

BRM-Parser: A tool for comprehensive analysis of BLAST and RepeatMasker results

Anjali Bajpai, Settu Sridhar, Hemakumar M. Reddy and Rachel A. Jesudasan*

Center for Cellular and Molecular Biology, Hyderabad - 500007, India

* Corresponding author

Email: rachel@ccmb.res.in

Edited by E. Wingender; received October 05, 2006; revised February 28, 2007; accepted March 31, 2007; published April 22, 2007

Abstract

BLAST and RepeatMasker Parser (BRM-Parser) is a service that provides users a unified platform for easy analysis of relatively large outputs of BLAST (Basic Local Alignment Search Tool) and RepeatMasker programs. BLAST Summary feature of BRM-Parser summarizes BLAST outputs, which can be filtered using user defined thresholds for hit length, percentage identity and *E*-value and can be sorted by query or subject coordinates and length of the hit. It also provides a tool that merges BLAST hits which satisfy user-defined criteria for hit length and gap between hits. The RepeatMasker Summary feature uses the RepeatMasker alignment as an input file and calculates the frequency and proportion of mutations in copies of repeat elements, as identified by the RepeatMasker. Both features can be run through a GUI or can be executed via command line using the standalone version.

Keywords: BLAST, RepeatMasker, repeats, parse, filter, merge, mutation

Introduction

The exponential increase in sequence data available in databases necessitates the development of adequate tools that allow easy handling of such large data to draw pertinent interpretations. BLAST [Altschul *et al.*, 1997] is one of the most popularly used alignment search tools, which generates outputs as alignments of matched sequences. Various BLAST parsers are available at BioPerl (<http://www.bioperl.org>), TIGR (<http://www.tigr.org>), SEALS (<http://www.ncbi.nlm.nih.gov/CBBresearch/Walker/SEALS/index.html>) etc. With the completion in sequencing of various genomes, analysis of repeats has become possible.

RepeatMasker (www.RepeatMasker.org) is among the most widely used software to identify and mask repeats providing a detailed output of the identified repeats and their alignments. With the currently available programs, BLAST parsing and RepeatMasker output analysis cannot be performed from a common platform. Although many of the features provided by BRM-Parser developed by us are available in most of the existing parsers, our program provides a combination of features, not available from any single platform, via a GUI. The BRM-Parser can parse BLAST outputs and RepeatMasker alignment files using the same platform. This provides the user with a powerful handle to study repeats with respect to their occurrence, periodicity and mutation frequency in the genome. The BRM-Parser is available as a window based GUI at <http://203.200.217.178/brm-parser/main.html>.

Program features

BRM-Parser can filter, sort and merge hits from a single or multiple BLAST output. It first summarizes the BLAST alignment output into a tabulated form, if it is not provided by the user. The summary generated contains the sequence identifier, coordinates for query and subject sequences, BLAST score, the length and orientation of the hits, the percentage identity and *E*-values, each of which is tab delimited (Fig. 1). The Filter feature allows the user to retrieve hits that satisfy user-defined criteria for length of hit, percentage identity and *E*-value either singly or in combination. Since the hits in the BLAST output file are in the order of the score, the Sort feature of BRM-Parser sorts the hits by the query length, or by subject or query coordinates. Long stretches of homology, interspersed by short regions of mismatch or low complexity regions are depicted as separate hits in the BLAST output. Merge option of the BRM-Parser merges BLAST hits present in same orientation, which satisfy a cutoff for hit length and gap between hits, both defined by the user. If the user does not define the parameters, the program sets a default length of 100 base pairs and a gap of 1/10th the length. For gaps greater than 100 bp, a 20% difference in the gaps between hits of the query and the subject is allowed, to account for variations in length of repeats. Regions of merged and unmerged matches that are repeated within a contig are documented in the BRM output as a supplementary file. Thus, the Merge feature provides a comprehensive overview of larger hits. Merge and Sort can be used in combination with the filter option to refine the analysis. BRM-Parser generates four output files, containing the summary, filtered hits, sorted hits, and merged hits, depending on the options used. Each file bears the input file name suffixed by the option used and can be downloaded. These features are exploited best while parsing large BLAST outputs such as whole chromosome(s) against a database or itself (Supplementary Tab. 1). The RepeatMasker Summary feature takes the RepeatMasker alignment file in the orientation of the repeat as an input file and sums up the mutations, at each nucleotide position for all copies, scored as transitions, transversions and deletions by RepeatMasker. The Summary file also provides the number of repeats of a subfamily, found in the direct and inverse orientation, besides the frequency and the proportion of mutations in repeats of the same family. The output files generated are tab-delimited and hence can be transferred to Excel sheets for easier handling and further analysis.

