"Deciphering Genetic Codes Across Species: A Machine Learning Approach with k-Means Clustering of DNA Sequences"

Deborah L. Young

DSC 680: Applied Data Science

Business Problem

Genomic sequencing is a burgeoning field, amassing nearly inconceivable amounts of data. These massive sets of information require intensive sorting to facilitate understanding. Machine learning serves as an excellent tool for this purpose. Machine learning methods are increasingly valuable in classifying DNA sequences, so we will attempt to do this across these distinct but related species (humans, chimpanzees, and dogs). In this research project, k-means clustering will be the method attempted in effort to arrange values into groups that can then be analyzed together. The goal is to identify patterns and parallels among nucleotide sequences, or genes. Because of the computational requirement for processing genetic sequences, this project elects to use sample sequences (as opposed to full genomes) to test various models' efficacy in classification, in hopes that its outcomes could inform broader genetic studies.

Background/History

The author has an ongoing interest in genetic sequencing and a moderate background in biology, microbiome health, and technical science. While her experience with DNA manipulation is limited, she has a foundational understanding and recognizes that this field – although very popular – is still widely unexplored due to its massive volume of data and complexity in processing. Genomic understanding stands to offer a wide array applications throughout human health and environmental sustainability.

Genomic sequencing technology has rapidly improved in recent years, leading to a notable expansion of DNA data availability. With so much information available, processing it quickly and effectively is tantamount in the field. Machine learning has the potential to be a powerful method in refining understanding about genes, their interactions, their effect on bodily mechanisms, their evolutionary implications, and more. It is currently applied via numerous approaches for DNA sequencing, including sequence alignment, classification, clustering, and pattern mining. Due to the author's novice in genomics, this project will attempt a straightforward k-means clustering approach.

Data Explanation

We are looking at the DNA sequences, made up of nucleotides (ACGT), that are essentially a language for directing the actions of the body. The sections (subsequences) of the language encode genes and families of genes. To interpret this language, it seems like NLP (natural language processing) would be an appropriate machine learning method, due to its ability to process to language.

NLP techniques can then be applied to identify patterns within k-mers, such as frequency counts, which can be analogous to word frequencies in text analysis. By using NLP methods like tokenization (segmenting sequences into k-mers), vectorization (converting k-mers into numerical format for machine learning), and classification algorithms, we can process and analyze genetic information to detect similarities, variations, and mutations across different organisms.

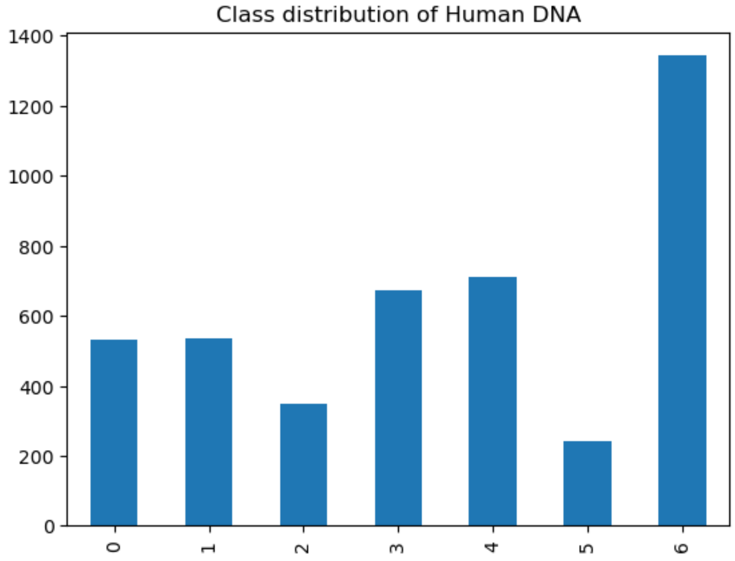
This project utilized the included datasets from a Kaggle notebook ("*Classifying DNA Sequence using ML"*) to save on computational space. To use full genome sequences, one would need to have considerably more power and space than a personal computer holds. For reference, the human genome consists of around 6 billion characters *("DNA Sequencing with Machine Learning"),* while this set is merely 5.5 million. These datasets include one set of sequences from each specie - human, chimpanzee, and dog. Insights into this process were garnered from several websites and white papers to support understanding of genomics, bioinformatics, and modeling, which are listed in the references section.

