Bhartiya Krishi Anushandhan Patrika, 36(1): 55-61 DOI: 10.18805/BKAP283

# Understanding the BLAST (Basic Local Alignment Search Tool) Program and a Step-by-step Guide for its use in Life Science Research

**Kailash Chandra Samal, Jyoti Prakash Sahoo, Laxmipreeya Behera, Trupti Dash**

Department of Agricultural Biotechnology, Odisha University of Agriculture and Technology, Bhubaneswar-751 003, Odisha, India.

Received: May 2021 Accepted: June 2021

## ABSTRACT

Bioinformatics is the new branch of science which deals with the acquisition, storage, analysis and dissemination of biological data with the help of computer science and information technology. It has the enormous ability to analyze a vast quantity of biological data quickly and cost-effectively. In the past decades, enormous sequence information has been generated due to the advances in DNA and protein sequencing techniques. Estimating similarities between biological sequences is becoming necessary to obtain hidden information present within the sequence and to trace evolutionary relationship exist within the sequences. This sequence comparison can be achieved by basic local alignment search tool (BLAST). So BLAST has become a fundamental tools of life science research. Hence it is essential to know how to do sequence comparison using BLAST and how to accurately interpret the BLAST output data. The present article aims to familiarize the biologists and researchers with different BLAST programs and their use in research program.

**Key words:** Bioinformatics, BLAST, Biological sequence, DNA, E-value, Protein.

## INTRODUCTION

Bioinformatics is the branch of science that employs computational and statistical approaches to analyze the enormous biological sequence information obtained through DNA and protein sequencing as well as other biological experiments (NIH, 2010). The most common bioinformatics tools used for data storing and similarity searching in the various database are BLAST (Basic Local Alignment Search Tool) (McGinnis and Madden, 2004) and FASTA. BLAST is an online search tool developed and maintained by NCBI (National Center for Biotechnology Information), USA. It is widely used to find/trace out “regions of similarity between biological (nucleotide or protein) sequences”. The BLAST tool takes up the query sequence (either DNA or protein sequence) provided by the researcher and compares that query sequences to the massive database of biological sequences stored in the NCBI database to find the most similar ones. During the search process, the entire genomic library is scanned for gene sequence of interest to find out identical or similar sequences in a matter of seconds (Syngai *et al.*, 2013).

### Blast: A sequence analysis tool

The Basic Local Alignment Search Tool (BLAST) program refers to an algorithm or software used for pairwise sequence alignment (Altschul *et al.,* 1990). It carries out sequence similarity searching to find the regions of local similarity to a query sequence. In bioinformatics, the word ‘similarity’ refers to the degree of likeness between two sequences and it expressed in percentage. However, the word ‘homology’ refers to the common evolutionary ancestry of two sequences based on an assessment of their similarity. The homology of biological sequences is estimated based on the degree of sequence similarity between two sequences to trace back the ancestral origin of the sequence. BLAST program is a pairwise sequence alignment tool that searches for regions of similarity between the two sequences as per the Smith-Waterman algorithm (Smith and Waterman, 1981). The pairwise local alignment aligns a query sequence substring to the substring of the target sequence database for getting the best possible alignment between the sequences. In this alignment method, the small regions of similarity between the sequences can be detected, which may be more biologically significant (Wootton and Federhen, 1993). In the pairwise sequence comparison, the similarity and differences among the unknown query sequence and the database sequences are estimated and traced. In addition, the program calculates the statistical significance of the matches.

The NCBI website includes a very user-friendly BLAST server through which pairwise sequence comparison between query sequence and database is made. It is the most popular, user-friendly, versatile sequence similarity searching tool. The program is used worldwide because of the flexibility of the search algorithm, reliability of the database, quality statistical report, progressive software development and the accuracy and speed of the searching method. In addition, the program uses an abundance of sequence data present in public databases such as Genbank, EMBL, DDBJ *etc*., during the search. The general objective of the BLAST tool is to establish a biological, structural, functional, phylogenetic and evolutionary relationship between the sequences.

