IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Comprehending the BLAST (Basic Local Alignment Search Tool) Program: Grasping the BLAST (Basic Local Alignment Search Tool) Program: A Detailed Guide for Its Application in Life Science Research

Raina Haque
Student
PPSIJC

Abstract : Bioinformatics is a rapidly evolving field that integrates computer science and information technology to manage, analyze, and share biological data. With advancements in DNA and protein sequencing, we now have access to vast amounts of sequence information. To uncover hidden patterns and understand evolutionary relationships within these sequences, it is crucial to compare them effectively. One of the key tools for this task is the Basic Local Alignment Search Tool (BLAST), which has become indispensable in life science research. Understanding how to use BLAST for sequence comparison and accurately interpret its results is essential for researchers. This article is designed to introduce biologists and researchers to the various BLAST programs and their applications in research.

INTRODUCTION

Bioinformatics is a scientific discipline that uses computational and statistical methods to analyze the vast amounts of biological sequence data generated through DNA and protein sequencing, as well as other biological experiments (NIH, 2010). Two of the most widely used bioinformatics tools for data storage and similarity searching are BLAST (Basic Local Alignment Search Tool) (McGinnis and Madden, 2004) and FASTA. BLAST, developed and maintained by the National Center for Biotechnology Information (NCBI) in the USA, is a powerful online tool designed to identify regions of similarity between biological sequences, whether DNA or protein. By comparing a given query sequence against NCBI's extensive database, BLAST quickly locates the most similar sequences, scanning the entire genomic library in seconds to find identical or closely related sequences.

BLAST

The Basic Local Alignment Search Tool (BLAST) is a software program designed for pairwise sequence alignment (Altschul et al., 1990). It performs sequence similarity searches to identify regions of local similarity between a query sequence and sequences in a database. In bioinformatics, 'similarity' denotes the degree of likeness between two sequences, expressed as a percentage, while 'homology' refers to the evolutionary relationship between sequences based on their similarity. BLAST estimates this homology by comparing sequences to infer their common ancestral origins.

Utilizing the Smith-Waterman algorithm (Smith and Waterman, 1981), BLAST aligns substrings of the query sequence with substrings from a target sequence database to find the most accurate match. This local alignment method identifies small but biologically significant regions of similarity. The program evaluates both the similarities and differences between the query sequence and database sequences, while also calculating the statistical significance of the matches.

The NCBI website offers a user-friendly BLAST server, making it a widely used and versatile tool for sequence similarity searching. Its popularity stems from its flexible search algorithm, reliable database, comprehensive statistical reporting, ongoing software improvements, and efficient search capabilities. BLAST leverages extensive sequence data from public databases such as GenBank, EMBL, and DDBJ. Its primary goal is to elucidate biological, structural, functional, phylogenetic, and evolutionary relationships between sequences.

BLAST ALGORITHM AND STATISTICS

The BLAST algorithm identifies regions of local alignment by dividing the query sequence into smaller sub-sequences. Initially, the query sequence, along with parameters such as the target database, word size, and expect value, is submitted through the BLAST input interface. The algorithm then scans the database to find matches with the query sequence. When a similar sequence is found, BLAST evaluates whether the alignment is sufficiently accurate to suggest a biological relationship or if it is likely due to chance (Eric et al., 2014).

To assess alignment quality, BLAST uses statistical metrics such as the bit score and expect value (E-value). For nucleotide alignments, a reward of +2 is given for identical aligned pairs, while a penalty of -3 is applied for non-identical pairs (Mount, 2004). Gaps introduced in the alignment to achieve optimal matches incur a negative gap penalty, with a lesser penalty for extending existing gaps (Mount, 2008).

The bit score measures sequence similarity independently of query length and database size, using a formula that considers aligned residues and gaps. A higher bit score indicates a better alignment. Amino acid alignments use substitution matrices like BLOSUM 62 or PAM, while nucleotide alignments use identity matrices. The PAM (Point Accepted Mutations) matrix, developed by Margaret Dayhoff, reflects accepted point mutations in closely related proteins, with one PAM unit representing one mutation per 100 amino acids. The BLOSUM (Blocks Substitution Matrix), created by Henikoff and Henikoff in 1992, is used for comparing conserved protein regions (Henikoff and Henikoff, 2004). These matrices are based on percentage identity values and similarity.