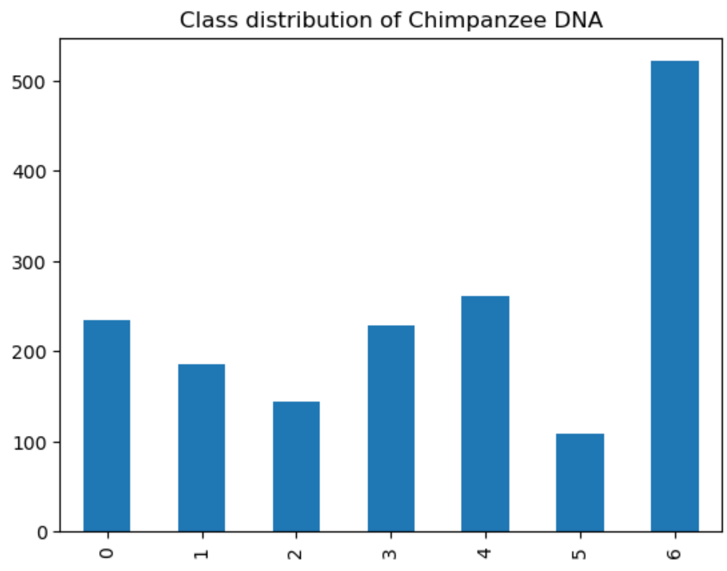
Methodology

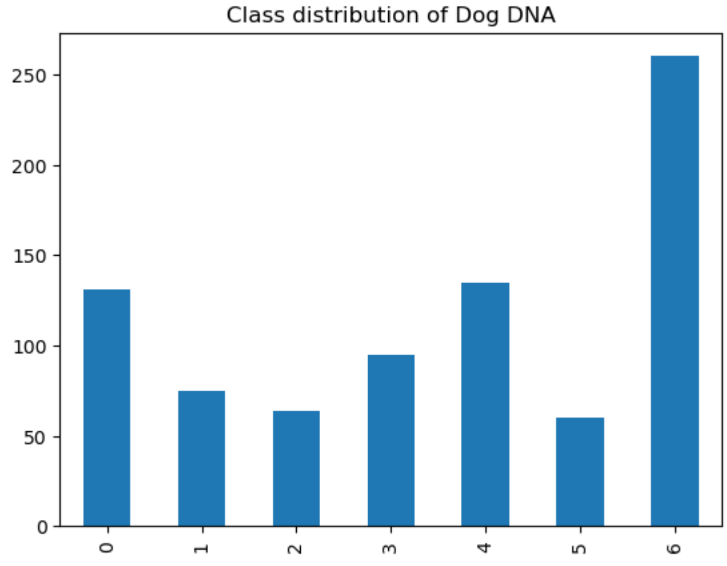
The methodology traverses from data importation and class distribution checks to k-mer conversion using Python's Pandas library. The pivotal k-mer counting converts DNA sequences into overlapping segments that are then vectorized into a numerical matrix—a critical step for subsequent machine learning. The modeling employed k-means and MiniBatchKMeans algorithms, refined through the elbow method for optimal cluster determination. The project aimed for visualization and model evaluation, which proved challenging.

Data Preprocessing:

1. Import datasets with Pandas
2. Review distribution of classes among datasets







There are similar distributions of classes so we will continue on with analysis.

3. Convert sequences into k-mers

- A function (Kmers_funct) is applied to sequences of human, chimpanzee, and dog DNA to create overlapping k-mers. A k-mer is a substring of length 'k' derived from a sequence, and in this case, it appears to be of length 6. For example, if we use "words" with a length of 6 (hexamers), "ACTCGAGTCA" becomes: 'ACTCGA', 'CTCGAG', 'CGAGTC', 'GGAGTCA'. Therefore, the example sequence is broken down into 4 hexamer words. These k-mers are essential in bioinformatics for comparing regions of genomes and understanding genetic similarities and differences

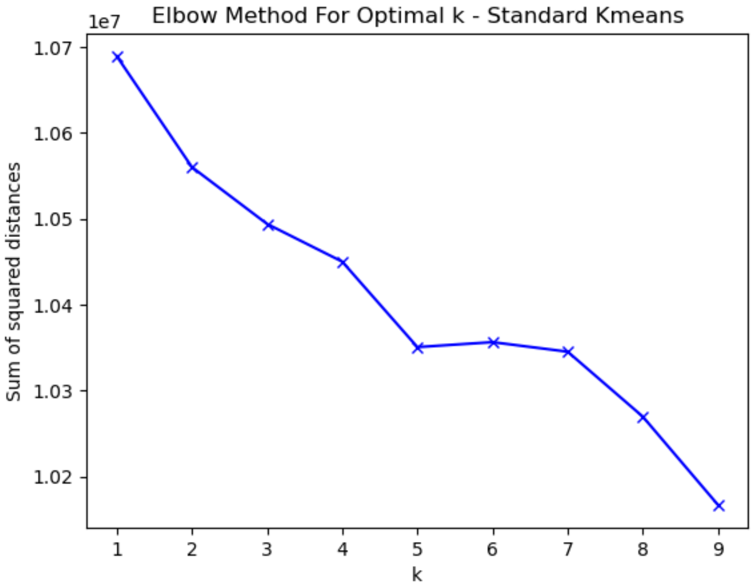
- The k-mers are then used to generate a list of strings for each species, which are likely used to represent DNA sequences as sentences where each word is a k-mer. This is a common technique in bioinformatics to transform sequence data into a format that can be used for further analysis, such as machine learning models

