BioBCDM: A novel integrated tool in Sequence alignment

[1] Deepalakshmi . R, [2] Dr. JothiVenkateswaran. C
[1] Research Scholar , [2] Research Supervisor, Associate Professor & Head
PG & Research Dept. of Computer Science, Presidency College - Chennai, India
[1]profdeepalakshmi@gmail.com, [2]jothivenkateswaran@yahoo.co.in

---

***Abstract*— Bioinformatics has emerged as an associate part of life sciences and biomedical analysis and most important in drug design analysis. Existing bioinformatics tools don't cross talk leading biologists to pay longer time in formatting the output from one tool as input for another tool. This results in huge loss of time and value. We therefore have created a platform that integrates the tools in a way that the output of one program may be directly used as an input of another and doesn't need any modifications. Tools for similarity search are needed in majority of all biological research. Thus, we tend to start integrating BLAST, ClustalW and Dotmatcher tools named BioBCDM tool which optimizes the time spend in browsing and downloading applications and is an interactive, effective and user friendly tool.**

***Index Terms*— Integrative, BLAST, ClustalW, Dotmatcher, BioBCDM tool, Bioinformatics**

Availability: http://biobcdm.in/blastdatabase.php , http://biobcdm.in/blastTool.php

I. INTRODUCTION

The recent advances in life sciences and technology have begun to turn out a large amount of data in a very quick and economical approach which requires the development of algorithms and parallel computing. Besides, biologists are sometimes non-programmers, therefore exacting intuitive computer applications that are simple to use by means that of a friendly Graphical User Interface. Many tools are generated over the past decade so as to cope with the data generation, however very little in integrating the tools and creating biologist friendly interfaces. It is therefore of most importance to beat such limitations, so that bioinformatics become far more widely used amongst biologists. Internet based interfaces are smart and common. They can access application program and perform needed analysis in a simple to use manner. The main goal of our project was to unify some of these existing bioinformatics applications and in one easy-to-use surroundings, freelance of the computing platform, being a concentrator resource tool with a friendly interface permitting intuitive bioinformatics tool usage. Our platform BioBCDM tool is a graphical interface integrating BLAST, ClustalW, and Dotmatcher tools permitting even non-programmer laboratory scientists to chain completely different processes into workflows and customize them without code writing. BLAST used as a single tool follows local alignment algorithm and does not necessarily return a complete match[1].

II. BIOBCDM: NEED OF THE DAY

Very few bioinformatics tools are available in recent day; each one has its own superiority and limitations. Only some tools perform multifunction. The new tool has the advantage of performing local and multiple sequence alignment. Simple gap function offers comparatively poor alignments. Alignments ought to be worked over retrospectively by another system. Output of aligned sequences is not in a customary format for input to alternative programs, thus porting the found information can generate a lot of fiddling. The poor alignments involving divergent sequence may be somewhat offset by manually increasing the gap penalty while reducing the gap extension penalty [2]. It is well confused by long stretches of unalignable sequences among otherwise well connected sequences. The user has got to use subjective criterion to make your mind decide when to cut off the search, or it's going to branch into another family connected solely by chance similarity. Often produces a form of false objectivity, where the user has fiddled with the program parameters to attain a subjectively pleasing result, instead of simply manually editing the alignment [2]. It has no applied math evaluatory properties. It'll produce an alignment whether or not the provided sequences are connected or not.

III. NEED FOR INTEGRATION

Rapid advances within the field of computers coupled with increasing computer literacy among professionals favour the implementation of computer applications in biological field. Further, the supply of various databases on the web has revolutionized the means by that a medical person devices a method for treatment [3]. Thus it is apt to conclude that the each tool has some disadvantages and when integrated will work better with a workflow. As several bioinformatics package tools are usually concerned in analysis tasks, scientists are more and more requiring that these heterogeneous bioinformatics tools be integrated in a uniform means. They are additionally requiring graphical user interfaces of these tools, and the ability to compose workflows without abundant programming effort. Online services based framework helps uniform integration of command-line bioinformatics software tools [4].

