BLAST Searches at the Cloud

Basic Local Alignment Search Tool (BLAST) is a popular sequence similarity search tool. The National Center for Biotechnology Information (NCBI) maintains a public BLAST server at blast.ncbi.nlm.nih.gov. The NCBI BLAST server is intended mainly for interactive use, with the expectation that users will perform a moderate number of searches (i.e., 10-20) per day. Some users have many more searches to perform than can be accommodated at the NCBI site. For such batch users, the NCBI offers the stand-alone BLAST+ package that can be run on local hardware. Running the stand-alone package requires that the user have access to sufficient compute resources (e.g., a local cluster) as well as some degree of familiarity with IT issues and command-line applications. Additionally, some users do not need to perform a large number of searches on a regular basis, but may need to do so occasionally. A cloud provider may offer an acceptable solution for these batch users.

The NCBI is experimenting with a BLAST+ installation at Amazon Web Services (AWS). The NCBI is providing an Amazon Machine Image (AMI) which can be run directly and via the AWS Marketplace. The BLAST AMI includes the BLAST+ applications, a "FUSE" client that can cache databases from the NCBI (during the first search), a PERL script that mimics a subset of the NCBI URL API, and a simplified BLAST search webpage. The instance runs an http daemon for the remote BLAST searches. The FUSE client can download a limited set of databases (see Appendix two).

The BLAST AMI supports three different search methods. First, users can ssh into the instance and run stand-alone BLAST at the command-line. Second, the instance has a simple webpage for search submission. Finally, the instance supports a subset of the BLAST URL API interface (see Appendix three). The latter two methods are remote searches that return a "Request ID".

The remote search methods allow one user in a group to register at AWS, configure the instance, and handle payment for the EC2 instances (NCBI does not charge for use). Other members of a group can then use the instance through the URL API or webpage without their own AWS account. The URL API interface is ideal for users familiar with the BLAST URL API at the NCBI. Currently, the BLAST AMI runs searches on only one instance. Many simultaneous users, through the remote interface, could overload a single instance. The NCBI is working on an approach that distributes searches to multiple instances.

Information and links to the latest BLAST AMI are available at
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=CloudBlast

Appendix one discusses setting up an instance at AWS. Appendix two lists the BLAST databases that can be downloaded by the FUSE client. Appendix three provides information on the URL API at AWS. Appendix four provides information

about using the BLAST+ applications at AWS. Finally, there is a section providing Disclaimer and Copyright information.

The NCBI welcomes feedback on this experiment at AWS. You may provide feedback at
http://www.ncbi.nlm.nih.gov/sites/ehelp/feedback?Ncbi_App=blastAWS&Data=JiraApp:HD;

Appendix 1: Setting up an AWS instance

An account is required to run searches at AWS. An account may be obtained at http://aws.amazon.com (requires a credit card). Documentation on using EC2 at AWS is at http://docs.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/. There are also useful videos at http://aws.amazon.com/getting-started/ Information on EC2 pricing is at https://aws.amazon.com/ec2/pricing/ AWS has a Command Line Interface that is documented at https://aws.amazon.com/cli/

BLAST can be started via the AWS marketplace at
https://aws.amazon.com/marketplace/

BLAST searches will not run efficiently on smaller instances. An instance with 32gb of memory and a minimum of 32gb SSD is required for reasonable performance. In Marketplace, you can start with a smaller instance (m3.2xlarge) but your requirements may drive you to a large instance.

The nucleotide nt database contains about 56 billion bases (04/14/2014), but the sequence data is compressed 4-to-1. Hence, the sequences require about 14 GB. The protein nr database contains about 14 billion residues (04/14/2014), so the sequences require 14 GB. Experiments at the NCBI have shown that the m3.2xlarge instance ("General Purpose" with 30 GB of memory and SSD) as well as the c3.4xlarge ("Compute Optimized" with 30 GB of memory and SSD) instances are good choices for the BLAST AMI. The BLAST AMI will concatenate and construct a filesystem at first boot for the BLAST databases from all the attached ephemeral storage. Therefore, an m3.2xlarge, which may have up to 2 80GB ephemeral drives attached, may have 160GB of BLAST database storage. The nr and nt databases will completely fill a 32 GB SSD (causing searches to fail), so a larger instance type than an m3.large should be selected if both protein and nucleotide databases will be searched. BLAST will not run on micro instances.

