

Biological sequence analysis

Samvedna Shukla¹, Bhawana Mishra², Himanshu Avashthi³ and Muktesh Chandra⁴

¹Molecular and Bioprospection Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India, ²Department of Chemistry, Central University of Haryana, Jant-Pali, Mahendergarh, Haryana, India, ³Division of Genomic Resources, ICAR National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi, India, ⁴Centre of Bioinformatics, Institute of Interdisciplinary Sciences, University of Allahabad, Prayagraj, India

3.1 Introduction

Information is transferred from one generation to another in an organism through the genetic material, that is, DNA/RNA (Li & Graur, 1991). The variations incurred in DNA lead to the divergence from a common ancestor (Carroll, Grenier, & Weatherbee, 2013). Therefore in biological research, sequencing of DNA/RNA becomes the primary requirement. Many efforts are made in developing tools for sequencing; initially, fewer numbers (tens or hundreds) of nucleotides were sequenced. Recent advancements in the next-generation sequencing (NGS) techniques and upgradation of machines large number (up to billions) of bases can be sequenced. Recent developments in NGS techniques (Liu, Li, & Li, 2012) reduced the cost of sequence production, and throughput increased many folds. Due to the development of such high-throughput techniques, data generated are enormous, and databases overflow with the data.

Contrary to sequence data, the structure determination methods are time-consuming and costly. The difference creates the gap between sequencing data and its corresponding structural information availability. Therefore the use of computational prediction tools is an excellent alternative to access useful information from sequences. Hence, bioinformatics is an interesting field to bridge the gap between raw sequence and meaningful sequence. To get the functional information, evolutionary hierarchy, and molecular phylogeny from sequence data, the first step is sequence alignment. This involves the aligning of unknown sequences with the sequences present in databases with known functions (Thompson, Linard, Lecompte, & Poch, 2011).

Sequence alignment may use DNA/RNA or protein to get meaningful information. In alignment, the arrangement of sequences is made so that maximum information can be predicted (Mount, 2004). The rearrangement is made by considering spaces or gaps, evolutionary insertions, deletions, or translocations. Furthermore, aligning short reads generated with the reference genome makes the entire length of the sequences. These full-lengths are then again aligned for gene annotation purposes. NGS techniques produce the short reads that need to be aligned, and both ends of the transcripts are covered. The reads sequenced are much smaller than the biological counterpart; hence, their alignment is vital to get maximum annotation. Many tools are developed to align the short sequences, such as Bowtie2 (Langmead & Salzberg, 2012), GSnap (Wu & Nacu, 2010), and Burrows–Wheeler Aligner's Smith–Waterman Alignment (BWA-SW) (Li & Durbin, 2010).

Pairwise sequence alignment (PSA) and Multiple sequence alignment (MSA) are two sequence alignment methods. PSA is used for aligning two sequences, while MSA aligns more than two sequences with generating an alignment score. MSA is considered more advantageous because it aligns multiple family sequences providing more biologically relevant information (Chowdhury & Garai, 2017). In macromolecular biological research, MSA is the starting point providing functional, evolutionary, and structural perspectives (Do & Katoh, 2009). Further sections will discuss MSA in detail. To identify the level of similarity and identity scoring matrices are used. In nucleotide alignment scoring, each identical base is assigned a positive score, while in protein alignment, identity/similarity score and amino acids having similar physicochemical properties are assigned positive scores (Stormo, 2009). Matrices used to identify these scores are Blocks Substitution Matrix (BLOSUM) (Mount, 2008a) and Point Accepted Mutation (PAM) (Mount, 2008b).

The traceback approach provides the optimum solution by the dynamic programming (DP) method (Eddy, 2002; Shyu, Sheneman, & Foster, 2004). Increasing the sequence number to more than two the biological relevance of

alignment results becomes low as considered in MSA (Notredame, 2002). Dynamic programming suffers from the dimensionality problem as each sequence corresponds to the number of dimensions. Moreover, the task becomes complex and requires more computational resources. It traces the path backwards in DP for optimum alignment; if more than one option is available, the complexity of tracing backwards exponentially increases. Hence, the usage of DP in MSA is less relevant in biological aspects. MSA follows the approximate heuristic approach that provides an optimal solution in a limited time. Available heuristic algorithms are approximate. Therefore there is a constant demand to develop MSA tools because of the growing size of databases; there is a need for high performable algorithms. Sequence alignment at the amino acid level is considered more relevant than nucleotide level as functional and/or structural information is carried by proteins, hence key functional biological molecules.

3.2 Sequence alignments: determining similarity and deducing homology

The regions of similarity in DNA/RNA or protein to access the structural, evolutionary, or functional information between the sequences are identified by sequence alignment.

3.2.1 Why construct sequence alignment?

Sequence alignment is used to carry out different types of sequence comparisons, searching, and annotation. For any study that involves several homologous sequences, the prerequisite will be the arrangement of homologous residues in columns to draw some inference (Wong & Tee, 2020). The biological relevance of sequence alignment lies in DNA/RNA or protein alignment, where conserved residues or residues with physiochemical relation are lined up to draw some functional importance of the unknown sequence/s (Popa, Oldenburg, & Ebenhöf, 2020). The relation between proteins and their family can be built up by sequence alignment strategy. A more confident conclusion about less similar protein/s can be made if conserved residues are present in the sequence despite being significantly less overall similar. In the homology modeling approach of proteins, alignment plays an important role. The target sequence with a certain amount of homology with X-ray/NMR-deduced tertiary structures can be used to predict the tertiary model of an unknown protein (Kelley & Sternberg, 2009).

Most related multigene family proteins have evolutionarily related domains from the same or different families. Protein families, especially multicellular organisms, are large enough to relate to many species having common domains. An evolutionary overview of these large protein families helps understand the subgrouping and functional diversity within the same or different organisms (Prjibelski, Korobeynikov, & Lapidus, 2019). The evolutionary relation is provided by the phylogenetic analysis, which is built based on sequence alignment. For the evolution of the complete genetic sequence, this becomes increasingly crucial.

3.2.2 Similarity of sequences

Comparison and detection of similarities between two or more biological sequences is the main idea of sequence alignment. The bioinformatics applications and predictions are based upon sequence similarity. To understand the similarity of sequences, let us take two small stretches of sequence with 11 bases. Matching symbols is the simplest way to attain similarity. The score thus obtained by matching the similarity of the symbols in sequences is known as the alignment score. The nonmatching symbols form the hamming distance, used to calculate the distance between the sequences, that is, the degree of dissimilarity (Bookstein, Kulyukin, & Raita, 2002) (Fig. 3.1A).

(A)
S1: ATGCATGCATG
S2: ATCGATGCGTG

FIGURE 3.1 (A) Two sequences with hamming distance 4. (B) Two sequences with edit distance of 3.

