are related by vertical descent from a common ancestor and encode proteins with the same function in different species. By contrast, paralogs are homologous genes that have evolved by duplication and code for a protein with similar, but not identical functions"* [9]. Other terms that are commonly used during sequences analysis are similarity and identity [10]. Unlike homology, similarity denotes the percentage of aligned residues with the same physicochemical properties that are easier to replace each other. It is pertinent to note that two sequences can be 70% similar but cannot share 70% homology. They are either nonhomologous or homologous [10]. In general, a shared ancestral relationship could be inferred if the sequence similarity level is very high. However, it is not really obvious at what similarity degree one should assume

homologous relationships. The solution depends on the sequence type and lengths under consideration [10]. For instance, proteins having high sequence identity and high structural similarity have similar functional and evolutionary relationships [11]. Identity corresponds to the proportion of matches between the two aligned sequences with the same amino acid residue [10].

Another term, namely gap, is common during sequence analysis. A gap can be defined as the absence of a segment in a certain sequence. Gaps are natural feature of biological sequences. A single mutational event can result in the addition or deletion of certain regions of sequences (predominantly in DNA), and thus the effective identification of gaps is an important step toward understanding the various biological phenomenon [12]. A variety of biological processes may lead to the formation of gaps in DNA sequences, like, large pieces of DNA may be replicated and inserted through a single mutational occurrence, and slippage during the replication of the DNA can allow the same region to be replicated many times as replication machine lose its position on the template [12]. Earlier it has been reported that instead of penalizing all editing operations individually, one must penalize the formation of a longer gap more severely than others [13].

7.3 Alignment Methods

To date, different alignment approaches like dynamic programming (DP), heuristic algorithms, or probabilistic methods have been developed [14].

7.3.1 Dynamic Programming

DP is an effective computing strategy implemented to a problem class that can be addressed recursively [15]. When Richard Bellman first developed the DP algorithm in 1953 for researching "multi-stage decision problems," he certainly did not expect its extensive usage within modern computer programming. Indeed, as Bellman has described in his comical autobiography [16], he wanted to employ the word "dynamic programming" as "an umbrella" for the mathematical research he carried out at RAND Corporation for protecting his boss, who was the Secretary of Defense Wilson and "had a pathological fear of word research." Since it is one of the first algorithms that were used in bioinformatics research and has since been widely applied [17], DP has become an inevitable algorithmic subject.

DP is indeed a normal preference for evaluating sequences. Needleman & Wunsch initially illustrated the use of bottom-up DP for calculating an optimal pairing amongst two protein sequences [3]. While this algorithm offers a comparative evaluation of sequences pair, it estimates the similarity throughout the complete sequences (a "global alignment algorithm"). Hence, this approach is time-consuming and computationally exhaustive [18]. To overcome this, Smith and Waterman adapted DP for performing local alignments in which alignment was made between similar parts of the input sequences [4]. DP provides an ideal

approach for PSA [18]. It is also widely employed to assembling DNA sequence data from fragments obtained from automated sequencing machines and for determining the exon/intron structure within eukaryotic genes [19]. It is also utilized for inferring proteins' function through homology study with other proteins having a known function [3, 4], and for predicting the secondary structure of functional RNA genes or regulatory elements [19].

7.3.2 Heuristic Algorithms

Though DP gives a more accurate result, it is slow [14]. Other efficient approaches, like heuristic algorithms or probabilistic methods, have been developed for large-scale database searching. The term "heuristic" means that the developed algorithm is faster than the classical method but may not be the optimum method [20]. Heuristic algorithms can be categorized into three subgroups, namely, progressive alignment (PA) approach, iterative alignment type, and block-based alignment type [10]. PA approach is the incremental strategy that generates a final MSA through conducting a set of PSA on successively less closely associated sequences. In this approach, we align the two closest-related sequences first and then align the closest-related sequence in the questionnaire to the alignment generated in the previous step. Although success is particularly dependent on the consistency of the initial alignment and dramatically deteriorates when all sequences in the set are related distantly, PA methods are enough to be implemented on a broad scale for several sequences [21]. The most commonly used PA methods are ClustalW (https://www.genome.jp/tools-bin/clustalw) and T-Coffee (https://www.ebi.ac.uk/Tools/msa/tcoffee/). However, it is not possible that the progressive approaches converge to optimal global alignment, and efficiency can be difficult to approximate. Additionally, its true biological importance may be unclear [21].

The iterative method is based on the premise that an ideal solution could be sought by adjusting current suboptimal solutions on a repeated basis. The process begins with a low-quality alignment and gradually improves it through well-defined procedures until no more improvement can be achieved on the alignment scores. Since the sequence order in each iteration is different, this method could mitigate the "greedy" problem of progressive strategy. Nevertheless, this approach is also heuristic in nature and has no promises for optimum alignment [10]. PRRN (https://www.genome.jp/tools-bin/prrn) is a web-based program that utilizes a double-nested iterative strategic plan for multiple alignments. The progressive as well as iterative alignment techniques are primarily global and thus cannot detect conserve motifs and domains amongst strongly diverging sequences of various lengths. A local alignment strategy must be employed for those divergent sequences that share only local similarities. This technique detects the ungapped alignment block that is present in all sequences, and hence this is called the block-based local alignment technique [10]. DIALIGN2 (http://dialign.gobics.de/) web-tools that employ block-based alignment for detecting local alignment.

7.3.3 Probabilistic Methods

Introduction of probabilistic modeling approaches, like profile secret Markov models (profile HMMs) as well as pair-HMMs [22] have advanced sequence similarity search. When variables are probabilities instead of random scores, objective statistical parameters refine them more readily. This helps to create more detailed, biologically relevant models with many parameters. For instance, profile HMMs employ position-specific deletion/insertion probabilities instead of the random, position-invariant gap expense of more conventional approaches like BLAST or PSI-BLAST [23], enabling profile HMMs to model the possibility that indels occur more frequently in certain sections of a protein than others (e.g., in surface loops than submerged core) [24].

The probability method has three primary benefits: (i) Any kind of analogy may be adjusted to the probabilities [e.g., The DNA error-prone reads against the genome]. The comparisons are supposed to be more precise. (ii) We may approximate the reliability, for instance, each column of every alignment part. This is helpful because alignments also have unknown sections owing to high inconsistencies or repeating sequences. (iii) A similarity between two integrated sequences over potential alignments may be calculated. This can more powerfully detect subtle connexions than single ideal alignments [25]. The probabilistic approach, however, also has significant disadvantages. Aside from a moderate computational drawback, the probabilistic method suffers from uncharacterized score statistics - unlike the local alignment of Smith-Waterman, for which at least the form of the ideal score distribution is defined from the null model, relatively little is known about the distribution of the log-like score in the local probabilistic random alignment. It is proven empirically that random usage of the z-score would not deliver really strong results [26].

7.4 Global and Local Alignment

Sequence alignment approaches typically fell into two categories: global and local alignments. While global alignment compares all character of query sequences, local alignments define similarity regions within long sequences that are typically divergent. The Needleman-Wunsch algorithm is a well-known global alignment algorithm designed on the basis of DP. Local alignments are always preferred but more challenging to quantify considering the additional difficulty of recognizing similarities regions. The Smith-Waterman algorithm is a general local alignments method based on the DP system, with added features for beginning and finishing in either place [14]. Most biologists think that local alignment is what really matters when we are looking for functional conservation. Local alignment is more important since certain proteins have roles that are controlled by their capability to attach to some other molecule (protein's ligand); therefore, the role would be maintained if this short portion becomes sustained via evolution, even if there is significant divergence in many other protein regions. As proteins are folded within their natural

form, these retained regions need not be continuous protein segments. Indeed, several researchers researching on lymphocyte antigen recognition specifically account for these discontinuities within binding domains (known as "non-linear" epitopes, where an epitope is the ligand of a lymphocyte) [12, 14].

In few cases of the global alignment mode, adding a distance in the leftmost location of the alignment might be needed, but we are not aware of the length of the next reference sequence factor to be already aligned. It is obvious from this scenario that an intermediate alignment is required between the local and global alignment (i.e., semiglobal alignment) [12]. A semiglobal alignment does not penalize starting or ending gaps in any global alignment so that the resultant alignment continues to overlap one end of a sequence with the end of the other [27]. A Parasail is a stand-alone tool that can be employed for performing global, local, and semi-global alignment [27]. Recently, Suzuki & Kasahara developed a semi-global alignment algorithm, namely, "difference recurrence relationships," that perform better than other available tools by 2.1 factor [28].

7.5 Pairwise Alignments

The most frequently employed mean of collecting information from protein and DNA sequences is a PSA. It is generally used to detect protein homolog, which diverged more than 2 billion years ago. For proteins that share statistically significant sequence similitudes, homology can be accurately inferred. If statistically meaningful similarities to a known sequence are observed, inferences may be made regarding the unknown sequence's function, structure, and biologically significant residues. Although the homology assumption [29] is very robust (i.e., proteins which share significant similarities within PSA often have similar features), a few of the more detailed preassumptions critically rely on the consistency of the alignment between the two sequences. For instance, functional inferences for protein sequences having more than 60% identity are typically very reliable. However, uncertainty in the alignment of badly conserved areas can lead to errors for more distantly linked proteins [30, 31].