*Corresponding author's E-mail: samalkcouat@gmail.com

**Understanding the BLAST Algorithm and BLAST Statistics**

The BLAST algorithm finds regions of local alignments by breaking the query sequence into the smaller sequence (sub-strings). First, a query sequence along with other input information such as the database to be searched, word size, expect value and so on is submitted through the input window of the BLAST program. Then the submitted sequence is scanned to find out the matches in the sequences present in the different biological sequence databases. During the search, once BLAST has found a similar sequence in the database against the query sequence, then the BLAST algorithm tests whether the alignment is "perfect or good enough" and whether it signifies a possible biological relationship or it is occurred by chance (Eric *et al.,* 2014). For this purpose, BLAST employs statistical measures such as bit score and expect value (E-value) for each sequence alignment pair. For nucleotide alignments, a reward of +2 was assigned for each aligned pairs of identical letters, whereas a penalty of "3" was assigned for each non-identical aligned pair (Mount, 2004). During the sequence alignment, gaps may be inserted in the sequence for optimal matching and the creation of a gap results in a negative gap penalty. Similarly, for the extension of the existing gap, a lesser penalty was assigned in the alignment scoring (Mount, 2008).

A bit score is a statistical indicator that measures sequence similarity independent of the length of the query sequence and database size. For the calculation of the bit score, a formula is used that takes into account the alignment of similar or identical residues as well as any gaps introduced during sequence alignment. The bit score value indicates how good the alignment is; the higher the score, the better the alignment. For this purpose, the substitution matrices (BLOSUM 62 or PAM) are used for amino acid alignments, whereas nucleotide sequences are compared using identity matrices.

PAM (Point Accepted Mutations) matrix was developed by Margaret Dayhoff observing the differences in closely related proteins. One PAM unit refers to one accepted point mutation per 100 amino acid residues. BLOSUM (Blocks Substitution Matrix) was developed by Henikoff and Henikoff in 1992 and used when conserved regions of the protein sequence were compared (Henikoff and Henikoff, 2004). These matrices are actual percentage identity values and they depend on similarity.

E-value or expect value is the measure of likeliness that indicates the statistical significance of a given pairwise alignment. It informs whether a given sequence pair match is purely a result of random chance or not. The lower E-value, justifies the more significant match. E-value is used to filter for the BLAST search technique to obtain significant matches. By default, the BLAST results are sorted by E-value.

**Types of the BLAST program**

In NCBI, different BLAST programs are available with respect to the query sequence submitted and sequence database searched for comparison. During the pairwise sequence alignment, the selection of the BLAST program depends on type of the query sequence and the objective of the comparison. The various types of popular BLAST programmes are BLASTp, BLASTn, BLASTx, tBLASTn, tBLASTx (Table 1).

**Table 1:** Commonly used BLAST programmes in life science research.

| BLAST programs | Query and target sequence and mode of alignment |
|---|---|
| BLASTn | Compares a nucleotide sequence against a nucleotide sequence database. |
| BLASTp | Compares a protein sequence against a protein sequence database. |
| BLASTx | Compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database. |
| tBLASTn | Compares a protein sequence against a six-frame translation of a nucleotide sequence database. |
| tBLASTx | Compares six-frame conceptual translation products of a nucleotide query sequence against six-frame translations of a nucleotide sequence database. It is commonly used to find very distant relationships between nucleotide sequences. |
| PSI-blast | Position-Specific Iterative BLAST program uses a specialized scoring matrix that assigns scores to each position (hence, position-specific) in the query sequence based on alignments defined by consecutive iterations of searches. It is commonly used to find distant relatives of a protein (Schaffer *et al*., 2001). |
| PHI-blast | Pattern-Hit Initiated BLAST uses an input sequence and a defined pattern to query a protein database. The pattern is defined in PROSITE format (http://ca.expasy.org/prosite/) and is used as the seed for the alignment. The pattern is used instead of the words that are usually generated for seeding alignments in BLASTp. |
| Primer-blast | A tool to design target-specific primers for polymerase chain reaction (Ye *et al.*, 2012). |
| IgBLAST | An immunoglobulin variable domain sequence analysis tool (Ye *et al.*, 2013). |
| MOLE-BLAST | An experimental tool that helps taxonomists find closest database neighbours of submitted query sequences. It computes a multiple sequence alignment (MSA) between the query sequences along with their top BLAST database hits and generates a phylogenetic tree. |
| DELTA-BLAST | It searches a database of pre-constructed PSSMs before searching a protein-sequence database, to yield better homology detection. It is a useful program for the detection of remote protein homologs. (Boratyn *et al*., 2012). |

**Sequence type and uses of different BLAST program**

A query sequence (nucleotide or protein) is always required for running a Blast program. The submitted query sequence is searched against (also called the target sequence) or a sequence database containing multiple such sequences. The BLAST algorithm will try to trace sub-sequences in the database which are similar to subsequences in the query. The details of query sequence, database type and uses of BLAST program are described in Table 2.