The E-value, or expect value, indicates the statistical significance of the alignment, showing the likelihood that the observed match is due to random chance. A lower E-value signifies a more significant match. BLAST results are typically sorted by E-value to highlight the most meaningful alignments.

TYPES OF BLAST PROGRAMS

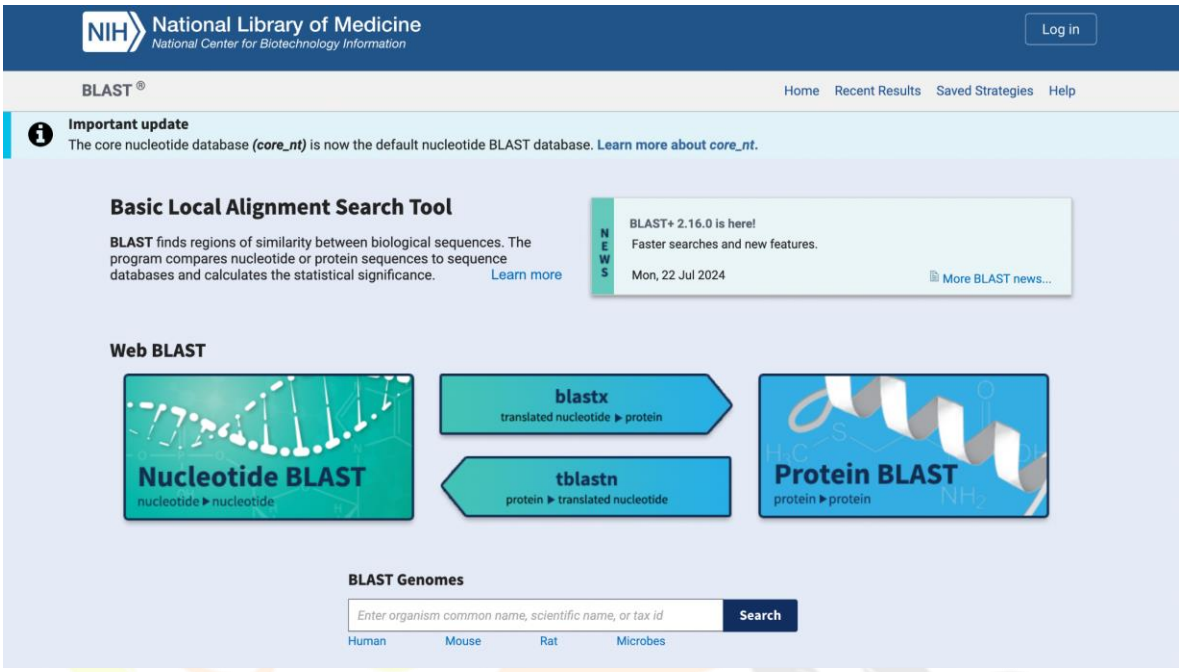
NCBI offers several BLAST programs tailored to different types of query sequences and comparison objectives. The choice of BLAST program for pairwise sequence alignment depends on the nature of the query sequence and the specific goals of the comparison. The most commonly used BLAST programs include BLASTp, BLASTn, BLASTx, tBLASTn, and tBLASTx.

