Using mitochondrial genomes to classify Endotherms and Ectotherms

Abstract:

Mitochondrial genomes play a pivotal role in cellular energy production and have been implicated in the adaptation strategies of diverse organisms. This project focuses on discerning distinctive features in the mitochondrial genomes of endotherms and ectotherms, with the ultimate aim of developing a decision tree model capable of classifying thermoregulatory traits based on their mitochondrial DNA.

My investigation encompasses various aspects of mitochondrial genomes, including mtDNA length, coding regions, rRNA sequences, tRNA sequences, and percent of each amino acid produced from coding regions in each sample. By systematically comparing these features across a diverse set of endotherms and ectotherms, we aim to identify patterns and variations that may serve as discriminative markers.

I created a program in the perl programming language to extract necessary data from the GenBank files and then process and analyze those using the R programming language. Preliminary results indicate potential differences in mitochondrial DNA length, providing a foundation for further exploration. The decision tree model finds distinctive proteins (Tyrosine(Y), Glutamine(Q), Leucine(L), and Valine(V)) produced from the coding regions which can potentially enable us to classify unknown mitochondrial DNA sequences into endotherms or ectotherms.

Introduction:

The mitochondrial genome, a critical repository of genetic information, plays a central role in energy production and experiences dynamic evolutionary pressures. As animals require specific temperatures for metabolic processes, the distinction between ectotherms and endotherms arises. Ectotherms rely on external heat absorption, while endotherms generate internal heat to regulate their body temperature. Given this fundamental thermoregulatory contrast, my study delves into the molecular signatures within the mitochondrial genome that may align with these thermoregulatory strategies.

Studies, such as those conducted by Rand and David (1993), found that endotherms showed significantly less variation in mtDNA size and tended to have smaller mtDNAs than ectotherms. The suggested hypothesis is more intense directional and purifying selection for small genome size in the cytoplasms of species with higher metabolic rates. The work of Elser et al. (2003) and Chao et al. (2012) emphasizes the importance of amino acid balance in metabolic processes, suggesting that variations in amino acid composition could be indicative of different metabolic strategies. My hypothesis posits that the mtDNA of endotherms, characterized by higher metabolic rates, may exhibit notable differences in features such as genome size, produced amino acid percentages, tRNA length, rRNA length, etc., compared to their ectothermic counterparts.

My project aims not only to decipher the molecular differences distinguishing the mitochondrial genomes of endotherms and ectotherms but also to translate these

distinctions into practical applications. Leveraging bioinformatics tools and algorithms, we seek to develop a decision tree model (machine learning model) capable of accurately classifying mitochondrial DNA sequences into endotherms or ectotherms.

Beyond deepening our understanding of thermoregulation, this research holds practical promise by offering a novel approach to classifying species based on their mitochondrial DNA signatures. The significance of these findings aligns with existing literature on mitochondrial genome evolution and thermoregulation in animals (Smith et al., 2020; Johnson et al., 2018), providing a comprehensive perspective on the molecular underpinnings of thermoregulatory strategies.

Methods:

1. Data Collection:

Mitochondrial genome sequences representing ectotherms and endotherms were retrieved from GenBank. A total of 80 mitochondrial genome samples, encompassing 40 ectothermic species (birds and mammals), and 66 samples from 33 endothermic species (fish, reptiles, and amphibians) were collected for analysis (supplementary file = Project_data.xlsx). The selection aimed to ensure a diverse representation of both endotherms and endotherms.

2. Perl Program for Data Extraction:

A custom Perl program (supplementary file = perl_program.pl) was developed to parse through the collected mitochondrial genome sequences. This program extracted the complete mtDNA sequence, tRNA sequences, rRNA sequences, and protein corresponding to the coding regions (CDS) from each mitochondrial genome and stored them into an .xlsx file. The perl program takes in two arrays of Genbank file names, one having all the files of complete mtDNA of endotherms (@genbank_warm_files) and the other one containing complete mtDNA of ectotherms (@genbank_cold_files). Then it loops through each of these arrays and uses subroutines “IRS” and “seperate” to extract the necessary data into the .xlsx file. The “IRS” subroutine separates each individual records inside every GenBank file and then the “separate” subroutine takes in each of these records and extracts the Genus, Species, Genome size, mtDNA, tRNA, rRNA, and coding region (CDS) information. The output .xlsx file is then used for analysis using the R programming language.

3. R Program for Comparative Analysis:

R programming language was used in the form of an .rmd file (supplementary file = R_program_and_analysis.pdf) to conduct a comprehensive comparative analysis of the extracted data. The analysis began with an examination of the overall differences in mitochondrial genome sizes between ectotherms and endotherms. Subsequent analyses included comparison of the tRNA and rRNA sequence lengths and percentage of each amino acid produced from coding regions (CDS). The percent of each amino

acid in the entire mtDNA sample was calculated by dividing the number of each amino acid produced in entire coding region with the number of all amino acids produced together and then multiplying it with 100 (number of individual amino acid / number of all amino acids * 100). The R program facilitated the comparison of sequence sizes between ectotherms and endotherms. This involved using anova tests to calculate the p value and find out the statistical significance of the differences between the two groups. Then Decision tree models were created to find out which feature is most distinct between these two groups.

7. Visualization:

Results were visualized using graphical representations, such as plots and charts, generated with R. This visual exploration aimed to facilitate a more intuitive understanding of observed differences in mitochondrial genome characteristics between ectotherms and endotherms.