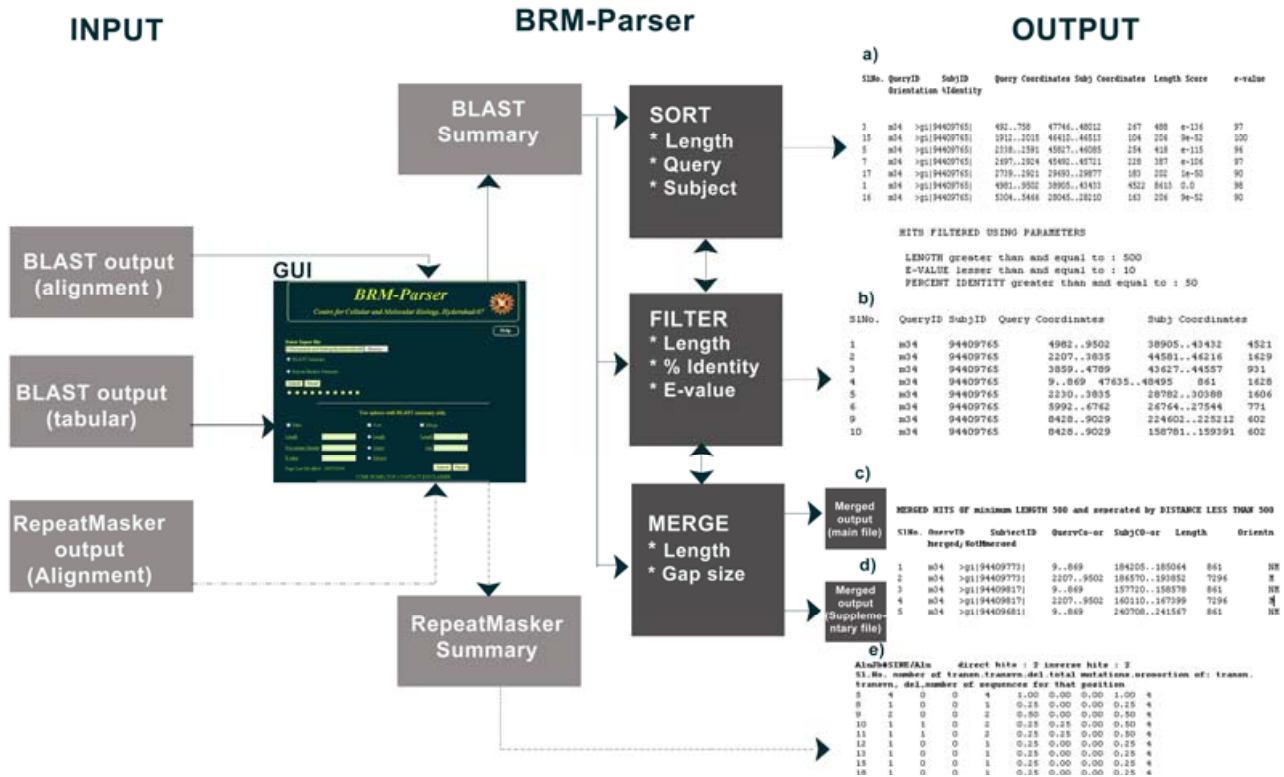


Figure 1: The workflow for BRM-Parser. The figure shows the two features provided in BRM-Parser, BLAST Summary and RepeatMasker Summary. The input files, i. e., [BLAST output (Alignment or Tabular format) and RepeatMasker output (Alignment file)] are given on the left-hand side of the GUI, given in the middle of the figure. The three options of BRM-Parser, Filter (by length, percent identity and *E*-value), Sort (by length, coordinates of Query and Subject), and Merge (by length and Gap size) are depicted on the right-hand side of GUI. The double-headed arrows indicate that these options can be used in combination. These options generate three output files, viz., filtered hits, sorted hits, and merged hits, depending on the options used (a, b, c). Regions of merged and unmerged matches that are repeated within a contig are documented in the BRM output as Merge supplementary file (d). The RepeatMasker Summary feature generates the RepeatMasker summary file (e).

