Concepts in Structural Biology & Bioinformatics.

JSBB .. ISSN 2830-8832

The journal of **Concepts in Structural Biology & Bioinformatics** (JSBB) is a unique open access scientific journal in that it publishes two types of articles; **Concept articles** which are of medium level of scientific-concepts content dedicated for providing young researchers and postgrad students with knowledge and original **Peer-reviewed articles** in current edge fields of science relevant to Bioinformatics, Structural Biology, Genomics, Proteomics, Biochemistry, Microbiology, Biotechnology, Bio-AI*, Biomathematics, Nutrition & Health and ultimately to Biology themes and applications.

* Artificial Intelligence

JSBB is a monthly journal that strives to publish one (1) peer-reviewed paper and four (4) concept articles each issue.

UNIVERSITY of SAIDA Dr MOULAY TAHAR

SUBMIT ARTICLES

jsbb@univ-saida.dz

Concept article: *Minireviews*

Sequence Alignment in Bioinformatics

RACHEDI Abdelkrim

Laboratory of Biotoxicology, Pharmacognosy and biological valorisation of plants, Faculty of Sciences, Department of Biology, University of Saida - Dr Moulay Tahar, 20100 Saida, Algeria.

E. mail: abdelkrim.rachedi@univ-saida.dz

Published: 21 March 2023

Abstract

Sequence alignment is a crucial tool in bioinformatics used to compare and analyse genetic sequences such as DNA and protein sequences. This technique involves aligning two or more sequences to identify similarities and differences, which can provide valuable information on the evolutionary relationship between species, identify disease-causing mutations, and design new drugs. This article provides a historical background of sequence alignment, its definition, and various algorithms, percentage and scoring matrices methods, online bioinformatics tools, and the motives for its use in biology, biotechnology, mutations discovery and disease fighting.

Key words

Primary structure, Sequence alignment, Bioinformatics, Phylogenetic trees, Mutation, PAM, BLOSUM, Biological Function

Popular Articles

01 Algerian Complete Genome sequences of SARS-CoV-2 strains detected in Blida province.
Jun 23, 2020 • 8 min read ★

02 التركيب الفراغي ثلاثي_الأبعاد لإنزيم "بروتياز" من فيروس كورونا الجديد
Feb 07, 2020 • 3 min read ★

03 Possible Molecular Basis behihind the late surge of COVID-19 in Algeria.
Dec 25, 2021 • 5 min read ★

04 FIMA-v.2; updated database version with Data Integration features
Mar 03, 2021 • 3 min read ★

05 Teixobactin an antibiotic from soil bacteria for fighting Multiple Antibiotic

Download the PDF version

Introduction

The advent of sequencing techniques has brought about a revolution in biology by allowing scientists to obtain and analyse genetic information on an unprecedented scale. However, sequencing a gene or protein alone does not provide enough information to fully understand its function or evolution.

Sequence alignment has a long history dating back to the 1960s when Margaret Dayhoff introduced the concept of protein sequence alignment (Dayhoff et al. ,1966). Dayhoff was a pioneer in the field of bioinformatics and developed the first comprehensive database of protein sequences, which led to the creation of the first scoring matrix for sequence alignment, the PAM matrix (Durbin et al., 1998). Since then, numerous algorithms have been developed to compare sequences, including dynamic programming, heuristic methods, and probabilistic methods.

Sequence alignment is a fundamental tool in bioinformatics used to compare and analyse genetic sequences such as DNA and protein sequences. This technique involves aligning two or more sequences to identify similarities and differences, which can provide valuable information on the evolutionary relationship between species, identify disease-causing mutations, and design new drugs.

This article provides a historical background of sequence alignment, its definition, and various algorithms, percentage and scoring matrices methods, online bioinformatics tools, and the motives for its use in biology, biotechnology, mutations discovery, and disease fighting.

There are two main types of sequence alignment: pairwise alignment and multiple sequence alignment. Pairwise alignment involves comparing two sequences, while multiple sequence alignment involves comparing more than two sequences. Dynamic programming, heuristic methods, and probabilistic methods are commonly used algorithms for sequence alignment. Percentage and scoring matrices methods are used to evaluate the quality of an alignment. The most widely used scoring matrices are BLOSUM and PAM matrices.

Bioinformatics tools, such as Basic Local Alignment Search Tool (BLAST, Altschul et al.,1990) and ClustalW, are widely used for sequence alignment. Sequence alignment has a wide range of applications in biology, biotechnology, and disease diagnosis. By aligning sequences, scientists can create phylogenetic tree to help infer evolutionary relationships, identify disease-causing mutations, design new drugs, and analyse the structure and function of proteins.

