Functional Annotation: Team 2
Background and Strategy

Di Zhou, Mansi Gupta, Bridget Neary, Tzu-Chuan Huang, Mingming Cao, Priyam Raut
Functional annotation is defined as the process of collecting information about and describing a gene's biological identity.

Enables you to take large lists of genes/proteins and turn them into a biologically useful model.

**Goals and Challenges**

- Accurately annotate as many genes as possible.
- Remove Duplication and redundancy for faster annotation.
Workflow

- Genome Assembly
- Gene Prediction
- Functional Annotation

...
Workflow

Genome Assembly

Gene Prediction

Protein-coding
- Domains/Motifs
- Transmembranes
- Signal Peptides

Non-coding RNA
- ncRNA
- Riboswitches
- CRISPR

Others
- Antimicrobial Resistance
- Operons
- Virulence Factors
- Plasmids
- etc.
**InterProScan**

*InterPro* is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites.

**Databases:** CATH-Gene3D, CDD, MobiDB, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SFLD, SMART, SUPERFAMILY, TIGRFAMs (*vs. NCBI-CDD: CDD, Smart, Pfam, PRK, TIGRFAM, COG and KOG*)

**Features**
- Largest source of automatic annotation of sequences in the UniProt Knowledgebase
- Provide optional prediction of signal peptides, transmembrane regions and coiled-coils, via the SignalP, Phobius, TMHMM and Coils software packages
- Flexible in output format choosing
- Time intensive

*The Newest Version: 72.0*
eggNOG-mapper is a tool for functional annotation of large sets of sequences based on fast orthology assignments using precomputed eggNOG clusters and phylogenies.

- **Database:** eggNOG itself
Approaches
1) Sequence Mapping (Choose one from A & B)
   A) HMM
      a. HMMER3 used to search for significant matches
      b. Next, each query protein is searched against the set of eggNOG proteins represented by the best matching HMM using the phmmer
      c. Finally, the best matching sequence for each query is stored as the query’s seed ortholog and used to retrieve other orthologs
   B) DIAMOND
      Faster but not sensitive as HMM
2) Orthology Assignment
   Best matching sequence used to retrieve a list of fine-grained orthology assignments.
3) Functional Annotation
Approaches

1) Sequence Mapping (Choose one from A & B)
   A) HMM
      a. HMMER3 used to search for significant matches
      b. Next, each query protein is searched against the set of eggNOG proteins represented by the best matching HMM using the phmmer
      c. Finally, the best matching sequence for each query is stored as the query’s seed ortholog and used to retrieve other orthologs
   B) DIAMOND
      Faster but not sensitive as HMM

2) Orthology Assignment
   Best matching sequence used to retrieve a list of fine-grained orthology assignments.

3) Functional Annotation
- **Test Report**

  - Compared to **BLAST** results reduced the rate of false positive assignments by 11%, and increased the ratio of experimentally validated terms recovered over all terms assigned per protein by 15%.
  
  - Compared with **InterProScan**, eggNOG-mapper achieved similar proteome coverage and precision while predicting.

  - eggNOG-mapper runs ~15x faster than BLAST and at least ~2.5x faster than InterProScan.
Signal Peptides

- Small amino peptide sequences that consist of positively charged region (n-region), central hydrophobic region (h-region) and a polar carboxy terminal domain (c-region).
- Highly Variable and Rapid evolving.
- Translocation of proteins; used to specify different cellular placement.
- 4 types of Signal Peptides - Sec, Tat, Lipo and pilin.
- Uncleaved SPs are called signal anchors.
- Non classical and leaderless secreted proteins.
SignalP (4.1 and 5.0)

Pros

- Deep recurrent neural network combined with conditional random field classification and optimized transfer learning.
- 5.0 can differentiate between various types of SPs.
- 4.1 is more customizable (D - cutoff Values, TM Regions in sequences, minimum sequence length, N-terminal truncation)

Cons

- Tat-SP cleavage site by SPase II. (PRED-TAT)
- Archaea SP prediction (PRED-SIGNAL)
- SignalP 5.0 (software yet to be released)
- 5.0 only gives a likelihood score of being a signal peptide.
- Cannot predict type IV Pilin type Signal Peptides (FlaFind).
- Non-customizable dataset (unlike DeepSig)
10-40% of proteins in prokaryotes are transmembrane proteins.

