Clustering Population Genetics Data using an Autoencoder Architecture (Genomic Auto Encoder - GAE) as an Alternative to PCA

Students Navneet Shankar,Prathamesh Dongritot, Iffat Khatib and Tad DiDio
Promoter Dr. Yue Wang

Abstract

Analyzing population or genetic sample structure is a very important step in many genetic studies that involve human, animal, and plant populations. Currently, principal component analysis (PCA) is most widely used for dimensionality reduction and visualization. However, PCA has a number of limitations, mainly in capturing complex, non-linear patterns in genetic data. In our project, we propose leveraging deep learning to perform nonlinear dimensionality reduction on Single Nucleotide Polymorphism (SNP) data using an auto-encoder called GAE that employs a convolutional neural network (CNN) architecture similar to U-NET [5], which consists of an encoder (down-sampler) which obtains features from the input, and a decoder (up-sampler) which decodes the data and gives spatial location to the features. While originally designed for Medical image segmentation, we intend to add a number of modifications to enable it to process sequential genetic data, we hope to demonstrate that our auto-encoder identifies population clusters and provide richer visual information compared to PCA by producing its own 'space' with custom vectors or components on which the data can be projected as opposed to the Eigen vectors used in PCA, with improved computational efficiency.

Project Aims

The specific aims of our project are as given below:

1. To develop and implement an auto-encoder that effectively performs dimensionality reduction on genome sequences and allows for visualization.
2. Compare and contrast the performance and output of our method utilizing Auto-Encoders to existing traditional methods such as PCA, t-SNP, and UMAP in terms of:
    - Ability to learn complex non-linear features that are present in genetic data
    - Preservation of Global Geometry i.e spatial information regarding the relative position of data-points with respect to each other, which enables a more meaningful representation of data
3. Evaluate our model using appropriate metrics for the following:
    - Detect fine scale population structure
    - Infer informative latent factors

- Identify clusters in the sample with similar genetic composition

4. Explore further applications of the autoencoder approach:
    - Lossless compression of genomic data
    - Dimensionality reduction for downstream analysis
    - Analysis of SNP data entropy across populations

Background

There are a wide variety of reasons that it is important to look at processing SNP genetic data. Information about these structures can shed light on topics related to population demographics, evolutionary progression and adaptation, tracing ancestry lines, and individual identification to name a few [1]. This type of data is widely available due to its prevalence and longevity [1] which makes it all the more important to be able to properly interpret the results of deep learning networks which transform SNP data. This is where an auto encoder can be used to great effect; it can dynamically learn to consolidate high dimensionality data into a much lower latent space [2] while preserving nearly all of the principle data that gives these strands of Nucleotides their individual meaning. This will allow a human to analyze the data and interpret it in a meaningful way because it will be reduced to a space which the human mind is capable of perceiving.

Similar research in this area has been conducted pertaining to diseases. Some papers discuss using auto encoders to detect and classify certain complex diseases such as different types of cancer [3]. Specifically [3] uses an auto encoder to both reduce the dimensionality of the data and also automate the process of feature extraction with impressive results. Another paper discusses using a self organizing auto encoder not to perform feature selection, but instead to organize preselected features into an optimal, learned structure for classification [4]. There are a wide variety of auto encoders [2], but experiments will show us which version is the best to use in our case.

Research Design and Method

We will investigate training an auto encoder network to produce a transform matrix that can be used to map zero-shot genomic data into a visualizable space of either 2 or 3 dimensions. Our hope is that when projecting our data into the lower dimensional space using this learned transform matrix, we will be able to visualize clustering of similar genomic sequences.

We plan to take heavy inspiration from the U-NET [5] architecture as we model our own network. We will follow the sequence of first encoding and compressing data, followed by upsampling. We intend to experiment to see whether the U-NET concept of skip connections [5] will be useful in our case. We are not sampling and operating on image data, so localization for the purpose of segmentation, such as in [5] is not necessary. However, we may find that adding these skip connections or modifying them to forward data to the bottom of the 'U' shape can give the model a better understanding of the global structure of the SNP data rather than just a intricate understanding of localized features. We also plan to use attention gates to help the model focus on important regions so that it can optimally reduce the dimensionality of the input without losing excessive amounts of important data.