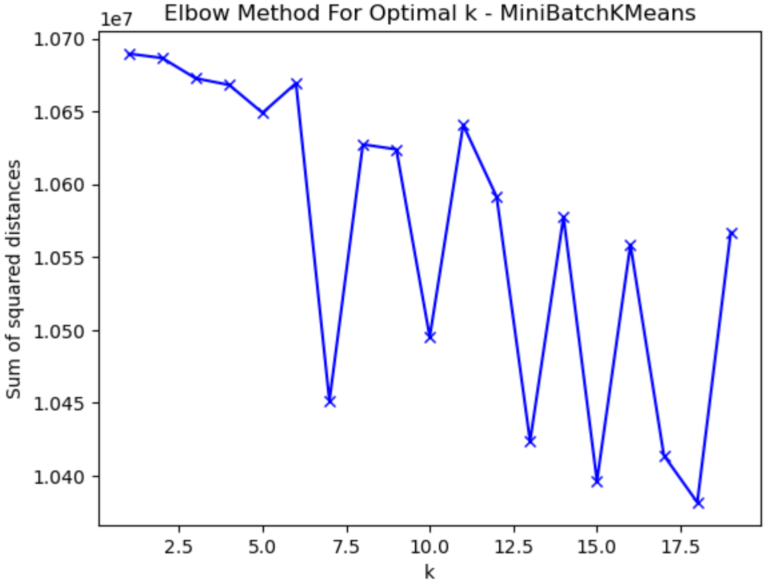
Data Tranformation:

4. A Bag of Words model is created using CountVectorizer, which is a method to convert a collection of text documents to a matrix of token counts. Here, it's used to analyze the k-mers. The ngram_range is set to (4,4), meaning that the model is looking at 4-gram k-mers (subsequences of 4 k-mers).
5. Sparse matrices for human, chimpanzee, and dog DNA sequences are generated. These matrices are efficient ways to store data when most of the elements are zero, which is common in text vectorization when many possible k-mers are not present in a given sequence. The matrices contain information about the frequency of each k-mer in the sequences.

Modeling:

