Comparison of features in protein residue interaction networks to infer relatedness among hemoglobins

Colton Avila, Jacob North, and Lucas Stella

CS446 Final Project

Fall 2020

Abstract

It is challenging to compare proteins with close mutants or other proteins using structural techniques, a fact which hampers studies of structural homology. To approach this ability, several homologous structures of the protein hemoglobin were featurized into residue interaction networks (RINs), split by interaction type (by interaction counts, 3-motifs, and 4- motifs for hydrogen-bonding networks, van der Waals interaction networks, π-π-stacking networks, and ligand interaction networks), then feature vectors were calculated using network methods. A pairwise comparison on the whole collection was calculated using cosine similarity and a single resulting feature average methods was obtained for all nine metrics. Using the null hypothesis that the differences in the distributions crossed zero for each pairwise comparison, Welch's t-test with $\alpha = 0.05$ confirmed that significant differences between organism classes (mammal, bird, fish, crustacean) existed for certain metrics. These significances hinged on the inclusion of the π-π-interaction network, which has the highest evolutionary fluctuation rate because it contains the fewest participating residues. Future work will center discovering which residues were critical for feature importance by node removal.

Motivation

- Hard to **compare two protein structures with differing sequence**
  - RMSD requires same sequence
- Protein structures can be **represented as residue interaction networks (RINs)**
  - Nodes are *residues* (or *atoms* on residues)
  - Edges are *interaction types*
  - A single residue interaction network may be split into several based on interaction
    - Hydrogen bonding network
    - Van der Waals interaction network
    - Ligand interaction network
    - π-π stacking interaction network

Definitions

Class = type of host animal

- E.g. fishes, mammals, crustaceans

Evolutionarily distant proteins = proteins in different classes

- E.g. perch (class = “fish”) and horse (class = “mammal”)

Evolutionarily similar proteins = proteins in the same class

- E.g. : perch (class = “fish”) and Antarctic fish (class = “fish”)

General hypothesis

If two proteins are evolutionarily distant, their sequences, structures and **their RINs should also differ**. Then their evolutionary distance may be estimated by **comparing the interaction and vertex motif counts between the two structures** using **cosine similarity.**

H_0 and H_1

H_0: The difference between different classes' similarity distributions **($\bar{x} \pm \sigma$) will overlap zero.**

H_1: The difference between different classes' similarity distributions **($\bar{x} \pm \sigma$) will not overlap zero.**

Will test using Welch's 2-sample t-test

The proteins

Variants of hemoglobin (Hb), a delightful oxygen-binding protein found in the blood which allows us to breathe

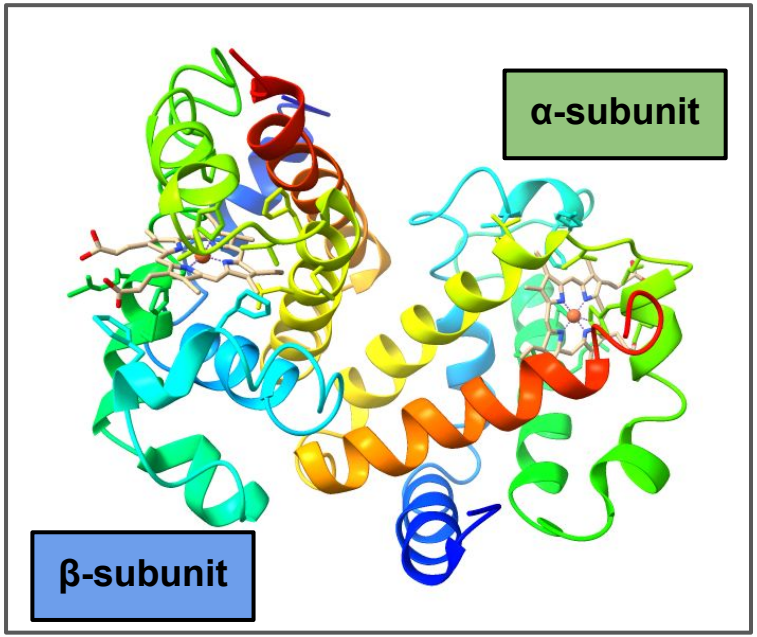


Figure 1: α and β subunits of type III hemoglobin from deer.