If you wish to allow remote searches, then you will need to enable HTTP access when you configure your instance. Please see Appendix 3. If you allow HTTP access, it is very important that you configure your security group properly. If the group is not restricted, anybody with knowledge of your public DNS may perform

remote searches on your instance or see all the BLAST results for remote searches. If possible, you should configure the security group of your instance to allow HTTP access only from a specific set of IP's (e.g. your lab's network). You should also only share the public DNS of your instance with people who have your permission to use it.

As a security precaution, the BLAST AMI blocks all ports but 22 and 80.

The BLAST results are kept on ephemeral storage. They will disappear when the instance is stopped or terminated. If you wish to keep them, then you should move them to permanent storage either at AWS or your institution.

Appendix 2: The FUSE client and BLAST databases

The FUSE client can download the BLAST databases listed in the table below. The client is configured to download the most recent version of these BLAST databases at the time of the first search, but these BLAST databases will not be automatically updated as newer databases become available. To force the client to switch to the newest database (and discard the last one downloaded), it is necessary to issue the following commands:

```
/sbin/remote-fuser-ctl.pl --stop
/sbin/remote-fuser-ctl.pl --start
```

The FUSE client downloads the BLAST database over the network, so the first BLAST search against a new database will be slow. The client caches the database on disk and subsequent searches will be faster.

Or, terminate the existing instance and start a new one. The new instance will cache the latest data.

Name	Title	Type
nt	Nucleotide collection	DNA
nr	Non-redundant	Protein
refseq_rna	NCBI Transcript Reference Sequences	DNA
refseq_protein	NCBI Protein Reference Sequences	Protein
swissprot	Non-redundant UniProtKB/SwissProt sequences	Protein
pdbaa	PDB protein database	Protein

pdbnt	PDB nucleotide database	DNA

Appendix 3: the URL API at AWS

You may run BLAST searches using an API like the NCBI BLAST URL API (http://www.ncbi.nlm.nih.gov/blast/Doc/urlapi.html). Every call to the URL API must include the CMD parameter, which can take one of four different arguments. You may use "Put" for search submission, "Get" for checking the status of a submission or retrieving results, "Delete" to remove a search and its results or "DisplayRIDs" to list all RIDs in the system. Each argument to CMD requires or allows certain other parameters as described in the table below.

To run a search, you should:

1.) Launch a BLAST AMI. The security group must allow access through HTTP. If possible, restrict the allowed IP's for HTTP so as to include only those of the intended users. See documentation under http://docs.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/

2.) Once the instance has launched, obtain the Public DNS from the EC2 Management Console. The Public DNS will look something like: ec2-54-82-43-97.compute-1.amazonaws.com though the exact numbers and letters will be depend upon your instance.

3.) Launch a search using your Public DNS plus cgi-bin/blast.cgi. For the example Public DNS given above, a URL to search the sequence of accession u00001 against nt would be: ec2-54-82-43-97.compute-1.amazonaws.com/cgi-bin/blast.cgi?QUERY=u00001&DATABASE=nt&PROGRAM=blastn&CMD=Put

4.) The instance will return a "Request ID" (RID). You may check the status of the search by invoking blast.cgi with the parameters CMD=Get, FORMAT_OBJECT=SearchInfo, and RID=VALUE, where VALUE is the RID returned by the server. The server will respond with a page containing a QBlastInfo block with the status set as one of WAITING, UNKNOWN, or READY.

