mGPS Algorithm Optimization

Course: Bioinformatics Research Project (BINP37), 15 credits

Student: Chandrashekar CR

(email: ch1131ch-s@student.lu.se)

Supervisor: Eran Elhaik

(email: eran.elhaik@biol.lu.se)

Lund University 2025

Abstract

mGPS (microbiome Geographic Population Structure) is a novel algorithm designed to analyze the geographic distribution of microbial populations. This report details the optimization of the mGPS algorithm, focusing on enhancing its computational efficiency and accuracy in processing large datasets. The optimization process involved refining the algorithm's core functions, improving data handling capabilities, and implementing parallel processing techniques. The results demonstrate significant improvements in processing speed and accuracy, making mGPS a more robust tool for microbiome research. Future work will explore further enhancements and applications of mGPS in various biological contexts.

In this work I mainly focused on optimzing the mGPS algorithm to improve its performance in analyzing the geographic distribution of microbial populations. I primarily focussed on the Metasub dataset, which consists of around 4070 samples from 2016 to 2017. I implemented several models to optimize the algorithm, and compared the results with the original mGPS algorithm.

This research presents an enhanced version of the microbial Global Population Structure (mGPS) algorithm for predicting geographical origins of microbial samples. While previous implementations of mGPS achieved 92% accuracy at city-level prediction with a median error distance of 137 km, these metrics were interdependent, potentially masking true performance. Our study implements an ensemble learning approach combining XGBoost, CatBoost, TabPFN, neural networks, and GrowNets with a threshold-filtered meta-model architecture to improve prediction accuracy across hierarchical geographical scales. Using a dataset of 4,070 samples collected from 40 cities across 7 continents, we address class imbalance through SMOTE techniques and implement 5-fold cross-validation with stratification to minimize bias. A key contribution is our corrected calculation of error metrics, providing a more accurate assessment of model performance by considering the weighted contribution of different prediction scenarios. This refined methodology offers a more robust framework for microbial geolocation, with potential applications in forensic investigation, public health surveillance, and ecological monitoring.

1. Introduction

1.1 [First subsection title]

[Your introduction text here. This section should provide background information on your research topic.] The ability to predict the geographical origin of microbial communities has significant implications across multiple disciplines, including biosurveillance, forensic investigation, and public health monitoring. Microorganisms present in environmental samples can serve as biological signatures of specific locations, reflecting local environmental conditions, human activity patterns, and ecological factors unique to particular geographical regions.

The microbial Global Population Structure (mGPS) algorithm represents an innovative approach to leverage these microbial signatures for geographical prediction. By analyzing the relative sequence abundance (RSA) of microorganisms in samples, mGPS can infer the likely origin of a sample at multiple geographical scales—from continent to precise coordinates. The original implementation employed a hierarchical prediction model using XGBoost, where continent classification preceded city prediction, followed by latitude and longitude coordinates. This approach achieved notable success with 92% accuracy at the city level and a reported median error distance of 137 km.

However, the original methodology contained notable limitations. First, the reported error metrics were interdependent—the distance error calculation benefited from the high accuracy of continent and city predictions, potentially obscuring the true predictive performance. Second, the hierarchical nature of the model meant that errors propagated through prediction levels, with early misclassifications affecting subsequent predictions.

Our research addresses these limitations through several methodological improvements. Instead of relying on a single algorithm, we implement an ensemble approach combining multiple machine learning models: XGBoost, CatBoost, TabPFN, neural networks, and GrowNets. We introduce a threshold-filtering mechanism where only models achieving specified accuracy thresholds contribute to meta-model predictions at each geographical level. This selective ensemble approach helps mitigate error propagation through the hierarchical prediction chain.

To address dataset imbalances, we employ Synthetic Minority Over-sampling Technique (SMOTE) particularly at the continent level, ensuring more balanced training representations. Our implementation also incorporates 5-fold cross-validation with stratification to maintain consistent class distributions across training and testing splits, reducing potential biases in model evaluation.

A significant contribution of our work is the implementation of corrected error calculations that provide a more comprehensive understanding of model performance. By computing the expected coordinate error as a weighted sum over different correctness

combinations (continent and city predictions), we offer a more nuanced view of prediction accuracy that accounts for the hierarchical nature of the geographical prediction task.

This research builds upon a dataset comprising 4,070 samples collected from train stations and major urban centers across 40 cities and 7 continents, each geo-tagged with precise latitude and longitude coordinates. Through our methodological enhancements, we aim to optimize the mGPS algorithm to provide more accurate and reliable geographical predictions based on microbial community profiles.

2. Materials and Methods

2.1 Data collection

The data used for this study is part of the Metasub dataset, which includes approximately 4070 samples collected from various locations between 2016 and 2017. Subways stations from 40 citites were collected by standardized protocols. 16sRNA sequencing was performed on the samples, and the resulting data was processed to extract relevant microbial population information. The metagenome dataset was processed according using the standardized protocols, which included quality control, filtering, and normalization of the sequencing data. The sequences were then converted to their relative sequence abundances, which were used for further analysis. The dataset for this study was obtained from the previous work on the mGPS algorithm, which was designed to predict the location of the samples based on their microbial composition.