(B)
S1: ATGC-AGCATG
S2: ATGCATGCA-G

The above method cannot be considered an alignment method since this position by position alignment process does not consider the important biological events, that is, deletions and insertions. The classical manner of sequence alignment considers all substitutions, deletions, and insertions to calculate the edit distance (Fig. 3.1B). A few approaches for calculating the alignment score while considering all of the criteria were developed in the 1990s. The algorithm for local and global alignment was developed for sequence analysis (Needleman & Wunsch, 1970; Smith & Waterman, 1981). However, these algorithms, due to their running time and memory allocation, become impractical for DNA/RNA sequences. These algorithms will be discussed in detail in the upcoming sections.

3.2.3 Homology of sequences

Sequences having similar or identical functions, that is, homologous sequences, if aligned with the new sequence/s, the probable function of sequence/s can be assigned reliably. Sequence similarity becomes the base for homology identification. The significant statistical homology is identified by many advanced tools (Stormo, 2009). Once any new sequence is determined, primary aim becomes to assign a function to that sequence or multiple functions to the subsequences. Homologous sequences linked to a common ancestor generally have the same or related functions. Methods developed are based on finding the sequence similarity that does not occur by chance, to make a significant prediction for evolving from a common ancestor (Ma & Wang, 2014).

3.2.4 Global sequence alignment

Two ways for aligning the sequences can be: aligning the sequence globally, that is, global alignment, and locally, that is, local alignment. The local alignment will be discussed in the section below. The global alignment algorithm is developed based on the DP method by (Needleman & Wunsch, 1970), also known as the Needleman–Wunsch algorithm. The global alignment score is predicted for the full length of the sequences. Global alignment strategy best works for PSA, but some techniques of MSA also use this approach. In MSA, difficulty arises when homologous regions are locally present in the sequence and domains are shuffled in the length of sequences.

In global alignment, each residue is aligned, best fitted when query and target sequences are almost similar or of the same size. The EMBL-European Bioinformatics Institute (EMBL-EBI) server incorporates the online tool (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) to perform a global alignment. In advancement, EMBOSS Stretcher provides optimal global alignment by using a modification of DP.

3.2.5 Local sequence alignment

Sequences are aligned that are having the highest density of matches. Subalignments or small islands of matches are generated in the aligned sequences. Local alignment is suitable for sequences having similar small regions in their length having a different overall length. It implies that local alignment aligns the more similar sequence regions. The method is more suitable for nonsimilar sequences that are said to contain similar regions/similar motifs.

Local alignment also follows the DP approach developed by (Smith & Waterman, 1981) and is also known as the Smith–Waterman algorithm. Local sequence alignment takes into account similar local regions among the sequences. In contrast to global alignment, local regions are aligned to find the optimum score. When sequence lengths vary, local alignment produces optimum results. In the more extensive sequence, the dissimilar sequences consist of motifs in local regions; hence, local alignment is beneficial. The server at EMBL-EBI incorporates the online tool (https://www.ebi.ac.uk/Tools/psa/emboss_water/) to perform a local alignment. In advancement, EMBOSS Matcher provides optimal local similarities using Bill Pearson's align application.

Some methods work in combining both methods, that is, local and global, known as “glocal.” These search for the best partial alignment that is best possible (Brudno, Malde, & Poliakov, 2003). The combination works best when the downstream portion of one sequence is aligned with the upstream portion of the other. Here, individual global methods could stretch the sequence while local may not find the similarity. One similar situation arises when one sequence is concise, and the other is very large. In both situations, global and local alignment individually becomes impractical; hence, their combination serves the purpose.

To improve the speed and accuracy of the alignment tools, research is happening worldwide. Alignments conducted with manual algorithms (not automated) were more accurate than automated algorithms (Do & Katoh, 2009). Automated computational algorithms found their way in many sequences as manual alignments become impossible for more than three sequences (Botta & Negro, 2009). Both global and local alignments serve the purpose. MSA becomes

typical for protein, which is much more challenging as it consists of 20 amino acids compared to nucleotide, which has only four nucleotides (Huang, Shah, & Yao, 2019).

3.2.6 Working of alignment algorithm

The two most important algorithms used for aligning the sequences are the Needleman–Wunsch algorithm and the Smith–Waterman algorithm. These algorithms are based on the DP method, that is, they find the optimal solution by a more quantitative method considering matches and mismatches. The alignment problem is solved by dividing the problem into individual subproblems. The dynamic programming matrix is defined with three different steps.

3.2.6.1 Initialization of the matrix

For two sequences of length m and n , a matrix is defined of dimensions $m + 1$ and $n + 1$. The scores of match, mismatch, and the gap can be user-defined, provided the gap penalty should be negative or zero. A gap score is defined as a penalty given to alignment when there is insertion or deletion.

3.2.6.2 Matrix filling with a maximum score

This step is most crucial for the algorithm, and filling starts from the upper left-hand corner of the matrix. To find the maximum score of each cell, it is required to know the neighboring scores (diagonal, left, and right) of the current position. A match or mismatch (assumed) score is added to the diagonal value from the assumed values. Similarly, the gap score added to the other neighboring values. Thus three different values are obtained, take the maximum among them, and fill the i th and j th positions with the scores obtained.

The filling of the matrix can be shown in the following manner:

$$M_{i,j} = [M_{(i-1,j-1)} + S_{(i,j)}, M_{(i,j-1)} + w, M_{(i-1,j)} + w]$$

3.2.6.3 Traceback approach to find appropriate alignment

The final step in the algorithm is the traceback for the best alignment. Traceback is followed till the 0th position to obtain the appropriate alignment. The best alignment among the alignments can be identified by using the maximum alignment score.

3.3 Scoring matrices: construction and proper selection

Nowadays due to the expansion of genetic data at a high pace challenges the algorithms developed for sequence alignment. Need for the development of efficient and high performing tools always remain the center point of research. Optical computing approaches have shown promising implementation with the advancement, yet their applicability is the point of discussion. In this section, we will be discussing the matrices used in the sequence alignment processes.

3.3.1 Scoring matrices

The scoring matrix is weight matrices used in programs for searching sequences against databases. These weight matrices add sensitivity and selectivity to the search. In daily tasks, most users choose default matrices that are PAM250 or BLOSUM62 specified by the developers (Wheeler, 2003). Maximum of the users do not know why these matrices are used; they follow the options specified for default search, but the fact is that by varying the matrices overall outcome of the research may vary to a certain extent.

Developers took some assumptions in BLOSUM (Henikoff & Henikoff, 1992) and PAM (Margaret Dayhoff) (Dayhoff, Schwartz, & Orcutt, 1978) scoring matrices; understanding of these assumptions may change the choice of the outcome. If we talk of proteins, certain amino acids undergo substitution during the evolutionary process. However, they perform similar functions in different species as the substitutions made are compatible with the function and structure of proteins. More common and less common changes in proteins will assist in sequence alignment algorithms. Margaret Dayhoff pioneered mutation analysis to develop the PAM matrix (Dayhoff et al., 1978). By analyzing the ancestor relationships in similar proteins, the amino acid changes commonly can be assessed. The following section will discuss the choice of scoring matrices.

