

A short guide to running PepNovo+ with MS-Blast

Ari M. Frank

November 17, 2010

This document describes the various steps that are required in order to identify proteins from MS/MS data using PepNovo+ and BLAST. If you run into problems or find error or omissions, please contact the author at arf@cs.ucsd.edu.

1 Installation

1. Install AB-Blast on your computer or gain access to a copy where you can run queries through the command line. See: <http://blast.advbiocomp.com/licensing>
2. After installing AB-BLAST, you need to edit the script `pnv_msblast.pl` and set the variable `$BLAST_DIR` to the path of the main directory of BLAST (this directory has the `blastp` binary in it). For example:
`$BLAST_DIR = '/home/bioinfo/tools/BLAST';`
3. Install PepNovo+ (see separate help file). Package is available at UCSD website:
<http://proteomics.ucsd.edu/>.
4. Make sure your operating system has PERL installed.

2 Creating a query for BLAST

2.1 De novo sequencing

The BLAST query is generated from de novo sequencing results. These results can be obtained by running PepNovo+ on each of the files with the additional flag “-msb_generate_query”. You must also supply a query name with “-msb_query_name”. For example:

```
PepNovo_bin -model CID_IT_TRYP -PTMs C+57:M+16 -file my_file.mzXML -msb_generate_query  
-msb_query_name output
```

Will run de novo sequencing on the file “my_file.mzXML” while considering a fixed modification on cystine and oxidation of methionine (see PepNovo help file for more options). Note that is is not advised to consider many PTMs (post-translational modifications) for de novo sequencing since this will dramatically reduce the quality of the results. Running the above command creates 3 output files:

1. `my_file_dnv.txt` - The regular de novo results for all the spectra.
2. `my_file_full.txt` - The full set of MS-BLAST sequences generated for all spectra.
3. `my_file_query.txt` - A formatted query made from the top scoring sequences.

There are a few other (optional) parameters that are specific to MS-BLAST (though usually you will not need to use them):

- `-msb_query_size` - The maximal number of amino acids in the query file (the default is 1000000).
- `-msb_num_solutions` - The number of peptides to generate from each spectrum (the default is 7).
- `-msb_min_score` - The minimal de novo score needed for the spectrum to be used. This is equivalent to the number of expected correct amino acids (the default is 3).

Note that for large experiments (involving tens of thousands and even millions of spectra, you will probably need to run several de novo jobs in parallel (e.g., using a grid) in order to save time (typically it takes 2 seconds to sequence each spectrum) . In such cases, you will need to create a file with the full paths to all the “my_file_full.txt” files (one per line). Let’s call this file “full_list.txt”. You then need to run PepNovo again in order to merge all these files (and possibly split them again into several queries). For example:

```
PepNovo_bin -model CID_IT_TRYP -PTMs C+57:M+16 -list full_list.txt -msb_merge_queries
-msb_query_name my_organism
```

This will create several query files (depending on the size of the input and value of “-msb_query_size”): `my_organism_pt_0.query`, `my_organism_pt_1.query`, ..., `my_organism_pt_n.query`.

2.2 Create a sequence database for BLAST

Before you can create a database for the blast query, you should merge the fasta files and add a decoy (it is good to also add sequences of common contaminants). This can be done with the provided script `create_fasta_with_decoy.pl`. For example the command:

```
perl create_fasta_with_decoy.pl fish.fa frogs.fa contaminants.fa > fasta_for_blast.fa
```

will create a single fasta file with the sequences from `fish.fa`, `frogs.fa`, and `contaminants.fa`. In addition, a shuffled version of each protein sequence will also be added as a decoy. The protein names in the decoy sequences all have the form “XXX_” and then an index number.

The next step is to convert the fasta file into a blast compatible database file. For example, the command (`xdformat` is located in the BLAST directory):

```
xdformat -p fasta_for_blast.fa -o blast_db
```

will create 3 files: `blast_db.xpt`, `blast_db.xps`, and `blast_db.xpd`. These three files makeup the formatted database the needs to be supplied to BLAST.