The output of our network should be an encoding matrix that we are able to use to project zero-shot data that the network has never seen before into a low dimensionality space to understand, interpret, and visualize emerging clusters. We plan to include a post processing step

to plot the genomic vectors in either 2 dimensional or 3 dimensional space so that a human is able to clearly view and interpret the results [9].

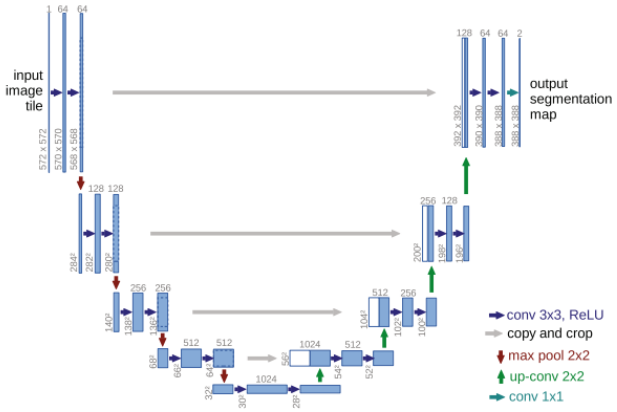

Figure 1: The U-NET architecture channels data through a pipeline which first encodes it into a minimal latent space representation for analysis, then upscales and decodes it and uses data available through the skip connections to reform the output segmentation image map.

Our network will also include a preprocessing step to format the data in a workable way. Typical SNP data is not given in a numerical format so a conversion is needed so that the network is able to process the data. We are interested in using mean encoding which is a way to center the data about the mean as determined by a target value. We will arbitrarily select a sample sequence to call the target and base the numerical representation of all other samples around this target sample. Once the data is formatted in a workable way, we will be able to feed it into our auto encoder network to begin training.

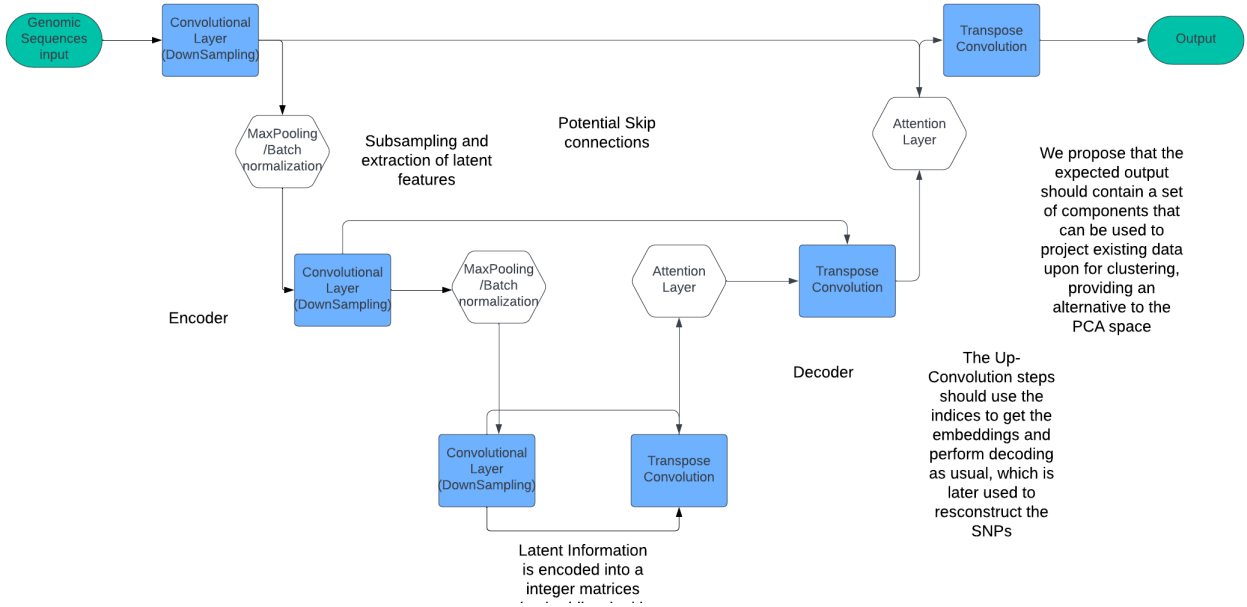

Figure 2: Our proposed architecture is constructed along the lines of U-Net , but with a number of key modifications, including the latent information being indexed , so that the decoder can use the indices to reconstruct the SNP, which could potential be a replacement for the skip connections that are generally used in U-NET. Attention gates are added as well.