**Step-by-step guideline to use BLAST programme**

One must have a query sequence for performing the BLAST program and the query sequence may be either a nucleotide or amino acid sequence. The sequence should be in FASTA format. The following steps should be followed to start a BLAST search,

1. **Open the BLAST homepage**

First navigate or open the blast homepage from the NCBI site (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

2. **Select the type of BLAST program for the query sequence to be submitted**.

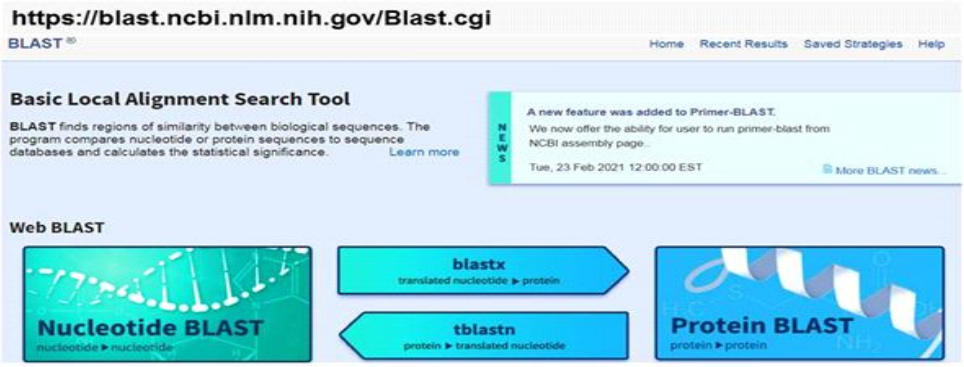
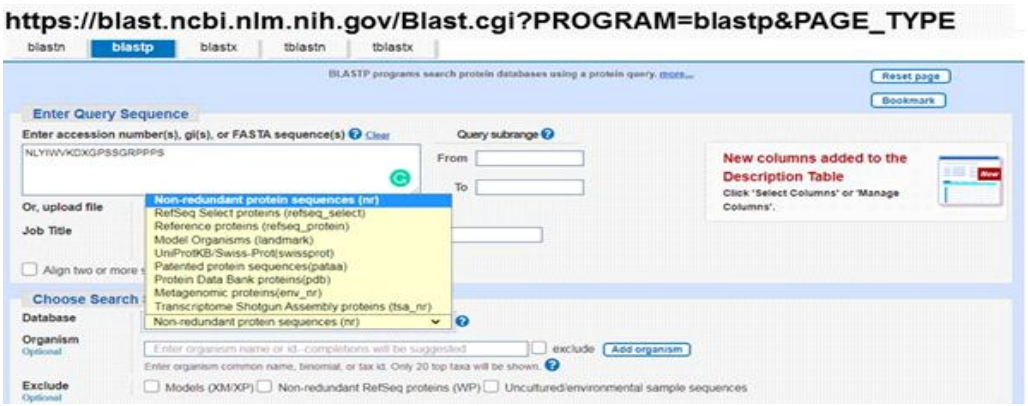



