Blast Parser - The Excel style viewer

Publication

Carreras Marco(1), Bosotti Roberta(2)*.
(1) Bioinformatic Software Developer, Pavia-Italy. (2) Genomics Unit, Department of Biotechnology, Nerviano Medical Sciences, Nerviano (MI)-Italy.
* Corresponding author (e-mail): "roberta.bosotti"@nervianoms.com (without quote)

- BITS Conference 2006 (Società di Bioinformatica Italiana). April 28-29, 2006 (Bologna, Italy).

Free Poster abstract (n.115) on BITS 2006.

NOTE. The Software was accomplished in collaboration with an institution external to GeneProject and is regulated by the Freeware License Agreement.

Introduction

Parsing a Blast output consists in extracting all the main information contained in the file and represent them in a different way, an Excel style way. Each row of the Excel representation contains the whole alignment including the statistical information, such: E-Value, Score, % of Identity, etc. plus several new added facilities obtained by a preliminary automated analysis.
This compact structure permits to see better and faster differences among alignments with the opportunity to filter, sort and/or group alignments by statistical values.

Features

The Software works under the Microsoft Windows environment. The interface helps in producing parsings by filtering the result, applying some preconditions **(e.g. level of % of Identity or Positivity/Similarity)** and choosing some options to facilitate further analyses, such Warnings or NCBI Accession Numbers extraction **(e.g. gi|21707502|gb|AAH33867.1|)**.

1) The extracted information

- **Query:** headers of sequences to analyze
- **Subject:** headers of sequences found in the database
- **Score:** a number representation **(e.g. 650)**
- **Score Text:** full text representation plus BITS **(e.g. 254 bits (650))**
- **Expect:** the E-Value as number **(e.g. 1e-66)**
- **Identities %:** a number representation **(e.g. 95)**
- **Identities Text:** full text representation plus characters matching **(e.g. 115/120 (95%))**
- **Positives %:** a number representation **(e.g. 97)**
- **Positives Text:** full text representation plus characters matching by similarity **(e.g. 117/120 (97%))**
- **Gaps %:** a number representation **(e.g. 9)**
- **Gaps Text:** full text representation plus voids **(e.g. 11/118 (9%))**
- **Strand:** the Nucleotide orientation **(e.g. "Plus / Minus")** valid only for the BLASTN program
- **Frame:** orientation of the translation ORF **(e.g. +2)**
- **Length Query:** the number **(e.g. 700)**
- **Length Subject:** the number **(e.g. 509)**
- **Position Query:** as text representation plus the length of the frame **(e.g. 341-700 (360))**
- **Position Subject:** as text representation plus the length of the frame **(e.g. 1-120 (120))**

- **Warnings:** represented by four independent letters **(i.e. LEIF)**
  - **L** = Warning on lengths mismatch
  - **E** = Warning on non 0 E-Value
  - **I** = Warning on non 100% Identity
  - **F** = Warning on fragmented alignment

- **Blast_ID:** a number (the original Blast order of the alignments)
- **Program:** **(e.g. "BLASTX 2.2.10 [Oct-19-2004]")**
- **Database:** **(e.g. "All non-redundant GenBank CDStranslations+PDB+SwissProt+PIR+PRF excluding environmental samples")**
- **Alignment:** (the whole contained in a cell)

```
Query: 23 LTESGPAVIKPGESHKLSCKASGFTFSSAY-MSWVRQAPGKGLEWVAYIYSGGSSTYYAQ 81
          LTESGP V KPGESHKL+C ASGFTFSS Y MSW+RQAPGKGLEW+AY YS  ++TYY+Q
Sbjct: 22 LTESGPVVKKPGESHKLTCTASGFTFSS-YEMSWIRQAPGKGLEWIAYSYS--TNTYYSQ 78

Query: 82 SVQGRFAISRDDSNSMLYLQMNSLKTEDTAVYYCARGGLGW----SLDYWGKGTMITV 135
          SVQGRF ISRDDS+S LYLQMNSLK+EDTAVYYCAR    W    + DYWG+GT++TV
Sbjct: 79 SVQGRFTISRDDSSSKLYLQMNSLKSEDTAVYYCAR---EWGAAAAFDYWGQGTIVTV 133
```

2) Recognized Programs (query vs. database)

- **BLASTN:** Nucleotide vs. Nucleotide
- **BLASTP:** Protein vs. Protein
- **BLASTX:** Translated vs. Protein
- **TBLASTN:** Protein vs. Translated
- **TBLASTX:** Translated vs. Translated

Also supported: PHI-BLAST, PSI-BLAST, Megablast, GEO and SNP BLAST.