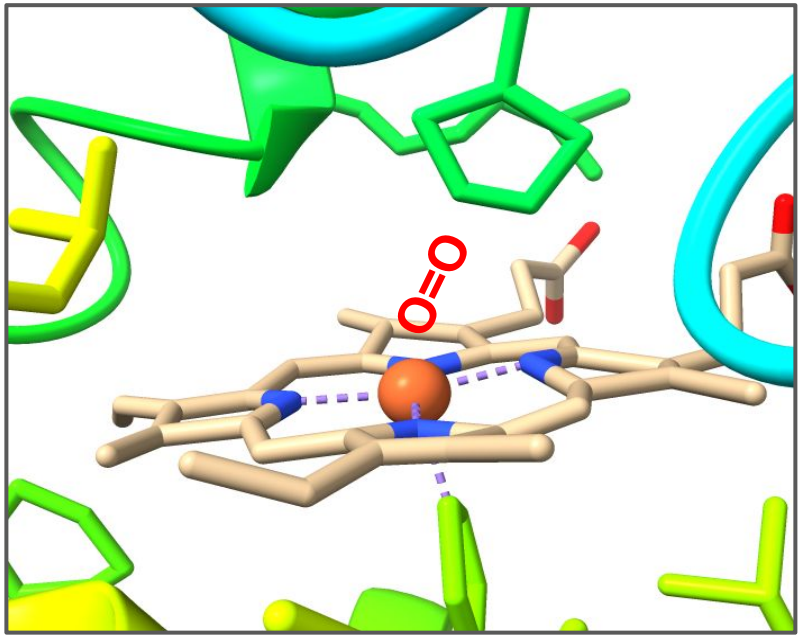


Figure 2: Heme ring of type III hemoglobin from deer with O_2 bound.

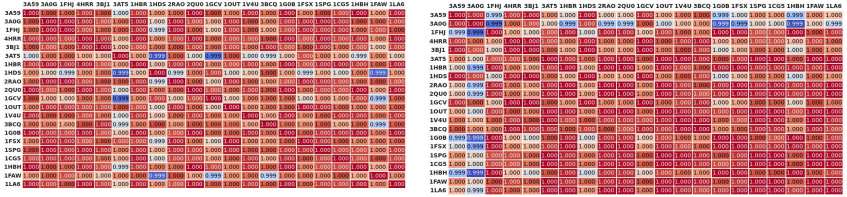
Methods

- Download homologous protein structures from the `RCSB PDB`[1]
- Retrieve residue interaction networks from `RING 2.0`[2] webserver
- Separate each network into four networks by interaction type
- Calculate network parameter vectors using `igraph`[3]
  - Interaction counts
  - 3-motifs
  - 4-motifs
- Pairwise calculate cosine similarity between all RINs
- Visualize by colormap
- Test significance by Welch's t-test

