QBl_Ast 's URL API. User's Guide

NCBI

September 27, 2001

Contents

1 Overview

The URLAPI is a standardized application program interface (API) for accessing the NCBI QBl_Ast system. It uses direct HTTP-encoded requests to NCBI web server.

These encoded requests should be directed to the NCBI cgi-bin program: `http://www.ncbi.nlm.nih.gov/blast/Blast.cgi`

A few of the highlighted features and advantages of using HTTP requests for the service are:

- no program should be downloaded from NCBI ftp site to use NCBI BLAST service,
- easy to keep backward compatibility of API,
- no need to transfer sequences from NCBI to local machines,
- no need to hack Web NCBI BLAST programs to run scripts,

Any HTTP browser or application capable to make HTTP requests can be utilized to use URLAPI. There are many tools around which can be used to make HTTP requests, for example - PERL modules, C libraries, lynx program, etc. Almost all examples in this document use 'lynx' browser by default, but 'netscape' and IE can be used for the same purposes.

Basically, in order to issue an URLAPI command your program will need to:

1. Make a connection to port 80 to NCBI web server (in UNIX: telnet www.ncbi.nlm.nih.gov 80)
2. Send the following commands to the port:

```
POST /blast/Blast.cgi HTTP/1.0
User-Agent: Hi_there
Connection: Keep-Alive
Content-type: application/x-www-form-urlencoded
Content-Length: 200

CMD=Put&QUERY=555... etc
```

2 Quick start

Searching the NCBI QBl_Ast consists of two major steps. The first step is called "Put", and it puts the query sequence with the appropriate search parameters into the QBl_Ast system. The second step is called "Get" and it formats the results with specified format parameters.

The following is an example how can this be done by using URLAPI.

Let's say we want to perform a nucleotide search of query with gi=555 against the 'nr' database. In addition we need to enable low-complexity filtering, set the expect value = 10, request HTML format of the output, use NCBI GI numbers in the output page, and show only the first 10 hits. In the URLAPI the first step will look like:

```
$ lynx -source \
"http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?QUERY=555\
&DATABASE=nr&HITLIST_SIZE=10&FILTER=L\
&EXPECT=10&FORMAT_TYPE=HTML&PROGRAM=blastn&CLIENT=web\
&SERVICE=plain&NCBI_GI=on&PAGE=Nucleotides\
&CMD=Put"
```

Please note that we use the "lynx" program in this example, and we have also split long lines in the shell as allowed in 'sh' syntax.

In the "url-encoded" format the '?' means the start of a list of parameters, which is followed by a list of name-pairs separated with '&'. The names of parameters basically describe themselves, but are also explained in more detail in the parameter section of this document. Note, that we set "CMD=Put" which means that we want to put this new search into QBl_Ast . It is also possible to specify a query sequence in FASTA format like "QUERY=acgtacgt". For those who work with the NCBI toolkit and are familiar with the Bioseq C structure, here is the code which converts NCBI Bioseq to FASTA format:

```
bsp = readdb_get_bioseq_ex(rdfp, oid, TRUE, TRUE);
if (!BioseqToFasta (bsp, fp, !is_prot)) {
    ErrPostEx(SEV_ERROR, 0, 0, "Can't convert Bioseq to FASTA");
}
BioseqFree(bsp);
```

The output of the 'Put' command will be a valid HTML page, the contents of which may be ignored except the following important section:

```
<!--QBlastInfoBegin
    RID = 954517067-8610-1647
    RTOE = 207
QBlastInfoEnd
-->
```

This portion of the output is special and contains the Request Identifier (RID) and the estimated Request Time of Execution in seconds (RTOE) for the search. The RID is different for every search, and is a mandatory parameter for the next step which is formatting the BLAST results.