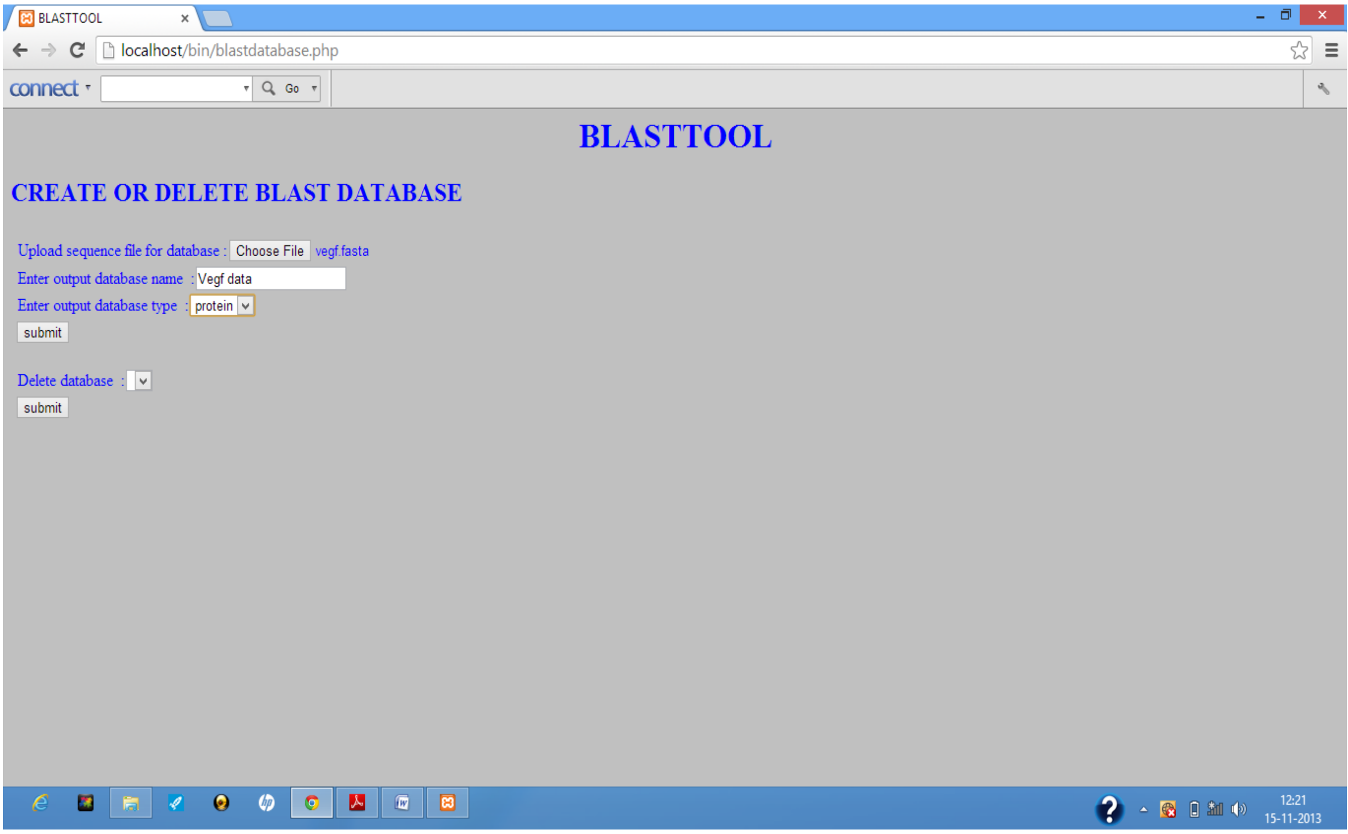
IV. IMPLEMENTATION

Architecture

- **HTML** stands for Hyper Text Mark up Language used to create web pages. Website authors use Hyper Text Mark-up Language to format text as titles and headings, to rearrange graphics on a webpage, to link to completely different pages among a website, and to link to various websites.
- **PHP:** PHP stands for Hypertext Pre-processor. It's a server-side scripting language that powers some of the most popular websites in the world, including Word Press and Face book. It is open source, which is easy to learn, and works well with MySQL, making it a good choice for web developers.
- **XAMPP** is an open source and free cross-platform web server solution stack package, consisting primarily of the Apache HTTP Server, MySQL database, and interpreters for scripts written in the PHP and Perl programming languages.

V. CREATION AND DELETION OF DATABASE

Database creation and deletion is of most pragmatic value to the BioBCDM tool. Fig.1 gives the front page for the user. Here one can upload the database which he has created. Sequences stored in database should be in FASTA format. We can store protein as well as DNA sequences. Unwanted databases can also be deleted in the same page. The database can be checked by the list box given.



BLASTTOOL

CREATE OR DELETE BLAST DATABASE

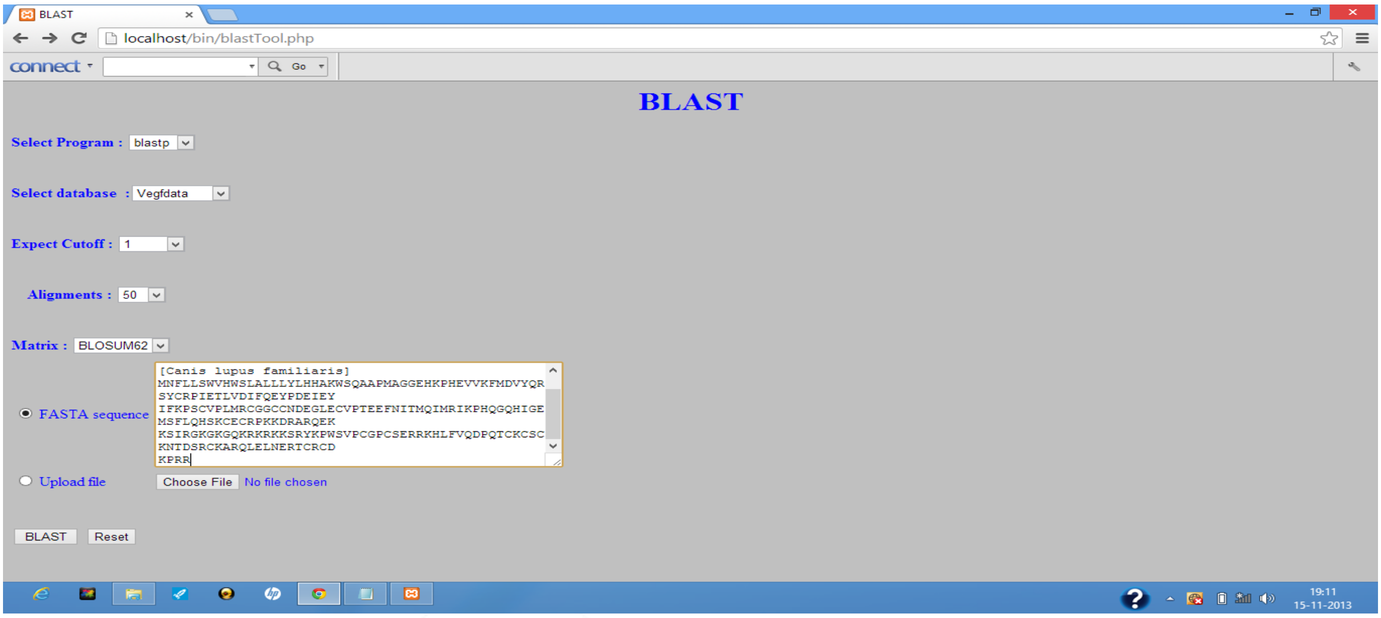
Upload sequence file for database : Choose File vegf.fasta
Enter output database name : Vegf data
Enter output database type : protein
submit

