Comparative Analysis of Protein, Nucleotide, and Structural Conservation Across Multiple Species

Divyadarshan Soni[1] **22b0387, Vikash Kumar**[3] **22b0359, Rana Das**[3] **22b0738**

[1]

1. Abstract

Genetic conservation provides critical insights into evolutionary mechanisms and functional significance of genes across different species. This study investigates conservation patterns by conducting comprehensive alignments of protein, nucleotide, and structural domains for genes across distinct species. By computing and comparing conservation scores across these three molecular domains, we aim to elucidate the correlation and variability of genetic conservation mechanisms. Our analysis reveals nuanced relationships between protein sequence, nucleotide composition, and structural elements, offering a multidimensional perspective on genetic preservation and evolutionary constraints.

2. Introduction

The fundamental understanding of genetic conservation has been a cornerstone of molecular evolutionary biology. While traditional analyses have often focused on singular molecular domains, comprehensive cross-domain comparisons remain relatively unexplored. Genetic conservation can manifest differently across protein sequences, nucleotide arrangements, and three-dimensional structural configurations, each reflecting distinct evolutionary pressures and functional constraints.
This research addresses this complexity by systematically examining conservation scores across protein, nucleotide, and structural domains. By analyzing 50 genes (with large, medium and small gene lengths) from five species, we employ a multifaceted approach to quantify and compare conservation levels. Our primary objectives include:

- Quantifying conservation scores for protein, nucleotide, and structural domains
- Determining inter-domain correlation patterns
- Identifying potential mechanistic insights into genetic preservation

The methodological framework involves precise molecular alignment techniques, followed by computational analysis to generate comparative conservation metrics. By integrating these diverse perspectives, we seek to provide a comprehensive understanding of genetic conservation mechanisms that transcend traditional single-domain investigations.

2.1. Algorithms

All the multiple sequence alignment for Protein Sequences and Nucleotide sequences is done using the functions available in the Align Module of the Bio Library package in Python Language. The spacial variation in the structure is done RMSD method.

The conservation score for the Protein and Nucleotide sequences is found using the Blosum62 scoring method which uses a matrix to compute the conservation score. The conservation score for the Structure part is found using the spatial variation and the score being 1/(spatial variation + 1).

3. Results

3.1. Result 1:

The selected Species are -

- Homo sapiens (Human)
- Mus musculus (Mouse)
- Danio rerio (Zebrafish)
- Pan troglodytes (Chimpanzee)
- Canis lupus (Dog)

The model was run on 50 different genes with varied gene lengths (Small - GAPDH, Medium - BRCA1, Large - PKD1). The model provided us with conservation score at each element of the MSA.

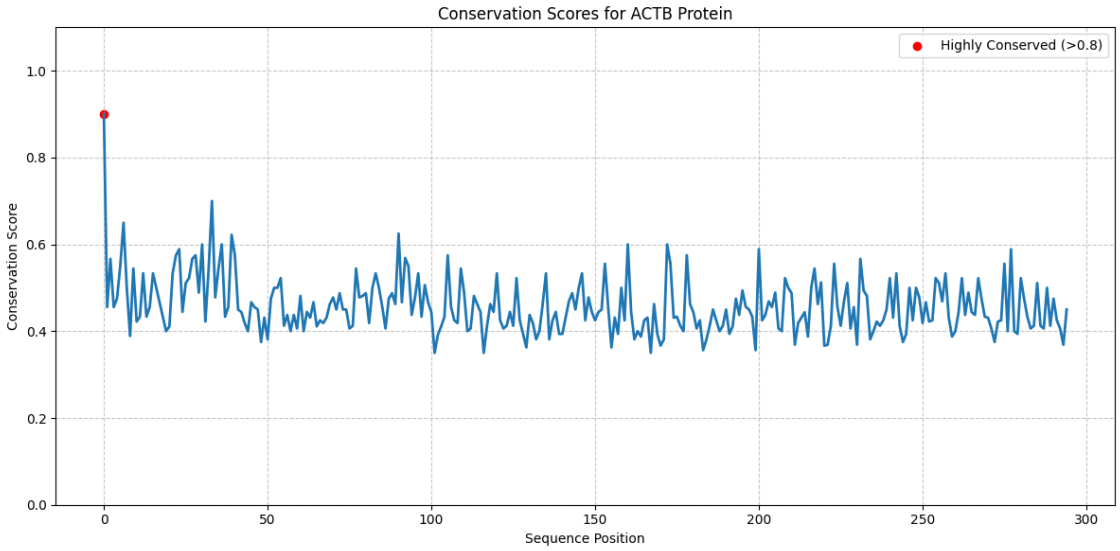

Figure shows conservation scores at all the alignment position of the MSA of Protein Sequences for ACTB gene

