Journal of Proteomics & Bioinformatics - Open Access

Short Communication
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000155

JPB/Vol.3 Issue 11

Mass Blaster V1.0 – A Perl Gui Tool for Mass Sequence Blast and Gene Prediction

Saravanan Vijayakumar

Department of Bioinformatics, SRM Arts and Science College, Kaatankulathr, Tamil Nadu, India

Abstract

Sequence similarity search is the principle technique, in sequence analysis, adapted for understanding the biological significance of a sequence. Of the various sequence similarity search tool, BLAST is the widely and commonly used tool for local alignment search. Even though several services like NCBI-BLAST, EBI-BLAST, and GENEBEE BLAST has the facility to perform a routine basic local alignment search for the given protein or nucleotide sequence, none of the service has equipped with mass sequence submission. To overcome this, we developed MASS BLASTER V1.0, which allows the user to submit multiple sequences at-a-time to the services like NCBI-Protein BLAST, NCBI-Nucleotide BLAST, GENEBEE BLAST, RNA NON-CODING BLAST, COGNITOR, GENEMARK HMM, and GLIMMER and saves the corresponding results to the user system without human interference.

Keywords: BLAST; Multiple sequence submission; GenemarkHMM; Cog; RNA Database

Introduction

There is a continuous need for similarity search and predictive tools for the isolation of protein function, coding region, non-coding region, genes, orthologs group and phylogenetic relation in modern biological research. Especially, similarity search tool like BLAST Altschul et al. (1990) and gene prediction tool GENEMARK HMM Lukashin and Borodovsky (1997) are very effective and trust worthy computational tools for locating domains and gene prediction respectively. Using such numerous tools in the discovery environment for routine analysis is often laborious and time-consuming. For example to perform a homology search for 200 different proteins for function identification, one has to submit the individual sequence to the BLAST service and then obtains the result. Some of the problems faced during this process are (1) Should do manually because most of the services do not have mass sequence submission for the BLAST search, even though NCBI BLAST has the multiple sequence submission facility, it has processing time limit restriction, also user has to wait for the individual sequence result until BLAST search is done for all the submitted sequences, (2) Error-prone – possibility of human error on doing repetitive task, (3) Time consuming – consume lots of time when done manually. So, MASS BLASTER (Figure 1) is developed keeping in mind that the process should be automated, to reduce time and error, and to produce the same result of what the service provider, without sacrificing the quality of the result.

Model development

The architecture of MASS BLASTER is shown in (Figure 2). The implementation of the system is carried out in PERL V5.12, perltk module is used to make Graphical User Interface (GUI) for the tool. The process flow of the tools is as follows: (1) the file containing multiple sequence[a] (FASTA formatted file) is loaded; (2) the sequences are parsed separately; (3) parameters for the corresponding tool is set; (4) connects to the service server and submits each and individual sequence to the service separately, therefore preventing the server side restriction for multiple sequence submission (5) obtains the result generated by the services and parse the information (E-Value..)[b] of the result; (6) the parsed information is aligned, tabulated and saved to the user desired location on the system in hypertext format. The services included in the MASS BLASTER are (1) NCBI BLAST – protein Altschul et al. (1990) (2) NCBI BLAST – Nucleotide Altschul et al. (1990) (3) GeneBee BLAST – Nucleotide Altschul et al. (1997) (4) RNA non-coding BLAST Altschul et al. (1997) (5) GENEMarkHMM – Prokaryotes and Eukaryotes Lukashin and Borodovsky (1997) (6) COgNitor and Tatusov et al. (2001) (7) GLIMMER Salzberg et al. (1998), Delcher et al. (1999).

Requirements

The tool requires PERL V.10.1 or higher, with the additional module "Tk" for graphical interface, "WWW::Mechanize" for connecting to the internet service. These two modules doesn't come with the standard PERL installer, but can be downloaded freely from the CPAN archive. The tool is compatible with both MS-Windows

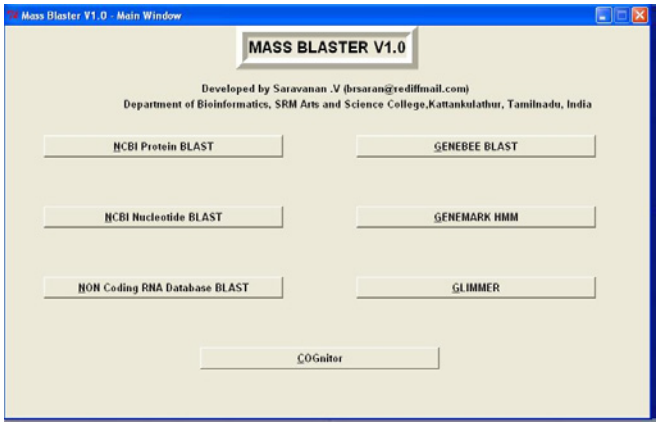

Figure 1: MASS BLASTER Main Window for accessing different services.