**Table 2:** Types of query sequence, database type and uses of BLAST program.

| BLAST Program | Query sequence | Target database sequence | Use of the programme |
|---|---|---|---|
| BLASTn | Nucleotide | Nucleotide | Mapping oligonucleotides, cDNAs and PCR products to a genome screening repetitiv elements cross-species sequence exploration annotating genomic DNA clustering sequencing reads |
| BLASTp | Protein | Protein | Identifying common regions between proteinscollecting related proteins for phylogenetic analyses |
| BLASTx | The nucleotide query sequence is first translated through six possible ORFs into six protein sequences and each translated protein sequence is served as a query sequence for comparison | Protein | Identifying protein-coding genes in genomic DNA determining if a cDNA corresponds to a known protein |
| tBLASTn | protein | Nucleotide present in the database translated through six possible ORFs into six different protein sequence and then the comparison is made between protein and protein sequence (withall possible ways) | Identifying transcripts, potentially from multiple organisms, similar to a given protein used in the identification of organism based on protein sequence similaritymapping a protein to genomic DNA |
| tBLASTx | The nucleotide query sequence is first translated through six possible ORFs into six protein sequences and each translated protein sequence is served as a query sequence for comparison | Nucleotide present in the database translated through six possible ORFs into six proteins and each is used for comparison with six translated protein sequence (with all possible ways) | Cross-species gene prediction at the genome or transcript level searching for genes missed by traditional methods or not yet in protein databases Cross-species gene identification |





3**. Submitting the query sequence in the window box**

Enter or Paste a query sequencein the window box. Otherwise, upload a file containing the query sequence in FASTA format. The user can also give the accession number or gi number or even a raw FASTA sequence. BLAST “query nucleotide sequences” are given as character strings (A, T, G, C). Similarly, BLAST “query protein sequences” are given as character stringsof single-letter amino acid codes. The query sequence is always preceded by a definition line, beginning with a “>” symbol and containing identifiers and descriptive information.

**4. Select database to search**

The user should have explicit knowledge regarding the availability of the databases to be searched and the type of sequences present in them. The default database is the non-redundant database (nr/nt). There are varied options available in the database pull-down menu from which one should choose a respective option as per the requirement. Available options are expressed sequence tags (est), sequence read archive (SRA), patent sequence (pat), wgs, TSA, HTGS *etc.* For protein sequence comparison, one should choose Uni-protKB/Swiss-prot from the database pull-down menu. Other options available in the database pull-down menu are patented protein sequence (pataa), protein data bank (pdb), metagenomic protein (env_nr), transcriptome shotgun assembly protein (tsa_nr). In addition, other optional options such as organism name and Id, exclude, limit to, entrez query *etc.* are available. One should select any options from this for a more organized and efficient search.

**5. Select the algorithm and the parameters of the algorithm for the search**.

The user has to choose the specific algorithm from the pull-down menu of the BLAST program. Nucleotide BLAST uses algorithms like BLASTn, which searches for somewhat similar sequences, whereas Mega BLAST searches for highly similar sequences. Similarly, BLASTp uses algorithms which searches for somewhat similar protein sequences. PSI-BLAST is a specific BLAST program that performs position-specific search iteratively (Altschul *et al.*, 1997), whereas PHI-BLAST searches for a particular pattern. If the user does not choose any algorithm, then the default algorithm will be used during the search.

6. **Running BLAST program**

The BLAST program is run by clicking the BLAST button at the end of the page. Then wait for the results. After few seconds, the results are displayed in three main headings: Graphic Summary, Descriptions and Alignments.

**BLAST result and its interpretation**

The biological, computer, statistical knowledge and practical expertise are required to interpret BLAST results efficiently. The BLAST results have the following fields:

E value: The E value (expected value) is a number that describes how many times you would expect a match by chance in a database of that size. The lower the E value is, the more significant the match.


**Per cent identity**

Per cent identity is a number that describes how similar the query sequence is to the target sequence (how many characters in each sequence are identical). The higher the percent identity is, the more significant the match.

**Query cover**

A query cover is a number that describes how much of the query sequence is covered by the target sequence. If the target sequence in the database spans the whole query sequence, then the query cover is 100%. The result page will appear along with the information like Query id, Description, Molecule type, Length of sequence, Database name and BLAST program. In the graphical representation of the BLAST results, the graph's top is a linear view of the query sequence, with the bars below indicating matches to it occurs. Each of the bars is coloured according to the score the alignment received. Grey areas in the bars represent areas that are not similar to the query sequence. Pointing the mouse over each of the bars will display the identifier of the aligned sequence. Clicking on a bar will take to the alignment of that sequence with the query and per cent identity within a similar region. Examining these values and the alignment itself is an essential step in deciding if your results are significant. The top line gives information about the type of BLAST program and its version) used. The citation of the research paper that describes BLAST is then mentioned, followed by the request ID, the query sequence definition line and a summary of the database searched. The Taxonomy reports link displays this BLAST result based on information in the Taxonomy database.