Methods

Algorithms:

Sequence alignment can be performed using different algorithms, each with its advantages and

Resistance phenomena
Feb 20, 2022 • 3 min read ★

06 The Omicron Wave - Why high transmissibility & vaccine's Booster Dose necessity?
Dec 26, 2021 • 3 min read ★

disadvantages. The most common algorithms include pairwise alignment, multiple sequence alignment, and global and local alignment.

Pairwise alignment compares two sequences to identify regions of similarity and difference. The algorithm is based on the dynamic programming method, which calculates the optimal alignment score by assigning scores to each match, mismatch, and gap. The Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) and the Smith-Waterman algorithm (Smith & Waterman, 1981) are two commonly used dynamic programming algorithms, Figure 1.

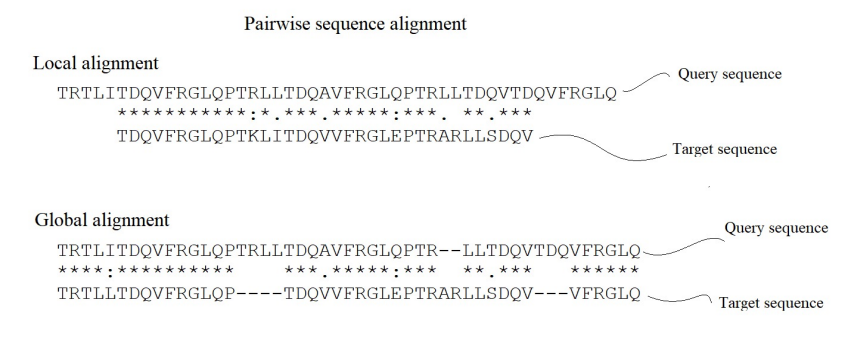


Figure 1. Pairwise sequence alignment in which two sequences are aligned. Types of algnment are highlighted; Local and Global.

Multiple sequence alignment, on the other hand, aligns more than two sequences and allows researchers to identify conserved regions among them. The algorithm is based on the progressive alignment method, which aligns sequences pairwise and then combines them into a multiple sequence alignment. The ClustalW and Clustal Omega algorithms are widely used for multiple sequence alignment, Figure 2.

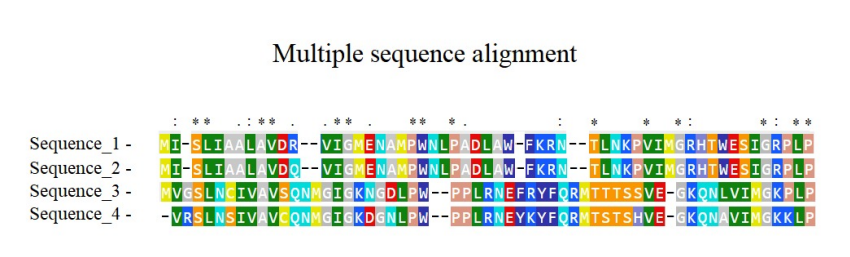


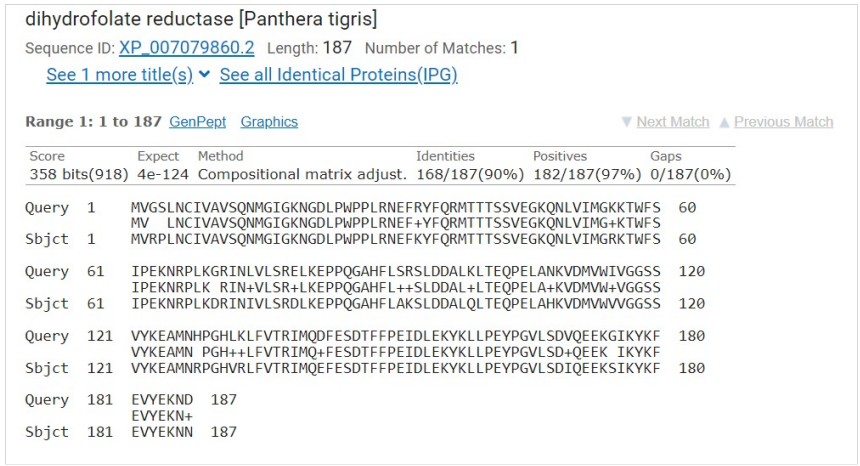
Figure 2. Multiple sequence alignment in which more than a pair of sequences are aligned.
This type of algnment exposes the positions and variations of amino acids or nucliec acids substitution between a number of seqeunces of proteins or genes. This helps to discover conserved regions in different species amongst other important conclusions.