The simplest way to get results for a given RID using the default format parameters is to use the following URL:

```
$ lynx -source "http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?\
RID=954517013-7639-11119&CMD=Get"
```

If the BLAST search is not yet complete, this will produce output in the following format with Status equal to "WAITING":

```
<!--QBlastInfoBegin
    Status=WAITING
QBlastInfoEnd
--><p>
```

If the results are completed, the output will show the formatted results with status information like the following, with Status=READY:

```
<!--QBlastInfoBegin
    Status=READY
QBlastInfoEnd
--><p>

... <formated output here>
```

If you search results has the Status of WAITING, it is advisable to wait for few seconds before trying to get results again.

Additional formatting options are also available, for example:

```
$ lynx -source "http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?\
    CMD=Get&RID=1234-566-897&FORMAT_OBJECT=Alignment\
    &FORMAT_TYPE=HTML&DESCRIPTIONS=100&ALIGNMENTS=200\
    &ALIGNMENT_TYPE=Pairwise&OVERVIEW=yes
```

This is interpreted as "get results for RID RID=1234-566-897, showing the alignments in pairwise form, produce HTML output, display at most 100 descriptions and 200 alignments, and show the graphical overview.

Fro more details please refer to the sections dealing with "Commands" and "Parameters". Individual descriptions for each command and parameter are available.

3 Commands

Every URLAPI request must have the following format:

```
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=<command>{&<name>=<value>}
```

Where; the <command> defines type of action to be performed, while <name> and <value> are parameters of the <command>. Some commands have mandatory parameters, for example the "RID" parameter for "CMD=Get" is required and must be a valid RID.

The following commands are supported: Web, Put, Get, Delete, Info.

3.1 Web

The 'Web' generates NCBI BLAST HTML pages.

Parameters accepted: PAGE and all parameters corresponding to HTML elements.

Example:

```
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&\
PAGE=Proteins&DATABASE=swissprot
```

3.2 Put

The 'Put' command - Sends search requests to QBl_Ast .

Parameters accepted: AUTO_FORMAT (p.7) , COMPOSITION_BASED_STATISTICS (p.8) , DATABASE (p.8) , DB_GENETIC_CODE (p.8) , ENDPOINTS (p.8) , ENTREZ_QUERY (p.9) , EXPECT (p.9) , FILTER (p.10) , GAPCOSTS (p.10) , GENETIC_CODE (p.11) , HITLIST_SIZE (p.11) , I_THRESH (p.11) , LAYOUT (p.11) , LCASE_MASK (p.11) , MATRIX_NAME (p.12) , NUCL_PENALTY (p.12) , NUCL_REWARD (p.12) , OTHER_ADVANCED (p.13) , PERC_IDENT (p.14) , PHI_PATTERN (p.14) , PROGRAM (p.14) , QUERY (p.14) , QUERY_FILE (p.15) , QUERY_BELIEVE_DEFLINE (p.15) , QUERY_FROM (p.15) , QUERY_TO (p.15) , SEARCHSP_EFF (p.16) , SERVICE (p.16) , THRESHOLD (p.16) , UNGAPPED_ALIGNMENT (p.17) , WORD_SIZE (p.17)

Example:

```
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?\
CMD=Put&QUERY=MKN&DATABASE=nr&PROGRAM=blastp&FILTER=L&HITLIST_SZE=500
```

3.3 Get

The 'Get' command gets formated results from QBl_Ast .

Parameters accepted: ALIGNMENTS (p.7) , ALIGNMENT_VIEW (p.7) , DESCRIPTIONS (p.8) , ENTREZ_LINKS_NEW_WINDOW (p.9) , EXPECT_LOW (p.9) , EXPECT_HIGH (p.9) , FORMAT_ENTREZ_QUERY (p.10) , FORMAT_OBJECT (p.10) , FORMAT_TYPE (p.10) , NCBI_GI (p.12) , RID (p.15) , RESULTS_FILE (p.16) , SERVICE (p.16) , SHOW_OVERVIEW (p.16)

Example:

```
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?\
CMD=Get&RID=954517013-7639-11119
```

3.4 Delete

The 'Delete' command deletes request and results corresponding to specified RID from the QBl_Ast .