The file to analyze (either be in DOS or in UNIX) must be in **text format** (better if in Plain text) and in a **Pairwise view**, so a web Blast output should be saved as a Text file, or copied as text and saved to Notepad, in order to avoid the HTML language format.

3) The Grid window

Once the file is given to the interface and options are set, the parsing can be started. A progressive bar indicates the progressive state of the work and if no error messages appear the result is shown in a Grid window.
The Grid window is a powerful Excel style viewer with which it is possible to start the analysis.

Some features of the Grid window:

- Set the individual row's checkbox
- Filter rows by proper checkbox or applying conditions to a single column using Boolean operators
- Sort columns in ascending or descending way. It differs among number, text and date. The multiple column sorting is available by holding the SHIFT key
- Columns are moveable and sizeable
- Selected cells can be copied and pasted directly to Excel and to any text editor
- An intelligent float number conversion is applied in accord to the Local Regional Settings **(e.g. 1.97 -> 1,97 or vice versa)**
- Perform a text search by opening the Find dialog **(i.e. pressing CTRL+F)**
- Show rows in white or colored by groups. The rows grouping is based on the same text found in a chosen column. Grouped rows are colored in the same way and groups are counted
- Generate the Report for each row = alignment in a new window by double-clicking on a row

The Grid window can be previewed, customized and printed.
Can also be exported to several formats: ASCII format (*.txt), CSV format comma (*.csv), CSV format semi-comma (*.csv), Excel format (*.xls), Excel Enhanced format (*.xls), HTML format (*.htm), Word format (*.doc).
Also Reports of the Grid window can be previewed, customized and printed and can be exported to these formats: Rich Text Format (*.rtf), ASCII format (*.txt).

When exporting to **Excel** remember that it has intrinsic limits:

- Worksheet size: [65536 rows x 256 columns]
- Characters length of a single cell content: [Declared=32767, Safe=31500]

So, here are some Tips:

Tip1: As the whole alignment is stored in a cell, when selected its content is shown in the formula bar at the top. To see it better change the formula bar font following: "Tools">"Options">"General">"Standard Character" and then choose "Courier New".

Tip2: Before starting the parsing enable the checkbox "Save output in multi CSV files". This will generate one or more CSV files that are limited to 65000 rows and/or 15 MB.