6. Perform k-means clustering utilizing Kmeans and MiniBatchKmeans
    - Determine best k value through elbow method

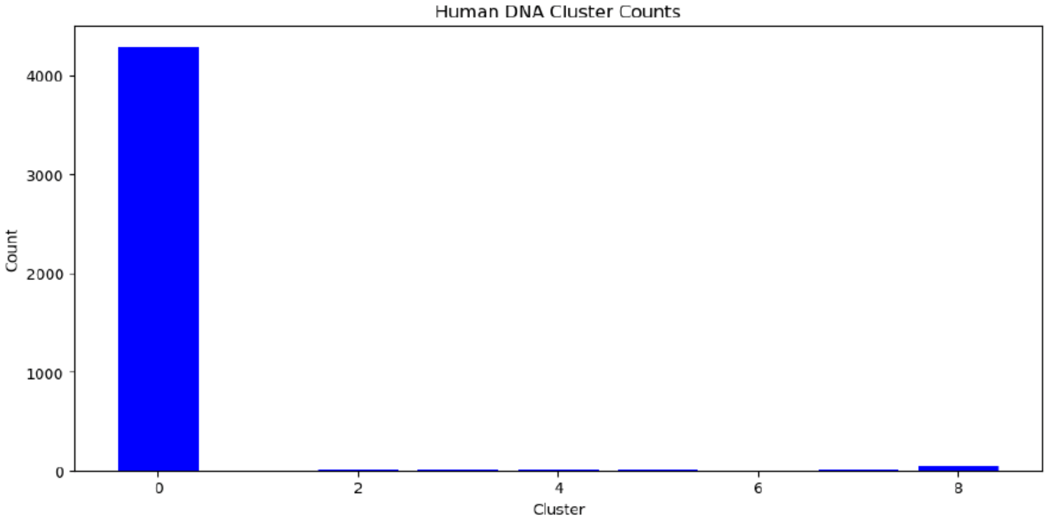


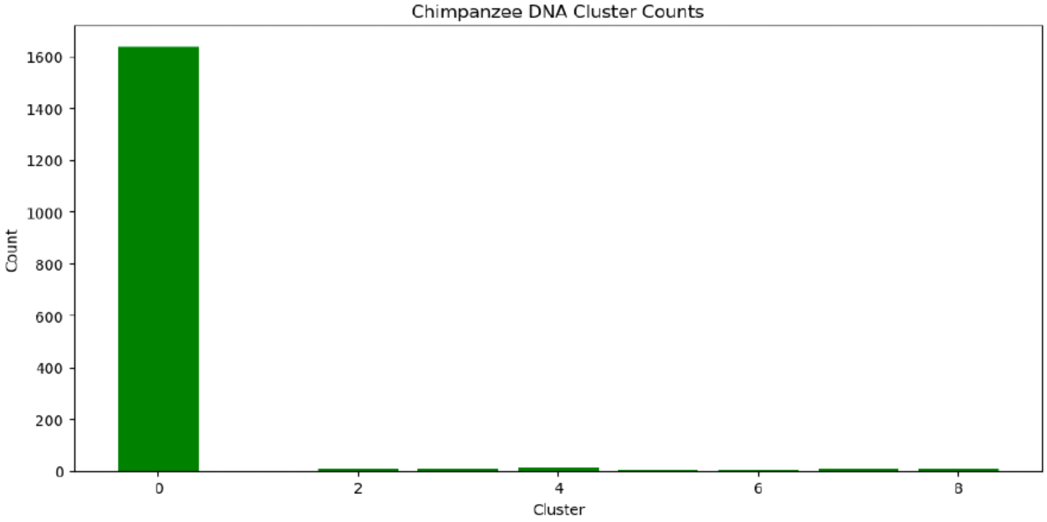


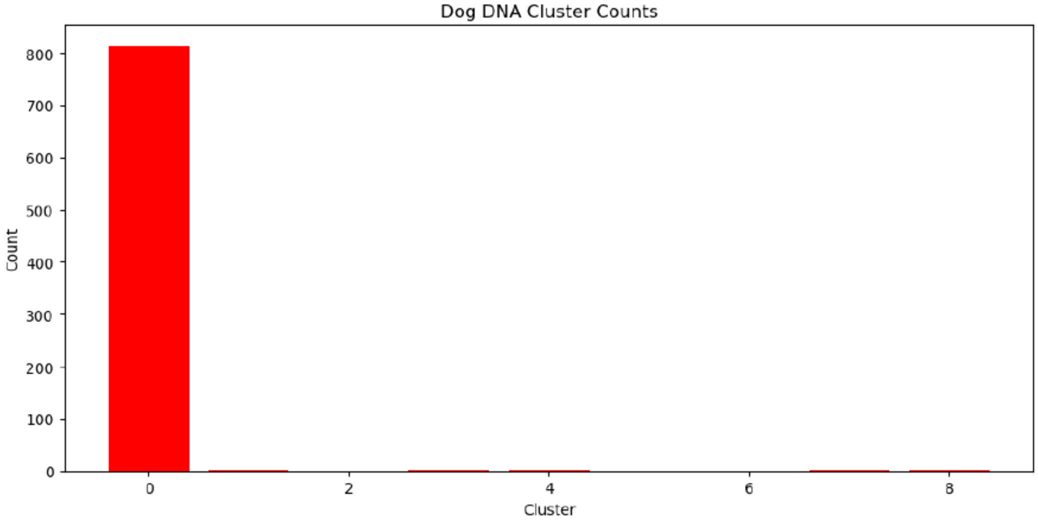
- Perform **KMeans** clustering with k=9 and **MiniBatchKmeans** with k=9 (both use random state = 42 for reproducibility)

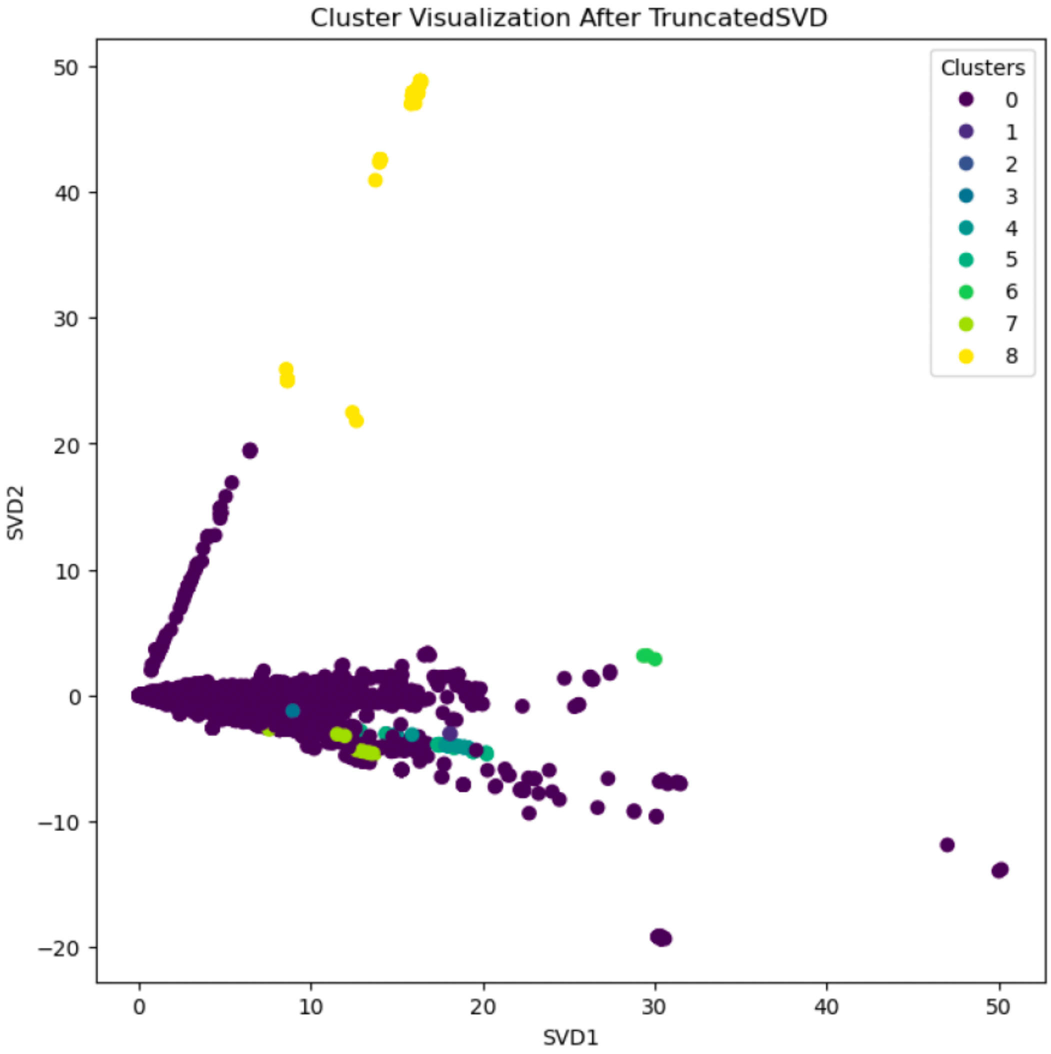
7. Perform **TruncatedSVD** for cluster visualization.
8. Visualize

