HEMOGLOBIN ANALYSIS

Name: Jason Sy
Date: 6/8/2023
Advisor: Dr. Jesse Zaneveld
Course: CSS 383 – Bioinformatics

UNIVERSITY *of* WASHINGTON | BOTHELL
SCHOOL OF SCIENCE, TECHNOLOGY, ENGINEERING & MATHEMATICS

Introduction:

Hemoglobin is responsible for transporting oxygen in blood. The study of hemoglobin protein sequences of different species can provide valuable insights into the evolution and adaptation of these organisms. Variations in the hemoglobin protein sequence can create differences in its structure that can affect an animal's ability to survive in different environments.

Which brings to question: How do physiological and evolutionary adaptations of hemoglobin and its genes influence how globin protein's function and evolve.

The goal of this research is to enhance our understanding of the globin protein across different species and how it evolved through physiological and evolutionary adaptations. This can offer important implications to broaden our understanding of the function and evolution of globin proteins which can have practical applications in fields such as medicine and biotechnology.

In the article by Nei and Gojobori (Nei and Gojobori), the authors present simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. They introduce the Nei-Gojobori method and the modified Nei-Gojobori method, which utilize codon evolution to estimate the ratio of nonsynonymous to synonymous substitutions (dN/dS ratio) and provide valuable tools for studying molecular evolution and natural selection.

The paper by Sergey Kryazhimskiy and Joshua B.Plotkin(Kryazhimskiy and Plotkin) focuses on the population genetics aspects of the dN/dS ratio. Kryazhimskiy and Plotkin

explore the interplay between the population genetics factors and the natural selection factors in shaping the dN/dS ratio. Kryazhimskiy and Plotkin discuss the influence of mutation rates, recombination, genetic drift, selection strength, and gene flow on the dN/dS ratio, and highlight the importance of accurately estimating this ratio for understanding the evolutionary dynamics of protein-coding genes.

The two articles build upon each other as Nei and Gojobori, along with Kryazhimkiy and Plotkin have a shared focus on the dN/dS ratio, although with a different perspective.

In the article by Dapeng Wang, et al. (Wang et al.), the authors investigate the user of the nonsynonmous substitution rate(Ka) as a parameter for defining fast-evolving and slow-evolving protein-coding genes. They demonstrate that Ka can serve as a relatively consistent indicator for distinguishing genes that evolve rapidly or slowly. They study analyzes large-scale genomic data across different organisms and highlights the associations between Ka values, functional categories of genes, and the influence of positive selection and purifying selection.

This article by Wang et al. relates back to the paper by Nei and Gojobori, as both studies explore measures related to the rate of genetic changes within protein-coding genes.

In the article by Hardison (Hardison 2012) investigates the evolution of hemoglobins and their genes, such as the multiple genes and gene families that encode human globins and the regulation of globin genes. The article goes into detail about the mix of conserved and lineage-specific DNA in the cis-regulatory modules controlling levels and timing of gene expression, suggesting a evolutionary constraint on core regulatory functions and adaptive fine-tuning in different orders of mammals.

Hemoglobin Analysis

In the article by Kapp et al. (Kapp et al., 1995) investigates the conservation and variation in the globin protein sequence, with a goal to analyze the relationship between sequence similarity. The study finds that despite variation in amino acid substitutions and sequence volumes, there is a continuum of sequences that assume the common three-on-three alpha-helical structure.

The article by Hardison and Kapp et al. discusses different aspects of hemoglobin proteins and their evolution. With Kapp et al. focusing on the extent of amino acid substitution and variation in volume amount 700 globin sequences, while Hardison discusses the evolution and regulation of human globin genes. Despite their different focuses, the two papers suggest that the globin protein and their genes are subject to both evolutionary constraint and adaptive fine-tuning. (Hardison 2012; Kapp et al., 1995)

In the article by Milo et al. (Milo et al., 2007) investigates the relationship between physiological and evolutionary adaptations using the hemoglobin molecule as a model system. The study compared measurements of oxygen saturation curves of 25 mammals with those of human hemoglobin under a range of physiological conditions and extracted microscopic parameters using the Monod-Wyman-Changeux model. The study finds that physiological and evolutionary act on different parameters, with the main parameter that changes in the physiology of hemoglobin being relatively constant in evolution. This relates back to the two previous articles as the study analyzes the evolution and function of globin protein relates to both physiological and evolutionary adaptations.