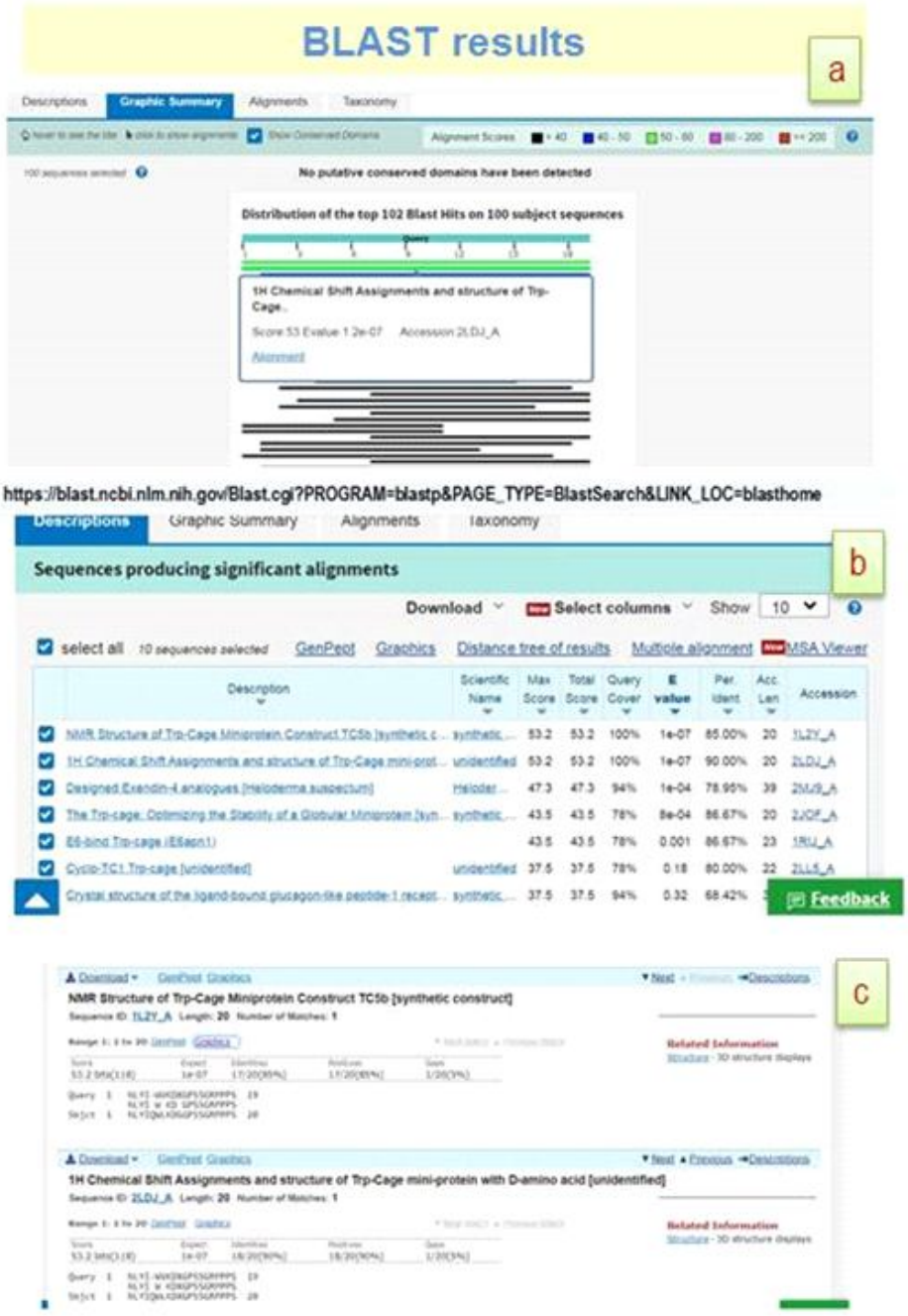
The query sequence represented as a numbered red bar below the colour key. Database hits are shown below the query (red) bar according to the alignment score. Among the aligned sequences, the most related sequences are kept near to the query sequence. The user can find more descriptions about these alignments by dragging the mouse to each coloured bar shown.

