CSIC5011 Mini-Project 1: Ancestry Prediction via Dimensionality Reduction Techniques on SNPs Data

Ruochen MA[1], Jihong TANG[1], Yuyan RUAN[2] , Zhi HUANG[2] {rmaam, jtangbd, yruanaf, zhuangdq}@connect.ust.hk
[1]: Division of Life Science, HKUST [2]: Department of Chemical and Biological Engineering, HKUST

Introduction

Single nucleotide polymorphisms (SNPs) are variations in a single DNA building block that occur within a population. SNP data can be used to identify genetic variations that are common across different populations.

However, SNP data can be high-dimensional which can present challenges for analysis, including increased computational requirements and the risk of overfitting.

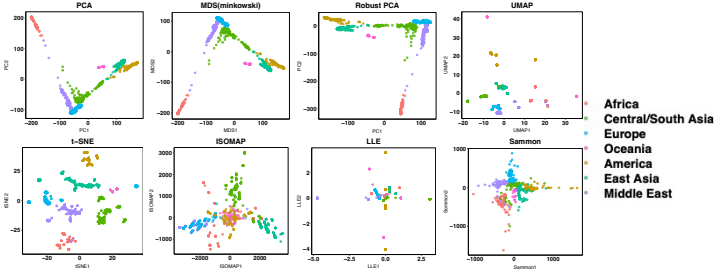
Dimensionality reduction techniques can reduce the number of features while preserving important patterns or relationships in the data. In this study, we utilized several popular dimensionality reduction techniques including PCA, MDS, t-SNE, ISOMAP, LLE, UMAP, robust PCA and random projection. We demonstrated that random projection performs the best when separating people from 7 different regions. Finally, using the exploratory dataset as a reference, we showed that the chosen method t-SNE can indeed be used for ancestry prediction in an additional 1000 Genomes Project dataset.

SNPs Dataset

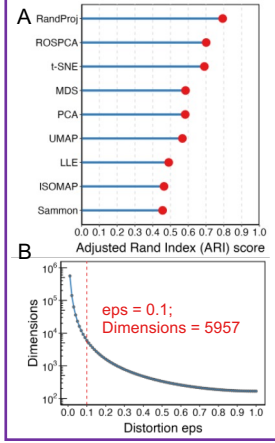
The exploratory cleaned dataset we used in this study is from Quanhua Mu and Yoonhee Nam. It incorporates 488,890 SNPs from 1043 samples, as well as the region information from patients. In the case study, we used the SNPs dataset from the 1000 Genomes Project[1], containing 11736 SNPs from 400 samples.

Results

Part1: SNPs data Dimension Reduction



We first visualized this dataset using various dimension reduction techniques. Figure1 shows that t-SNE and linear methods like PCA and MDS can well separate the whole dataset into 7 regions. However, other methods especially those based on nearest neighborhood, like ISOMAP and LLE, show relatively bad results.

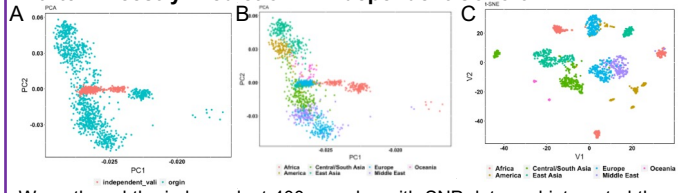


Part2: Comparison of Different Methods

We used different dimension reduction methods to reduce the data dimension, and then separated all samples into 7 clusters using kNN based on the low dimension data. Finally, different methods are compared based on the Adjusted Rand Index (ARI) score between known regions and predicted clusters. ARI[2] measures the similarity of the true labelling and the clustering labelling.

Figure 2A shows the ARI score lollipop figure. Among the compared 9 methods, random projection (RandProj) ranks top 1 with the highest ARI score. As shown in Figure 2B, the reduced dimension was set as 5957 for random projection considering the sample number and epsilon. For the left 8 methods, the clustering are all based on reduced 2 dimensions, and robust sparse PCA (ROSPCA) and t-SNE both rank top with high ARI scores.

Part3: Ancestry Prediction in Independent Cohort



We gathered the independent 400 samples with SNP data and integrated them into the original dataset. We first use PCA to check whether batch effect exists. Figure 3A shows that the batch effect isn't obvious. Figure 3B shows region annotated PCA result and the region cluster is obvious. We then employed t-SNE, which has been previously demonstrated to be effective, to cluster all the data. Figure 3C demonstrates that t-SNE can more effectively segregate the clusters. Finally, we utilized the same approach to get the ARI score in the independent cohort. The ARI score is 0.709, which shows the prediction power.

Conclusion

Dimensionality reduction techniques enable ancestry prediction from SNPs data. In this study, we utilized and compared several popular dimensionality reduction techniques. According to the ARI, we showed that random projection performs the best when separating people from 7 different regions. Finally, we demonstrated that the chosen method t-SNE can indeed be used for ancestry prediction in an additional 1000 Genomes Project dataset.

References

1. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Rand, W. M. Objective criteria for the evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**, 846–850 (1971).

Contribution

Ruochen MA: Part2 PCA, ISOMAP, t-SNE, MDS and Poster
Jihong TANG: Part2 RandProj, ROSPCA, LLE, UMAP, Sammon, and Poster
Yuyan RUAN: Part1 PCA, MDS, ISOMAP, t-SNE,UMAP, Sammon, and Poster
Zhi Huang: Part3 Extra dataset pre-process, PCA, T-SNE, K-means predicition