Delete database :
submit

Fig1: Tool where we can create or delete the database. Here the user uploads his own database.

VI. BLAST TOOL

The first step in interweaving the tools is to create the input and output file formats. The input can be Uploaded or Pasted in Fasta format. Fig2. below shows the homepage of Blast tool from where other tools are interlinked [5]. At first program can be selected of our choice, it can be Blastp, Blastx, Blastn, tblastp or tblastx. Database which we have created and uploaded is selected. Expected cut off and alignment values are entered as required by the user. Matrix values such as PAM30, PAM70, BLOSUM45,62 or 70 is selected. Fasta sequence which is to be compared with the database is pasted or uploaded from a file. Now the Blast is performed with the given input.



BLAST

Select Program : blastp

Select database : Vegfdata

Expect Cutoff : 1

Alignments : 50

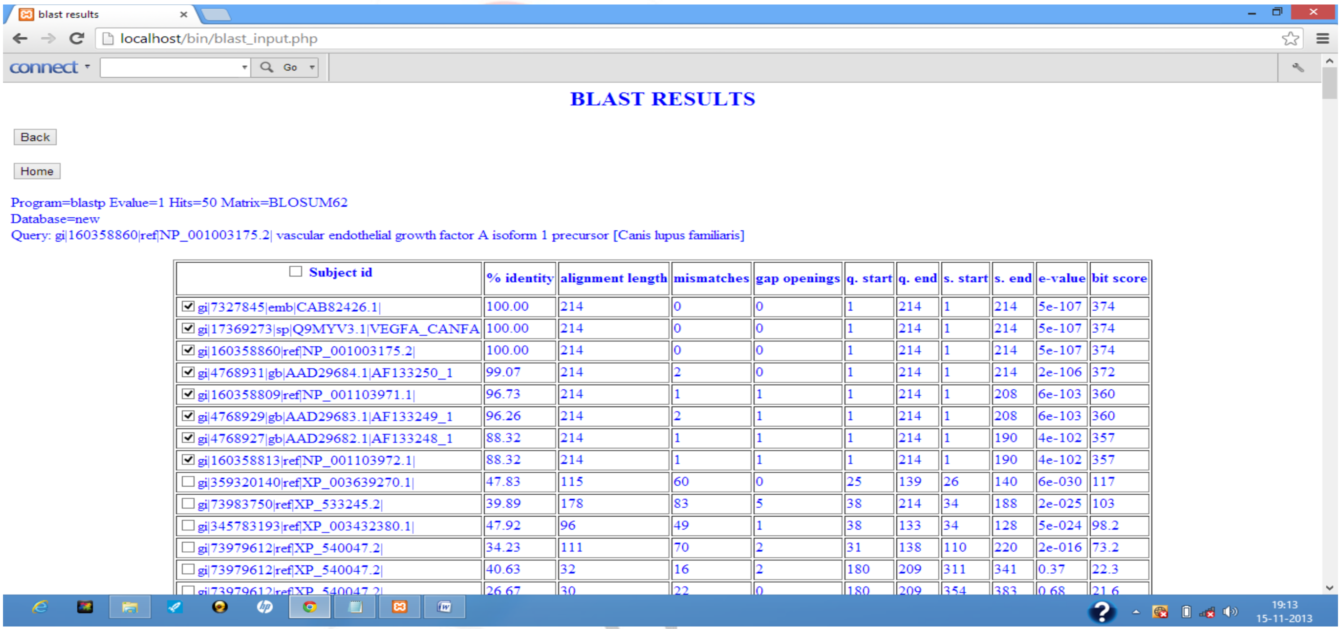
Matrix : BLOSUM62

● FASTA sequence

```
[Canis lupus familiaris]
MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQR
SYCRPIETLVDIFQEYPDEIEY
IFKPSCVPLMRCGGCCNDEGLECVPTEEFNITMQIMRIKPHQGQHIGE
MSFLQHSKCECRPKKDRARQEK
KSIRGKGKGQKRKRKKSRYKPWSVPCGPCSERRKHLFVQDPQTCKCSC
KNTDSRCKARQLELNERTCRCD
KPRR
```

○ Upload file Choose File No file chosen

Fig 2: BLAST homepage with the sequence pasted for comparison.