4) Last considerations

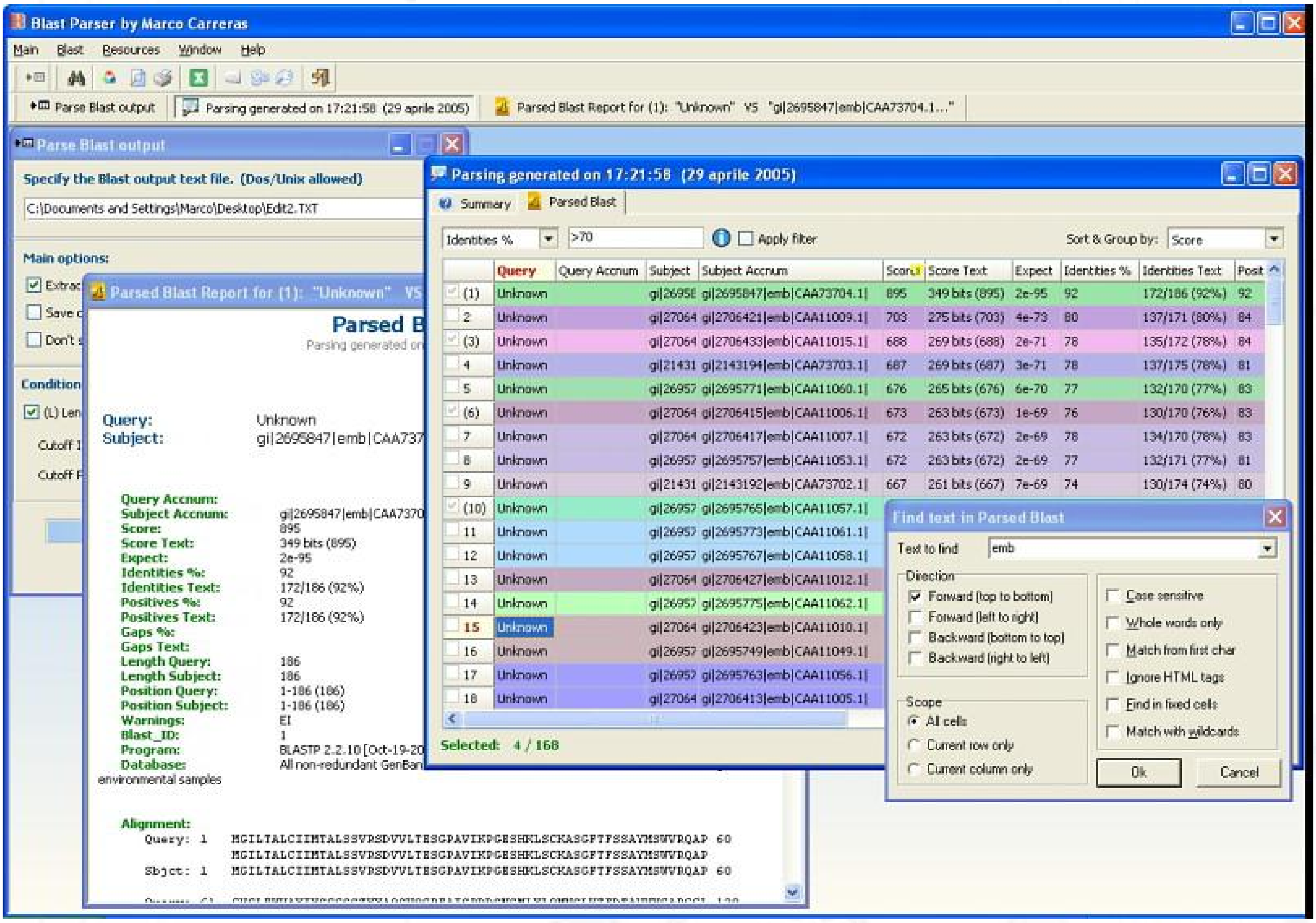
A set of resources is provided as web links **(e.g. NCBI home page)** and external applications links **(e.g. SpotFire, if locally installed)**.

The Software comes with a built-in Memory Optimizer that starts automatically when the memory limit goes below 20%. Swapping unnecessary blocks of memory leaves the CPU to focus only on running processes. On occurrence it can be also launched manually.

In event of Critical Messages **(i.e. when a procedure fails)** the Software produces a detailed report of the failure making also a screenshot. All files are stored on the local directory. A message informs about what happened giving the choice to freely send me by e-mail these Bug report files.

You are welcome to send me suggestions, feedbacks and... Bug reports.

The Interface


Blast Parser by Marco Carreras
Main
Blast
Resources
Window
Help
Parse Blast output
Parsing generated on 17:21:58 (29 aprile 2005)
Parsed Blast Report for (1): "Unknown" VS "gi|2695847|emb|CAA73704.1..."
Parse Blast output
Main options:
Extrac
Save c
Don't s
Condition
(L) Len
Cutoff 1
Cutoff P
Parsing generated on 17:21:58 (29 aprile 2005)
Summary
Parsed Blast
Apply filter
Sort & Group by:
Parsed Blast Report for (1): "Unknown" VS
Parsed B
Parsing generated or
Find text in Parsed Blast
Ok
Cancel

Specify the Blast output text file. (Dos/Unix allowed)