3.3.2 PAM matrices

The acronym stands for Point Accepted Mutation, developed by Margeret Dayhoff in 1978. It is a model developed on the probability of replacing amino acids in related proteins to the expected replacement frequency of the distant proteins. The basis of this algorithm lies in the understanding that some amino acids are replaced more frequently compared to others. These replacements are more common in related protein sequences rather than distant sequences. Charge, size, hydrophobicity, and other characteristics are conserved in the substituted amino acids. During evolution, the likelihood of substituting amino acids in related proteins is tracked by the PAM matrix. Despite being based on a small dataset, it is a reliable tool for sequence alignment and is the only scoring matrix to date based on sound evolutionary principles.

For homologous proteins that diverged less from each other in a short period are still similar up to 50% or more, matrix gives changes expected. In addition, it provides expected changes for sequences that remained only 20% similar over a long period. Predictions mentioned are used to get the optimal alignment score between two protein sequences, considering the assumption that substitutions that occurred over a short period can be extrapolated on longer duration evolutionary substitutions.

The assumption that amino acid change at position one is independent of previous changes at that position forms base for PAM matrix function (Dayhoff et al., 1978). Changes in protein sequence by amino acid substitution are viewed as a Markov model. As the amino acid substitutions considered in the matrix were from closely related proteins, they do not affect the protein's overall function. Therefore, amino acid substitution those accepted by natural selection are known as accepted mutations.

George, Barker, & Hunt (1990) discussed the derivation of PAM matrices. PAM computes the frequency of a single substitution per 100 amino acids known as PAM1. For higher PAM matrices, PAM1 is multiplied by a defined number to itself. Thus PAM 80 means that the PAM1 matrix is multiplied to itself 80 times and so on. Fifty substitutions in 100 amino acids are considered as PAM50 matrix, while 25 amino acid replacements for each site come under the PAM250 matrix. This may be due to the condition that one amino acid (suppose alanine) may be changed to second (suppose glycine) and third (suppose valine) and then again to first amino acid (i.e., alanine). These mutations, known as silent mutations, are derived from the families and superfamilies of the protein (Tomii & Kanehisa, 1996).

3.3.3 BLOSUM matrices

The acronym stands for BLOcks SUBstituion Matrix (BLOSUM) developed by Henikoff and Henikoff (1992). BLOSUM considers the blocks database developed by them, consisting of multiple alignments of conserved regions in protein families. It consists of an ungapped alignment of the protein sequence. A varying similarity may be present in different blocks. Unlike the PAM matrix derivation sequences represented multiple times were included in the calculation. Derivation of BLOSUM matrices avoids the highly related sequences and sequences occurring frequently. As derivation was formulated in the early 1990s, representation of nonglobular and hydrophobic proteins was high. This makes many identical sequences in the same blocks. These identikits may create bias in the alignment process. To avoid this bias, similar sequences in a block were clustered, and sequences present in the cluster count fractional in the selection. As a result, raising the percentage of distantly related sequences reduces the impact of related sequences. The BLOSUM matrix is called by the cluster threshold values, for example, BLOSUM80 has an 80 percent threshold value. Default used matrix BLOSUM62 (Clustering at 62%) reduces the contribution of blocks up to 25%. In BLOSUM62, the blocks data are reduced but still 1.25×10^6 amino acid pairs are included in the final matrix calculation. As the threshold of clustering decreases, the contribution of far-off related sequences increases in final matrix calculations (Mount, 2008a). BLOSUM62 is the most widely used matrix. This is because it functions well with 70% different proteins or will predict similarity for as low as 30% similar protein. BLOSUM62 is considered more accurate and superior to the PAM250 matrix. BLOSUM62 works well for dissimilar proteins or fragments of a larger protein. BLOSUM matrix is also useful for constructing a phylogenetic tree and determining the relationship with other proteins (Eddy, 2004).

3.3.4 PAM versus BLOSUM

DNA/RNA or protein alignments are influenced by choosing a scoring matrix that includes substitutions, matches, mismatches, insertions, and deletions (INDEL). In related proteins, the probability of changing one amino acid to another amino acid is utilized to calculate matches and mismatch scores. In homologous proteins, the likelihood of one amino

acid substituted by another during evolution is considered in PAM matrices; thus an evolutionary aspect of protein is investigated. On the other hand, BLOSUM looks into scoring substitutions over a series of evolutionary periods considering the threshold clustering value. The derivation of PAM and BLOSUM matrices differs significantly (Wheeler, 2003).

PAM matrix derivation assumes that mutational changes at one site are independent of changes that occurred previously at that site, that is, Markov process. Scoring is done for 85% similar sequences. The first changes in the protein during evolution differ from its ancestor; PAM is based on these changes (Mount, 2008a,b).

Contrary to PAM matrices, BLOSUM matrices do not consider the evolutionary model. Derivation occurs by observing the changes that occurred at aligned regions of related family proteins, and it does not consider the overall similarity of the proteins. These proteins share a common ancestry as they are related biochemically. The starburst model of protein evolution in which the distal versus closer relationships are not considered despite proteins of the exact origin is followed. In the case of the PAM matrix, scoring is based on the position of amino acid in related sequences, whereas scoring in BLOSUM is done considering the conserved positions and substitutions in blocks representing the common regions in proteins (Mount, 2008a).

3.4 Basic Local Alignment Search Tool

In bioinformatics, the Basic Local Alignment Search Tool (BLAST) algorithm compares biological sequences (DNA/RNA or protein) to the database present at NCBI (Fig. 3.2A). BLAST performs the comparison of sequences using local alignment. The approach used in BLAST is very successful as it focuses on local regions, and active sites are preserved in these regions. Statistically significant analysis conducted through a rigorous statistical approach. These statistical approaches take advantage and minimize the alignment time; it reduces the change to change the threshold at which chance similarity could occur. BLAST produces results much quicker than similar kinds of heuristic algorithms (Altschul, Gish, & Miller, 1990). BLAST takes input in FASTA format or Gen Bank format, and output is generated in various formats for user convenience, such as HTML, XML, and text (Fig. 3.2B). The recent addition of “SUM” statistics has increased the BLAST sensitivity; it adds the highest scores for multiple hits by sequence (Chen, Ye, Zhang, & Xu, 2015).

