enTAP Documentation

Release 0.5.0

Alex Hart, Jill Wegrzyn

Nov 20, 2017

Contents

CHAPTER 1

EnTAP Introduction

The Eukaryotic Non-Model Transcriptome Annotation Pipeline (*EnTAP*) is designed to improve the accuracy, speed, and flexibility of functional gene annotation for de novo assembled transcriptomes in non-model eukaryotes.

This software package addresses the fragmentation and related assembly issues that result in inflated transcript estimates and poor annotation rates. Following filters applied through assessment of true expression and frame selection, open-source tools are leveraged to functionally annotate the translated proteins.

Downstream features include fast similarity search across multiple databases, protein domain assignment, orthologous gene family assessment, Gene Ontology term assignment, and KEGG pathway annotation.

The final annotation integrates across multiple databases and selects an optimal assignment from a combination of weighted metrics describing similarity search score, taxonomic relationship, and informativeness. Researchers have the option to include additional filters to identify and remove potential contaminants and prepare the transcripts for enrichment analysis. This fully featured pipeline is easy to install, configure, and runs much faster than comparable functional annotation packages. It is developed to contend with many of the issues in existing software solutions.

EnTAP is optimized to generate extensive functional information for the gene space of organisms with limited or poorly characterized genomic resources.

1.1 Pipeline Stages:

- Transcriptome Filtering: designed to remove assembly artifacts and identify true CDS (complete and partial genes)

1. Expression Filtering (RSEM)
2. Frame Selection (GeneMARKS-T)

- Annotation

3. Similarity Search: optimized search against user-selected databases (DIAMOND).
4. Contaminant Filtering and Best Hit Selection: selects final annotation and identifies potential contaminants

5. Orthologous Group Assignment: independent assignment of translated protein sequences to gene families (eggNOG). Includes protein domains (SMART/Pfam), Gene Ontology (GO) terms, and KEGG pathway assignment.

All of the software integrated into this pipeline are packaged within the EnTAP repository with the exception of GeneMarkS-T. Installation and usage of EnTAP is documented in this guide.

1.2 Citations:

[1] B. Buchfink, Xie C., D. Huson, "Fast and sensitive protein alignment using DIAMOND", Nature Methods 12, 59-60 (2015).

[2] eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Nucl. Acids Res. (04 January 2016) 44 (D1): D286-D293. doi: 10.1093/nar/gkv1248

[3] Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Jaime Huerta-Cepas, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering and Peer Bork. Submitted (2016).

[4] Li, B., Dewey, C., & Liu, P. RSEM. In.

1.2.1 Software contained or used within this pipeline:

- RSEM
- DIAMOND
- EggNOG
- GeneMarkS-T

CHAPTER 2

Installation

EnTAP is packaged with all of the software necessary to fully annotate a set of transcripts. It is optimized to allow a single-command execution for all steps in the pathway, including paramterization by the user. EnTAP does not have a graphical user interface but it does generate visual summaries for the user at each stage as well as detailed summary files and logs.

1. *System Requirements*
2. *Dependency Check*
3. *Pipeline Software*
4. *EnTAP*

Before full EnTAP installation, dependencies must be checked to see if they are included in your system (many are by default) and the accompanying pipeline software will need to be installed (unless is already present on the system).

2.1 Dependency Check

Before continuing on in the installation process, ensure that the following dependencies are fully installed on your system:

- C++11 compiler (GCC 4.8.1 or later)
- CMake
- **Boost C++ Libraries (1.50 or later)**
  - Ensure this is installed with the C++11 compiler
- Perl
- **Python with support for the following modules**
  - SQLITE
  - Matplotlib (figures generated by EnTAP)

- Unix wget (generally included in most distros)
- Unix gzip (generally included in most distros)

2.2 Pipeline Software

EnTAP leverages several software distributions within the pipeline to provide the best quality annotations. Not all of the software is required, but it is suggested. However, similarity searching and orthologous group assignment should not be skipped.

Note: If the software is already installed on your system, this stage can be skipped

Software:

- RSEM (Expression Filtering with alignment file)
- GeneMarkS-T (Frame Selection)
- DIAMOND (Similarity Search)
- EggNOG-Emapper (Orthologous Group Assignment)

If you have downloaded the full repository from the GitLab page, each of these (with the exception of GeneMarkS-T) are contained within the /libs directory. GeneMarkS-T must be acquired from the website linked previously due to licensing (free for academic use).