The fundamental law for sequence alignment is the structural alignment amongst two proteins known to have a 3D structure. The 3D-structure comprises more information relative to the 1-D sequence as well as diverges at a very slow rate. Thus, distant evolutionary correlations may also be established amongst sequences which do not display statistically significant similarities. Even directly relevant proteins with major sequence similarities may elicit sequence alignments that differ from the most accurate structural alignments. Since it is not possible to identify the three-dimensional structure of each protein, researchers are continually seeking for strategies for producing structurally correct homology models for sequences with unknown structure. The most common as well as successful methods, are to find a template for constructing the model within the set of established structures. This feature is relatively trivial in the case of high sequence similitude (i.e., $> 60\%$ identity) because both sequences, as well as structural alignments, are typically very

near to this range. However, in this zone, there are just a few sequences; in the so-called "twilight zone," there are several more sequences (i.e., ~20–40 percent sequence identity) where divergent yet clearly homologous protein may be hard to match. Although the precision of the end 3D model is dependent on the degree of alignment of the unspecified sequence to the structural template, researchers are mainly concentrating on enhancing the quality of alignment between proteins that share statistically relevant similarities and have 20% to 40% sequence identity [49, 50]. Dot-matrix techniques, DP, and Word techniques are the most widely used methods for PSA.

7.5.1 DOT Matrix Plot

Since visualization of alignment of character of hundreds or more sequences can be troublesome, scientists created a more visually understandable approach called the dot matrix approach. This sequence alignment process, which was first carried out manually and then computationally, allows the more apparent mapping of similarities for visual inspection. In this process, a sequence is shown on the top and one on the side of the matrix and a mark on the crossroads of the corresponding character pairs [51]. A dot matrix pattern will have a continuous array of dots running along the middle diagonal of the matrix for a pair of exactly matched sequences (Fig. 7.1). However, this trend is hardly used. Sometimes, without further processing, diagonal patterns are hard to recognize. Thus, a number of filters are also added to the results, as well as the use of color and other methods to highlight matching sequences. For instance, typical filtering is a stringency/window combination. The window represents the number of points evaluated at a time, while the minimum number of matches needed in each window is the stringency [51].

The study of the dot matrix is extremely valuable in recognizing recurring characters or short sequences within one sequence, as is the case for the mapping the recurrent regions of entire chromosomes. Repeats of the same character produce artificially high scores and complicate sequence alignment. Methods of dot matrix are most appropriate for single PSA problems, particularly for relatively high similitudes. Sequences with a lower similarity and MSA need more efficient methods [51]. Even though window stringency values are always heuristically determined, they could be dependent on dynamic averages, matched scores in aligned protein groups, or different methods for calculating the amino acid similarity. For example, score matrices establish alignment scores in the aligned protein families depending on their statistical frequency. These matrices may be used to construct a sliding window, where only scores above an average scoring may appear in the matrix, as defined in the following section [51].

To date, various algorithms and computer software tools were created for performing the dot-matrix plot. While several of these tools accommodate 100 kb of sequences, the study of the genome sequences above 10 Mb on a microcomputer remains to be inoperative considering the length of time needed for execution as well as computer memory [53]. In 2004, Huang and Zhang created two dot matrix

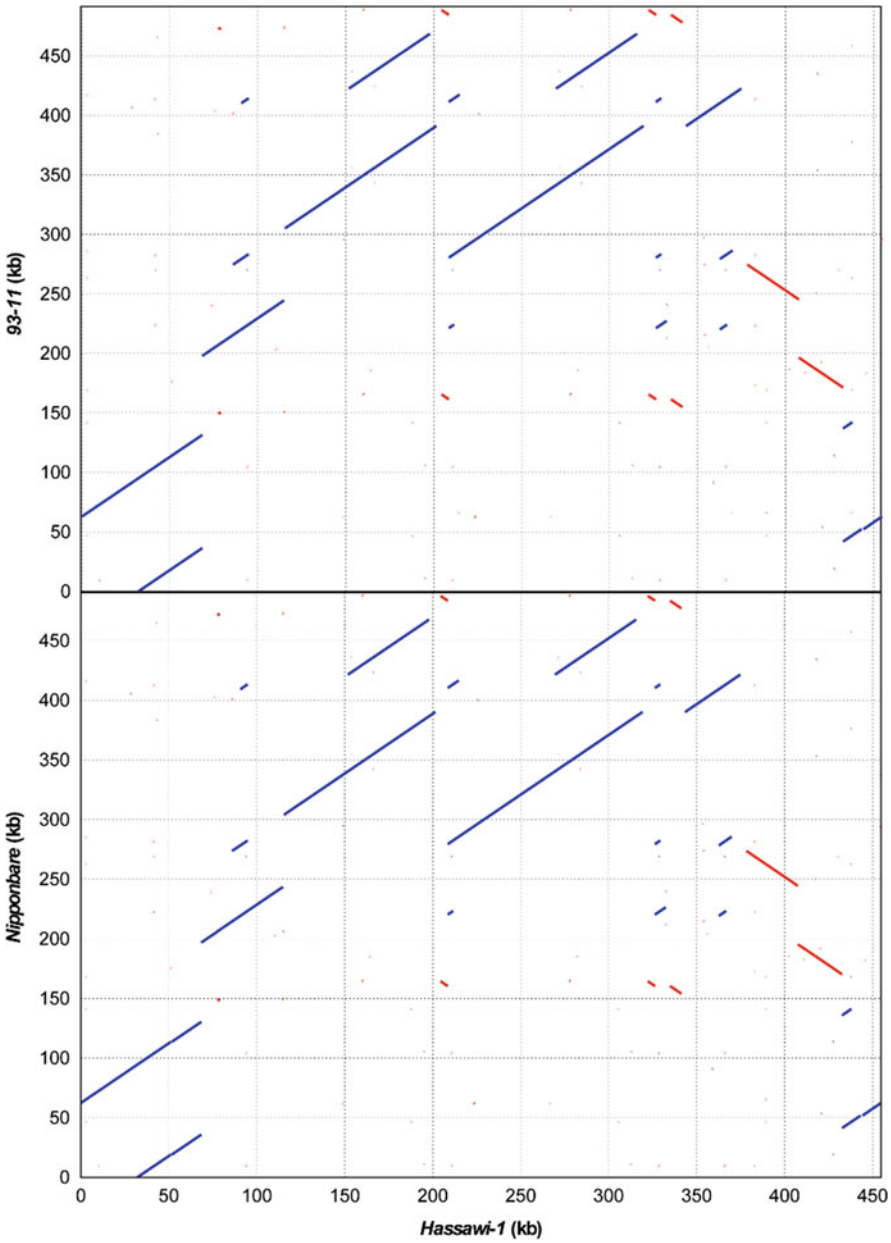

Fig. 7.1 The dot-plot of the alignment for human chromosomes 2, 7, and 14 and mouse chromosome 12. The x-axis indicates the positions of mouse chromosome 12, and y-axis indicates the positions of human chromosomes 2, 7, and 14. The orthologous landmarks are plotted based on the pairwise alignments between the three human chromosomes and mouse chromosome 12 (Adapted from [52]).

Table 7.1 Softwares and tools used for PSA (Adapted from https://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

Name	Description	Alignment type[a]	Sequence type[b]	References
ACANA	Fast heuristic anchor-dependent PSA	Both	Both	[32]
AlignMe	Membrane PT sequences alignment	Both	PT	[33]
Bioconductor biostrings:: pairwiseAlignment	DP	Both + ends-free	Both	[34]
BioPerl dpAlign	DP	Both + ends-free	Both	https://metacpan.org/pod/release/CJFIELDS/BioPerl-1.6.924/Bio/Tools/dpAlign.pm
BLASTZ, LASTZ	Seeded pattern-matching	LL	Nucleotide	[35]
DNASTAR Lasergene molecular biology suite	Align RNA, DNA, PT, or PT + DNA sequences	Both	Both	https://www.dnastar.com/
FEAST	Posterior-dependent LL extension having descriptive evolution model	LL	Nucleotide	[36]
G-PAS	GPU-based DP with backtracking	LL, SemiGL, GL	Both	http://gpualign.cs.put.poznan.pl/gpas20.html
GapMis	Does PSA with one gap	SemiGL	Both	[37]
Genome magician	Software for ultra-fast LL DNA sequence motif scan as well as PSA of high-throughput data in both FASTA and FASTQ format.	LL, SemiGL, GL	DNA	https://science.do-mix.de/software_genomemagician.php
GGSEARCH, GLSEARCH	GL:LL (GL) and GL: GL (GG) alignment with statistics	GL in query	PT	[38]
JAligner	Java-based techniques of Smith-Waterman	LL	Both	http://jaligner.sourceforge.net/

(continued)

Table 7.1 (continued)