The findings by Milo et al. presents the idea that physiological and evolutionary adaptions act on different parameters of the hemoglobin molecule, which directly relates back to Kapp et al, as it discussed amino acid substitutions and volume variations. (Milo et al., 2007; Kapp et al., 1995)

All in all, these articles explore the relationship between physiological and evolutionary adaptations in context of hemoglobin protein and their genes. Together highlighting the multidisciplinary approach necessary to understand the function and evolution of globin protein and their genes.

Through this investigation, we will attain more insight into physiological adaptations of hemoglobin in various species to understand more about how different organisms have evolved to optimize oxygen transport. And paint a better picture about the diversification of the globin gene to broaden our understanding about the relation between different species on the evolutionary tree.

Hypothesis: The dN/dS ratio of hemoglobin proteins varies between organisms based on their evolutionary distance. With closely related species having dN/dS ratio closer to 1 (neutral selection).

This is because dN/dS ratio closer to 1 indicates that the hemoglobin protein has undergone fewer amino acid changes during evolution. With values of dN/dS ratio further away from 1 suggesting the protein has undergone more amino acid changes over time, with a possible explanation being stronger selective pressures imposed by respective environments. My goal project is to analyze hemoglobin protein to see if selective pressure and evolutionary distance correlated.

Method:

I will test this hypothesis by using data in FASTA format collected from NCBI for hemoglobin DNA sequences from a range of organisms spanning a variety of evolutionary distances. And align the sequences using sequence alignment tools such as Clustal Omega. Then perform evolutionary distance calculation for each comparison of hemoglobin protein sequences and compare results with existing dN/dS tools to test our hypothesis to gauge the correlations between dN/dS ratio and evolutionary distance.

The current method will test different species to see how well programs measure evolutionary distance and create an evolutionary tree. Then branching to further investigation towards dN/dS ratio calculation tools such as FEL and CBU, to analyze how selective pressures imposed by respective environments influence changes in amino acids and see if dN/dS ratio is correlated with evolutionary distance.

- Collect FASTA data from UniProt for hemoglobin protein sequences from a range of organisms spanning a variety of evolutionary distances.
    - 2 ape species (chimpanzee, human)
    - 2 monkey species (crab-eating macaque, rhesus money)
    - 2 bovine species (cattle, water buffalo)
    - 2 rodent species (mouse, rat)
- Leveraging tools from Biopython Align the sequences using a multiple sequence alignment tool such as Clustal Omega.

Hemoglobin Analysis

```python
from Bio import AlignIO
from Bio.Align.Applications import ClustalOmegaCommandline
import os

# Input file containing sequences in FASTA format
input_file = "sequences_test1.fasta"

# Run Clustal Omega for multiple sequence alignment
output_file = "alignment.fasta"

# Delete the output file if it already exists ---> encountering errors if files sti
if os.path.isfile(output_file):
    os.remove(output_file)

clustalo_cline = ClustalOmegaCommandline(infile=input_file, outfile=output_file, ve
clustalo_cline()

# Parse the alignment output
alignment = AlignIO.read(output_file, "fasta")

# Print the alignment
print(alignment)
```

[82]

```
Alignment with 8 rows and 1713 columns
------------------ACATTTGCTTCTGACACAACTGTGT...CAA Human
--------ATCTATTGCTTACATTTGCTTCTGACACAACTGTGT...CAA Chimpanzee
------------------ACACTTGCTTCTGACACAACTGTGT...CAA Macaque
---------------------------------------------...--- Rhesus
---------------------------------------------...--- Cattle
GCCGGGCCAGCTGCTGCTTACACTTGCTTCTGACACAACCGTGT...A-- Buffalo
-----------------------TGCTTCTGACATAGTTGTGT...C-- Rat
------------------ACACTTGCTTTTGCACACTTGAGAT...TG- Mouse
```