   - Standard Kmeans:

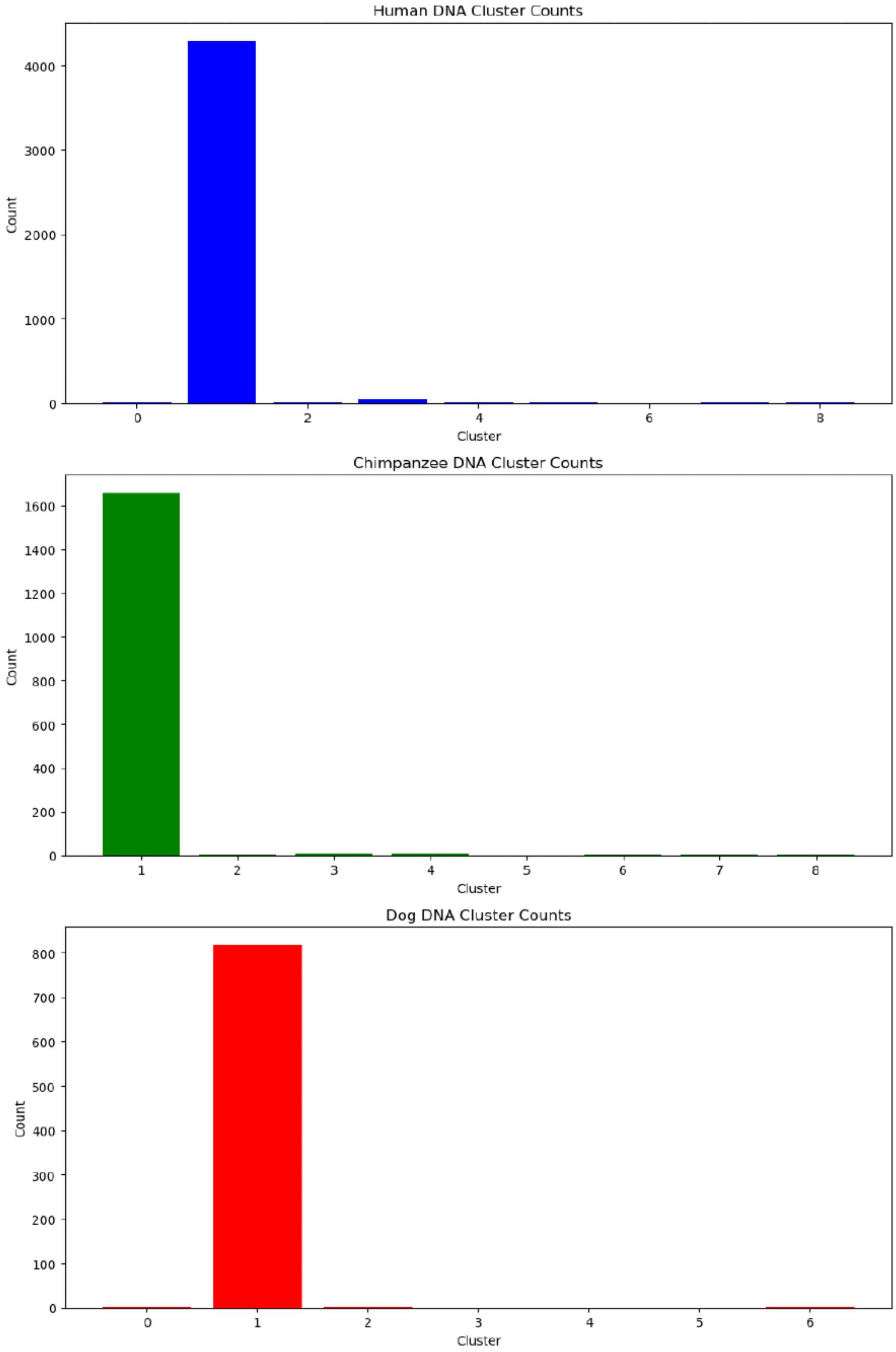

Human DNA Cluster Counts
Count
0
1000
2000
3000
4000
Cluster
0
2
4
6
8

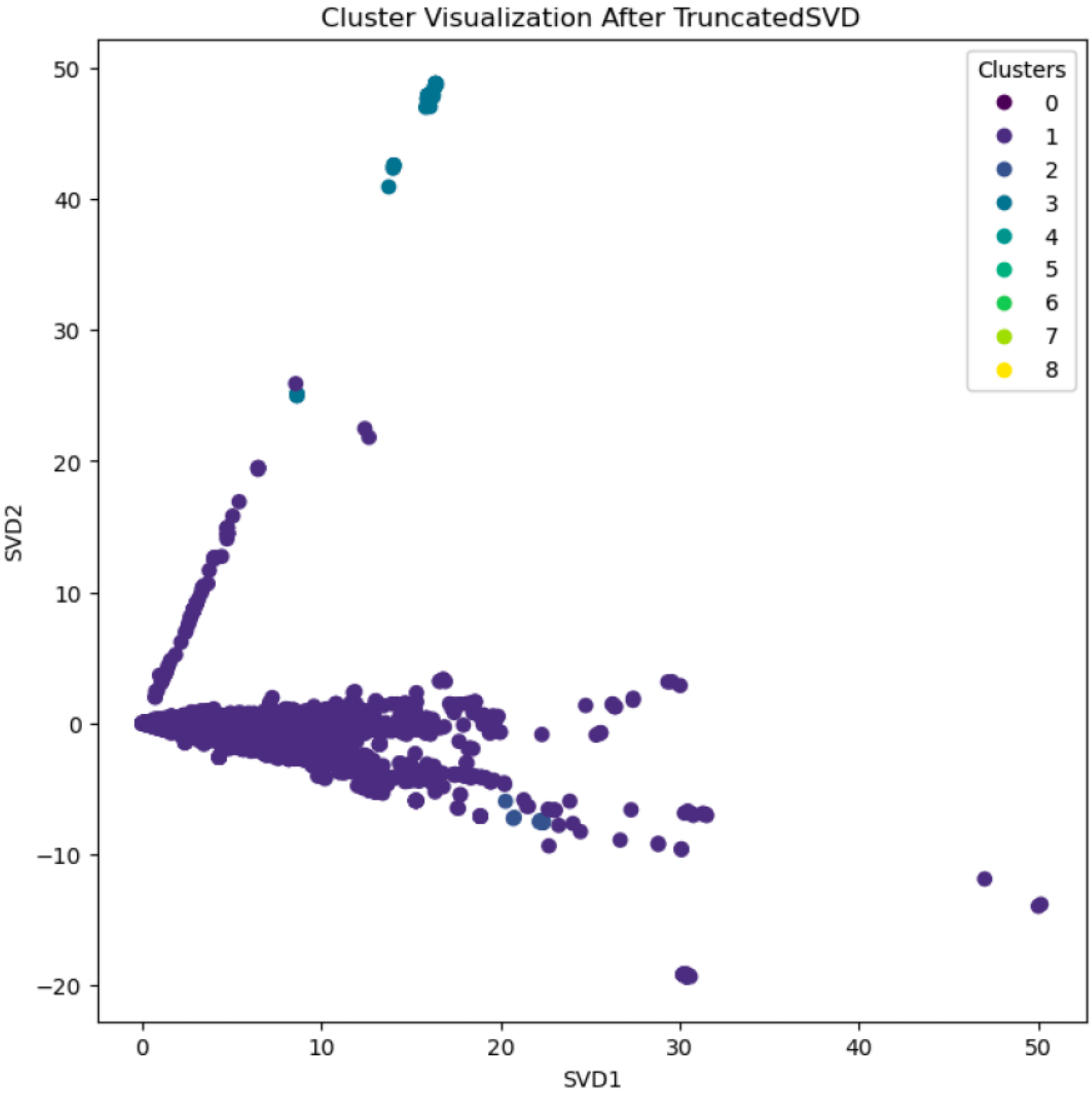

Chimpanzee DNA Cluster Counts
Count
0
200
400
600
800
1000
1200
1400
1600
Cluster
0
2
4
6
8


Dog DNA Cluster Counts
Count
0
100
200
300
400
500
600
700
800
Cluster
0
2
4
6
8



- MiniBatchKmeans:


Human DNA Cluster Counts
Count
4000
3000
2000
1000
0
0
2
4
6
8
Cluster
Chimpanzee DNA Cluster Counts
Count
1600
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
Cluster
Dog DNA Cluster Counts
Count
800
700
600
500
400
300
200
100
0
0
1
2
3
4
5
6
Cluster



Analysis

Upon examining the scatter plots generated from the singular value decomposition (SVD) of the genetic sequences for humans, dogs, and chimps, we observe distinct clustering patterns that reveal insights into the genetic similarities and differences among these species.

Human Genetic Clustering: The human genetic data shows a dominant cluster, which we've labeled as Cluster 1, encompassing a significant majority of the sequences with 4298 instances. This suggests a high degree of conservation in certain genetic regions within the human samples. The presence of smaller clusters and isolated points indicates variability in other genetic regions, possibly corresponding to less conserved sequences or those that are subject to greater evolutionary pressures.

Chimpanzee Genetic Clustering: The chimpanzee data also exhibits a primary cluster, with Cluster 1 containing 1646 sequences. The distribution and count of the clusters are less dense compared to humans, reflecting the genetic diversity present within the chimpanzee sequences. The scattered nature of the smaller clusters suggests a variety of genetic functions or differences within the chimp genome.