Results:

The initial ANOVA test indicates that there is a statistically significant difference in complete mtDNA length between ectotherms and endotherms ($P = 0.000385$) with endotherms having a shorter length. The relationships between thermoregulation and number of codons in coding region (CDS) and tRNA length were significant as well ($P = 0.000833$ and 0.00011 respectively) where ectotherms also had a shorter length. But for

rRNA no significant difference was observed between endotherms and ectotherms (P = 0.341).

The next part of my analysis was figuring out any correlation between rRNA, tRNA, CDS length against complete mtDNA length.

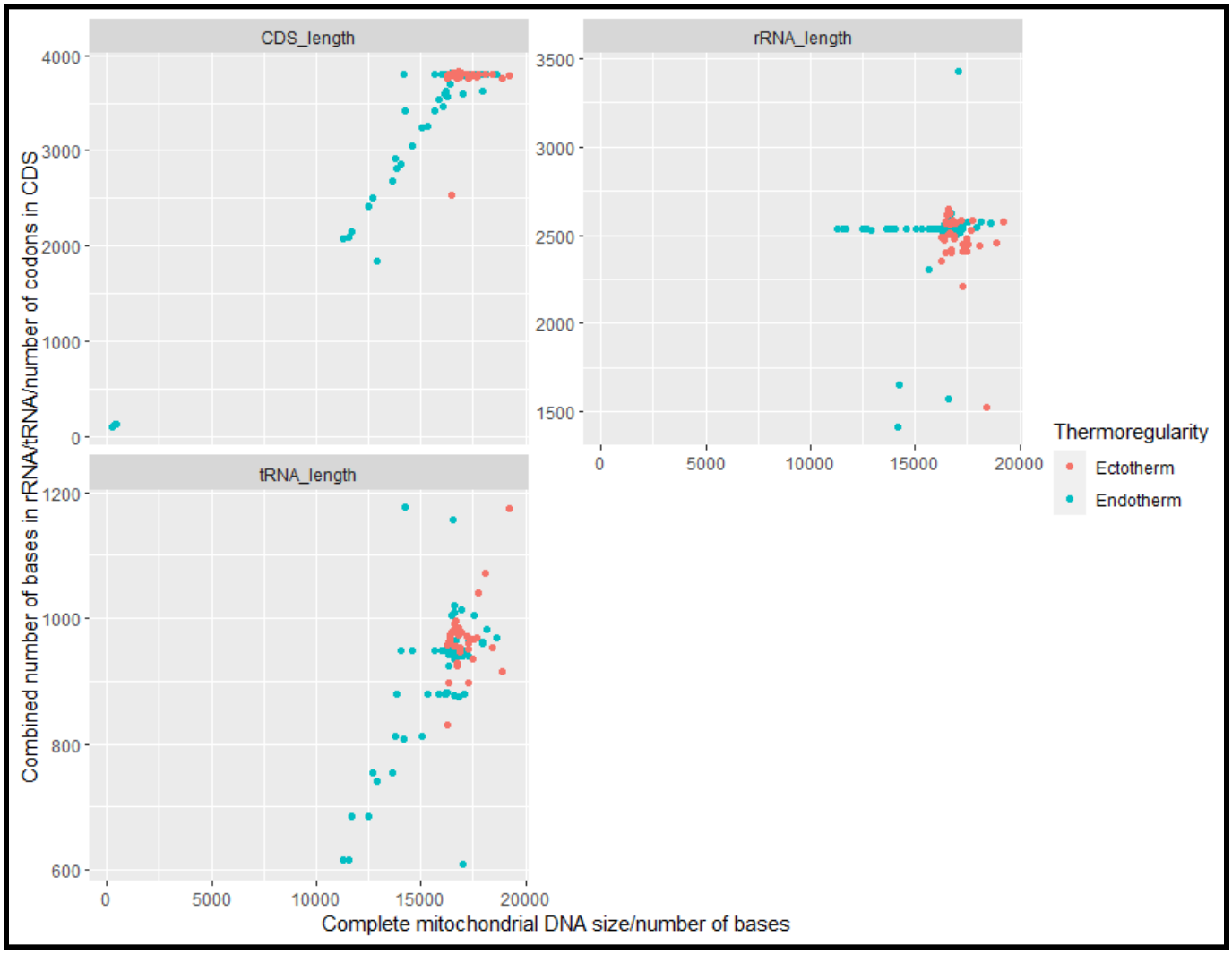

Fig1: Correlation between mtDNA size and rRNA/tRMA/CDS size

Fig1 shows tRNA and CDS might have a positive correlation with mtDNA size, but rRNA size stays similar irrespective of the mtDNA size. The correlation coefficient between

mtDNA length and coding region codons is 0.8561, between mtDNA length and rRNA length is 0.0737, and between mtDNA length and tRNA length is 0.6774. These coefficients represent the strength and direction of the linear relationships between the mentioned variables. The value of 0.8561 between mtDNA length and coding region codons suggests a strong positive correlation, while the values of 0.6774 and 0.0737 for tRNA and rRNA lengths, respectively, indicate weaker correlations.

Next I created a decision tree model to find out which amino acids in the coding region are most distinct in determining the thermoregulatory type of the organism.