Cross-correlation matrices

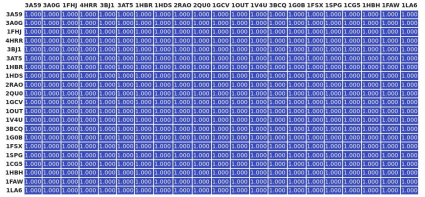
Interaction counts

Hydrogen bonding

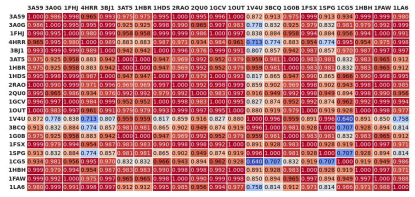


van der Waals

Ligand

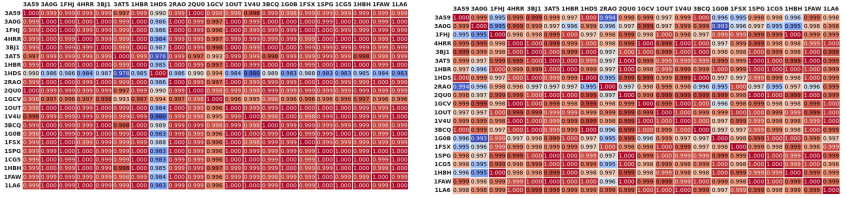


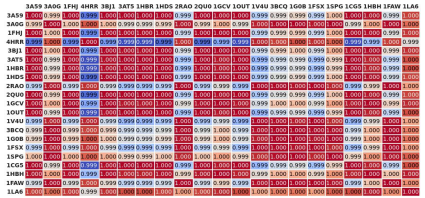
π-π stacking



3-motifs

4-motifs





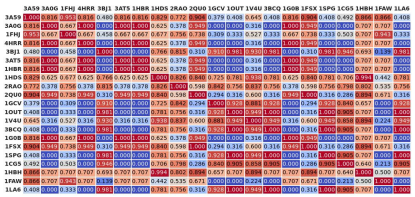
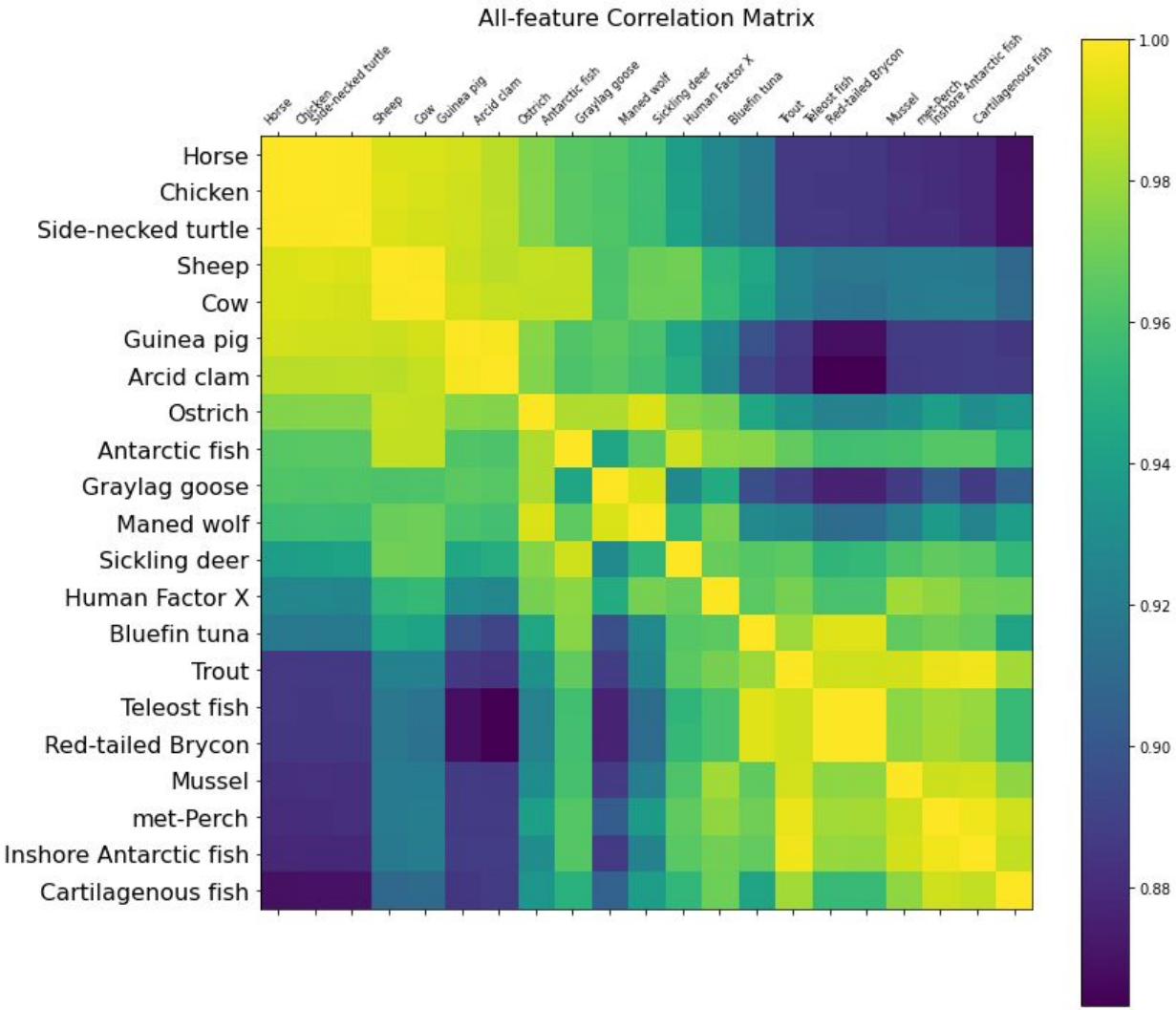


