Automatic evolutionary inference using Generative Adversarial Networks

Sara Mathieson
Assistant Professor

NIH, NHGRI

Machine Learning in Genomics Workshop
April 13, 2021
Central question in population genetics: data -> quantify evolution

**INPUT**

Sites or SNPs

```
100001000100000101100001010010111000010
10001110110111111111111111110110100111
0100101010010000000000111111111110001
11011000011101100000001010001100011101
0010110110000000000110101111010100001100
0100010000000111011000011000001011111000
0000001010111110101111010101010110000100
000000010111000011000001011110111100011
101110011010101001111110110000011111111
100010101001000000011111001011111111111
000111101100000001110010110111100111000
0000000111000111000111110001111100000101
010011100001110000000001011111111111111
0100101010001000011100000001011111111111
000001010100111100000001111111111111111
```

**OUTPUT**

- Recombination hotspots
- Genomic Position
- Recombination Rate

- Population size changes
- Natural selection
- Mutation rate variation
- Migration, admixture, introgression
- Heritable traits and diseases

Images: wikipedia

Sara Mathieson
Central question in population genetics: data -> quantify evolution

- Population size changes
- Natural selection
- Mutation rate variation
- Migration, admixture, introgression
- Heritable traits and diseases
Outline

• Shift to machine learning in population genetics

• Shift away from summary statistics to “raw” data

• GANs and adversarial training
  • pg-gan algorithm for creating realistic simulated data

• Results on human data from Africa, Europe, and East Asia
Outline

• Shift to machine learning in population genetics

• Shift away from summary statistics to “raw” data

• GANs and adversarial training
  • pg-gan algorithm for creating realistic simulated data

• Results on human data from Africa, Europe, and East Asia
2013: Using machine learning to infer selection

Learning Natural Selection from the Site Frequency Spectrum
Roy Ronen, Nitin Udpa, Eran Halperin and Vineet Bavna
GENETICS September 1, 2013 vol. 195 no. 1 181-193; https://doi.org/10.1534/genetics.113.152587
2013: Using machine learning to infer selection

**Method:**
support vector machines (SVM)

Neutral regions (simulated)

Regions under selection (simulated)

Image from: “Towards Data Science”
Which summary statistics to use?

- Number of segregating sites: 3 stats
- Tajima’s $D$: 3 stats
- Folded site frequency spectrum (SFS): 150 stats
- Length distribution between segregating sites: 48 stats
- Identity-by-state (IBS) tract length distribution: 90 stats
- Linkage disequilibrium (LD) distributions: 48 stats
- Haplotype frequency statistics: 3 stats

2016: deep learning with summary statistics

Outline

• Shift to machine learning in population genetics

• **Shift away from summary statistics to “raw” data**

• GANs and adversarial training
  • pg-gan algorithm for creating realistic simulated data

• Results on human data from Africa, Europe, and East Asia
Can we do better? Convolutional neural networks (CNNs)

**Issues**

1. Image CNNs are optimized for different local features
2. For unstructured populations, sample (row) order doesn’t matter

---


2018: CNN for “raw” population genetic data

$n$ samples (exchangeable)

SNPs and Distances

```
0 1 1
1 1 0
0 0 1
0 0 1
```

SNP data
2018: CNN for “raw” population genetic data

$n$ samples (exchangeable)

SNPs and Distances

Conv + ReLU
2018: CNN for “raw” population genetic data

- Conv + ReLU
- Additional Conv + ReLU
- Permutation-Invariant Function
- SNPs and Distances
- $n$ samples (exchangeable)

Sara Mathieson
2018: CNN for “raw” population genetic data

- Permutation-Invariant Function
- Conv + ReLU
- Flatten + Fully Connected
- Additional Conv + ReLU

Inputs:
- $n$ samples (exchangeable)
- SNPs and Distances

Additional FC layers

Output: evolutionary parameter of interest
Impact of permutation-invariant architecture (recombination hotspots)

Testing Accuracy for Human Recombination Maps

- Exchangeable Architecture
- Nonexchangeable, Patch Height = 10
- Nonexchangeable, Patch Height = 3
- Nonexchangeable, Patch Height = 7
Outline

• Shift to machine learning in population genetics

• Shift away from summary statistics to “raw” data

• GANs and adversarial training
  • pg-gan algorithm for creating realistic simulated data

• Results on human data from Africa, Europe, and East Asia
Even using good simulation programs, it is difficult to match real data.

High-quality simulated data is crucial!

- Develop intuition
- Validate methods
- Provide training data for machine learning methods
- Popular simulators: SLiM, msprime

Real identity-by-descent lengths

Simulated identity-by-descent lengths

Number of SNPs for simulated/real datasets

YRI: Yoruba in Ibadan, Nigeria
Idea behind GANs (Generative Adversarial Networks)

Which is “real” and which is “fake”?