RSEM and DIAMOND both require compilation from source code while EggNOG-Emapper does not. To compile these, run the script within the main directory:

```
./setup.sh
```

This will make+install DIAMOND and make RSEM.

Warning: Ensure that DIAMOND was properly installed (global installation required by EggNOG-Emapper)

If there are any problems with the setup script, installation steps can be found on the GitHub pages for each.

2.3 EnTAP Installation

Once dependencies and pipeline software have been installed, you can now continue to install EnTAP! Within the main directory, execute the following command:

```
cmake
```

This will generate a MakeFile. Then execute:

```
make
```

Or to install:

```
make install
```

This will complete the installation process. You are ready to start using EnTAP!

CHAPTER 3

Basic Usage

EnTAP has two stages of execution, *configuration* and *run*. Configuration should be completed before the first run and everytime any of the source databases have been updated by the user. This should also be run if you would like to include the latest version of the NCBI Taxonomy database and the Gene Ontology database. All of these are updated regularly at the source and you can ensure you have the most recent version by running configuration before your annotation runs.

3.1 Configuration

Configuration is the first stage of EnTAP that will download and configure the necessary databases for full functionality. This is run if you would like to change/update the databases that EnTAP is reading from. I'll break this up into two sections, *folder hierarchy* and *usage*. The folder hierarchy section will just describe how everything is 'typically' setup with EnTAP, however these paths can be easily changed in the entap_config.txt (more on that later!). The usage section will go over the basic usage during the Configuration stage of EnTAP.

3.1.1 Folder Hierarchy

The EnTAP folder organization is referred to as the execution directory where all files will be made available. This is essentially the hierarchy that was downloaded from the repository.

The /EnTAP directory contains:

- /EnTAP /libs
- /EnTAP /src
- /EnTAP /bin (created during configuration)
- /EnTAP /databases (created during configuration)

In addition to some other files/directories.

Recognition of EnTAP databases, src files, and execution accompanying pipeline software can rely on this 'default' directory hierarchy. However, any necessary files/directories can be changed from the default with the entap_config.txt file (by specifying the paths flag).

Warning: Ensure you are pointing to the correct paths if not using the defaults!

The entap_config.txt file mentioned above has the following defaults:

- diamond_exe_path=/EnTAP/libs/diamond-0.8.31/bin/diamond
- rsem_exe_path=/EnTAP/libs/RSEM-1.3.0 (this is a path to the directory)
- genemarkst_exe_path=/EnTAP//libs/gmst_linux_64/gmst.pl
- eggnog_exe_path=/EnTAP/libs/eggnog-mapper/emapper.py
- eggnog_download_exe=/EnTAP/libs/eggnog-mapper/download_eggnog_data.py
- eggnog_database=/EnTAP/libs/eggnog-mapper/data/eggnog.db (downloaded during Configuration)
- entap_tax_database=/EnTAP/bin/ncbi_tax_bin.entp (binary version, downloaded during Configuration)
- entap_tax_download_script=/EnTAP/src/download_tax.pl
- entap_go_database=/EnTAP/bin/go_term.entp (binary version, downloaded during Configuration)
- entap_graphing_script=/EnTAP/src/entap_graphing.py

These can be changed to whichever path you would prefer. If something is globally installed, just put a space " " after the '='. EnTAP will recognize these paths first and they will override defaults.

This configuration file will be automatically detected if it is in the same directory as the EnTAP .exe, otherwise the path to it can be specified through the paths flag.

Note: Be sure you set the paths before moving on (besides the databases that haven't been downloaded yet)!

3.1.2 Usage

All source databases must be provided in FASTA format so that they can be indexed for use by DIAMOND. This can be completed independent of EnTAP with DIAMOND or as part of the configuration phase of EnTAP. While any FASTA database can be used, it is recommended to use NCBI (Genbank) sourced databases such as RefSeq databases or NR. In addition, EnTAP can easily accept EBI databases such as UniProt/SwissProt. EnTAP can read the species information from these header formats. If the individual FASTAs in a custom database do not adhere to one of these two formats, it will just not be possible to weight examine taxanomic or contaminant status from them.