Figure 1: Sequence Alignment using ClustalOmega

- Calculate the evolutionary distance between the organisms using tools from Biopython

```python
from Bio.Phylo.TreeConstruction import DistanceCalculator
import pandas as pd

# Calculate the evolutionary distance
calculator = DistanceCalculator('identity')
dm = calculator.get_distance(alignment)

# Convert the distance matrix to a pandas DataFrame
df = pd.DataFrame(dm.matrix, index=dm.names, columns=dm.names)

# Format the output for better readability
pd.set_option('display.float_format', '{:.6f}'.format)  # Set float formatting

# Print the formatted distance matrix using pandas DataFrame
print(df.to_string(na_rep='', justify='right'))
```

```
              Human  Chimpanzee  Macaque   Rhesus   Cattle  Buffalo      Rat    Mouse
Human      0.000000
Chimpanzee 0.026270    0.000000
Macaque    0.079977    0.093403 0.000000
Rhesus     0.165207    0.178634 0.095155 0.000000
Cattle     0.623468    0.631057 0.624051 0.531816 0.000000
Buffalo    0.442499    0.440747 0.436661 0.485698 0.500876 0.000000
Rat        0.506713    0.517221 0.503795 0.535318 0.673672 0.564507 0.000000
Mouse      0.565090    0.577350 0.565090 0.579685 0.672504 0.575598 0.561588 0.000000
```

Figure 2: Evolutionary Distance based on Alignment

- Using Biopython and based on the calculated evolutionary distance plot a tree to show evolutionary relationships between species.

```python
from Bio import Phylo
from Bio.Phylo.TreeConstruction import DistanceTreeConstructor

# Build the tree using the neighbor-joining method
constructor = DistanceTreeConstructor()
tree = constructor.nj(dm)

# Print the tree
Phylo.draw(tree)
```

✓ 1.2s

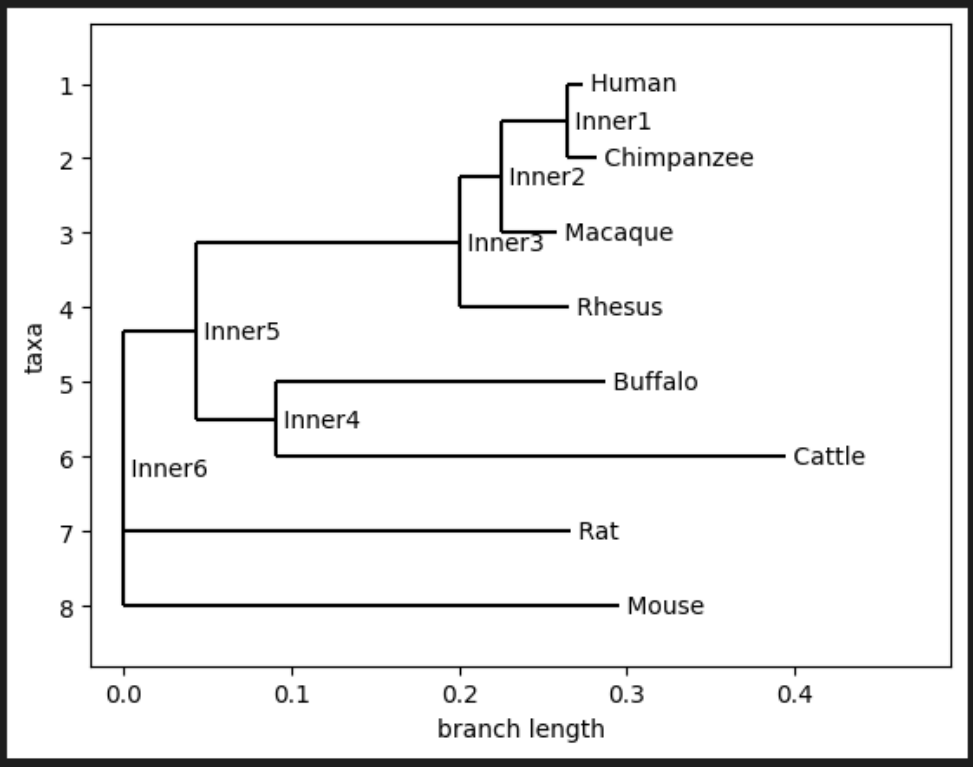

Figure 3: Evolutionary Tree, based on distance

- Calculate the dN/dS ratio of species using FEL and CBU to measure selection strength.

