Project: DNA classifier

Author: Aaryan Garg

Objective: Use DNA base sequence to determine the superpopulation group of a person

Data source: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/
Has DNA base sequences for 2950 samples from 5 superpopulations.

Project repository: https://github.com/ryngrg/DNA_classifier

Files:

- **master.py** - Main file of the project, run this. This calls everything. Should have no complex code.
- **downloadData.py** - Downloads all sequence reads (.filt.fastq.gz) of all samples and stores it into ./phase3_data/sampleName/
- **unzipAllGz.py** - Unzips all .filt.fastq.gz files
- **parseSequence.py** - Reads the data in all unzipped data files (.filt.fastq) and encodes the raw sequences in a binary file (.bin)
- **ML.py** - This file contains the LSTM creation and training function.
- **trainingData.py** - Contains generator function: Reads .bin files and stores data in one hot vectors.
- **samples_population.csv** - CSV file summarizing the samples of 3115 people, their genders, population code and the superpopulation code.

Figure 1 File structure

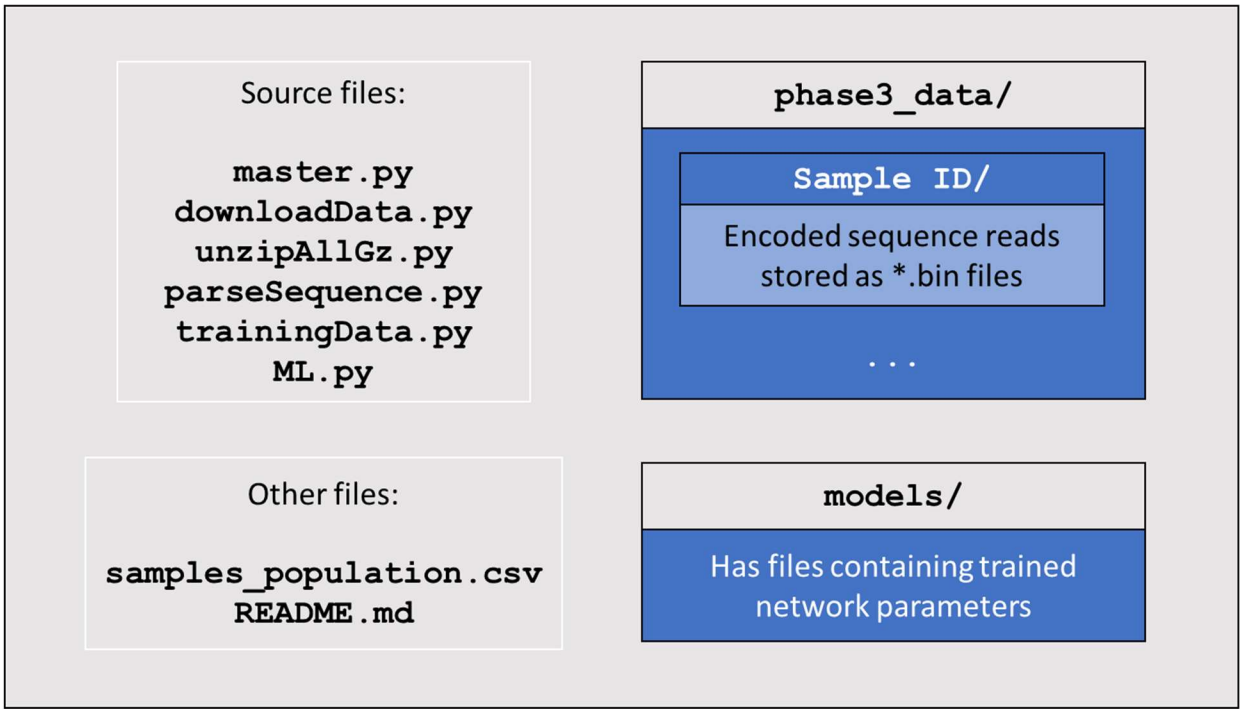
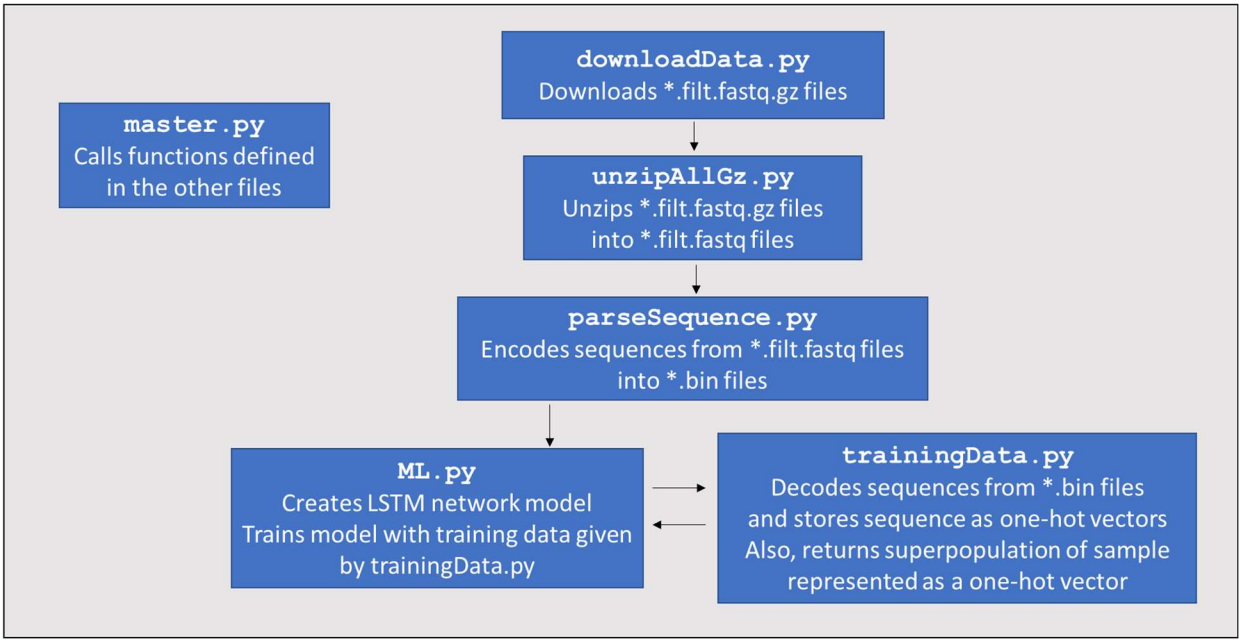