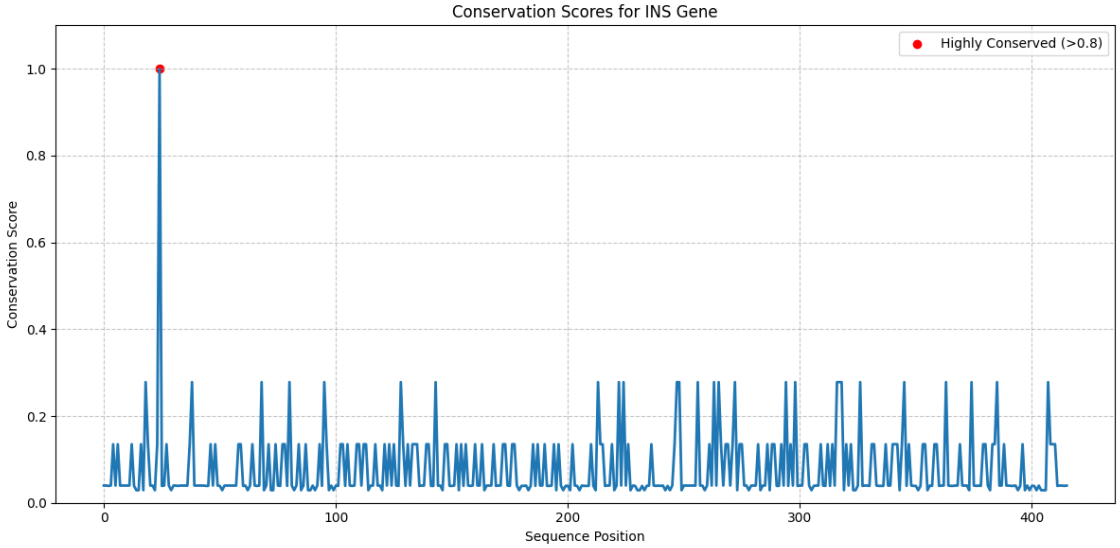

Figure shows conservation scores at all the alignment position of the MSA of Nucleotide Sequences for INS gene

Mean of those conservation score is used to get a overall conservation score for that gene in both Nucleotide and Protein sequences.

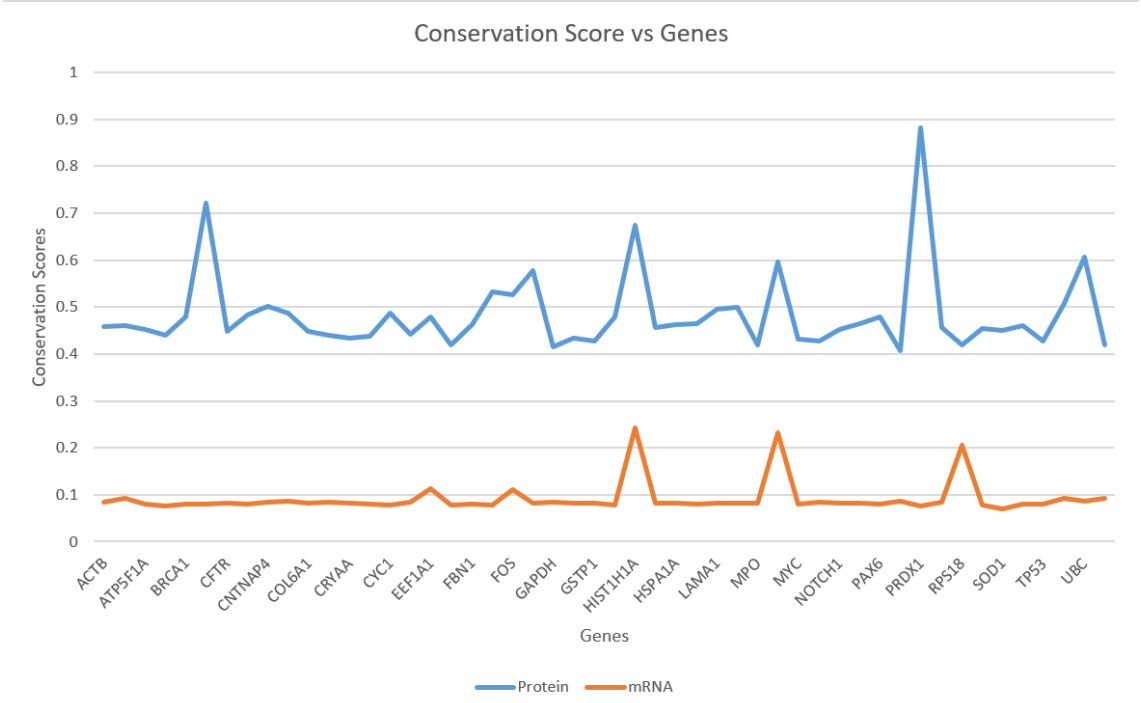

Figure shows mean conservation scores for both nucleotide and protein sequences

3.2. Result 2:

Protein Sequences of all the genes and species are uploaded to AlphaFold which uses its first model to predict a structure for that sequence from which we will give the spatial variation and then the conservation score. Mean score is used for further analysis.