C:\Documents and Settings\Marco\Desktop\Edit2.TXT

☑ Extrac

☐ Save c

☐ Don't s

☑ (L) Len

Identities % >70 ☐ Apply filter Sort & Group by: Score

	Query	Query Accnum	Subject	Subject Accnum	Score	Score Text	Expect	Identities %	Identities Text	Posit
(1)	Unknown		gi|26958	gi|2695847|emb|CAA73704.1|	895	349 bits (895)	2e-95	92	172/186 (92%)	92
2	Unknown		gi|27064	gi|2706421|emb|CAA11009.1|	703	275 bits (703)	4e-73	80	137/171 (80%)	84
(3)	Unknown		gi|27064	gi|2706433|emb|CAA11015.1|	688	269 bits (688)	2e-71	78	135/172 (78%)	84
4	Unknown		gi|21431	gi|2143194|emb|CAA73703.1|	687	269 bits (687)	3e-71	78	137/175 (78%)	81
5	Unknown		gi|26957	gi|2695771|emb|CAA11060.1|	676	265 bits (676)	6e-70	77	132/170 (77%)	83
(6)	Unknown		gi|27064	gi|2706415|emb|CAA11006.1|	673	263 bits (673)	1e-69	76	130/170 (76%)	83
7	Unknown		gi|27064	gi|2706417|emb|CAA11007.1|	672	263 bits (672)	2e-69	78	134/170 (78%)	83
8	Unknown		gi|26957	gi|2695757|emb|CAA11053.1|	672	263 bits (672)	2e-69	77	132/171 (77%)	81
9	Unknown		gi|21431	gi|2143192|emb|CAA73702.1|	667	261 bits (667)	7e-69	74	130/174 (74%)	80
(10)	Unknown		gi|26957	gi|2695765|emb|CAA11057.1|						
11	Unknown		gi|26957	gi|2695773|emb|CAA11061.1|						
12	Unknown		gi|26957	gi|2695767|emb|CAA11058.1|						
13	Unknown		gi|27064	gi|2706427|emb|CAA11012.1|						
14	Unknown		gi|26957	gi|2695775|emb|CAA11062.1|						
15	Unknown		gi|27064	gi|2706423|emb|CAA11010.1|						
16	Unknown		gi|26957	gi|2695749|emb|CAA11049.1|						
17	Unknown		gi|26957	gi|2695763|emb|CAA11056.1|						
18	Unknown		gi|27064	gi|2706413|emb|CAA11005.1|						

Selected: 4 / 168

Text to find: emb

Direction
☑ Forward (top to bottom)
☐ Forward (left to right)
☐ Backward (bottom to top)
☐ Backward (right to left)

Scope
● All cells
○ Current row only
○ Current column only

☐ Case sensitive
☐ Whole words only
☐ Match from first char
☐ Ignore HTML tags
☐ Find in fixed cells
☐ Match with wildcards

Query: Unknown
Subject: gi|2695847|emb|CAA737

Query Accnum:
Subject Accnum: gi|2695847|emb|CAA7370
Score: 895
Score Text: 349 bits (895)
Expect: 2e-95
Identities %: 92
Identities Text: 172/186 (92%)
Positives %: 92
Positives Text: 172/186 (92%)
Gaps %:
Gaps Text:
Length Query: 186
Length Subject: 186
Position Query: 1-186 (186)
Position Subject: 1-186 (186)
Warnings: EI
Blast_ID: 1
Program: BLASTP 2.2.10 [Oct-19-20
Database: All non-redundant GenBan
environmental samples

Alignment:

```
Query: 1   MGILTALCIIMTALSSVRSDVVLTESGPAVIKPGESHKLSCKASGFTFSSAYMSWVRQAP  60
           MGILTALCIIMTALSSVRSDVVLTESGPAVIKPGESHKLSCKASGFTFSSAYMSWVRQAP
Sbjct: 1   MGILTALCIIMTALSSVRSDVVLTESGPAVIKPGESHKLSCKASGFTFSSAYMSWVRQAP  60
```

Click here to see the Image Gallery

Download

System Requirements:

- Microsoft Windows XP, 2000, NT (not tested on 9x family)
- 10 MB of disk space. Memory RAM > 128 MB and CPU processor > 133 MHz
- (With the minimum configuration, performance and functionality may be less than expected)

Blast Parser v1.2.6.14	Executables and doc files (.zip)	2.7 MB
Freeware License Agreement	PDF document	101 KB
Help Manual (this web page)	PDF document	327 KB
Support & Donation	PDF document	126 KB

By downloading the Software you agree ALL the terms of conditions defined in the ’Freeware License Agreement’.