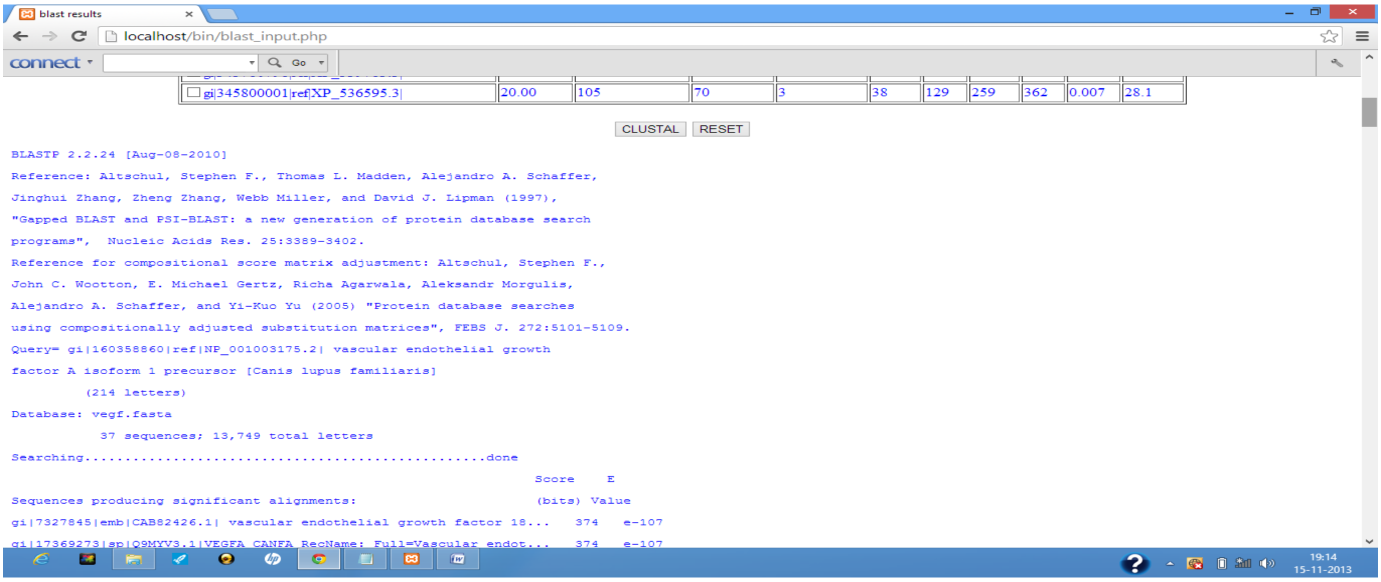


BLAST RESULTS

Program=blastp Evalue=1 Hits=50 Matrix=BLOSUM62
Database=new
Query: gi|160358860|ref|NP_001003175.2| vascular endothelial growth factor A isoform 1 precursor [Canis lupus familiaris]

☐ Subject id	% identity	alignment length	mismatches	gap openings	q. start	q. end	s. start	s. end	e-value	bit score
☑ gi|7327845|emb|CAB82426.1|	100.00	214	0	0	1	214	1	214	5e-107	374
☑ gi|17369273|sp|Q9MYV3.1|VEGFA_CANFA	100.00	214	0	0	1	214	1	214	5e-107	374
☑ gi|160358860|ref|NP_001003175.2|	100.00	214	0	0	1	214	1	214	5e-107	374
☑ gi|4768931|gb|AAD29684.1|AF133250_1	99.07	214	2	0	1	214	1	214	2e-106	372
☑ gi|160358809|ref|NP_001103971.1|	96.73	214	1	1	1	214	1	208	6e-103	360
☑ gi|4768929|gb|AAD29683.1|AF133249_1	96.26	214	2	1	1	214	1	208	6e-103	360
☑ gi|4768927|gb|AAD29682.1|AF133248_1	88.32	214	1	1	1	214	1	190	4e-102	357
☑ gi|160358813|ref|NP_001103972.1|	88.32	214	1	1	1	214	1	190	4e-102	357
☐ gi|359320140|ref|XP_003639270.1|	47.83	115	60	0	25	139	26	140	6e-030	117
☐ gi|73983750|ref|XP_533245.2|	39.89	178	83	5	38	214	34	188	2e-025	103
☐ gi|345783193|ref|XP_003432380.1|	47.92	96	49	1	38	133	34	128	5e-024	98.2
☐ gi|73979612|ref|XP_540047.2|	34.23	111	70	2	31	138	110	220	2e-016	73.2
☐ gi|73979612|ref|XP_540047.2|	40.63	32	16	2	180	209	311	341	0.37	22.3
☐ gi|73979612|ref|XP_540047.2|	26.67	30	22	0	180	209	354	383	0.68	21.6

Fig 3: User can check the box for which the particular sequences alone will be compared.