Name	Description	Alignment type[a]	Sequence type[b]	References
K*sync	PT sequence to structure alignment that comprises of secondary structure, structure-derived sequence profiles, structural conservation, and consensus alignment scores	Both	PT	[39]
LALIGN	Multiple, nonoverlapping, LL similarity	LL nonoverlapping	Both	https://www.ebi.ac.uk/Tools/psa/lalign/
mAlign	Modeling alignment; models the information content of the sequences	Both	Nucleotide	[40]
Matcher	Waterman-Eggert LL alignment (dependent on LALIGN)	LL	Both	https://www.ebi.ac.uk/Tools/psa/emboss_matcher/
MCALIGN2	Explicit models of indel evolution	GL	DNA	[41]
MUMmer	Suffix tree-dependent	GL	Nucleotide	[42]
NW-align	Standard Needleman-Wunsch DP algorithm	GL	PT	https://zhanglab.ccmb.med.umich.edu/NW-align/
Needle	Needleman-Wunsch DP	SemiGL	Both	https://www.ebi.ac.uk/Tools/psa/emboss_needle/
Ngila	Logarithmic as well as affine gap costs and explicit models of indel evolution	GL	Both	[43]
Parasail	C/C++/python/Java SIMD DP library for SSE, AVX2	GL, ends-free, LL	Both	[27]
Path	Smith-Waterman on PT back-translation graph (detects	LL	PT	[44]

(continued)

Table 7.1 (continued)

Name	Description	Alignment type[a]	Sequence type[b]	References
	frameshifts at PT level)			
PatternHunter	Seeded pattern-matching	LL	Nucleotide	[45]
SABERTOOTH	Alignment employing predicted "connectivity profiles"	GL	PT	[46]
Satsuma	Parallel whole-genome synteny alignments	LL	DNA	Genome-wide synteny through highly sensitive sequence alignment
SPA: Super pairwise alignment	Fast pairwise GL alignment	GL	Nucleotide	[47]
SWIFOLD	Smith-Waterman acceleration on Intel's FPGA with OpenCL for long DNA sequences	LL	Nucleotide	[48]
UGENE	Opensource Smith-Waterman for SSE/CUDA, suffix array-based repeats finder and dotplot	Both	Both	http://ugene.net/

[a]Alignment type: Global(GL)/Local(LL)
[b]Sequence type: Nucleotide (NT)/Protein(PT)

comparison methods for studying large sequences. Initially, the methods identify similarity regions amongst two sequences using a rapid word search algorithm and explicitly compare these regions. Because several random matches are omitted from the initial sampling, the estimation duration is decreased dramatically. These approaches yield good quality plots of the dot matrix with low background noise. Spatial criteria are linear, so genome scaling sequences can be compared by algorithms. Highly repetitive sequence structures of eukaryote genomes may impact the computational speed. In the 80s, with a 1GHz personalized machine, a dot matrix complot was developed for the yeast genome (12 Mb) for both strands [53].

7.5.2 Dynamic Programming

The most widely employed algorithm of PSA is DP, initially introduced by Needleman and Wunsch [3]. The DP ensures an optimum algorithmic alignment

with unique parameters and sequences. However, an optimum sequence alignment score would not assure the structural consistency of the alignment. Additionally, there are no natural mechanisms under which two proteins align together. Therefore "optimum" alignments of the sequence may vary greatly from ideal structural alignments [31]. Moreover, distant-related proteins also have several optimal alignments and a significant number of sub-optimal alignments with scores quite similar to the optimal score [50, 54, 55]. If one moves further from the desired score, the number of alternatives alignment also keeps increasing. Therefore, one must sample the suboptimal alignment space for holding the number of alignments computationally trackable [50, 54, 55].

While a structure-based alignment is the "gold standard" against which sequence alignments are measured, structural alignment may vary, and no optimum structural alignment algorithm is possible [56]. As the number of structures appear to be smaller than the number of sequences, the structural alignment variations are minimal relative to the sequence-structural alignment variations. Although this definitely refers to quite distantly linked proteins that have no meaningful similitude (and therefore cannot be substantially aligned with sequence data alone), the structural and sequence alignment precision of proteins that share statistically significant similarities has not been closely studied [56]. Given that structurally correct alignments frequently include suboptimal alignment scores, researchers have been researching the alternate alignments and wondering whether they include details about precise structural alignments. Jaroszewski et al. [50] have studied alternate alignments, both based on an almost ideal algorithm for alignment generation and by combining score parameters (i.e., substitution matrix and gap penalties), and have found that alignment in the sets is much similar to the structural alignment. Their inference was that the two alternate alignment methods, namely, alternatives and sub-optimizing alignments, had complementary information (in contrast to redundant information) because the combination of the two sets created much higher alignments than any of the sets. The exactness of the optimal sequence alignment was also investigated by Holmes and Durbin [57]. They developed a technique for calculating the expected accuracy. In an algebraic approach, Zhang and Marr [58] used alternate alignments with maximal alignments in the neighborhood.

Various scholars also took the help of a probabilistic approach for producing alternate alignment sets. In 1995, Miyazawa [59] measured alignment likelihoods relying on alignment score exponent and, subsequently, compared the resulting likelihoods of matched amino acids throughout alignment with the respective protein structure alignments. Yu and Hwa investigated the statistically significant of alignments made using a pairwise Hidden Markov Model (HMM) [26]. Knudsen and Miyamoto [60] designed a pairwise HMM alignment approach that provided an explicit indel evolutionary model. Eventually, Mückstein and the team [61] constructed a sampling alignment procedure on the basis of statistical weighting employing partition function overall plausible two-sequence alignments.

Although it is of theoretical interest to compare individual sequence and structure sets in the absence of any structural information, it is only of practical use if the alignment of the sequence can be determined correctly. One approach to resolving

this issue is to calculate the accuracy of a certain aligned residual pair (that we term an edge, using the norm for determining the optimum score in the dynamic programming path graph, aligned residues, insertions, and deletions along the edge) [31]. Cline and the team examined four strategies for forecasting the accuracy of a particular pair of aligned residues [62] and concluded that the most improved alignment quality was the method proposed by Yu and Smith [63] for retrieving near-optimal alignments from the HMM profile. The association between both the edge probabilities and structural alignment was studied by Knudsen & Miyamoto [60] and Mückstein et al. [61] and Miyazawa [59]. However, in the former two cases, only in the context of a limited number of protein pairs, usually considered a strong correspondence amongst them. In another study, Mevissen and Vingron [64] have evaluated the feasibility of an edge reliability index known as robustness that Chao and the team had previously defined [65]. They found that an edge's robustness predicted correctly if the edge was still aligned in structural alignment. In another study, Sierka and the team improvised the robustness analysis by adding extra details on alignment consistency and creating a logistic regression model that returns the likelihood that a given edge is embedded in a structural alignment [31].

7.5.3 The Word or K-Tuple (Ktup) Method

It is the heuristic process, which offers greater alignment than DP. Currently, with massive datasets, DP cannot be used. This is why we use the K-tuple approach when searching for a specific question along with a large database. K Tuple corresponds to a series of k words. For instance, for nucleotide and protein, K is defined as 11 and 3, respectively. The K system has been introduced in the family of FASTA and BLAST.

7.5.3.1 FASTA

FASTA is a rapid alignment application for protein and DNA sequence pairs. Rather than comparing individual residues in both sequences, FASTA looks for matching sequence patterns or terms called k-tuples. In both sequences, these patterns contain k consecutive matches of letters. Based on these word matches, the algorithm then tries to establish a local alignment. FASTA is useful for regular database searches of this kind because of the ability of the algorithm to locate similar sequences in a sequence database with high-speed. FASTA programs offer a detailed range of simple similarity search resources (fasta36, fastx36, tfastx36, fasty36, and tfasty36), comparable to those offered by the BLAST tool, as well as programs for local, slower, optimal, as well as global similarity searches (search36, ggsearch36) and oligonucleotide and short peptide searches (fasts36, fastm36). fasta36 employs the FASTA algorithm developed by Pearson alone and Pearson & Lipman and compare protein (or nucleotide) sequence to protein (or nucleotide) sequence database [66, 67]. With the ktup (word size) parameter, search speed and selectivity are regulated. By default, ktup = 2 for protein comparisons; ktup = 1 is more sensitive but slower. By default, ktup = 6 for DNA comparisons; ktup = 3 or ktup = 4 allows

maximum sensitivity. fastx36/fasty36 compares the translated nucleotide sequence into three frames and allowing gaps and changes, fastx36 compares a nucleotide sequence to a protein sequence base. Fastx36 uses a faster and simplified alignment algorithm, which only allows the frameshift between codons. However, fasty36 is slower, but better alignments are possible because frame shifts inside codons are permitted [68]. tfastx36/ tfasty36 compares a protein sequence with a nucleotide sequence database and measures comparisons for forward and reverse directed frames-shifts [68]. ssearch36 employs the Smith-Waterman algorithm [4] for comparing a nucleotide (or protein) sequence against a nucleotide (or protein) sequence database. The Fasta36 is just 2–5 times faster than Farrar SSE2 [69]. ggsearch36/ glsearch36 compares a protein (or nucleotide) sequence to a protein (or nucleotide) sequence database, employing an optimal global algorithm: global: local (glsearch36) or global (ggsearch36). fasts36/ tfasts36 compares collection of small peptide fragments as collected from mass-spec, protein research, against nucleotide (tfasts) or protein (fasts) databases [70]. fastm36 compares ordered short nucleotide sequences (or peptides) to a nucleotide (or peptides) database.