Parameters accepted: RID (p.15)

3.5 Info

The 'Info' provides up-to-date information about NCBI Blast.

Parameters accepted: TARGET (p.17)

4 Parameters

Here we describe all parameters used in the URL API. The following descriptors are used:

Descr Text description of the parameter

Values Allowed values of the parameter

Default Default value

Commands Which commands recognize the parameter

blastall Analog of the parameter in NCBI blastall program

blastpgp Analog of the parameter in NCBI blastpgp program

The parameters sorted alphabetically:

4.1 ALIGNMENTS

Descr:	Number of alignments
Values:	integer value
Default:	500
Commands:	Get
blastall:	'-b'

4.2 ALIGNMENT_VIEW

Descr:	Type of alignment view (FORMAT_OBJECT=Alignment only)
Values:	Pairwise, QueryAnchored, QueryAnchoredNoIdentities, FlatQueryAnchored, FlatQueryAnchoredNoIdentities, Tabular
Default:	Pairwise
Commands:	Get
blastall:	'-m'

4.3 AUTO_FORMAT

Descr:	Determines whether to format results automatically
Values:	Off, Semiauto, Fullauto
Default:	Off
Commands:	Put

4.4 CMD

Descr:	A command name
Values:	Web, Put, Get, Delete, Info
Default:	Web

4.5 COMPOSITION_BASED_STATISTICS

Descr:	Do search with tweak parameter set to true
Values:	yes, no
Default:	no
Commands:	Put
blastpgp:	-t

4.6 DATABASE

Descr:	Database name
Values:	valid database name
Default:	nr
Commands:	Put
blastall:	'-d'

4.7 DB_GENETIC_CODE

Descr:	Database genetic code (PROGRAM=tblast[nx] only)
Values:	integers: 1..16,21,22
Default:	1
Commands:	Put
blastall:	'-D'

4.8 DESCRIPTIONS

Descr:	Number of descriptions
Values:	integer value
Default:	500
Commands:	Get
blastall:	'-v'

4.9 ENDPOINTS

Descr:	Get only alignment endpoints in results (megablast only)
Values:	yes, no
Default:	no
Commands:	Put
megablast:	'-D 0'

4.10 ENTREZ_LINKS_NEW_WINDOW

Descr:	Add TARGET to Entrez links in formatted results
Values:	yes or no
Default:	empty
Commands:	Get
blastall:	does not support

4.11 ENTREZ_QUERY

Descr:	Entrez query to limit Blast search
Values:	Entrez query format
Default:	empty
Commands:	Put
blastall:	does not support

4.12 EXPECT

Descr:	Expect value
Values:	double type value
Default:	10.0
Commands:	Put
blastall:	'-e'

4.13 EXPECT_LOW

Descr:	Low expect value threshold for formatting
Values:	double type value
Default:	0
Commands:	Get
blastall:	does not support

4.14 EXPECT_HIGH

Descr:	High expect value threshold for formatting
Values:	double type value
Default:	equal to EXPECT
Commands:	Get
blastall:	does not support

4.15 FILTER

Descr:	Sequence filter identifier
Values:	"L" for Low Complexity, "R" for Human Repeats, "m" for Mask for Lookup
Default:	not set
Commands:	Put
blastall:	'-F'

It is possible to specify more than one filter in the URL request.

