DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4648

ISSN No. 0976-5697

Volume 8, No. 8, September - October 2017

International Journal of Advanced Research in Computer Science

TECHNICAL NOTE

Available Online at www.ijarcs.info

xmlBLASTparser V1.1 — A PHP BASED NCBI BLAST XML OUTPUT PARSER

T. Ashok Kumar
Department of Bioinformatics
Noorul Islam College of Arts and Science
Kumaracoil, Thuckalay, India

B. Rajagopal
Department of Zoology
Government Arts College
Dharmapuri, Tamil Nadu, India

Abstract: xmlBLASTparser is a lightweight PHP library for parsing an XML formatted output of NCBI BLAST sequence alignment and rendering into attractive web page. The biological database accession numbers present in each sequence alignment hit have properly hyperlinked to the original source. Moreover, hit ids in the description summary is anchor hyperlinked to the corresponding sequence alignment section. The xmlBLASTparser library can be easily embedded or integrated in a web page at server-side through standalone NCBI BLAST software or RESTful web service of NCBI BLAST. The output of xmlBLASTparser has the same flavour of the online NCBI BLAST. xmlBLASTparser is freely available under terms of GNU General Public License version 3 (GPLv3), at https://github.com/AshokHub/xmlBLASTparser.

Keywords: xmlBLASTparser; PHP library; Sequence alignment; XML output parser; NCBI BLAST

I. INTRODUCTION

Sequence alignment is one of the well-known and most widely used method in bioinformatics/molecular biology for a broad range of applications such as sequence analysis, phylogenetic analysis, homology modeling, structural motif prediction, domain prediction, molecular fingerprint prediction, pattern matching, function prediction, genome fragment assembly, SNP analysis, biodata integration, etc. BLAST is a popular local sequence alignment tool originally developed by Steve Altschul and his team members at the National Institutes of Health (NIH) during 1990 [1-3]. Due to the high demand of BLAST tool, NCBI has released various BLAST programs based on the demand of life scientists. The different types of the BLAST program include BLASTN, BLASTP, BLASTX, TBLASTN, TBLASTX, IgBLAST, SmartBLAST, BLAT, MOLE-BLAST, WU BLAST, PSI-BLAST, PHI-BLAST, MegaBLAST, DELTA-BLAST, RPS BLAST, AB BLAST, CaBLAST, Parcel BLAST, BLASTZ, VecScreen, CDART, CD-search, GEO, Primer-BLAST, etc. [4]. Moreover, NCBI has extended their service through several modes such as Web BLAST, Stand-alone command line BLAST, WWW BLAST, Cloud BLAST, BLAST URL API, Remote BLAST+, and C++ BLAST API [5].

Through NCBI BLAST tool, we can able to perform three types of nucleotide or protein sequence comparisons: (*i*) pairwise sequence comparison, (*ii*) query sequence against the set of sequences (local database), and (*iii*) query sequence against the large set of sequences from the external biological databases. The commonly used database for sequence comparisons are NR, ENV NR, NCBI GenBank/RefSeq, DDBJ, RCSB PDB, SwissProt/UniProt, EBI EMBL, PIR, PRF, PAT, EST, and DBSTS. The NCBI BLAST delivers output of sequence alignment result in various types of file format which include ASN.1 (Text), ASN.1 (Binary), Hit Table (CSV), Hit Table (Text), HTML, JSON Seq-align, Multiple-file JSON, Multiple-file XML2, SAM, Single-file JSON, Single-file XML2, Text, and XML for download. Among them ASN.1, CSV, JSON, Text, and XML file format are program specific outputs used for analyzing the result using any programming languages. The default file format of the sequence alignment output is HTML [4,5]. xmlBLASTparser is a small PHP program used to parse the output of the NCBI BLAST sequence alignment result and generate a rich webpage with well formatted alignment.

II. METHODS

The output files of NCBI BLAST sequence alignment result are programming language specific and can be easily parsed for various sequence analysis. Most of the server-side applications have own user interactive graphical wrapper to execute and retrieve output of the sequence alignment result. BLASTphp is a simple PHP library used to wrap or embed NCBI BLAST tool in a web page and retrieve the output in various file formats [6]. In general, the output of sequence alignment can be categorized into four sections: (*i*) header section – consist of details of BLAST program, query definition, database, and program parameters; (*ii*) descriptive summary – list of matching hits, subject definition, bit score, and E-value; (*iii*) sequence alignment – sequence length, bit score, E-value, identities, positives, and gaps; and (*iv*) footer section – number of sequences searched, length of database, and algorithm scores. The programming language specific BLAST outputs and accessing methods are given in the Table I bellow.

Table I. Purpose of various BLAST outputs.