The FASTA systems employ an empiric approach for approximating statistical importance that is consistent with a variety of similarities in scores and gap penalties and increases alignment of boundary precision as well as search sensitivity. FASTA systems can generate "BLAST-like" alignment as well as tabular results for ease of integrating analytics pipelines and can scan for small, descriptive datasets and afterward report findings for larger sequences employing small dataset connexions. FASTA systems operate in a wide range of database formats, like PostgreSQL and MySQL databases. Recently, Pearson has developed programs that lay out a strategy for incorporating domain as well as active site annotations into alignments and emphasizing the mutation status of functionally important residues. These protocols also explain how FASTA systems can classify protein and nucleotide sequences through protein: DNA, protein: protein, and DNA: DNA comparative study [71].

7.5.3.2 BLAST

The "Basic local alignment search tool" (BLAST) is a sequence similarity search software which could be employed either as a stand-alone tool or through a web interface for comparing all combinations of protein (or nucleotide) sequence to a protein (or nucleotide) sequence database [72]. BLAST is a heuristic approach that finds short matches between two sequences and tries to initiate alignment from these "hot spots." BLAST also offers statistical details about alignment in addition to executing alignments [72]. The E-value contains details on the probability of a sequence being matched by sheer chance. The smaller the E-value, the less probable the database match is to be attributed to random chance, and thus the more important the match. If $E < 1e^{-50}$ (or 1×10^{-50}), there should be an exceptionally strong conviction that matching the database is the product of a homologous partnership. If E is between 0.01 and $1e - 50$, matching can be viewed as a consequence of homology. If E is between 0.01 and 10, the match is assumed to be nonsignificant but could suggest a possible remote homology relationship. Additional proof is required to validate the partnership. If $E > 10$, the sequences within evaluation are

either unrelated or associated with incredibly remote relationships that fall far below the detection limit of the current system [10]. Although the E-value is proportionally influenced by the size of the database, an apparent concern is that as the database expands, the E-value often increases for a given sequence match. Since the true evolutionary relationship between the two sequences remains unchanged, as the database expands, the decline in the sequence match's credibility means that one will "lose" homologs previously observed as the database enlarges. Consequently, an alternative to E-value calculations is needed [10].

BLAST is a family of services that comprises BLASTN, BLASTX, BLASTTP, TBLASTX, and TBLASTN. BLASTN searches nucleotide sequences in the nucleotide sequence database. BLASTP employs protein sequences as requests to scan a database of protein sequences. BLASTX employs nucleotide sequences as inputs and converts them into all six reading frames to generate translated protein sequences that are used to query the protein sequence database. TBLASTN requests protein sequences to a nucleotide sequence database, with sequences encoded into all six reading frames. TBLASTX employs nucleotide sequences that are interpreted into all six frames to scan a nucleotide sequence database that has all the sequences interpreted into six frames. In addition, also there is a bl2seq program that executes a local alignment of two user-provided input sequences. The graphic production involves horizontal bars as well as a diagonal in a two-dimensional diagram displaying the total degree of the matching between the two sequences [10].

7.6 Multiple Sequence Alignment

MSA is an alignment between more than two biological sequences. In most scenarios, the input sequences are believed to have a shared ancestor. Sequence homology can be derived from the subsequent MSA, and a phylogenetic study can be carried out to determine the common ancestral roots of the sequences. Visual alignment representations, as seen in the Fig. 7.2, demonstrate mutation occurrences like point mutations (single nucleotide or amino acid changes) that occur as distinct symbols within a single alignment column and insertion/deletion of mutations (indels or gaps) that occur as hyphens in one or more alignment sequences. MSA can also be used to determine sequence conservation of protein domains, tertiary as well as secondary structures, as well as specific amino acids or nucleotides [73–75].

Since MSA of three or more lengthy sequences may be complicated and are often time-consuming to be aligned by hand, statistical algorithms are often used for generating and evaluating alignments. MSAs need more advanced approaches than PSA since they are more computationally complicated. Many MSA programs use heuristic approaches rather than global optimization since it is prohibitively costly to determine the optimum alignment amongst more than a few sequences of moderate length. On the other side, heuristic approaches usually refuse to guarantee the consistency of the answer, with heuristic strategies sometimes found to be well below the ideal solution in the case of benchmarks [73–75].

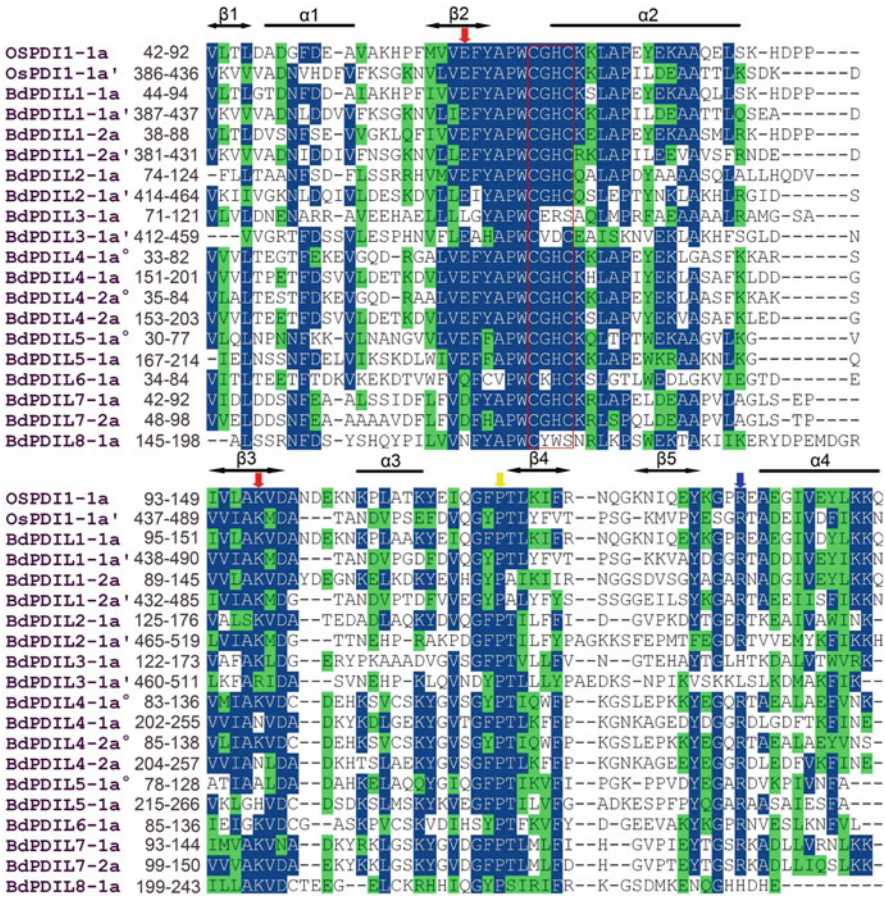

Fig. 7.2 "Multiple sequence alignment of *a*-type domains of *B. distachyon* PDI and PDI-like proteins and a typical rice PDI. These thioredoxin-like domains of the *B. distachyon* were annotated in Phytozome database, and comparative analysis used BioEdit software. Residues highlighted in deep blue and green show they were identical and similar, respectively. Open bars and arrowheads represent the α helices and β strands, respectively. The red box indicates the -CxxC- catalytic site, and red arrows indicate the glutamicacid–lysine charged pair. Blue and yellow arrows represent the conserved arginine (R) and the *cis* pralines (P) near the active site, respectively" (Adapted from [76]).

7.6.1 Dynamic Programming

The complex programming algorithms, namely, Smith-Waterman and Needleman-Wunsch, that are employed for a PSA, can also be used for evaluating the optimum alignment of over two sequences. Nevertheless, the difficulty of this algorithm is much shoddier than that of PSA. For performing PSA, the running period of the algorithm is proportionate to $m \times n$, where m and n are the lengths of two aligned sequences. If $n \geq m$, the argument is generalized to indicate that the algorithm's

Table 7.2 Softwares and tools used for MSA (Adapted from https://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

Name	Description	Alignment type[a]	Sequence type[b]	Alignment type[a]	References
ABA	A-Bruijn alignment	GL	PT	GL	[77]
CHAOS, DIALIGN	Iterative alignment	LL (preferred)	Both	LL (preferred)	[78]
ClustalW	PA	LL or GL	Both	LL or GL	[79]
CodonCode aligner	MSA; ClustalW and Phrap support	LL or GL	NT	LL or GL	https://www.codoncode.com/aligner/
Compass	COmparison of multiple PT sequence alignments through statistical assessement	GL	PT	GL	[80]
DECIPHER	Progressive-iterative alignment	GL	Both	GL	[81]
DIALIGN-TX and DIALIGN-T	Segment-based method	LL (preferred) or GL	Both	LL (preferred) or GL	[82]
DNA baser sequence assembler	MSA; full automatic sequence alignment; automatic ambiguity correction; internal base caller; command line seq alignment	LL or GL	NT	LL or GL	https://www.dnabaser.com
DNADynamo	Linked DNA to PT MSA with MUSCLE, Smith-Waterman and Clustal	LL or GL	Both	LL or GL	https://www.bluetractorsoftware.com/
DNASTAR Lasergene molecular biology suite	Software to align RNA, DNA, PT, or DNA + PT sequences via pairwise and MSA algorithms	LL or GL	Both	LL or GL	https://www.dnastar.com/
FAMSA	PA for extremely huge PT families	GL	PT	GL	[83]
FSA	Sequence annealing	GL	Both	GL	http://fsa.sourceforge.net/
Geneious	Progressive-iterative alignment; ClustalW plugin	LL or GL	Both	LL or GL	https://www.geneious.com/
Kalign	PA	GL	Both	GL	[84]