4.16 FORMAT_ENTREZ_QUERY

Descr:	Entrez query to limit formatting of Blast results
Values:	Entrez query format
Default:	empty
Commands:	Get
blastall:	does not support

4.17 FORMAT_OBJECT

Descr:	Specifies object to get
Values:	Alignment, Neighbors, PSSM
	SearchInfo
	TaxBlast, TaxblastParent, TaxBlastMultiFrame
Default:	Alignment
Commands:	Get

4.18 FORMAT_TYPE

Descr:	Type of formatting
Values:	HTML, Text, ASN.1, XML
Default:	HTML
Commands:	Get
blastall:	partially '-T'

4.19 GAPCOSTS

Descr:	Gap open and gap extend costs
Values:	Space separated float values
Default:	"5 2" for nuc-nuc, "11 1" for proteins, non-affine for megablast
Commands:	Put
blastall:	'-G' and '-E'

4.20 GENETIC_CODE

Descr:	Query genetic code
Values:	integers: 1..16,21,22
Default:	1
Commands:	Put
blastall:	'-Q'

4.21 HITLIST_SIZE

Descr:	Number of hits to keep
Values:	integer value
Default:	500
Commands:	Put
blastall:	is max('-v','-b')

4.22 I_THRESH

Descr:	Threshold for extending hits (PSI BLAST only)
Values:	float
Default:	0.001
Commands:	Put
blastpgp:	'-h'

4.23 LAYOUT

Descr:	Windows layout when use it in a browser
Values:	OneWindow, TwoWindows
Default:	OneWindow
Commands:	Put

4.24 LCASE_MASK

Descr:	Enable masking of lower case in query
Values:	yes, no
Default:	no
Commands:	Put
blastall:	'-U'

4.25 MEGABLAST

Descr:	Use Mega BLAST algorithm (nucleotide search only)
Values:	yes, no
Default:	no
Commands:	Put
blastall:	'-n'

4.26 MATRIX_NAME

Descr:	Matrix name (protein search only)
Values:	Valid matrix name
Default:	BLOSUM62
Commands:	Put
blastall:	'-M'

4.27 NCBI_GI

Descr:	Show NCBI GI
Values:	yes, no
Default:	no
Commands:	Get
blastall:	'-I'

4.28 NUCL_PENALTY

Descr:	Penalty for a nucleotide mismatch (blastn only)
Values:	negative integer value
Default:	-3
Commands:	Put
blastall:	'-q'

4.29 NUCL_REWARD

Descr:	Reward for a nucleotide match (blastn only)
Values:	integer value
Default:	1
Commands:	Put
blastall:	'-r'

4.30 OTHER_ADVANCED

Descr:	Other advanced options
Values:	A string combining the options: -G Cost to open a gap, see GAPCOSTS (p.10) -E Cost to extend a gap, see GAPCOSTS (p.10) -r Reward for match, see NUCL_REWARD (p.12) -q Penalty for mismatch, see NUCL_PENALTY (p.12) -e Expectation value (E), see EXPECT (p.9) -W Word size, see WORD_SIZE (p.17) -y Dropoff (X) for blast extensions in bits (default if zero) (Integer) default = 20 for nuc-nuc 7 for other programs Not applicable for megablast -X X dropoff value for gapped alignment (in bits) (Integer) default = 30 for nuc-nuc (blastn and megablast), 15 for other programs -Z Final X dropoff value for gapped alignment (in bits) (Integer) 50 for nuc-nuc (blastn), 25 for other programs megablast - not applicable -P 0 for multiple hits 1-pass, 1 for single hit 1-pass (Integer) Does not apply to blastn or megablast -A Multiple Hits window size (zero for single hit algorithm)(Integer) -I Number of database sequences to save hits for, see HITLIST_SIZE (p.11) -b Number of database sequences to show alignments for, see ALIGNMENTS (p.7) -v Number of database sequences to show one-line descriptions for, see DESCRIPTIONS (p.8) -Y Effective length of the search space, see SEARCHSP_EFF (p.16) -z Effective length of the database (use zero for the real size)(Real), default=0 -c Pseudocount constant for PSI-BLAST (Integer), default=7 -F Filtering directives, see FILTER (p.10)
Default:	empty
Commands:	Put

4.31 PAGE

Descr:	Type of the web page
Values:	Proteins, Nucleotides, Translations, MegaBlast
Default:	Proteins
Commands:	Web

