DNA Genome Analysis with PCA and K-Means

Suzanne Pittman, Ph.D.

June 18, 2018

I Introduction

Modern DNA sequencing has brought forth an the extensive amount of genomic data. Computational methods, such as FASTA [8], BLAST [1], Hidden Markov Models[3, 2], Interpolated Markov Models [9] and Information Theory [4, 7, 6, 5], provide efficient and reliable means towards analyzing these data sets and gene prediction. Whereas the BLAST and FASTA algorithms can identify new genes with similar homology to known genes, the statistical methods are able to determine genes with unknown homology, without context. Statistical methods are particular useful for prokaryotic genomes, which are typically compact (10-20% noncoding DNA). Finding the correct phase-shift is often challenging for these organisms.

In this project, principle component analysis (PCA) and K-Means cluster techniques are used to locate the genes within a DNA segment of the bacteria Pseudomonas Aeruginosa. Additionally, a seven cluster structure is revealed when the 64 dimensional space of codon probability distribution is projected onto the first two and three principle components. Such structures have been identified in over 143 different bacterial DNA sequences with $G-C$ rich content [5].

II Theory

II.A Background on DNA and Genetic Code

DNA is a polymer that encodes genetic information about living organisms. It consists of double stranded chain of nucleotides that are wound into a double helix. Each nucleotide consists of a pentose sugar (2-deoxyribose), a phosphate group, which is attached off the fifth carbon (5') of the sugar, and one of the four nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Along a single strand, the nucleotides form a covalent bond between the 3' hydroxyl group of one nucleotide and the 5' phosphate group of the adjacent nucleotide with a phosphodiester bond. This creates the sugar-phosphate backbone of DNA. The directionality of DNA is from the 5'-end, which has a free phosphate group from the first nucleotide, to the 3'-end, which has a free hydroxyl group from the terminating nucleotide.

The bases along each polynucleotide chain lie on the inside of the backbone of DNA. The two polynucleotide chains of DNA connect together by hydrogen bonds between the bases, according to the following rules: A bonds with T and C bonds with G. This means that the two strands of DNA are complementary, and contain essentially the same genetic information. The two strands are antiparallel to each other, where one starts and ends in the 5'→3' direction and the other starts and ends in the 3'→5' direction.

Genes are sequences of nucleotides in DNA that can be expressed to form functional gene products, such as proteins. For protein synthesis, gene expression involves transcription of DNA to RNA, followed by the translation of the RNA into a chain of amino acids (a polypeptide). During transcription, a copy of a DNA segment in the 5'$\rightarrow 3'$ direction is copied to RNA, which produces a strand of messenger RNA (mRNA) in the case of a protein. During translation, the mRNA is decoded into a sequence of amino acids. While there are around 500 naturally occurring amino acids, only 20 amino acids appear in proteins.

According to the genetic code, non-overlapping triplets of nucleotides in DNA (mRNA) sequences, called codons, translate into the 20 different amino acids. While there are $4^3 = 64$ different codons, only 61 code to amino acids. These 61 codons are unique and always translate to specific amino acids. However, their correspondence to the 20 amino acids is degenerate, meaning that more than one codon codes into a specific amino acid. The two exceptions to this is the start codon ATG (AUG in mRNA), which codes to methionine, and TGG (UGG in mRNA) Tryptophan. The redundancy in the genetic code makes the process of gene expression more resilient to mutations in which a single base is changed. The remaining three codons consist of stop codons: TAA, TAG, and TGA. Stop codons mark the end of a protein. A gene begins with a start codon[1] and is translated in the 5' to 3' direction of the mRNA strand. The gene terminates with any of the three stop codons.

Within the genome, there are coding regions and noncoding regions. Coding regions contain genes that encode proteins. Noncoding regions contain non-protein genes that transcribe into functional RNA, such as transfer RNA (tRNA), and transcription factors, which regulate gene expression. The noncoding regions additionally contain sequences that do not encode any genetic information. The percentage of coding versus noncoding regions in the genome varies widely amongst living organisms. However, the percent of coding regions in a given genome is often anticorrelated with complexity– the more complex the organism, the smaller the percent of coding regions in the genome will be.