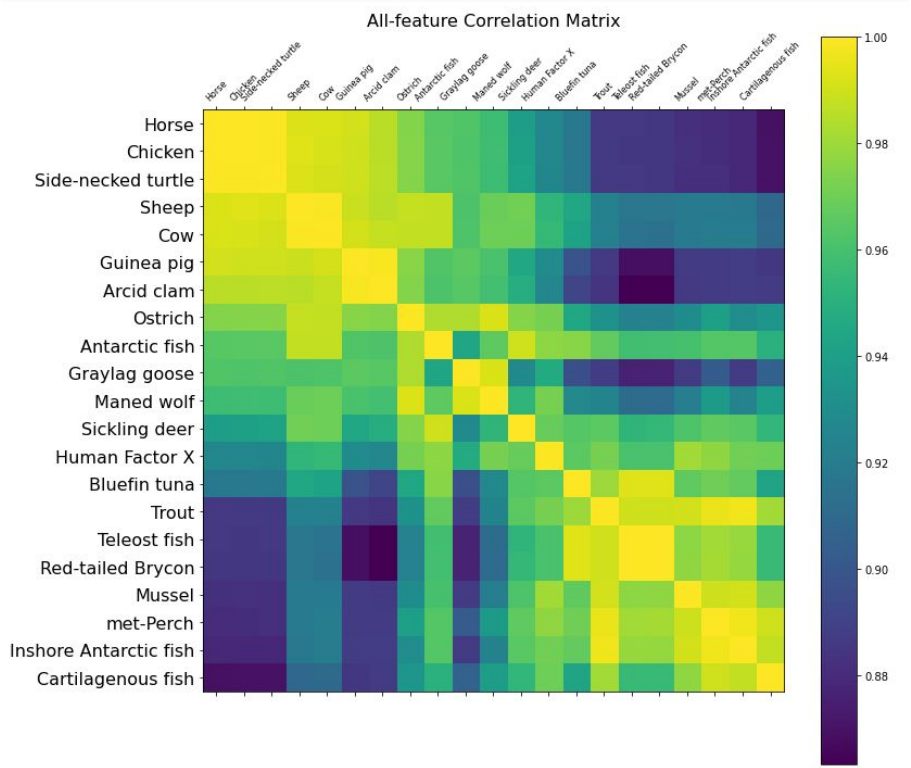
Figure 3: Cross-correlation matrices calculated from interaction counts for the whole network, and 3-motifs and 4-motifs for each of hydrogen bonding, van der Waals interactions, ligand interactions, and π-π stacking interactions, creating nine matrices total. Most showed great similarities between proteins with very small deviations (from 0.99-1.00) whereas π-π stacking showed the greatest deviation between individuals. Some faint clustering behavior is visible, most of which (aside from π-π stacking) did not correspond to differences between defined classes.



Cosine correlation Feature average

Figure 4: Average Cross-correlation matrix calculated from interaction counts for the whole network, and 3-motifs and 4-motifs for each of hydrogen bonding, van der Waals interactions, ligand interactions, and π-π stacking interactions. The resulting average dataframe was sorted with distance from Horse hemoglobin, producing the pattern visible to the left. Yellow squares around the diagonal indicate rough relatedness between species, i.e. class differences: for example, mammals and land-dwelling or air-breathing animals are separated from the fishes and water-dwelling animals. Intermediary species include wild animals, while more extremes include domesticated animals.

Structural (correlation matrix) and evolutionary (cladogram)