4.32 PERC_IDENT

Descr:	Percent of identity cut-off threshold (megablast only)
Values:	integer between 0 (no cut-off) and 100 (exact matches only)
Default:	99
Commands:	Put
megablast:	'-p'

4.33 PHI_PATTERN

Descr:	Phi Blast pattern
Values:	string, valid PHI BLAST pattern
Default:	empty
Commands:	Put

4.34 PROGRAM

Descr:	Blast program name
Values:	blastn, blastp, blastx, tblastn, tblastx
Default:	blastn
Commands:	Put
blastall:	'-p'

4.35 PSSM

Descr:	PSI BLAST checkpoint
Values:	string encoded checkpoint
Default:	?
Commands:	Put
blastpgp:	'-R'

4.36 QUERY

Descr:	Sequence query (queries if MegaBlast page)
Values:	Accession(s), gi(s), or FASTA sequence(s)
Default:	mandatory option unless QUERYFILE is provided
Commands:	Put
blastall:	'-i'

4.37 QUERY_FILE

Descr:	File with sequence queries (MegaBlast only)
Values:	Accessions, gis, or FASTA sequences
Default:	empty
Commands:	Put
megablast:	'-i'

4.38 QUERY_BELIEVE_DEFLINE

Descr:	Whether to believe defline in FASTA query
Values:	yes, no
Default:	no
Commands:	Put

4.39 QUERY_FROM

Descr:	Start of subsequence (one offset)
Values:	integer value
Default:	0
Commands:	Put
blastall:	'-L'

4.40 QUERY_TO

Descr:	End of subsequence (one offset)
Values:	integer value
Default:	0, that means not to use subsequence
Commands:	Put
blastall:	'-L'

4.41 RID

Descr:	Request ID
Values:	Valid request ID
Default:	this is a mandatory field for Get
Commands:	Get

4.42 RESULTS_FILE

Descr:	Allow to download megablast results as a gzip-compressed file
Values:	yes, no
Default:	no
Commands:	Put

4.43 SEARCHSP_EFF

Descr:	Effective length of the search space
Values:	integer value
Default:	0
Commands:	Put
blastall:	'-Y'

4.44 SERVICE

Descr:	Blast service which needs to be performed
Values:	plain, psi, phi, rpsblast, megablast
Default:	plain
Commands:	Put, Get
blastall:	supports plain service
blastpgp:	supports psi and phi services

4.45 SHOW_OVERVIEW

Descr:	Show graphical overview
Values:	yes, no
Default:	yes
Commands:	Get

Corresponding to FORMAT_OBJECT viewer is to be used (if any).

4.46 THRESHOLD

Descr:	Threshold for extending hits
Values:	integer value
Default:	???
Commands:	Put
blastall:	'-f'

4.47 UNGAPPED_ALIGNMENT

Descr:	Should the ungapped alignment be performed ?
Values:	yes, no
Default:	no
Commands:	Put
blastall:	'-g'

4.48 WORD_SIZE

Descr:	Word size
Values:	integer value
Default:	3 for proteins, 11 for nuc-nuc, 28 for megablast
Commands:	Put
blastall:	'-W'

4.49 TARGET

Descr:	Sets target for the Info command
Values:	QBlast, databases
Default:	empty
Commands:	Info

5 Compatibility

We will attempt to make changes to the API in a backward compatible manner.

All commands produce HTML output which is human readable if loaded into a browser. However, NCBI reserves the right to change the format of BLAST results as we see fit and the format of the returned BLAST report may change with little or no notice.

6 The best way to use QBl_Ast

One should use only one or two threads to submit jobs, but you can submit a new job as soon as you get back the RID from the server. If the server is sending back RID's this means that it is up, healthy, and running. Since it usually takes less than a second to send back an RID you can actually put quite a few jobs into the queue in a short period of time. Please DO NOT start 50 threads to expedite the process! Flooding the server can lead to many problems and eventually we may be forced to block access from sites which flood the severs with no warning. We strongly suggest you limit your scripts to not send a request until you receive a RID from the server. Alternatively please introduce a "sleep" command in your script which sends request no less then one once per three seconds.

