**Identifying population structure using a neural network**

One of the common problems in genomics that we covered in class was identifying population structure/sub-structure, and how unsupervised learning can be utilized to solve that task effectively. The approach that we took in the assignments used principal component analysis and PCA loadings to cluster genes together, and that revealed the underlying population structure hidden in the sample set. There are also other techniques for finding population structure, but we didn’t explore them much in class. One type of machine learning algorithms that is almost never used to identify population structure is an artificial neural network. Neural networks are the hype nowadays because they’ve shown record performance on certain problems such as image and speech recognition, and we’re wondering if they would perform well at identifying population structure. Finding population structure is essentially a clustering problem, and certain types of neural networks are designed for unsupervised learning for clustering purposes. One such example is a self-organizing map (SOM), which has a set of neurons connected to form topological grid (usually rectangular). When some pattern is presented to SOM, the neuron with the closest weight vector is considered a winner and its weights are adapted to pattern, as well as weights of its neighbor. This way, SOM naturally finds data clusters. Thus, we’re proposing a new approach to solve a common problem. We’d like to investigate SOMs and find out how well they perform at finding population structure. There are multiple implementations online that we can use right away, but we’d have to first read about SOMs and survey the literature to figure out why neural networks aren’t used for finding population structure.

The data we’ll use is the GEUVADIS SNP genotype dataset. Unfortunately, it doesn’t come with population structure labels, so we have no “golden standard” to judge performance against. Instead, our proposed evaluation metric is to compare the neural network’s prediction to a traditional PCA approach, and perhaps some other common approach as well. While this will give us a subjective evaluation metric (i.e. if the predictions are slightly different than the baseline approach, we can’t really tell if the NN prediction is more right or more wrong than the baseline), it would show us how close the neural network’s predictions are to the baseline approaches. Perhaps we can extrapolate a population structure from 2-3 different traditional approaches, and then compare the NN vs. PCA approaches on the extrapolated population structure. This is slightly more objective, but it’s still a rough approximation of the performance.