MAFFT	Progressive-iterative alignment	LL or GL	Both	LL or GL	[85]
MARNA	MSA of RNAs	LL	RNA	LL	[86]
MAVID	PA	GL	Both	GL	[87]
MSAProbs	DP	GL	PT	GL	[88]
MULTALIN	DP-clustering	LL or GL	Both	LL or GL	[89]
Multi-LAGAN	Progressive DP alignment	GL	Both	GL	[90]
MUSCLE	Progressive-iterative alignment	LL or GL	Both	LL or GL	[91]
Opal	Progressive-iterative alignment	LL or GL	Both	LL or GL	[92]
Pecan	Probabilistic consistency	GL	DNA	GL	[93]
Phylo	A human computing framework for comparative genomics to solve MSA	LL or GL	NT	LL or GL	[94]
Praline	Progressive-iterative-consistency-homology-extended alignment with preprofiling and secondary structure prediction	GL	PT	GL	[95]
PicXAA	Nonprogressive, maximum expected accuracy alignment	GL	Both	GL	[96]
POA	Partial order/HMM	LL or GL	PT	LL or GL	[97]
Probalign	Probabilistic/consistency with partition function probabilities	GL	PT	GL	[98]
ProbCons	Probabilistic/consistency	LL or GL	PT	LL or GL	[99]
PROMALS3D	PA/HMM/secondary structure/3D structure	GL	PT	GL	[100]
PRRN/PRRP	Iterative alignment (especially refinement)	LL or GL	PT	LL or GL	https://www.genome.jp/tools-bin/prrn
PSAlign	Alignment preserving nonheuristic	LL or GL	Both	LL or GL	[101]
RevTrans	Combines DNA and PT alignment, by back translating the PT alignment to DNA.	LL or GL	DNA/PT (special)	LL or GL	[102]

(continued)

Table 7.2 (continued)

Name	Description	Alignment type[a]	Sequence type[b]	Alignment type[a]	References
StatAlign	Bayesian co-estimation of alignment and phylogeny (MCMC)	GL	Both	GL	[103]
Stemloc	MSA and secondary structure prediction	LL or GL	RNA	LL or GL	[104]
T-coffee	More sensitive PA	LL or GL	Both	LL or GL	[105]
UGENE	Supports MSA with MUSCLE, KAlign, Clustal, and MAFFT plugins	LL or GL	Both	LL or GL	http://ugene.net/
GLProbs	Adaptive pair-HMM based approach	GL	PT	GL	[106]

[a]Alignment type: GL (GL)/LL(LL)
[b]Sequence type: Nucleotide (NT)/proteins(PT)

execution time is n^2. The exponent in the n^2 definition derives from the presumption that, during PSA, if we presume that our sequences length is n, then $n \times n$ cells need to be filled within the dynamic programming matrix. If we were to employ either Needleman-Wunsch or Smith-Waterman algorithm to three sequences, we would need to build a 3-dimensional array for measuring and monitoring the alignment. Therefore, for sequences having n length, we will have $n \times n \times n$ cells for filling in (http://readiab.org/book/0.1.3/2/3). Runtime for MSA employing complete DP algorithms increases dramatically with the sequences number to be aligned. If s and n are the sequence number and sequence length, respectively, then the execution time will be ns. However, in PSA, $s = 2$, which makes the problem handier (http://readiab.org/book/0.1.3/2/3).

7.6.2 Progressive Alignment

PA is a heuristic approach and does not optimize any obvious alignment score. The aim is to accomplish a series of PSA that begins with aligning nearest identical sequence pairs and subsequently aligning least similar ones [22, 107]. The PA method reduced the overall computational difficulty to polynomial-time by splitting the MSA problem into a set of PSA guided by a tree reflecting the evolutionary sequence relation [108]. Today, most popular alignment programs that employ the progressive approach are ClustalW [79], Mafft ("Multiple sequence alignment based on Fast Fourier Transform") [109], "Multiple sequence comparison by log-expectation" (MUSCLE) [91], and T-Coffee [110].

7.6.2.1 ClustalW

ClustalW is currently the most commonly deployed alignment software, and the oldest of the modules examined. The program conducts a PA, first using PSA through computing the distance matrix that retains the sequence's discrepancy. Just after the matrix is collected, a guided tree is created utilizing Neighbor-Joining algorithms, accompanied by a final stage where the sequences are aligned as per the branching order within the guide tree. In its alignment procedure, the software utilizes two gap penalties: gap expansion and gap opening, during polypeptides availability, a total amino acid weight matrix. These distance penalties rely strongly on variables like sequence length, similarity, and weight matrix. In a simple scenario, Clustal W will exactly match the related domains and sequences of established secondary or tertiary structures but can be seen as a strong starting point for more refinement in more complicated cases (Fig. 7.3a) [73, 79].

7.6.2.2 Mafft

Mafft is a program that can be employed with different alignment methods, either PA alone (with Fast Fourier Transform) or iteratively aligned PA. Mafft's basic run requires up to three stages, but the default procedure performs the first two steps. The first stage is to create a PA centered on each sequence pair's rough distance, on the basis of the mutual 6-tuples. The unweighted pair group method with arithmetic

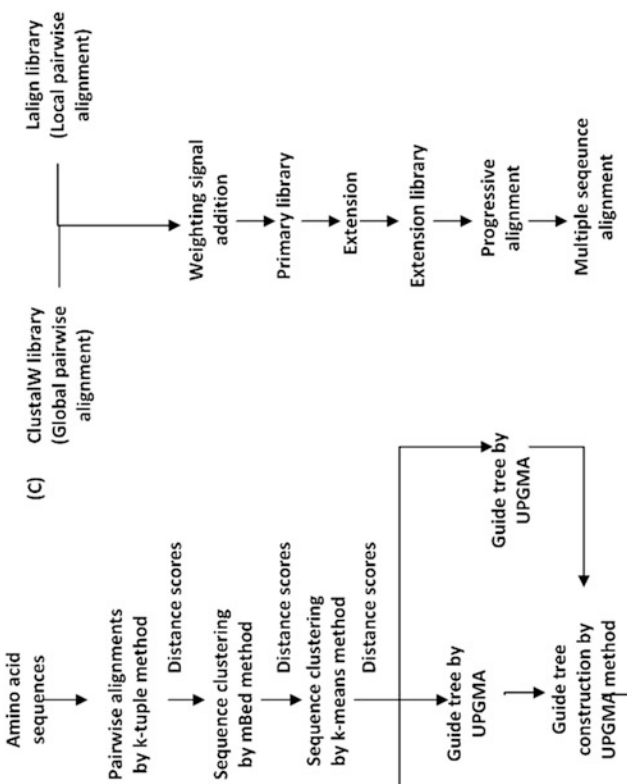

Fig. 7.3 Steps for generating MSA via (a) ClustalW, (b) Clustal omega, and (c) T-Coffee (Adapted from [75])

Table 7.3 Softwares and tools used for motif scanning (Adapted from https://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

Name	Description	Sequence type[a]	Reference
BASALT	Multiple motif and regular expression scan	Both	http://www.proteinguru.com/toolbox/basalt/
BLOCKS	Ungapped motif prediction from BLOCKS database	Both	https://www.genome.jp/tools/motif/
CUDA-MEME	GPU accelerated MEME (v4.4.0) algorithm for GPU clusters	Both	https://cuda-meme.sourceforge.io/homepage.htm#latest
eMOTIF	Extraction and prediction of shorter motifs	Both	http://motif.stanford.edu/distributions/emotif/
FMM	Motif scan and prediction (can get also positive and negative sequences as input for enriched motif scan)	NT	[128]
Gibbs motif sampler	Stochastic motif extraction by statistical likelihood	Both	[129]
HMMTOP	Prediction of transmembrane helices and topology of PTs	PT	[130]
MEME/MAST	Motif prediction and scan	Both	[125]
MERCI	Discriminative motif prediction and scan	Both	[131]
PHI-blast	Motif scan and alignment tool	Both	[132]
Phyloscan	Motif scan tool	NT	[133]
PMS	Motif scan and prediction	Both	[134]
PRATT	Pattern production for use with ScanProsite	PT	https://www.ebi.ac.uk/Tools/pfa/pratt/
ScanProsite	Motif database scan tool	PT	https://prosite.expasy.org/scanprosite/
TEIRESIAS	Motif extraction and database scan	Both	[135]

[a]Sequence type: Protein (PT) or Nucelotide (NT)

mean (UPGMA) guide tree is then generated with the changed linkage, and the sequences are then aligned with the tree branch order (the so-called FFT-NS-1 strategy). In the second phase, the distance matrix is recalculated based on the knowledge obtained from the previous stage, and the PA is reassessed using a tree from the existing matrix as the starting point (till this process, the technique is known as FFT-NS-2 and is the preferred approach used by the software). The final step is the iterative refinement, which optimizes the "Gotoh weighted pair sum" (WSP) score [111], the "group-to-group alignment" [85], and "the tree-dependent constraint partition technique" [112]. The method is referred to as FFT-NS-i, where all three steps are used, which indicates that it employs the FFT method to conveniently distinguish the homologous regions throughout the sequences followed by the refining iterative process. The FFT converts an amino acid inside a sequence into

a vector describing volume and polarity that is key to replacement instances, allowing the software to accurately predict these events [73].