II.B Reading Frames

For a given segment of DNA, a sequence of nucleotides can be read in 6 different ways. Consider the following segment of DNA of an oligopeptide and its complementary strand,

```
5’ - GACTATGCTCATATTGGTCCTTTGACAATGCAGTTGGGCCATTAG - 3’ (forward)
3’ - CTGATACGAGTATAACCAGGAAACTGTTAGGTCAACCCGGTAATC - 5’ (reverse)
```

The forward strand can be read three different ways,

```
1st Frame: 5’ - GAC TAT GCT CAT ATT GGT CCT TTG ACA ATG CAG TTG GGC CAT TAG - 3’
2nd Frame: 5’ - G ACT ATG CTC ATA TTG GTC CTT TGA CAA TGC AGT TGG GCC ATT AG - 3’
3rd Frame: 5’ - GA CTA TGC TCA TAT TGG TCC TTT GAC AAT GCA GTT GGG CCA TTA G - 3’
```

[1] For bacteria, transcription can be initiated by GTG (GUG in mRNA) or TTG (UUG in mRNA). However, the first amino acid created during translation is always methionine

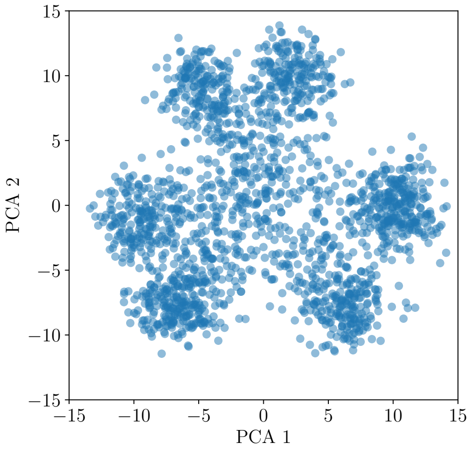


Figure 1: Projected data along first two principle components for the DNA sequence of Pseudomonas Aeruginosa. A flower-like pattern emerges when the 64-dimension manifold variable space is projected onto the principle components.

where start codons are marked in red and stop codons are marked in blue. In the first and second reading frames, two completely different peptides are embedded in the short DNA segment. However, in the second and third reading frame, it isn't complete clear whether these nucleotide sequences encode a gene or not, in the case a extended portion of the DNA segment was included.

The backward strand can be read in 3 different ways as well (in the 5' to 3' direction)

```
1st Frame: 3’ - CTG ATA CGA GTA TAA CCA GGA AAC TGT TAG GTC AAC CCG GTA ATC  - 5’
2nd Frame: 3’ - C TGA TAC GAG TAT AAC CAG GAA ACT GTT AGG TCA ACC CGG TAA TC - 5’
3rd Frame: 3’ - CT GAT ACG AGT ATA ACC AGG AAA CTG TTA GGT CAA CCC GGT AAT C   -
```

In this case, no complete peptides are embedded within the DNA segment. Even with such a small segment of DNA, it is apparent how difficult finding the correct frameshift for a DNA sequence can be. This is especially so when the gene is an unknown homolog and no context is available to suggest a particular frameshift.

III Results and Discussion

The DNA sequence used in this analysis is from the bacteria Pseudomonas Aeruginosa. The genomic data comes from GenBank,

`GCF_000006765.1_ASM676v1_genomic.fna`

To verify that this pattern is not just an visual artifact of the method, PCA is performed on the same DNA sequence and fragment length $N = 300$, but varying

