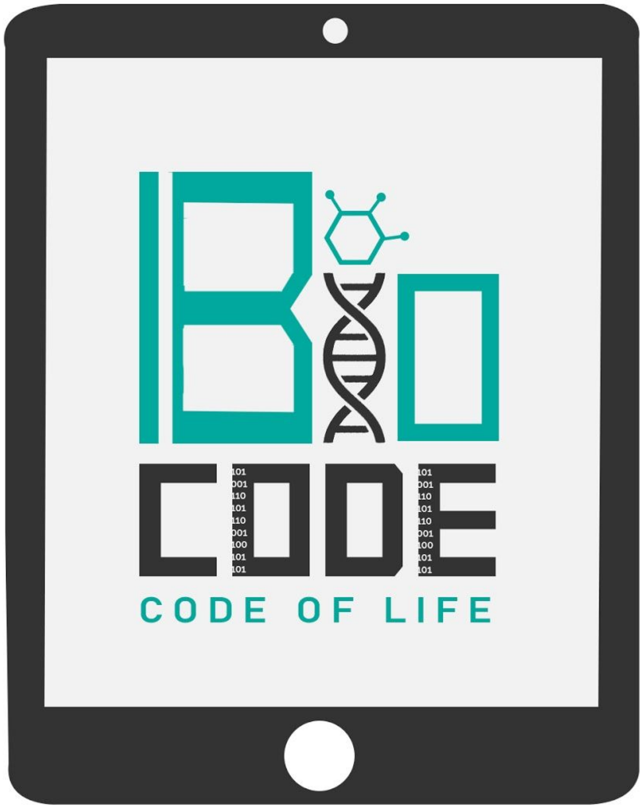

LEARN PYTHON & R FOR BIOINFORMATICS

Prerequisite Terminologies:

In order to have thorough understanding of the main topic, you should have the basic concept of following terminologies:

- **Gene Prediction:** The process of identifying the regions of genomic data that encode genes.
- **GFF Format (General Feature Format):** File format used for describing genes and other features of DNA, RNA and protein sequences.
- **GenBank Format:** Flat file format, stores sequence and its annotation together.

Introduction:

Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm) is a gene prediction software, that is utilized for prokaryotic genomes (runs efficiently on finished genome, draft genomes and metagenomes). Prodigal is utilized through command-line such as CMD on windows and terminal or BASH on Linux.

Characteristics:

- Prodigal is a fast and lightweight open source program (analyzes the E-Coli k-12 genome in 10 seconds on a modern MacBook Pro) which achieves efficient results compared to other gene prediction methods.
- It provides accurate protein-coding gene prediction in GFF3, GenBank and Sequin table format.
- Most advantageous fact about prodigal is that it is an **unsupervised machine learning algorithm**. It does not really need to be provided with any training data or data set, instead it automatically learns the properties of the given genome file including RBS motif presence, start codon usage and coding statistics.
- In case, when you're dealing with shorter genomes, you have to provide a dataset for training of the prodigal that can be a known gene of the particular related genome. Prodigal will compare the predicted genes with available data sets.
- It can handle gaps and partial genes as well. It'll ignore N's (repeated mask regions) to predict the genes out of the particular genome. The user can specify if prodigal should build genes across runs of N's as well as how to handle genes at the edges of contigs.
- It can identify correct translation sites and can output information about every potential start site in the genome, including confidence score, RBS motif and much more.

Modes:

There are three particular modes within prodigal:

- **Normal Mode:**
  In which you provide genome sequence, prodigal will study it, learn its properties and predict genes based on these properties.
- **Anonymous Mode:**
  In which prodigal applies pre-calculated training files to the provided input sequence and predict genes based on the best results. As discussed, this mode is utilized when the genome size is smaller than 300kb or 500kb (Bacterial genomes can be that small), keeping in mind that prodigal works with only prokaryotes.
- **Training Mode:**
  It works as normal mode but in this mode prodigal saves a training file for future use.
  [For example: If you've determined gene prediction of any particular genome using training mode, it would not require any data set related to the particular genome but save those statistics for the later analysis.**]**

➢ Which mode should be used depends on what type of data set you're analyzing.

Normal Mode	In case, when you've sufficient data (100kb+ for good 3' prediction, 500kb+ for good 5' prediction), Used on finished genomes, draft genomes and big viruses.
Anonymous Mode	Used on metagenomic data sets, Low quality draft genomes, Small viruses & small plasmids.
Training Mode	Primarily useful when you wish to train on a different sequence than the one you wish to analyze (as it saves training files for future analysis).

Parameters:
As prodigal is a command-line tool so it requires basic parameters that should be defined in the command.

