Limiting a Search with a List of Identifiers

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

BLAST can also limit a database search by a list of identifiers (e.g.: accessions), which should be specified one per line in a file. These identifiers, referencing the sequences to include or exclude in the BLAST search, should not contain any whitespace and should be retrievable from the BLAST database.

Starting with BLASTDB version 5, an accession list must be pre-processed before it can be used in a search. This process checks that the accessions appear to be real and produces a file optimized for use with BLAST. It is also possible to confirm that all the accessions are actually in the target database. The examples below demonstrate this functionality:

```
# 9606.pacc is a text file with protein accessions. This command produces a file called
9606.pacc.bsl
$ blastdb_aliastool -seqid_file_in 9606.pacc

# This command searches nr limited to the accessions in the file 9606.pacc.bsl
$ blastp –db nr –query QUERY.fsa -outfmt "7 std taxid" -seqidlist 9606.pacc.bsl
```

Additionally, one may use the -negative_seqidlist option to exclude sequences by accession from the BLAST search.

When the search is limited by a list of IDs the statistics of the BLAST database are re-calculated to reflect the actual number of sequences and residues/bases included in the search.