2.2 Preprocessing of the dataset

The preprocessing of the dataset involved several steps to ensure the data was suitable for analysis. This included quality control to remove locations that had less samples and corrections of co-ordinates where the samples were collected. This was done exactly the same way as described in the original mGPS paper. The code written in the orignial paper was in R, the similar code was written in Python to preprocess the dataset.

2.3 Optimization of the mGPS algorithm

The mGPS algorithm was optimized by implementing several strategies to enhance its performance. Since neural networls are considered to be the best approach for this kind of problem, I implemented several neural network models to optimize the algorithm. The following models were implemented:

- **Model 1: Seperate Neural Networks** A completely disconncected neural network model which was built separately on prediciting contient. Then another model was built to predict the city based on the predicted probabilities of the continent as augmanted features. Then similary the predicted probabilities of city were used to predict the exact location of the sample, i.e latitude and longitude. The model was trained accroding to the 60 20 20 split of the dataset, where 60% of the data was used for training, 20% for validation, and 20% for testing.

- **Model 2: Combined Neural Network** A single neural network model that was built to predict the continent, city, and exact location of the sample in a single pass. The loss function was designed to optimize the predictions for all three levels simultaneously.

- **Model 3: Grownet** A neural network that was built using the Grownet architecture, which is designed to handle hierarchical predictions. This is also a bossting model which uses shallow neural networks to predict the continent, city, and exact location of the sample in a single pass. The model was trained using the same 60 20 20 split of the dataset.

- **Model 4: XBNet** This is a neural network that was built using the XBNet architecture, This model uses a combination of XGBoost and neural networks to predict the exact location fo the sample. This model was only trained to predict the exact location of the sample, i.e latitide and lonngitude. The classificaiton model requires to see all the different classes in a batch, which is not possible in this case because the dataset is not balanced. Hence, this model was only limited to prediction the latitude and longitude for each sample. This uses XGBoost's feature importance to select the most important features for the prediction, to reweight the hidden layers of the neural network.

- **Model 5: Ensemble Model**This model consists fo neural networks for classificaiton of continent and city, and a regression model for predicting the latitude and longitude of the sample. A classificaiton model in Grownet, XGBoost, CatBoost, and LightGBM was used to predict the continent and city of the sample. The same models were also use to predict the latitude and longitude of the sample, but this time the output was a regression model. The final output of the continent layer was the predicted probabilities of the contientn, which was then used to train a meta-model to get the best preidtion for continent and city. The same was done for the latitude and longitude of the sample. A caveat was added that if each model perfroms at least at 93% accuracy, then the model will be used to predict the continent and city of the sample. or above, then the predicted probabilities will be added to the final prediction using the meta-model.

2.4 Ensemble Learning Approach

To further improve predictive accuracy, we implemented a comprehensive ensemble learning approach combining multiple machine learning algorithms: XGBoost, CatBoost, LightGBM, TabPFN, Neural Networks and GrowNet. This approach was designed to enhance prediction capabilities particularly for underrepresented classes in the dataset.

The ensemble framework was structured as a hierarchical prediction system:

2.4.1 Continent Classification

Each base model was trained on the relative sequence abundances to generate probability distributions across continents. To optimize ensemble performance, we implemented a

quality-based filtering criterion where only models achieving individual accuracy above 94% contributed their probability outputs to the meta-model. This threshold was established to prevent lower-performing models (e.g., those with 85% accuracy) from introducing noise. The filtered probability outputs were then fed into an XGBoost meta-learner to produce the final continent predictions.

2.4.2 City Classification

For city-level prediction, we augmented the original microbial abundance features with the probability distributions from the continent classification stage. The same ensemble methodology was applied, with a filtering threshold of 92% accuracy (selected based on benchmark performance from previous mGPS studies). The filtered model outputs were again combined using an XGBoost meta-learner to generate city probability distributions.

2.4.3 Coordinate Regression

For the final coordinate prediction stage, we used different strategies depending on the model architecture:

- **Tree-based models:** For XGBoost, CatBoost, and LightGBM, we directly predicted raw latitude and longitude values without scaling. Latitude predictions were used as additional features for longitude prediction to capture geographic relationships.
- **Neural network models:** For TabPFN and other neural network-based approaches, we transformed geographic coordinates to Cartesian coordinates to mitigate gradient vanishing issues. The following transformations were applied:

$$x = \cos(\text{lat}) \times \cos(\text{lon}) \tag{1}$$

$$y = \cos(\text{lat}) \times \sin(\text{lon}) \tag{2}$$

$$z = \sin(\text{lat}) \tag{3}$$

The final coordinate predictions were generated by combining outputs from all qualified regression models using an XGBoost meta-model. This hierarchical ensemble approach enabled more accurate location prediction by leveraging strengths of diverse algorithms while systematically filtering out poor performers at each geographic resolution level.