Frequently Asked Questions (FAQ)

Q. What are the most common error messages during the "BLAST procedure"?

A.

- *Parsing Database*: the "Database" name has not been detected inside the first block of 20000 characters of the file.
- *Program unknown*: no BLAST Program has been detected among those recognized inside the first block of 20000 characters of the file.
- *Parsing Source*: the header of the "Query" relative to the given sequence is missing or exceeding the 20000 characters.
- *Parsing Found*: the header of the "Subject" relative to the found sequence is missing or exceeding the 20000 characters.
- *Parsing Score*: the term "Score" has not been detected in any alignment.
- *Parsing Query*: the term "Query=" relative to the given sequence has not been detected inside the whole file.
- *Parsing Query2*: the term "Query" relative to the alignment match has not been detected.
- *Parsing Sbjct*: the term "Sbjct" relative to the alignment match has not been detected.
- *The Output file seems to be truncated*: keywords with which BLAST outputs end (e.g. Database, Matrix, etc.) have not been detected.
- *Parsing Length1*: the number relative to the length of the given Query is non-numerical.
- *Parsing Length2*: the number relative to the length of the found Subject is non-numerical.
- *Parsing Score2*: the format of the "Score" is incorrect.
- *Parsing Expect*: the number relative to the "E-Value" is non-numerical.
- *Parsing Identities*: the format of the "Identities" is incorrect.
- *Parsing Positives*: the format of the "Positives" is incorrect.
- *Parsing Gaps*: the format of the "Gaps" is incorrect.
- *Parsing Frame*: the format of the "Frame" is incorrect.
- *Parsing Strand*: the format of the "Strand" is incorrect.
- *Parsing Pos1a, Pos1b, Pos2a, Pos2b*: the numbers relative to the position respectively of the given Query and/or the found Subject is non-numerical.
- *Parsing Alignment*: the length of the Alignment is less than 10 characters.
- *Out of Memory*: the memory can’t handle further parsing data. Please, close the not necessary applications and use the Memory Optimizer.

Q. What are the most common error messages during the "FILE procedure"?

A.

- *Error #2*: the file has not been found.
- *Error #5*: the access to the file is denied. If the file is located on a remote server try to mapping it.
  Error #32: the file is already in use by another application (Sharing violation).

History

February 15, 2008	Blast Parser v1.2.6.14 (Release 6)
(*)	Updated the NCBI Accession Numbers identification algorithm.
March 28, 2007	Blast Parser v1.2.5.13 (Release 5)
(+)	Supported versions from BLAST v2.1.3 [Apr-1-2001] to BLAST v2.2.16 [Mar-11-2007].
(+)	Ability of extraction directly from HTML files.
October 23, 2006	Blast Parser v1.1.4.12 (Release 4, beta)
(+)	Updated to the version of BLAST 2.2.14 [May-07-2006].
April 26, 2006	Blast Parser v1.1.3.10 (Release 3, beta)
(+)	Eduction of Statistics (Matrix, Lambda, ...). Added a new grid.
(+)	Eduction of Query, Consensus and Subject Sequences from the alignment.
(+)	Enhanced the PSI Blast compatibility (Method, Iteration).
(+)	Optional conversion to Local Regional Settings. Increased the parsing speed. Grid export to XML.
June 18, 2005	Blast Parser v1.0.2.6 (Release 2)
(+)	Added the compatibility with NCBI BLAST v2.2.11 (June 13, 2005). Supported the new SNP BLAST.
(+)	Added new web links: ArrayExpress, Database of Interacting Proteins, Entrez Gene, Gene Expression Omnibus
May 19, 2005	Blast Parser v1.0.1.5 (Release 1)