Group 1 report: protein folding

Saketh Marrapu and Matthew Russell

December 12, 2025

1 Background

We begin by summarizing the background information in our proposal. Proteins are a class of biomolecules made up of up to twenty types of amino acids arranged in one or more chains. This suggests a network structure: each individual amino acid will be a node (often referred to as a *residue* in the field), and two nodes will be adjacent in the network if the corresponding residues have backbone C_α atoms that are closer than 7 Å. This is the method used by Chakrabarty and Parekh [3]. The cutoff of 7 Å comes from that paper, but roughly speaking, this is the maximum distance that two backbone C_α atoms from different residues can be to allow for the possibility of interaction between the side chain atoms.

2 Centrality of amino acids

For this part of the project, we chose a collection of proteins contained in the Protein Data Bank [1], available at `https://www.rcsb.org/`. Initially, we had considered only myoglobin proteins, as there are several hundred myoglobin proteins available in the PDB. These would come from different species, or there could be multiple variants of myoglobin from the same species (due to mutations). However, the differences between them was relatively small, and thus this was not a good direction to go in.

Instead, we considered six families of proteins: myoglobin, lysozyme, calmodulin, carbonic anhydrase, phosphoglycerate kinase (PGK), and Alpha-hemoglobin-stabilizing protein (AHSP). From each family, we chose five examples of each from the Protein Data Bank, giving us a total of thirty proteins to consider. From each of these 30 proteins, we computed the 10 most central residues (using eigenvector centrality).

At this stage, for each of the twenty amino acids, we computed the expected number of times that it would appear among the 10 most central residues in each protein if the amino acids for that protein were randomly assigned to the nodes (once the underlying network was fixed). For one protein, this would be $\dfrac{10 \cdot \#(\text{occurrences of amino acid in protein sequence})}{\text{total length of protein sequence}}$; simply sum that up over all 30 proteins. We then created a scatterplot of the actual number of times each amino acid was one of the most central against this expected number of times it would be most central. The online NAPS tool [3] was used to create the networks and determine the eigenvalue centralities. Results are given in Figure 1. The code to determine expected values and to create the scatterplot is in `central.R`.

Compared to the baseline, the three amino acids that were most likely to be central in our analysis were glycine, leucine, and valine. Meanwhile, the four amino acids that were least likely to be central were asparagine, aspartate, lysine, and proline. This is interesting from a biochemistry perspective: three of the four least likely to be central proteins (asparagine, aspartate, and lysine) are all electrically

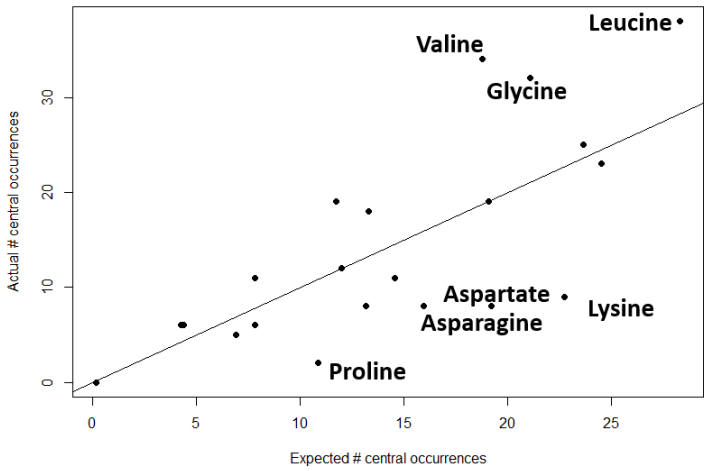

Figure 1: Frequency of occurrence of amino acids as very central residues

charged and/or polar. However, the three that were most likely to be central (glycine, leucine, and valine) are all both neutral in charge and nonpolar.

3 Comparison of hemoglobin and myoglobin

Hemoglobin (Hb) and myoglobin (Mb) are two globin proteins that both deal with oxygen. More specifically, hemoglobin is responsible for the transport of oxygen through blood to cells all throughout the body, while myoglobin stores oxygen molecules until needed. Despite both dealing with oxygen, there are two key differences to them.

The first big difference between hemoglobin and myoglobin is an ability unique to Hb: letting go of oxygen. Hemoglobin is a transporter protein, meaning it must at some point let go and deposit oxygen; however, myoglobin has no need to do this. This means there must be some structure or residues in hemoglobin that allows it to create an alternate state that does not solely hold onto oxygen.