USES OF DIFFERENT BLAST PROGRAMS

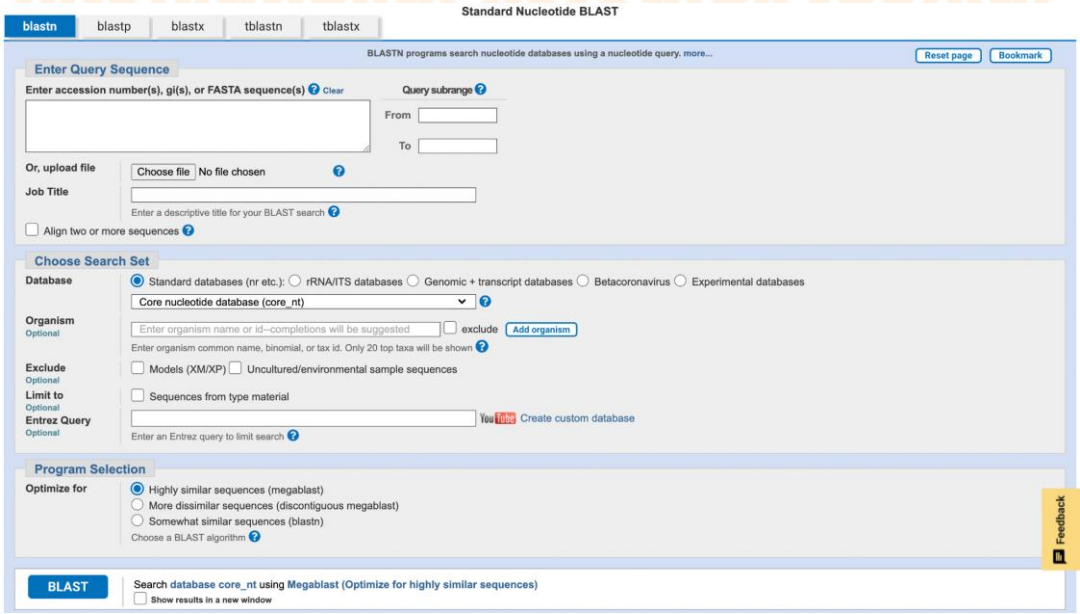
To run a BLAST program, a query sequence (either nucleotide or protein) is essential. This query sequence is compared against a sequence database, which contains numerous potential matches. The BLAST algorithm searches the database to identify sub-sequences that are similar to those in the query sequence.

STEP-BY-STEP GUIDELINE TO USE BLAST PROGRAMME

1. Access the BLAST Homepage
   - Begin by visiting the BLAST homepage on the NCBI website at https://blast.ncbi.nlm.nih.gov/Blast.cgi.



2. Choose the appropriate BLAST program based on the type of query sequence you are submitting.

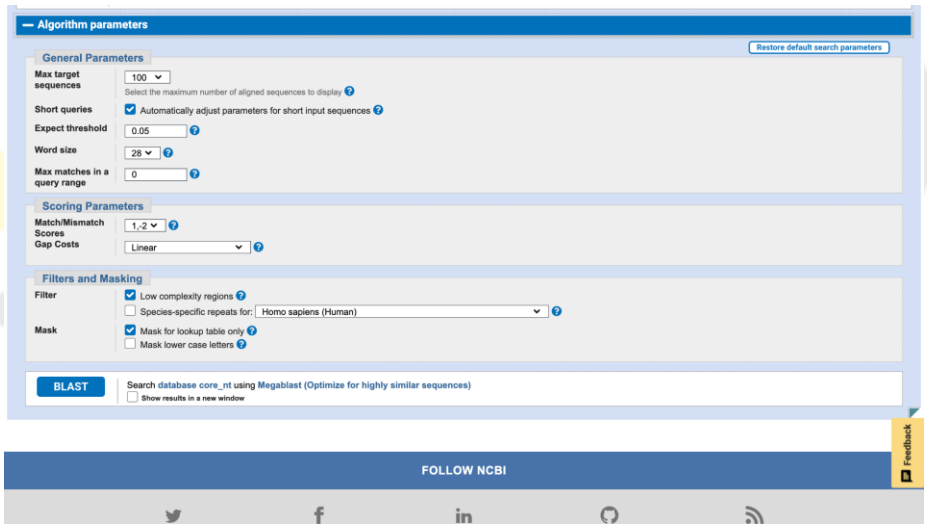


3. Submitting the query sequence in the window box
    - Submit the query sequence in the input box by either typing or pasting it directly. Alternatively, you can upload a file with the query sequence in FASTA format. You may also provide an accession number, GI number, or a raw FASTA sequence. For nucleotide queries, use character strings (A, T, G, C), and for protein queries, use single-letter amino acid codes. The query sequence should start with a definition line, indicated by a ">" symbol, which includes identifiers and descriptive information.