3 Running BLAST

To run BLAST you need to use the script `pnv_msblast.pl`. This script runs BLAST with a configuration that is suitable for aligning short peptides to large sequence databases. It also post-processes the results in order to provide false discovery rates, which are widely used in proteomics. There are several flags that can be used (run the script without arguments to get the full list). The command line for each job looks something like the following:

```
perl pnv_msblast.pl -Q queriesmy_organism_pt_0.query -O outmy_organism_pt_0
-D databasesorganism_db
```

Where:

- `-Q` - the path to the query file generated by PepNovo (either directly from running de novo or from merging outputs for several files).
- `-O` - the path and prefix name for the output files (there will be `.query` `.res` and `.sum` - see below)
- `-D` - the path and prefix name of the BLAST-formatted databases (this is the full path and file name without the ".xps" extension).

Note that if you have several split queries, you will need to run the script with each query files. Be sure use the same output prefix for all files (the paths and names should be the same up to the `pt_` part), since they will need to be integrated in a final run (see below).

There are additional parameters that can be supplied to the script (run script without parameters to get full list and default values):

- `-R` - maximum number of results for BLAST search (sets `-h`, `-V`, `-B`, to this value), default `-R=25000`.
- `-V` - `V` parameter for `blastp` (number of descriptions), default `V=25000`
- `-B` - `B` parameter for `blastp` (number of alignments), default `B=25000`
- `-S` - `S` parameter for `blastp`, default `S=34`
- `-E` - `E` (Expect) parameter, default `E` is 500000. **This parameter might need to be tuned depending on the number of hits to the decoy database (see discussion below).**
- `-m` - matrix name (from options installed by BLAST in `$BLAST_DIR/matrix` (default `blosum62`))
- `-f` - filter for BLAST runs, default=`seg`, (typical options: `seg`, `xnu`, `seg+xnu`, etc.)
- `-h` - `hspmax` for BLAST, default `h=5000`
- `-C` - decoy protein name prefix (default created by `create_fasta_with_decoy.pl` is `XXX`)
- `-F` - FDR level for protein ids (default 0.05)
- `-a` - additional BLAST parameter, use format `NAME1=value1,NAME2=values2,...` (see BLAST manual for full list)

The perl script `pnv_msblast.pl` launches a BLAST job. It creates three files (assuming the `-O` parameter is out):

1. `out.query` - the query sequence for BLAST
2. `out.res` - the results file from BLAST (including all protein ids and alignments between de novo sequences and protein sequences).
3. `out.sum` - a summary file that lists the protein and peptide ids, along with the false discovery rates.

3.1 BLAST warnings and adjusting the `-E` parameter

When BLAST runs it might emit certain warnings (to be found in the `.res` file). One warning is:

WARNING: Use of the `hspsepQmax` parameter should be considered with long query sequences, to improve the biological relevance of the HSP groups that are assembled and to improve the statistical discrimination of these groups from random background.

This warning can be ignored since in MS-BLAST there is no importance to the locality of the query sequences (since each is a de novo result).

Also BLAST might complain that some matches were not shown due to low values of V and B. If the results show a sufficient number of hits to the decoy database (i.e., you see hits to proteins with the name XXX_nnn and there is no warning on the top of the .sum file), then you can ignore this warning. However, if you only have hits to good proteins, then you should re-run the script and supply higher values for -B, -V and -h (e.g., 10000 instead of 5000). This will retain more good hits in the results file.

Finally, pnv_blast.pl might issue a warning as follows:

```
### WARNING: too few hits were made to decoy database! It is suggested rerun BLAST  
with increased E-value (-E X,  $X \geq 30000$ ) ###
```

This warning is issued if there are too few hits to the decoy, which means that many good hits are also not included in the results file. In such cases you need to supply a higher value for the -E parameter (e.g., double or triple the current value) until the warning does not appear.

3.2 Final round of BLAST

For large jobs in which several query files were created (pt_0.txt,...,pt_n.txt) you need to run one final BLAST run using all the results files in order to integrate the results and create a single file with all identified proteins and peptides. This can be done by using the -L flag (instead of the -Q). The -O flag should point to the prefix of the paths of all the output files from the previous rounds. This run will create files with the suffixes `.final_query`, `.final_res`, and `.final_sum` which will hold the combined results.

3.3 Acknowledgements

The author would like to thank Andrej Shevchenko, Shamil Sunyaev, and Ivan Adzhubey for their helpful discussions regarding running MS-Blast and for providing the original MS-Blast scripts.