- **Mode Parameter:**

-p, --mode: Specify mode (normal, anonymous or training)

Normal	Single genome, any number of sequences (by default).
Anonymous	Anonymous steps, analyze using preset training files, ideal for metagenomes or short sequences.
Training	Do only training, input should be multiple FASTA of one or more closely related genomes.
Meta	(Deprecated) same as anonymous.
Single	(Deprecated) same as normal.

- **Input/Output Parameters:**

-i, --input_file	Specify input file (Default stdin)
-o, --output_file	Specify output file (Default stdout)
-a, --protein_file	Specify protein translation file
-d, --mrna_file	Specify nucleotide sequences file
-s, --start_file	Specify complete starts file
-w, --summ_file	Specify summary statistics file
-f, --output_format	Specify output format ● **gbk:**Genbank format(Default) ● **gff:**GFF format ● **sqn:**Sequin feature table format ● **sco:**Simple coordinate output
-q, --quiet	Run quietly (Suppress logging output) (when you have quite a bigger genome)

- In this video, we've discussed only normal mode, other modes will be discussed in other sections of this video.
  The basic command for gene prediction through normal mode:

  ```
  $ prodigal -i my.genome.fna -o gene.coords.gbk -a protein.translations.faa
  ```

  prodigal: to call in the prodigal software.
  -i my.genome.fna: the input file of your particular genome and its format
  -o gene.coords.gbk: the output file of predicted genes and its format.
  -a protein.translation.faa: the output file of protein translations of the genes and its format.

Steps:

Installation:

- Use the link: https://github.com/hyattpd/prodigal to download the prodigal software.
- After downloading, you'll find the option 'Installing Prodigal' which will open up another page from where you can install it on Mac OS X, Generic Unix or on Windows.
- Installing prodigal on Linux is quite easy, just run the command:
  '*sudo apt install prodigal'*

Analyzing a Particular Genome:

- To analyze a particular genome, you can download it from NCBI but mostly when you perform gene prediction you don't really download a particular genome, you should've your own sequenced genome.

  Note: In this video, we've performed gene prediction analysis on the genome sequence of *'Lactobacillus Fermentum'*. You can use any other prokaryotic genome for your analysis.
- To have a look at the sequence of this particular genome:

  ```
  head  -n 10  seq*
  ```

  [You'll see the first 10 lines of the sequence of Lactobacillus fermentum. As it was downloaded in FASTA format, you'll see the accession number in one line definition which you can use to download it.]
- To count the exact number of letters and lines within the sequence:

  ```
  wc  seq*
  ```

  [29984 lines and 2128733 letters will be the count of this genome.]

  Note: It'll vary when you perform analysis on other prokaryotic genomes.

Gene Prediction:

- List the genomic file of the given genome by command ***'ls'***.
- To predict the genes out of your genome, use the basic command which was discussed earlier.

  ```
  prodigal  -i sequence.fasta -o output_genes.gbk -a output_protein.faa
  ```

  [Call in the prodigal tool and provide the input file of the genome, output gene file and its format that you want (GenBank format which is by default) and the protein translations file of the predicted genes and its format.**]**

  [It'll start analyzing the sequence, as discussed it is quite efficient in predicting a particular genome.]
- Run ***'ls'*** to list all the files present in the prodigal.

Visualizing the Genes File:

-
  ```
  cat  output_gene.gbk
  ```

  [You'll see all the genes that have been predicted form this genome.]
- Counting the exact number of genes:

  ```
  grep  -c "CDS" output_genes.gbk
  ```

  [2094 numbers of predicted genes.]

Visualizing the Protein Translation File:

-
  ```
  cat  output_protein.faa
  ```

  [Prodigal will start showing the translated proteins. * represents the stop codon in the Sequence, where the protein translation terminated. You'll see the accession number of the genome we analyzed, total of 2094 proteins have translated, starting position and ending position, the strands where genes are present, start type mentioned as ATG and RBS motif and GC content is also mentioned.]
- To predict mRNA out of your genome, following changes should be made in the basic command:

  ```
  prodigal  -i sequence.fasta -o  output_genes.gff -a output_protein.faa -d output_mRNA.fa -f gff
  ```

[Gene files will be created in both Genbank and Fasta format, also protein translation file, mRNA file and the original genome sequence file.]

Visualizing mRNA file:

- ***cat *mRNA.fa***

  [Prodigal will show you the mRNA sequences of the particular genome.]

Visualizing GFF Format of Gene file:

- ***cat *.gff***

  [GFF format will be displayed, it can be utilized for the annotation and gene structure display as well.]

Summary:

In this video, we learned to utilize the Prodigal gene prediction software. We've gone through with its several properties, different modes, parameters used in it and its advantages over other tools. We learned to perform the analysis of a particular genome and predicted genes out of it utilizing prodigal software.