The following FTP sites contain common reference databases that enTAP can recognize:

- RefSeq:
- Arthropod RefSeq:
- Plant RefSeq:
- Mammalian RefSeq:
- NR:
- SwissProt:
- UniProt:

It is generally recommended that a user select at least three databases with varying levels of NCBI curation. Unless the species is very non-model (i.e. does not have close relatives in databases such as RefSeq, it is not necessary to use the full NR database which is less curated).

To run configuration with a sample database, the command is as follows:

```
EnTAP --config -d path/to/database
```

This stage must be done at least once prior to *running*. Once the database is configured, you need not do it again unless you updated your original database or plan on configuring several others.

Note: If you already have DIAMOND (.dmnd) configured databases, you can skip the configuration of that database. Although, due to other EnTAP database downloading (taxonomy and ontology), configuration must still be ran at least once without any flags.

Configuration can be ran without formatting a database as follows:

```
EnTAP --config
```

Note: This is the only stage that requires connection to the Internet.

3.1.3 Flags:

Required Flags:

- **(- - config)**
  - The only required flag.
  - Although in order to run the full EnTAP pipeline, you must have a .dmnd configured database.

Optional Flags:

- **(-d/ - - database)**
  - Specify any number of FASTA formatted databases you would like to configure for EnTAP
  - Not necessary if you already have DIAMOND configured databases (.dmnd)
- **(- - paths)**
  - Point to entap_config.txt for specifying paths
- **(- - database-out)**
  - Specify an output directory for the databases to be sent to
  - This will send the Taxonomic Database, GO Database, and any DIAMOND databases to this location
  - EggNOG database will not be sent here as it must remain in the EggNOG directory
- **(- t/ - - threads)**
  - Specify thread number for Configuration

3.1.4 Memory Usage:

Memory usage will vary depending on the number of databases you would like configured. Although, EnTAP will download several other databases as well:

- Gene Ontology References: 6Mb
- NCBI Taxonomy: 400Mb
- EggNOG Database: 30Gb

3.2 Run

The run stage of *EnTAP* is the main annotation pipeline. After configuration is ran at least once, this can be ran continually without requiring configuration to be ran again (unless more databases will be configured).

3.2.1 Input Files:

Required:

- .FASTA formatted transcriptome file (either protein or nucleotide)
- .dmnd (DIAMOND) indexed databases, which can be formatted in the :ref:‘configuration<config-label>‘stage.

Optional:

- .BAM/.SAM alignment file. If left unspecified expression filtering will not be performed.

3.2.2 Sample Run:

A specific run flag (**runP/runN**) must be used:

- runP: Indicates protein input transcripts. Selection of this option will skip the frame selection portion of the pipeline.
- runN: Indicates nucleotide input transcripts. Selection of this option will cause frame selection to be ran.

An example run with a nucleotide transcriptome:

```
EnTAP --runN -i path/to/transcriptome.fasta -d path/to/database.dmnd -d path/to/
↪database2.dmnd -a path/to/alignment.sam
```

With the above command, the entire EnTAP pipeline will run. Both frame selection and expression filtering can be skipped if preferred by the user. EnTAP would require protein sequences (indicated by –runP) in order to avoid frame selection. If there is not a short read alignment file provided in SAM/BAM format, then expression filtering via RSEM will be skipped.

3.2.3 Flags:

Required Flags:

- **(- -runP/- -runN)**

  - Specification of input transcriptome file. runP for protein (skip frame selection) or runN for nucleotide (frame selection will be ran)

- **(-i/- -input)**
  - Path to the transcriptome file (either nucleotide or protein)
- **(-d/- -database)**
  - Specify up to 4 DIAMOND indexed (.dmnd) databases to run similarity search against

Optional Flags:

- **(-a/- -align)**
  - Path to alignment file (either SAM or BAM format)
  - **Note:** Ignoring this flag will skip expression filtering
  - If you have ran alignment with paired end reads be sure to use the - -paired-end flag as well
- **(- - contam)**
  - Specify *contaminant* level of filtering
  - Multiple contaminants can be selected through repeated flags
- **(- - species)**
  - This flag will allow for *taxonomic* 'favoring' of hits that are closer to your target species or lineage. Any lineage can be used as referenced by the NCBI Taxonomic database, such as genus, phylum, or species.
  - Format **must** replace all spaces with underscores ('_') as follows: "- -species homo_sapiens" or "- -species primates"
- **(- - level)**
  - Specify Gene Ontology levels you would like to normalize to
  - Any amount of these flags can be used
- **(- - tag)**
  - Specify output folder labelling.
  - Default: /outfiles
- **(- - fpkm)**
  - Specify FPKM cutoff for expression filtering
  - Default: 0.5
- **(-e)**
  - Specify minimum E-value cutoff for similarity searching
  - Default: 10E-5
- **(- - tcoverage)**
  - Specify minimum target coverage for similarity searching
  - Default: 50%
- **(- - qcoverage)**
  - Specify minimum query coverage for similarity searching