Secondly, hemoglobin is a tetrameric protein. This means that a typical it is comprised of four chains: two α chains and two β chains. Myoglobin, on the other hand, just has one chain. However, one single hemoglobin chain is pretty similar in size and structure to a myoglobin molecule. Thus, we ended up comparing the α_1 chain of hemoglobin to a single molecule of myoglobin. We attain this by looking at a "shared core" of residues that are aligned by taking the sequential order of residues for each chain (the α_1 of Hb and the singular Mb chain) and matching the residues that exist in both and aren't filler or a gap. The code to obtain this shared core is in `datacollection.py`.

We first created network representations of the hemoglobin α_1 chain and the myoglobin molecule. These networks are depicted in Figures 3 and 4; descriptive statistics for the networks can be found in Table 2. Then, we performed spectral clustering on each protein separately. In each case, this revealed five clusters of residues. For hemoglobin, these five clusters were:

1. $\alpha_1\beta_1$ **interface**: This "holds" the protein together. It is where the α_1 chain is adjacent with the β_1 chain (remember that a hemoglobin molecule is comprised of four separate chains).

2. **Heme pocket**: Heme is an iron-containing molecule that binds to oxygen. The heme pocket is the part of the hemoglobin molecule where the heme (carrying oxygen) fits in.

3. **Allosteric switch**: Hemoglobin exists in two conformational states: the T state (tense) that rejects oxygen and the R state (relaxed) that accepts oxygen. This is the part of the protein that controls which of these states it is in.

4. **Scaffold**: This provides structural stability to the chain.

5. **Bridge**: This is the part of the chain that connects the scaffold to the active site.

See Figures 5 and 6 for more detail. For myoglobin, the five clusters are:

1. **E-F-A Corner** - This region is the entrance into the heme pocket, and stabilizes it as oxygen enters the chain.

2. **Scaffold**: As in the hemoglobin chain above, this provides structural stability.

3. **Heme-Pocket** - Much like in hemoglobin, this is the part of the molecule where the heme binds to the oxygen, attaching it to the protein.

4. **Hinge** - This chain of myoglobin acts as a hinge that allows myoglobin to be more flexible. Notably, in hemoglobin this cluster acts as an anchor to stabilize it.

5. **Exposed Surface** - This is an exposed surface that interacts with the solvent (usually water) and doesn't do much. (In hemoglobin, these residues make up the $\alpha_1\beta_1$ interface.)

See Figures 7 and 8 for more detail.

By looking at this shared core between the Hb and Mb chains, we get two networks with similar structures. In fact, biologically, they share a structure known as a globin fold. The shared globin structure could be considered as the common invariant subspace (V) while the specific residue changes that allow for differing functionality could be $R^{(Hb)}$ and $R^{(Mb)}$ in a common invariant subspace independent edge (COSIE) model. This would mean $A^{Hb} \approx R^{Hb}V^T$ and $A^{Mb} \approx R^{Mb}V^T$, which satisfies the assumptions for the COSIE model, and from this we can use spectral methods. We use Multiple Adjacency Spectral Embeddings (MASE) here specifically so that it aligns the geometries of both embeddings in order to use the Euclidean distance to find differing functions between specific residues.

The visualization of both embeddings in the same space is pictured in Figure 2. Many of the corresponding residues between the two proteins had similar latent positions. However, our results indicated that there were some specific residues in the corresponding protein sequences where there was a large distance between the latent positions of the corresponding residues. Interestingly, all three of these outliers have interesting biochemical properties:

- **Residue 137** - *Tyrosine-140 (hemoglobin) vs Tyrosine-146 (myoglobin)* - This residue is useless in myoglobin, but in hemoglobin it is particularly significant because it enables important function in Allostery by triggering the T state as it forms a hydrogen bond [2].

- **Residue 54** - *Lysine-56 (hemoglobin) vs Lysine-62 (myoglobin)* - In myoglobin this residue faces the water and does little. In hemoglobin this is part of the E-helix that forms the $\alpha_1\beta_1$ interface, essentially working as the glue to make it a tetrameric protein.

- **Residue 82** - *Serine-84 (hemoglobin) vs Alanine-90 (myoglobin)* - These correspond to the residue that pulls the F-helix when oxygen is bound. In myoglobin this movement is small and local, but in hemoglobin this movement is transferred to other chains and thus has to be more connected in order to transmit this signal to other chains.