5.) Once the search is finished, you may retrieve results by invoking blast.cgi with the parameters CMD=GET and RID=VALUE, where VALUE is the RID the system returned. An example would be "CMD=Get&RID=6NTRF1YLO8". You may also specify the report type, but need to include FORMAT_OBJECT=Alignment (e.g, add "FORMAT_OBJECT=Alignment&FORMAT_TYPE=Tabular").

The BLAST instance at AWS supports a subset of the NCBI URL parameters. The table below lists the supported parameters. An asterisk (*) in the CMD column indicates that the parameter is required for that command.

Parameter	Definition	Type	CMD	Allowed values
QUERY	Search query	String	Put *	Accession, GI, or FASTA.
DATABASE	BLAST database	String	Put *	Database from appendix 2 or one uploaded to blastdb_custom (see appendix 4)
PROGRAM	BLAST program	String	Put *	One of blastn, megablast, blastp, blastx, tblastn, tblastx
FILTER	Low complexity filtering	String	Put	F to disable. T or L to enable. Prepend "m" for mask at lookup (e.g., mL)
FORMAT_TYPE	Report type	String	Put, Get	HTML, Text, XML, or Tabular. HTML is the default. Must be used with FORMAT_OBJECT if the CMD is Get.
EXPECT	Expect value	Double	Put	Number greater than zero
NUCL_REWARD	Reward for matching bases (BLASTN and megaBLAST)	Integer	Put	Integer greater than zero.
NUCL_PENALTY	Cost for mismatched bases (BLASTN and megaBLAST)	Integer	Put	Integer less than zero
GAPCOSTS	Gap existence and extension costs	String	Put	Pair of positive integers separated by a space such as "11 1"
HITLIST_SIZE	Number of database sequences to keep.	Integer	Put, Get	Integer greater than zero.
DESCRIPTIONS	Number of descriptions to print (applies to HTML and Text)	Integer	Put, Get	Integer greater than zero.
NCBI_GI	Show NCBI GI's in report	String	Put, Get	T or F
RID	Results to retrieve	String	Get *, Delete *	The Request ID (RID) returned when the search was submitted.
THRESHOLD	Neighboring score for initial words	Integer	Put	Positive integer (BLASTP default is 11). Does not apply to BLASTN or megablast.
FORMAT_OBJECT	Object type.	String	Get *	SearchInfo (status check) or Alignment (report formatting)
NUM_THREADS	Number of virtual CPU's to use.	Integer	Put	Integer greater than zero. Default is 1.

If a parameter is not required and not provided, then the default value will be used. That default value may depend upon the BLAST search you are running.

If you have not restricted your security group, please also note that anybody with knowledge of your public DNS can use CMD=DisplayRIDs to see all remote searches available on the system

Appendix 4: Running BLAST+ at AWS

Users familiar with command line interfaces can run the BLAST+ applications directly on their AWS EC2 instance by logging on via ssh and issuing commands at the terminal prompt. The BLAST AMI uses Ubuntu LINUX, so you should use the user name "ubuntu" rather than "ec2-user" for the ssh connection.

The BLAST AMI already has the BLAST+ applications installed and will retrieve the databases listed earlier upon demand. Additionally, users may manually download other databases from the NCBI FTP site using the update_blastdb.pl program as follows:

```
$ cd /blast/blastdb_custom
$ update_blastdb.pl –-decompress –-passive <dbname>
```

These (and any other) BLAST databases installed in the /blast/blastdb_custom directory in the AWS instance will be immediately available for searching via the URL API, web interface, and BLAST+ command line applications on that instance.

The BLAST+ manual contains instructions for running the applications and is at http://www.ncbi.nlm.nih.gov/books/NBK1763

DISCLAIMER

Although all reasonable efforts have been taken to ensure the accuracy and reliability of the software and data, The NLM and the U.S. Government do not and cannot warrant the performance or results that may be obtained by using this software or data. The NLM and the U.S. Government disclaim all warranties, express or implied, including warranties of performance, merchantability or fitness for any particular purpose.