BLAST Output	***File Usage Description***	
	Accessing Method	***Available Data***
Text	Regular Expresion	Full report, Descriptive summary, Formatted sequence alignment, and Status tracking
CSV	Comma and New line Seperators	Descriptive summary
JSON	JavaScript Object Notation	Full report, and Descriptive summary
XML	Human or Machine Readable Formatted Tags	Full report, and Descriptive summary
ASN.1	Abstract Syntax Notation	Descriptive summary

Only for data extraction through programming language

There are other output file formats such as SAM, ASN.1 (Binary), and HTML which cannot be parsed using

programming languages, instead those files can be read using a suitable viewer. HTML file formatted output is a standard type to view through a web browser. It is similar to the Text file format except the clickable hyperlinks. Text file formatted output can be simply viewed through any ASCII/Unicode code supported text editor.

III. RESULTS AND DISCUSSION

XML (eXtensible Markup Language) is a software or hardware independent and customizable (except HTML tags) markup language designed for storing and retrieving data. In XML, tags were arranged in hierarchical order similar to HTML, where XML tag names act as *variables* and content between the tags are *values* [7]. The XML file formatted output of NCBI BLAST sequence alignment result consists of a major section known as `<Iteration>` which contains a brief description of the matching sequences, HSP score parameters, and the sequence alignment of each hit [8] (Figure 1).

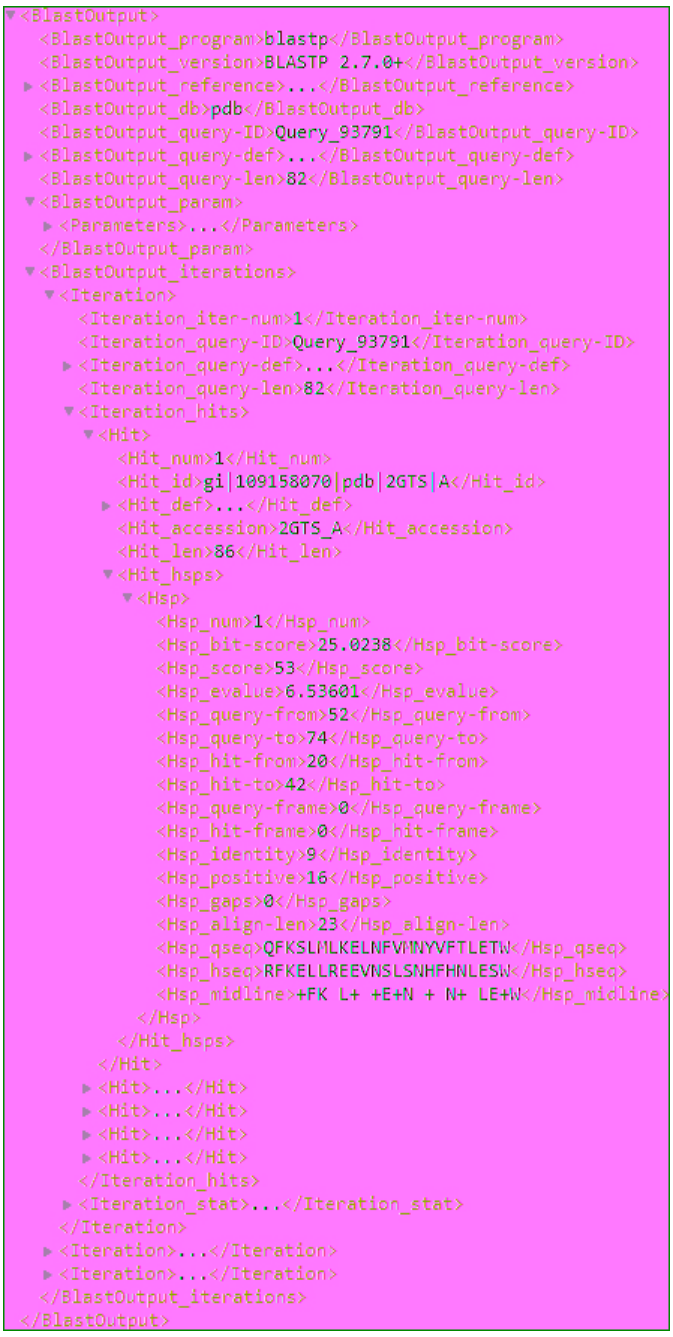

Figure 1. Minimized XML file formatted NCBI BLAST output.

The `<Hit_id>` and `<Hit_def>` tags were used to annotate database accession number with a hyperlink to the original source through sequence identifiers (gi, gb, pdb, etc.) using regular expression. Similarly, the tags `<Hsp>` and `<Hit_def>` tags were used to generate the descriptive summary of sequence alignments and the hits ids were annotated with anchor hyperlinks to the corresponding sequence alignment section using regular expression (Figure 2).