4. Select database to search
    - Users should be well-informed about the available databases and the types of sequences they contain. The default database is the non-redundant database (nr/nt). The database drop-down menu offers a variety of options, including expressed sequence tags (EST), sequence read archive (SRA), patent sequences (PAT), whole genome shotgun (WGS), transcriptome shotgun assembly (TSA), and high-throughput genome sequences (HTGS). For protein sequence comparisons, select UniProtKB/Swiss-Prot from the database menu. Other available options include patented protein sequences (PATA), protein data bank (PDB), metagenomic protein sequences (ENV_NR), and transcriptome shotgun assembly proteins (TSA_NR). Additionally, there are options for specifying organism names, IDs, exclusions, and limitations, which can help refine and enhance the search process.

5. Choose the algorithm and its parameters for the search.
    - The user must select the specific algorithm from the BLAST program's drop-down menu. Nucleotide BLAST offers algorithms like BLASTn for searching somewhat similar sequences and Mega BLAST for highly similar sequences. BLASTp is designed for searching somewhat similar protein sequences. PSI-BLAST performs iterative, position-specific searches (Altschul et al., 1997), while PHI-BLAST looks for specific patterns. If no algorithm is chosen, the search will default to the pre-set algorithm.

6. Executing the BLAST program
    - To initiate the BLAST program, click the 'BLAST' button at the bottom of the page and wait for the results. After a short delay, the results will appear under three main sections: Graphic Summary, Descriptions, and Alignments.



BLAST OUTPUT AND ITS ANALYSIS

To effectively interpret BLAST results, a combination of biological, computational, statistical knowledge, and practical experience is necessary. The BLAST results include the following field:

- E value: The E value (expected value) indicates how frequently a match might occur by random chance in a database of that size. A lower E value signifies a more statistically significant match.
- Per cent identity: Per cent identity is a measure of how closely the query sequence matches the target sequence, reflecting the proportion of identical characters in both sequences. A higher percentage identity indicates a more significant match.
- Query cover: Query cover indicates the extent to which the target sequence overlaps with the query sequence. A query cover of 100% means the target sequence encompasses the entire query sequence. The results page displays various details including the Query ID, Description, Molecule Type, Sequence Length, Database Name, and BLAST Program used. In the graphical representation, the top line shows a linear view of the query sequence, with bars below representing matches. Each bar is color-coded based on the alignment score, and gray areas indicate regions of no similarity. Hovering over a bar reveals the identifier of the aligned sequence, and clicking on a bar provides access to the alignment details, including percentage identity in similar regions. Assessing these values and the alignment is crucial for determining the significance of your results. The top line of the results page provides information about the BLAST program type and version, followed by the citation of the research paper that describes BLAST, the request ID, the query sequence definition, and a summary of the database searched. The Taxonomy Reports link presents the BLAST results in the context of the Taxonomy database. The query sequence is shown as a numbered red bar, with database hits displayed below according to alignment score. Sequences most related to the query are positioned closest to it. You can obtain more details about these alignments by hovering over each colored bar. Each alignment includes sequence identities, definition line, matched sequence length, score, and E-value. Additional information on identical residues, positive matches, and alignment gaps is also provided. Finally, the alignment displays the query sequence at the top and the database sequence below, with numbers indicating the position of amino acids or nucleotides in the sequence.

USES OF BLAST IN THE FIELD OF BIOLOGICAL SCIENCES

The BLAST tool has a broad range of applications in biological sciences, including:

- Identifying similarities between sequences.
- Detecting domains or conserved regions in protein sequences and determining the protein family of a query sequence.
- Mapping DNA to known chromosomes and assisting in genome sequence assembly.
- Identifying homologous gene candidates across different genomes (Lu et al., 2006).
- Comparing species by identifying similar genes across various organisms (Holton, 2004).
- Conducting comparative gene prediction by searching across two genome sequences for both sensitive and specific gene predictions (Parra et al., 2003).
- Determining functional properties and biological roles of nucleotide sequences within genomes (Moriya et al., 2007).
- Assisting in prokaryotic genome sequence assembly through contig mapping (van Hijum et al., 2005).
- Exploring the evolutionary history of genes and understanding the genomic evolution (Zhang et al., 2006).
- Creating datasets for phylogenetic analysis (Dereeper et al., 2010) and constructing phylogenetic trees from protein sequences (Kelly and Maini, 2013).