Three additional refining algorithms are also provided by Mafft: L-INS-i, G-INS-i, and E-INS-I [113]. These strategies improve the number of steps required to align the MSA to five. In such instances, the first step would also entail the formation of a distance matrix, not employing six-fold. In comparison to the FFT-NS- * solution, the UPGMA tree is not rebuilt, and the program continues into the second step, splitting gap-free segments and store the scoring arrays from sequence to sequence for each gap-free segment. Mafft subsequently calculates the "importance" value of the segment score and stores the residue in other segments. All "importance" values are then obtained in step three of the "importance" matrix, which is rapidly followed by a group-to-group alignment of scores and a weighting scheme based on the Needleman-Wunsch algorithm [79]. The final stage refines the alignments obtained, increases the WSP score, and the fixed "importance" values. All "importance" values are then obtained in step three of the "importance" matrix, which is rapidly accompanied by a group-to-group alignment of scores and a weighting scheme centered on the Needleman-Wunsch algorithm [79]. The final stage refines the alignments obtained, strengthens the WSP score, and the prescribed "importance" values.

7.6.2.3 Muscle

The muscle uses a pairwise alignment technique to the profile. First, the program establishes a progressive alignment, which is then refined and configured in two following stages. After the similarity of the sequence, the PA is produced, the distance estimation and the UPGMA tree are calculated. Muscle utilizes two distance measurements: a km distance for unaligned series pairs and a Kimura distance for ordered pairs [91]. A new tree with the already defined Kimura distance matrix is generated by the optimization stage of PA, which guarantees a stronger alignment centered on this improved tree. The last step of refinement uses the restricted partition variant tree-dependent [112]. This approach eliminates one of the tree edges, splits the orientation, and eliminates the profiles of the two partitions, which would then be re-aligned with the profile-profile alignment. Each tree edge will be iteratively visited and the alignment with the updated description score of each sequence pair will be preserved. The edges are inspected to minimise the gap from the root by reshaping each sequence and moving to similarly associated sequence classes [91].

7.6.2.4 Clustal Omega

Clustal Omega is the Clustal family's new MSA algorithm [75]. This algorithm is used only for aligning protein sequences (though nucleotide sequences are likely to be introduced in time). The precision of Clustal Omega is comparable to other high-quality aligners on limited numbers of sequences; moreover, Clustal Omega surpasses other MSA algorithms in terms of completion time as well as overall quality of alignment on large sequence sets. In a few hours, Clustal Omega is able to align 190,000 sequences on a single process. By firstly generating pairwise

alignments using the k-tuple form, the Clustal Omega algorithm generates a multiple sequence alignment. Then, employing the mBed method, the sequences are clustered. This is accompanied by the clustering process of k-means. Next, the guide tree is built using the UPGMA method. Finally, using the HHalign module, which aligns two profile hidden Markov models (HMM) as seen in Fig. 7.3b, the multiple sequence alignment is made.

7.6.2.5 T-Coffee

T-Coffee has a radical approach to match sequences. The software first builds a library from two separate sources: Clustal W's global alignment and Lalign's local alignment [114]. Global alignments and pairwise local alignments for each pair of sequences are generated from the top ten nonoverlapping segments. The software processes global and local information and assigns weights to all PSA according to sequence identity [115]. This is accompanied by a mixture of groups that converge into a single repository. This consolidated library has an extension phase, such that the final weight of any pair of residues constitutes part of the information contained in the library. The ultimate step involves calculating the distance matrix and the neighboring joint tree by aligning the two nearest weight sequences on the tree with the stored weight of the consolidated library with a PA. The initial pair is then fixed, and no other gap can be consequently transmitted. The PA will proceed until all sequences fit [73].

Irrespective of their uses, earlier researchers have detected that the majority of PA programs employ the Neighbor-Joining algorithm for inferring a guided tree. Neighbor-Joining's O(N 3) time complexity renders it a bottleneck when large data sets are aligned. The Relaxed Neighbor-Joining algorithm relaxes the joining nodes and decreases standard time complexity to O(N 2 log N) without any major qualitative results [47]. In 2008, Sheneman explored the relationship between the topology of the guide tree and the alignment reliability. He developed two different genetic algorithms, each of which enhances the population of tree guide topologies utilizing stochastic crossover and mutation operators. One genetic algorithm, EVALYN, generates highly accurate scores when evaluated against established reference samples. Nevertheless, we find that the disruptive crossover of EVALYN restricts the genetic algorithm to a stochastic hill climb (Fig. 7.3c).

7.6.3 Probabilistic Alignment

7.6.3.1 PRANK

PRANK [116] is one of the best examples of a probabilistic MSA tool. In comparison to other alignment systems, PRANK uses phylogenetic knowledge to identify alignment differences created through deletions or insertions and then treats the two forms of events differently. As a by-product of the proper handling of inserts and deletions, PRANK will also have assumed ancestral sequences as part of the production and label the alignment gaps differently based on their origin in the insertion or deletion incident. As the algorithm infers the ancestral history of the

sequences, PRANK could be vulnerable to errors in the phylogeny guide as well as a violation of basic assumptions about the origin as well as the pattern of the gaps [116].

7.6.3.2 PSAR

In 2014, Kim and Ma developed a new metric, known as PSAR [117], that can metric the reliability of the MSA by agreeing to probabilistically sample Suboptimal Alignments (SAs). The SAs offer extra information which cannot be obtained by optimizing alignment on its own, particularly when the ideal alignment is not too far preferable to the SAs [117].

7.6.3.3 ProbPFP

Recently, Zhan and the team developed ProbPFP that incorporates HMM configured with partition function by particle swarm. The PSO algorithm was used to refine the parameters of the HMM. Subsequently, the posterior likelihood obtained by the HMM was compared with that retrieved through the partition function, and hence the integrated substitution score for the alignment was determined. To test the effectiveness of ProbPFP, 13 excellent or classical MSA methods were compared. The results show that the alignments obtained by ProbPFP have the highest mean SP and TC values for both SABmark and OXBench data sets, as well as the second highest mean TC scores and mean SP scores for BAliBASE. ProbPFP is also compared with four other excellent approaches by restoring phylogenetic trees spanning six protein families in the TreeFam database based on alignments achieved across these five approaches. The results show that the reference trees are like the phylogenetic trees rebuilt from the ProbPFP alignments compared with other approaches [118].

7.6.3.4 ProbCons

ProbCons is a modification of the regular pair-score approach and also provides a secret PA algorithm based on the pair-hidden Markov model. The alignment method is divided into the following steps, starting with the calculation of the reverse likelihood matrices for each pair of sequences. The alignment method is split into the following steps, starting with the calculation of the posterior-probability matrices for each pair of sequences. This is accompanied by a complex software calculation of each PSA's expected accuracy. The probabilistic quality transition is then used to reassess the match's accuracy. A hierarchical clustering determines the guiding tree by the similarities defined by the weighted average of the values between the sequences of every cluster. The guidance tree is employed for matching sequences with a progressive strategy. There is also a postprocessing phase in which random bipartitions of the generated alignment are realigned to find better regions for alignment. ProbCons varies from other alignment systems because it does not implement biological principles like evolutionary tree construction, role-specific gap score, and other features typically utilized with other packages [99].

7.7 Motif Search

Motif exploration is an application layer sequence analysis problem and one of the main obstacles while developing bioinformatics applications. Sequence motifs are constant in size, frequently repetitive and conserved, but at the same moment are small (approximately 6–12 Bp) and very long and are also highly variable in intergenic regions that make the motif discovery a difficult task. A motif is also known as regulatory elements in eukaryotic genes and occurs in the Regulatory Region (RR). These patterns play a crucial role in the identification of the Transcription Factor Binding Sites (TF-BSs), which aid in the understanding of gene expression regulation mechanisms [119, 120]. Motifs are broadly categorized into various forms, namely, sequence motifs, planted motifs, gapped motifs, structured motifs, and network motifs [119]. There are two major forms of algorithms for motif discovery, i.e., enumeration approach probabilistic technique. Enumeration method looks for consensus sequences; motifs are projected dependent on word counts and word similitudes; thus, this method is often named as word enumeration approach to solving Motif problem with panted Motif Problem with motif length and a maximum number of mismatches [120]. The algorithms focused on the word enumeration method extensively scan the entire search field for classifying the ones with potential substitutes, and then normally locate the global optimum. This implies, though, that they are exponential time algorithms that take long for detecting the larger one and inefficient to accommodate hundreds of sequences, and are thus only appropriate for the short motif. Additionally, these algorithms require several user-defined parameters, including the length of the motif, the number of mismatches permitted, and a minimum of sequences the motif requires to appear in [121]. The method to word enumeration can be accelerated by utilizing various data structures, like parallel processing or suffix trees. CisFinder (https://lgsun.grc.nia.nih.gov/CisFinder/), DREME [122], Weeder [123], and MCES [124] are common algorithms based on this method. A second group is a probabilistic method. This constructs a probabilistic model known as Position-Specified Weight Matrix (PSWM) or Motif Matrix, which describes a base distribution to differentiate motifs from nonmotifs for each position of TFBS and needs few search parameters [124]. MEME [125], EXTREME [126], and BioProspector [127] are the most common methods focused on probabilistic approaches. The third form, the nature-inspired approach, incorporates the core attributes of the first two approaches. This method is a basic idea and a global scan but can work with large data and long motifs concurrently. It has a dynamic intention representation, contributing to an infinite range of degenerated positions. The final form is the combinatory method, which depends on the hybrid algorithms which shape the appropriate algorithm.

