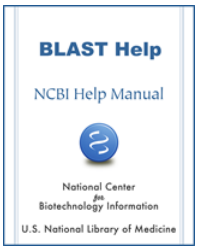
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Search with database masking enabled. 2008 Jun 23 [Updated 2021 Jan 7].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Search with database masking enabled

Created: June 23, 2008; Updated: January 7, 2021.

Database masking has two modes. The first is known as "soft-masking", and BLAST uses the database mask only during the (initial) word-finding phase of BLAST. The second is known as "hard-masking", and BLAST uses the database mask during all phases of the search. Here, we look at both types of masking.

To enable database masking during a BLAST search, we use the –info parameter of blastdbcmd to discover the masking Algorithm ID. For the database generated in the previous cookbook entry, we can use the following command line to activate the windowmasker soft masking:

```
$ blastn -query HTT_gene -task megablast -db hs_chr -db_soft_mask 30 \
  -outfmt 7 -out HTT_megablast_softmask.out -num_threads 4
```

Here, we search a nucleotide query, HTT_gene* (-query HTT_gene), with the megablast algorithm (-task megablast) against the database hs_chr (-db hs_chr). We use soft masking (-db_soft_mask 30), set the result format to tabular output (-outfmt 7), and save the result to a file named HTT_megablast_softmask.tab (-out HTT_megablast_softmask.tab). We also activated the multi-threaded feature of blastn to speed up the search by using 4 CPUs[$] (-num_threads 4).

For the database generated in the previous cookbook entry, we can use the following command line to activate the windowmasker hard masking:

```
$ blastn -query HTT_gene -task megablast -db hs_chr -db_hard_mask 30 \
  -outfmt 7 -out HTT_megablast_hardmask.out -num_threads 4
```

The options are similar to the ones for soft masking, except that we use –db_hard_mask rather than –db_soft_mask. Additionally, we changed the name of the output file.

Hard masking is much more aggressive than soft masking. In interspersed or simple repeats, soft masking normally provides the best results. Hard masking may be warranted to remove vector or other contamination from the BLAST results.

*This is a genomic fragment containing the HTT gene from human, including 5 kb up- and down-stream of the transcribed region. It is represented by NG_009378.

[$] The number to use under in your run will depend on the number of CPUs your system has.

In a test run under a 64-bits Linux machine, the search with soft masking took about 1.5 seconds real time, and the search with hard masking took about 2.5 seconds real time. The search without database masking took about 31 minutes.