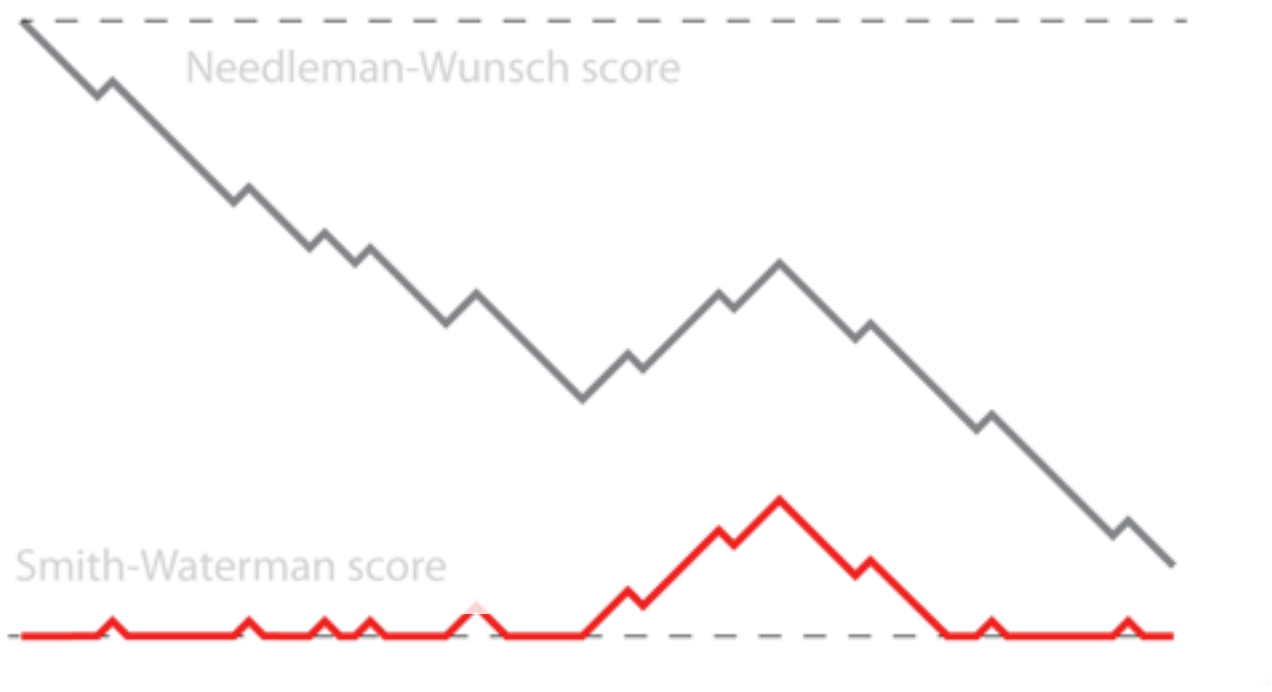
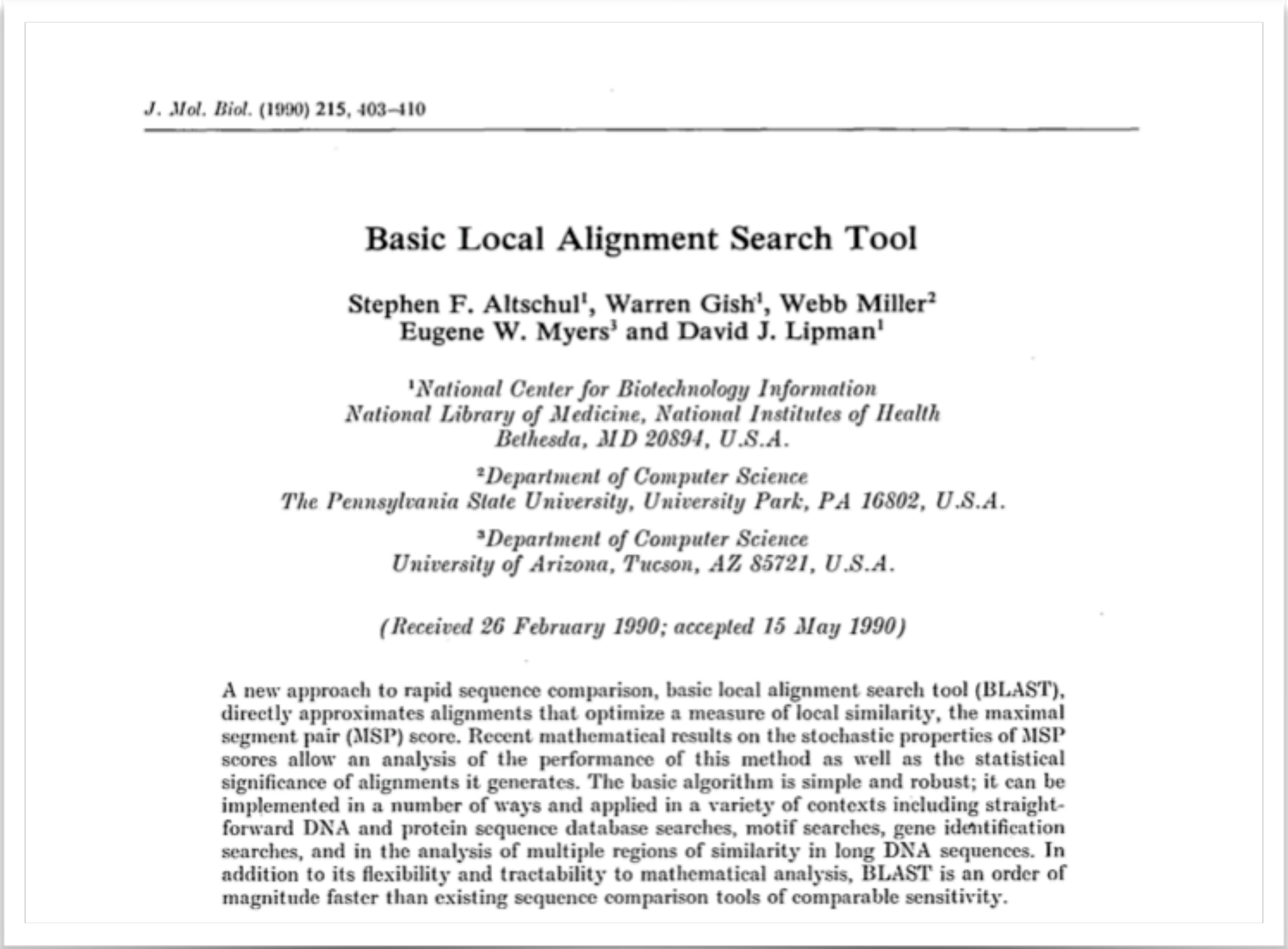