Datasets and Experiments

0.1 Datasets

We primarily plan to evaluate our model on a number of datasets, these are (tentatively): The 1000 Genomes human dataset [6], which has 2504 individuals from 26 populations from all continents, The human genome diversity project [7] adding 929 samples from 54 distinct population groups, and The Simons Genome Diversity Project [8], provides genomes from 300 individuals from 142 diverse populations. We are pruning the datasets to ensure it only consists of people with a single ancestry (i.e all four grandparents reported having the same ancestry). Each sequence would contain a maternal and paternal copy (corresponding to haplogroups). The dataset after pruning is split into 3 non-overlapping groups with proportions 80%, 10% and 10%, to generate training, validation and test sets. During forward propagation we supply our model with data using Wright-Fischer online simulation , which provides new samples on the fly for each population separately. To construct whole genome sequences, we apply to techniques known as Minor Allele Frequency Filtering (MAF) where we remove SNP positions with a frequency less than 0.01, and LD Pruning, where we eliminate SNPs that have an R^2 Pearson coefficient greater than 0.01 with any other SNP in a 50-SNP sliding window.

0.2 Evaluation and Metrics

We employ a Loss function for training our model which consists of two components, the generative Loss term and the Latent Loss term. The Generative loss aims to minimize the differences between a reconstructed SNP position and the original one. Essentially we seek to maximize the probability of a reconstructed SNP position belonging to a particular distribution with a set of parameters. An SNP can be modeled as a Bernoulli distribution, therefore we seek to maximize the Likelihood function of a Bernoulli distribution [9] by minimizing cross-entropy loss. Let θ represent the parameters of our model, and let the dataset be denoted as D with D dimensions, the likelihood is given by :

$$P(D_x|\theta) = \Pi_{n=1}^{N} p(x_n|\theta)$$

assuming each dimension of the SNP is independent we can approximate the likelihood function as a Bernoulli distribution as given below:

$$P(D_x|\theta) = \Pi_{n=1}^{N} \Pi_{i=1}^{d} f_\theta(x_n)_i^{x_n} (1 - f_\theta(x_{ni})_i)^{(1-x_{ni})}$$

When we take the logarithm of the above expression we can obtain the expression for Binary Cross Entropy (BCE) loss :

$$log(p(D_x|\theta) = \Sigma_{n=1}^{N} \Sigma_{i=1}^{d} [x_{ni} log(f_\theta(x_n)_i) + (1 - x_{ni}) log(1 - f_\theta(x_n)_i))]$$

$$= \Sigma_{n=1}^{N} BCE(x_n, f_\theta)$$

In addition we also need to use some clustering metrics in order to have a solid comparison with PCA , the three metrics we could potentially use are the pseudo F statistic (also known as Calinski-Harabasz index) , the Davies-Bouldin index (DBI) , and the silhouette coefficient (SC) [9]

The Pseudo - F statistic for example , is given by :

$$CH((Y)) = BSS(Y)/WSS(Y).\alpha(Y)$$

Where BSS is Between Cluster Dispersion and WSS is Within Cluster Dispersion , both can be described as the sum L2 norms of the clusters : $\Sigma||\mu_k - x^-||^2 \quad \alpha(y) = (N - |y|)/(|Y| - 1)$, where $|Y|$ denotes the number of populations and N is the number of samples.

