CDS GENE PREDICTION WORKFLOW

Overview of tools and pipeline

Barnap
Non-Coding RNA Prediction workflow

50 Genomes provided by the genome assembly team → ARAGORN → 50 Fasta files containing tRNA genes → MERGE → 50 Fasta files containing tmRNA, tRNA and rRNA genes → 50 Genomes provided by the genome assembly team → BARRNAP → 50 Fasta files containing rRNA genes → MERGE → 50 Fasta files containing tmRNA, tRNA and rRNA genes
Validation of Tools

- Make local databases using genes from known strains downloaded from NCBI. (DB)
- Try ab-initio tools to predict genes using corresponding genomes. (query)
- Blastn for 5 known strains, query predicted genes against their own database of genes respectively.
- Gain the $\frac{TP}{TP+FP}$, $\frac{TP}{TP+FN+FP}$, $\frac{TP}{TP+FN}$ of every tool.

Sensitivity : $\frac{TP}{TP + FN}$  Positive Predictive Value: $\frac{TP}{TP+FP}$  Accuracy: $\frac{TP}{TP+FN+FP}$

TP: True Positives; FP: False Positives; FN: False Negatives
Evaluation of tools--GeneMarkS2, Prodigal, Glimmer3

For every strain:

TP: predicted seqs that have hits in blast results

FP: predicted seqs that do not have hits in blast results

FN: sequences that did not have hits in the blast db.
Positive Predictive Value

<table>
<thead>
<tr>
<th>Sample</th>
<th>0104</th>
<th>0157</th>
<th>IAI39</th>
<th>k12</th>
<th>UMN026</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure Names</td>
<td>GeneMarkS-2</td>
<td>Glimmer</td>
<td>Prodigal</td>
<td>GeneMarkS-2</td>
<td>Glimmer</td>
</tr>
<tr>
<td>Value</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

GeneMarkS-2, Glimmer and Prodigal for each Sample. Color shows details about GeneMarkS-2, Glimmer and Prodigal.
Sensitivity

GeneMarkS-2, Glimmer and Prodigal for each Sample. Color shows details about GeneMarkS-2, Glimmer and Prodigal.
GeneMarkS-2, Glimmer and Prodigal for each Sample. Color shows details about GeneMarkS-2, Glimmer and Prodigal.
The PRODIGAL gene returns!!

Prodigal (**Prokaryotic Dynamic Programming Genefinding Algorithm**) is a microbial (bacterial and archaeal) gene finding program developed at Oak Ridge National Laboratory and the University of Tennessee.

Prodigal is an extremely fast gene recognition tool (written in vanilla C). It can analyze an entire microbial genome in 30 seconds or less.

Prodigal is a highly accurate gene finder. It correctly locates the 3' end of every gene in the experimentally verified Ecogene data set (except those containing introns).

It possesses a very sophisticated ribosomal binding site scoring system that enables it to locate the translation initiation site with great accuracy (96% of the 5' ends in the Ecogene data set are located correctly).

```
prodigal -i input_file -a protein -d nucleodite -o gff_file -f gff
```
Sum of F2 for each Sequences. The view is filtered on Sequences, which excludes sequence_.fna.
GeneMark S-2 is based on a hidden Markov model (HMM) framework that follows the logic of the genetic structure of the bacterial genome.

It can deal simultaneously with direct and reverse DNA strands using the functional units of bacterial genomes.
Sum of Number of Genes for each Sequence.
Sum of F2 for each Sequence.
Let's have a BLAST!! (Validation)

Database: swissprot_v5
Taxonomy: Escherichia coli
Max_target_seq: 1
E-value: 1e-6

Version_5 database supports a faster searching method depending on taxonomic node.