Corresponding author: Saravanan Vijayakumar, Department of Bioinformatics, SRM Arts and Science College, Kaatankulathr, Tamil Nadu, India, E-mail: brsaran@rediffmail.com

Received October 23, 2010; **Accepted** November 15, 2010; **Published** November 17, 2010

Citation: Saravanan V (2010) Mass Blaster V1.0 – A Perl Gui Tool for Mass Sequence Blast and Gene Prediction. J Proteomics Bioinform 3: 302-304. doi:10.4172/jpb.1000155

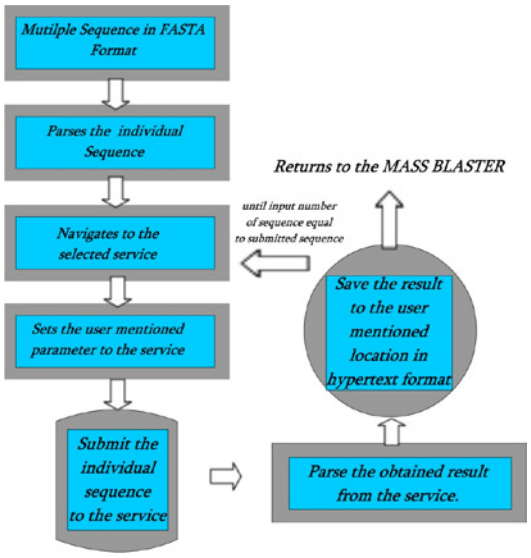

Figure 2: Architecture of MASS BLASTER.

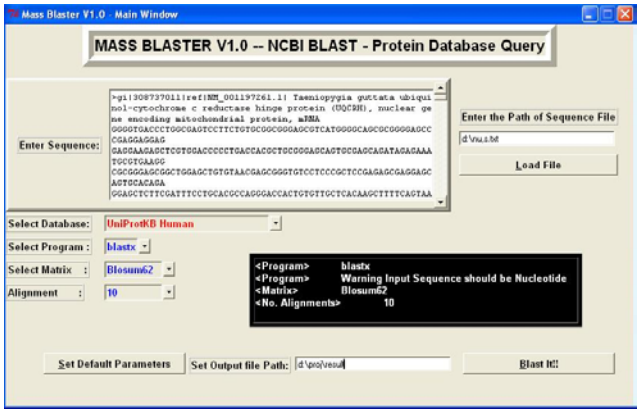

Figure 3: Snapshot of NCBI BLAST Protein service in MASS BLASTER.

platform [XP (service pack 2 or higher), VISTA, and Windows 7] as well as LINUX platform.

Input

The tool accepts both protein and nucleotide sequence in standard FASTA format. The number of sequence loaded on to the tool is not restricted, but it depends on the user systems memory and processor. With a minimum configuration (1GB RAM, Pentium IV or Higher processor, at least 10GB physical memory) at a time, up to 10,000[a] sequences can be loaded and submitted. User can select and set the various parameters provided by the corresponding services, like Database, program type (blastx/blastn), Matrix, etc. (Figure 3) The modification made to services will be indicated in the message window of the tool.

Output

Results from the service, for each submitted sequence, are hypertext formatted (hypertext format is preferred because as the tool being executed in multiple platform, HTML format can be easily viewed in any browser and no special software required for viewing the result) and stored in the system in user desired location. For the NCBI BLAST services (both nucleotide and protein) the information's like Hit ID, Accession, Definition, Length, Bit Score, E Value, Query From, Query To, Hit From, Hit To, Query Frame, Hit Frame, Identity, Positive, Align Length and Alignment are parsed and tabulated, for easy understanding, in hypertext format and saved in the system (Figure 4). For other services the result page content is stored as-such in hypertext format. The status of the submitted sequence and the progress of the results are displayed in status window (Figure 5).

