Estimating Population Genetic Parameters Using Neural Networks and Genetic Simulations

Jinho Park, Instructed by Nicholas Miller Ph.d

Abstract

Advancements in neural networks have opened new possibilities in DNA sequence analysis, offering innovative methods for interpreting complex genetic data. In this study, we explore the feasibility of using neural networks to predict population genetic parameters, specifically the mutation rate scaled by effective population size (θ), from pooled DNA sequences. This approach provides a cost-effective and efficient solution for monitoring genetic variation, which is crucial for understanding evolutionary dynamics and informing strategies in various biological fields. Our findings demonstrate the potential of machine learning techniques to enhance DNA sequence analysis by accurately estimating key population genetics parameters, thereby contributing to a deeper understanding of genetic diversity and evolution.

Introduction

Background

Understanding genetic diversity within populations is fundamental to the fields of evolutionary biology. Population genetic parameters, such as the mutation rate scaled by effective population size (θ), play a critical role in describing the genetic structure and evolutionary potential of species. Accurate estimation of θ allows researchers to infer historical population sizes, assess genetic variation, and predict future evolutionary trajectories.

Traditional methods for estimating θ involve extensive sampling and sequencing of individual genomes, followed by statistical analyses based on population genetics models. While effective, these approaches can be time-consuming, labor-intensive, and costly, especially when dealing with large populations or requiring frequent monitoring over time. Moreover, they may not efficiently capture the full spectrum of genetic diversity present in a population.

Recent advancements in neural networks (NNs) and machine learning have shown significant promise in analyzing complex biological data, including high-throughput DNA sequences. Neural networks are capable of modeling non-linear relationships and uncovering intricate patterns within large and complex datasets.

In the context of population genetics, NNs offer a novel approach to estimate θ by learning from simulated genetic data where population parameters are known.

By applying NNs to pooled DNA sequences—a composite sample representing the genetic material of multiple individuals—we can leverage the power of machine learning to infer population genetic parameters from aggregate data. This method has the potential to revolutionize genetic analyses by reducing the need for extensive individual sampling and sequencing. It offers a more scalable and cost-effective means of monitoring genetic variation, which is particularly beneficial for studies involving non-model organisms, endangered species, or large-scale population surveys.

The integration of machine learning techniques into population genetics can enhance our ability to detect subtle genetic changes, understand evolutionary processes, and make informed decisions

---

Objectives

The primary objective of this study is to investigate the potential of neural network models in predicting the population genetic parameter θ from pooled DNA sequences. To achieve this, we aim to:

1. Develop Neural Network Models: Construct and train neural networks using simulated genetic data with known population parameters to predict θ effectively.
2. Evaluate the Impact of Different Features: Assess how various genetic features—basic statistics, derived statistics, and their combinations—affect the performance of the neural network models in estimating θ.
3. Analyze Feature Importance: Utilize correlation matrices to understand the relationship between each feature and θ, identifying which statistics contribute most significantly to accurate predictions.
4. Address Challenges in Data Scaling and Simulation Parameters: Tackle issues related to the small magnitude and limited variability of θ values by scaling and adjusting simulation parameters to enhance model training and prediction accuracy.
5. Contribute to Advances in DNA Sequence Analysis: Demonstrate how machine learning approaches can enhance the analysis of genetic data, offering improved methods for estimating key population genetics parameters and deepening our understanding of genetic diversity and evolution.

By exploring these objectives, we aim to provide a foundation for integrating neural network methodologies into population genetics, ultimately contributing to more efficient and insightful DNA sequence analysis across various biological disciplines.

- **Primary Goals:**
    - Develop neural network models to estimate θ from simulated genetic data.
    - Explore the impact of different features and the number of datasets on model performance.
    - Analyze feature importance using correlation matrices.
- **Secondary Goals:**
    - Investigate the effect of scaling θ values on model training.
    - Adjust simulation parameters to increase the variability of θ.

Materials and Methods

Simulation of Genetic Data

We utilized **msprime**, a coalescent simulation tool, to generate synthetic genetic data. The simulations modeled sequences of a specified length (e.g., 50,000 base pairs) for a set number of diploid samples (e.g., 50 individuals). We varied key population genetic parameters, such as the effective population size (N) and the mutation rate per generation (μ), to generate a range of theta (θ) values, where $\theta = 4N\mu$.

Feature Extraction

From the simulated genetic data, we extracted various features to serve as inputs for our neural network models:

- **Basic Statistics**
    - **Number of Segregating Sites (S):** Total number of polymorphic sites.
    - **Nucleotide Diversity (π):** Average number of nucleotide differences per site.
    - **Singleton Count:** Number of variants observed only once.
    - **Allele Frequency Variance, Skewness, and Kurtosis:** Statistical measures of allele frequency distribution.
    - **Rare Variant Proportion:** Proportion of variants with minor allele frequency below a threshold (e.g., 0.05).
    - **Inter-SNP Distance Measures:** Mean and variance of distances between SNPs.