How to run BLAST locally

Jarek Bryk

Huddersfield, 29th March 2017

J. Mol. Biol. (1990) 215, 403-410



J. Mol. Biol. (1990) **215**, 403–410

Basic Local Alignment Search Tool

**Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]**

[1]*National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.*

[2]*Department of Computer Science
The Pennsylvania State University, University Park, PA 16802, U.S.A.*

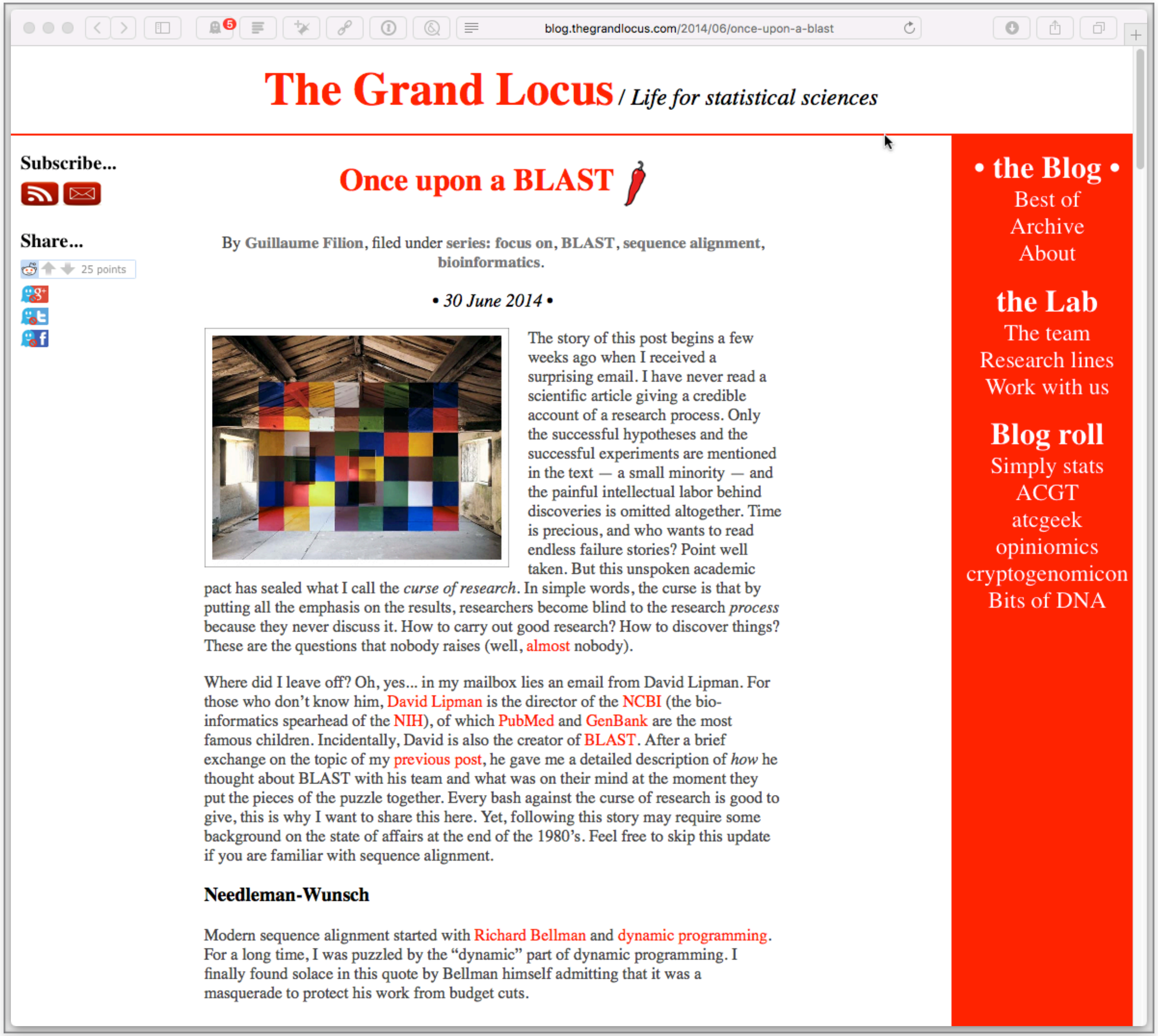
[3]*Department of Computer Science
University of Arizona, Tucson, AZ 85721, U.S.A.*

(Received 26 February 1990; accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

blog.thegrandlocus.com/2014/06/once-upon-a-blast

(the origin story of BLAST)



The Grand Locus / *Life for statistical sciences*

Once upon a BLAST

By Guillaume Filion, filed under series: focus on, BLAST, sequence alignment, bioinformatics.