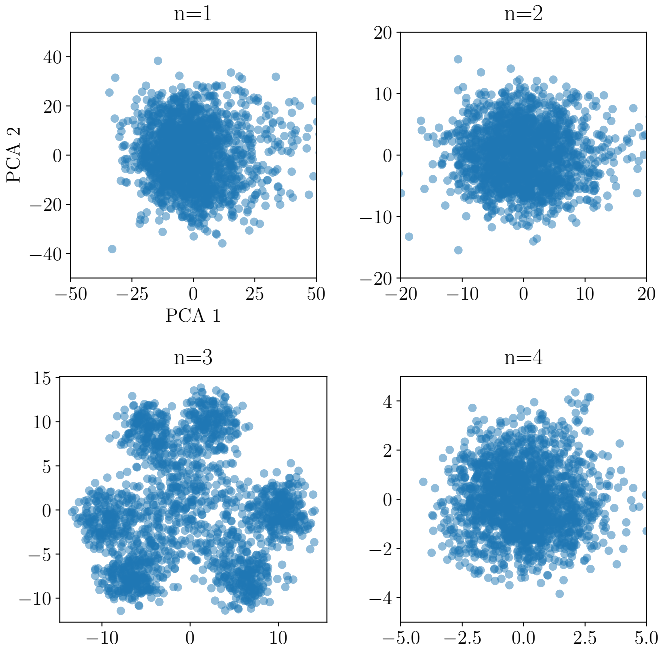
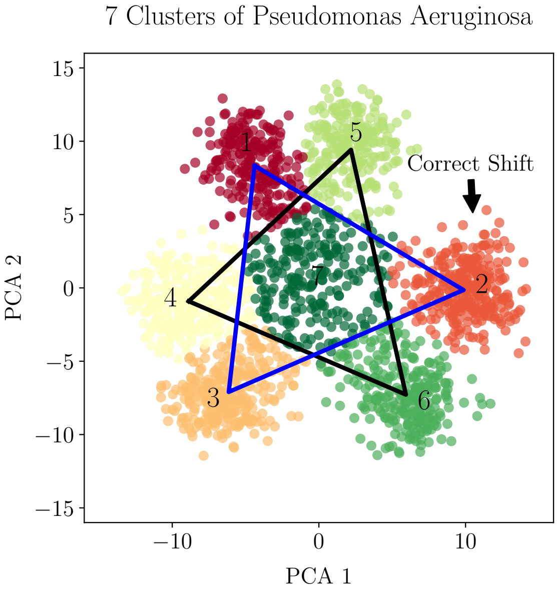


Figure 2: Projected data along first two principle components for varying codon length of $n = \{1,2,3,4\}$. Only the overlapping triplets case (n=3) has an underlying structure and symmetry.

codon length $n \in 1,2,3,4$. The frequency for each the 4^n variables is then measured for each fragment, and PCA projects the 4^n dimensional manifold variable space onto the two principle axes, as shown in Figure 2. Figure 2 clearly shows that DNA has an underlying structure and symmetry only in the $n = 3$ case. This is consistent with the genetic code, which describes how information is encoded in nonoverlapping triplets. Therefore, it is highly likely the PCA provides a meaningful representation of the data and that the flower-like structure correlates with the information embedded within the DNA sequence.

K-Means is then used to identify seven distinct clusters in the flower-like structure, as shown in Figure 3(upper). The location of each of these data points (e.g. each fragments) in the DNA sequence is in Figure 3(below). When the first three principle components of the data are plotted (see Figure 4), the six clusters on the outisde lie of two separate, nearly parallel, planes. Connecting the centers of each of the clusters reveals an orbit with approximate C_3 symmetry. These orbits have been identified in Ref. [6, 5] as corresponding to the 3 different phase shifts in the forward and backward strands of the DNA.