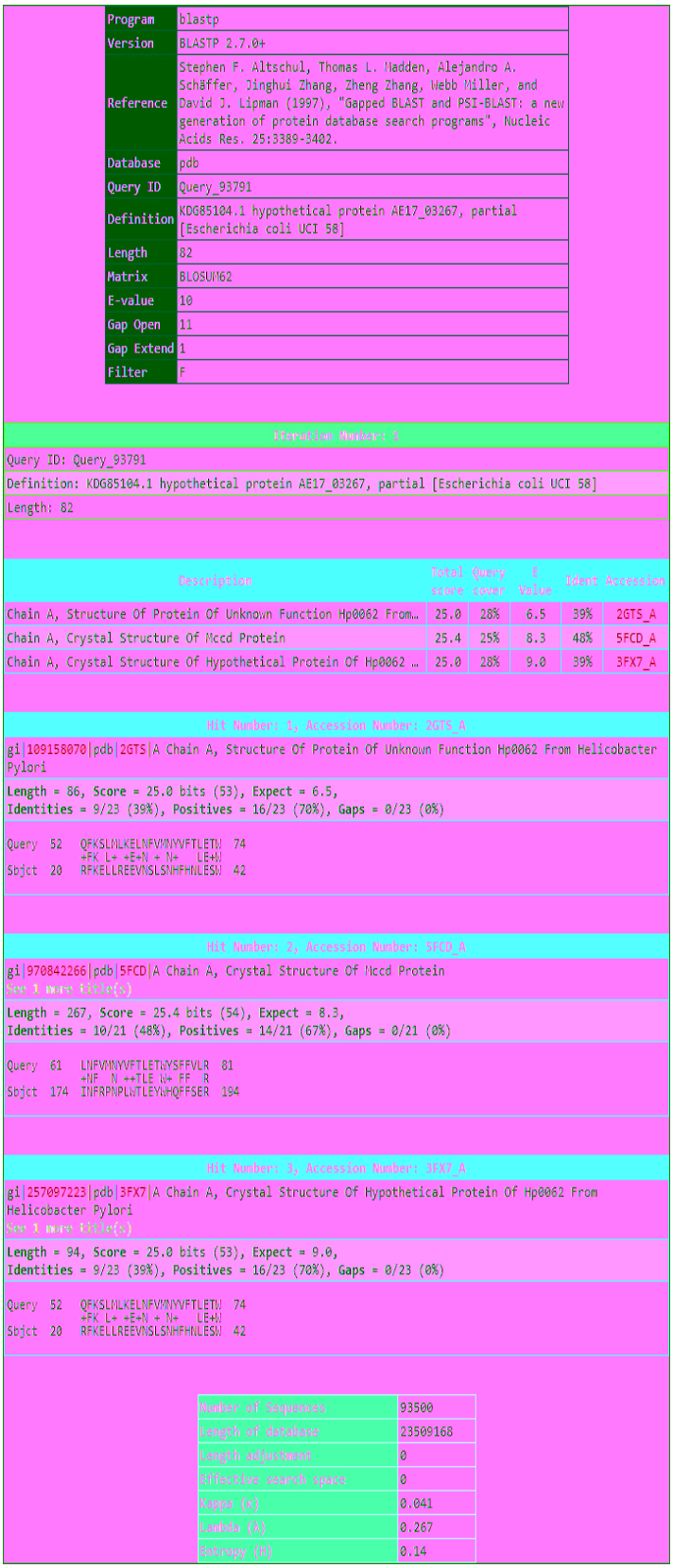

Figure 2. Model output of xmlBLASTparser program.

A. Utility

A standardized XML file formatted output obtained from the sequence alignment result of NCBI BLAST is used as the input for parsing through xmlBLASTparser. The different types of methods adopted to retrieve the XML file are given bellow:

- *Online tool* – It is a simple method to download an XML file from the online NCBI BLAST tool after performing the sequence alignment. Alternatively, we can also obtain the XML file through command line execution of Remote BLAST or Local BLAST using the standalone NCBI BLAST+ tool. The PHP script to read the XML file is

```
$xml = simplexml_load_file("output.xml");
```