Dog Genetic Clustering: The canine sequences demonstrate a very distinct pattern, with a primary cluster (Cluster 1 with 811 sequences) and several minor clusters or single instances. The spread of clusters is indicative of the genetic diversity within the dog population, possibly due to the extensive breeding and variation seen in domesticated dogs.

The clustering patterns provide a visual representation of the genetic relationships within and between species. It is evident from the plots that there is a significant overlap in genetic sequences between humans and chimps, which is to be expected given their close evolutionary relationship. The dog sequences, while distinct, still show a pattern of clustering, which might correspond to conserved sequences that are essential across mammalian species.

Conclusion

Overall, the clustering analysis has the potential to identify regions of genetic similarity and divergence, which can further our understanding of genetic evolution and the underlying biological functions of these sequences. Future analysis could involve annotating these clusters with biological data to ascertain the functional similarities within clusters and the evolutionary significance of the sequences that fall outside of the primary clusters.

This project's approach, at the intersection of genomics and computational linguistics, underscores the innovative potential of NLP in biological sciences. By adapting language processing techniques to genetic data, we open a new avenue for analyzing biological information, which could have far-reaching implications for medical research, evolutionary biology, and beyond. Nevertheless, we must proceed with an understanding of the limitations posed by computational capabilities, novice capabilities of the author, and the representativeness of our data samples.

Assumptions

This analysis is based on the assumptions that:

- Genetic sequences (nucleotides consisting of ACTG) can be broken down into over-lapping by k-mers to be converted and vectorized for running through a clustering model, such as k-means, to demonstrate potential similarities.
- The bag-of words model was an appropriate transformation technique. The values retrieved by this method are assumed to be representative of the data.
- SVD was used to reduce the dimensionality of a similarity matrix derived from genetic sequences, and that the clusters represent groups of similar sequences. The counts of sequences in each cluster suggest the prevalence of certain genetic motifs within the population. However, to provide a more detailed analysis, additional biological context would be required.

Limitations

The use of full genomes for comparative analysis necessitates significantly greater computational power and storage capacity than currently available. This study, therefore, relies on a reduced dataset comprising sample sequences, which, while informative, may limit the extrapolation of findings to entire genomes.

The initial analysis suggests that the preprocessing step could be influencing the clustering outcomes significantly. Moreover, the characteristics of our dataset may necessitate the use of alternative clustering algorithms that do not impose the same assumptions as k-means and MiniBatchKMeans. Further investigations will include revising the preprocessing pipeline and experimenting with other clustering techniques to find the most representative clustering solution for our data.

This project navigates a unique form of data—sequences represented by letters, requiring meticulous encoding before analysis to perform properly with machine learning models, which are traditionally designed for numeric or categorical data. Marrying the intricacies of biology with the precision of machine learning into an interesting and intelligible presentation demands intentional and creative approaches. This is an ambitious proposal for this project, yet, despite its limitations, it has the potential to serve as valuable experience for the author and could benefit the greater understanding of fellow scholars interested in this exciting intersection of fields.

Challenges:

As previously discussed, to perform a comparative analysis like this using full genomes, computational capacity would need to be much more robust. The data used in this project is limited and, due to its nature, needs to be encoded in a specific manner, then converted again. The practice of feature selection, conversion, and vectorization leaves a lot of room for error.

The initial analysis suggests that the preprocessing step could be influencing the clustering outcomes significantly. Moreover, the characteristics of our dataset may necessitate the use of alternative clustering algorithms that do not impose the same assumptions as k-means and MiniBatchKMeans. Further investigations will include revising the preprocessing pipeline and experimenting with other clustering techniques to find the most representative clustering solution for our data.

Future Uses/Additional Applications

Machine learning has the potential to be a powerful method in refining understanding about genes, their interactions, their effect on bodily mechanisms, their evolutionary implications, and more. These are the building blocks for breakthroughs in health, wellness, biotechnology, and possibly even immortality. If this method is proven to be accurate and reproducible, it could be applied to other animals, plants, bacteria, and fungi. These applications could guide environmental conservation, sustainability, and technology.

Recommendations

Ideally, this method would be reproduced with larger datasets (full genomes) and data sourced from more subjects. As it is refined, it could be useful to compare clustering of genotypical sequences against abnormal sequences to see where they vary. It could also be useful to compare them with other animals. For each attempt, feature selection, conversion, and vectorization could be refined. Being able to visualize a fuller assessment of the clustering is also recommended. Having a subject matter expert present during these explorations and experiments would also be ideal.

Implementation Plan

While the project was able to achieve its objectives within the scope of the provided datasets, the findings are preliminary. To expand on this research, one would need to consider more sophisticated bioinformatics techniques, access to full genomes, and the computational resources to process them. Furthermore, collaboration with geneticists would be invaluable in interpreting the biological significance of the clusters identified by the machine learning models.

Ethical Assessment

The ethical implications of genomic research are expansive and multifaceted, most notably in the privacy of data. The sources of the DNA sequences are anonymized, which is imperative for ensuring rights of subjects and protecting against the misuse of genetic information. Anonymizing the data removes any personal identifiers so that sensitive information cannot be traced back to a subject, which could be used to determine the identity of and/or discriminate against a subject.