Figure 2 Bird's eye view of the project



FASTQ file parsing implementation details:

Figure 3a

Step #1: Reading sequence from .fastq file

a) files with letter space representation

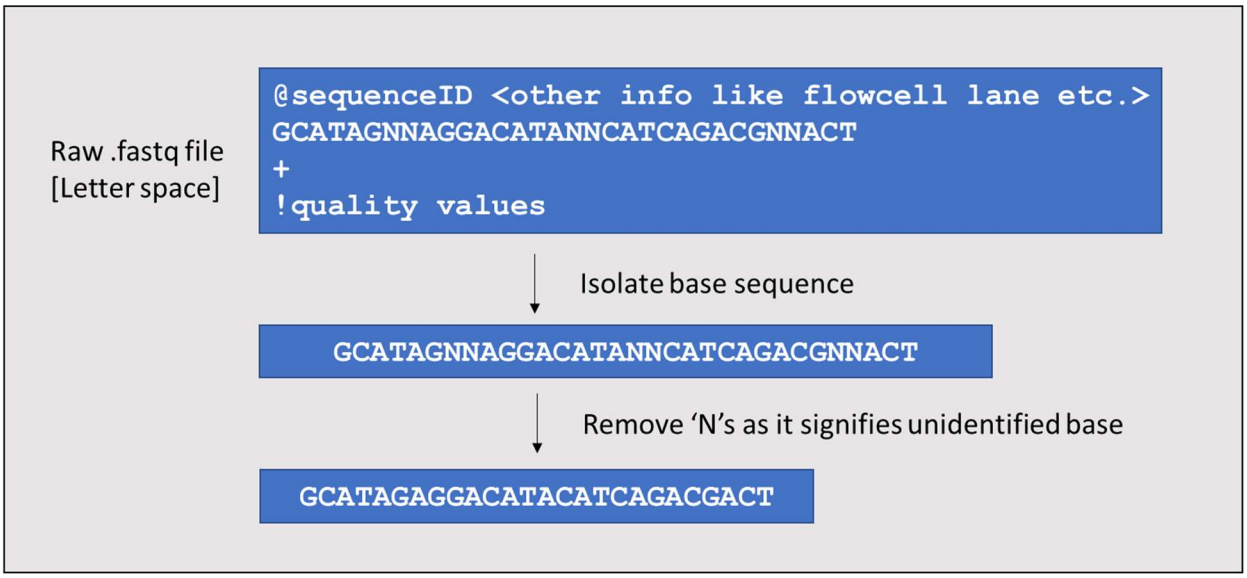


Figure 3b

b) files with color space representation

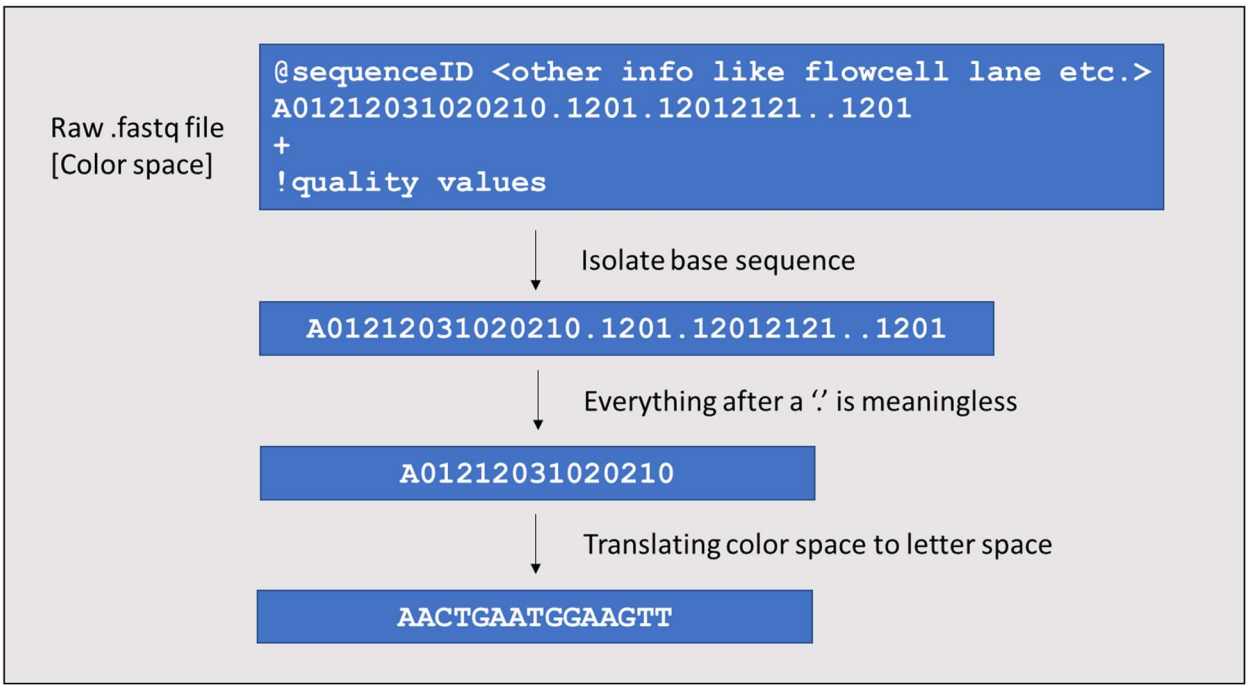
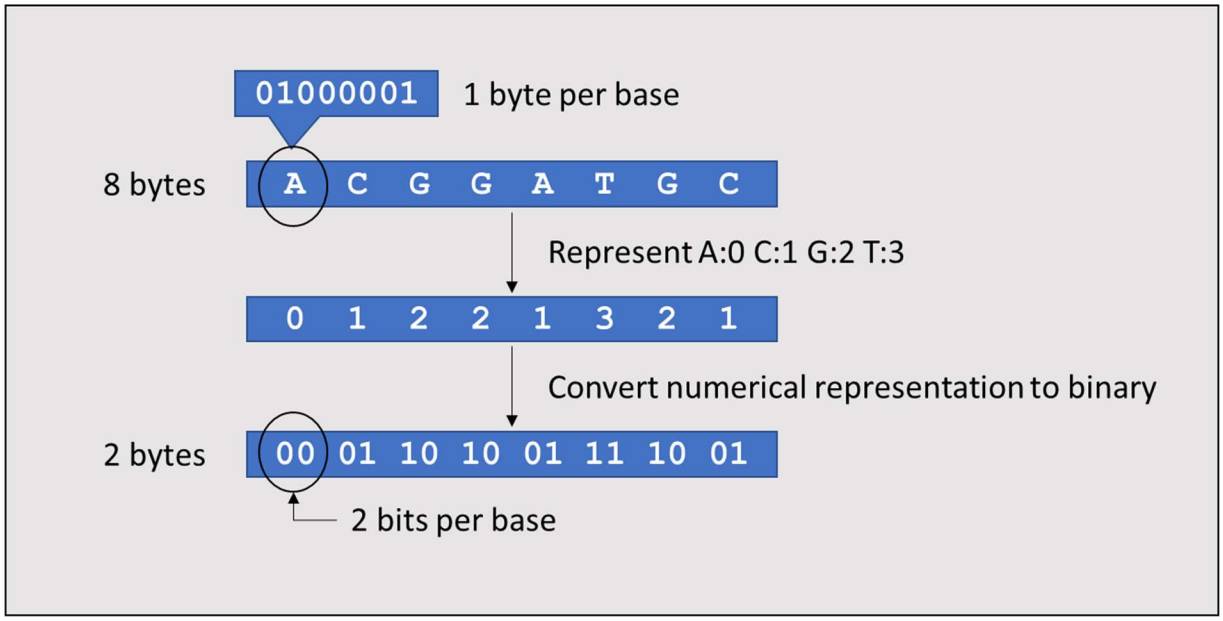


