Comparison of Variants of BLAST

Harpreet Kaur, Amandeep Singh, Pardeep Singh

***ABSTRACT*-Nowadays, large quantities of gene sequences of related species of plants, animals and microorganisms show complex patterns of similarity to one another and many molecular biologists are convinced that an understanding of sequence evolution is the first step towards understanding the evolution itself. There are varieties of different tools available to perform sequence analysis. Blast is a successful tool to compare biological sequences. Now a days Large amount of biological data is available, So Standalone Blast is not sufficient to handle all types of queries related to sequence similarities, so different variants (BlastX, BlastP, BlastN, TblastN, TblastX, PSI_Blasts) have been developed. Each variant has limitations and advantages. Every tool is made to handle with different purposes. In this paper comparison of variants of BLAST(BlastX, BlastP, BlastN, TblastN, TblastX, PSI_Blasts) is done on the basis of different factors. This has been also analyzed that how we can increase the efficiency of these variants.**

Keywords- **pattern, sequence alignment**

1.0 INTRODUCTION

Bioinformatics is the recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence (genes and RNAs), protein sequence and structural information. This includes databases of the sequences and structural information as well methods to access, search, visualize and retrieve the information. Bioinformatics concern the creation and maintenance of databases of biological information whereby researchers can both access existing information and submit new entries. Bioinformatics includes Sequence analysis used by geneticists, cell biologists, molecular biologists, Molecular modeling used by crystallographers, cell biologists, biochemists, Molecular phylogeny/evolution, Ecology and population studies.

Harpreet Kaur Author is with the Baba Banda Singh Bahadur Engg College(BBSBEC),Lecturer Computer Science and Engg Deptt., Fatehgarh Sahib, INDIA (Mobile: 98140-08879, Email: harpreet_khasria@rediffmail.com)

Amandeep Singh Author is with the Baba Banda Singh Bahadur Engg College(BBSBEC), Lecturer Mechanical Deptt., Fatehgarh Sahib, INDIA (Email: amandeep_singh81@rediffmail.com)

Pardeep Singh Author is with National Institute of Technology(NIT), Hamirpur, Lecturer Computer Science and Engg. Deptt., INDIA (Email: p_singh79@rediffmail.com)

The most pressing tasks in bioinformatics involve the analysis of sequence information. **Computational Biology** is the name given to this process, and it involves the following:

- Finding the genes in the DNA sequences of various organisms
- Developing methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.
- Clustering protein sequences into families of related sequences and the development of protein models.
- Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships.

Blast is a successful tool to compare biological sequences. Now a day's large amount of biological data is available, but Standalone Blast is not sufficient to handle all types of queries related to sequence similarities, so different variants (BlastX, BlastP, BlastN, TblastN, TblastX, PSI_Blasts) have been developed. Each variant has limitations and advantages. Every tool is made to handle with different purposes. So the user should have knowledge in which situation to use which tool. Comparison is needed between these variants different to know thoroughly about these tools

2.0 COMPARISON ON THE BASIS OF PARAMETERS

All variants of BLAST run on same algorithm followed by Main Blast Program. There are some differences occur between these variants, due to which the functionality differs. All the parameters are same for all variants, which are used for MAIN BLAST program. But still there are some parameters which can be present in some variants, or the absence of which can make other tools to advantageous one over the other.

2.1 Conserved Domain Search Is Not Applied To Blastn, It Is Applicable To Blastp.

Proteins often contain several domains, each with a distinct function (membrane binding, signal peptide, etc.) .As species evolve; the functional parts of important proteins remain relatively constant over time, and may even be copied and adapted for use by other proteins. Such domains have evolved as modules that are combined in various arrangements to produce proteins of unique function. Conserved domains are structural modules that have been reused frequently during the process of evolution. NCBI's new Conserved Domain Search (CD-Search) service can be used to identify conserved domains in a protein sequence.