7.8 Conclusion and Future Perspective

In conclusion, sequence alignment serves as a basic requirement for most of the biological research ranging from phylogenetics construction to protein design. Sequence alignment also employed for motif search in biological sequence, which in turn plays a key role in understanding the regulation of various biological phenomenon. However, because of the continuous increase of sequence amount, there is an urgent requirement of developing novel tools and techniques which can improvise the accuracy of the sequence analysis, including motif search, result obtained. Earlier several researchers have suggested that a successful tool for motif discovery can be constructed from different suggested motif discovery methods. The tool should be fitted with these features: (1) all models should be identified, (2) the overall search feature should be optimized, (3) the parallel processing abilities are needed, (4) optimized data structures should be accessible, (5) the overall search function should be able to locate both long and short motifs, (6) several motif discovery capabilities at the same time, i.e., without elimination of the discovered motif to find another motif. This research would then establish a new algorithm for motif discovery, which incorporates the key characteristics of enumerative and probabilistic approaches and utilizes them as a seed to a naturally inspired algorithm, taking into account the above-noted variables [120].

Conflict of Interest None.

References

1. Saeed U, Usman Z. Biological Sequence Analysis. In: Husi H, editor. Computational Biology [Internet]. Brisbane: Codon Publications; 2019. [cited 2020 Oct 13]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK550342/.
2. Prjibelski AD, Korobeynikov AI, Lapidus AL. Sequence Analysis. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of Bioinformatics and Computational Biology [Internet], Academic Press. Oxford; 2019. p. 292–322. [cited 2020 Oct 11]. Available from: http://www.sciencedirect.com/science/article/pii/B9780128096338201064.
3. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.
4. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147:195–7.
5. Wang Y, Wu H, Cai Y. A benchmark study of sequence alignment methods for protein clustering. BMC Bioinformatics. 2018;19:529.
6. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Sci Am Assoc Adv Sci. 2008;319:473–6.
7. Rosenberg MS. Sequence alignment: Concepts and history. Sequence Alignment: Methods, Models, Concepts, and Strategies. California: University of California Press; 2009. p. 1–22.

8. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Ann Rev Genet. 2005;39:309–38.
9. Koonin EV, Mushegian AR, Bork P. Non-orthologous gene displacement. Trends Genet. 1996;12:334–6.
10. Xiong J. Essential bioinformatics. Cambridge: Cambridge University Press; 2006.
11. Hark Gan H, Perlow RA, Roy S, Ko J, Wu M, Huang J, et al. Analysis of protein sequence/structure similarity relationships. Biophys J. 2002;83:2781–91.
12. Barton C, Flouri T, Iliopoulos CS, Pissis SP. Global and local sequence alignment with a bounded number of gaps. Theor Comput Sci. 2015;582:1–16.
13. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol. 1982;162:705–8.
14. Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms Mol Biol. 2011;6:25.
15. Ye Y, Tang H. Dynamic Programming Algorithms for Biological Sequence and Structure Comparison. Bioinform Algorithms [Internet]. 2007:7–28. [cited 2020 Oct 15]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470253441.ch2
16. Bellman R. Eye of the hurricane. Singapore: World Scientific Publishing Co Pte Ltd; 1984.
17. Sankoff D. The early introduction of dynamic programming into computational biology. Bioinformatics. 2000;16:41–7.
18. Nalbantoğlu ÖU. Dynamic programming. In: Russell DJ, editor. Multiple sequence alignment methods [internet]. Totowa: Humana Press; 2014. p. 3–27. . [cited 2020 Oct 15]. https://doi.org/10.1007/978-1-62703-646-7_1.
19. Giegerich R. A systematic approach to dynamic programming in bioinformatics. Bioinformatics. 2000;16:665–77.
20. Mukhopadhyay CS, Choudhary RK, Iquebal MA. Basic Applied Bioinformatics. Wiley-Blackwell, Hoboken; 2017.
21. Saeed F, Khokhar A. An Overview of Multiple Sequence Alignment Systems. arXiv:09012747 [cs, q-bio] [Internet]. 2009. [cited 2020 Oct 15]; Available from: http://arxiv.org/abs/0901.2747
22. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press; 1998.
23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
24. Eddy SR. A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. PLOS Comput Biol. 2008;4:e1000069.
25. Frith MC. How sequence alignment scores correspond to probability models. Bioinformatics. 2020;36:408–15.
26. Yu YK, Hwa T. Statistical significance of probabilistic sequence alignment and related local hidden Markov models. J Comput Biol. 2001;8:249–82.
27. Daily J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. BMC Bioinformatics [Internet]. 2016;17. [cited 2020 Oct 16], Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4748600/
28. Suzuki H, Kasahara M. Introducing difference recurrence relations for faster semi-global alignment of long sequences. BMC Bioinformatics. 2018;19:45.
29. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A. 1998;95:6073–8.
30. Venclovas C. Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. Proteins. 2003;53(Suppl 6):380–8.
31. Sierk ML, Smoot ME, Bass EJ, Pearson WR. Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. BMC Bioinformatics. 2010;11:146.

32. Huang W, Umbach DM, Li L. Accurate anchoring alignment of divergent sequences. Bioinformatics. 2006;22:29–34.
33. Stamm M, Staritzbichler R, Khafizov K, Forrest LR. AlignMe—a membrane protein sequence alignment web server. Nucleic Acids Res. 2014;42:W246–51.
34. Aboyoun P. Pairwise Sequence Alignments. p. 34.
35. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human–mouse alignments with BLASTZ. Genome Res. 2003;13:103–7.
36. Hudek AK, Brown DG. FEAST: sensitive local alignment with multiple rates of evolution. IEEE/ACM Trans Comput Biol Bioinform. 2011;8:698–709.
37. Flouri T, Frousios K, Iliopoulos CS, Park K, Pissis SP, Tischler G. GapMis: a tool for pairwise sequence alignment with a single gap. Recent Pat DNA Gene Seq. 2013;7:84–95.
38. Pearson WR. FASTA Search Programs. eLS [Internet]. American Cancer Society; 2014 . [cited 2020 Dec 12]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0005255.pub2
39. Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res. 2006;34:e112.
40. Wheeler WC, Gladstein DS. MALIGN: A Multiple Sequence Alignment Program. J Hered. 1994;85:417–8.
41. Wang J, Keightley PD, Johnson T. MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. BMC Bioinformatics. 2006;7:292.
42. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology. 2018;14:e1005944.
43. Cartwright RA. Ngila: global pairwise alignments with logarithmic and affine gap costs. Bioinformatics. 2007;23:1427–8.
44. Girdea M, Noe L, Kucherov G. Back-translation for discovering distant protein homologies in the presence of frameshift mutations. Algorithms Mol Biol. 2010;5:6.
45. Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search. Bioinformatics. 2002;18:440–5.
46. Teichert F, Bastolla U, Porto M. SABERTOOTH: protein structural alignment based on a vectorial structure representation. BMC Bioinformatics. 2007;8:425.
47. Sheneman LJ. The limits of progressive multiple sequence alignment [phd]. [USA]: University of Idaho; 2008.
48. Rucci E, Garcia C, Botella G, De Giusti A, Naiouf M, Prieto-Matias M. SWIFOLD: Smith-Waterman implementation on FPGA with OpenCL for long DNA sequences. BMC Syst Biol. 2018;12:96.
49. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol. 2001;8:559–66.
50. Jaroszewski L, Li W, Godzik A. In search for more accurate alignments in the twilight zone. Protein Sci. 2002;11:1702–13.
51. Bergeron BP. Bioinformatics computing. Prentice Hall Professional; 2003.
52. Lin H-N, Hsu W-L. GSAlign: an efficient sequence alignment tool for intra-species genomes. BMC Genomics. 2020;21:182.
53. Huang Y, Zhang L. Rapid and sensitive dot-matrix methods for genome analysis. Bioinformatics. 2004;20:460–6.
54. Waterman MS, Byers TH. A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. Math Biosci. 1985;77:179–88.
55. Zuker M. Suboptimal sequence alignment in molecular biology. Alignment with error analysis. J Mol Biol. 1991;221:403–20.
56. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng. 1994;7:1059–68.
57. Holmes I, Durbin R. Dynamic Programming Alignment Accuracy. J Comput Biol. 1998;5:493–504.