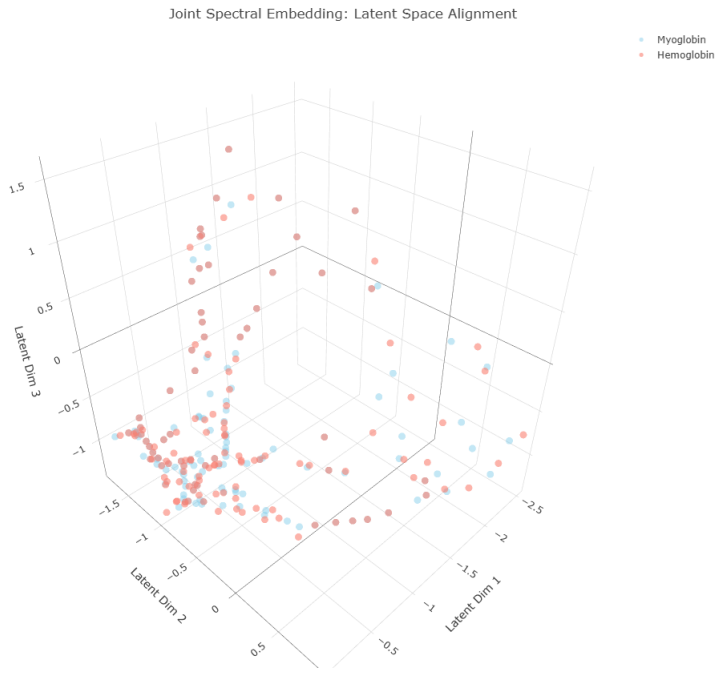

Figure 2: Joint spectral embedding: latent space alignment

There were other residues with differences in position, but these were the most important ones of the top resulting distances. For the most part these non-mentioned residues also fell into the functions mentioned above, as protein function isn't defined solely singular residues. This MASE ended up identifying residues that were part of both of the differing functions mentioned earlier between Hb and Mb. All of the code for this spectral analysis is in `spectral.R`.

4 Small-world property

One final part of our project (not covered in our presentation) was determining whether these protein networks tended to follow the small-world property. This was introduced by Watts and Strogatz [6]. Two important network statistics relating to small-world networks, as noted by Watts and Strogatz, are the network transitivity C [1] and the mean minimum path length over all pairs of vertices L. Newman and Watts [5] then state graph G with transitivity C_g and mean path length L_g is a small-world network if $L_g \geq L_{\text{rand}}$ and $C_g \gg C_{\text{rand}}$, where L_{rand} and C_{rand} are the expected values of the corresponding statistics for an Erdős-Rényi random graph G_{nm} with the same number of vertices and edges as the network in question. (Note that this is not the more common of the two definitions of an Erdős-Rényi random graph.) Humphries and Gurney [4] suggest combining the two inequalities into one, defining $S_g = \dfrac{C_g}{C_{\text{rand}}} \cdot \dfrac{L_{\text{rand}}}{L_g}$, and then simply stating that the network is a small-world network if $S_g \gg 1$. In their analysis of real networks, Humphries and Gurney classified a network as small-world if $S_g > 3$, and performed Monte Carlo simulation to make further decisions in the case $1 < S_g \leq 3$.

We carried out this process for three proteins (Paramecium tetraurelia calmodulin: PDB 1CLM,

[1] To be more precise, Watts and Strogatz's definition looked at the mean value of the clustering coefficient over all nodes in the network. We are following Humphries and Gurney [4], who compute their statistics in two ways: one along the lines of Watts and Strogatz, and the other using the transitivity. For the sake of simplicity, we only use the definition involving transitivity.

Protein	C_g	C_{rand}	L_g	L_{rand}	S_g
Calmodulin	0.5544	0.0534	7.2548	2.6552	3.7996
PGK	0.4968	0.0189	11.8430	3.1502	6.9833
Myoglobin	0.5568	0.0512	6.7603	2.6580	4.2834

Table 1: Small-world network statistics

yeast PGK: PDB 3PGK, and dwarf sperm whale myoglobin: PDB 6BMG[2]). Our code for this section is in `smallworld.R`.