The mutual information is used to determine the correct phase shift, as well as determine the information content in the central cluster. The mutual information



```
ACTTCTTCGCCAGCAATGACTACCGCTCGGTGTCGCTGCTCGGCGACCAGCTGAACAGCCTGCTGGAAGACGGCGCCTACGTGCAGAAGGGTGAGCGCAAGCGCCCGATCAGCGCCT
TCAAGGACGGCCTGGACTGGCTGATGGCCGAAGGTACCAAGCGCCACAGCATCCAGCGATACAAGGGGCTGGGCGAGATGAACCCTGAGCAGCTGTGGGAAACCACCATGGATCCGA
ACGTCCGGCGCATGCTCAAGGTGACCATCGAGGATGCCATCGCCGCCGACCAGATCTTCAACACCCTGATGGGCGATGCCGTGGAGCCGCGCCGCGACTTCATCGAAAGCAACGCGC
TGGCGGTGTCGAACCTGGACGTGTGACAGGTCGGCAGACCGACCCTCATGGAAACCCCGGCCTGGCGCCGGGGTTTTCTTTTTGCGCCAGGTAGCCTGGATACGGGCGCCAGGGGTG
CCTTGCCAAGGTGTGGCCGAGCCCCGGTTACTCGCCGGCATTGGCACCCTGGACCCGGGCCTCCTGGCGGCGTGGCACCGCTTCTGTGGGCAGCGGTCGCCGGTCGGTTCGGTCGAT
CGTTGCCAGCCGGCCTTGCTTTGCGCCTCGCGTGAAAACAAGAAGCCCCGCTATGGCGGGGCTTCTTTTATCGAATCGGCGCACAGTACGCCTTGCAGGGGGCTTTGCGGCCCTTCA
CGAGACCACCGACGGCTCCGGATGGCTGACCCGCTGCTGGATGGGGCTGATCTCGGCCATGGTCTCGCTGACCCAGGCTTCGGCGCGCTGGTTTAGCTCGGCGATGGCGCGCGGGCC
TTCGCCTTCGGCGTGCATGGCCGGGCCGATCACCACCTGGATGGTGCCCGGGTACTTGGCCCAGCCGGCCTTGGGCCAATACTGCCCGGCGTTGTGGGCGATCGGCAGTACCGGTAG
CCCGGCGTTGACCGCCAGGGCGGTGCCGCCGCGGGAGAACTTGCCCATCTGCCCCACCGGAATACGCGTGCCTTCCGGGAAGATCAGCACCCAGGCGCCTTTCTTCAGGCACTCGTC
GCCCTGCTTGGCCAGTTGCTTGAGGGCCAGCTTGGGCTGGCTGCGGTCGATGGCGATGGGCTTGAGCAGGGCCAGGGCCCAGCCGAAGAACGGCACGTAGAGCAGCTCGCGCTTGAG
TACCTGGCTGAGTGGCTCGAAGAAGCCGGAGAGGAAGAAGGTTTCCCAGGTGCTCTGGTGCTTGGAGAGGATCACGCAGGGCTTTTCCGGGATGTTCTCCAGTCCGCGCACCTCGTA
GCGGATGCCGGCGACCACGCGGGTCAGCCAGATCGCGAAGCGGCACCAGTTCTGTACCACGAAGCGGTAGCGGGCGCGGAACGGCAGGATCGGCGCGATGAAGAAGCTGAGGGTGCC
CCAGACGAACGCGCTGGCGGACAGCAGCAGGTAAAAGAGGACGGTTCTGATGGCCTGCACTGTCGACATGTATTCTGACCTTACTGAAGTAATGCGCTGGCGACTGCCGCCAGATCG
TCGAATATCAGGGTGCCCTCTGGCAAGGGCTTGCCCAGCGTACGTACACCTTTTCCGGTCTTTACCAATACCGGCTGACAATCGACGGCCCGCGCCGCCTCCAGGTCACCGATGCTG
TCGCCGACGAACCAGATACCCGACAGATCGACCCCGTAGTGCTCGCCGATCTGCCGCAGCATACCCGGCTTCGGCTTGCGGCAGTCGCAACCGTCGTCCGGTCCATGCGGACAATAG
ACGATGAGGCCGACCTCGCCGCCCTGCTCCGCGACCAGTTCGCGCAAGCGCGCATGCATGGCCTCGAGCACTGCCAGGTCGTAATAGCCACGGGCGATGCCGGACTGGTTGGTAGCC
ACCGCGACGGTCCAGCCGGCCTGGCTCAGGCGGGCGATGGCCTCGATCGAGCTGGGGATGGGGATCCACTCGTCGAGGGTCTTGATGTAATCGTCGGAGTCGAGGTTGATGACTCCA
TCGCGGTCGAGAATCAGCAGGGAACGGGACATCGATGCGCAGTTGGCCATGAACGGAAAGGATCGATTCTACCTCAGCCAGCCGGAGCATCGCGGATGGATCGACGAGCCCCCTACG
CTGGACAATCTGTACTAGATTCAATAATGGCAATGGAATGGCAATCCCAGGCCAGGGAGTCATTTGGCACGCAGCAGTGGGTGTCTCGGAGTACGTGCGAACCCCTTGGCGGGTGGA
TCACCGGAGTAACGCAGTAAGCGCCTATCGCATGCCAGCACGGCCGACCGGCCGTGTTTCCACCACAGGTGGAGCGGGCGTGACGACGGGGAGTTCCGGGGCGTCTGCCCGTTCCCA
AGGACGGTAACCGTGAAACGCCTGAAAAAGACACTGCACCTTTCAAGCTTGTCCCTCGCTTCCCTGGCTCTTTCTTCCGCCGCCCTGGCGGCCGCTCCGGTCATGCTCGACCAGGGC
AAGGAATGGACCGAAAGCCACCGCCAGGACTTCTACAGCCGCGACCAGGGCTCGCAGGTGATGCCCCTGCCCTGGCTCAAGGCGTTGCGACAGCCGGATGGAACGCCTTTCCTCGCC
GACAGCCTGGCCCGCTACGGCTATTTGCCCAACCCCAAGGCGCCCGCGGAAGGCCTGCCGGTGGGCTTCACCGTAGCCGGCACGGGCGCCCGGCAGATGGTCGGCATGACCTGTTCG
GCCTGCCATACCCGGCAGATCGAGGTGAAGGGCACTGCCTATCGGATCGACGGCGGTCCGGCGATCGTCGACTTCCAGGCATTCCTCGCCGACCTCGATCGGGCCGTGGACCGCTG
```