- 2 types - alpha-helix and beta barrel.
- Acts as gateways of permit the transport of proteins.
- Usually anchored in the phospholipid bilayer.
- TM HMM (Hidden Markov Model)
- TOPCONS2 (Consensus)
- MemSAT-SVM (Support Vector Machine)
- Results to be compared with homology based tools.

[Image: Transmembrane-Topology of OmpA]
Non-coding RNA Annotation

- Non-Coding RNA Annotation Tools
  - Infernal
  - tRNAscan-SE
  - ARAGORN
Infernal - INFERence of RNA ALignment

- Identifies RNA structure and sequence similarities in DNA databases
- It uses *Covariance Model (CM)* which scores a combination of sequence consensus and RNA secondary structure consensus.
- Uses CMs to search for new family members in sequence databases.
- Uses HMM filters to speed up the homology search.
tRNAscan-SE

- Used for tRNA gene identification and functional prediction, which is linked to Genomic tRNA database and UCSC genome browsers.
- Uses Infernal in background to search DNA sequences for tRNA-like structure and sequence similarities.
- Combines speed of heuristic algorithms with sensitivity and selectivity of covariance models.
- Benchmark algorithm for tRNA detection.
ARAGORN
● Identifies tRNA and tmRNA
● Employs heuristic algorithms and a modified version of BRUCE program for tRNA and tmRNA respectively.
● It also uses covariance models to predict tRNA secondary structures

Advantages:
● User friendly
● Easy to interpret results

Disadvantages:
● Computationally intensive
● Selectivity is less than tRNAscan-SE
● Number of false positives increase with increasing G+C content in input genomes
A **Riboswitch** is a regulatory segment in mRNA that causes change in protein production upon binding with a small molecule (like metabolites).

Tools for finding riboswitch gene are based on motif searching mechanisms:

**RibEx - Riboswitch Explorer:**
- It is a web server for searching sequences known for riboswitches.
- It can identify more types of riboswitches than RiboSW.

**RiboSW:**
- It is a web server based tool for searching putative riboswitches in a sequence.
- A systematic method for identifying 12 kinds of riboswitches.
● Clustered Regularly Interspersed Short Palindromic Repeats
● Found in prokaryotes like bacteria and archaea.
● Derived from DNA fragments from viruses that infected the system previously.
● Plays key role in antiviral defense system.

PILER-CR
- It is designed to identify the characteristic signature of CRISPR repeats.
- Extremely fast program (completes a 5Mb genome in ~5 secs).
- High sensitivity and specificity

CRT - CRISPR Recognition Tool
- k-mer based approach
- Improved performance for recall and quality than PilerCR
- Faster than PilerCR if genomes have large number of repeats
DOOR²: Database of proOkaryotic Operons

- Complete and reliable operon database covering 2072 bacteria genomes and with overall accuracy of ~90%
- The biggest operon database available nowadays.
- Massive alternative TUs (transcription units) from experiments and predicted by RNA-seq.
- Comprehensive operon-centered resources
- Support exploring all data in an intuitive visualization manner.
Operon

- **Required Data:**
  1. Gene locations (ptt file)
  2. Protein sequences (faa file)
  3. Chromosome sequences (fna file)
- **Classifiers:**
  1. Linear logistic function based classifier
  2. Non-linear decision tree based classifier
### Sample Output:

**Operon**

You are here: **Home** &gt; **Escherichia coli str. K-12 substr. MG1655** &gt; **NC_000913** &gt; **Operon: 2996**

#### Basic information

<table>
<thead>
<tr>
<th>Operon ID</th>
<th>Operon Size</th>
<th>Protein gene number</th>
<th>RNA gene number</th>
<th>Similar operon number</th>
<th>GI</th>
<th>Start</th>
<th>End</th>
<th>Strand</th>
<th>Synonym</th>
<th>COG Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>Row 2</td>
<td>Row 3</td>
<td>Row 4</td>
<td>Row 5</td>
<td>2996</td>
<td>190</td>
<td>255</td>
<td>+</td>
<td>b0001</td>
<td>thr operon leader peptide</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>fused aspartokinase I and homoserine dehydrogenase I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>homoserine kinase</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16127998</td>
<td>3734</td>
<td>5020</td>
<td>+</td>
<td>b0004</td>
<td>threonine synthase</td>
</tr>
</tbody>
</table>