For each of those proteins, we used R to compute C_g and L_g. In addition, we sampled 1000 Erdős-Rényi random graphs G_{nm} (with the appropriate values of n and m) to obtain estimates for L_{rand} and C_{rand}. Our numeric results are in Table 1. It is apparent that $S_g > 3$ for all three proteins, so, following Humphries and Gurney, we conclude that all three have the small-world property. The largest value of S_g belonged to the PGK molecule, which also was comprised of the greatest number of residues ($n = 415$ for our PGK molecule, $n = 144$ for the calmodulin molecule, and $n = 154$ for the whale myoglobin). According to Humphries and Gurney, this is unsurprising, as they observed a linear scaling of S_g with n for their real-world networks that did have the small-world property.

This suggests a couple of possible avenues for further work. First, this S_g statistic could be computed for many more proteins. We would expect that most, if not all, of the proteins that we would find from the PDB would satisfy the small-world property above. However, some proteins may have fewer internal node adjacencies. In addition, seeing if there are any proteins with remarkably large or small observed S_G statistics (among proteins of similar sizes) could lead to some intriguing observations about the biophysical properties of those proteins. A second possibility would be to see just how close these proteins would be modeled by the Watts-Strogatz model, when you start with a $k-$ring lattice for some value of k and then rewire some edges at random.

References

[1] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[2] Sandeep Chakane, Vijay Markad, Kisan Kodam, and Leif Bülow. The penultimate tyrosine residues are critical for the genotoxic effect of human hemoglobin. *Adv Exp Med Biol*, 977:351–357, 2017.

[3] Broto Chakrabarty and Nita Parekh. NAPS: Network analysis of protein structures. *Nucleic Acids Research*, 44(W1):W375–W382, 05 2016.

[4] Mark D. Humphries and Kevin Gurney. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLOS ONE*, 3(4):1–10, 04 2008.

[5] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60:7332–7342, Dec 1999.

[6] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998.

[2]The choice of the species for calmodulin and PGK comes from biochemistry tradition — they are well-studied — while one of us has a personal connection to the researcher responsible for the whale myoglobin PDB entry.

5 Additional charts and graphs

Network	Nodes	Edges	Density	Avg Deg	Trans	Avg Path	Diam
Myoglobin	138	534	0.0565	7.739	0.571	4.64	12
Hemoglobin	138	536	0.0567	7.768	0.558	4.516	10
Difference	0	-2	-0.000212	-0.029	0.0123	0.124	2