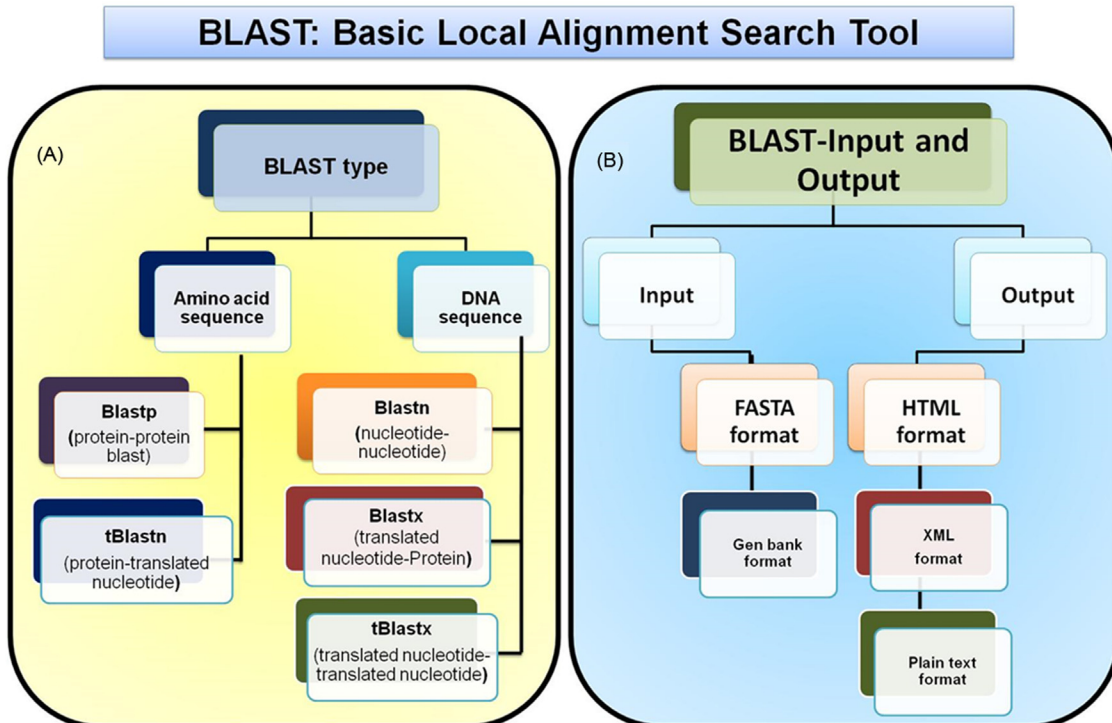


FIGURE 3.2 (A) Type of Basic Local Alignment Search Tool programs. (B) Input and output file formats for Basic Local Alignment Search Tool.

High-scoring segment pair (HSP) is the center point of BLAST output. Local maximum alignment at equal lengths of the two sequences according to the scoring system is known as HSP. BLAST standalone version on the command line provides the option of controlling the significance of HSPs and database sequences (Faradonbeh & Monjezi, 2017). BLAST is a successful example of the client-server framework, with the server at NCBI providing access to client programs on user machines. A service designated “experimental” has been offered since 1992 to allow users to search databases at the NCBI. This service transfers data between the client and the server as the BLAST report. BLAST-based on heuristic search tool has many application functional annotation (Chandra, Kushwaha, & Sangwan, 2020), orthology mapping (Benzekri, Armesto, & Cousin, 2014), gene prediction (Chandra et al., 2020), secondary and tertiary structure prediction (Lins, 2005), and genome variation analysis (Asano, Tanaka, & Yang, 2005).

BLAST algorithm undergoes three major steps.

3.4.1 Seeding step

In this step, query sequences are scanned through subject sequences to find the heuristic points for significant local alignments. A lookup table is build consisting of coordinates for query and subject sequences.

3.4.2 Ungapped extension step

Extensions are made to the seeds to make longer ungapped alignments. These ungapped alignments are verified for the location of interest, and the rest of the seeds are discarded. In this step, maximum of the seeds are discarded.

3.4.3 Gapped extension step

Ungapped alignments are now used in this step to create local gapped alignments for improving the score further. The gapped alignments that cross the expected value threshold (e-value) are considered for the final set.

Each of these steps takes 30% of the time. Major studies are ongoing to improve the speed of the BLAST. Improving the sensitivity of BLAST longer seeds could serve the task, but memory limitations would make it impractical with good running time (Shiryev, Papadopoulos, Schäffer, & Agarwala, 2007).

3.4.4 Types of BLAST

BLAST searches query sequences against subject sequences to produce local alignments. To make it user friendly and make it result oriented, BLAST has five types, as mentioned in Table 3.1.

Based on sequence input (nucleotide and/or protein) and subject database used (nucleotide and/or protein), the BLAST family consists of many versions (Fig. 3.3).

3.4.5 BLAT versus BLAST

BLAT (BLAST-like alignment tool) (Kent, 2002) is a new algorithm similar to BLAST. BLAST scans by taking into account HSPs. BLAST scans by indexing the query sequence against the database while BLAT scans by indexing the database. An extension is triggered in BLAST when hits occur close to each other; in BLAT, extension is triggered on perfect hits. Each area of homology is shown in BLAST between two sequences while they are taken together for larger

TABLE 3.1 Types of BLAST programs.

Program	Query	Subject/database	Default matrix
Blastn	Nucleotide	Nucleotide	–
Blastp	Protein	Protein	BLOSUM62
Blastx	Nucleotide (translated into six frames)	Protein	BLOSUM62
tblastn	Protein	Nucleotide (translated into six frames)	BLOSUM62
tblastx	Nucleotide (translated into six frames)	Nucleotide (translated into six frames)	–

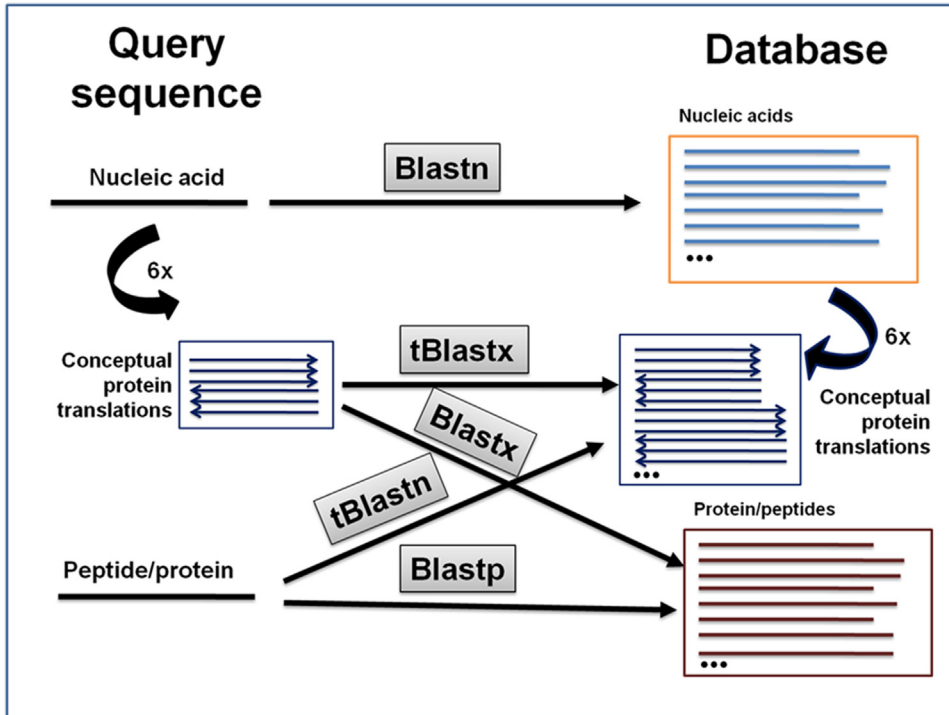


FIGURE 3.3 Basic Local Alignment Search Tool types with respect to the database and query type.

alignment in BLAT. BLAT unspliced mRNA on to genome providing alignment that uses mRNA bases only once positioning splice sites through its unique code. On the other hand, BLAST delivers exons sorted by size with an extended edge of each exon.

