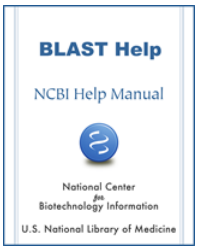
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Extracting data from BLAST databases with blastdbcmd.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Extracting data from BLAST databases with blastdbcmd

Extract lowercase masked FASTA from a BLAST database with masking information

If a BLAST database contains masking information, this can be extracted using the blastdbcmd options –db_mask and –mask_sequence as follows:

```
$ blastdbcmd -info -db mask-data-db
Database: Mask data test
        10 sequences; 12,609 total residues

Date: Feb 17, 2009  5:10 PM     Longest sequence: 1,694 residues

Available filtering algorithms applied to database sequences:

Algorithm ID  Algorithm name      Algorithm options
    20        seg                 default options used
    40        repeat              -species Desmodus_rotundus

Volumes:
        mask-data-db
$ blastdbcmd -db mask-data-db -mask_sequence_with 20 -entry 71022837
>gi|71022837|ref|XP_761648.1| hypothetical protein UM05501.1 [Ustilago maydis 521]
MPPSARHSAHPSHHPHAGGRDLHHAAGGPPPQGGPGMPPGPGNGPMHHPHSSYAQSMPPPPGLPPHAMNGINGPPPSTHG
GPPPRMVMADGPGGAGGPPPPPPPHIPRSSSAQSRIMEAaggpagpppagppastspavQklslANEaawvsIGsaaetm
EdydralsayeaalrhnpysvpalsaiagvhrtldnfekavdyfqrvlnivpengdTWGSMGHCYLMMDDLQRAYTAYQQ
ALYHLPNPKEPKLWYGIGILYDRYGSLEHAEEAFASVVRMDPNYEKANEIYFRLGIIYKQQNKFPASLECFRYILDNPPR
PLTEIDIWFQIGHVYEQQKEFNAAKEAYERVLAENPNHAKVLQQLGWLYHLSNAGFNNQERAIQFLTKSLESDPNDAQSW
YLLGRAYMAGQNYNKAYEAYQQAVYRDGKNPTFWCSIGVLYYQINQYRDALDAYSRAIRLNPYISEVWFDLGSLYEACNN
QISDAIHAYERAADLDPDNPQIQQRLQLLRNAEAKGGELPEAPVPQDVHPTAYANNNGMAPGPPTQIGGGPGPSYPPPLV
GPQLAGNGGGRGDLSDRDLPGPGHLGSSHSPPPFRGPPGTDDRGARGPPHGALAPMVGGPGGPEPLGRGGFSHSRGPSPG
PPRMDPYGRRLGSPPRRSPPPPLRSDVHDGHGAPPHVHGQGHGQGHGQGHGQGHGQGHGQSHGHSHGGEFRGPPPLAAAG
PGGPPPPLDHYGRPMGGPMSERERERMEWERERERERERERERQAARGYPASGRITPKNEPGYARSQHGGSNAPSPAFGRPPVY
GRDEGRDYYNNSHPGSGPGGPRGGYERGPGAPHAPAPGMRHDERGPPPAPFEHERGPPPPHQAGDLRYDSYSDGRDGPFR
GPPPGLGRPTPDWERTRAGEYGPPSLHDGAEGRNAGGSASKSRRGPKAKDELEAAPAPPSPVPSSAGKKGKTTSSRAGSP
WSAKGGVAAPGKNGKASTPFGTGVGAPVAAAGVGGGVGSKKGAAISLRPQEDQPDSRPGSPQSRRDASPASSDGSNEPLA
ARAPSSRMVDEDYDEGAADALMGLAGAASASSASVATAAPAPVSPVATSDRASSAEKRAESSLGKRPYAEEERAVDEPED
SYKRAKSGSAAEIEADATSGGRLNGVSVSAKPEATAAEGTEQPKETRTETPPLAVAQATSPEAINGKAESESAVQPMDVD
GREPSKAPSESATAMKDSPSTANPVVAAKASEPSPTAAPPATSMATSEAQPAKADSCEKNNNDEDEREEEEGQIHEDPID
APAKRADEDGAK
$
```

Custom data extraction and formatting from a BLAST database

The following examples show how to extract selected information from a BLAST database and how to format it:

```
Extract the accession, sequence length,
and masked locations for GI 71022837:
$ blastdbcmd -entry 71022837 -db Test/mask-data-db  -outfmt "%a %l %m"
XP_761648.1 1292 119-139;140-144;147-152;154-160;161-216;
```

Extract different sequence ranges from the BLAST databases

The command below will extract two different sequences: bases 40-80 in human chromosome Y (GI 13626247) with the masked regions in lowercase characters (notice argument 30, the masking algorithm ID which is available in this BLAST database) and bases 1-10 in the minus strand of human chromosome 20 (GI 14772189).

```
$ printf "%s %s %s %s\n%s %s %s\n" 13626247 40-80 plus 30 14772189 1-10 minus \
| blastdbcmd -db GPIPE/9606/current/all_contig -entry_batch -
>gi|13626247|ref|NT_025975.2|:40-80 Homo sapiens chromosome Y genomic contig, GRCh37.p10
Primary Assembly
tgcattccattctattctcttctACTGCATACAatttcact
>gi|14772189|ref|NT_025215.4|:c10-1 Homo sapiens chromosome 20 genomic contig, GRCh37.p10
Primary Assembly
GCTCTAGATC
$
```

Display the locations where BLAST will search for BLAST databases

This is accomplished by using the -show_blastdb_search_path option in blastdbcmd:

```
$ blastdbcmd -show_blastdb_search_path
:/net/nabl000/vol/blast/db/blast1:/net/nabl000/vol/blast/db/blast2:
$
```

Display the available BLAST databases at a given directory

This is accomplished by using the -list option in blastdbcmd:

```
$ blastdbcmd -list repeat  -recursive
repeat/repeat_3055 Nucleotide
repeat/repeat_31032 Nucleotide
repeat/repeat_35128 Nucleotide
repeat/repeat_3702 Nucleotide
repeat/repeat_40674 Nucleotide
repeat/repeat_4530 Nucleotide
repeat/repeat_4751 Nucleotide
repeat/repeat_6238 Nucleotide
repeat/repeat_6239 Nucleotide
repeat/repeat_7165 Nucleotide
repeat/repeat_7227 Nucleotide
repeat/repeat_7719 Nucleotide
repeat/repeat_7955 Nucleotide
repeat/repeat_9606 Nucleotide
repeat/repeat_9989 Nucleotide
$
```

The first column of the default output is the file name of the BLAST database (usually provided as the –db argument to other BLAST+ applications), the second column represents the molecule type of the BLAST database. This output is configurable via the list_outfmt command line option.