#### Genes

<table>
<thead>
<tr>
<th>GI</th>
<th>Start</th>
<th>End</th>
<th>Strand</th>
<th>Synonym</th>
<th>COG Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>16127995</td>
<td>190</td>
<td>255</td>
<td>+</td>
<td>b0001</td>
<td>thr operon leader peptide</td>
</tr>
<tr>
<td>16127996</td>
<td>337</td>
<td>2799</td>
<td>+</td>
<td>b0002</td>
<td>fused aspartokinase I and homoserine dehydrogenase I</td>
</tr>
<tr>
<td>16127997</td>
<td>2801</td>
<td>3733</td>
<td>+</td>
<td>b0003</td>
<td>homoserine kinase</td>
</tr>
<tr>
<td>16127998</td>
<td>3734</td>
<td>5020</td>
<td>+</td>
<td>b0004</td>
<td>threonine synthase</td>
</tr>
</tbody>
</table>

#### Species name

- **Escherichia coli str. K-12 substr. MG1655**

#### NC name

- **NC_000913**

#### NC description

- **Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome.**

#### ODB info

- No information available in ODB.

#### VIMSS info

- No information available in VIMSS operon database.

#### Reference

- No literature information available.
Operon-mapper:

- A web server for accurately predicting the operons of any bacterial or archaeal genome sequence in a very easy and direct way.
- The operon predictions are based on the intergenic distance of neighbouring genes as well as the functional relationships of their protein coding products.
Operon

**Input:**
Required file: a Fasta sequence or upload a Fasta file
Optional file: GFF and GenBank file

**Databases:**
Uniprot Knowledgebase consisting of non-fragmented proteins that are from Bacteria or Archaea

**Output:**

<table>
<thead>
<tr>
<th>Output files</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Predicted operonic gene pairs</td>
<td></td>
</tr>
<tr>
<td>2 Predicted operons</td>
<td></td>
</tr>
<tr>
<td>3 Predicted ORFs coordinates</td>
<td></td>
</tr>
<tr>
<td>4 DNA sequences of the predicted ORFs</td>
<td></td>
</tr>
<tr>
<td>5 Protein sequences of the translated predicted ORFs</td>
<td></td>
</tr>
<tr>
<td>6 COGs assignments</td>
<td></td>
</tr>
<tr>
<td>7 ORFs functional descriptions</td>
<td></td>
</tr>
<tr>
<td>8 Compressed file with all the above 48325.tar.gz</td>
<td></td>
</tr>
</tbody>
</table>
**PlasmidSeeker:**

- A k-mer based program for the identification of known plasmids from bacterial whole genome sequencing reads
- A command line tool
- A database with 8514 plasmids and approximate time with 8,514 Refseq plasmids, k=20 with 32 cores and 512GB RAM was 11 minutes.
Plasmid


- Options:
  - -i - Input FASTQ file location
  - -o - Output file name (default is "plasmidseeker_result.txt")
  - -d - Path to plasmid database directory
  - -b - Closest reference bacterium FASTA file location
  - -t - Number of threads used (default 32)
  - -f - Minimum threshold F - at least this fraction of unique k-mers that has to be found for a plasmid (default 80)
  - --verbose - Prints out more of the working process
  - --ponly - Assumes that reads contain only plasmid sequences (use for extracted plasmids)

- outputs:

Reference plasmids which share more than 80% of k-mers are presented in a single cluster and ordered by the percentage of unique k-mers found.
Virulence factors (VFs) are elements that help bacteria replicate or evade host defenses more effectively.

Detection of related genes can help determine how dangerous or pathogenic a bacterial strain is expected to be.

Antibiotic resistance determinants (ARDs) or antibiotic resistance genes (ARGs) are elements that confer resistance to one or more families of antimicrobial agents.

