Tools for Maintenance and Preparation of FASTA Protein Databases

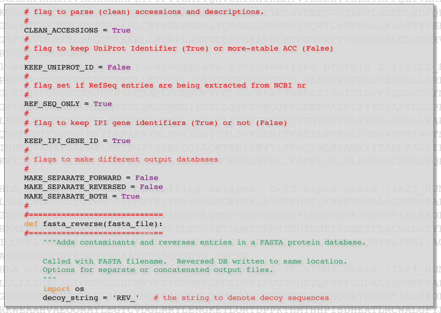
Phillip A. Wilmarth and Larry L. David
Department of Biochemistry and Molecular Biology, Oregon Health & Science University, Portland OR

Background:

- Explosive growth in number of sequenced genomes.
- Keeping protein databases up-to-date is challenging.
- Few tools available to simplify preparing FASTA databases for proteomics analyses.

Overview:

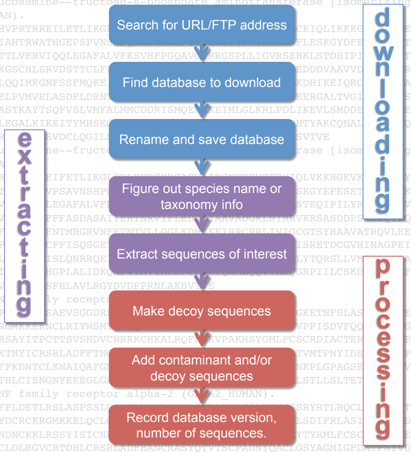
- Utility programs written in Python v2.6 to support NCBI, UniProt, and IPI databases.
- Standard libraries support FTP file transfers and compressed files.
- Release notes, version numbers, and taxonomy information downloaded.
- File/folder naming scheme keeps downloads organized.
- Taxonomy numbers used to extract species-specific protein databases.
- Creation of decoy sequences and adding contaminant sequences supported.



```
# flag to parse (clean) accessions and descriptions.
#
CLEAN_ACCESSIONS = True
#
# flag to keep UniProt Identifier (True) or more-stable ACC (False)
#
KEEP_UNIPROT_ID = False
#
# flag set if RefSeq entries are being extracted from NCBI nr
#
REF_SEQ_ONLY = True
#
# flag to keep IPI gene identifiers (True) or not (False)
#
KEEP_IPI_GENE_ID = True
#
# flags to make different output databases
#
MAKE_SEPARATE_FORWARD = False
MAKE_SEPARATE_REVERSED = False
MAKE_SEPARATE_BOTH = True
#
#==============================
def fasta_reverse(fasta_file):
#==============================
    """Adds contaminants and reverses entries in a FASTA protein database.

    Called with FASTA filename.  Reversed DB written to same location.
    Options for separate or concatenated output files.
    """
    import os
    decoy_string = 'REV_'   # the string to denote decoy sequences
```

Readable source code. Python source code is extensively commented with any program control flags located at the top of text files.



Search for URL/FTP address
Find database to download
Rename and save database
Figure out species name or taxonomy info
Extract sequences of interest
Make decoy sequences
Add contaminant and/or decoy sequences
Record database version, number of sequences.
downloading
extracting
processing

Steps involved in getting protein databases. Tasks fall into three categories: downloading FASTA-formatted database files, unpacking and extracting sequences of interest (usually different species), and processing of sequences for analyses.

Function	Programs
Downloading	nr_get_analyze.py, sprot_get_analyze.py, uniprot_get_analyze.py, ipi_get_all.py
Extracting	nr_extract_taxon.py, uniprot_extract_from_one.py, uniprot_extract_from_both.py, extract_by_string.py
Processing	add_extras_and_reverse.py, reverse_fasta.py, remove_duplicates.py
Other	count_fasta.py, check_for_duplicates.py, taxon_group_analyzer.py

Utilities are listed above by function. Programs contain FTP addresses and use file/folder naming to organize databases. Files remain compressed. Species analysis files and taxonomy files allow extraction by taxonomy numbers. Taxonomy nodes can be expanded (e.g. "rodents"). Reversed databases can be created. Other tools provide diagnostic information.

Extraction criteria or download source	Number of proteins
"Homo sapiens" or "Human" (case insensitive)	553,645
"Homo sapiens" (case sensitive)	217,201
"Homo sapiens" then "|ref|" (case sensitive)	31,880
Taxon=9606, all sequences	231,498
Taxon=9606, RefSeq only	31,855
Taxon=9606 downloaded from NCBI	521,247 (266,432 duplicates)
Taxon=9606 RefSeq downloaded from NCBI	38,789 (6,934 duplicates)

Many ways to get human protein sequences from NCBI. The number of sequences varies more than 17-fold. The RefSeq project (www.ncbi.nlm.nih.gov/RefSeq/) makes a dramatic difference. Ironically, taxon=9606 sequences downloaded from NCBI (last two rows) were redundant databases.

Taxon #	Species Name	Sequences	RefSeqs
4932	Saccharomyces cerevisiae	11900	5822
545124	Saccharomyces cerevisiae AWRI1631	5466	
643680	Saccharomyces cerevisiae EC1118	5989	
574961	Saccharomyces cerevisiae JAY291	5198	
285006	Saccharomyces cerevisiae RM11-1a	5374	
559292	Saccharomyces cerevisiae S288c	5	
307796	Saccharomyces cerevisiae YJM789	5901	
41870	Saccharomyces cerevisiae var. diastaticus	8	
11008	Saccharomyces cerevisiae virus L-A (L1)	14	3

Taxonomy node complications. Many organisms, such as yeast, have genomes of strains. The RefSeq counts clearly show the correct yeast taxonomy number. A string extraction using "Saccharomyces cerevisiae" would be a poor choice.

Summary:

- Genomic sequence information is changing rapidly.
- Tools for FASTA protein databases have not kept pace.
- Comprehensive suite of Python utilities developed.
- Freely available for non-commercial use at http://www.ProteomicAnalysisWorkbench.com.

Supported by NIH grants EY007755 and EY010572.
Author contact: wilmarth@ohsu.edu