[a]virtually unlimited sequence can be loaded, and it depends on system memory and processor
[b]for NCBI Blast services alone

Results and Discussion

This tool is used and tested while carrying out the Genome-wide analysis of intergenic regions in 11 species of *Mycoplasma*. To explore the coding regions in the 6840 intergenic sequence from the 11 genomes of *Mycoplasma* species, all the intergenic sequences are extracted from the genome and subjected to similarity search with BLASTX Altschul et al. (1990) program to explore the coding segment present in it. Also, the sequences are subjected to GENEMARK.HMM Lukashin and Borodovsky (1997) for the prediction of potential gene activity in the intergenic region. During the course of this work all the 6840 intergenic sequences are extracted from the corresponding genome sequence of *Mycoplasma* species and fed as input to the MASS BLASTER to perform the BLASTX and GENEMARK-HMM. It took approximately 22 hours and 42 minutes for performing the translated BLASTX and 4 hours 50 minutes for performing the GENEMARK-HMM for all the 6840 intergenic regions, with a internet speed of 1Mbps and Pentium IV processor 2 GB RAM running Microsoft Windows XP as operating system. The results of all the individual sequences are automatically stored as file in the system (existence of result files for all the 6840 input is verified manually) by the MASS BLASTER for further analysis. This clearly indicates the tool performance and usability in speeding up the process carried out using Bioinformatics prediction tool.

Blast Type: blastp; Blast Database: uniprotkb_pdb;
Blast Query Definition: AAA51411.1 AAA51411.1 albumin [Bos taurus]

Hit Number:	1
Hit ID:	lcl|SP:ALBU_HUMAN
Hit Accession:	SP:ALBU_HUMAN
Hit Definition:	P02768 Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2
Hit Length:	609
Bit Score:	992.645
E value:	0
Query From:	1
Query To:	606
Hit From:	1
Hit To:	607
Query Frame:	1
Hit Frame:	1
Identity:	465
Positive:	536
Align Length:	607

```
1     MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFKGLVLIA     50
      MKWVTFISLL LFSSAYSRGVFRRD HKSE+AHRFKDLGEE+FK LVLIA
1     MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIA     50

51    FSQYLQQCPFDEHVKLVNELTEFAKTCVADESHAGCEKSLHTLFGDELCK     100
      F+QYLQQCPF++HVKLVNE+TEFAKTCVADES    C+KSLHTLFGD+LC
51    FAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCT     100

101   VASLRETYGDMADCCEKQEPERNECFLSHKDDSPDLPKL-KPDPNTLCDE     150
      VA+LRETYG+MADCC KQEPERNECFL HKDD+P+LP+L +P+ + +C
101   VATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMCTA     150

151   FKADEKKFWGKYLYEIARRHPYFYAPELLYYANKYNGVFQECCQAEDKGA     200
      F  +E+ F  KYLYEIARRHPYFYAPELL++A +Y   F ECCQA DK A
151   FHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAA     200
```

Figure 4: Snapshot of result (NCBI BLAST Protein) parsed by MASS BLASTER.

Short Communication
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000155
JPB/Vol.3 Issue 11

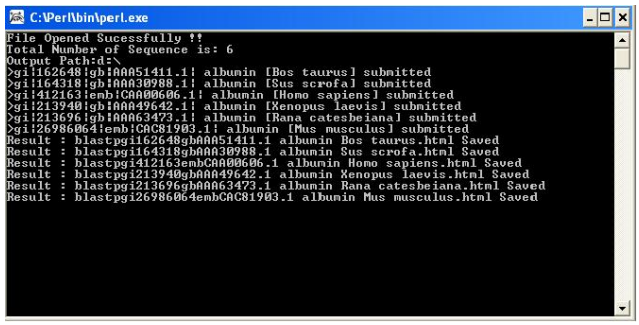

Figure 5: Status window of MASS BLASTER showing the progress of submitted sequence.

Conclusion

MASS Blaster is designed and developed in PERL and framed in a way to perform simple repetitive task in an ambiguity free manner. The tool is supplied as open source software, and hence one can study and change the software for further improvement. The intention of the work is to automate the regular and routine basic sequence analysis process, there by speeding up the biological research process that adapts Bioinformatics prediction tools for analysis.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
3. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636-4641.
4. Lukashin A, Borodovsky M 1997 GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26: 1107-1115.
5. Salzberg S, Delcher A, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26: 544-548.
6. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631-637.
7. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22-28.