3.5 Multiple sequence alignment (MSA)

MSA is the most important technique used in molecular biology, computational biology, and bioinformatics (Shyu et al., 2004). Day by day, with advancements in NGS techniques, the data generation speed has reached another level. Databases are flooded with the data, with the generation of massive data. Therefore it is a challenge for MSA algorithm developers to keep pace with the massive amount of data generation and inculcate the data clusters present in the computing cloud (Daugelaite, O'Driscoll, & Sleator, 2013). MSA algorithms and cloud computing have increased the quality, speed, and reliability for handling many sequences. Constructing reliable MSA is a biologically complex task and computationally intense; to date, no MSA algorithm can be considered to produce biologically perfect results. This lacuna makes this field promising, active, and aiming to develop tools that can align thousands of lengthy sequences and produce reliable results in a reasonable time. Computational complexities along with speed are affected negatively by the increase in the sequence number. The growth of high-throughput sequencing technologies at an exponential rate can be understood by recent breakthroughs, such as Human Genome Project (Sawicki, Samara, Hurwitz, & Passaro, 1993), Genome 10K Project (Koepfli, Paten, Scientists, & Brien, 2015), and 1000 Genomes Project (Siva, 2008).

3.5.1 Need for MSA

MSA is a widely used technique in computational biology for biological sequence analysis, bioinformatics, and allied subjects (Popa et al., 2020). Either phylogenetic reconstruction or structure analysis initial step is the comparison of homologous proteins by MSA. MSA plays a vital role in new member identification of protein on comparing with similar sequences. Biologically good and accurate alignments show homology and relationships to deduce meaningful information. The accuracy of MSA plays a vital role as further analysis depends on the results of the MSA. Hence, the development of reliable and accurate MSA tools always remains a promising task.

TABLE 3.2 Common used tools of biological sequence alignment.

Name	Sequence type	Alignment type	Reference
BLAST	Both	Database search	Atschul (1991)
DIAMOND	Protein	Database search	Buchfink, Xie, and Huson (2015)
PSI-BLAST	Protein	Database search	Altschul (1997)
PSI-Search	Protein	Database search	Li, McWilliam, and Goujon (2012)
SWIMM	Protein	Database search	Rucci, García, and Botella (2015)
SWIMM2.0	Protein	Database search	Rucci, Garcia Sanchez, Botella, and Juan (2019)
BAlI-Phy	Both + codons	Global	Suchard and Redelings (2006)
Kalign	Both	Global	Madeira, Park, and Lee (2019)
MAFFT	Both	Local or global	Katoh, Rozewicki, and Yamada (2019)
MUSCLE	Both	Local or global	Edgar (2004)
T-Coffee	Both	Local or global	Notredame et al. (2000)
ClustalW	Both	Local or global	Thompson et al. (2003)

3.5.2 Multiple sequence alignment algorithm

The development of an accurate and reliable MSA algorithm is always a daunting task. Still, MSA algorithms are increasing, with approximately 1–2 algorithms published per month ([Naznooshadat, Elham, & Ali, 2020](#); [Phillips, Janies, & Wheeler, 2000](#)). These modern tools are improved with better computational complexity and accuracy in results. However, no tool/algorithm can be called biologically relevant to produce exact/accurate results ([Higgins, 2001](#)). Some of the most commonly used tools for MSA are Mafft ([Katoh & Standley, 2013](#)), T-Coffee ([Notredame, Higgins, & Heringa, 2000](#)), MUSCLE ([Edgar, 2004](#)), and Kalign ([Lassmann & Sonnhammer, 2005](#)) ([Table 3.2](#)). The Clustal family of tools is most accurate and used widely, such as ClustalW, ClustalV, and Clustal Omega ([Chenna, Sugawara, & Koike, 2003](#)). Clustal family tools are also bundled in the commercial tools as a black box to perform the MSA, such as DNASTAR and MEGA ([Table 3.2](#)). Let us discuss the most common and accurate tool used for MSA studies, that is, ClustalW ([Thompson, Gibson, & Higgins, 2003](#)).

3.5.3 ClustalW

Clustal has series of MSA tools developed initially by Des Higgins. Series comprises of many tools, such as Clustal, ClustalW, ClustalV, Clustal Ω , ClustalX, and Clustal2. ClustalW, introduced by [Thompson et al. \(2003\)](#), is the most used MSA tool due to improved alignment, speed, and sensitivity than other available tools. ClustalW algorithm incorporates a weighing scheme and position-specific scoring scheme for downweighing the overrepresented group of sequences. In the ClustalW algorithm, the first step involves the pairwise alignment of sequences by Wilbur and Lipman's k-tuple method for nucleotide or protein sequences or DP-based Needleman and Wunsch algorithm if sequences are less ([Wilbur & Lipman, 1983](#)). The second step is to guide tree generation using the neighbor-joining (NJ) method (discussed in phylogenetic [Section 3.6](#)) using the distance table. Third step MSA generation based on the guide tree constructed in the previous step by using a progressive alignment approach ([Fig. 3.4](#)). The following subsections will deal with each of these sections in detail.

3.5.3.1 Pairwise alignment

For pairwise aligning the sequences, the “best guess” method known as the k-tuple method is used if the number of sequences is large. k-tuple similarity matches, that is, 2–4 for nucleotide and 1–2 for proteins used to calculate the alignment score ([Clark & Kalita, 2014](#)). To calculate the score, matches are divided by the paired residue similarity score sum. From similarity score, penalties for gaps are subtracted, and this score is divided by 100 to obtain distance score and to get differences per site, it is subtracted from 1.0. A Hash table is made by locating all the k-tuples of two

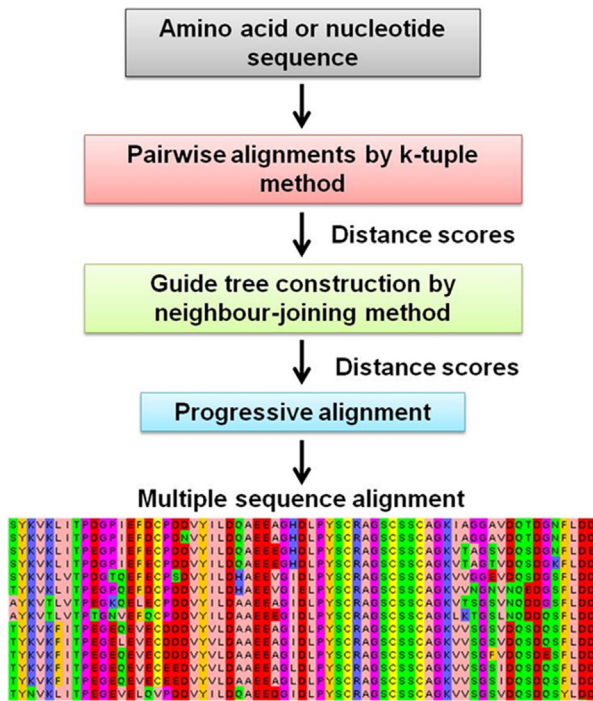


FIGURE 3.4 Workflow of ClustalW showing three steps: (1) k-tuple, (2) guide tree generation, and (3) progressive alignment.

sequences. In the dot matrix plot, the dot represents the k-tuple match. For a particular window size, the matches found diagonally are marked. These diagonals are considered as the similarity regions between two sequences. Like the Needleman–Wunsch method, the k-tuple method matches the dots for a particular window size, and the score obtained is the highest score. The gap penalties are subtracted from the score to get the final score (Tamulevičius, Juodėnas, & Klinavičius, 2018).

