HOW TO USE THE NAIVE BAYES CLASSIFIER (NBC) FOR METAGENOMIC DATA ANALYSIS

The Naive Bayes Classifier (NBC) is an C++ program from the EESI Lab at Drexel University. It uses the Naive Bayes algorithm to classify metagenomes using k-mer frequencies. In the instruction, you'll be guided on how to use NBC to train and classify metagenomes.

WHAT YOU NEED:

1: THE NBC PROGRAM:
https://github.com/EESI/Naive_Bayes.git

2: .METAGENOME SEQUENCE FILES FOR TRAINING AND CLASSIFY

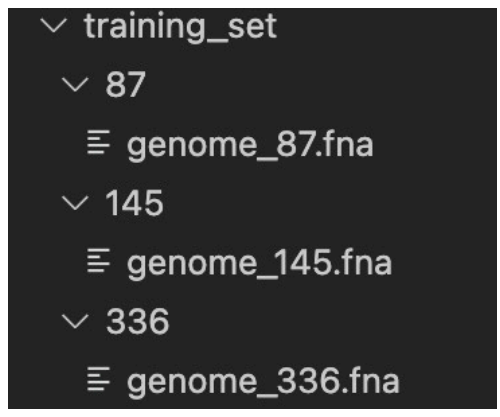
TRAINING WITH NBC

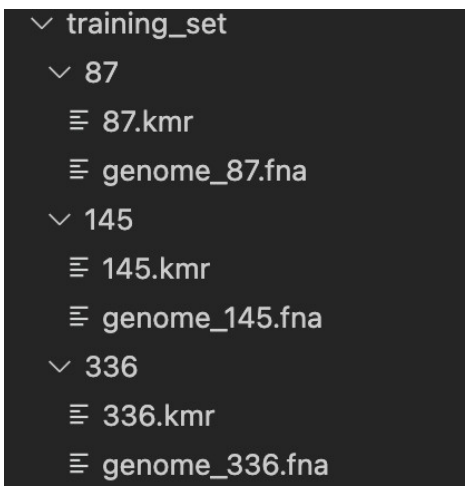
1: Prepare Training Data:

1A: Create a directory called "training_set" as the root directory.

1B: Within "training_set," make sub-directories for each class using taxonomy IDs as names.

1C: Place the corresponding metagenome sequence files for training into their respective directories.





2: Generate k-mer Counting Files:

2A: Use external kmer counting tool **Jellyfish** to create k-mer count files. You can install jellyfish and learn how to use it on: **https://github.com/gmarcais/Jellyfish.git**

You can also use the provided: **inbc/jellyfish_gen_new.bash** to run Jellyfish.

```
% bash ./inbc/jellyfish_gen_new.bash ./training_set 15 false
```

Make sure to set to "**false**" for the canonical counting of the kmers, which NBC uses.

2B: Ensure all k-mer count files are stored in the correct sub-directories and have a consistent file extension, different from non-k-mer count files.

3: Execute NBC Training:

3A: Run NBC with the following command structure:

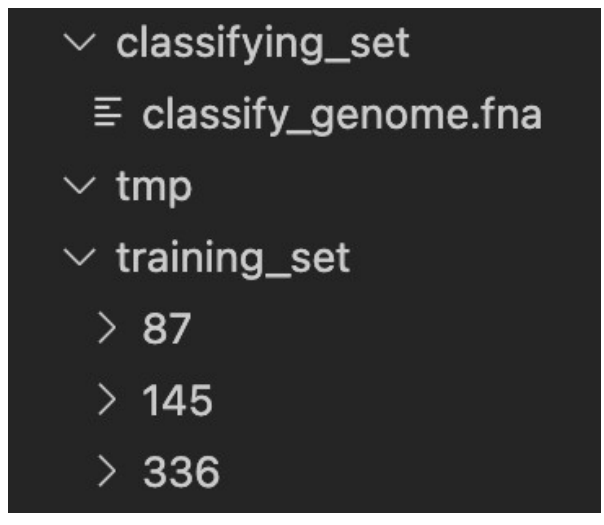
NB.run train [training_set] -s [model_save_directory] -k [k-mer_size] -m [memory_limit_in_MB] -t [threads] -e [file_extension]

Replace placeholders with your specific parameters, including the k-mer size and file extension used in your k-mer count files.

Example:

```
$ ./NB.run train ./training_set/ -s ./training_set/save -k 15 -m 2000 -t 48 -e .kmr
```

CLASSIFYING



1: Prepare Classification Data:

1A: Create a directory named "classifying_set" for the metagenome file to classify (**separate** "classifying_set" from "training_set").

1B: Store the metagenome file in "classifying_set". If there are multiple files, concatenate them into **one**.

1C: Make an **empty** directory named "tmp" for storing temporary files.

2: Execute NBC Classification:

2A: Run NBC with the following command structure:

NB.run classify [classifying_set] -s [trained_model] -k [kmer_size] -m [memory_limit_in_MB] -t [threads] -o [output_prefix] -d [temporary_directory] -r [output_max_row] -c [output_max_col] [-f]

Replace placeholders with your specific parameters, ensuring the k-mer size matches the one used in training. If the flag **-f** is **NOT** specified, NBC will only output the max likelihood of each genome. If the flag **-f** is specified, NBC will output the max likelihood and the full result of each genome against all the classes.

Example:

full result + max:

```
% ./NB.run classify ./classifying_set -s ./training_set/save -k 15 -m 2000 -t 48 -o result_set_1 -d ./tmp -r 20000 -c 20000 -f
```

max only:

```
% ./NB.run classify ./classifying_set -s ./training_set/save -k 15 -m 2000 -t 48 -o result_set_1 -d ./tmp -r 20000 -c 20000
```

The default of output_max_row is: 450000.
output_max_col is: 20000.

Please enter threads >= 2 for classifying

Please keep in mind:

1: Ensure that the directory structure is prepared exactly as outlined in the instructions.

2: Verify that the temporary directory (-d [temporary_directory]) is empty before classifying with NBC.

Upon completion of the process, NBC will output the results in the program's current directory.

For issues or questions, contact the EESI Lab support at eesi@drexel.edu.