Hemoglobin Analysis

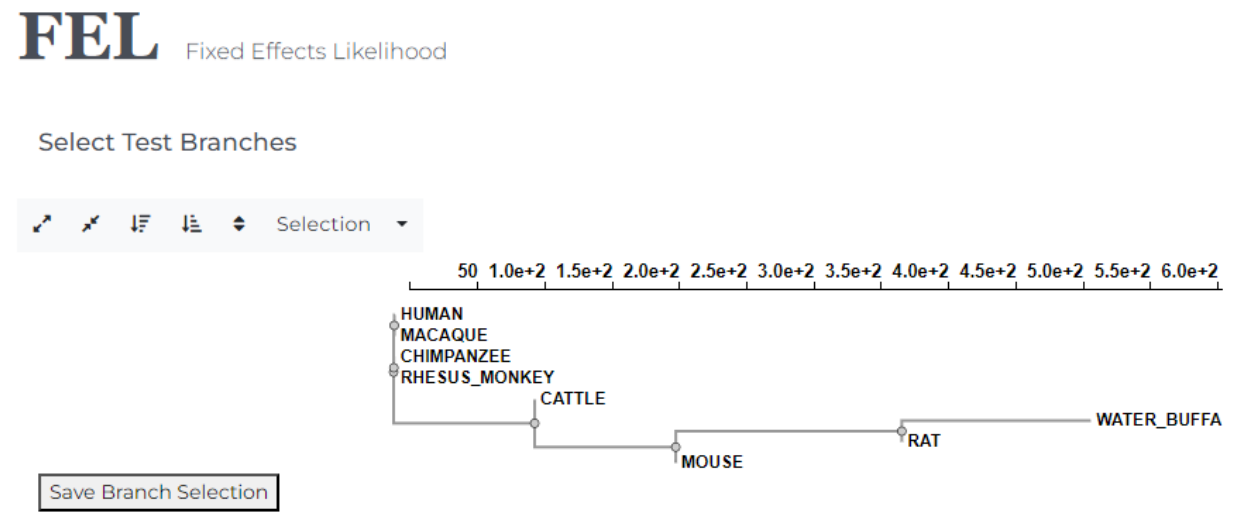

Figure 4: dN/dS tree from FEL

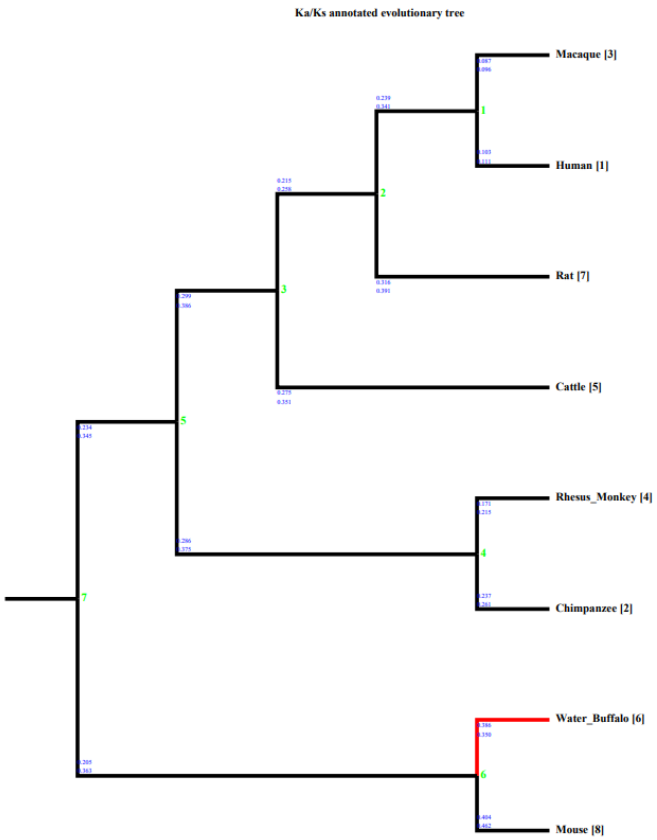

Figure 5: Ka/Ks ratio tree from CBU

- Compare trees to test our hypothesis to see if selective pressure and evolutionary distance are correlated.

Results:

In our analysis, we utilized BioPython and Clustal Omega to perform multiple sequence alignment, as shown in Figure 1 above. This allowed us to generate a text file containing the results of the alignment, capturing the similarities and differences among the sequences.

Leveraging these tools, we computed the evolutionary distances based on the aligned sequences. This distance measurement provided us with a quantitative representation of how genetically distinct or related the species are. To organize the data, we used Pandas, to visualize and interpret the outputs.

We were also able to generate an evolutionary tree based on those distances, as depicted in Figure 3 above. BioPython's tools enabled us to construct this tree, which visually represents the branching patterns and relationships between the species based on their evolutionary distances. By examining the tree, we could gain a clearer understanding of the evolutionary history and connections among the species under investigation.