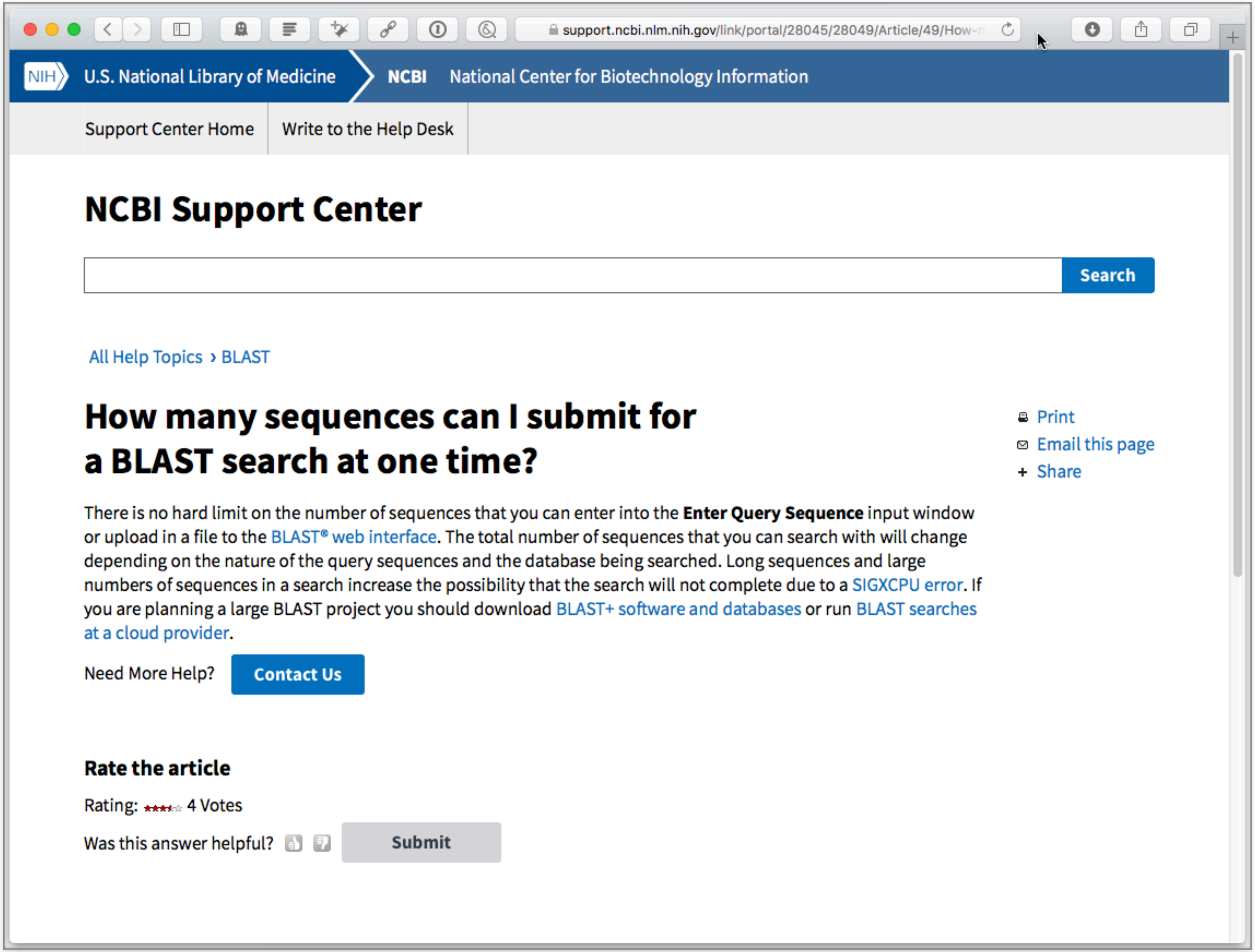
• *30 June 2014* •

The story of this post begins a few weeks ago when I received a surprising email. I have never read a scientific article giving a credible account of a research process. Only the successful hypotheses and the successful experiments are mentioned in the text — a small minority — and the painful intellectual labor behind discoveries is omitted altogether. Time is precious, and who wants to read endless failure stories? Point well taken. But this unspoken academic pact has sealed what I call the *curse of research*. In simple words, the curse is that by putting all the emphasis on the results, researchers become blind to the research *process* because they never discuss it. How to carry out good research? How to discover things? These are the questions that nobody raises (well, almost nobody).

Where did I leave off? Oh, yes... in my mailbox lies an email from David Lipman. For those who don't know him, David Lipman is the director of the NCBI (the bioinformatics spearhead of the NIH), of which PubMed and GenBank are the most famous children. Incidentally, David is also the creator of BLAST. After a brief exchange on the topic of my previous post, he gave me a detailed description of *how* he thought about BLAST with his team and what was on their mind at the moment they put the pieces of the puzzle together. Every bash against the curse of research is good to give, this is why I want to share this here. Yet, following this story may require some background on the state of affairs at the end of the 1980's. Feel free to skip this update if you are familiar with sequence alignment.

Needleman-Wunsch

Modern sequence alignment started with Richard Bellman and dynamic programming. For a long time, I was puzzled by the "dynamic" part of dynamic programming. I finally found solace in this quote by Bellman himself admitting that it was a masquerade to protect his work from budget cuts.


support.ncbi.nlm.nih.gov/link/portal/28045/28049/Article/49/How-
NIH
U.S. National Library of Medicine
NCBI
National Center for Biotechnology Information
Support Center Home
Write to the Help Desk
Search
All Help Topics › BLAST
Print
Email this page
Share

NCBI Support Center

How many sequences can I submit for a BLAST search at one time?