Figure 3: (Upper) K-Means separates the flower-like structure into seven distinct clusters. The correct phase-shift of the data is identified by looking at the mutual information (Eqn. 1) of each fragment. (Lower) Shows the cluster number of each fragment within the DNA sequence.

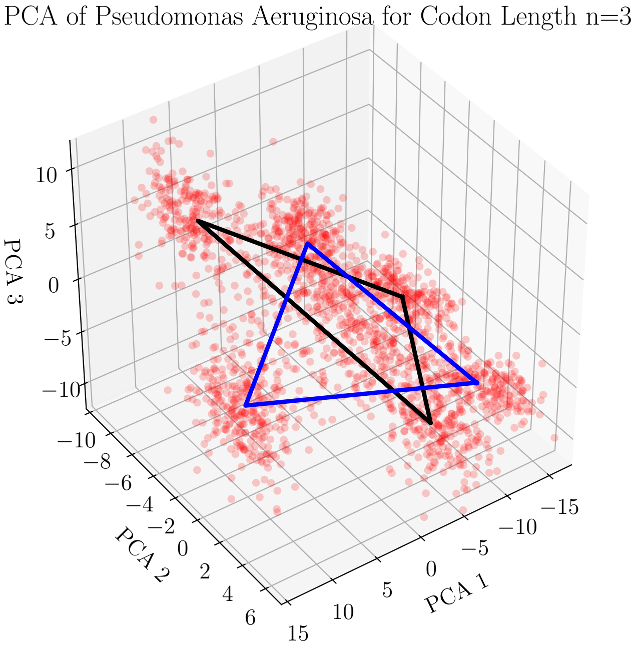


Figure 4: Plotting the first three principle components shows two separate orbits with approximate C_3 symmetry.

measures the information in each fragment,

$$M = \sum_{ijk} f_{ijk} \log \left[\frac{f_{ijk}}{p_i p_j p_k} \right] \tag{1}$$

where p_k for $k \in \{A, C, G, T\}$ corresponds to the probability of observing the k-th nucleotide and f_{ijk} is the probability of observing the codon triplet ijk for $i, j, k \in \{A, C, G, T\}$. The mutual information ranges from $M = [0, \ln(27)]$. The mutual information is maximized when at least one nucleotide uniquely determines a codon in the fragment. When p_i's are independent random variables, each codon probability is simply the product of each nucleotide probability, $f_{ijk} = p_i p_j p_k$. In this limit, all information in regards to nucleotide position is lost, and the mutual information is trivial, $M = 0$. The non-coding regions of bacterial genomes are well represented by this state of maximal entropy.