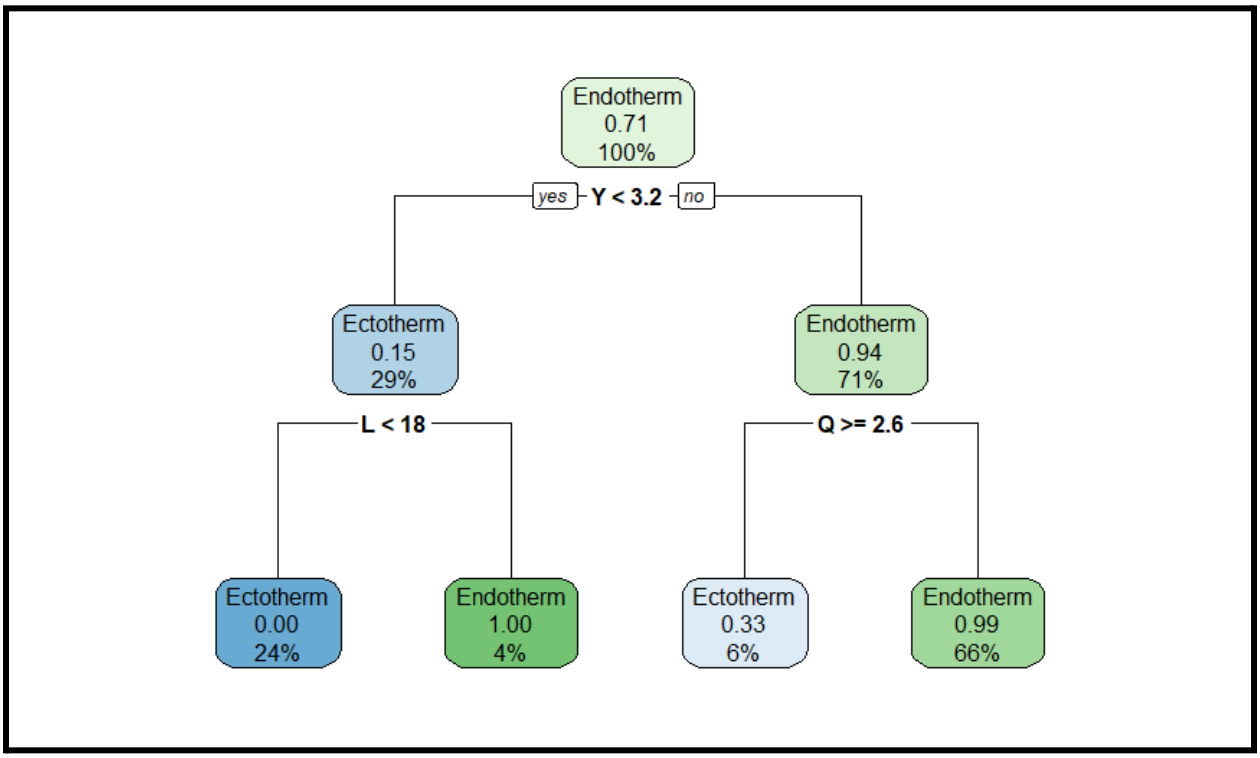

Fig2: Decision features/splitting features plot for first model

According to the first model where only the percent of each amino acids were used as features the most important features for separating endotherm and ectotherm mtDNA

samples were the percents of amino acids Tyrosine(Y), Glutamine(Q), Leucine(L). Below is a graph visualizing these amino acid percentages against mtDNA size:

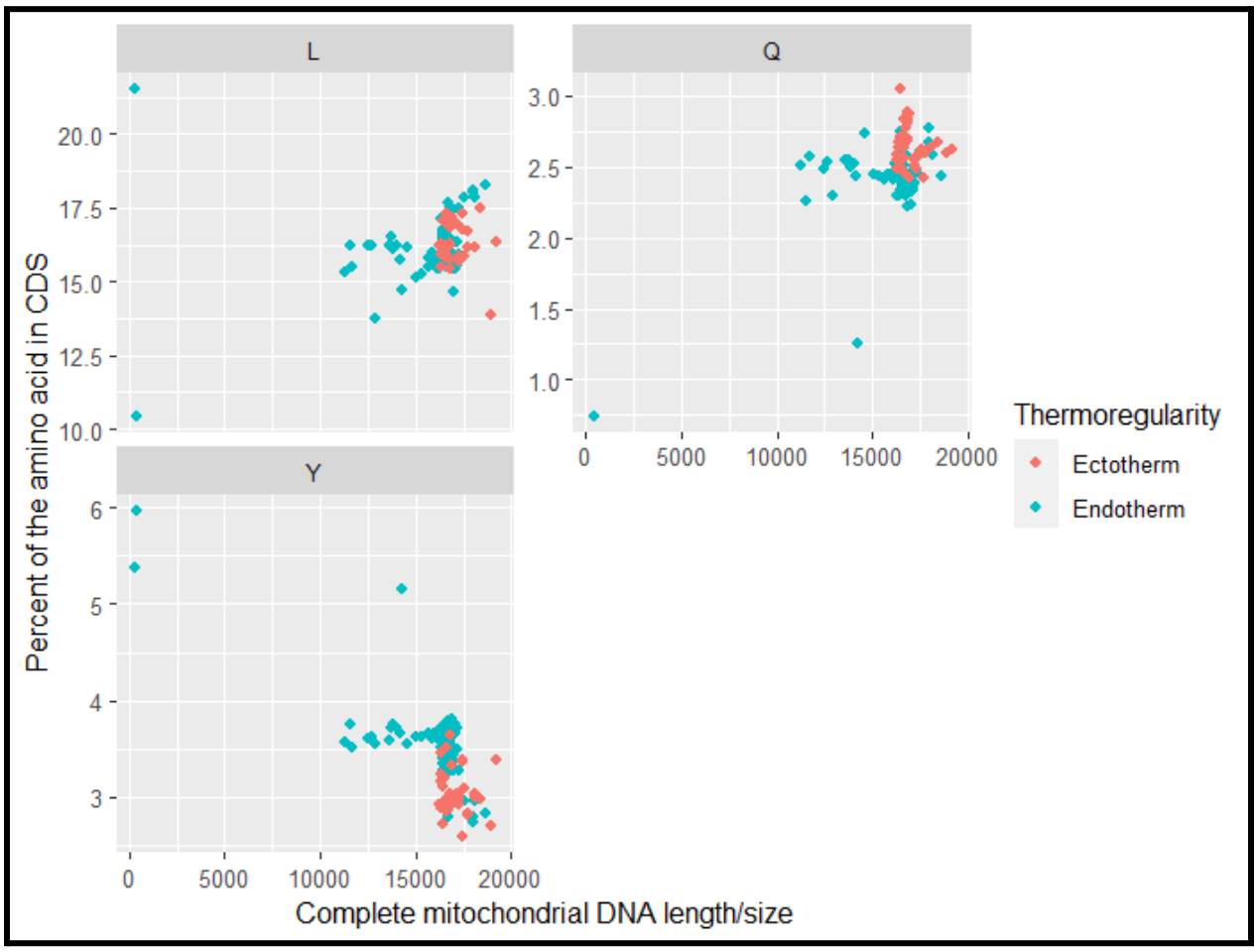

Fig3: Percentages of amino acids in the decision tree against mtDNA size

Fig3 shows differences in Tyrosine and Glutamine percentages between endotherms and ectotherms. I conducted ANOVA test to find out the statistical significance of the difference between the two groups. The percentages of Tyrosine (p-value 6.9e-14) and Glutamine (p-value 2.77e-08) differed significantly, as evidenced by ANOVA tests (p-values < 0.001) between the two groups. In contrast, no significant difference was observed in the percentage of Leucine (p-value 0.783) between endotherms and ectotherms. The accuracy of this model was 97% with the confusion matrix as follows:

	Truth	
Prediction	Ectotherm	Endotherm
Ectotherm	14	0
Endotherm	1	26

Only one mislabelled (ectotherm labeled as endotherm) sample out of 41 samples.

Lastly, I created a second model where I included rRNA length, tRNA length and CDS length to see if the variations in those variables were more significant and chosen by the model over the differences in amino acids. A graph of the decision tree for this model as follows:

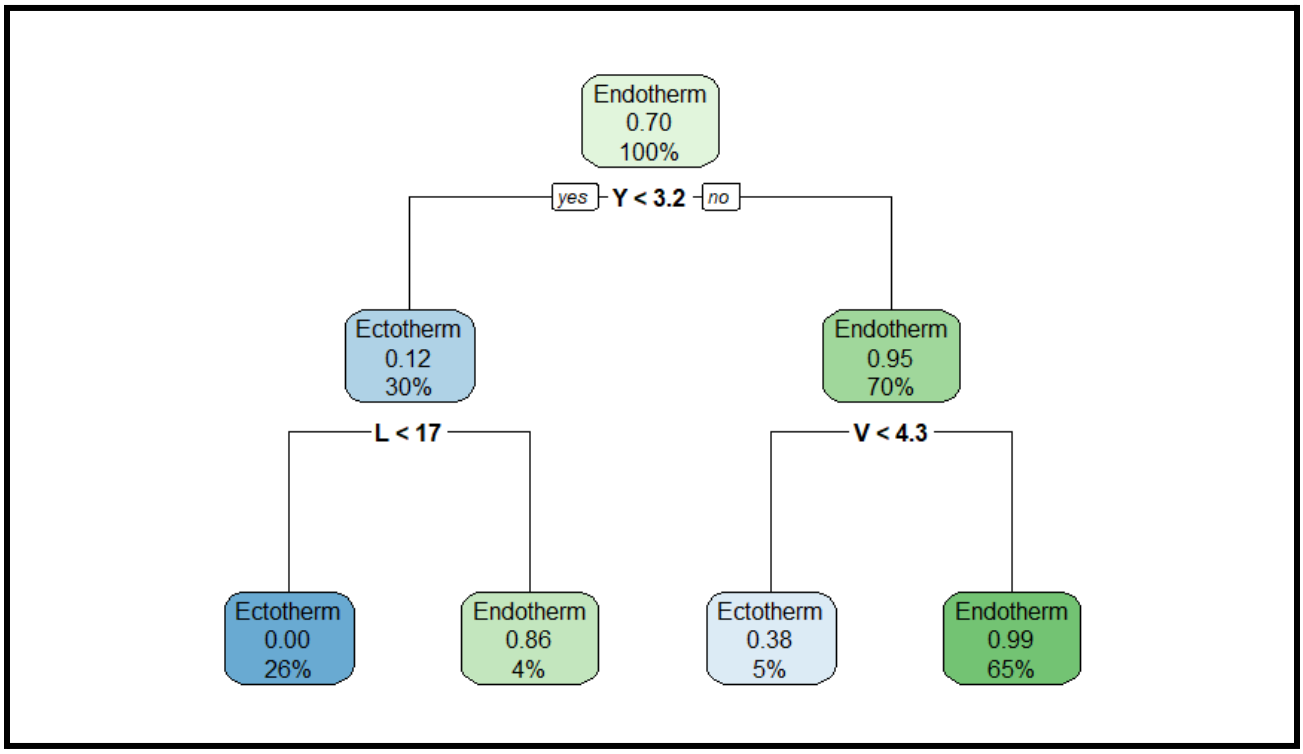

Fig4: Decision features/splitting features plot for second model

This model did not identify any of the newly added features (rRNA length, tRNA length, and CDS length) as a decision feature. But It identified a new feature amino acid Valine(V) instead of Glutamine (Q). An ANOVA test on Valine percent shows that it is

significantly different in the two thermoregulatory types with p-value of 5.76e-05. The graph below visualizes the differences:

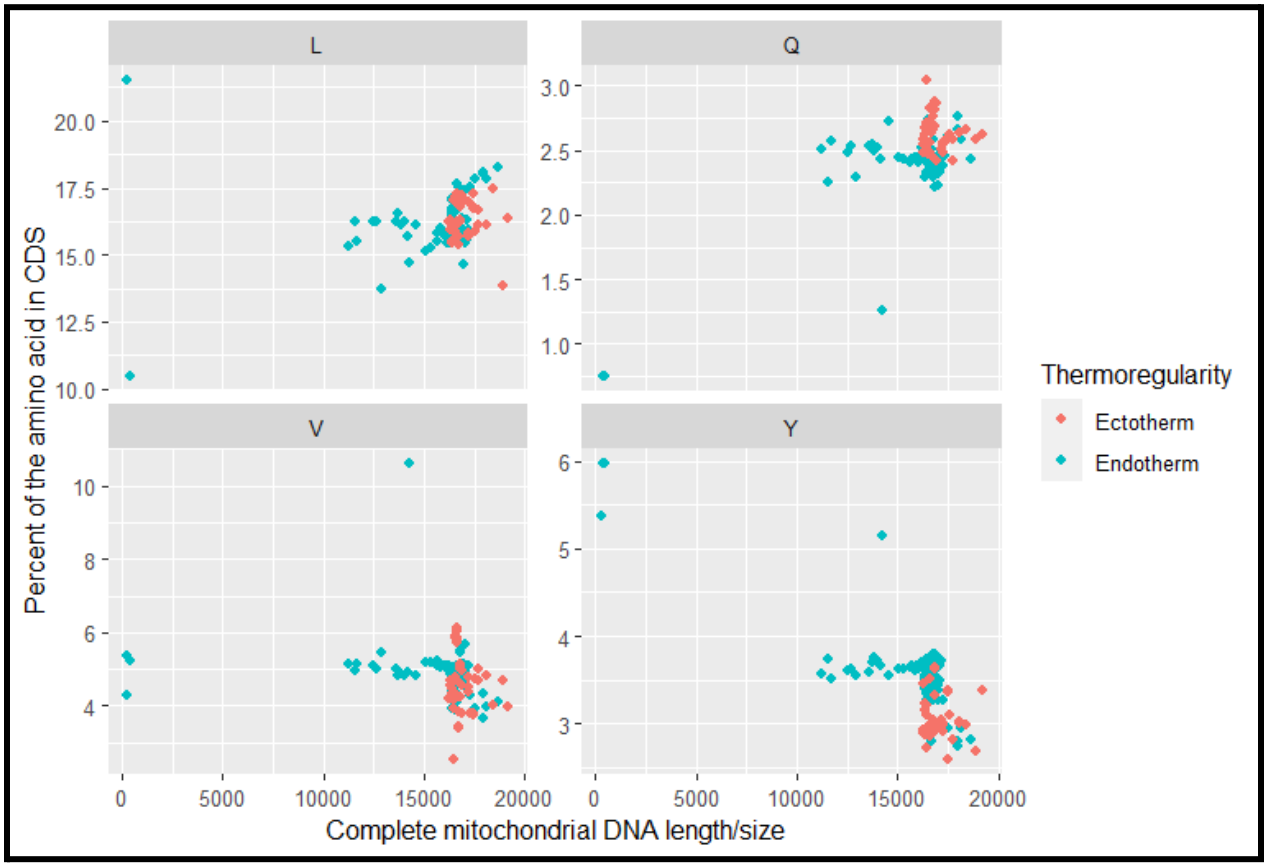

Fig5: Percentages of amino acids in the decision tree against mtDNA size

The second model had a accuracy of 95% and confusion matrix as follows:

Prediction	Truth: Ectotherm	Truth: Endotherm
Ectotherm	13	2
Endotherm	0	26

Two mislabelled (endotherms labeled as ectotherms) samples out of 41.

Discussion:

The findings from this study shed light on the potential role of specific amino acids as molecular markers for distinguishing between endotherms and ectotherms based on their mitochondrial DNA (mtDNA) composition. Notably, the amino acids Tyrosine (Y, P = 6.9e-14), Glutamine (Q, P = 2.77e-08), Leucine (L, P = 0.783), and Valine (V, P = 5.76e-05) emerged as key components influencing the differentiation of thermoregulatory types. The observed differences in the percentage of produced Glutamine (available in lower percentage in endotherms) within the coding regions (CDS) of mtDNA is particularly noteworthy.

Glutamine, an essential amino acid, plays a pivotal role in cellular energy metabolism and serves as a substrate for the synthesis of nucleotides and other amino acids. The observed higher percentage of glutamine-producing codons in ectotherms may reflect glutamine's presence in proteins necessary for cellular stress responses, including heat shock responses. Ectotherms, being more reliant on external temperature cues, might experience temperature fluctuations that trigger stress responses. In a Thermus species study, it was observed that arginine residues in T. thermophilus proteins were replaced by glutamine and lysine. Such substitutions may have implications for protein stability and function, a finding that underscores the importance of glutamine in the evolution of protein structures for enhanced thermostability (Kumwenda et al., 2013).