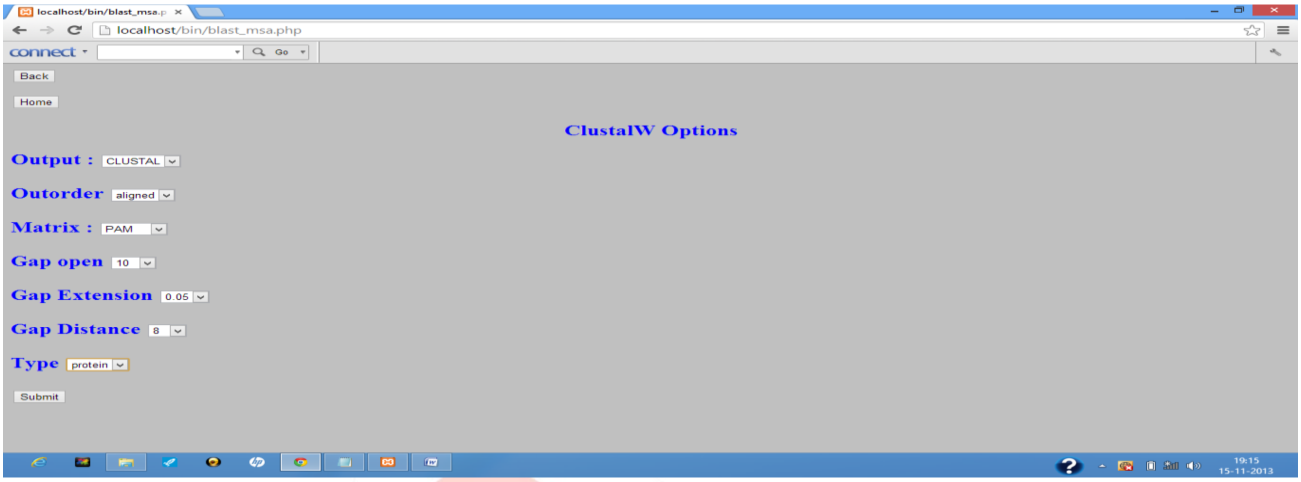
Subject id	% identity	alignment length	mismatches	gap openings	q. start	q. end	s. start	s. end	e-value	bit score
☐ gi|345800001|ref|XP_536595.3|	20.00	105	70	3	38	129	259	362	0.007	28.1

```
BLASTP 2.2.24 [Aug-08-2010]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.
Reference for compositional score matrix adjustment: Altschul, Stephen F.,
John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis,
Alejandro A. Schaffer, and Yi-Kuo Yu (2005) "Protein database searches
using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.
Query= gi|160358860|ref|NP_001003175.2| vascular endothelial growth
factor A isoform 1 precursor [Canis lupus familiaris]
         (214 letters)
Database: vegf.fasta
           37 sequences; 13,749 total letters
Searching..................................................done
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value
gi|7327845|emb|CAB82426.1| vascular endothelial growth factor 18...   374   e-107
gi|17369273|sp|Q9MYV3.1|VEGFA_CANFA RecName: Full=Vascular endot...   374   e-107
```

Fig 4:Blast results are displayed and can view the clustal result by clicking the button

VII. INTEGRATING WITH CLUSTALW AND DOTMATCHER

BLAST **[6]**, ClustalW **[7]** and Dotmatcher **[8]** were chosen for integration using PHP coding. Of the 3 tools only BLAST requires database storage. Separate page is created for Database addition and deletion. Here the user can create own database and start adding it to the BLAST database. The interlinking of the tools was done using PHP programs. Fig 5, 6 and 7 below shows the CLUSTALW comparison and the output displayed.



ClustalW Options

Output : CLUSTAL

Outorder aligned

Matrix : PAM

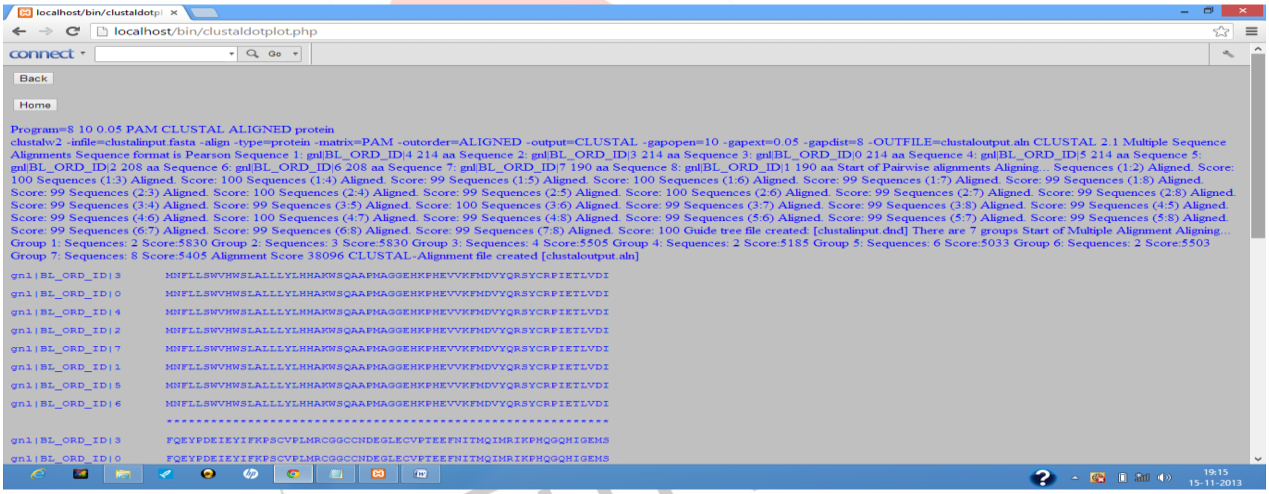
Gap open 10

Gap Extension 0.05

Gap Distance 8

Type protein

Fig 5 Matrix, Gap open, Extension, Distance and type of sequence are entered in Clustal homepage