Knowing the susceptibility of a pathogen to treatment allows more effective control of its spread.
Virulence and Antibiotic Resistance

- **VFDB/VFanalyzer**
  - A comprehensive, curated database of virulence factors of bacterial pathogens
  - New web tool VFanalyzer - iterative BLAST and gene cluster detection
  - We will use the database locally and implement a simplified algorithm

- **Victors**
  - Another manually curated database of VF sequences using BLAST

- **ResFinder**
  - Curated database of acquired antimicrobial resistance genes and mutations - BLAST

- **RGI + CARD**
  - Resistance Gene Identifier for the Comprehensive Antibiotic Resistance Database
  - Database and detection-model tool for reference antibiotic resistance gene sequences
Virulence and Antibiotic Resistance

- **DeepARG**
  - Deep learning using ARD sequences and other protein sequences to predict whether a novel sequence could confer antibiotic resistance
  - Uses a dissimilarity matrix created from known ARGs

- **PCM**
  - Pairwise comparative modeling
  - Uses protein homology modeling for 3D structure
  - Machine learning classifies based on 3D structure whether a novel gene could be ARG
● A wrapper, it collects together several pieces of software.
● Pros: Accurate and fast (smaller core databases, parallel processing, ~10 min).
● Input: a set of scaffold sequences produced by de novo assembly software.
● Two-step process for the annotation of protein coding regions
  ○ First, Prodigal identifies the coordinates of candidate genes
  ○ Second, predicted by similarity to proteins in one of many protein or protein domain databases.
● Hmmscan is implemented to search hidden Markov model profile database, including Pfam and TIGRFAMs.
● BLAST+ is used to search against sequence database(RefSeq, Uniprot...).
Prokka pipeline
● A web-based environment that allows users to upload a genome, annotate the genome, edit the annotations and compare the genome with other sequenced genomes in the SEED database.
● RAST provides a mapping of genes to subsystems and a metabolic reconstruction.
● All annotated features can be viewed and downloaded from the View Features page.
● Cons: Normally, the annotated genome is available around 12-24 hours of submission.
● Pros: RASTtk is available and it allows users to build their own annotation pipelines.
RASTtk pipeline

- **Calling genes:**
  - Offers the option with Glimmer3, GeneMarkS, and Prodigal.

- **Annotating proteins with k-mers:**
  - First searches against CoreSEED (limited number of more stable annotation.)
  - If not found, searches against the larger collection of FIGfam from the PubSEED.

- **Annotating proteins missed by k-mers:**
  - Combination of BLAT and BLASTP is performed against a set of non-redundant genus-specific protein databases for the organism's genu.
The pipeline is provided by NCBI.
PGAP combines *ab initio* gene prediction algorithms with homology based methods.
Pan-genome approach is used to annotate protein for a specific clade.
Cons: It takes around 1 day to complete the annotation.
Pros: Designed by NCBI, the result is highly approved.
Third party software are used:
  ○ tRNAScan-SE v.1.21
  ○ hmmer v.3.1b2
  ○ CRISPR v.1.02
  ○ AntiFam v.3.0
  ○ Rfam v.12.0
  ○ infernal v.1.1.1
  ○ TIGRfam 15.0 (for naming)
GeneMarkS2+ now being used for *ab initio* gene prediction
Output file format: annot.fna, annot.faa, annot.gbk, annot.gff, annot-gb.ent
BEACON provides detailed comparison of gene function annotations by different annotation methods (AMs) and generates extended annotations through combination of individual ones.

For two annotation X and Z, the similarity score is calculated:

- Identical: Genes from annotation X that completely overlap with genes in annotation Z.
- Similar: Genes share some common locations between the two annotations more than a given offset (e.g. 2%).

\[
\text{SimilarityScore} = \frac{\text{Identical} + \text{Similar}}{\text{Total}_X + \text{Total}_Z} \times 2 \times 100
\]
Proposed pipeline

Assembly contig (.fasta)
Gene prediction results (.gff)

Clustering

Protein-coding region
- Domain/motif
- InterProScan
- BLAST
- EggNOG
- Transmembrane region
- THHMM
- MemSVM
- Signal peptide
- SignalP
- DeepSig

Non-coding RNA
- Riboswitches
- Riboswitches
- Riboswitches
- Riboswitches
- tRNA
- tRNA
- tRNA
- tRNA
- sRNA
- CRISPER
- PlierCR

Others
- Antibiotic Resistance Genes
  - RGI
  - CARD
  - MEGARes
  - PCM
  - DeepARG
- OPERON
- DOOR2
- Virulence Factors
- VFDB
- Plasmid
- PlasmidSeeker

BEACON

Prokka, RAST, PGAP validation

Merge Annotation (.gff, .gbk)
References

References(contd.)

- Laslett, Dean, and Bjorn Canback. "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences." Nucleic acids research 32.1 (2004): 11-16.
- InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research, Jan 2019; doi: 10.1093/nar/gky1100