CONCLUSION

NCBI-BLAST (Basic Local Alignment Search Tool) is an online software used to compare primary sequences—whether amino acid sequences of proteins or nucleotide sequences of DNA and/or RNA—with those in a database. Pairwise sequence alignment is crucial in biology for uncovering hidden information within biological sequences. It also aids in identifying various features, such as predicting the 3D structure of protein molecules, studying molecular interactions, and deriving valuable insights from biological data. Since protein sequences are generally more conserved than nucleotide sequences, BLAST programs like tBLASTn, tBLASTx, and BLASTx often yield more reliable and accurate results when analyzing coding DNA.

ACKNOWLEDGEMENTS

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic alignment search tools. *Journal of Molecular Biology*, 215, 403-410.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389-3402.
- Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., Madden, T.L. (2012). Domain Enhanced Lookup Time Accelerated BLAST. *Biology Direct*, 7, 12. https://doi.org/10.1186/1745-6150-7-12.
- Dereeper, A., Audic, S., Claverie, J.-M., Blanc, G. (2010). BLASTEXPLORER: a tool for building datasets for phylogenetic analysis. *BMC Evolutionary Biology*, 10, 8.
- Donkor, E.S., Dayie, N.T.K.D., Adiku, T.K. (2014). Bioinformatics with BLAST and FASTA. *Journal of Bioinformatics and Sequence Analysis*, 6, 1-6.
- Henikoff, S., Henikoff, J.G. (2000). Amino acid substitution matrices. *Advances in Protein Chemistry*, 54, 73-97.
- Holton, W.C. (2004). The Path to Species Comparison. *Environmental Health Perspectives*, 112(12), A672. https://blast.ncbi.nlm.nih.gov.
- Kelly, S., Maini, P.K. (2013). Dendro BLAST: Approximate Phylogenetic Trees without Multiple Sequence Alignments. *PLOS ONE*, 8(3), e58537.
- Lu, G., Jiang, L., Helikar, R.M.K., Rowley, T.W., Zhang, L., Chen, X., Moriyama, E.N. (2006). Genome Blast: a web tool for small genome comparison. *BMC Bioinformatics*, 7(Suppl 4), S18.
- McGinnis, S., Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(Web Server issue), W20-W25.
- Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35, W182-W185.
- Mount, D.W. (2004). Alignment of pairs of sequences. In *Bioinformatics: Sequence and Genome Analysis*, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Mount, D.W. (2008). Using gaps and gap penalties to optimize pairwise sequence alignments. *Cold Spring Harbor Protocols*, 2008, pdb.top40.
- National Institutes of Health (2010). NIH Working Definition of Bioinformatics and Computational Biology. Bethesda, USA. http://www.bisti.nih.gov/docs/CompuBioDef.pdf.
- Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W., Guigo, R. (2003). Comparative gene prediction in human and mouse. *Genome Research*, 13, 108-117.
- Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14), 2994-3005.
- Smith, T.F., Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195-197.
- Syngai, G.G., Barman, P., Bharali, R., Dey, S. (2013). BLAST: An introductory tool for students to bioinformatics applications. *Kenean Journal of Science*, 2, 67-76.
- van Hijum, S.A.F.T., Zomer, A.L., Kuipers, O.P., Kok, J. (2005). Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Research*, 33, W560-W566.
- Wootton, J.C., Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers in Chemistry*, 17, 149-163.
- Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., Madden, T.L. (2012). Primer-BLAST: a tool for designing target-specific primers for PCR. *BMC Bioinformatics*, 13, 134.
- Ye, J., Ma, N., Madden, T.L., Ostell, J.M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41, W34-W40.

- Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P.M., Gerstein, M. (2006). PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*, 22(12), 1437-1439.