AAC51230 (human MEN1 protein) against nr
Validation of the results from the Union

<table>
<thead>
<tr>
<th>Fasta file</th>
<th>Blast result</th>
<th>Two fasta file</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq1</td>
<td>seq2</td>
<td>TP seq2 seq3 seq4 seq5 seq6</td>
</tr>
<tr>
<td>seq2</td>
<td>seq3</td>
<td></td>
</tr>
<tr>
<td>seq3</td>
<td>seq4</td>
<td></td>
</tr>
<tr>
<td>seq4</td>
<td>seq5</td>
<td></td>
</tr>
<tr>
<td>seq5</td>
<td>seq6</td>
<td></td>
</tr>
<tr>
<td>seq6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq7</td>
<td></td>
<td>FP seq1 seq7</td>
</tr>
</tbody>
</table>
Sum of % of Blast hits for each Sequence.
Non-coding region

- Aragorn (For tRNA and tmRNA prediction)

```
aragorn -l -t -gc1 -w input_file -o tRNAtxt # The input is fasta file, and the output is aragon default txt format.
aragorn -l -m -gc1 -w input_file -o tmRNAtxt
aragorn -t input_file -fo tRNAfasta # Print the output in fasta file only.
aragorn -m input_file -fo tmRNAfasta
```

- Barrnap (For rRNA prediction)

```
barrnap --outseq filename 1020_scaffolds.fasta
```
Aragorn

Sum of F2 for each sequences.
Sum of F2 for each sequences.
Pipeline

Requirements

- python3
- Latest Perl
- bedtools
- samtools
- Latest Prodigal
- Latest GeneMarkS-2
- Latest Aragorn
- Latest Barrnap
- Latest Biopython (If running Bedtools)

Quick Start

- f : Path to file input directory (Required)
- p : Run Prodigal prokaryotic mRNA gene prediction tool
- g : Run GeneMarkS-2 prokaryotic mRNA gene prediction tool
- nc : Run Aragorn and Barrnap to predict tRNA/tmRNA and rRNA (respectively) (optional)
- ncs : Separate Aragorn and Barrnap results into two distinct sets of nucleotide fasta files

Default behavior will still require -f and will run both Prodigal and GeneMarkS-2 with Bedtools.
Bedtools will run if both Prodigal and GeneMarkS-2 are run, and includes a union folder of both tools.
Example usage: ./geneprediction_pipeline_t1.py -f <input_dir>

Output Description

Prodigal and GeneMarkS-2 run individually will be found in their respective folders, ./prodigalresults or ./gms2results.
Output files are split into three folders. One for GFF format, fna and faa.
If Prodigal and GeneMarkS-2 are run in tandem, then the combined output will also be in ./prodigal-geneMark

Aragorn and Barrnap results are joined by default into single .fna files by assembly, located in ./arabarr

Nucleotide and Amino acid fasta formats may be used with BLAST homology validation as described below.
**BLAST validation**

**Requirements**
- Version_5 database (required)
- taxonomic_id list (required)
- EDirect command-line utility (required)
- Latest Perl (required)
- python3 (required)
- blast+ (required)

**Quick start**

For downloading database:

Use `./update_blastdb.pl --blastdb_version 5 --showall` to see the option.
Use `./update_blastdb.pl --blastdb_version 5 [Database] --decompress` to download.

For getting the taxonomy_idlist:

Use `get_species_taxids.sh -n [organism]`

For blastp (amino acid):

```
./blastp.py -d [queried_fold] -t [taxonomy_idlist] -o [outputfolder]
```

or blastx (DNA seqs):

```
./blastx.py -d [queried_fold] -t [taxonomy_idlist] -o [outputfolder]
```

**Output Description**

There will be two folders in your output folder:
- knownprotein/: The fasta files in this folder have got rid of the sequences that do not have hit in blast.
- novelgene/: The fasta files in this folder do not have hit in blast.
References

https://biopython.org/wiki/Download
http://www.vicbioinformatics.com/software.barrnap.shtml
http://mbio-serv2.mbioekol.lu.se/ARAGORN/Downloads/
http://topaz.gatech.edu/GeneMark/license_download.cgi
http://compbio.ornl.gov/prodigal/
https://bedtools.readthedocs.io/en/latest/content/installation.html
http://samtools.sourceforge.net/
https://www.python.org/
http://www.perl.org/get.html
https://www.ncbi.nlm.nih.gov/books/NBK179288/