Analysis of Gene Expression Using PCA

Sarah Thuman and Mitchell Friess

Abstract

Principal component analysis is a common method used in genome analysis. PCA reduces the dimensionality of large data sets while preserving the maximum amount of information. This allows for the identification of differences and similarities in gene expression across different genomes, either across species or different members of the same species. In this paper we focused on comparing differences in gene expression between different species using PCA for analysis.

Introduction

PCA has a wide range of applications including image compression, facial recognition, and quantitative finance. It has also become a commonly used technique in genome analysis in part due to its ability to preserve information. A few applications of PCA in genetics are detection of weak expression patterns, detection and quantifying genetic structures of populations, and gene expression. For this reason, we have chosen PCA for comparing gene expression across species. Through our analysis we hope to show that the major contributors to gene expression varied across species. To this end, we chose different species from different orders to compare.

We will discuss how the experiment was set up, overview of species selected, how the data was pre-processed, and how the data was analyzed using PCA. In the observations section, we will describe the results of the experiment. The conclusion discusses the results along with possible improvements that could be made to the experiment.

Methods

We selected nine species' genomes to analyze ranging from a virus genome to a fungi genome. The complete list of species is given in appendix A. Due to time restrictions and computing power, we selected species with genomes with storage size of 15 MB or lower with one exception, Chlorella sorokiniana, with a genome storage size of about 39 MB. We also only selected species with completed genomes. One other restriction we imposed on our analysis was that we only analyzed one chromosome from the larger genomes, also due to time and computing power restrictions.

The first step in our analysis was to convert the strings of nucleotides, A, T, C, and G, into a numerical representation that could be used in PCA computations. This was done by searching for a start codon, ATG, and a stop codon, either TAG, TAA, or TGA, to separate the different genes within the genome. Once the genes were found, they were parsed in groups of three nucleotides to identify the amino acids that comprised the genes. Each amino acid was given a weight based on its prevalence in the gene. This data was put into a gene matrix with the

rows representing genes, the columns representing amino acids, and the elements representing the proportion of the gene comprised of each amino acid.

Once the data was processed, PCA was performed on the gene matrix. The principal components were sorted from the largest eigenvalue to the smallest eigenvalue and then plotted in a scree plot and biplot. Figures 1 and 2 show one of the scree plots and biplots, respectively. For the biplots, only the top three principal components were included.