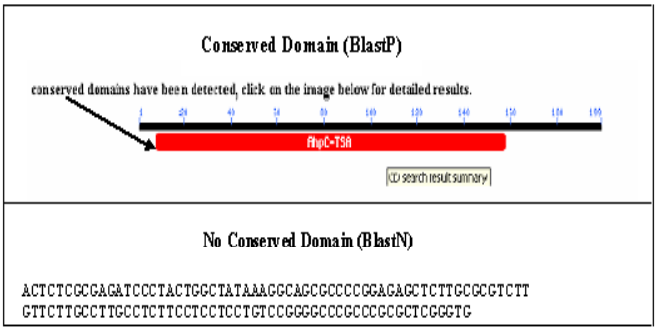

Fig. 1.1 **Conserved Domain For BlastN and BlastP**

Influence of absence of CDD Search: Conserved Domain Search is applicable only to proteins. Because it is based on PSSMs (Position Specific Score Matrices) which is applied only on proteins. By applying PSSMs, specific functional areas with in proteins can be searched. The searched functional domains are used in future for further research.
Because PSSM is not applied on nucleotides so if there are specific functional areas exist in nucleotides, no search option is available for that.
Conserved domain will not work for nucleotide as -it is based on PSSM which does not apply to nucleotide.

2.2 The Default Word Size Is 11 Characters For Blastn. The Default Word Size Is 3 For BLASTP, due To Which BLASTP Searches Run Slower Than BLASTN.

Word size (seed) strongly affects the database searching. Speed of the algorithm is inversely proportional to the word size. By decreasing the word size the sensitivity increases but speed of the search program decreases. Word size for BlastP is very small as compared to BlastN. Word size (seed) in case of BlastP is of 3-residues.It is seen for BlastP, during the second step of algorithm, large no of hits are found in the database. This is because of the small size of the seed. So more time is spent on the search. But in case of BlastN, seed is of 11-nucleotides.It is difficult to find more number of exact matches for such large seed size. Results are displayed in lesser time as compared to BlastP and less number of hits are found. But sensitivity decreases in BlastN.

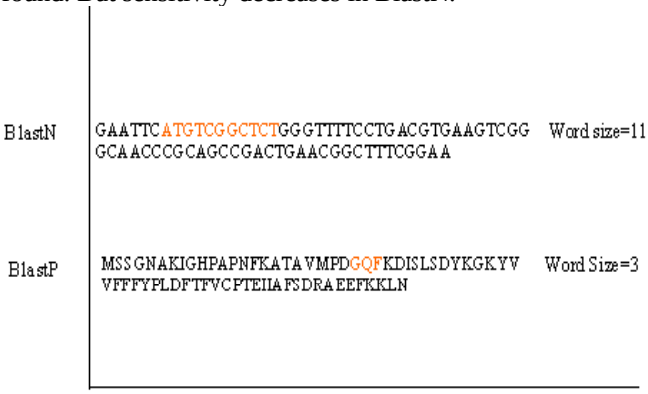

Fig. 1.2 **Different Word Sizes For BlastN and BlastP**

2.3 Unlike Nucleotide BLAST, There Is No Comparable MEGABLAST For Protein Searches.

MegaBlast is optimized for aligning sequences that differ slightly as a result of sequencing or other similar "errors". MegaBlast is also able to efficiently handle much longer DNA sequences than the blastn program of traditional BLAST algorithm. When larger word size is used (see explanation below), it is up to 10 times faster than more common sequence similarity programs. Mega BLAST is also able to efficiently handle much longer DNA sequences than the blastn program of traditional BLAST algorithm.

Influence of absence of Mega Blast: MegaBlast is an improvement to existing BlastN algorithm, but for proteins there is no such program exists. No batch queries can be run in case of protein sequence searching. Longer sequence searches cant be applied so efficiently. To improve the speed of the protein searches by speed, and to handle long sequence searches MegaBlast like program should be developed for proteins, Which can run large protein sequence and batch sequences at a time.

2.4 The Two-Hit Algorithm Isn't Used In BLASTN, Because Word Hits Are Generally Rare With Large Identical Words.

The two-hit algorithm isn't used in original version. BLASTN the statistical alignments which are found using main BLAST algorithm are based on threshold value 'T' and drop-off score X.
Influence of absence of two-hit algorithm: Two-hit algorithm is not used for BlastN, because the word size for BlastN is large (11 nucleotide). Word hits are the identical words. It is rare and difficult to find word hits with large word size. It is easy to find identical matches for one or two nucleotide in a given database.