- **Derived Statistics**
    - **Tajima's D:** A summary statistic that compares π and S to detect selection.
- **Combined Features**
    - A concatenation of basic and derived statistics.

Neural Network Models

We developed several neural network architectures tailored to the features extracted in each experiment:

- **Feedforward Neural Networks**: For experiments utilizing summary statistics as input features.
- **Convolutional Neural Networks (CNNs)**: For experiments where input features had a spatial or sequential structure, such as binned allele frequency spectra.

Data Preparation

- Dataset Generation:
    - Simulated a total of 100,000 datasets (adjusted as needed for experiments).
- Data Splitting:
    - Used consistent indices to split data into training (70%), validation (15%), and test (15%) sets.
    - Ensured alignment of features and labels across all datasets.
- Feature Scaling:
    - Applied standard scaling to input features using StandardScaler when necessary.

Model Training and Evaluation

Models were trained to predict θ using the extracted features. We split the data into training, validation, and test sets to evaluate model performance. Key metrics used included Mean Squared Error (MSE), Mean Absolute Error (MAE), and visual assessments of predicted versus true θ values.

- Evaluation Metrics
    - **Mean Squared Error (MSE):**
        - Measures the average squared difference between predicted and true θ values.
    - **Mean Absolute Error (MAE):**

        - Measures the average absolute difference between predicted and true θ values.
    - **Coefficient of Determination (R^2):**
        - Indicates the proportion of variance in θ explained by the model.

Experiments

Experiment 1: Binning Approach

Objective

The primary goal of this initial experiment was to determine whether a neural network could predict the population genetic parameter θ from simulated genetic data using a straightforward approach. By applying a binning strategy to the allele frequencies across the genome, we aimed to create a simple model to assess the feasibility of our overall objective before delving deeper into more complex methods.

Methods

- Data Simulation: We simulated genetic data using msprime with varying N and μ to produce a range of θ values.
- Feature Extraction: The genome was divided into bins, and allele frequencies were categorized into frequency bins. Counts of alleles falling into each genomic and frequency bin combination were calculated, forming a 2D histogram representing the allele frequency spectrum.
- Model Architecture: A neural network model was constructed to accept the binned allele frequency data as input and predict θ.
    - Input Layer: Accepts a 2D histogram of binned allele frequencies.
    - Hidden Layers: Fully connected layers with activation functions (e.g., ReLU).
    - Output Layer: Single neuron for regression output (predicting θ).
- Features Utilized:
    - Binned Allele Frequency Spectrum: Counts of alleles in genomic bins and frequency bins, forming a 2D representation.

Results

The model demonstrated the potential to predict θ, indicating that neural networks could extract meaningful information from binned allele frequency data. However, several issues were observed:

- Increased Variance with Higher θ: As the true θ value increased, the variance in the model's predictions also increased. Predicted θ values were more widely scattered around the true values for higher θ.
- Underestimation of θ: The model tended to underestimate θ for larger true θ values. Predictions skewed towards lower values as the true θ increased, indicating a bias in the model.

These results suggested that while the binning approach provided a foundation, further refinement was necessary to improve prediction accuracy, especially for higher θ values.

Experiment 2: Impact of Fixed and Variable Parameters

Objective

We aimed to investigate whether fixing either the mutation rate (μ) or the effective population size (N) would influence the model's ability to predict θ. By comparing scenarios where one parameter was held constant while the other varied, we sought to determine if the variability of specific parameters affected model performance.

Methods

- Data Simulation:
    - Scenario 1: Fixed N with varying μ.
    - Scenario 2: Fixed μ with varying N.
    - Scenario 3: Both N and μ varied.
- Feature Extraction: Similar to Experiment 1, using the binned allele frequency approach.
- Model Training: Separate models were trained for each scenario, using consistent architectures and training procedures.
- Model Architecture:
    - Similar to Experiment 1 for consistency.
- Features Utilized:

- Scenario-Specific Data: Depending on the scenario (fixed N or μ), the simulations reflected the fixed or variable parameters.
- Same Binned Allele Frequency Spectrum as in Experiment 1.

Results

The comparison revealed no significant differences in model performance across the three scenarios. The models showed similar predictive accuracy and exhibited the same tendencies observed in Experiment 1, such as increased variance and underestimation at higher θ values.

These findings indicated that fixing either N or μ did not substantially impact the model's ability to learn and predict θ. This suggested that the neural network was primarily responding to the combined effect of N and μ encapsulated in θ, rather than the individual variability of these parameters.

Experiment 3: Binning Approach with Additional Features

Objective

Building on the initial binning approach, we aimed to enhance the model by incorporating additional features commonly used in traditional population genetics methods for estimating θ. The goal was to determine if these extra features could improve the model's predictive performance and address the issues identified in Experiment 1.