Table 2: Descriptive statistics for hemoglobin and myoglobin networks

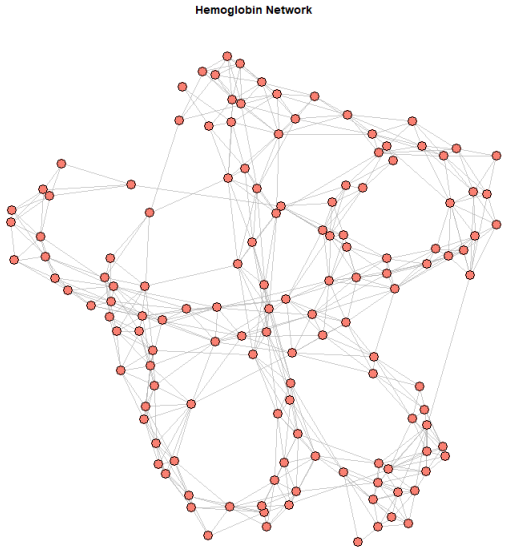

Figure 3: Network representation of a hemoglobin α_1 chain

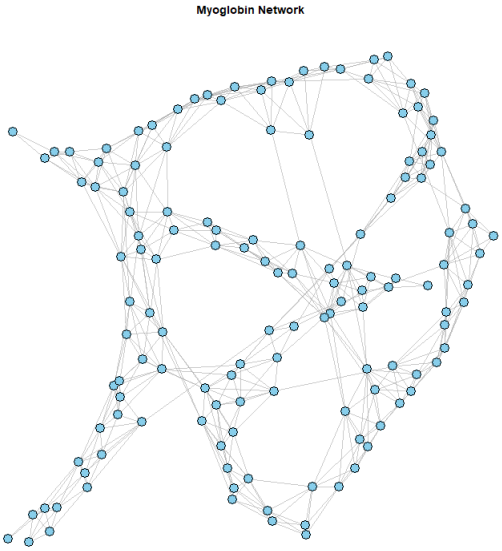

Figure 4: Network representation of a myoglobin molecule

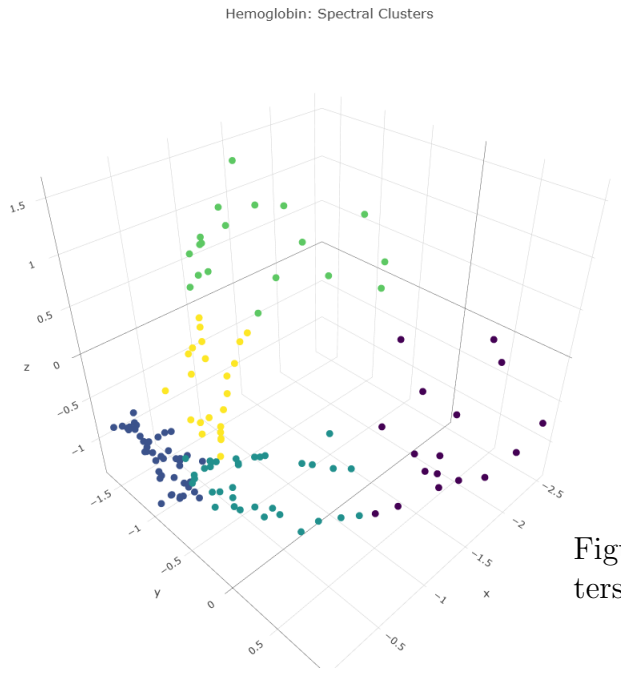

Figure 5: Hemoglobin α_1 chain clustering

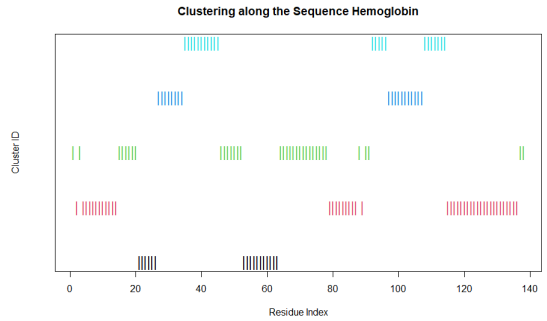

Figure 6: Sequential order of hemoglobin α_1 chain clusters

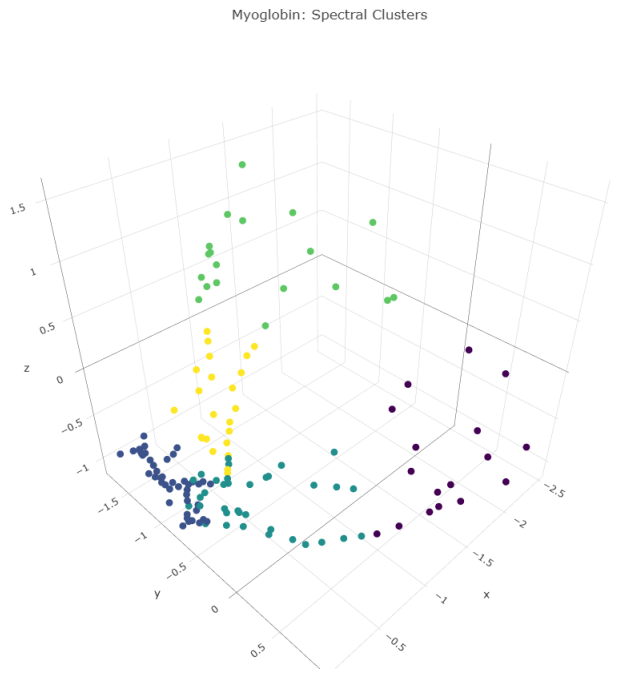

Figure 7: Myoglobin clustering

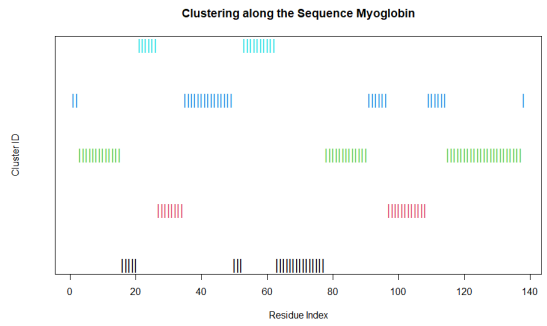

Figure 8: Sequential order of myoglobin clusters