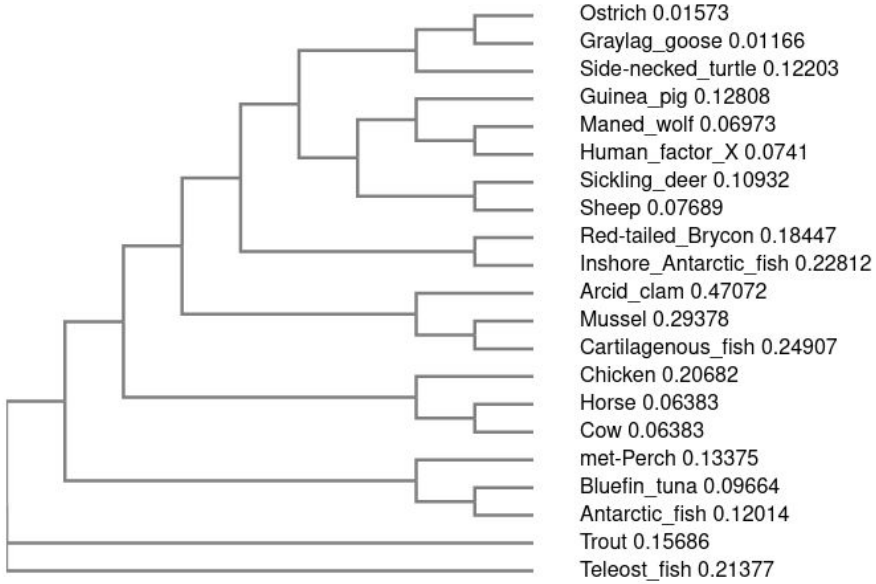
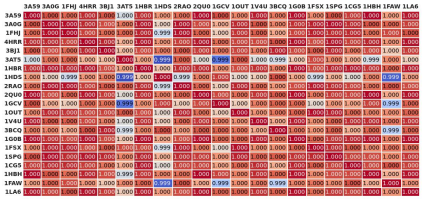
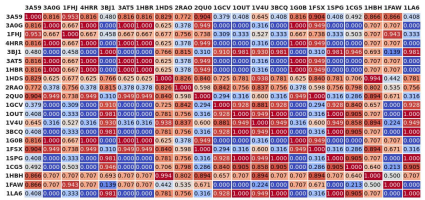


Figure 5: Comparison of average cross-correlation graph with cladogram for these proteins. Two rough groups are observed (mammals and fish) with structural analysis, but these differences are less obvious for the cladogram.

Hypothesis Testing





What is the probability that the differences in means between all classes will be zero?

Welch's 2 Sample T-test : Found significant evidence (p-values) that there are differences in all- feature cosine similarities between classes, but not within classes

	3-motif π-π stacking	4-motif π-π stacking	Total Correlation	3-motif hydrogen bond
Bird - Fish	**0.0143**	**2.627 E -9**	**5.494 E -9**	**0.956**
Bird - Mammal	**0.775**	**0.802**	**0.913**	**0.719**

Differences stem from the pi-pi stacking and the 4-motif analysis. For instance, only the Fish - Turtle comparison is rejected by the test when comparing three-motif hydrogen bonding cosine similarities between the classes.

Conclusions

- Less-frequent interactions (π-π) caused more pronounced dissimilarities
  - Higher evolution rate for this type of interaction
  - May or may not be useful for structural differences between less-related proteins
- More complex (4-motif) schemes lead to higher dissimilarities
  - May signify slight differences in secondary structures
- Comparisons of fishes with other classes generally yields significant comparisons
- Smaller difference between 3-motifs structures between species & classes
- Deer sickling mutation may have caused the structural differences in interaction number and 4-motif hydrogen bond counts
  - Homology modelling showed little structural perturbation, so further investigation is necessary