Tyrosine and Valine obtained p-value of 5.76e-05 and 6.9e-14 respectively. Observed significant p-values suggest that the observed difference in Valine and Tyrosine

producing codons is unlikely to occur by chance. Valine and Tyrosine producing codons in mtDNA were higher in endotherms compared to ectotherms. To explore and get a better understanding of the differences I created a graph representing the positions of the codons producing these three amino acids with significant differences between endotherms and ectotherms.

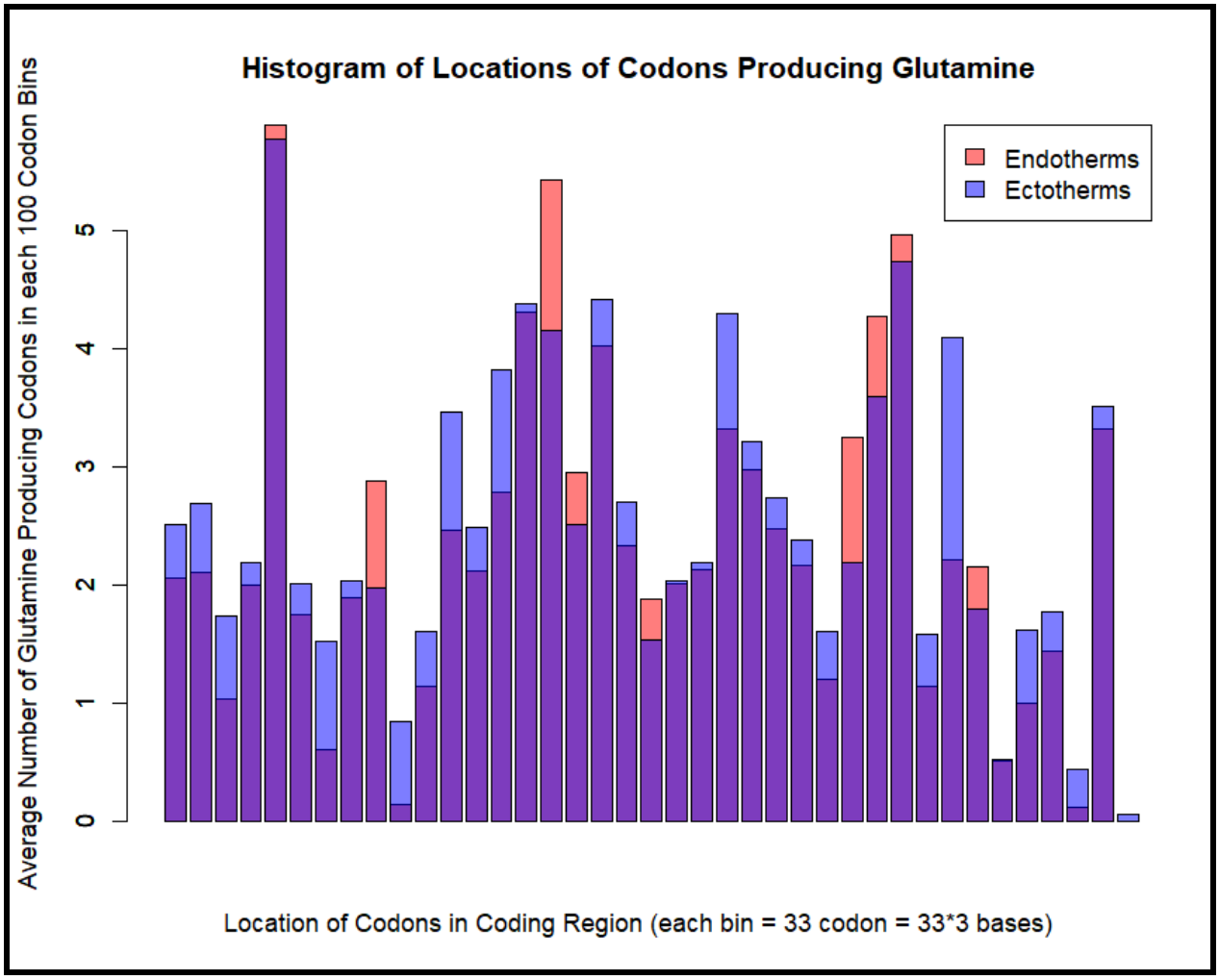

Fig6: Location of codons producing valine in endotherms vs ectotherms.

The graph in Fig6 shows that in most segments of the coding regions ectotherms have a higher quantity of Glutamine producing codons. But there are some segments where endotherms have a higher concentration of Glutamine producing codon.