To further explore and test our hypothesis, we employed additional tools such as FEL and CBU. These tools allowed us to compare the evolutionary tree, derived from evolutionary distance, with patterns of the tree based on dN/dS ratio. By examining potential correlation between dN/dS and evolutionary distance, we aimed to uncover any possible links between selective pressure and evolutionary distance.

Hemoglobin Analysis

From the three trees we gather, it seems dN/dS is not very accurate to outline evolutionary relationships. And from this we can infer that evolutionary distance and dN/dS ratios are not strongly correlated. The findings emphasize the importance of considering multiple factors and conducting further investigations to unravel the intricate dynamics of molecular evolution and the relationship between species.

Discussion:

We utilized BioPython library to leverage tools to test evolutionary distance and created a tree. Which shows that if we have a sequence of amino acids, we can determine the evolutionary relationship among species and construct an evolutionary tree based on evolutionary distance. Also, from our findings dN/dS ratio may not be a good indicator for measuring genetic divergence between species.

Based on the results we obtained from the analysis and further research into the literature, dN/dS ratio and evolutionary distance are not likely to be correlated. From corroborations from my literature, the two measures capture different aspects of the evolutionary process and may not directly relate to each other.

The dN/dS ratio focuses on evaluating the selective pressure acting on a specific gene, particularly the balance between nonsynonymous and synonymous substitutions. On the other hand, evolutionary distance provides a measure of the overall genetic divergence between sequences or species.

My research findings align with previous studies, which have also highlighted that dN/dS and evolutionary distance are independent measures that offer insights into different aspects of molecular evolution.

Based on this understanding, if I were to conduct this project again, I would propose using species that are specific to different environments, such as artic fish and tropical fish or high altitude and low altitude animals. By comparing the dN/dS ratio and evolutionary distance in these distinct ecological contexts, I can further explore and gauge the utility of dN/dS tools and their relationship with evolutionary distance.

This approach could provide valuable insights into how environmental factors shape genetic variation, selection pressures, and overall evolutionary processes. By examining species with contrasting adaptations to diverse environments, we may uncover new patterns and associations between dN/dS ratios, evolutionary distance, and environmental influences.

Some limitations currently, is the limited set of species. So, the evolutionary relationships inferred from this sample size may not accurately represent the broader picture across all species.

Also, a gap in dN/dS ratio is that different genes experience different selective pressures, so the dN/dS ratio calculation may not fully paint a full picture of the full selective pressure and may need the integration of additional biological information.

Overall, these findings highlight the importance of considering the specific context and nature of the organisms under investigation when interpreting evolutionary

relationships and the relevance of different molecular evolution measures like dN/dS and evolutionary distance.

References:

Hardison RC. Evolution of hemoglobin and its genes. Cold Spring Harb Perspect Med. 2012 Dec 1;2(12):a011627. doi: 10.1101/cshperspect.a011627. PMID: 23209182; PMCID: PMC3543078. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3543078/

Kapp OH, Moens L, Vanfleteren J, Trotman CN, Suzuki T, Vinogradov SN. Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume. Protein Sci. 1995 Oct;4(10):2179-90. doi: 10.1002/pro.5560041024. PMID: 8535255; PMCID: PMC2142974. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2142974/

Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet. 2008 Dec;4(12):e1000304. doi: 10.1371/journal.pgen.1000304. Epub 2008 Dec 12. PMID: 19081788; PMCID: PMC2596312.

Milo R, Hou JH, Springer M, Brenner MP, Kirschner MW. The relationship between evolutionary and physiological variation in hemoglobin. Proc Natl Acad Sci U S A. 2007 Oct 23;104(43):16998-7003. doi: 10.1073/pnas.0707673104. Epub 2007 Oct 17. PMID: 17942680; PMCID: PMC2040440. https://pubmed.ncbi.nlm.nih.gov/17942680/

Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986 Sep;3(5):418-26. doi: 10.1093/oxfordjournals.molbev.a040410. PMID: 3444411.

Sergei L. Kosakovsky Pond , Simon D. W. Frost, Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection, *Molecular Biology and Evolution*, Volume 22, Issue 5, May 2005, Pages 1208–1222, https://doi.org/10.1093/molbev/msi105

Wang D, Liu F, Wang L, Huang S, Yu J. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. Biol Direct. 2011 Feb 22;6:13. doi: 10.1186/1745-6150-6-13. PMID: 21342519; PMCID: PMC3055854