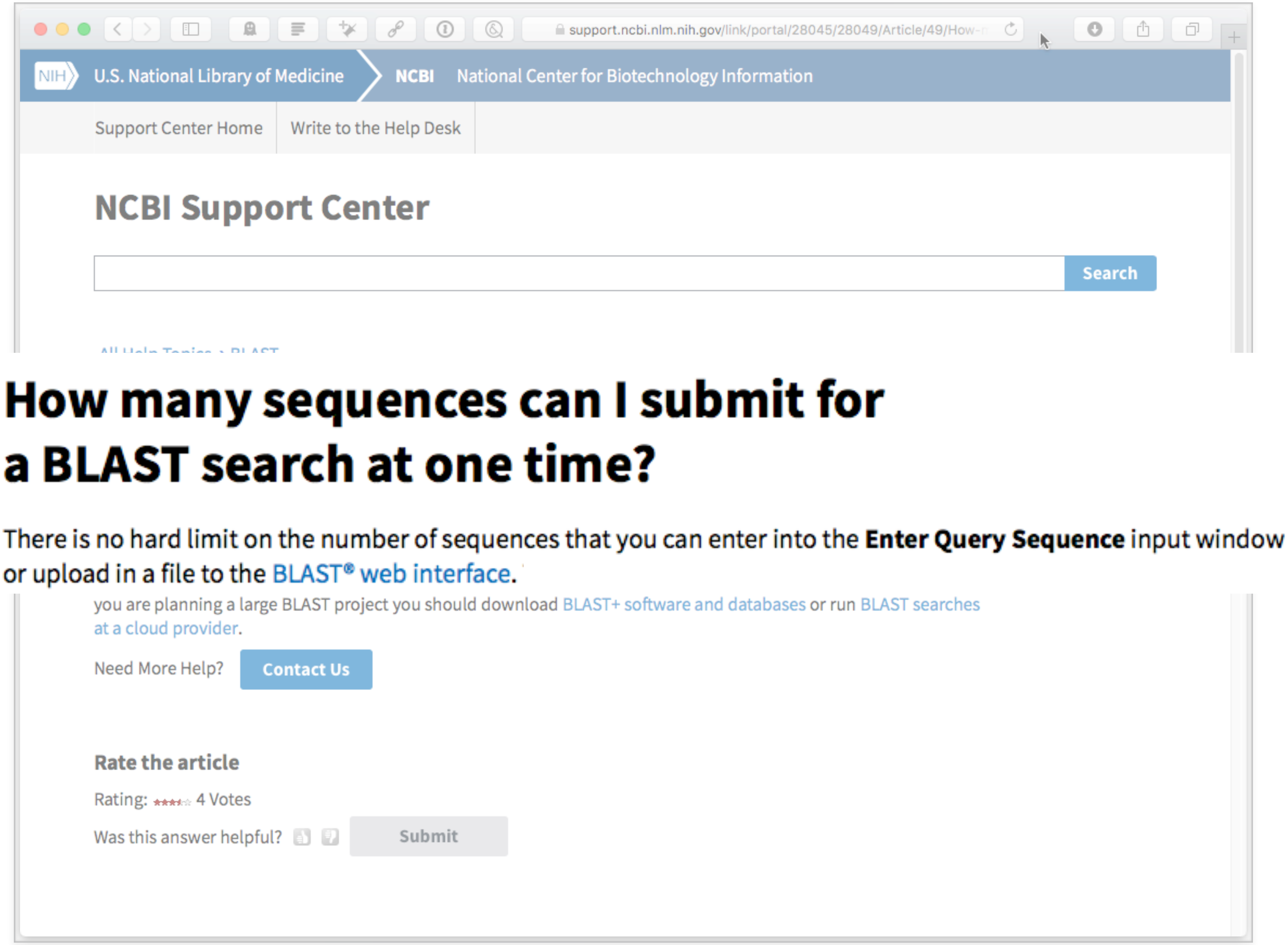
There is no hard limit on the number of sequences that you can enter into the **Enter Query Sequence** input window or upload in a file to the BLAST® web interface. The total number of sequences that you can search with will change depending on the nature of the query sequences and the database being searched. Long sequences and large numbers of sequences in a search increase the possibility that the search will not complete due to a SIGXCPU error. If you are planning a large BLAST project you should download BLAST+ software and databases or run BLAST searches at a cloud provider.

Need More Help?

Rate the article

Rating: ★★★★☆ 4 Votes

Was this answer helpful?



support.ncbi.nlm.nih.gov/link/portal/28045/28049/Article/49/How-

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information

Support Center Home | Write to the Help Desk

NCBI Support Center

Search

All Help Topics > BLAST

How many sequences can I submit for a BLAST search at one time?

There is no hard limit on the number of sequences that you can enter into the **Enter Query Sequence** input window or upload in a file to the BLAST® web interface.

you are planning a large BLAST project you should download BLAST+ software and databases or run BLAST searches at a cloud provider.