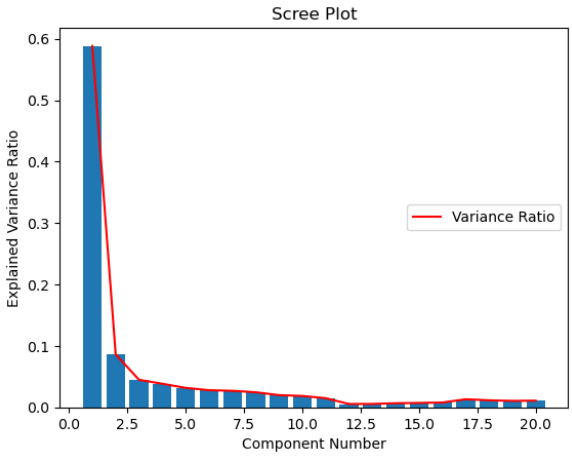

Figure 1: Biplot of Chlorella sorokiniana

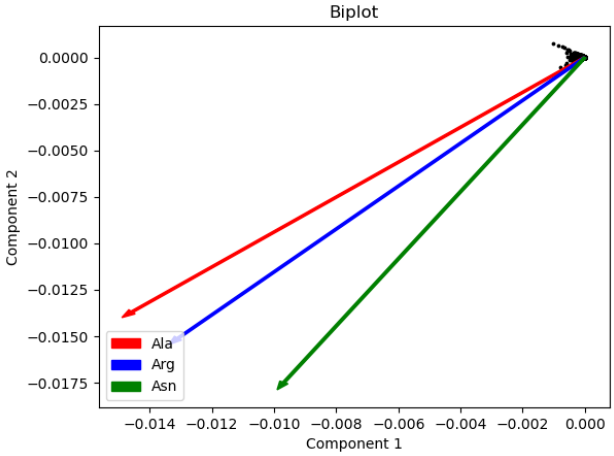

Figure 2: Scree plot of Chlorella sorokiniana

Observations

One observation from the analysis is that the amino acids that consistently accounted for the most variance in the genomes were Alanine, Arginine, and Asparagine. While there was little change in the order of the top three amino acids, the order of the amino acids explaining the remaining variance varied from species to species. The other major observation was that Alanine accounted for about 80% (see Figure 3) of the variance in every species. The one exception to this was Chlorella sorokiniana, where it explained about 60% of the variance (see Figure 1). The difference in explained variance between Chlorella sorokiniana and the other species could be due to the fact that the Chlorella sorokiniana genome was the largest genome we examined and the fact that we only performed PCA on one chromosome.

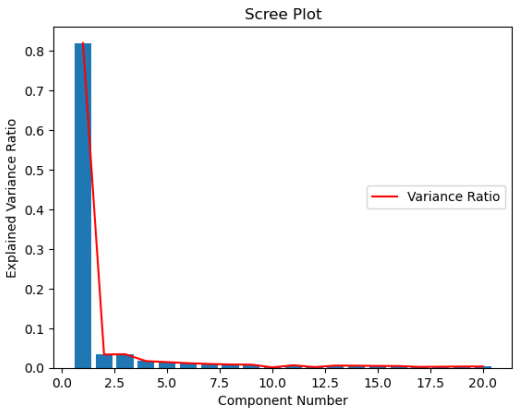

Figure 3: Scree plot of Catharanthus roseus

Conclusion

In this experiment we hoped to show different amino acids affected gene expression differently in different species using PCA. Our analysis showed that about 85% of the variance in gene expression can be explained by the same three amino acids, Alanine, Arginine, and

Asparagine. This was consistent across the nine species' genomes we analyzed. Out of these three amino acids, Alanine explained the majority of the variance in every case.

A few factors that may have affected our analysis and skewed the results are the sample size, the sizes of the genomes, and limiting the number of chromosomes that were analyzed. For our analysis we only compared the genomes of nine different species which may not have been a large enough sample population for difference to be expressed. Also, while the species were chosen at random, they were randomly selected from the fist page of a list of different species genomes that had thousands of entries. It is entirely possible, while the species selected all expressed their genes in similar ways, that they may not be an adequate representation of the entire population. The fact that we selected genomes that were comparatively small could also have contributed to the consistency in gene expression that we saw in our experiment. It is of note, even though Chlorella sorokiniana had the same three amino acids explain the majority of the variance as the other species, the variance explained by Alanine was about 25% lower in Chlorella sorokiniana than other species. Given that Chlorella sorokiniana was the largest genome analyzed, this may indicate that in more complex species, different amino acids would explain the variance in gene expression. Further research would need to be done to evaluate this. The other factor that might have skewed the results is that only the first chromosome in the more genetically complex species was analyzed. It is possible that analyzing every chromosome would result in different gene expressions.

Finally, we assumed that every start codon represented the start of a gene within the genome. Gene expression is a heavily regulated process and so this is not an accurate representation of what would happen within a cell. Accurate gene expression would require laboratory analysis. Combination of the PCA method described and a more accurate representation of the species gene expression would lead to better analysis of the genome.

Appendix A: List of species

1. Catharanthus roseus – bacteria
2. Echinacea purpurea – bacteria
3. Acetobacter ascendens – bacteria
4. Acidianus ambivalens – sulfolobales
5. Brettanomyces nanus – fungi
6. Clavispora lusitaniae – fungi
7. Chlorella sorokiniana – green algae
8. Cryptosporidium parvum – apicomplexan
9. Acanthamoeba polyphaga mimivirus – virus