Global alignment aligns the entire sequence, while local alignment aligns only the most similar regions. The global alignment algorithm is used to compare two sequences with similar lengths and regions of high similarity. The local alignment algorithm is used to identify similar regions in sequences with different lengths, see also Figure 1.

Percentage and Scoring Matrices Methods:

Scoring matrices are used to calculate the similarity between sequences (Durbin et al., 1998). The most widely used scoring matrices are the PAM and BLOSUM matrices. The PAM matrix was developed by Margaret Dayhoff in the 1970s and is based on the evolutionary distance between protein sequences. The BLOSUM matrix, on the other hand, is based on the observation that highly conserved regions in proteins are more functionally important than less conserved regions.

Percentage identity is a measure of the similarity between two sequences, expressed as a percentage of identical residues in the aligned sequences. It is calculated by dividing the number of identical residues by the total number of residues in the aligned sequences, Figure 3.


dihydrofolate reductase [Panthera tigris]

Sequence ID: XP_007079860.2 Length: 187 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 187 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
358 bits(918)	4e-124	Compositional matrix adjust.	168/187(90%)	182/187(97%)	0/187(0%)

```
Query  1    MVGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQNLVIMGKKTWFS  60
            MV  LNCIVAVSQNMGIGKNGDLPWPPLRNEF+YFQRMTTTSSVEGKQNLVIMG+KTWFS
Sbjct  1    MVRPLNCIVAVSQNMGIGKNGDLPWPPLRNEFKYFQRMTTTSSVEGKQNLVIMGRKTWFS  60

Query  61   IPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSS  120
            IPEKNRPLK RIN+VLSR+LKEPPQGAHFL++SLDDAL+LTEQPELA+KVDMVW+VGGSS
Sbjct  61   IPEKNRPLKDRINIVLSRDLKEPPQGAHFLAKSLDDALQLTEQPELAHKVDMVWVVGGSS  120

Query  121  VYKEAMNHPGHLKLFVTRIMQDFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKGIKYKF  180
            VYKEAMN PGH++LFVTRIMQ+FESDTFFPEIDLEKYKLLPEYPGVLSD+QEEK IKYKF
Sbjct  121  VYKEAMNRPGHVRLFVTRIMQEFESDTFFPEIDLEKYKLLPEYPGVLSDIQEEKSIKYKF  180

Query  181  EVYEKND  187
            EVYEKN+
Sbjct  181  EVYEKNN  187
```

Figure 3. BLAST alignment output showing a number of alignment evaluation terms including identifies and postives percentage and score value based on BLOSUM scoring matrix.

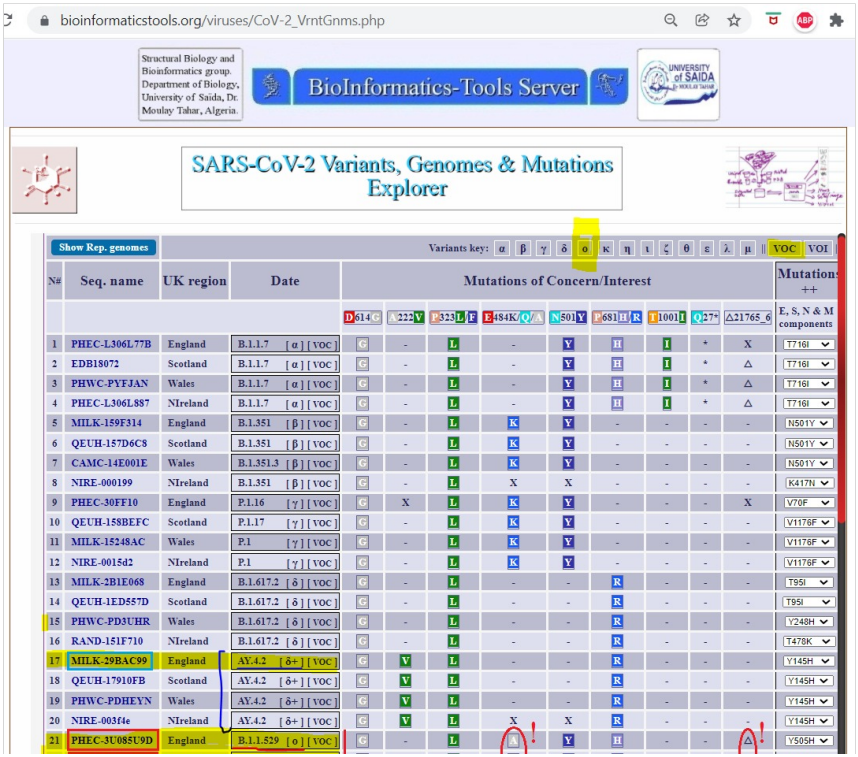
Motives for Biology, Biotechnology, and Disease Fighting:

Sequence alignment has numerous applications in biology, biotechnology, and disease fighting. In biology, sequence alignment is used to study the evolutionary history of species and identify conserved regions in proteins that may have functional significance. In biotechnology, sequence alignment is used to design new drugs, enzymes, and vaccines based on conserved regions in proteins. In disease fighting, sequence alignment is used to diagnose genetic diseases and identify potential targets for drug development.

Looking at the recent COVID-19 pandemic, the case of SARS-CoV-2, the virus responsible for the disease, sequence alignment has played a critical role in both understanding the virus and developing vaccines and antiviral drugs.

Firstly, sequence alignment has helped researchers to identify the genetic makeup of SARS-CoV-2 and compare it to other related viruses, such as SARS-CoV-1 and MERS-CoV. By aligning the genetic sequences of these viruses, researchers have been able to identify similarities and differences between them, which has provided insights into the origins, transmission, and virulence of SARS-CoV-2 (Grubaugh et. al., 2020).

Furthermore, sequence alignment has been used to identify the specific proteins and epitopes (i.e., small protein fragments that can elicit an immune response) that are most important for developing effective vaccines and antiviral drugs against SARS-CoV-2 (Walls et. al., 2020). For example, by aligning the sequences of the SARS-CoV-2 spike protein with those of related viruses, researchers have been able to identify mutations in the regions of the spike protein that are most likely to elicit an immune response, and these regions have been targeted in the development of several COVID-19 vaccines (Rachedi, 2020), Figure 4.



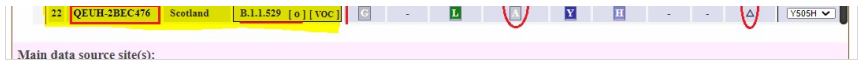


Figure 4. Table highlighting a number of mutations in the Spike gene that distincted the different variants of the SARS-CoV-2 responsible for the COVID-19 disearse

See: https://bioinformaticstools.org/viruses/CoV-2_VrntGnms.php

In addition to vaccine development (Callaway, 2020), sequence alignment has also been used to identify potential drug targets for SARS-CoV-2 (Zhou et. al., 2020). For example, by aligning the genetic sequences of SARS-CoV-2 with those of related viruses, researchers have been able to identify conserved regions of the virus that may be vulnerable to targeting by antiviral drugs. One example is the main protease (Mpro) enzyme, which is essential for the replication of SARS-CoV-2 and is highly conserved among coronaviruses. By aligning the Mpro sequences from different coronaviruses, researchers have been able to identify potential drug candidates that can inhibit Mpro and block viral replication.

Phylogenetic Trees:

Sequence alignment plays a crucial role in creating phylogenetic trees, which are diagrams that illustrate the evolutionary relationships among different species or groups of organisms. By comparing the DNA or protein sequences of different species, scientists can construct phylogenetic trees to study the evolutionary history of those species (Yang, 1997).

Phylogenetic trees are created by analyzing the similarities and differences in the DNA or protein sequences of different species. This analysis involves comparing the sequences and identifying similarities or differences between them. Once the similarities and differences are identified, scientists can use this information to construct a tree that shows how the different species are related to each other.

Phylogenetic trees generated using sequence alignment can provide important insights into the evolutionary relationships among different species. For example, these trees can be used to trace the origin of a particular trait or characteristic across different species, to identify common ancestors, and to understand how different species have evolved over time.

The implementation of sequence alignment in creating phylogenetic trees allows scientists to study the evolutionary relationships among different species in a systematic and quantitative way, providing important insights into the origins and diversification of life on Earth.

Sequence Alignment and Structure Alignment:

Sequence alignment and structure alignment are two complementary techniques that are widely used in bioinformatics. Sequence alignment is the process of comparing two or more sequences of nucleotides or amino acids to identify regions of similarity or homology. Structure alignment, on the

other hand, involves the comparison of the three-dimensional structures of proteins or other macromolecules to identify similarities and differences (Russell & Barton, 1992).

While sequence alignment is primarily used to compare the primary structure of proteins or nucleic acids, structure alignment is used to compare their three-dimensional structures. Structure alignment can reveal similarities that are not apparent from sequence comparison alone, as it takes into account the spatial arrangement of atoms in the molecule.