For a high resolution picture click here: [Figure 1 \(2562 KB\)](#)

Testing BRM-Parser

To test the utility of the BRM-Parser we parsed BLAST outputs of large repeating segments of 9.5 kb and 300 kb on mouse Y chromosome. BLAST of a 9.5 kb repeat, M34 (GenBank accession, DQ907163) from the long arm of mouse Y chromosome [Singh *et al.*, 1994], against the whole mouse Y chromosome (NCBI build 36.1) yielded 1122 hits across 29 contigs. To restrict our analysis to hits of significance, we filtered the BLAST output, defining length of 100 bp and percentage identity of 90%, which yielded 218 hits across 24 contigs. Sorting the filtered hits by the query coordinate provided an insight into the regions of the query, repeated on the Y chromosome, which ranged from 100 bp to 3 kb. A 3 kb region of the query was present on 3 contigs. Two segments, 1 kb and 2 kb, from the 3 kb region were repeated 14 times and 17 times, respectively, on the Y chromosome. Smaller stretches of 100 to 200 bp were repeated 20 to 50 times (Supplementary Tab. 2). Sorting the filtered hits by the subject co-ordinates helped us understand the arrangement and the orientation of the repeating segments of the query on the Y chromosome. We observed that ~9 kb of M34 (492-9502) followed by two short stretches of 183 and 163 bp from within M34, were present 17 times, in both direct and inverse orientations on the Y chromosome (Supplementary Tab. 3).

Further, BLAST of the 300 kb stretch from the Y contig, gi: 38087949, (NCBI version 32.1) containing M34 against the mouse Y chromosome resulted in 71,281 hits. To assess the utility of the Merge option of BRM-Parser, small discontinuous hits obtained from the above output were merged, to obtain larger stretches of continuous matches by using different gap sizes. The Merge option was applied to 2615 hits that satisfied length of 1000 bp and percentage identity of 90%, defined using the Filter option. The longest hit obtained was 9.2 kb. LINEs and Retroviral sequences [Smit, 1996] masked in the query sequence, could give rise to larger gaps; therefore the discontinuous hits were merged using gaps of 100 to 1000 bp. Using the default gap of 100 bp, 372 hits could be merged to obtain 177 hits, while 2243 hits remained unmerged (Tab. 1). Of the 177 merged hits 27 hits were repeated more than once in the same contig and were documented in the supplementary file (Fig. 1). The longest hit obtained by merging 3 hits of 9.26, 1.25 and 2.42 kb was 12.99 kb. The 12.99 kb region was repeated in forward orientation in 3 contigs, and inverse orientation in 1 contig. On increasing the threshold of hit length to 2000 bp the number of merged hits decreased to 37, on further increasing the length to 5000 bp no merged hits were obtained, therefore we limited the minimum length of hits to be merged to 1000 bp, and varied the gap size. On increasing the gap length to 200 bp, the number of merged hits increased to 254. Further increase of gap size to 500 bp resulted in 348 merged hits. Allowing for a larger gap did not further increase the number of merged hits, although, the number of merged hits greater than 10 kb increased from 12 to 34. To accommodate large gaps due to the presence of LTRs/Retroviral elements, we increased the gap size to 1000 bp, which resulted in the longest merged hit of 23.68 kb, by merging 7 hits. To assess the reliability of the Merge option to identify large discontinuous matches we allowed a gap length of 5000 bp, which resulted in 50 hits larger than 25 kb, the longest of which were 60.83, 53.5 and 45 kb. The corresponding sequence from the query and subject were retrieved and the sequence similarity was confirmed using blast2 alignment. Query and subject sequences for the 60.83 kb match were subjected to RepeatMasker, which

confirmed the presence of the same repeat elements in both, except for varying lengths of low complexity repeats. Overall there were 74 repeat elements in this region occupying 34.59 kb in the query and 34.54 kb in the subject sequence.

Table 1: Number of merged and unmerged hits obtained for different parameters of hit length and gap size provided to Merge option. Parsing a BLAST output of a 300kb region of mouse-Y against the mouse genome

Parameters							Results		
Filter			Merge		Number of hits		Size of the largest merged hit (kb)	Number of merged hits	
Length-(bp)	% Identity	Length-(bp)	Gap size(bp)	Non-merged	Merged	>10 kb		>25 kb	
1	1000	90	1000	100	2243	177	12.98	11	Nil
2	1000	90	1000	200	2064	254	12.98	12	Nil
3	1000	90	1000	500	1813	348	14.83	12	Nil
4	1000	90	1000	1000	1689	337	23.68	34	Nil
5	1000	90	1000	2000	1366	404	27.05	67	5
6	1000	90	1000	5000	1094	330	60.83	134	50
7	1000	90	2000	200	487	37	9.26	Nil	Nil
8	1000	90	5000	500	56	Nil	Nil	Nil	Nil