Methods

- Feature Extraction:
  - Retained the binned allele frequency data from the initial approach.
  - Added Features: Included Watterson's θ (θw) and nucleotide diversity (π) calculated per genomic bin.
- Model Architecture: Modified the neural network to accommodate the additional features, potentially increasing its capacity to learn complex patterns.
  - Modified to accommodate additional input features.
  - May include additional layers or neurons to process the extra information.
- Features Utilized:
  - Binned Allele Frequency Spectrum.
  - Additional Features:

- Watterson's θ (θw): Calculated per genomic bin.
- Nucleotide Diversity (π): Calculated per genomic bin.

Results

The inclusion of traditional population genetic features led to a slight improvement in the model's performance. There was a modest reduction in prediction variance for higher θ values, and the underestimation bias was somewhat less pronounced.

However, these improvements were limited. The core issues of increased variance and underestimation at higher θ values remained unresolved, indicating that while the additional features contributed positively, they were insufficient to fully address the model's shortcomings.

Experiment 4: Effect of Dataset Size on Model Performance

Objective

We hypothesized that increasing the size of the dataset might enhance the model's ability to learn and generalize, particularly for higher θ values where the model previously underperformed. By expanding the dataset on a logarithmic scale, we aimed to observe the impact of dataset size on the model's predictive accuracy.

Methods

- Data Simulation: Generated datasets of varying sizes, ranging from 1,000 to 100,000 simulations. Due to computational constraints, we were unable to test larger dataset sizes.
- Model Training: Trained models on each dataset size using consistent architectures and training procedures to enable fair comparisons.
- Model Architecture:
    - Consistent with previous experiments to isolate the effect of dataset size.
- Features Utilized:
    - Binned Allele Frequency Spectrum (same as in Experiment 1).
- Dataset Sizes:
    - Ranged from 1,000 to 100,000 simulations, with a logarithmic scale

Results

Increasing the dataset size resulted in slight improvements in the model's performance:

- Reduction in Variance: There was a noticeable decrease in prediction variance for higher θ values as the dataset size increased.
- Less Underestimation: The tendency to underestimate θ at higher true values was reduced but not entirely eliminated.

These results suggested that while larger datasets could enhance model performance, the improvements were incremental. The fundamental issues observed in earlier experiments persisted, indicating that simply increasing dataset size was insufficient to overcome the model's limitations.

Experiment 5: Extracting Relevant Core Features

Objective

Considering that neural networks excel at pattern recognition, we explored the idea of feeding the model with features traditionally used in population genetics for estimating θ. By focusing on core summary statistics and derived metrics, we aimed to determine if the model could better predict θ using these informative features.

Methods

- Feature Extraction:
    - Compiled a comprehensive set of basic and derived population genetic statistics, including Tajima's D, number of segregating sites, nucleotide diversity, and others.
    - Feature Selection:
    - Correlation Matrix Analysis: Calculated the correlation between each feature and θ.
    - Feature Selection Strategy: Selected features with the highest correlation to θ while avoiding multicollinearity by excluding features highly correlated with each other.
    - Model Training: Trained neural networks using the selected features.
- Model Architecture:
    - Input Layer: Accepts a feature vector of selected summary statistics.
    - Hidden Layers: Multiple dense layers with activation functions.
    - Output Layer: Single neuron for regression output.
- Features Utilized:

- Core Summary Statistics:
  - Basic Statistics: Number of segregating sites, nucleotide diversity, singleton counts, allele frequency variance, skewness, kurtosis, etc.
  - Derived Statistics: Tajima's D, and potentially others.
- Feature Selection:
  - Based on correlation analysis, selecting features most predictive of θ.

Results

Note: At the time of reporting, we are in the process of fixing issues encountered during this experiment, and complete results are not yet available.

Preliminary observations indicate that:

- The selected features showed stronger correlations with θ compared to previous feature sets.
- The expectation is that using these core features will enhance the model's ability to predict θ accurately.

Further analysis and model training are ongoing, and we anticipate that this approach may address the issues of increased variance and underestimation observed in prior experiments.

Conclusion

Through these experiments, we systematically explored different approaches to predict the population genetic parameter θ using neural networks trained on simulated genetic data. Initial models using simple binning strategies provided a foundational understanding but exhibited limitations, particularly in predicting higher θ values.

Incorporating additional features and increasing dataset sizes led to incremental improvements but did not fully resolve the core issues. Our ongoing work in Experiment 5, focusing on extracting and utilizing core population genetic features, holds promise for enhancing model performance. By leveraging features with strong predictive power and refining our models accordingly, we aim to overcome previous challenges and achieve more accurate predictions of θ.

Note: Further details on model configurations, training procedures, and complete results will be provided upon completion of ongoing experiments and analyses.