  - Default: 50%

- **(- - overwrite)**
  - All previously ran files will be overwritten if the same - -tag flag is used
  - Without this flag EnTAP will *recognize* previous runs and skip things that were already ran
- **(- - paired-end)**
  - Signify your reads are paired end for RSEM execution
- **(- - graph)**
  - This will check whether or not your system has graphing functionality supported
  - If Python with the Matplotlib module are installed on your system graphing should be enabled!
  - This can be specified on its own
- **(-t/ - - threads)**
  - Specify the number of threads of execution
- **(- - state)**
  - Precise control over execution *stages*. This flag allows for certain parts to be ran while skipping others.
  - Warning: This may cause issues depending on what you plan on running!

3.2.4 Taxonomic Favoring and Contaminant Filtering

Taxonomic contaminant filtering (as well as taxonomic favoring) is based upon the NCBI Taxonomy database. In saying this, all species/genus/lineage names must be contained within this database in order for it to be recognized by EnTAP.

Contaminant Filtering:

Contaminants can be introduced during collection or processing of a sample. A contaminant is essentially a species that is not of the target species you are collecting. Some common contaminants are bacteria and fungi that can sometimes be found within collected samples. If a query sequence from your transcriptome is found when matching against a similarity search database, it will be flagged as such (but NOT removed automatically). Oftentimes, researchers would like to remove these sequences from the dataset.

An example of flagging bacteria and fungi as contaminants can be seen below:

```
EnTAP --runN -i path/to/transcriptome.fasta -d path/to/database.dmnd -c fungi -c␣
↪bacteria
```

Taxonomic Favoring

During best hit selection of similarity searched results, taxonomic consideration can utilized. If a certain lineage (such as sapiens) is specified, hits closer in taxonomic lineage to this selection will be chosen. Any lineage such as species/kingdom/phylum can be utilized as long as it is contained within the Taxonomic Database

3.2.5 Picking Up Where You Left Off

In order to save time and make it easier to do different analyses of data, EnTAP allows for picking up where you left off if certain stages were already ran and you'd like analyze data with different contaminant flags or taxonomic favoring. As an example, if similarity searching was ran previously you can skip hitting against the database and analyze the data to save time. However, the - - overwrite flag will not allow for this as it will remove previous runs and not recognize them.

In order to pick up and skip re-running certain stages again, the files that were ran previously **must** be in the same directories and have the same names. With an input transcriptome name of 'transcriptome' and example database of 'complete.protein':

- **Expression Filtering**
  - transcriptome.genes.results
- **Frame Selection**
  - transcriptome.fasta.faa
  - transcriptome.fasta.fnn
  - transcriptome.fasta.lst
- **Similarity Search**
  - blastp_transcriptome_complete.protein.faa.out
- **Gene Family**
  - annotation_results.emapper.annotations
  - annotation_results_no_hits.emapper.annotations

Since file naming is based on your input as well, the flags below **must** remain the same: * (-i / - - input)

- (-a / - - align)
- **(-d / - - database)**
  - Do not necessarily need to remain the same. If additional databases are added, EnTAP will recognize the new ones and run similarity searching on them
- (- - qcoverage)
- (- - tcoverage)

3.2.6 State Control

CHAPTER 4

Interpreting the Results

EnTAP provides many output files at each stage of execution to better see how the data is being managed throughout the pipeline:

1. *Expression Filtering*
2. *Frame Selection*
3. *Similarity Searching*
4. *Orthologous Groups/Ontology*
5. *Final Annotation Results*

4.1 Expression Filtering

The */expression* folder will contain all of the relevant information for this stage of the pipeline.

More documentation coming soon!

4.2 Frame Selection (GeneMarkS-T)

The */frame_selection* folder will contain all of the relevant information for the frame selection stage of the pipeline. This folder will contain the *main files* (results from frame selection software), files *processed* from EnTAP, and *figures* generated from EnTAP.