Need More Help? Contact Us

Rate the article

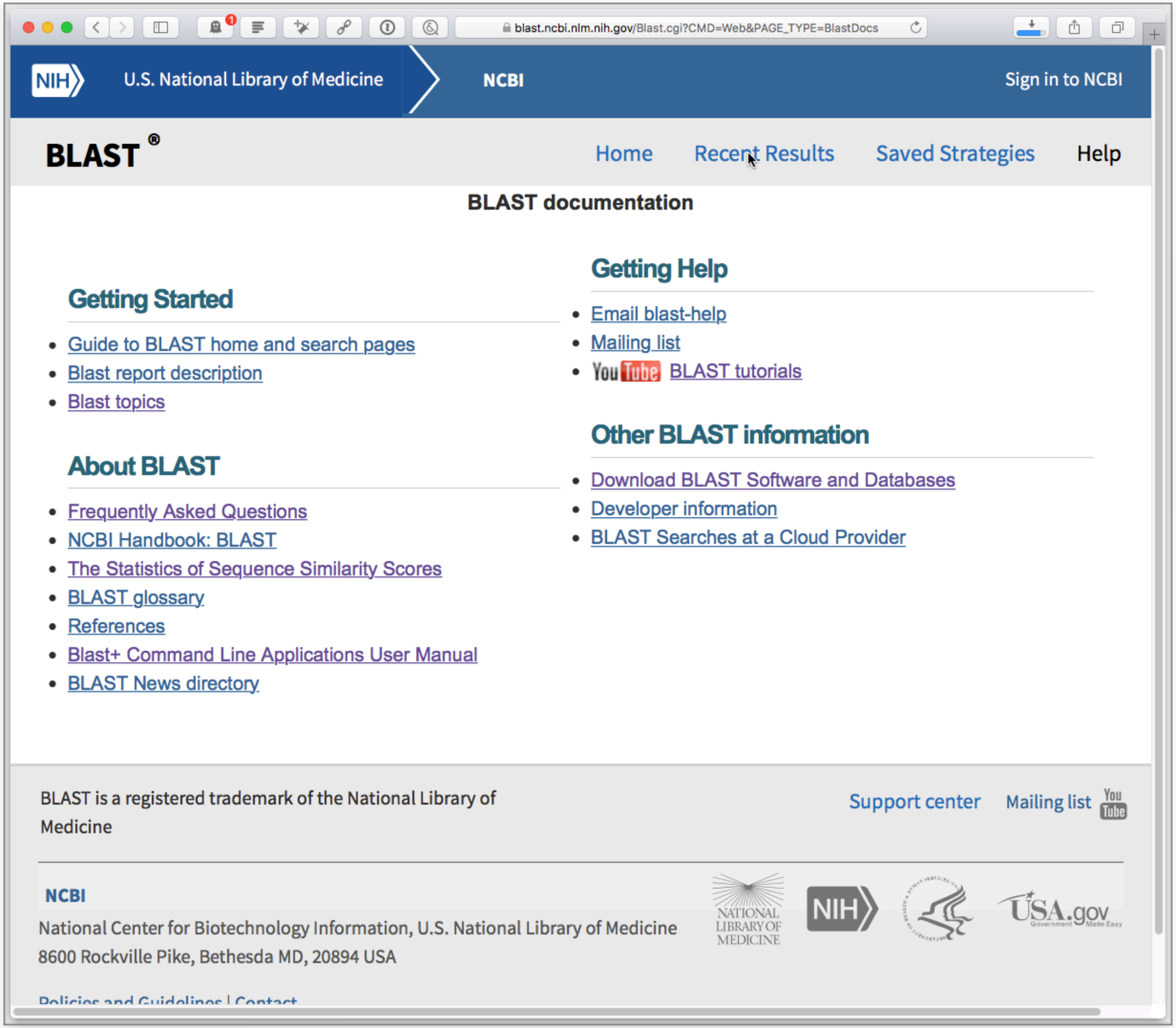
Rating: ★★★★☆ 4 Votes

Was this answer helpful? Submit

>100?

If you have a lot of sequences to BLAST and/or need to be able to modify the BLAST output, BLAST+ is the only way to do it

blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs


blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs
NIH
U.S. National Library of Medicine
NCBI
Sign in to NCBI
BLAST®
Home
Recent Results
Saved Strategies
Help

BLAST documentation

Getting Started

- Guide to BLAST home and search pages
- Blast report description
- Blast topics

About BLAST

- Frequently Asked Questions
- NCBI Handbook: BLAST
- The Statistics of Sequence Similarity Scores
- BLAST glossary
- References
- Blast+ Command Line Applications User Manual
- BLAST News directory

Getting Help

- Email blast-help
- Mailing list
- YouTube BLAST tutorials

Other BLAST information

- Download BLAST Software and Databases
- Developer information
- BLAST Searches at a Cloud Provider

BLAST is a registered trademark of the National Library of Medicine

NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

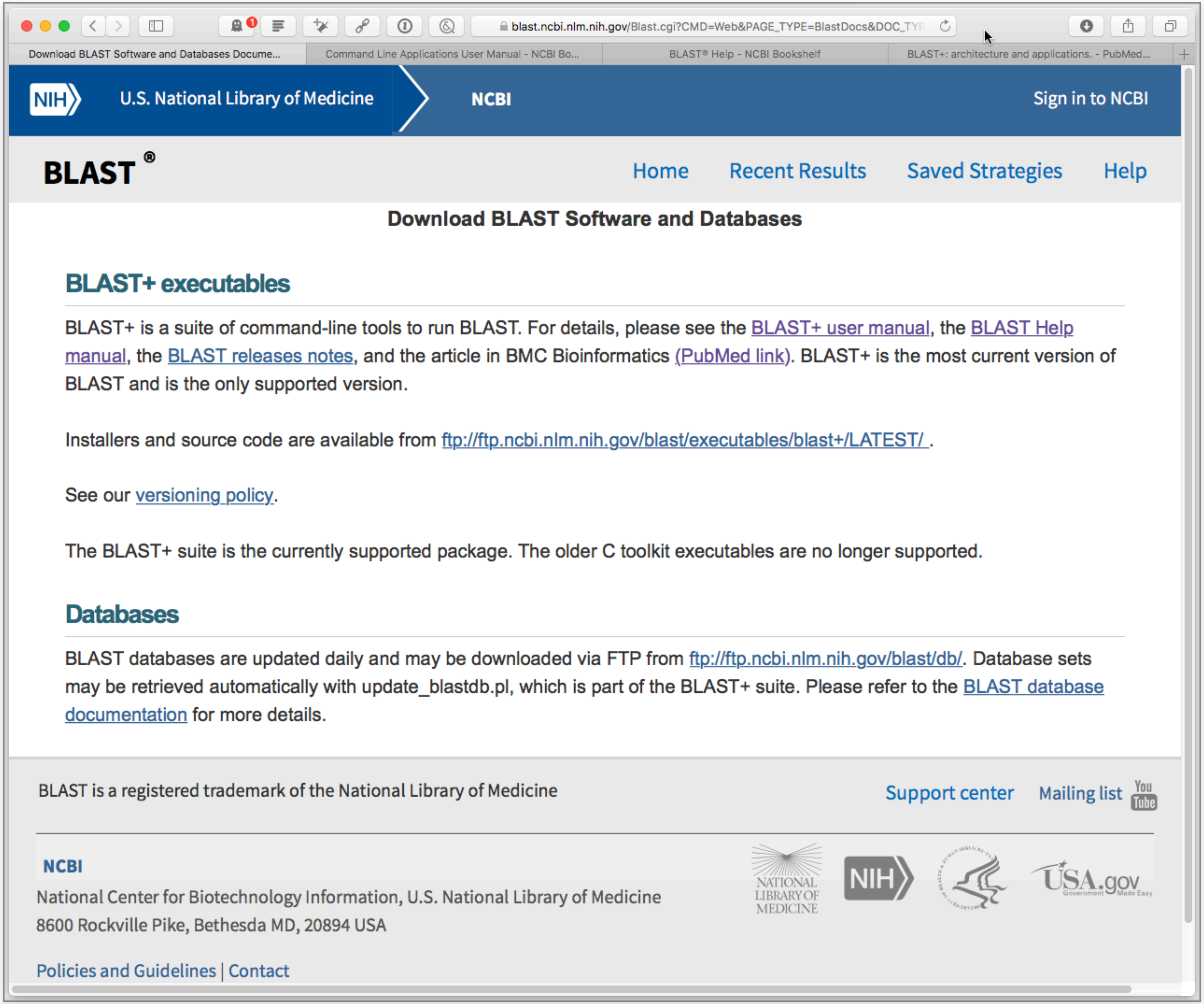
NATIONAL LIBRARY OF MEDICINE
NIH
USA.gov
Policies and Guidelines | Contact

Install BLAST+

Download or make the database

Prepare your query

Run BLAST with desired options


blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYP
Download BLAST Software and Databases Docume...
Command Line Applications User Manual - NCBI Bo...
BLAST® Help - NCBI Bookshelf
BLAST+: architecture and applications. - PubMed...
NIH
U.S. National Library of Medicine
NCBI
Sign in to NCBI
BLAST®
Home
Recent Results
Saved Strategies
Help

Download BLAST Software and Databases

BLAST+ executables

BLAST+ is a suite of command-line tools to run BLAST. For details, please see the BLAST+ user manual, the BLAST Help manual, the BLAST releases notes, and the article in BMC Bioinformatics (PubMed link). BLAST+ is the most current version of BLAST and is the only supported version.