The alignment is preceded by the sequence identities, along with the definition line, length of the matched sequence, followed by the score and E-value. The line also contains information about the identical residues in alignment (identities), number of positivity's, number of gaps used in the alignment. Finally, it shows the actual alignment and the query sequence on the top and the database sequence below the query. The number on either side of the alignment indicates the position of amino acids/ nucleotides in the sequence.

**Applications of BLAST in biological sciences**

The BLAST tool finds its use in a wide range of biological applications. Identification of similarity between the sequences.

- Looking for domains or conserved region in the protein sequence and finding the protein family of the query sequence.
- Mapping DNA to a known chromosome of an organism and genome sequence assembling.
- Identification of homologous gene candidates across diverse genomes (Lu *et al*., 2006).
- Looking for species and species comparison by identifying similar genes in different organisms (Holton, 2004).
- Comparative gene prediction involves searching for two genome sequences to provide both sensitive and specific gene predictions (Parra *et al*., 2003).
- Identification of functional properties and biological roles of the nucleotide sequence in the genomes (Moriya *et al*., 2007).
- Prokaryotic genome sequence assemblies with help of Contig mapping (van Hijum *et al*., 2005).
- Determining the evolutionary history of genes and Understanding the evolutionary history of the genome (Zhang *et al*., 2006).
- Building datasets for phylogenetic analysis (Dereeper *et al*., 2010) and constructing phylogenetic dendrograms/trees from protein sequences (Kelly and Maini, 2013).




Designing target-specific primers for polymerase chain reaction (Ye *et al.*, 2012).

## CONCLUSION

NCBI-BLAST (basic local alignment search tool) is an online software program for comparing primary sequences (amino-acid sequences of proteins or the nucleotides of DNA and/ or RNA sequences) with the database sequences. Pairwise sequence alignment has become a vital part ofbiology for extracting hidden information present in the biological sequences. It is also used to identify various features such as predicting 3D structure of protein molecules, studying molecular interactions and extracting useful information from the biological data. As the protein sequences are more conserved than nucleotide sequences, the BLAST programs such as tBLASTn, tBLASTx and BLASTx, produce more reliable and accurate results when dealing with coding DNA.

## ACKNOWLEDGEMENT

We thank the authors of the various papers cited in this review article.

## REFERENCES


Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic alignment search tools. J. Mol. Biol. 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. and 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). BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BioMed Central Evolutionary Biology. 10(8): pp. 1-6.

Eric S. Donkor, Nicholas, T.K.D. Dayie and Theophilus K. Adiku, (2014). Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). Journal of Bioinformatics and Sequence Analysis. 6(1): 1-6.

Henikoff, S., Henikoff, J.G. (2000). Amino acid substitution matrices. Adv. Protein Chem. 54: 73-97

Holton, W.C. (2004). The Path to Species Comparison. In: Environmental Health Perspectives. 112(12): A 672.

https://blast.ncbi.nlm.nih.govBLAST: Basic Local Alignment Search Tool.

Kelly, S., Maini, P.K. (2013). Dendro BLAST: Approximate Phylogenetic Trees in the absence of Multiple Sequence Alignments. PLOS ONE. 8(3): e58537 pp. 1-11.

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. BioMed Central Bioinformatics. 7(Suppl 4): S18: 1- 9.

McGinnis, S. and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 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, by David W. Mount. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.

Mount, D.W. (2008). Using gaps and gap penalties to optimize pairwise sequence alignments. CSH Protoc, 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. and 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, SF. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29(14): 2994-3005.

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol Biol. 147(1): 195-197.

Syngai, G.G., Barman, P., Bharali, R. and Dey, S. (2013). BLAST: An introductory tool for students to Bioinformatics Applications. Keanean Journal of Science. 2: 67-76.

vanHijum, 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. and 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., and Madden, T.L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 13: 134.

Ye, J., Ma, N., Madden, T.L., and Ostell, J.M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013 Jul; 41: W34-W40.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., Madden, T.L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BioMed Central Bioinformatics. 13(134): 1-11.

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.