4.2.1 GeneMarkS-T Files: */frame_selection*

The files within the root */frame_selection* directory contain the results from the frame selection portion of the pipeline. More information can be found at GeneMarkS-T. With a generic transcriptome input of "Species.fasta", these files will have the following format:

- Species.fasta.fnn
  - Nucleotide fasta formatted frame selected sequences
- Species.fasta.faa
  - Amino acid fasta formatted frame selected sequences
- Species.fasta.lst
  - Information on each sequence (partial/internal/complete/ORF length)
- .err and .out file
  - These files are will contain any error or general information produced from the GeneMarkS-T run

4.2.2 EnTAP Files: */processed*

Files within the */processed* are generated by EnTAP and will contain ORF information based on the GeneMarkS-T execution.

- complete_genes.fasta
  - Amino acid sequences of complete genes from transcriptome
- partial_genes.fasta
  - Amino acid sequences of partial (5' and 3') sequences
- internal_genes.fasta
  - Amino acid sequences of internal sequences
- sequences_lost.fasta
  - Nucleotide sequences in which a frame was not found. These will not continue to the next stages of the pipeline

4.2.3 EnTAP Files: */figures*

In addition to files, EnTAP will generate figures within the */figures* directory. These are some useful visualizations of the information provided by GeneMarkS-T

- frame_results_pie.png
  - Pie chart representing the transcriptome (post expression filtering) showing complete/internal/partial/and sequences in which a frame was not found
- frame_selected_seq.png
  - Box plot of sequence length vs. the sequences that were lost during frame selection and the sequences in which a frame was found

4.3 Similarity Search (DIAMOND)

The */similarity_search* directory will contain all of the relevant information for the similarity searching stage of the pipeline. This folder will contain the *main files* (results from similarity search software), *files* analyzing hits from each database, *overall* results combining the information from each database, and *figures* generated from EnTAP.

4.3.1 DIAMOND Files: */similarity_search*

The files within the */similarity_search* directory contain the results from the similarity searching portion of the pipeline against each database you select. More information can be found at DIAMOND. With running blastp (protein similarity searching), a generic transcriptome input of "Species.fasta", with a database called "database" the files will have the following format:

- blastp_Species_database.out
  - This contains the similarity search information provided in the format from DIAMOND
  - Header information (from left to right):
    - Query Sequence ID
    - Subject Sequence ID
    - Percentage of Identical Matches
    - Alignment Length
    - Number of Mismatches
    - Number of gap openings
    - Start of alignment in query
    - End of alignment in query
    - Start of alignment in subject
    - End of alignment in subject
    - Expect (e) value
    - Bit score
    - Query Coverage
    - Subject Title (pulled from database)
- blastp_Species_database_std.err and .out
  - These files are will contain any error or general information produced from DIAMOND

4.3.2 EnTAP Files: */processed*

Files within the */processed* are generated by EnTAP and will contain information based on the hits returned from similarity searching against each database. This information contains the *best hits* (discussed previously) from each database based on e-value, coverage, informativeness, phylogenetic closeness, and contaminant status.

The files below represent a run with the same parameters as the section above:

- All the TSV files mentioned in this section will have the same header as follows (from left to right):
  - Query sequence ID
  - Subject sequence ID
  - Percentage of identical matches
  - Alignment length
  - Number of mismatches
  - Number of gap openings

  - Start of alignment in query
  - End of alignment in query
  - Start of alignment in subject
  - End of alignment in subject
  - Expect (e) value
  - Query coverage
  - Subject title
  - Species (pulled from hit)
  - Origin Database
  - ORF (taken from frame selection stage)
  - Contaminant (yes/no the hit was flagged as a contaminant)
- database/best_hits.faa and .fnn and .tsv
  - Best hits (protein and nucleotide) that were selected from this database
  - This contains ALL best hits, including any contaminants that were found as well as uninformative hits
  - The .tsv file contains the header information mentioned above of these same sequences
  - Note: Protein or nucleotide information may not be available to report depending on your type of run (these files will be empty)
- database/best_hits_contam.faa/.fnn/.tsv
  - Contaminants (protein/nucleotide) separated from the best hits file. As such, these contaminants will also be in the _best_hits.faa/.fnn.tsv files
- database/best_hits_no_contam.faa/.fnn/.tsv
  - Sequences (protein/nucleotide) that were selected as best hits and not flagged as contaminants
  - With this in mind: best_hits = best_hits_no_contam + best_hits_contam
  - These sequences are separated from the rest for convenience if you would like to examine them differently
- database/no_hits.faa/.fnn/.tsv
  - Sequences (protein/nucleotide) from the transcriptome that did not hit against this particular database.
  - This does not include sequences that were lost during expression filtering or frame selection
- database/unselected.tsv
  - Similarity searching can result in several hits for each query sequence. With only one best hit being selected, the rest are unselected and end up here
  - Unselected hits can be due to a low e-value, coverage, or other properties EnTAP takes into account when selecting hits