58. Zhang MQ, Marr TG. Alignment of molecular sequences seen as random path analysis. J Theor Biol. 1995;174:119–29.
59. Miyazawa S. A reliable sequence alignment method based on probabilities of residue correspondences. Protein Eng Des Sel. 1995;8:999–1009.
60. Knudsen B, Miyamoto MM. Sequence alignments and pair hidden Markov models using evolutionary history. J Mol Biol. 2003;333:453–60.
61. Mückstein U, Hofacker IL, Stadler PF. Stochastic pairwise alignments. Bioinformatics. 2002;18(Suppl 2):S153–60.
62. Cline M, Hughey R, Karplus K. Predicting reliable regions in protein sequence alignments. Bioinformatics. 2002;18:306–14.
63. Yu L, Smith TF. Positional statistical significance in sequence alignment. J Comput Biol. 1999;6:253–9.
64. Mevissen HT, Vingron M. Quantifying the local reliability of a sequence alignment. Protein Eng. 1996;9:127–32.
65. Chao KM, Hardison RC, Miller W. Locating well-conserved regions within a pairwise alignment. Comput Appl Biosci. 1993;9:387–96.
66. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85:2444–8.
67. Pearson WR. Effective protein sequence comparison. Methods Enzymol. 1996;266:227–58.
68. Zhang Z, Pearson WR, Miller W. Aligning a DNA sequence with a protein sequence. J Comput Biol. 1997;4:339–49.
69. Farrar M. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics. 2007;23:156–61.
70. Mackey AJ, Haystead TAJ, Pearson WR. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. Mol Cell Proteomics. 2002;1:139–47.
71. Pearson WR. Finding protein and nucleotide similarities with FASTA. Curr Protoc Bioinformatics. 2016;53:3.9.1–25.
72. Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. Nucleic Acids Res. 2006;34:W6–9.
73. Nuin PA, Wang Z, Tillier ER. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics. 2006;7:471.
74. Thompson JD, Linard B, Lecompte O, Poch O. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. Plos One. 2011;6:e18093.
75. Daugelaite J, O' Driscoll A, Sleator RD. An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics [Internet]. Hindawi: ISRN Biomathematics; 2013. p. e615630. [cited 2020 Oct 17]. Available from: https://www.hindawi.com/journals/isrn/2013/615630/?utm_source=google&utm_medium=cpc&utm_campaign=HDW_MRKT_GBL_SUB_ADWO_PAI_DYNA_JOUR_X_PCUPS&gclid=CjwKCAjwia X8BRBZEiwAQQxGx2v_vI4i9kMbWescOdwJwv8fn0RGzfe3dBlNeNp-D_OfmWBKpzMnNhoCQ28QAvD_BwE
76. Zhu C, Luo N, He M, Chen G, Zhu J, Yin G, et al. Molecular Characterization and Expression Profiling of the Protein Disulfide Isomerase Gene Family in Brachypodium distachyon L. Plos One. 2014;9:e94704.
77. Raphael B, Zhi D, Tang H, Pevzner P. A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res. 2004;14:2336–46.
78. Brudno M, Steinkamp R, Morgenstern B. The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. Nucleic Acids Res. 2004;32:W41–4.
79. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
80. Low A, Rodrigue N, Wong A. COMPASS: the COMPletely arbitrary sequence simulator. Bioinformatics. 2017;33:3101–3.

81. Wright ES. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics. 2015;16:322.
82. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics. 2005;6:66.
83. Deorowicz S, Debudaj-Grabysz A, Gudyś A. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Sci Rep. 2016;6:33964.
84. Lassmann T, Sonnhammer EL. Kalign – an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005;6:298.
85. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66.
86. Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics. 2005;21:3352–9.
87. Bray N, Pachter L. MAVID: constrained ancestral alignment of multiple sequences. Genome Res. 2004;14:693–9.
88. González-Domínguez J, Liu Y, Touriño J, Schmidt B. MSAProbs-MPI: parallel multiple sequence aligner for distributed-memory systems. Bioinformatics. 2016;32:3826–8.
89. Mitchell C. MultAlin–multiple sequence alignment. Bioinformatics. 1993;9:614.
90. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Program NCS, et al. LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 2003;13:721–31.
91. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.
92. Wheeler TJ, Kececioglu JD. Multiple alignment by aligning alignments. Bioinformatics. 2007;23:i559–68.
93. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18:1814.
94. Kawrykow A, Roumanis G, Kam A, Kwak D, Leung C, Wu C, et al. Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. PLOS ONE. 2012;7:e31362.
95. Simossis VA, Heringa J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res. 2005;33: W289–94.
96. Sahraeian SME, Yoon B-J. PicXAA-web: a web-based platform for non-progressive maximum expected accuracy alignment of multiple biological sequences. Nucleic Acids Res. 2011;39:W8–12.
97. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics. 2002;18:452–64.
98. Roshan U, Livesay DR. Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics. 2006;22:2715–21.
99. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 2005;15:330–40.
100. Pei J, Kim B-H, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36:2295–300.
101. Sze S-H, Lu Y, Yang Q. A polynomial time solvable formulation of multiple sequence alignment. J Comput Biol. 2006;13:309–19.
102. Wernersson R, Pedersen AG. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 2003;31:3537–9.
103. Arunapuram P, Edvardsson I, Golden M, Anderson JWJ, Novák Á, Sükösd Z, et al. StatAlign 2.0: combining statistical alignment with RNA secondary structure prediction. Bioinformatics. 2013;29:654–5.
104. Bradley RK, Pachter L, Holmes I. Specific alignment of structured RNA: stochastic grammars and sequence annealing. Bioinformatics. 2008;24:2677–83.

105. Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang J-M, et al. T-coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 2011;39:W13–7.
106. Ye Y, Cheung DW, Wang Y, Yiu S-M, Zhang Q, Lam T-W, et al. GLProbs: aligning multiple sequences adaptively. IEEE/ACM Trans Comput Biol Bioinformatics. 2015;12:67–78.
107. Feng D-F, Doolittle RF. [21] Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. In: Methods in Enzymology [Internet]. London: Academic Press; 1996. p. 368–82. [cited 2020 Oct 17]. Available from: http://www.sciencedirect.com/science/article/pii/S0076687996660236.
108. Maiolo M, Zhang X, Gil M, Anisimova M. Progressive multiple sequence alignment with indel evolution. BMC Bioinformatics. 2018;19:331.
109. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
110. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J Thornton. J Mol Biol. 2000;302:205–17.
111. Gotoh O. A weighting system and algorithm for aligning many phylogenetically related sequences. Comput Appl Biosci. 1995;11:543–51.
112. Hirosawa M, Totoki Y, Hoshida M, Ishikawa M. Comprehensive study on iterative algorithms of multiple sequence alignment. Comput Appl Biosci. 1995;11:13–8.
113. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005;33:511–8.
114. Huang XQ, Hardison RC, Miller W. A space-efficient algorithm for local similarities. Comput Appl Biosci. 1990;6:373–81.
115. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991;9:56–68.
116. Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008;320:1632–5.
117. Kim J, Ma J. PSAR-align: improving multiple sequence alignment using probabilistic sampling. Bioinformatics. 2014;30:1010–2.
118. Zhan Q, Wang N, Jin S, Tan R, Jiang Q, Wang Y. ProbPFP: a multiple sequence alignment algorithm combining hidden Markov model optimized by particle swarm optimization with partition function. BMC Bioinformatics. 2019;20:573.
119. Bataineh MA, Al-qudah Z, Al-Zaben A. Iterative sequential Monte Carlo algorithm for motif discovery. IET Signal Proc. 2016;10:504–13.
120. Hashim FA, Mabrouk MS, Al-Atabany W. Review of different sequence motif finding algorithms. Avicenna J Med Biotechnol. 2019;11:130–48.
121. Zhang Y, Wang P, Yan M. An Entropy-Based Position Projection Algorithm for Motif Discovery. Biomed Res Int [Internet]. 2016. [cited 2020 Oct 19]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5110948/
122. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27:1653–9.
123. Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 2004;32:W199–203.
124. Yu Q, Huo H, Chen X, Guo H, Vitter JS, Huan J. An efficient algorithm for discovering motifs in large DNA data sets. IEEE Trans Nanobioscience. 2015;14:535–44.
125. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39–49.
126. Quang D, Xie X. EXTREME: an online EM algorithm for motif discovery. Bioinformatics. 2014;30:1667–73.
127. Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput. 2001:127–38.

128. Sharon E, Lubliner S, Segal E. A Feature-Based Approach to Modeling Protein–DNA Interactions. PLoS Comput Biol [Internet]. 2008;4. [cited 2020 Dec 13], Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2516605/
129. Thompson W, McCue LA, Lawrence CE. Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences. Curr Protoc Bioinformatics. 2005. Chapter 2:Unit 2.8.
130. Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics. 2001;17:849–50.
131. Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences. Bioinformatics. 2011;27:1231–8.
132. Zhang Z, Miller W, Schäffer AA, Madden TL, Lipman DJ, Koonin EV, et al. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res. 1998;26:3986–90.
133. Carmack CS, McCue LA, Newberg LA, Lawrence CE. PhyloScan: identification of transcription factor binding sites using cross-species evidence. Algorithms Mol Biol. 2007;2:1.
134. Dinh H, Rajasekaran S, Kundeti VK. PMS5: an efficient exact algorithm for the (ℓ, d)-motif finding problem. BMC Bioinformatics. 2011;12:410.
135. Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics. 1998;14:55–67.