[Image of two paintings]
Idea behind GANs (Generative Adversarial Networks)

Which is “real” and which is “fake”?

https://webartacademy.com/fake-picasso
Idea behind GANs (Generative Adversarial Networks)

Which is “real” and which is “fake”?

Photo: Courtesy International Foundation for Art Research (IFAR).
Idea behind GANs (Generative Adversarial Networks)

Generator ("forger") tries to create realistic artwork

Discriminator ("art critic") tries to identify real vs. fake

GAN diagram: Adapted from Kevin McGuinness

(Source: Lost in the Louvre)
pg-gan algorithm overview

Generate parameters for an evolutionary model (e.g. $N_1$, $N_2$, $N_3$)

Simulator (i.e. msprime)

Generated Data (nxSx2 matrices)

Real Data (nxSx2 matrices)

Discriminator

Binary classification: synthetic or real

pg-gan algorithm overview

Simulator (i.e. msprime)

Generated Data (nxSx2 matrices)

0 0 1 0 1 0 1
1 1 0 1 1 0 0
0 1 0 0 0 1 0

Real Data (nxSx2 matrices)

0 1 0 0 1 1 0
1 0 0 1 0 1 0
1 1 0 1 0 1

Discriminator

Binary classification: synthetic or real

Simulated annealing algorithm
Temperature cools linearly
Change one parameter each iteration

Generate parameters for an evolutionary model (e.g. $N_1$, $N_2$, $N_3$)

pg-gan algorithm overview

Generate parameters for an evolutionary model (e.g. $N_1$, $N_2$, $N_3$)

Simulator (i.e. msprime)

Generated Data (nxSx2 matrices)

Real Data (nxSx2 matrices)

Discriminator

Binary classification: synthetic or real

pg-gan discriminator architecture: extend to multiple populations

Population 1 (e.g. YRI)
- Conv, ReLU
- Conv, ReLU
- permutation-invariant function
- flatten
- dense layers, softmax, output probabilities
- synthetic/real

Population 2 (e.g. CEU)
- Conv, ReLU
- Conv, ReLU
- permutation-invariant function
- flatten

YRI: Yoruba in Ibadan, Nigeria
CEU: Utah residents with Northern and Western European ancestry
Example of failed GAN training

- Discriminator classifies everything as real
- Generator cannot learn and reduce loss
Example of successful GAN training

Generator and discriminator are balanced

Generator not fooling discriminator

Discriminator easily able to tell training from simulated

Discriminator is often confused
Outline

• Shift to machine learning in population genetics

• Shift away from summary statistics to “raw” data

• GANs and adversarial training
  • pg–gan algorithm for creating realistic simulated data

• Results on human data from Africa, Europe, and East Asia
Sara Mathieson

**CHB: 1-param model**

- Frequency per region
- Inter-SNP distances
- LD ($r^2$)
- Distance between SNPs
- Pairwise heterozygosity ($h$)
- Number of haplotypes

**CHB: 5-param model**

- Frequency per region
- Inter-SNP distances
- LD ($r^2$)
- Distance between SNPs
- Pairwise heterozygosity ($h$)
- Number of haplotypes

**Models**

- **Single Population, growth (EXP)**
- **Isolation with Migration (IM)**
- **Out-of-Africa 2 (OOA2)**
- **Post out-of-Africa split (POST)**
Simulated data under our GAN-inferred model matches real data

Out-of-Africa 2 (OOA2)

- Single Population, growth (EXP)
- Isolation with Migration (IM)
- Post out-of-Africa split (POST)

YRI real data
YRI sim data
CEU real data
CEU sim data

N_1
N_2
N_3
T_1
T_2
mig
anc

frequency per region
inter-SNP distances
minor allele count (SFS)
LD (r^2)
distance between SNPs
Tajima's D
pairwise heterozygosity (n)
number of haplotypes
Hudson's Fst
pairwise heterozygosity (n)
number of haplotypes
• Time measured in generations

• Out-of-African bottleneck apparent
Conclusion for Machine Learning in Population Genetics

Future directions for pg-gan

- Apply to understudied populations
- Overcome data imbalance

Where are we going?

- Keep the data in mind
- ML methods need to be more interpretable
- Combine ML with evolutionary modeling
- Unsupervised learning
Thank you!

- Jeffrey Chan
- Nhung Hoang
- Paul Jenkins
- Michael Kourakos
- Hunter Lee
- Iain Mathieson
- Valerio Perrone
- Yun S. Song
- Jeffrey Spence
- Zhanpeng Wang
- Jiaping Wang
END