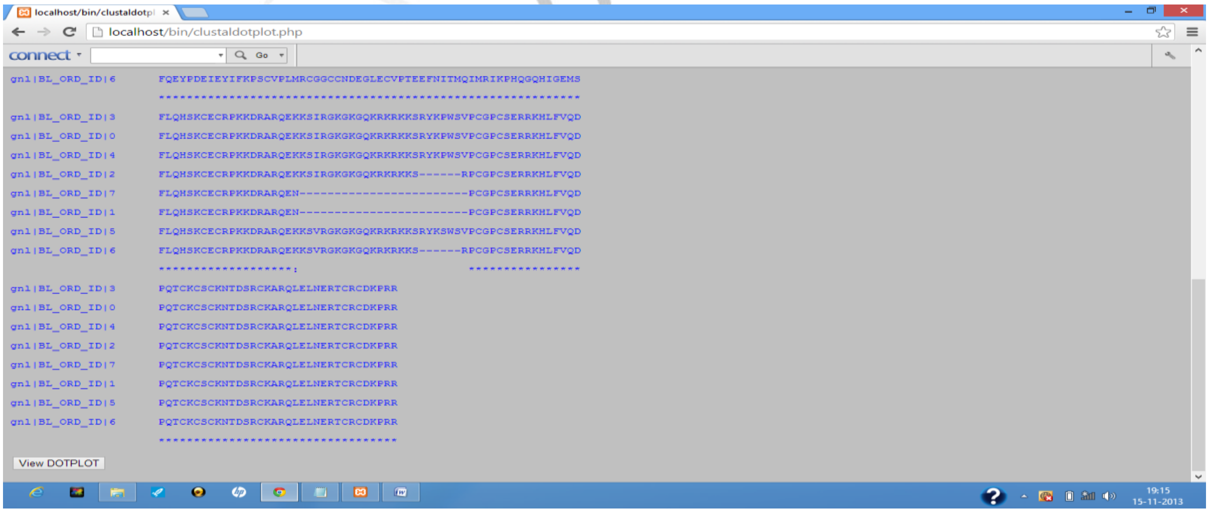


```
Program=8 10 0.05 PAM CLUSTAL ALIGNED protein
clustalw2 -infile=clustalinput.fasta -align -type=protein -matrix=PAM -outorder=ALIGNED -output=CLUSTAL -gapopen=10 -gapext=0.05 -gapdist=8 -OUTFILE=clustaloutput.aln CLUSTAL 2.1 Multiple Sequence Alignments Sequence format is Pearson Sequence 1: gnl|BL_ORD_ID|4 214 aa Sequence 2: gnl|BL_ORD_ID|3 214 aa Sequence 3: gnl|BL_ORD_ID|0 214 aa Sequence 4: gnl|BL_ORD_ID|5 214 aa Sequence 5: gnl|BL_ORD_ID|2 208 aa Sequence 6: gnl|BL_ORD_ID|6 208 aa Sequence 7: gnl|BL_ORD_ID|7 190 aa Sequence 8: gnl|BL_ORD_ID|1 190 aa Start of Pairwise alignments Aligning... Sequences (1:2) Aligned. Score: 100 Sequences (1:3) Aligned. Score: 100 Sequences (1:4) Aligned. Score: 99 Sequences (1:5) Aligned. Score: 100 Sequences (1:6) Aligned. Score: 99 Sequences (1:7) Aligned. Score: 99 Sequences (1:8) Aligned. Score: 99 Sequences (2:3) Aligned. Score: 100 Sequences (2:4) Aligned. Score: 99 Sequences (2:5) Aligned. Score: 100 Sequences (2:6) Aligned. Score: 99 Sequences (2:7) Aligned. Score: 99 Sequences (2:8) Aligned. Score: 99 Sequences (3:4) Aligned. Score: 99 Sequences (3:5) Aligned. Score: 100 Sequences (3:6) Aligned. Score: 99 Sequences (3:7) Aligned. Score: 99 Sequences (3:8) Aligned. Score: 99 Sequences (4:5) Aligned. Score: 99 Sequences (4:6) Aligned. Score: 100 Sequences (4:7) Aligned. Score: 99 Sequences (4:8) Aligned. Score: 99 Sequences (5:6) Aligned. Score: 99 Sequences (5:7) Aligned. Score: 99 Sequences (5:8) Aligned. Score: 99 Sequences (6:7) Aligned. Score: 99 Sequences (6:8) Aligned. Score: 99 Sequences (7:8) Aligned. Score: 100 Guide tree file created: [clustalinput.dnd] There are 7 groups Start of Multiple Alignment Aligning... Group 1: Sequences: 2 Score:5830 Group 2: Sequences: 3 Score:5830 Group 3: Sequences: 4 Score:5505 Group 4: Sequences: 2 Score:5185 Group 5: Sequences: 6 Score:5033 Group 6: Sequences: 2 Score:5503 Group 7: Sequences: 8 Score:5405 Alignment Score 38096 CLUSTAL-Alignment file created [clustaloutput.aln]

gnl|BL_ORD_ID|3      MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQRSYCRPIETLVDI
gnl|BL_ORD_ID|0      MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQRSYCRPIETLVDI
gnl|BL_ORD_ID|4      MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQRSYCRPIETLVDI
gnl|BL_ORD_ID|2      MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQRSYCRPIETLVDI
gnl|BL_ORD_ID|7      MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQRSYCRPIETLVDI
gnl|BL_ORD_ID|1      MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQRSYCRPIETLVDI
gnl|BL_ORD_ID|5      MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQRSYCRPIETLVDI
gnl|BL_ORD_ID|6      MNFLLSWVHWSLALLLYLHHAKWSQAAPMAGGEHKPHEVVKFMDVYQRSYCRPIETLVDI
                     ************************************************************

gnl|BL_ORD_ID|3      FQEYPDEIEYIFKPSCVPLMRCGGCCNDEGLECVPTEEFNITMQIMRIKPHQGQHIGEMS
gnl|BL_ORD_ID|0      FQEYPDEIEYIFKPSCVPLMRCGGCCNDEGLECVPTEEFNITMQIMRIKPHQGQHIGEMS
```