There are a number of known issues and limitations to the approach described here. They are listed below.

There is no load balancing in the current AMI. A number of users running searches at the same time could overload an instance. The NCBI hopes to address this in a future version.

It is possible that searches at an AWS instance could produce different results than those at the NCBI BLAST website. The different results could be caused by differences in the version of the software, BLAST database, or other issues beyond the control of the NCBI.

A running instance will download BLAST databases from the NCBI. The instance will normally place minimal demands on the NCBI website since the database is cached after the first download. Use of an instance in a manner that places heavy demands on the NCBI site may result in the instance being blocked.

The BLAST results and databases are kept on ephemeral storage at AWS. User error (e.g., premature stopping or termination of an instance) could result in loss of data. Problems at AWS or with NCBI software could also result in the loss of data.

COPYRIGHT

With the exception of certain third-party files summarized below, this software is a "United States Government Work" under the terms of the United States Copyright Act. It was written as part of the authors' official duties as United States Government employees and thus cannot be copyrighted. This software is freely available to the public for use. The National Library of Medicine and the U.S. Government have not placed any restriction on its use or reproduction.

Please cite the authors in any work or product based on this material [1, 2].

1: NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2013 Jan;41(Database issue):D8-D20. doi: 10.1093/nar/gks1189. Epub 2012 Nov 27. PubMed PMID: 23193264; PubMed Central PMCID: PMC3531099.

2: Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PubMed Central PMCID: PMC2803857.

EXCEPTIONS (in all cases excluding NCBI-written makefiles):

Location: configure
Authors: Free Software Foundation, Inc.
License: Unrestricted; at top of file

Location: config.guess, config.sub
Authors: FSF
License: Unrestricted when distributed with the Toolkit; standalone, GNU General Public License [gpl.txt]

Location: {src,include}/dbapi/driver/ftds*/freetds
Authors: See src/dbapi/driver/ftds*/freetds/AUTHORS
License: GNU Library/Lesser General Public License [src/dbapi/driver/ftds*/freetds/COPYING.LIB]

Location: include/dbapi/driver/odbc/unix_odbc
Authors: Peter Harvey and Nick Gorham
License: GNU LGPL

Location: {src,include}/gui/widgets/FLU
Authors: Jason Bryan
License: GNU LGPL

Location: {src,include}/gui/widgets/Fl_Table
Authors: Greg Ercolano
License: GNU LGPL

Location: include/util/bitset
Author: Anatoliy Kuznetsov
License: MIT [include/util/bitset/license.txt]

Location: {src,include}/util/compress/bzip2
Author: Julian R Seward
License: BSDish [src/util/compress/bzip2/LICENSE]

Location: {src,include}/util/compress/zlib
Authors: Jean-loup Gailly and Mark Adler
License: BSDish [include/util/compress/zlib/zlib.h]

Location: {src,include}/util/regexp
Author: Philip Hazel
License: BSDish [src/util/regexp/doc/LICENCE]

Location: {src,include}/misc/xmlwrapp
Author: Peter J Jones at al. [src/misc/xmlwrapp/AUTHORS]

License: BSDish [src/misc/xmlwrapp/LICENSE]

Q. Our product makes use of the NCBI source code, and we did changes and additions to that version of the NCBI code to better fit it to our needs. Can we copyright the code, and how?

A. You can copyright only the *changes* or the *additions* you made to the NCBI source code. You should identify unambiguously those sections of the code that were modified, e.g. by commenting any changes you made in the code you distribute. Therefore, your license has to make clear to users that your product is a combination of code that is public domain within the U.S. (but may be subject to copyright by the U.S. in foreign countries) and code that has been created or modified by you.

Q. Can we (re)license all or part of the NCBI source code?

A. No, you cannot license or relicense the source code written by NCBI since you cannot claim any copyright in the software that was developed at NCBI as a 'government work' and consequently is in the public domain within the U.S.