Figure 4

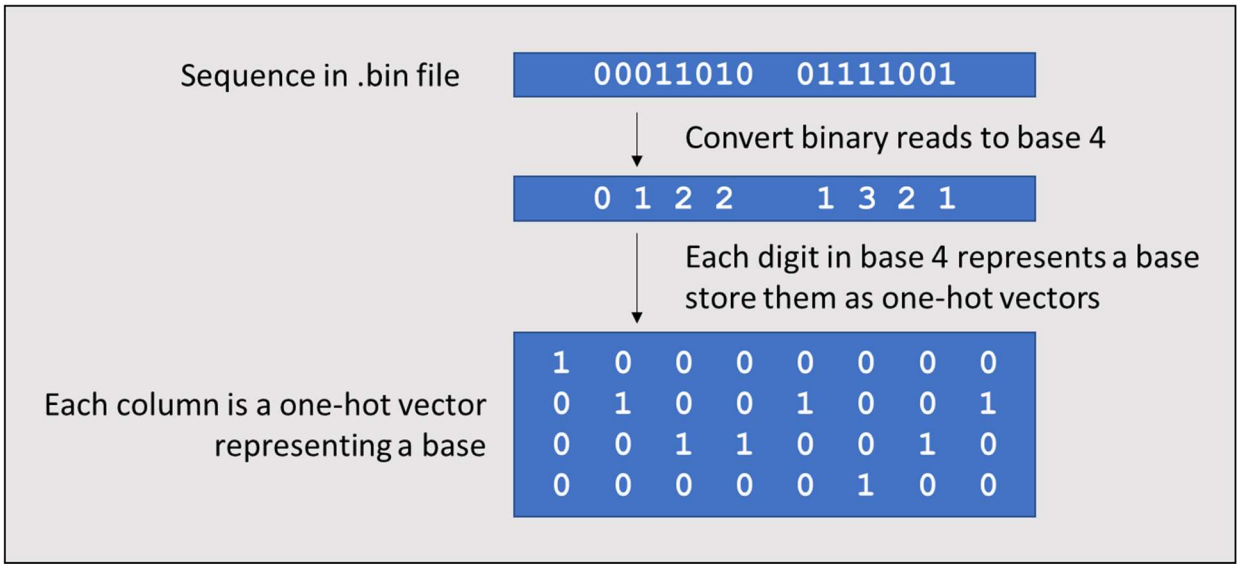
Step #2: Encoding base sequence



Training data generator implementation details:

Figure 5

Decoding base sequence



LSTM model used:

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, None, 64)          17664
_________________________________________________________________
dropout (Dropout)            (None, None, 64)          0
_________________________________________________________________
lstm_1 (LSTM)                (None, 10)                3000
_________________________________________________________________
dropout_1 (Dropout)          (None, 10)                0
_________________________________________________________________
dense (Dense)                (None, 5)                 55
=================================================================
Total params: 20,719
Trainable params: 20,719
Non-trainable params: 0
_________________________________________________________________
```