Fig 6 Comparison of sequences done using CLUSTAL tool

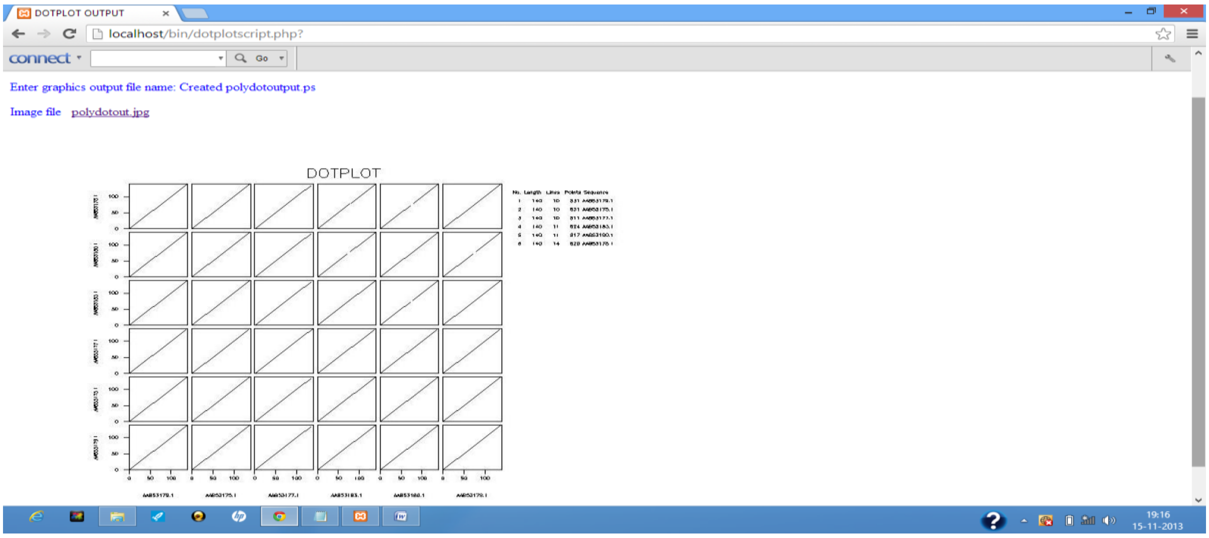


```
gnl|BL_ORD_ID|6      FQEYPDEIEYIFKPSCVPLMRCGGCCNDEGLECVPTEEFNITMQIMRIKPHQGQHIGEMS
                     ************************************************************

gnl|BL_ORD_ID|3      FLQHSKCECRPKKDRARQEKKSIRGKGKGQKRKRKKSRYKPWSVPCGPCSERRKHLFVQD
gnl|BL_ORD_ID|0      FLQHSKCECRPKKDRARQEKKSIRGKGKGQKRKRKKSRYKPWSVPCGPCSERRKHLFVQD
gnl|BL_ORD_ID|4      FLQHSKCECRPKKDRARQEKKSIRGKGKGQKRKRKKSRYKPWSVPCGPCSERRKHLFVQD
gnl|BL_ORD_ID|2      FLQHSKCECRPKKDRARQEKKSIRGKGKGQKRKRKKS------RPCGPCSERRKHLFVQD
gnl|BL_ORD_ID|7      FLQHSKCECRPKKDRARQEN------------------------PCGPCSERRKHLFVQD
gnl|BL_ORD_ID|1      FLQHSKCECRPKKDRARQEN------------------------PCGPCSERRKHLFVQD
gnl|BL_ORD_ID|5      FLQHSKCECRPKKDRARQEKKSVRGKGKGQKRKRKKSRYKSWSVPCGPCSERRKHLFVQD
gnl|BL_ORD_ID|6      FLQHSKCECRPKKDRARQEKKSVRGKGKGQKRKRKKS------RPCGPCSERRKHLFVQD
                     *******************:                        ****************

gnl|BL_ORD_ID|3      PQTCKCSCKNTDSRCKARQLELNERTCRCDKPRR
gnl|BL_ORD_ID|0      PQTCKCSCKNTDSRCKARQLELNERTCRCDKPRR
gnl|BL_ORD_ID|4      PQTCKCSCKNTDSRCKARQLELNERTCRCDKPRR
gnl|BL_ORD_ID|2      PQTCKCSCKNTDSRCKARQLELNERTCRCDKPRR
gnl|BL_ORD_ID|7      PQTCKCSCKNTDSRCKARQLELNERTCRCDKPRR
gnl|BL_ORD_ID|1      PQTCKCSCKNTDSRCKARQLELNERTCRCDKPRR
gnl|BL_ORD_ID|5      PQTCKCSCKNTDSRCKARQLELNERTCRCDKPRR
gnl|BL_ORD_ID|6      PQTCKCSCKNTDSRCKARQLELNERTCRCDKPRR
                     **********************************
```

Fig 7 After verifying CLUSTAL result user can view the DOTPLOT by clicking the button.