4.3.3 EnTAP Files: */overall_results*

While the */processed* directory contains the best hit information from each database, the */overall_results* directory contains the overall best hits combining the hits from each database.

4.3.4 EnTAP Files: */figures*

In addition to files, EnTAP will generate figures within the */figures* directory for each database. These are some useful visualizations of the information provided by similarity searching.

Here, there will be several figures:

- species_bar.png / species_bar.txt
  - Bar graph representing the top 10 species that were hit within a database
  - Text file representing the data being displayed
- contam_bar.png / contam_bar.txt
  - Bar graph representing the top 10 contaminants (within best hits) that were hit against the databast
  - Text file representing the data being displayed

4.4 Orthologous Groups/Ontology (EggNOG)

The */ontology* directory will contain all of the relevant information for the EggNOG stage of the pipeline. This folder will contain the *EggNOG files*, *files* analyzing the annotation from EggNOG, and *figures* generated from EnTAP.

4.4.1 EggNOG Files: */ontology*

Files within the */ontology* are generated by EggNOG and will contain information based on the hits returned from EggNOG against the orthologous databases. More information can be found at EggNOG.

- annotation_results.emapper.annotations
  - EggNOG results for sequences that previously hit against DIAMOND databases in similarity searching
- annotation_results_no_hits.emapper.annotations
  - EggNOG results for sequences that previously did NOT hit against DIAMOND databases in similarity searching

4.4.2 EnTAP Files: */processed*

4.4.3 EnTAP Files: */figures*

4.5 Final Annotations

The final EnTAP annotations are contained within the */outfiles* directory. These files are the summation of each stage of the pipeline and contain the combined information. So these can be considered the most important files!

All .tsv files in this section will have the following header information (from left to right)

- Query sequence ID
- Subject sequence ID
- Percentage of identical matches
- Alignment length
- Number of mismatches

- Number of gap openings
- Start of alignment in query
- End of alignment in query
- Start of alignment in subject
- End of alignment in subject
- Expect (e) value
- Query coverage
- Subject title
- Species (DIAMOND)
- Origin Database (DIAMOND)
- ORF (GeneMarkS-T)
- Contaminant (yes/no the hit was flagged as a contaminant)
- Seed ortholog (EggNOG)
- Seed E-Value (EggNOG)
- Seed Score (EggNOG)
- Predicted Gene (EggNOG)
- Taxonomic Scope (EggNOG, tax scope that was matched)
- OGs (EggNOG, orthologous groups assigned)
- Description (EggNOG)
- KEGG Terms (EggNOG)
- Protein Domains (EggNOG)
- GO Biological (Gene Ontology normalized terms)
- GO Cellular (Gene Ontology normalized terms)
- GO Molecular (Gene Ontology normalized terms)

Gene ontology terms are normalized to levels based on the input flag from the user (or the default of 0,3,4). A level of 0 within the filename indicates that ALL GO terms will be printed to the annotation file. Normalization of GO terms to levels is generally done before enrichment analysis and is based upon the hierarchical setup of the Gene Ontology database. More information can be found at GO.

- final_annotations_lvlX.tsv
  - As mentioned above, the 'X' represents the normalized GO terms for the annotation
  - This .tsv file will have the headers as mentioned previously as a summary of the entire pipeline
- final_annotated.faa / .fnn
  - Nucleotide and protein fasta files containing all sequences that either hit databases through similarity searching or through the ontology stage
- final_unannotated.aa / .fnn
  - Nucleotide and protein fasta files containing all sequences that did not hit either through similarity searching nor through the ontology stage

CHAPTER 5

Indices and tables

- genindex
- modindex
- search