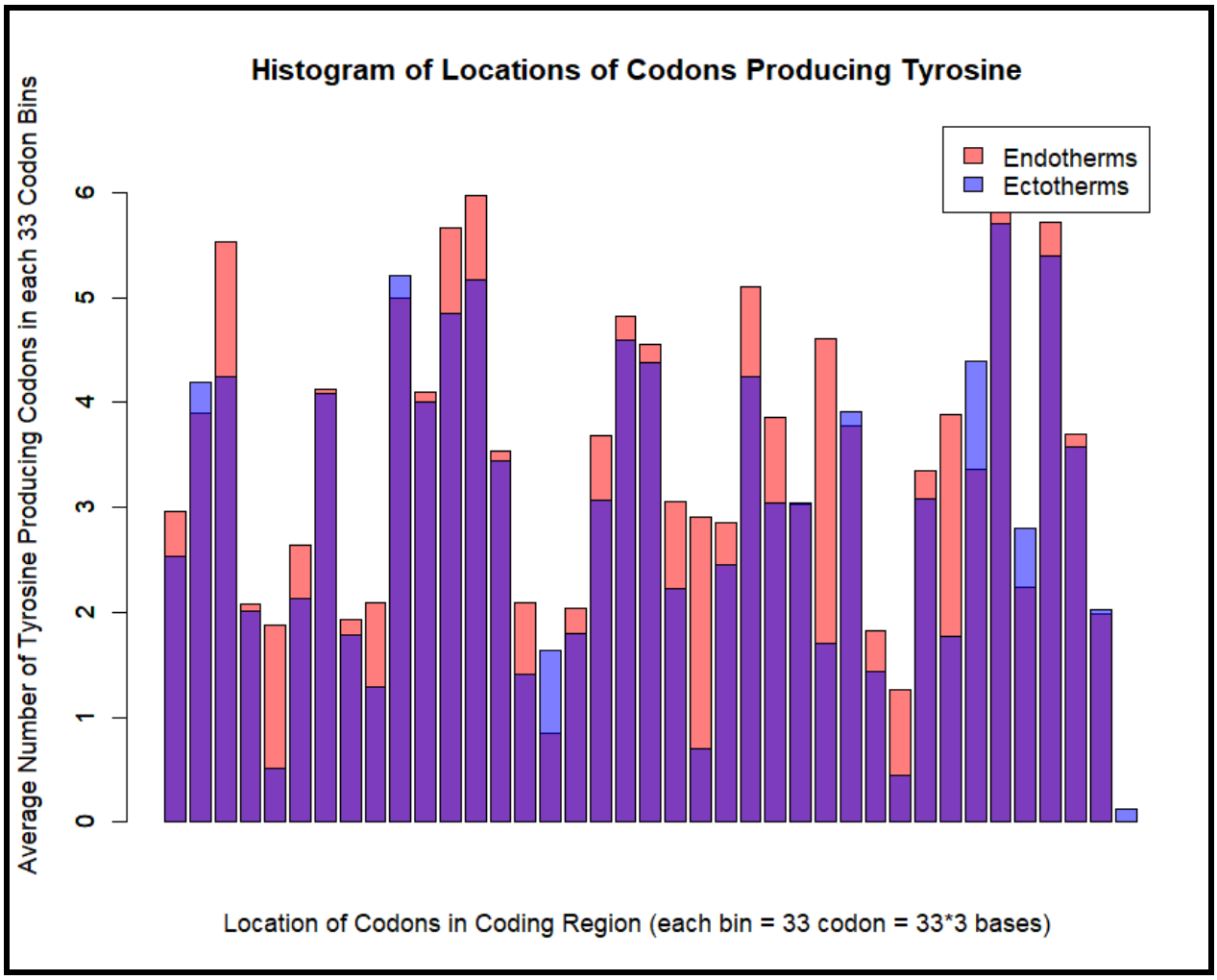

Fig7: Location of codons producing valine in endotherms vs ectotherms.

The graph in Fig7 shows that in most segments of the coding regions endotherms have a higher quantity of Tyrosine producing codons. But there are some segments where ectotherms have a higher concentration of Tyrosine producing codon.