Also, if you get 'Status=WAITING' when you try to get results it really means that it is waiting, the request is not lost, please do not resubmit this search again. If query is lost for some unknown reason (which happens very rarely) you would get 'Status=UNKNOWN'. In this case, stop your script and contact blast-help@ncbi.nlm.nih.gov ASAP and we will look into the problem. Also if you are receiving no results from the BLAST server, do not rerun your scripts, contact us and we will investigate the matter.

7 More examples

7.1 Using telnet with URLAPI

Sending the "CMD=Info" request to QBl_Ast :

```
$ telnet www.ncbi.nlm.nih.gov 80
Trying 130.14.22.107...
Connected to www.ncbi.nlm.nih.gov.
Escape character is '^]'.
POST /blast/Blast.cgi HTTP/1.0
Content-type: application/x-www-form-urlencoded
Connection: Keep-Alive
User-Agent: qblastapp
Content-Length: 10

CMD=Info
HTTP/1.0 200 OK
Date: Thu, 29 Jun 2000 02:33:42 GMT
Server: Apache/1.3.9 (Unix) mod_fastcgi/2.2.2 mod_ssl/2.4.0....
IPConnectionLimit/1.0
Content-Type: text/html

<html>
22

# exlclusive databases
```

```
nr              1       TRUE
nr              2       FALSE
.....
```

7.2 Using lynx with URLAPI

Place url-encoded request into a file, say, urlenc.in, and run lynx:

```
    $ lynx "http://www.ncbi.nlm.nih.gov/blast/Blast.cgi" \
-post_data  -mime_header < urlenc.in
```

7.3 Web

Generate web page with default parameters:

```
CMD=Web
```

Generate protein search web page with default parameters.

```
CMD=Web&PAGE=Proteins
```

Generate nucleotides search web page with preset database name 'nr' and query GI number 555:

```
CMD=Web&PAGE=Nucleotides&DATABASE=nr&QUERY=555
```

7.4 Put

7.4.1 Using QUERY_BELIEVE_DEFLINE

If it is necessary to set a SeqId for a search query, this SeqId should be encoded into definition line (defline) of the input query in FASTA format. All invalid symbols should be replaced with URL hex codings, like:

```
while (<QUERY_FILE>) {
    s/>/%3E/g;
    s/ /+/g;
    s/\n/%0D%0A/g;
    s/\|/%7C/g;
    print;
}
```

To ask server to use that defline the QUERY_BELIEVE_DEFLINE (p.15) is to be set to 'yes'.

```
$ echo "CMD=Put&QUERY_BELIEVE_DEFLINE=yes&QUERY=%3Elcl|\
GB:D15320_1+(D15320)%0D%0AGACGCCGCCGCCACCACCGCCACCGCCGC\
AGCAGAAGCAGCGCACCGCAGGAGGGAAGATGCCGGCGGGGCACGGGCTGCGGGC\
GCGGACGGCGACCTCTTCGCGCGGCCGTTCCGCAAGAAGGGTTACATCCCGCTCA\
CCACCTACCTGAGGACGTACAAGATCGGCGATTACGTNGACGTCAAGGTGAACGG\
TG&DATABASE=nr&PROGRAM=blastn" \
| lynx "http://iblast1/blast/Blast.cgi" -post_data -mime_header | grep RID

RID = 962308435-15495-3269

$ echo "CMD=Get&RID=962308435-15495-3269&FORMAT_TYPE=ASN.1" | \
lynx "http://iblast1/blast/Blast.cgi" -post_data -mime_header
```

```
...
Seq-annot ::= {
...
        segs
          denseg {
            dim 2 ,
            numseg 3 ,
            ids {
              local
                str "GB:D15320_1" ,
```

```
CMD=Delete&RID=938103916-13276-29714
```

7.5 Info

Index