3.5.3.2 Guide tree construction

The NJ method is used for building the tree by ClustalW. NJ algorithm gives priority to nodes rather than taxa or their cluster. Matrix is prepared by using the scores from the previous k-tuple method. The matrix consists of values for nodes which are adjusted according to the average divergence from all nodes. The guide tree is built by using these matrix values. The terminal nodes are removed from the tree to add linked nodes to the common ancestor node. NJ method produces the unrooted tree with proportional divergence along the branch. On the equal distance of branches, the root is placed (Daugelaite et al., 2013).

3.5.3.3 Progressive alignment

As the name suggests, this method progressively adds successively less related sequences to the most similar sequences during alignment. This step continues until all the sequences in the query are incorporated. In ClustalW, pairwise alignment is done according to the branching order generated in the NJ method guide tree. The procedure moves toward the root, starting at the tip. DP algorithm is used at each step with BLOSUM as a weighted matrix with opening and extending penalties (Daugelaite et al., 2013).

3.6 Phylogenetic analysis

The study of relatedness among species, populations, and a group of organisms is known as phylogenetic analysis. In other words, it is the study of the evolutionary relationship of the species, that is, to determine how a family has evolved. Phylogenetic analysis is understood in the light of MSA (Patwardhan, Ray, & Roy, 2014). The alignment methods that minimize the nonhomology are favored. Hence, the alignment method that satisfies the phylogenetic optimality criterion should be considered. The treatment of alignment gaps as penalties *viz.* transversion, transition, insertion, and/or deletion for phylogenetic reconstruction (Kapli, Yang, & Telford, 2020).

3.6.1 Types of trees

The arrangement of biological sequences among various species based on relatedness in two-dimensional plots is a phylogenetic tree. It is a drawing consisting of lines that represent the evolutionary distance. A phylogenetic tree can be represented in three forms: Cladogram, Phylogram, and Dendrogram (Roy, Dasgupta, & Bagchi, 2014).

3.6.2 Algorithms for phylogenetic analysis

Algorithms are designed either by character-based methods or distance-based methods. Distance-based methods are unweighted paired group method with arithmetic mean (UPGMA) (Gronau & Moran, 2007), NJ (Saitou & Nei, 1987), Fitch-Margoliash method (FM) (Fitch & Margoliash, 1967), and minimum evolution method (ME) (Rzhetsky & Nei, 1993). The character-based method provides the actual data pattern optimization. These methods are the maximum likelihood (ML) method (Felsenstein, 1981) and the maximum parsimony (MP) method. To compare the tree-building methods, important criteria should be considered: consistency of estimated topology, computational speed, frequency of obtaining the right topology, phylogenetic trees statistical consistency, and reliability of estimated branch length. These parameters of tree-building methods depend upon the algorithms used for their development. Based on the computation speed, NJ method is superior to all others. Bootstrap tests incorporated in these methods allow them to handle large sequences with ease. On the other hand, tools, such as MP, ML, and ME, consider all the topologies. In large datasets, with the increased number of sequences, topologies also increase; hence, it becomes impractical to use these methods. The need for algorithms that may be easy and simple for these methods is a matter of research. ME and NJ methods are consistent if no bias is applied; in contrast, MP method results were inconsistent. ML method is flexible in choosing the evolutionary model but comes with the drawback of time-consuming and lengthy (Patwardhan et al., 2014).

3.6.3 Terminology of phylogenetic tree

Phylogenetic trees are of different forms may be inverted, oriented sideways, curved radial, or circular. These are how trees are represented; besides these, the tree infers the evolutionary ancestry and divergence pattern. The tree consists of nodes, either terminal or internal, and branches; taxa are represented on the terminals (Fig. 3.5).

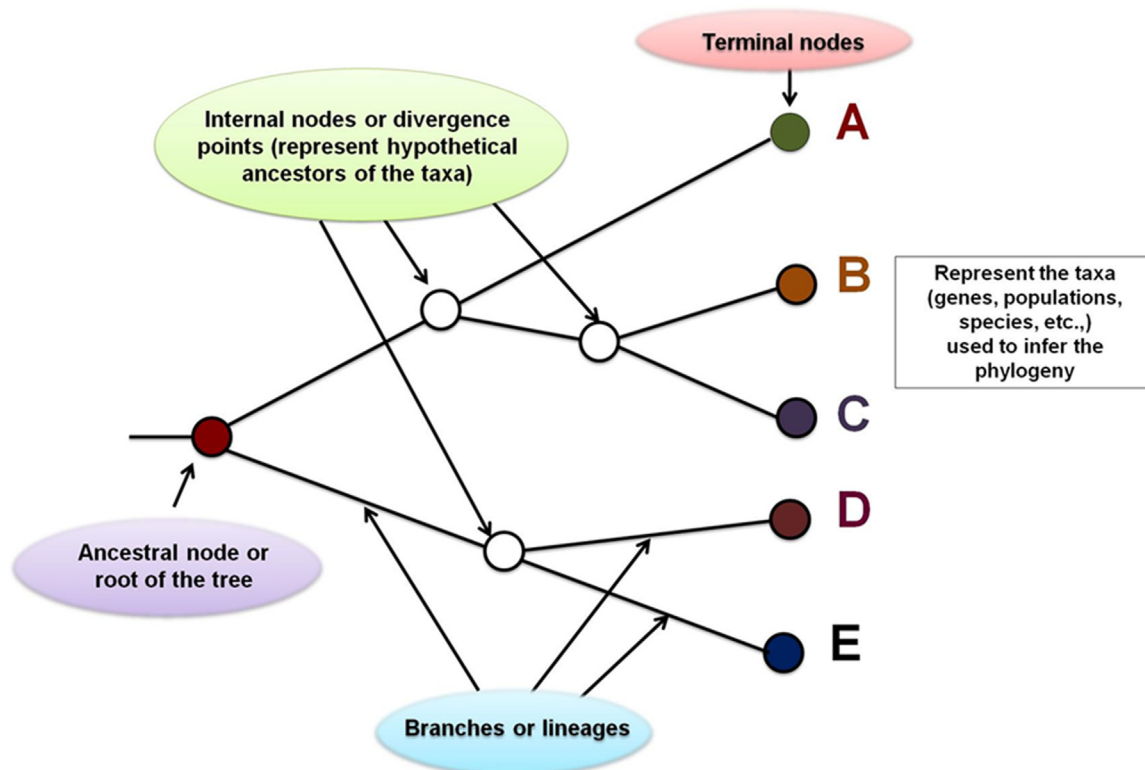


FIGURE 3.5 Common phylogenetic tree representation.