Installers and source code are available from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ .

See our versioning policy.

The BLAST+ suite is the currently supported package. The older C toolkit executables are no longer supported.

Databases

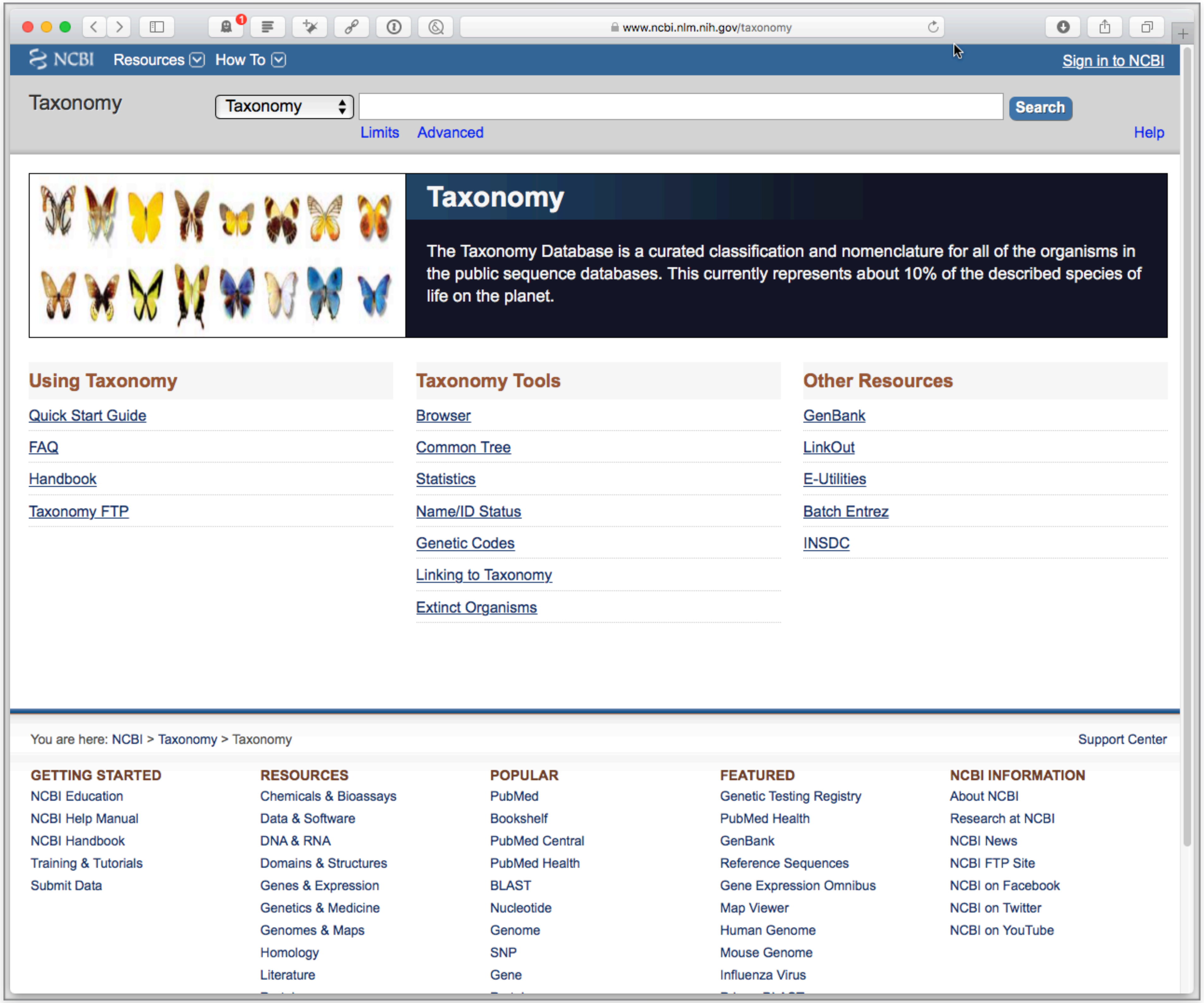
BLAST databases are updated daily and may be downloaded via FTP from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. Database sets may be retrieved automatically with update_blastdb.pl, which is part of the BLAST+ suite. Please refer to the BLAST database documentation for more details.

BLAST is a registered trademark of the National Library of Medicine

NCBI
National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

NATIONAL LIBRARY OF MEDICINE
NIH
USA.gov Government Made Easy
Policies and Guidelines | Contact

www.ncbi.nlm.nih.gov/taxonomy



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

- Quick Start Guide
- FAQ
- Handbook
- Taxonomy FTP

Taxonomy Tools

- Browser
- Common Tree
- Statistics
- Name/ID Status
- Genetic Codes
- Linking to Taxonomy
- Extinct Organisms