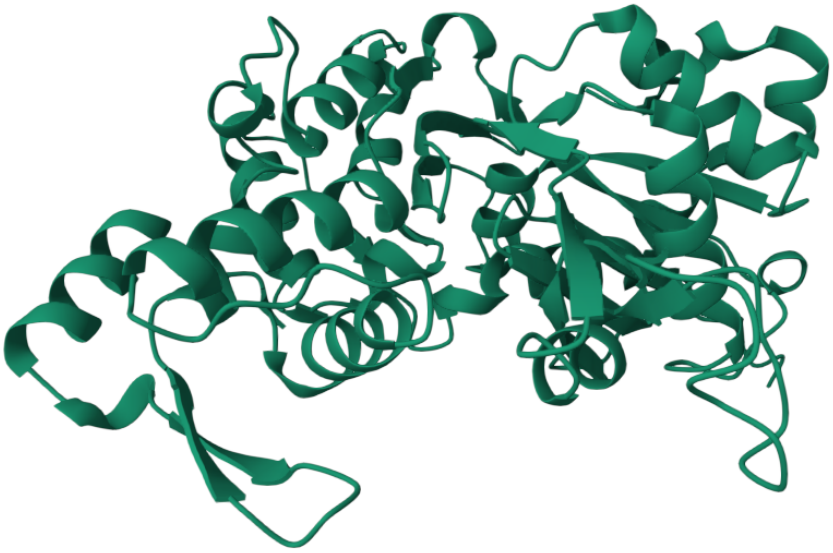

Figure shows the predicted structure for Danio rerio ACTB protein sequence

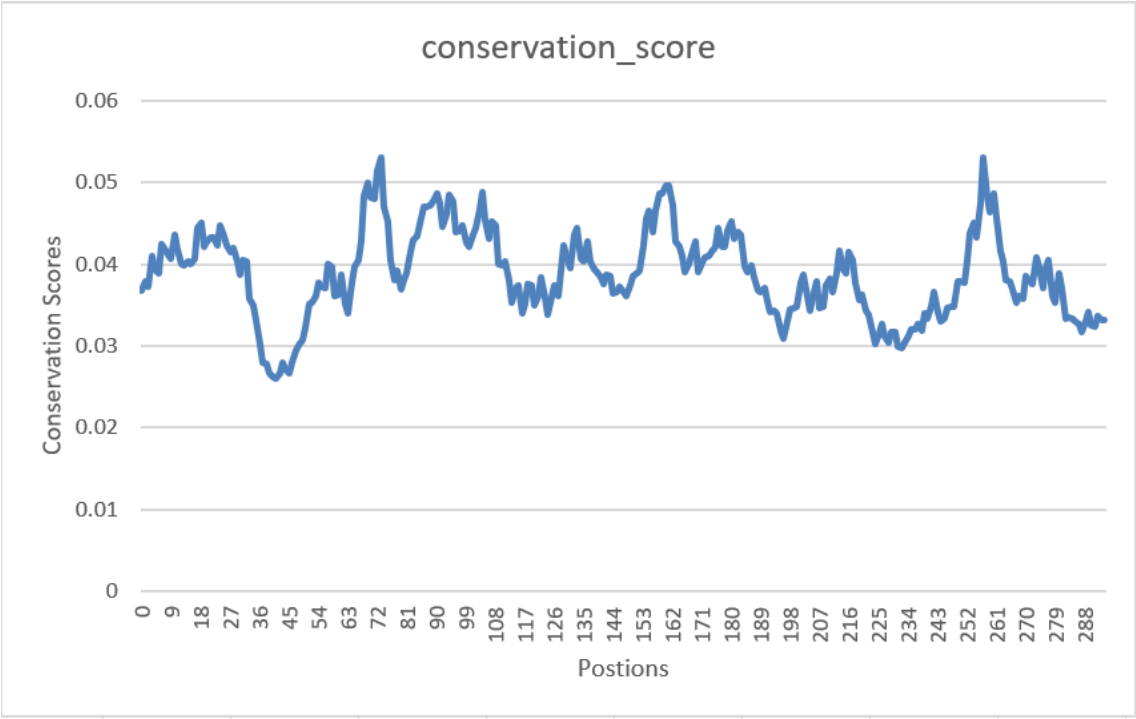

Figure shows the conservation score for the structure at every position for ACTB

3.3. Result 3:

Checking for correlations between Protein, Nucleotide and Structure we get to know that all are positively correlated to each other. Also there is more correlation between protein and structure, than structure and nucleotide, which is to be expected as protein sequences are the one used to predict the structure.