In determining the significant substitutions of amino acids, the ancestral information related to protein/peptide leading to an alteration in function during evolution plays a crucial role. In bioinformatics, MSA for predicting ancestral protein sequences is a proper technique, not a simple process. In evolutionary sequence analysis, these predictions are used to map substitutions to a particular lineage. Also, more functional representative sequences are generated than obtained by an alignment. The alignment precision and phylogeny analysis decide the correct prediction of ancestral protein. Algorithms based on MP or ML are many, yet they have limited implementation of recent advancements, such as sequence fragmentation and INDEL events. Overrepresentation of related sequences in large clades is also a significant setback in strict consensus methods. The contribution of large clades in sequences leads to the increased number of sequences in consensus. This problem is prevailed over by the MP method. The MP method disregards the agreement with terminal sequences; instead, it minimizes the mutational steps. However, some parsimonious predictions are not distinguished by this method. FASTML and CODEML give an accurate balance between precision and speed (Patwardhan et al., 2014). The phylogenetic analysis is an essential tool as it addresses the various biological queries, such as the spread and origin of viral infection relationships among genes or species, evolutionary patterns, migration patterns and demographic changes of species. The methods of phylogenetic analysis, such as likelihood, parsimony, distance, and Bayesian, have different weaknesses and strengths.

Albeit distance phylogeny methods are helpful for large data sets of highly similar sequences, while Bayesian and likelihood methods are specially used for inferring deep phylogenies, and these methods are often more robust and powerful. An accurate phylogenetic tree analysis is fundamental for various biological events and studies that reinforce understanding the major events and transitions during species evolution. These events or transitions include the origin of new genes, new body plans or metabolism, molecular adaptation, demographic changes, the emergence of several morphological adaptations, and species evolution in recently diverged species. The ML tree, MP tree, and NJ tree are desirable to determine the confidence in the point estimate for the true phylogeny.

Bootstrapping is the most common method for this purpose, which is applied to assess confidence in estimated ML tree, parsimony tree, and NJ methods. This method generates the same size pseudo-datasets. Resampling with alignment sites leads to the generation of original sites. These pseudo-datasets are considered original datasets and analyzed. Based on the bootstrap datasets, the values are appended for clades, that is, frequency at which it is recovered, followed by a phylogenetic tree. In other statistical applications, the bootstrap methods are not present. The accuracy of trees obtained by least-squares, NJ, and minimum evolution methods are done by available solid statistical tests. However, the methods for parsimony and likelihood trees still need refinement (Kapli et al., 2020).

3.7 Application of sequence alignment

In the modern world of data and high-throughput, sequencing alignment plays a crucial role and becomes indispensable for biological research. A few of the applications are mentioned below:

- *Genomics*: In both personal and comparative genomics studies, genomics determine the sequence conservation and variance of biological sequence alignment.
- *Proteomics*: Proteomics study the similarity between the distinct protein sequences and assigning the function to the unknown.
- *Transcriptomics*: Transcriptomics uses biological sequence alignment to examine gene expression, de novo analysis, and other topics.
- *Structure prediction*: MSA gives optimal homolog structure for predicting 3D structure.
- *DNA regulatory elements*: Regulatory alignment, such as binding sites, can be traced using the alignment data.
- *Pattern identification*: To identify the functional regions by analyzing the conserved regions or sites.
- *Evolutionary studies*: Tracing the evolution of origin using a phylogenetic tree is a common strategy based on the sequence alignment.

3.8 Conclusion

If one has obtained a novel sequence, then assigning a function to that sequence is essential. This first strategy is to identify the homologous sequences. For identifying the related sequences, biological sequence alignment plays a crucial role. The chapter discusses the various methods and algorithms for biological sequence alignment. Alignment can be pairwise for two sequences or MSA for more than two sequences. Alignment scores, including the INDEL and penalties, are calculated to find the relatedness of the sequences. Tools, such as BLAST and FASTA, play a crucial role in

sequence analysis. For evolutionary analysis, various character-based and distance-based methods are discussed. A combined approach of biological sequence alignment, phylogenetic analysis is most effective in identifying homologous sequences and evolutionary analysis.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Altschul, S. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*, 3389–3402.
- Altschul, S. F., Gish, W., Miller, W., et al. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410.
- Asano, T., Tanaka, N., Yang, G., et al. (2005). Genome-wide identification of the rice calcium-dependent protein kinase and its closely related kinase gene families: Comprehensive analysis of the CDPKs gene family in rice. *Plant & Cell Physiology*, *46*, 356–366.
- Atschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *Molecular Biology*, *219*, 555–665.
- Benzekri, H., Armesto, P., Cousin, X., et al. (2014). De novo assembly, characterization and functional annotation of *Senegalese sole (Solea senegalensis)* and common sole (*Solea solea*) transcriptomes: Integration in a database and design of a microarray. *BMC Genomics*, *15*, 952.
- Bookstein, A., Kulyukin, V. A., & Raita, T. (2002). Generalized hamming distance. *Information Retrieval*, *5*, 353–375.
- Botta, M., & Negro, G. (2009). Multiple sequence alignment with genetic algorithms. In *International meeting on computational intelligence methods for bioinformatics and biostatistics* (pp. 206–214). Springer.
- Brudno, M., Malde, S., Poliakov, A., et al. (2003). Glocal alignment: Finding rearrangements during alignment. *Bioinformatics (Oxford, England)*, *19*, i54–i62.
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*, 59–60.
- Carroll, S. B., Grenier, J. K., & Weatherbee, S. D. (2013). *From DNA to diversity: Molecular genetics and the evolution of animal design*. John Wiley & Sons.
- Chandra, M., Kushwaha, S., & Sangwan, N. S. (2020). Comparative transcriptome analysis to identify putative genes related to trichome development in *Ocimum* species. *Molecular Biology Reports*, *49*(9), 6587–6598.
- Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High-speed BLASTN: An accelerated MegaBLAST search tool. *Nucleic Acids Research*, *43*, 7762–7768.
- Chenna, R., Sugawara, H., Koike, T., et al. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, *31*, 3497–3500.
- Chowdhury, B., & Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, *109*, 419–431. Available from <https://doi.org/10.1016/j.ygeno.2017.06.007>.
- Clark, C., & Kalita, J. (2014). A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics (Oxford, England)*, *30*, 2351–2359.
- Daugelaite, J., O'Driscoll, A., & Sleator, R. D. (2013). An overview of multiple sequence alignments and cloud computing in bioinformatics. *ISRN Biomathematics*, *2013*, 1–14.
- Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). 22 a model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, *5*, 345–352.
- Do, C. B., & Katoh, K. (2009). Protein multiple sequence alignment. In J. D. Thompson, C. Schaeffer-Reiss, & M. Ueffing (Eds.), *Functional proteomics. Methods in molecular biology* (pp. 379–413). Totowa, NJ: Humana Press.
- Eddy, S. R. (2002). A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, *3*, 18.
- Eddy, S. R. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*, *22*, 1035–1036.
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797.
- Faradonbeh, R. S., & Monjezi, M. (2017). Prediction and minimization of blast-induced ground vibration using two robust meta-heuristic algorithms. *Engineering with Computers*, *33*, 835–851.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*, 368–376.
- Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, *155*, 279–284.
- George, D. G., Barker, W. C., & Hunt, L. T. (1990). Mutation data matrix and its uses. In R. F. Doolittle (Ed.), *Methods in enzymology* (183, pp. 333–351). Academic Press.
- Gronau, I., & Moran, S. (2007). Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters*, *104*, 205–210.
- Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*, 10915–10919.
- Higgins, D. (2001). Alignment problem. In S. Brenner, & J. H. Miller (Eds.), *Encyclopedia of genetics* (pp. 29–35). New York: Elsevier.
- Huang, M., Shah, N. D., & Yao, L. (2019). Evaluating global and local sequence alignment methods for comparing patient medical records. *BMC Medical Informatics and Decision Making*, *19*, 263.

- Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21, 428–444.
- Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, 20, 1160–1166.
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30, 772–780.
- Kelley, L. A., & Sternberg, M. J. E. (2009). Protein structure prediction on the web: A case study using the phyre server. *Nature Protocols*, 4, 363–373.
- Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12, 656–664.
- Koepfli, K.-P., Paten, B., Scientists, G. 10K. C. of, O., & Brien, S. J. (2015). The Genome 10K Project: A way forward. *Annual Review of Animal Biosciences*, 3, 57–111.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357.
- Lassmann, T., & Sonnhammer, E. L. L. (2005). Kalign—An accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6, 1–9.
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, England)*, 26, 589–595.
- Li, W.-H., & Graur, D. (1991). *Fundamentals of molecular evolution*. Sinauer Associates.
- Li, W., McWilliam, H., Goujon, M., et al. (2012). PSI-Search: Iterative HOE-reduced profile SSEARCH searching. *Bioinformatics (Oxford, England)*, 28, 1650–1651.
- Lins, R. (2005). Homology modeling functional genomics. *Proteins*, 8, 1–18.
- Liu, L., Li, Y., Li, S., et al. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, 12, 1–11.
- Ma, J., & Wang, S. (2014). *Algorithms, applications, and challenges of protein structure alignment* (1st ed.). Elsevier Inc.
- Madeira, F., Park, Y., Lee, J., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, 47, W636–W641.
- Mount, D. W. (2004). *Bioinformatics: Sequence and genome analysis* (2nd ed.). Cold Spring Harbor: Cold Spring Harbor Laboratory Press.
- Mount, D. W. (2008a). Using BLOSUM in sequence alignments. *Cold Spring Harbor Protocols*.
- Mount, D. W. (2008b). Using PAM matrices in sequence alignments. *Cold Spring Harbor Protocols*.
- Naznooshadat, E., Elham, P., & Ali, S. Z. (2020). FAME: Fast and memory efficient multiple sequences alignment tool through compatible chain of roots. *Bioinformatics (Oxford, England)*, 36, 3662–3668.
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48, 443–453.
- Notredame, C. (2002). Recent progress in multiple sequence alignment: A survey. *Pharmacogenomics*, 3, 131–144.
- Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302, 205–217.
- Patwardhan, A., Ray, S., & Roy, A. (2014). Molecular markers in phylogenetic studies—A review. *Journal of Phylogenetics and Evolutionary Biology*.
- Phillips, A., Janies, D., & Wheeler, W. (2000). Multiple sequence alignment in phylogenetic analysis. *Molecular Phylogenetics and Evolution*, 16, 317–330.
- Popa, O., Oldenburg, E., & Ebenhöf, O. (2020). From sequence to information. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 375, 20190448.
- Prijbelski, A. D., Korobeynikov, A. I., & Lapidus, A. L. (2019). Sequence analysis. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (pp. 292–322). Oxford: Elsevier.
- Roy, S. S., Dasgupta, R., & Bagchi, A. (2014). A review on phylogenetic analysis: A journey through modern era. *Computational Molecular Bioscience*, 04, 39–45.
- Rucci, E., García, C., Botella, G., et al. (2015). An energy-aware performance analysis of SWIMM: Smith-Waterman implementation on Intel’s Multicore and Manycore architectures. *Concurrency and Computation Practice and Experience*, 27, 5517–5537.
- Rucci, E., Garcia Sanchez, C., Botella, Juan, G., et al. (2019). SWIMM 2.0: Enhanced Smith-Waterman on Intel’s multicore and manycore architectures based on AVX-512 vector extensions. *International Journal of Parallel Programming*, 47, 296–316.
- Rzhetsky, A., & Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10, 1073–1095.
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4, 406–425.
- Sawicki, M. P., Samara, G., Hurwitz, M., & Passaro, E., Jr (1993). Human genome project. *American Journal of Surgery*, 165, 258–264.
- Shirye, S. A., Papadopoulos, J. S., Schäffer, A. A., & Agarwala, R. (2007). Improved BLAST searches using longer words for protein seeding. *Bioinformatics (Oxford, England)*, 23, 2949–2951.
- Shyu, C., Sheneman, L., & Foster, J. A. (2004). Multiple sequence alignment with evolutionary computation. *Genetic Programming and Evolvable Machines*, 5, 121–144.
- Siva, N. (2008). 1000 Genomes Project.
- Smith, T., & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147, 195–197.
- Stormo, G. D. (2009). An introduction to sequence similarity (“homology”) searching. *Current Protocols in Bioinformatics*, 27, 1–7.

- Suchard, M. A., & Redelings, B. D. (2006). BAli-Phy: Simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*.
- Tamulevičius, T., Juodėnas, M., Klinavičius, T., et al. (2018). Dot-matrix hologram rendering algorithm and its validation through direct laser interference patterning. *Scientific Reports*, 8, 1–11.
- Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, 2–3.
- Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives. *PLoS One*, 6, e18093.
- Tomii, K., & Kanehisa, M. (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engineering, Design & Selection*, 9, 27–36. Available from <https://doi.org/10.1093/protein/9.1.27>.
- Wheeler, D. (2003). Selecting the right protein-scoring matrix. *Current Protocols in Bioinformatics*, 3.5.1–3.5.6.
- Wilbur, W. J., & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences of the United States of America*, 80, 726–730.
- Wong, T. S., & Tee, K. L. (2020). *Sequence analysis. A practical guide to protein engineering* (pp. 11–27).
- Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, 26, 873–881.