Enter graphics output file name: Created polydotoutput.ps

Image file polydotout.jpg

```
No. Length Lines Points Sequence
  1    140    10    531 AAB53179.1
  2    140    10    531 AAB53175.1
  3    140    10    511 AAB53177.1
  4    140    11    524 AAB53183.1
  5    140    11    517 AAB53180.1
  6    140    14    520 AAB53178.1
```

Fig 8 Multiplotgraph is displayed showing the similar sequence by a perfect diagonal line and non similar sequences are denoted by small gaps in diagonal line.

VIII. COMPARISON WITH OTHER INTEGRATED TOOLS

So far we have only a handful of integrated BioTools. Database Creation and Deletion are not included in any of the integrated tools and this is one of the added advantages of using BioBCDM. To list Bioparisodhana[**9**] has integrated Blast, Clustal and Primer but has not created any space for database creation. Some of the earlier tools like BioExtract**[10]** Discovery Net **[11]** and in GAP **[12],** where the tools exist as individual tools and work flows can be created by the users have provided lot of help. We believe that BioBCDM platform can give the much needed simpler and easier approaches for biologists to use the tools and analyze the data without much effort.

IX. CONCLUSION

Bioinformatics is a rapid processing field. Both the experimental technologies and the computer based methods are in dynamic phase of development. While some years ago human experts would check every program output, nowadays sequence analysis routines are being applied in an automatic fashion creating annotation that is included in various databases. Many of the Bioinformatics tools exist individually and the need for integrated tool arises mainly to save time for the users and to facilitate easy pavement for the biologists to get the required output in minimal interval of time. Although the quality of many existing tools has increased dramatically, the possibility of error and in particular its perpetuation by further automatic methods exists. Certainly, the BioBCDM Tool will be an optimum tool for the biologists. As there is plenty of tool emerging in Bioinformatics, these types of integrated tools will be of great use in minimising the work. Our future work will include integrating other useful tools.

X. REQUIREMENTS

Project name : BioBCDM
Project home page : http://biobcdm.in/blastdatabase.php
Operatingsystem(s): Win XP,Win 7 or Win 8

REFERENCES

[1] Jian Ye1*, George Coulouris1, Irena Zaretskaya1, Ioana Cutcutache2, Steve Rozen2and Thomas L Madden1, Primer-BLAST: "A tool to design target-specific primers for polymerase chain reaction", BMC Bioinformatics 2012
[2] biochem.uthscsa.edu/~hs_lab/frames/molgen/tutor/compare.html, March 22,2003.
[3] M. Madan Babu, Need for integration,Centre for Biotechnology, 2010.
[4] Badidi, Serhani, Bouktif Innovations in Information Technology, IIT Conference, 2008.
[5] Stajich J et al. "The Bioperl toolkit: Perl modules for the life sciences". [PMID: 12368254]Genome Res. 2002 **12:** 1611,
[6] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, "Basic Local Alignment Search Tool", Mol Biol, 1990 Oct 5;215(3):403-10,[PMID: 2231712]
[7] Chenna R et al. "Multiple Sequence alignment with the Clustal series of programs", Nucleic Acids Res. 2003 **31**: 3497 [PMID:12824352]
[8] Steve Rozen& Helen J, Humana Press. 2000 pp 365-386
[9] RajaniKanthVangala*, Lucky Singh$ & Ravi Prakash Gupta, " BioParishodhana: A novel graphical interfaceintegrating BLAST, ClustalW, primer3 and restriction digestion tools" Pubmed 2012.
[10] Lushbough CM, Bergman MK, Lawrence CJ, Jennewein D, Brendel V, "Implementing Bioinformatic Workflows within the Bioextract Workflows Within The Bioextract Server". Int J ComputBiol Drug Des. 2008 **1**: 302, 2008;1(3):302-12. [PMID: 20054995]

[11] Rowe A, Kalaitzopoulos D, Osmond M, Ghanem N, Guo Y, "The discovery net system for high throughput bioinformatics, 2003;19 Suppl 1:i225-31.Bioinformatics. 2003 **19**: 225 [PMID: 12855463]

[12] Qi J, Zhao F, Buboltz A, Schuster SC, "Ingap: an integrated next-generation genome analysis pipeline". Jan 1;26(1):127-9. doi: 10.1093/bioinformatics/btp615. Epub 2009 Oct 30. [PMID: 19880367].

Authors Biography

Dr. C. JothiVenkateswaran working as a Head and Associate Professor , PG & Research Dept. of Computer Science, Presidency College, Tamilnadu, India. He has been serving more than 25 years of teaching experience and more than 13 years of research experience in the field of Data mining, Algorithm Analysis, Geographical Data mining, Bioinformatics and Image Mining. He served different positions as academician and successfully completed different projects. He has published many articles in the National and International Journals and has presented papers in many conferences.

R.Deepalakshmi received the B.Sc Mathematics in 1999 and MCA in 2002 both from Madras University, India and M.Phil in Computer science from Madurai Kamaraj University .She is working as a Head & Assistant Professor, Department of Computer Science & Applications, Sir Theagaraya College , Chennai -21. She has 10 years of teaching experience. Her research interest includes Data Mining & Bioinformatics.