Figure 5 shows the mean mutual information per cluster. Cluster 7 (e.g. the central cluster) has the minimum mutual information, which suggests the data in this region corresponds to fragments in the noncoding regions of the DNA sequence. While the MI isn't identically zero for cluster 7, it is likely that the fragment length isn't long enough to fully sample a uniform random distribution. To verify this, the mutual information was measured for a randomly generated strand of length $l = \{300, 3000, 30000\}$. As expected, the mutual information decreased with fragment length, where $MI = \{0.346, 0.032, 0.003\}$ respectively. Figure 5

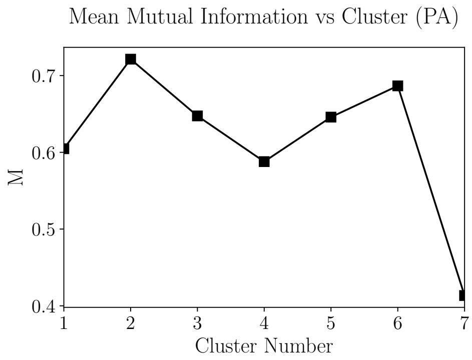


Figure 5: The average mutual information per cluster. The figure suggests that the correct phase shift is either cluster 2 or 6, and the non-coding region is likely cluster 7.

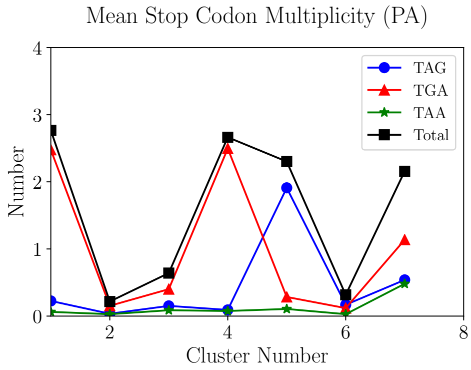


Figure 6: The average mutual information per cluster. The figure suggests that the correct phase shift is either cluster 2 or 6, and the non-coding region is likely cluster 7.

also shows that clusters 2 and 6 have the highest amount of mutual information. To verify which cluster is the correct phase shift, the average number of stop codons per cluster is measured (Fig. 6). It is clear from this figure that cluster 2 is the correct shift of the forward strand, as it has the minimum number of total stop codons.

Bibliography

[1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[2] Stéphane Audic and Jean-Michel Claverie. Self-identification of protein-coding regions in microbial genomes. *Proceedings of the National Academy of Sciences*, 95(17):10026–10031, 1998.

[3] Mark Borodovsky and James McIninch. Genmark: parallel gene recognition for both dna strands. *Computers & chemistry*, 17(2):123–133, 1993.

[4] N. N. Bugaenko, A. N. Gorban, and M. G. Sadovsky. Maximum entropy method in analysis of genetic text and measurement of its information content. *Open Systems & Information Dynamics*, 5(3):265–278, Sep 1998.

[5] A Gorban, Tatiana Popova, and Andrei Zinovyev. Four basic symmetry types in the universal 7-cluster structure of 143 complete bacterial genomic sequences. 5:0039, 10 2004.

[6] Alexander N Gorban, Andrei Y Zinovyev, and Tatyana G Popova. Seven clusters in genomic triplet distributions. *In silico biology*, 3(4):471–482, 2003.

[7] AN Gorban, TG Popova, and MG Sadovsky. Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy. *Open Systems & Information Dynamics*, 7(1):1–17, 2000.

[8] David J Lipman and William R Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985.

[9] Steven L Salzberg, Arthur L Delcher, Simon Kasif, and Owen White. Microbial gene identification using interpolated markov models. *Nucleic acids research*, 26(2):544–548, 1998.