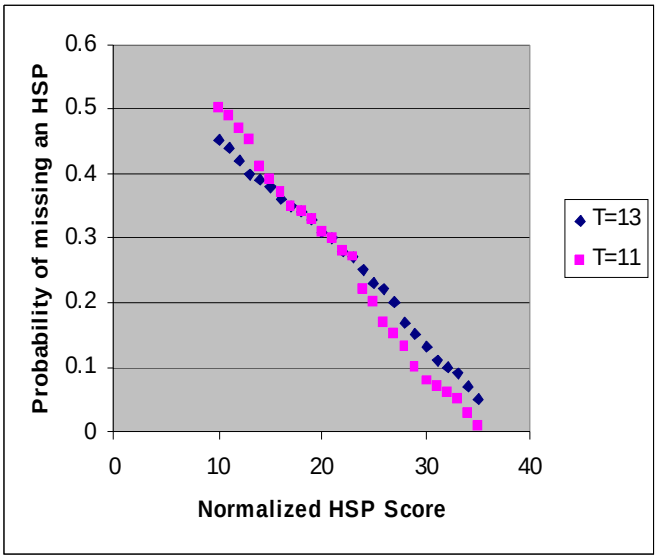

Fig. 1.3 **shows the empirically estimated probability that an HSP is missed by this method, as a function of its normalized score**

But it is very rare that we find exactly same nucleotide sequence with the seed of 11 bp. Therefore two-hit algorithm is not used.

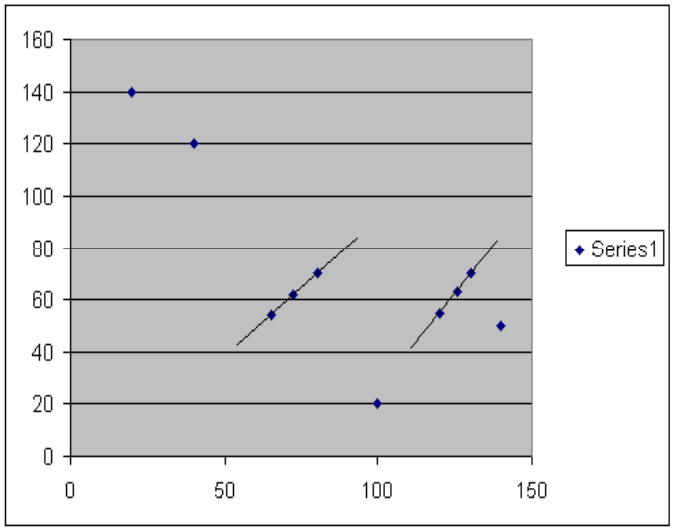

Fig. 1.4 **Speeds of the one-hit and two-hit methods**

Improvement: If two-hit algorithm will be applied to blastn, The sensitivity of BlastN will increased and more accurate sequence similarity will be obtained. This can be done by decreasing the word size of BlastN. Because with large words size it is difficult to find the same matches regularly at two positions. But with short word size it is easy to find the exact matches at more than one position.

2.5 Extension in BlastN is different from BlastP and other protein based programs.

Extension for BlastN is different from Blastp. This is because of the Proteins and Nucleotides. Different Scoring matrices are used for scoring of neighborhood during extension. Different scoring matrices yields separate drop-off(X) score for BlastN and BlastP.But in BlastN there are 11-nucleotides for which the whole score has to be evaluated. It will take more time to calculate as compared to BlastP because the word size for BlastP is small as compared to BlastN.

3.0 COMPARISON ON THE BASIS OF PERFORMANCE

Every tool is efficient in different conditions and to different input queries. Performance of variants is measured on the basis of following criteria.
Performance of various variants of Blast is measured on the basis of :

- Expect Value
- Word Size
- Time

3.1Comparison On The Basis Of - Varying Expect Values

A **BlastN** was performed using the mRNA sequence of PRDX1 against the non-redundant database. To observe the effect of the "expect value" parameter, values of 10, 0.1, and 1e-30 were used, keeping the wordsize (11) and the filter (low complexity) constant. The following table show the results:

Table I **No of hits for varying expect values**

Expect value (e)	BlastN	BlastP	BlastX	TblastN	Tblastx	PSI-Blast
10	163	100	100	100	101	501
0.1	157	100	100	101	100	501
1e-30	65	80	58	75	98	480

The results from expect=10 returned 163 hits, expect=0.1 returned 157 hits, and expect=1e-30 returned only 65 hits. The expect value is the measure of how many times the sequence could hit another by chance. By decreasing this value, the blast becomes more stringent and less results are returned.
In the same manner, the protein sequence of PRDX1 was blasted against the non-redundant protein databases, **BlastP, BalstX, TblastX, TblastN and PSI-Blast**. Again, the expect value was varied while keeping the word size (3) constant. The results from the expect values of 10 and 0.1 both returned almost 100 hits nad in PSI-BLAST it gives 501 hits, meaning that a decrease in stringency by 100x yields no difference. However, when an expect value of 1e-30 was used, only 58 hits were returned. The protein sequences in the database aligned so well with the PRDX1 protein sequence that only very low expect values altered the output.

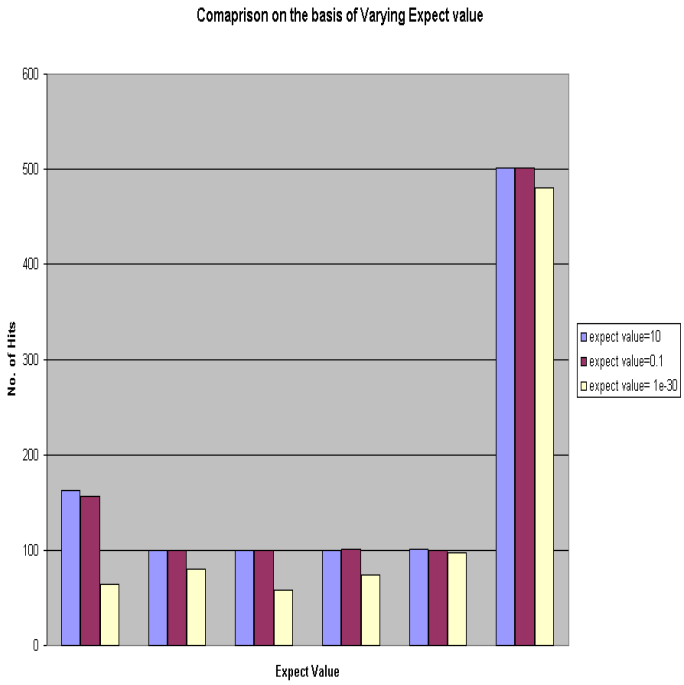

Fig. 1.5 **Comparison - Varying Expect Values**

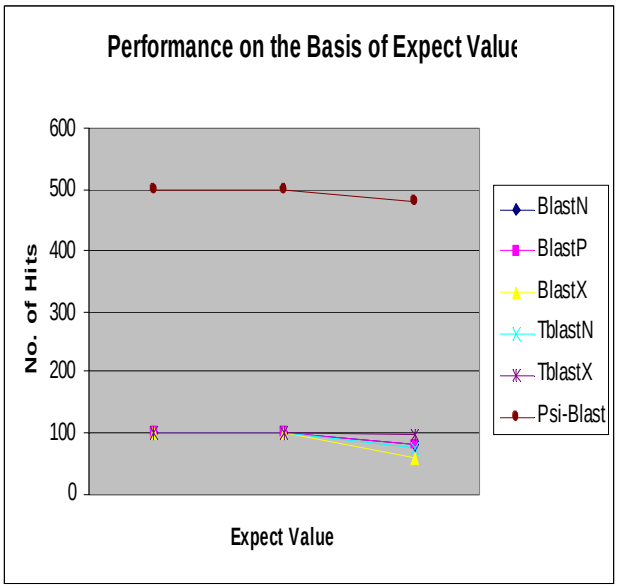

Fig. 1.6 **Comparison - Varying Expect Values**

By lowering the value by just 100th does not make much difference in number of hits in BlastP, BalstX, TblastX, BlastP.But variation comes when the expect value is reduced by a large factor.But as it can be seen from the graph , irrespective of the same input parameters given to all the variants,PSI-BLAST and BLASTN gives the maximum output.

