**The Manual of ProtExcluder**

**Application**

ProtExcluder was designed to exclude the portion of DNA sequence (as well as certain length of flanking sequence – given by the user, default = 50 bp) matching subjects in a protein database.

**Input files**

Two input files are required by ProtExcluder. The first one is blastx output which contains sequence alignment between the query and the subject. The second one is the fasta file of the DNA sequences used in blastx search.

The package was developed using blastx output from ncbi-blast-2.2.28+.

**Steps in ProtExlcuder**

The Blastx output is first filtered by a low complexity filter which excludes alignments where 3 individual amino acids account for 60% or more of the matches between the query and the subject. This is because even if the “low complexity” filter for BLAST is on, there are still alignments apparently arisen due to low complexity sequences.

Subsequently, the regions matching proteins in the query sequences are examined. If two regions are within certain distance (given by user, default = 50 bp), the two regions will considered as a single piece. All the pieces match proteins as well as their flanking sequences (again given by users, default = 50 bp) will be excluded. After exclusion if a single sequence is less than 50 bp in length (excluding NNN), the entire entry is removed.

The resulting sequences and sequences without matching proteins are combined together and put in “XXXnoProtFinal”

**Other software required by ProtExlcuder**

Hmmer-3.0 (http://hmmer.janelia.org/)

**Installation**

% tar zxf ProtExcluder1.0.tar.gz

% cd ProtExcluder1.0

% ./Installer.pl -m HmmerPath -p ProtExcluderPath

-m the full path to hmmer folder

-p the full path to ProtExcluder folder

For example “./Installer.pl -m /home/jiang/bin/hmmer-3.0-linux-intel-x86\_64/ -p /home/jiang/bin/ProtExcluder1.0/”

Please note the path must end with “/”.

**Operation**

% ProtExcluderPath/ProtExcluder.pl -option blastxoutput fastasequencefile

Blastxoutput is the blastx output file generated by the relevant DNA sequences searching against the protein database

Fastasequencefile is the fast file containing the relevant DNA sequences

For example “/home/jiang/bin/ProtExcluder1.0/ProtExcluder.pl -f 100 AT\_RM.fa.blastx AT\_RM.fa”

 “-f 100” indicates excluding the 100 bp flanking the matched region on each side. This is optional. The default value is 50 bp. If you don’t want any flanking excluded, please use “-f 0”.

**Acknowledgement**

I thank Yujun Han and Zhirong Bao. The development of this package was inspired by their work.