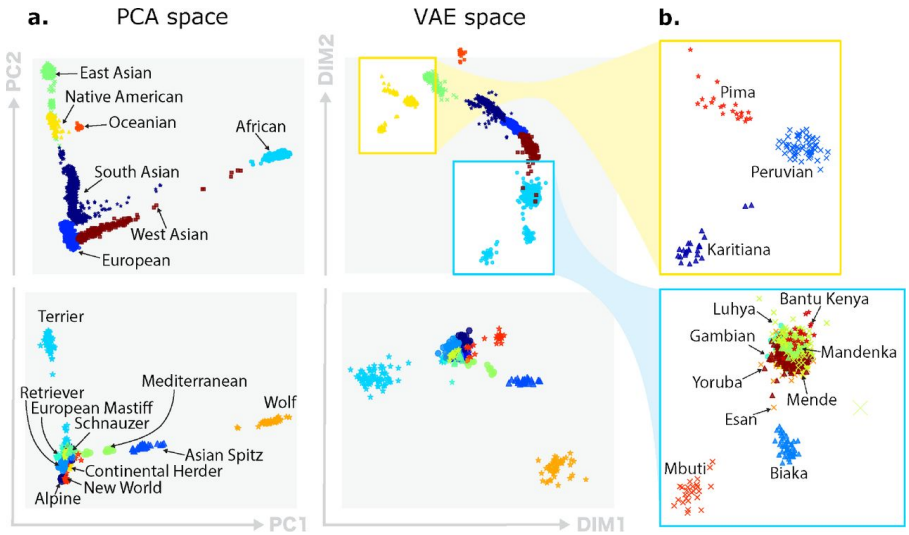

Figure 3: The difference in results between PCA and VAE spaces as obtained by [9]

Timeline

We have a lot of areas to explore with this project. In October, our main goals will be obtaining and downloading datasets and understanding their formatting. We will also map out several preliminary model versions and understand which avenues seem most promising. In November, we will implement one or more of our planned model versions and train them using the data sets listed above. For expediency, we will not perform extensive training on every model, but instead attempt to cull the total amount of training and focus on the most promising results. This month will include looking into data representation within the network, and the use of skip connections and attention gates. The end of November will be used to compare our results to existing methods as well as summarize where additional research on this topic could occur to improve it. December will be used to write the final report describing our method, results, and anything that was unexpected or has changed since this proposal.

References

[1] A. D. Leaché and J. R. Oaks, "The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics," *Annual Review of Ecology, Evolution, and Systematics*, vol. 48, no. 1, pp. 69–84, Nov. 2017, doi: https://doi.org/10.1146/annurev-ecolsys-110316-022645.

[2] J. Zhai, S. Zhang, J. Chen and Q. He, "Autoencoder and Its Various Variants," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 2018, pp. 415-419, doi: 10.1109/SMC.2018.00080. keywords: Decoding;Gallium

nitride;Mathematical model;Training;Generative adversarial networks;Data models;Computational modeling;autoencoder;decoder;deep learning;feature learning;generative model,

[3] Saeed Pirmoradi, Mohammad Teshnehlab, Nosratollah Zarghami, Arash Sharifi, A Self-organizing Deep Auto-Encoder approach for Classification of Complex Diseases using SNP Genomics Data, Applied Soft Computing, Volume 97, Part B, 2020, 106718, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2020.106718.(https://www.sciencedirect.com/science/article

[4] D. Karthika, M. Deepika, N. Radwan, and H. M. Alzoubi, "Genetic Algorithm-Based Feature Selection and Self-Organizing Auto-Encoder (Soae) for Snp Genomics Data Classifications," *Studies in Big Data*, pp. 167–181, 2024, doi: https://doi.org/10.1007/978-3-031-55221-2_10.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lecture Notes in Computer Science*, vol. 9351, pp. 234–241, 2015, doi: https://doi.org/10.1007/978-3-319-24574-4_28.

[6] The 1000 Genomes Project Consortium: A global reference for human genetic variation. Nature 526(7571), 68–74 (2015) doi:10.1038/nature15393 . Accessed 2021-05-29 (2020)

[7] Bergströmom, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al: Insights into human genetic variation and population history from 929 diverse genomes. Science 367(6484)

[8] Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al: The simons genome diversity project: 300 genomes from 142 diverse populations. Nature 538(7624), 201–206 (2016)

[9] Margarita Geleta, Daniel Mas Montserrat, Xavier Giro-i Nieto, and Alexander G. Ioannidis. Deep variational autoencoders for population genetics. bioRxiv, 2023