Resources

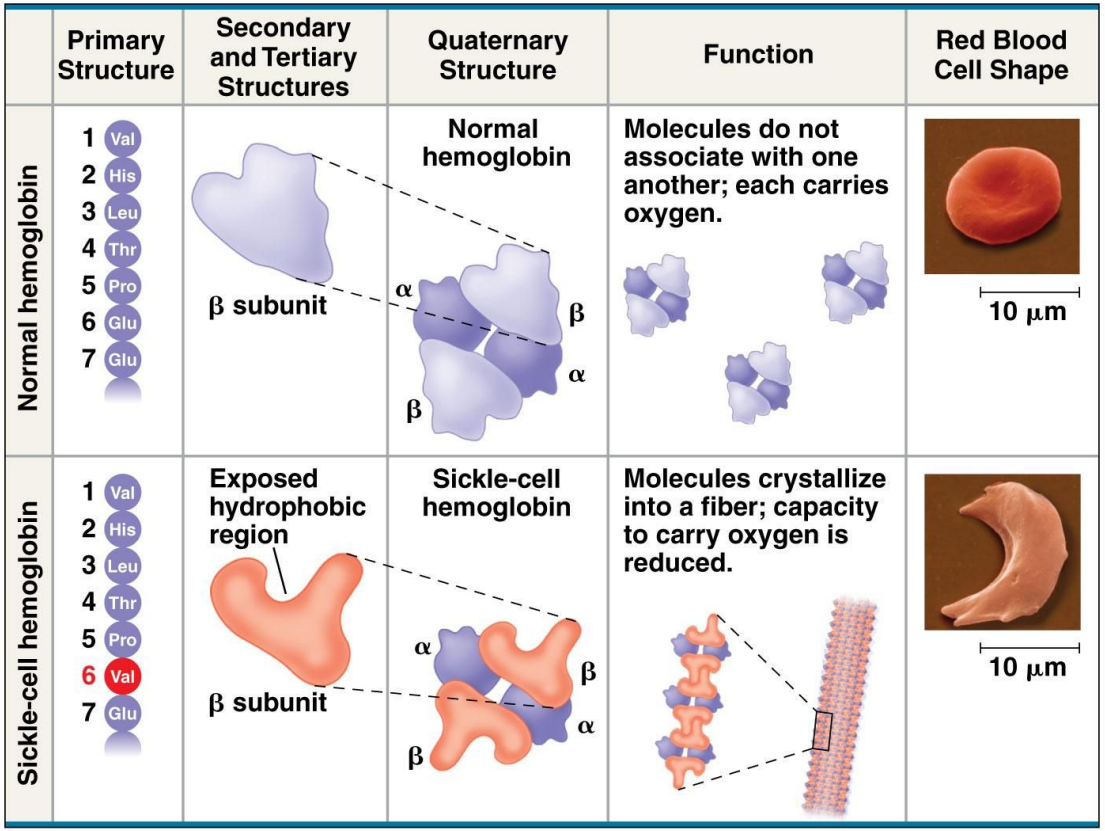
1. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. Nucleic Acids Res 2000, 28 (1), 235–242. https://doi.org/10/c7g.

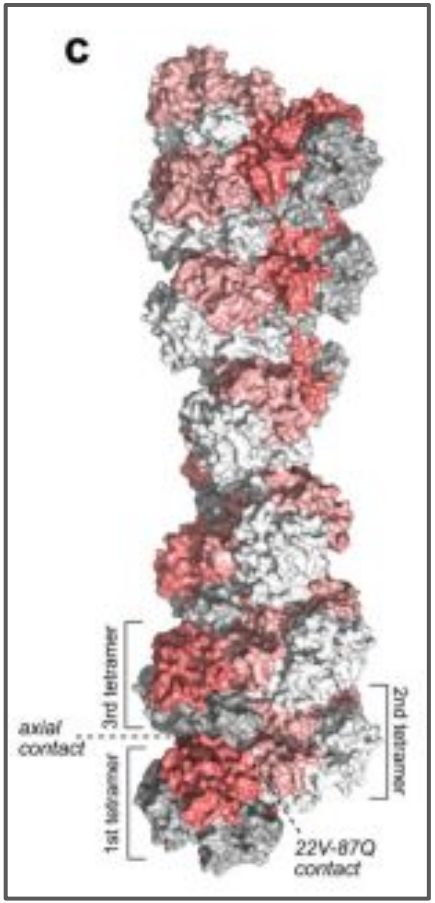
2. Piovesan, D.; Minervini, G.; Tosatto, S. C. E. The RING 2.0 Web Server for High Quality Residue Interaction Networks. Nucleic Acids Res 2016, 44 (W1), W367–W374. https://doi.org/10.1093/nar/gkw315.