3.2 Comparison On The Basis of – Word Size

Similar to the above experiment, a **BlastN** was performed using PRDX1 mRNA. This time, the expect value was held constant at 10 while the word size was changed (7, 11, 15). Also, other variables such as the nr database and the low complexity filter were similarly used. The following results were observed.

Table II **No of hits for varying expect values BlastN**

Word Size (w)	BlastN
7	163
11	163
15	139

The results showed that both a wordsize of 7 and 11 returned 163 hits while a wordsize of 15 returned only 139 hits. Wordsize is a measure of how many items, nucleotides in this case, are taken and compared to the database. In a wordsize of 11, a group of 11 sequential nucleotides are compared with the database. The larger the wordsize, the more stringent the analysis. That is why a wordsize of 15 returned less results

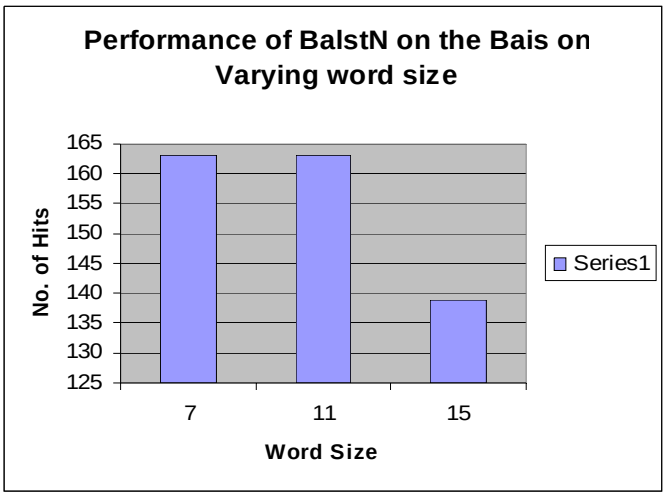

Fig. 1.7 **Varying Expect Values for BlastN**

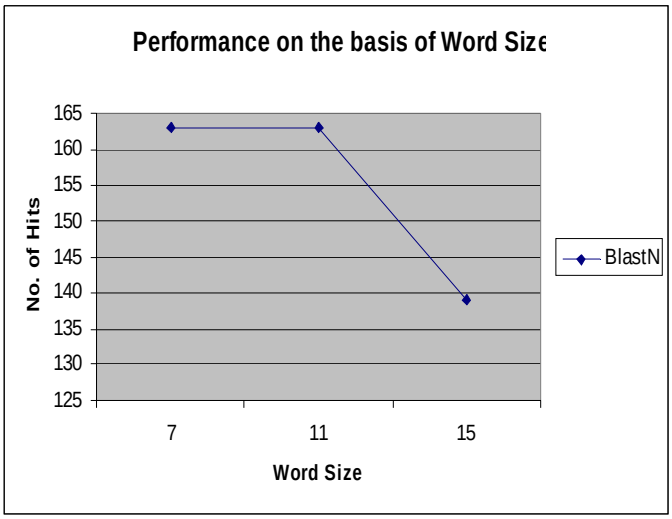

Fig. 1.8 **Varying Expect Values BlastN**

Wordsize can also be varied in a **BlastP, BalstX, TblastX, TblastN and PSI-Blast**. In the next comparison, PRDX1 protein was blasted against the protein database using a constant expect value (1e-70), database (nr), and filter (low complexity). Wordsize was varied between 2 and 3.

Table III **No of Hits for Varying Word Size**

Word size (w)	BlastP	BlastX	TblastX	TblastN	PSI
2	58	100	100	115	501
3	58	100	57	115	501

Varying word size does not affect the performance of BlastP,BlastN,TBlastN,TBlastP and PSI_Blast.But it only affects the performance of TblastX.Performance of TBlastX declines with the increase of word size.