Other Resources

- GenBank
- LinkOut
- E-Utilities
- Batch Entrez
- INSDC

You are here: NCBI > Taxonomy > Taxonomy

Support Center

GETTING STARTED

NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
Submit Data

RESOURCES

Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature

POPULAR

PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene

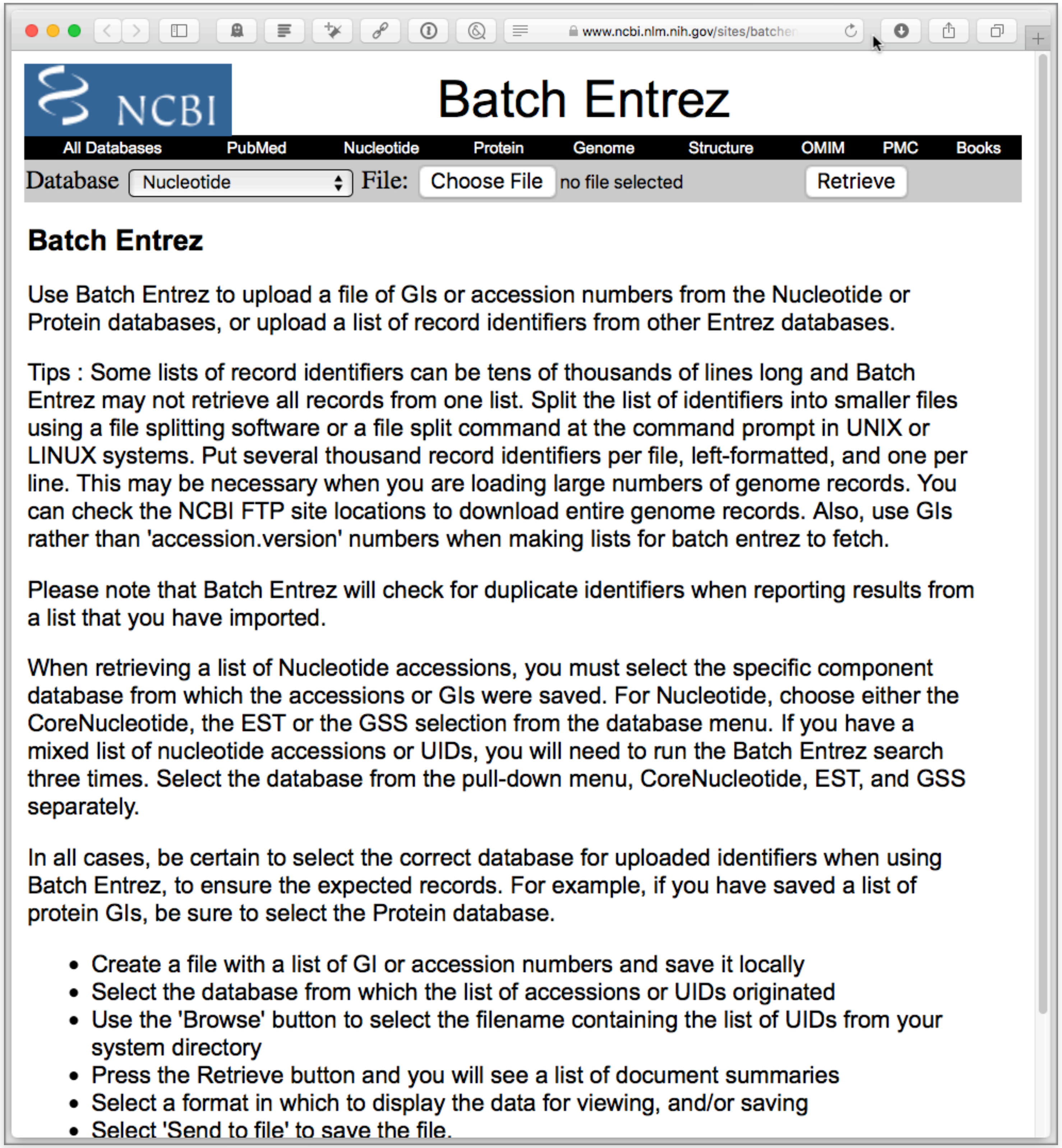
FEATURED

Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
Gene Expression Omnibus
Map Viewer
Human Genome
Mouse Genome
Influenza Virus

NCBI INFORMATION

About NCBI
Research at NCBI
NCBI News
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube

www.ncbi.nlm.nih.gov/sites/batchentrez



Batch Entrez

Use Batch Entrez to upload a file of GIs or accession numbers from the Nucleotide or Protein databases, or upload a list of record identifiers from other Entrez databases.

Tips : Some lists of record identifiers can be tens of thousands of lines long and Batch Entrez may not retrieve all records from one list. Split the list of identifiers into smaller files using a file splitting software or a file split command at the command prompt in UNIX or LINUX systems. Put several thousand record identifiers per file, left-formatted, and one per line. This may be necessary when you are loading large numbers of genome records. You can check the NCBI FTP site locations to download entire genome records. Also, use GIs rather than 'accession.version' numbers when making lists for batch entrez to fetch.

Please note that Batch Entrez will check for duplicate identifiers when reporting results from a list that you have imported.

When retrieving a list of Nucleotide accessions, you must select the specific component database from which the accessions or GIs were saved. For Nucleotide, choose either the CoreNucleotide, the EST or the GSS selection from the database menu. If you have a mixed list of nucleotide accessions or UIDs, you will need to run the Batch Entrez search three times. Select the database from the pull-down menu, CoreNucleotide, EST, and GSS separately.

In all cases, be certain to select the correct database for uploaded identifiers when using Batch Entrez, to ensure the expected records. For example, if you have saved a list of protein GIs, be sure to select the Protein database.