To verify the utility of RepeatMasker Summary of the BRM-parser, we analyzed 165 L1 elements from mouse Y chromosome, identified as L1-mur1-orf2 by the RepeatMasker. The alignment file was parsed through the BRM-parser using the RepeatMasker Summary option. Most of the elements were partial, with 120 elements retaining sequences towards the 3 prime end of the orf2, whereas 40 sequences contained regions from 5' end. We observed certain regions of mutation hotspots. In a 1.12 kb region of the orf 2, 186 bases showed proportion of mutation greater than 0.5. It was seen that the number of transitions and transversions were comparable at 21 positions only, while mutations were either transition or transversion at 132 positions. Most of the deletions also occurred in pockets, such as a 18 bp region from positions 3818 to 3835, 15 bp stretch from position 4058 to 4072 and 6 bp from 4264 to 4269.

Discussion

A number of BLAST parsers are available each designed to suit particular needs, such as speed, scalable features, interactive GUIs, and common platform to summarize and filter BLAST results. The NCBI (<http://www.ncbi.nlm.nih.gov/BLAST/>) itself allows users to retrieve BLAST output in tabulated format, although BLAST services such as the bl2seq and IgBlast and BLAST used by other sites such as the UTRsource (<http://www.ba.itb.cnr.it/UTR/>) and the Eukaryotic Promoter Database available at

<http://www.epd.isb-sib.ch/> do not give tabulated outputs. More recently, the NCBI introduced a new feature to sort hits by score, percent identity and coordinates of query and subject. However, these are currently not available to all BLAST options and are restricted to few of the output formats only. Other parsers such as MSPcrunch (<http://bioweb.pasteur.fr/seqanal/interfaces/mspcrunch.html>) [Sonnhammer, 1994], BLASTaid (<http://search.cpan.org/~twylie/BLASTaid-v0.0.3/lib/BLASTaid.pm>), Boulder::blast (<http://stein.cshl.org/software/boulder/docs/Boulder/Blast.html>) provide features to parse, summarize and filter hits, without a direct option to sort and merge hits. BRM-Parser on the other hand provides a filter by additional criteria such as length and percent identity besides *E*-value and sorts by query length besides coordinates of query and subject. The Merge feature of the BRM-Parser is not commonly seen in other parsers. It gives flexibility to the user to set limits according to ones discretion in order to look for longer matches. On increasing gap size to 2000 bp we retrieved 5 hits of 25 kb and more, and on further increasing the gap size to 5000 bp we obtained 50 hits of 25 kb and more. SEALS (<http://www.ncbi.nlm.nih.gov/CBBresearch/Walker/SEALS/index.html>) package provides blast2blast, which enables filtering of BLAST outputs by score, *E*-value etc., and blast2bounded sorts hits by co-ordinates of query and subject; SEALS provides merged option as well besides others however, it does not provide a GUI. Programs such as Reputer (<http://bibiserv.techfak.uni-bielefeld.de/reputer/>) [Kurtz, 2001] compute and visualize repeated regions in whole genomes/chromosomes and PILER [Edgar, 2005] identifies and classifies novel genomic repeats, however these do not provide a direct option to score the repetitiveness of a segment of interest in a larger sequence, such as the M34 on the mouse Y chromosome. BRM-parser has an advantage of generating information of periodicity of repeating segments from a BLAST output itself. Therefore it allows the user to elicit maximum inferences from a single output. The RepeatMasker Summary is an efficient tool to calculate the mutation frequency and the extent of variations within repeat elements. This is especially useful when analyzing large number of repeats of any kind for their mutations in functionally or structurally distinct landscapes of genomes.

Acknowledgements

Fellowship to AB and HMR from CSIR is duly acknowledged. We thank Dr. Shrish Tiwari for the discussions.

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
- Edgar, R. C. and Myers, E. W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, Suppl 1:i152-i158.
- Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29**, 4633-4642.
- Singh, L., Panicker, S. G., Nagaraj, R. and Majumdar, K. C. (1994). Banded krait minor-satellite (Bkm)-associated Y chromosome-specific repetitive DNA in mouse. *Nucleic Acids Res.* **22**, 2289-2295.
- Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**, 743-748.
- Sonnhammer, E. L. L. and Durbin, R. (1994). A workbench for large scale sequence homology analysis. *Comput. Applic. Biosci.* **10**, 301-307.