Despite these differences, sequence alignment and structure alignment are often used in combination to gain a more complete understanding of the evolutionary relationships between proteins and their functions. For example, sequence alignment can be used to identify conserved regions of a protein that are critical for its function, while structure alignment can be used to compare the 3D structures of proteins with similar sequences to identify structural features that are important for their function.

Online Bioinformatics Tools:

Online bioinformatics tools have made sequence alignment accessible to researchers worldwide. These tools include BLAST, Clustal Omega, and MUSCLE, among others. BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) is a widely used tool for identifying homologous sequences in a database (Altschul et al.,1990). Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) and MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/) are used for multiple sequence alignment and can align thousands of sequences in a matter of minutes.

Discussion

Sequence alignment has many applications in biology and biotechnology. It is commonly used in evolutionary studies to infer relationships between species by comparing their genetic sequences. Sequence alignment is also used in mutation detection to identify disease-causing mutations. By comparing the sequences of a patient's genes to a reference sequence, scientists can identify mutations that may be responsible for a particular disease. This information can be used to develop new diagnostic tests and treatments for that disease. For example, sequence alignment has been used to identify mutations responsible for cystic fibrosis, a genetic disease that affects the lungs and digestive system (Margoliash 1963).

In biotechnology, sequence alignment is used to design new drugs and biologics. By aligning the sequences of proteins that are involved in disease processes, scientists can identify potential drug targets and design drugs that specifically target those proteins. For example, sequence alignment of the original genomic sequence of the SARS-CoV-2 virus with new variants of virus let to deeper understandng of the virus's mode of infection and ulimately to the developement of vaccines. Another example is the human immunodeficiency virus (HIV) protease protein led to the development of protease inhibitors, which are now used to treat HIV infection (De Clercq, 2010).

Furthermore, sequence alignment can be used to analyse the structure and function of proteins. By aligning the sequences of proteins with known structures and functions, scientists can infer the structure and function of related proteins. This information can be used to design new proteins with specific functions, such as enzymes with improved catalytic activity.

Conclusion

Sequence alignment is a powerful tool in bioinformatics that allows scientists to compare and analyse genetic sequences. By identifying similarities and differences between sequences, scientists can infer evolutionary relationships, identify disease-causing mutations, design new drugs, and analyse the structure and function of proteins. With the development of new sequencing technologies and bioinformatics tools, sequence alignment is becoming an increasingly important tool in biology and biotechnology.

References

Callaway E. (2020) The race for coronavirus vaccines: a graphical guide. Nature. Oct;586(7830):506-508.

Dayhoff, M. O., & Eck, R. V. (1966). Atlas of Protein Sequence and Structure, 1967-1968. National Biomedical Research Foundation, Washington, D.C.

De Clercq, E. (2010). The discovery of antiviral agents: ten different compounds, ten different stories. Medical Research Reviews, 30(3), 533-582.

Dong E, Du H, Gardner L. (2020) An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. May;20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1.

Durbin R. et al. (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

Gibrat, J. F., Madej, T., & Bryant, S. H. (1996). Surprising similarities in structure comparison. Current opinion in structural biology, 6(3), 377-385.

Grubaugh ND, Hanage WP, Rasmussen AL. (2020) Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell. Aug 20;182(4):794-795. doi: 10.1016/j.cell.2020.06.040.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences, 89(22), 10915-10919.

Katoh, K., & Standley, D. M. (2016). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 33(7), 2499-2504.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., ... & Collins, F. S. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science, 245(4922), 1073-1080.

Margoliash, E. (1963). Primary structure and evolution of cytochrome c. Proceedings of the National Academy of Sciences, 50(4), 672-679.

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(3), 443-453.

Taylor, W. R., & Orengo, C. A. (1989). Protein structure alignment. Journal of molecular biology, 208(1), 1-22.

Rachedi A. (2020) Mutations in the SARS-CoV-2 complete genome sequences from strains isolated in Blida province, Algeria. JNBGP: Journees Nationales virtuelles de bioinformatique: Genomique et proteomique, ctober 09-10. doi: 10.13140/RG.2.2.20838.45123.

Russell, R. B., & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins: Structure, Function, and Bioinformatics, 14(2), 309-323.

Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. Apr 16;181(2):281-292.e6. doi: 10.1016/j.cell.2020.02.058.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences: CABIOS, 13(5), 555-556.

Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. (2020) Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov. Mar 16;6:14. doi: 10.1038/s41421-020-0153-3.