- Create a file with a list of GI or accession numbers and save it locally
- Select the database from which the list of accessions or UIDs originated
- Use the 'Browse' button to select the filename containing the list of UIDs from your system directory
- Press the Retrieve button and you will see a list of document summaries
- Select a format in which to display the data for viewing, and/or saving
- Select 'Send to file' to save the file.

main function

```
blastn -task megablast -db db/refseq_rna
-query test_query.txt -dust no -max_target_seqs 1
-outfmt "6 qseqid sseqid evalue pident stitle"
```

variant of the function

```
blastn -task megablast -db db/refseq_rna
-query test_query.txt -dust no -max_target_seqs 1
-outfmt "6 qseqid sseqid evalue pident stitle"
```

location of the database

```
blastn -task megablast -db db/refseq_rna
-query test_query.txt -dust no -max_target_seqs 1
-outfmt "6 qseqid sseqid evalue pident stitle"
```

```
blastn -task megablast -db db/refseq_rna
-query test_query.txt -dust no -max_target_seqs 1
-outfmt "6 qseqid sseqid evalue pident stitle"
```

location of the query (multi FASTA)

```
blastn -task megablast -db db/refseq_rna
-query test_query.txt -dust no -max_target_seqs 1
-outfmt "6 qseqid sseqid evalue pident stitle"
```

mask low complexity regions in query

```
blastn -task megablast -db db/refseq_rna
```

number of hits returned per query

```
-query test_query.txt -dust no -max_target_seqs 1
-outfmt "6 qseqid sseqid evalue pident stitle"
```

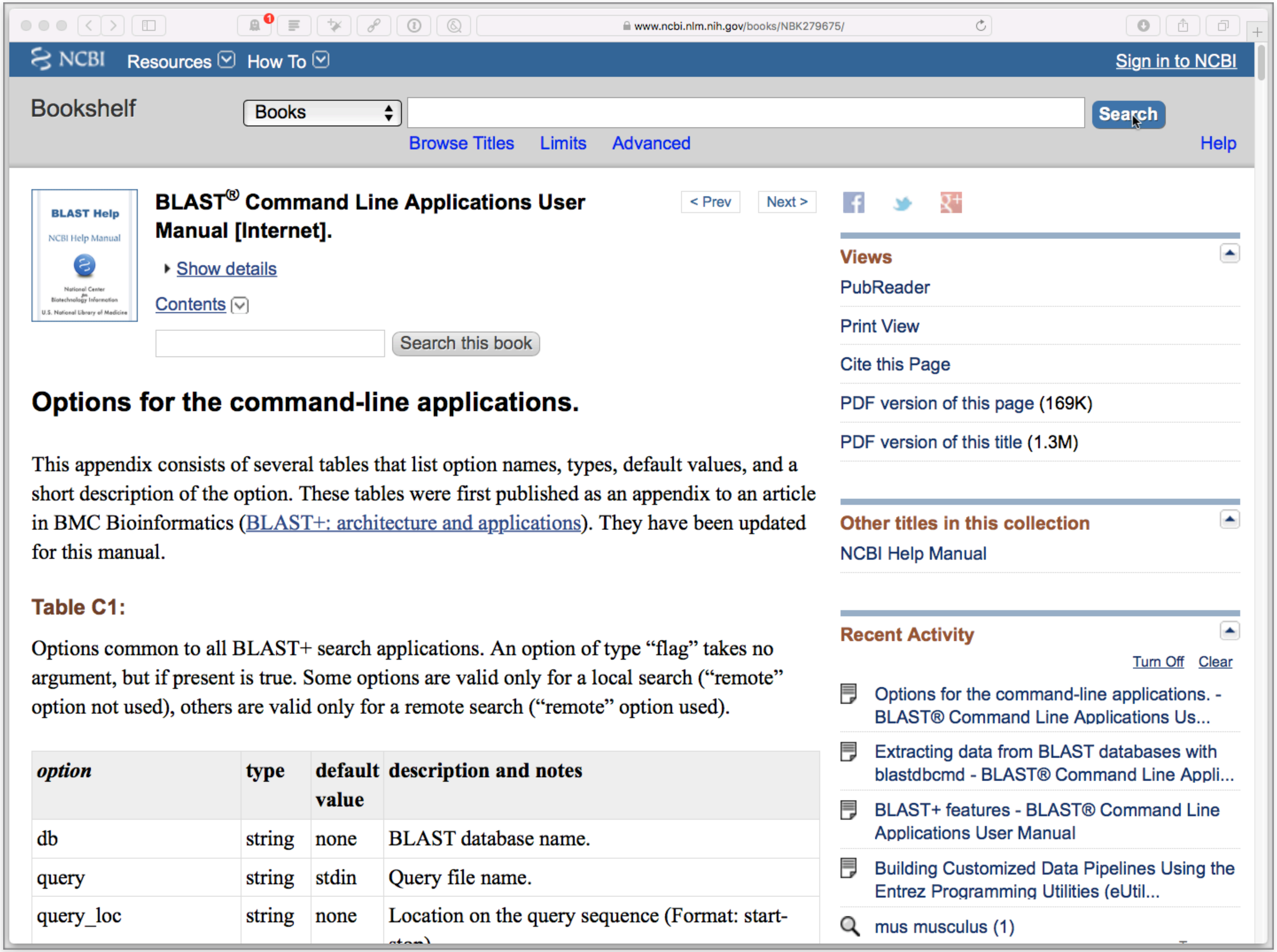
```
blastn -task megablast -db db/refseq_rna

-query test_query.txt -dust no -max_target_seqs 1
```

formatting options for hits

```
-outfmt "6 qseqid sseqid evalue pident stitle"
```

www.ncbi.nlm.nih.gov/books/NBK279675



BLAST® Command Line Applications User Manual [Internet].

Options for the command-line applications.

This appendix consists of several tables that list option names, types, default values, and a short description of the option. These tables were first published as an appendix to an article in BMC Bioinformatics (BLAST+: architecture and applications). They have been updated for this manual.

Table C1:

Options common to all BLAST+ search applications. An option of type "flag" takes no argument, but if present is true. Some options are valid only for a local search ("remote" option not used), others are valid only for a remote search ("remote" option used).

option	type	default value	description and notes
db	string	none	BLAST database name.
query	string	stdin	Query file name.
query_loc	string	none	Location on the query sequence (Format: start-stop)

Views
PubReader
Print View
Cite this Page
PDF version of this page (169K)
PDF version of this title (1.3M)
Other titles in this collection
NCBI Help Manual
Recent Activity
Turn Off
Clear
Options for the command-line applications. - BLAST® Command Line Applications Us...
Extracting data from BLAST databases with blastdbcmd - BLAST® Command Line Appli...
BLAST+ features - BLAST® Command Line Applications User Manual
Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtil...
mus musculus (1)

```
blastn -task megablast -db db/refseq_rna

-query test_query.txt -dust no -max_target_seqs 1

-outfmt "6 qseqid sseqid evalue pident stitle”
```

name and desired location of the output file

```
-out outputfile.txt
```