- *RESTful service* – It is a widely used method by software developers to obtain an XML file from online at back end through any server-side programming languages. There are many O|B|F bioprogramming modules such as BioPerl [6], BioRuby [7], BioJava [8], BioPython [9], and BioConductor [10] were used for similar functionality. BLASTphp [11] is a simple PHP script which allows remote execution of either Online or Cloud BLAST and stream the sequence alignment result. Through combining BLASTphp with xmlBLASTparser, we can easily perform sequence alignment in addition to XML file parsing. The PHP script to extract the XML file using RID of NCBI BLAST is

```
$out = file_get_contents("https://blast.ncbi.nlm.
nih.gov/blast/Blast.cgi?CMD=Get&FORMAT_TYPE=XML&
FORMAT_OBJECT=Alignment&RID=$rid");
$xml = new SimpleXMLElement($out);
```

- *Command line* – It is an excellent choice for database developers to perform a homologous sequence comparison between the sequences in the database. Moreover, the command line standalone NCBI BLAST+ with BLASTphp and xmlBLASTparser is a good alternative for WWW BLAST tool. An XML file can be obtained by executing NCBI BLAST+ using `exec()`, `passthrough()`, `shell_exec()`, or `system()` functions in the PHP. The PHP script for command line execution of standalone NCBI BLAST+ and XML file extraction is

```
exec('blastp.exe ‐db pdb ‐query seq.fa ‐remote ‐o
utfmt 5 ‐out output.xml');
$xml = simplexml_load_file("output.xml");
```

IV. CONCLUSION

xmlBLASTparser is a simple PHP script which consumes very less bandwidth and resource on the web server. It can be easily integrated with any NCBI BLAST applications and sequence alignment information can be parsed from the XML file formatted output. The current version of xmlBLASTparser generates tabular formatted rich web content with annotations. Through combining BLASTphp and xmlBLASTparser library into a PHP web form can able to build a sequence alignment tool analogue to Web NCBI BLAST. The sequence alignment generated by xmlBLASTparser is well formatted and identical to the NCBI BLAST sequence alignment result. The current version xmlBLASTparser v1.1 provides a brief summary of matching hits with detailed alignment scores. xmlBLASTparser is still under development, as we are currently focused on generating CDS region prediction, and graphical descriptive summary of sequence alignment from XML output using jQuery and CSS in addition to the xmlBLASTparser PHP library.

V. REFERENCES

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” J. Mol. Biol., vol. 215, no. 3, pp. 403–410, Oct. 1990, doi:10.1016/S0022-2836(05)80360-2.

[2] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” Nucleic Acids Res., vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[3] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, “BLAST+: architecture and applications,” BMC Bioinformatics, vol. 10, no. 1, p. 421, 2009, doi:10.1186/1471-2105-10-421.

[4] NCBI BLAST - https://blast.ncbi.nlm.nih.gov/Blast.cgi

[5] T. Madden, “The BLAST Sequence Analysis Tool,” 15 Mar. 2013, In: The NCBI Handbook, 2nd ed., Bethesda (MD): National Center for Biotechnology Information (US), 2013. Available from:

https://www.ncbi.nlm.nih.gov/books/NBK153387/

[6] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney, “The Bioperl toolkit: Perl modules for the life sciences,” Genome Res., vol. 12, no. 10, pp. 1611–1618, Oct. 2002, doi: 10.1101/gr.361602

[7] N. Goto, P. Prins, M. Nakao, R. Bonnal, J. Aerts, and T. Katayama, “BioRuby: bioinformatics software for the Ruby programming language,” Bioinformatics, vol. 26, no. 20, pp. 2617–2619, Oct. 2010. doi: 10.1093/bioinformatics/btq475

[8] R. C. G. Holland, T. A. Down, M. Pocock, A. Prlić, D. Huen, K. James, S. Foisy, A. Dräger, A. Yates, M. Heuer, and M. J. Schreiber, “BioJava: an open-source framework for bioinformatics,” Bioinformatics, vol. 24, no. 18, pp. 2096–2097, Sep. 2008. doi: 10.1093/bioinformatics/btn397

[9] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” Bioinformatics, vol. 25, no. 11, pp. 1422–1423, Jun. 2009. doi: 10.1093/bioinformatics/btp163

[10] R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, Eds., “Bioinformatics and Computational Biology Solutions Using R and Bioconductor,” Springer, New York, 2005, doi: 10.1007/0-387-29362-0

[11] T. Ashok Kumar and B. Rajagopal, “BLASTphp: a PHP wrapper for NCBI BLAST API,” Int. J. Comp. Bio., vol. 6, no. 1, pp. 31-33, Jul. 2017.

[12] Extensible Markup Language (XML), In: W3C, Retrieved from https://www.w3.org/XML/, Accessed 25 Aug. 2017.

[13] NCBI BLAST Output DTD, In: NCBI, Retrieved from https://www.ncbi.nlm.nih.gov/dtd/NCBI_BlastOutput.mod.dtd, Accessed 25 Aug. 2017.