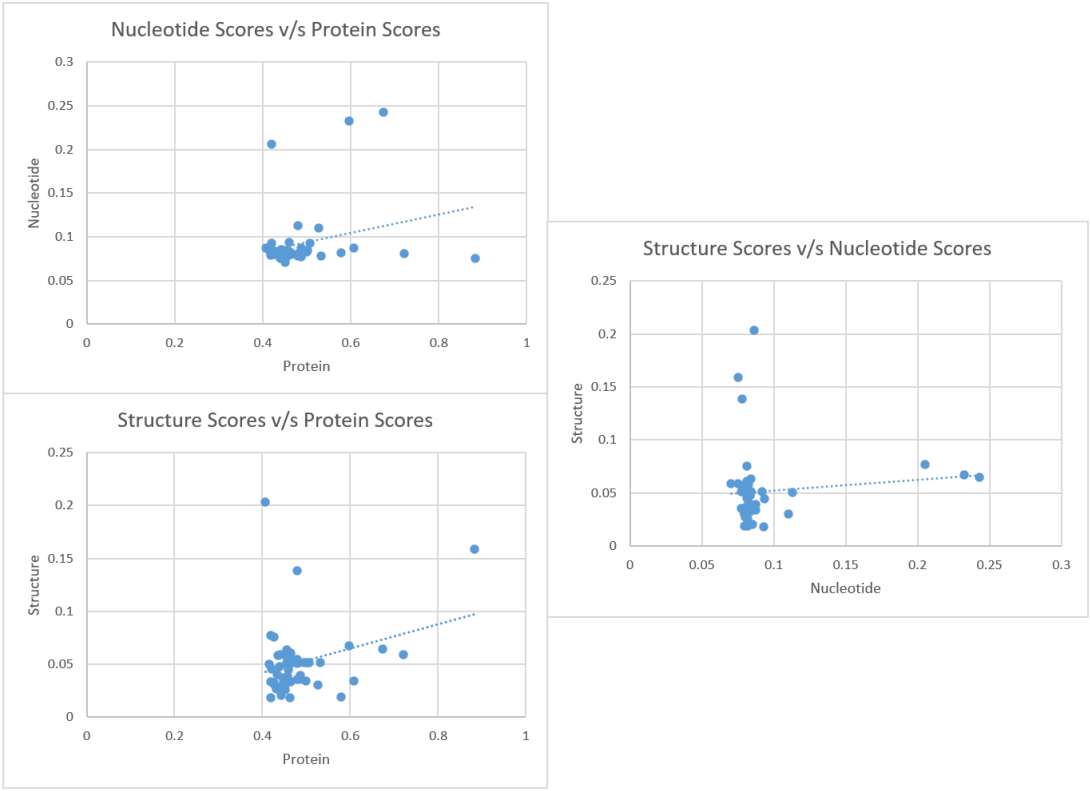

Figure shows nucleotide vs protein, structure vs nucleotide and structure vs protein

This is the correlation between all of them.

	Protein	*mRNA*	*Structure*
Protein	1		
mRNA	0.25816	1	
Structure	0.2902	0.105	1

Figure shows correlation between Protein, Nucleotide and Structure

4. Discussions

- Blosum62 should be used for sequences with similarities less than 62% which is followed in alomst all the cases.
- The model prediction come with a error of around 2 percent which minor may result in false results.
- Structure is more related to protein sequence than mRNA(Nucleotide).
- During structure prediction som sequences could not predict a structure so that case the conservation score is taken to be mean of the rest.

5. Appendix

5.1. Procedure

- Get Protein and Nucleotide sequences in FASTA format.
- Align the sequences and find the conservation score at every position.
- Upload the protein sequences to AlphaFold to predict their structures.
- Find Spatial Variation for the structures and then find the score for structures.
- Plot Protein, Nucleotide and Structure Scores against each other to know how one changes with respect to the other.
- Find the correlation between all of them.

5.2. References

- Softwares: RStudio, Excel, Vs-Code
- Datasets: NIH - National Library of Medicine, AlphaFold, UniProt
- Additional Resources: ALphaFold2 Colab File, Mol PDB viewer