3. Csardi, G.; Nepusz, T. The Igraph Software Package for Complex Network Research.

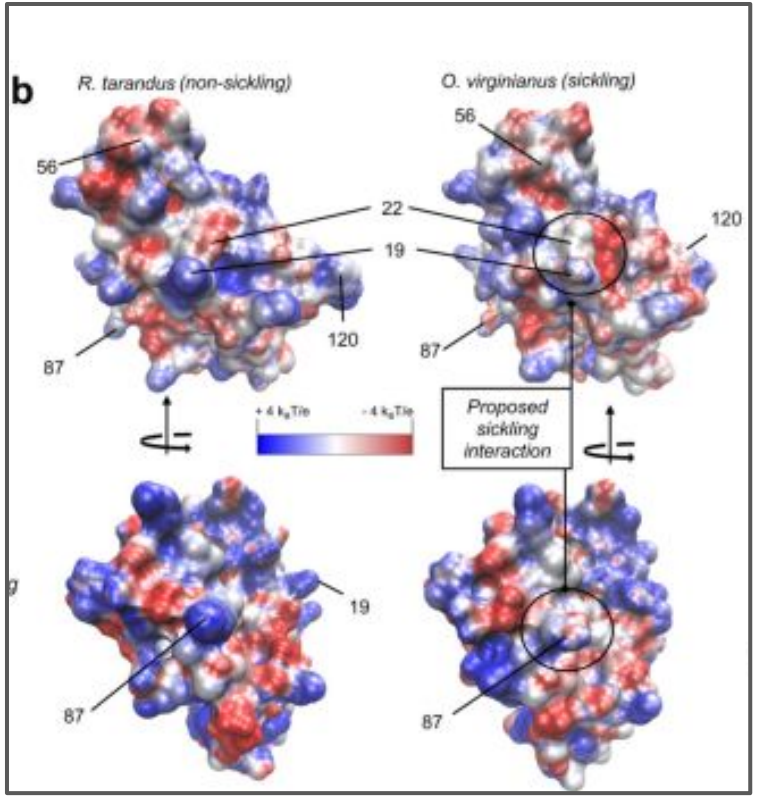
4. Esin, A.; Bergendahl, L. T.; Savolainen, V.; Marsh, J. A.; Warnecke, T. The Genetic Basis and Evolution of Red Blood Cell Sickling in Deer. Nat Ecol Evol 2018, 2 (2), 367–376. https://doi.org/10.1038/s41559-017-0420-3.

Supplementary slides





Esin *et al.* (2018)[4]



Esin *et al.* (2018)[4]

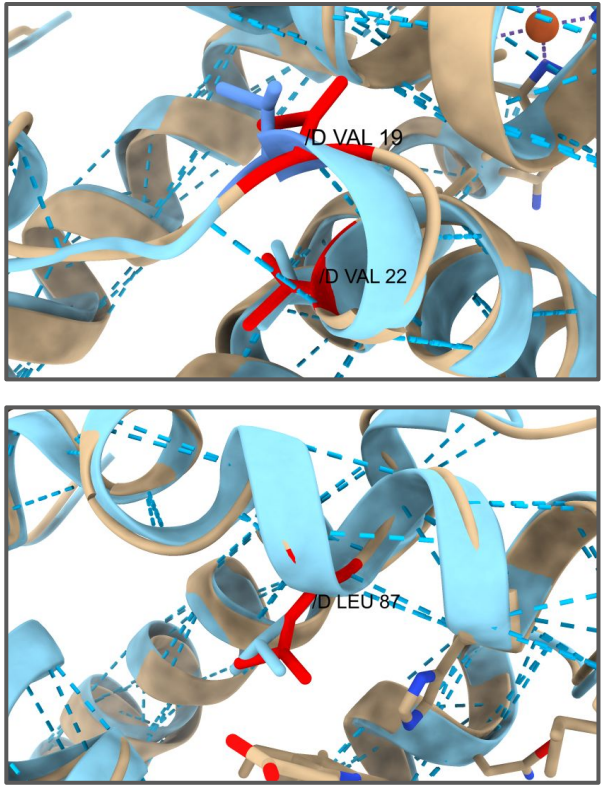


Figure 6: A structural view of three mutations implicated in enhanced sickling proclivity in deer (teal) versus a homology model from other proteins (khakhi). A lack of significant deviation in local hydrogen bonding suggests this may not indicate local structural changes but rather another feature causing hydrogen-bond motif and interaction count dissimilarity. RMSD: 0.951 Å. Calculated with UCSF Chimera.

Future directions

- **Identify important residues for each feature using node removal which minimizes cosine similarity**
- **Scale dataset and automate conversion of input data**
  - Integrate RING 2.0 standalone into Python script
- **Broader analyses of protein structure**
  - Analyze representative folds from SCOP database to investigate similarities and differences between diverse protein structures
- **Measure similarities with additional network metrics that are less noisy by virtue of frequency in the structure**