BlastP BlastX TbalstX TbalstN PSI

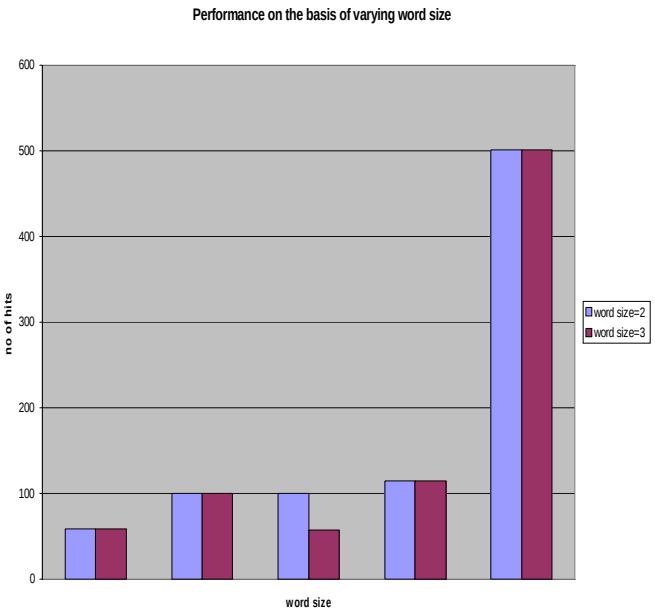

Fig. 1.9 **Varying Expect Values for variants**

3.3 Comparison on the Basis of Execution Time

All the variants were executed on 32-bit and 64-bit processors and their performance was compared in terms of seconds and number of processors, which is shown below**.**

Table IV **Varying Execution Time**

TEST	NUMBER OF CPUs	32-BIT TIME (in seconds)	64-BIT TIME (in seconds)
blastX	1	1516	1085
blastX	2	751	550
blastN	1	297	252
blastN	2	153	132
tblastX	1	4999	3545
tblastX	2	2761	1940

The observations are represented in the graph as shown below:

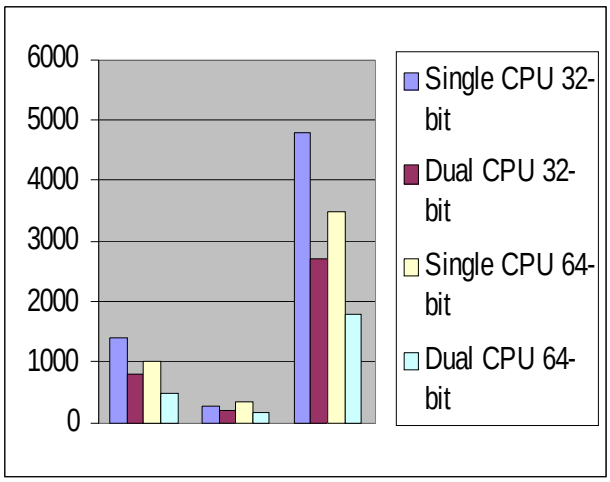

Fig. 1.10 **Compares the performance of BLAST compiled with 32-bit and 64-bit processor**

4.0 CONCLUSION AND FUTURE SCOPE

Variants of Blast (BlastN, BlastP, BlastX, TblastN, and TbalstX) run on different parameters, different algorithms, and each tool have different performance criteria. The performances differ on the basis of parameters like Word Size, Expect Value, and Databases Available. By selecting different values the efficiency of each tool can be improved. In this chapter the performance is being checked on the basis of execution time, and varying parameters and algorithm comparison. On the performance we can make decision that in which situation which tool is to be used.

Over the past decade many biological tools have been developed, but still improvements are needed in these tools , to improve the speed and accuracy. Research for improvements of existing tools is carrying on. Examinations of the problems arising from the use of biological tools is being noticed. Professionals are analyzing the effects of the execution of the code, how the execution effects the performance. Modifications have to be done to the source code.

By doing modifications to the existing parameters and source code, speed will increase and the field of bioinformatics will emerge with and more dynamic scope.

"***Measurement and Analysis is the key to Development and Improvement"***

So with continuous evaluations of existing versions of biological tools, further improvements will be possible.

5.0 REFERENCES

[1] By blast-help group, NCBI User Service, "BLAST Program Selection Guide**",** NCBI, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894

[2] Dan E. Krane, Michel L. Raymer Fundamentals concepts of Bioinformatics",PearsonEducation, 2003.

[3] Jean Michel Claverie and Cedric Notredame "Bioinformatics A Beginners Guide", Wiley Publishing, Inc. 2003.