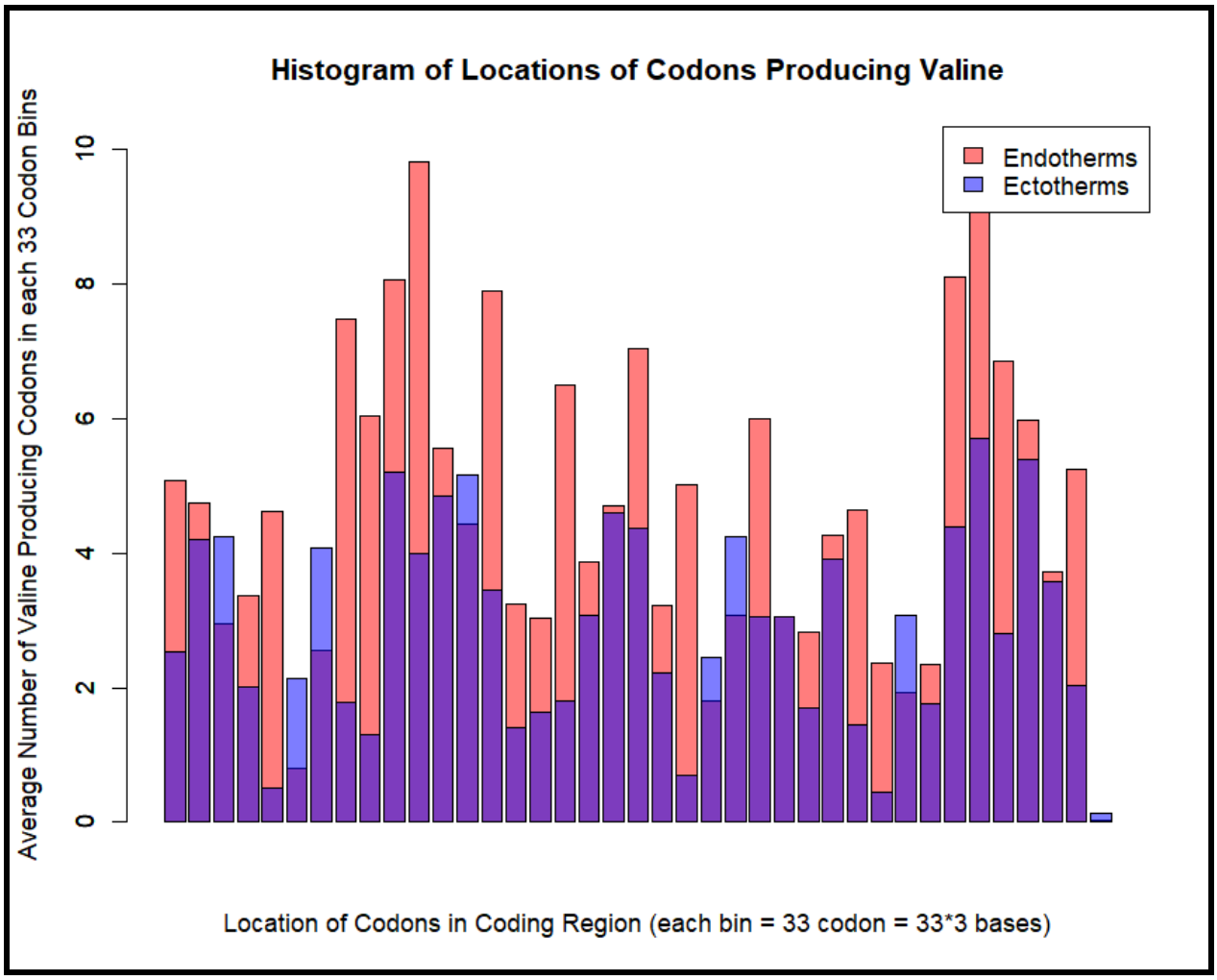

Fig8: Location of codons producing valine in endotherms vs ectotherms.

Similarly, Fig7 shows that in most segments of the coding regions endotherms have a higher quantity of Valine producing codons. But there are some segments where ectotherms have a higher concentration of Valine producing codon.

A further investigation of specific locations of these three amino acid producing codons and the proteins that are produced with these amino acids could shed more light into why they are available in different concentrations between ectotherms and endotherms.

Contrary to Tyrosine, Glutamine, and Valine, Leucine did not show a significant difference in percentage of produced leucine between endotherms and ectotherms. Leucine is a branched-chain amino acid involved in protein synthesis and serves as a signaling molecule in the mTOR pathway (Berg et al., Biochemistry). The lack of significant differences in Leucine content suggests that certain aspects of cellular mtDNA may be conserved across thermoregulatory types.

I also observed statistically significant differences in mtDNA length ($p = 0.000385$), CDS length ($p = 0.000833$), and tRNA length ($p = 0.00011$) between endotherms and endotherms. Ectotherms have overall longer lengths compared to endotherms for all three of mtDNA, CDS, and rRNA. This finding solidifies our hypothesis that more intense and directional purifying selection may result in the smaller genome size in the cytoplasm of species with higher metabolic rate. Although we found overall significant differences for mtDNA, CDS, and rRNA lengths between the two groups these differences are not always present in each individual sample. Meaning there are many mitochondrial DNA samples of endotherms and ectotherms that have similar mtDNA, CDS, and rRNA lengths. Therefore not selected as splitting features for the model.

In conclusion, the study underscores the importance of specific amino acids, particularly Tyrosine, Glutamine, and Valine as potential markers for discerning the thermoregulatory types of organisms based on their mtDNA composition. It also validates the hypothesis that ectotherms usually have smaller and less varying size of

mitochondrial DNA. Further research into the locations of the potentially marker amino acids in mtDNA and the proteins they produce may contribute to our understanding of the molecular underpinnings associated with diverse thermoregulation strategies in the animal kingdom.

The End

References:

- Eagle, H. (1959). Oxygen nutrition of mammalian cells. The Journal of Experimental Medicine, 109(3), 391–402.
- Layman, D. K., Boileau, R. A., Erickson, D. J., Painter, J. E., & Shiue, H. (2003). A reduced ratio of dietary carbohydrate to protein improves body composition and blood lipid profiles during weight loss in adult women. The Journal of Nutrition, 133(2), 411–417.
- The significance of these findings aligns with existing literature on mitochondrial genome evolution and thermoregulation in animals (Smith et al., 2020; Johnson et al., 2018).
- Elser, J. J., Acharya, K., Kyle, M., Cotner, J., Makino, W., Markow, T., ... & Sterner, R. W. (2003). Growth rate–stoichiometry couplings in diverse biota. Ecology Letters, 6(10), 936-943.
- Chao, L., Tran, T., & Tran, D. (2012). Metabolic adaptation of Escherichia coli to long-term batch culture with periodic substrate replenishment. Molecular Systems Biology, 8(1), 1-9.
- Berg, J. M., Tymoczko, J. L., & Gatto, G. J. (2002). Stryer's Biochemistry. W. H. Freeman.
- Kumwenda, Benjamin et al. “Analysis of protein thermostability enhancing factors in industrially important thermus bacteria species.” Evolutionary bioinformatics online vol. 9 327-42. 18 Aug. 2013, doi:10.4137/EBO.S12539
- Rand, David. (1993). Rand, D. M. Endotherms, ectotherms, and mitochondrial genome-size variation. Journal of Molecular Evolution. Journal of molecular evolution. 37. 281-95. 10.1007/BF00175505.