All collaborators must do their best to be mindful of any regulatory standards and procedures that would be present in an actual bioinformatics setting. The design of this project must include foresight about the engagement this research would have among subject-matter experts, policymakers, subjects, and the public.

Transparency is tantamount in any research project, but especially when the outcome of the research may be used to influence human health. Detailed documentation is essential for reproducibility as well as peer review.

The goal of this project is to contribute positively to genomic research, so the author and future collaborators must uphold their responsibility to ethically use the information and tools at our fingertips.

References

A Comprehensive Introduction To Your Genome With the SciPy Stack | Toptal®. (n.d.). Toptal Engineering Blog. Retrieved December 2, 2023, from https://www.toptal.com/python/comprehensive-introduction-your-genome-scipy

Classifying DNA Sequence using ML. (n.d.). Retrieved December 3, 2023, from https://kaggle.com/code/tarunsolanki/classifying-dna-sequence-using-ml.

DNA Sequencing with Machine Learning. (n.d.). Retrieved December 2, 2023, from https://kaggle.com/code/singhakash/dna-sequencing-with-machine-learning.

Edwards, D. J., & Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microbial Informatics and Experimentation, 3(1), 2. https://doi.org/10.1186/2042-5783-3-2

FASTA Format for Nucleotide Sequences. (n.d.). Retrieved December 3, 2023, from https://www.ncbi.nlm.nih.gov/genbank/fastaformat/

OpenAI. (2024). ChatGPT [Large language model]. https://chat.openai.com

Sriram, R. (2021). How data science is driving genomics in the pharmaceutical industry. https://www.drugtargetreview.com/article/79054/how-data-science-is-driving-genomics-in-the-pharmaceutical-industry/

Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., & Zhang, L. (2020). Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Frontiers in Bioengineering and Biotechnology*, *8*. https://www.frontiersin.org/articles/10.3389/fbioe.2020.01032

10 Questions an Audience May Have

1. *Will you explain more about Natural Language Processing and its mechanisms?*

   NLP is a field at the intersection of computer science, artificial intelligence, and linguistics. It involves programming computers to process and analyze large amounts of natural language data. The goal is to understand and generate human language in a valuable way. Mechanisms of NLP include syntactic analysis, semantic analysis, and, more recently, machine learning models that can capture the subtleties of language through large datasets and complex algorithms.

2. *Could you elaborate on the implications of the skewed data observed in your analysis?*

   Skewed data can lead to biased or inaccurate models since machine learning algorithms assume that the patterns they learn from the training dataset apply to unseen data. If the training data is not representative, the model's predictions could be systematically off-target, especially for underrepresented groups within the data. Addressing skewed data often involves techniques such as resampling, reweighting, or collecting more balanced data.

3. *What approaches could improve the efficacy of this methodology?*

   To improve the efficacy of the methodology, one could employ more sophisticated data preprocessing, feature selection, and normalization techniques. Enhancing the dataset size and quality, using advanced algorithms, and incorporating domain expertise into the model training process could also lead to improvements.

4. *Why are the assumptions listed critical for the methodology of this project?*

   Assumptions are critical because they set the stage for how the methodology is applied and interpreted. They often include presumptions about data distribution, relevance, and noise levels. Assumptions must be clear to ensure that conclusions drawn are valid and that the methodology can be appropriately adapted or replicated in future studies.

5. *How might the findings of successful k-means clustering contribute to the existing body of knowledge in genomic research and machine learning?*

   Successful k-means clustering can reveal natural groupings or patterns in genomic data that may not be apparent through human analysis alone. This can contribute to a more nuanced understanding of genetic structures and relationships, aiding in the identification of genetic markers for diseases or traits and refining machine learning models in bioinformatics.

6. *Has this methodology been implemented on a larger scale in other studies?*

   Machine learning methods, including clustering algorithms, are commonly used in large-scale genomic studies due to their ability to handle vast datasets and uncover complex patterns.

7. **What other types of clustering methods would be applicable to this data?**

   Besides k-means may still be appropriate, but other clustering methods that could be applied to genomic data include hierarchical clustering, DBSCAN, and Gaussian mixture models. Each has its strengths and can reveal different aspects of the data depending on the underlying distribution and the research question at hand.

8. *What additional machine learning methods could be attempted with this data?*

   Beyond clustering, other machine learning techniques such as supervised learning models (e.g., support vector machines, random forests), neural networks, and dimensionality reduction methods (e.g., PCA, t-SNE) could be applied to discover predictive patterns or reduce the complexity of genomic data.

9. *What are the potential real-world applications of this research on human health and environmental sustainability?*

   This research can have several applications in human health, such as personalized medicine, where genomic data helps tailor treatments to individual genetic profiles. In environmental sustainability, understanding the genetic factors that contribute to species adaptation can inform conservation strategies and agricultural practices.

10. *What are the broader ethical implications of using machine learning in genomic sequencing, particularly concerning data privacy?*

    Using machine learning in genomic sequencing raises important ethical concerns, particularly around data privacy and consent, as genetic data is highly sensitive and personal. There's also the risk of misuse of genetic information and the need for equitable access to the benefits of such research. Ethical frameworks and regulations need to be in place to guide the responsible use of these technologies.