**Comparison of Phenylalanine v. Tyrosine rotamers with**

**respect to alpha helix structure using 3D modeling software**

1. **Introduction**

Since Crick determined the central dogma (Crick, 1958), researchers have been working endlessly to determine how proteins work. We currently know that structure determines function (Berg, Tymoczko, & Stryer, 2002). So in this long process to understand how proteins work, we are currently trying to understand how the structure works. Once we determine the structure, we will have a better understanding of how it works in its own unique environment. There are a handful of ways to determine the structure of a protein, and according to PDBs protein structure submission results from the past current year, the most popular methods of determining structure are x-ray crystallography and nuclear magnetic resonance, usually called NMR (N/A, 2017). However, there is another emerging technique that has been growing the past couple of decades, and it is a hot research field in bioinformatics, cheminformatics, and computational biology: it is molecular modeling using computers.

Protein modeling with computers has been a challenge area in the field of bioinformatics since at least the 90’s. David Searls argues that perfect protein prediction is the goal of bioinformatics itself (Salzberg, Searls, & Kasif, 1998). There’s a couple of ways that protein structure is predicted, but the most popular way is using homology modeling, so basically looking at previously determined structures and finding close matches in primary sequence that will most likely cause an unknown protein to fold similarly to a previously determined one.

Making a successful method that can predict protein structure with 100% accuracy would speed up research efforts exponentially (Huang, Boyken, & Baker, 2016). A functioning solution would be able to predict any protein structure based on a primary sequence. It will be a critical step in computer aided drug discovery, simulating protein-protein interactions, and so on (?).

The critical assessment of protein structure prediction(CASP) is a biannual competition that researchers enter their protein structure prediction programs and attempt to determine the structure of a new protein. The better models predicted with 80% accuracy or above. Similar software available is HHpred, we used this in phage lab by searching for similar sequences. Matching sequences sometimes had previously discovered domains which aids in prediction of the function of our novel gene. Dr. Kellogg, my mentor, has done similar research, while his research isn’t strictly based on developing protein structure prediction software, his findings can definitely be used to forward the cause. Dr. Kellogg’s program quantifies the hydropathic interactions within molecules (Ahmed, Koparde, Safo, Neel Scarsdale, & Kellogg, 2015).

Dr. Kellogg uses a slightly different approach to the homology approach. He uses a program that analyses the hydropathic interactions in a given molecule. Kellogg’s paper uses his program called HINT! (hydropathic interactions). It measures the hydropathic interactions within molecules. Hydropathic interactions are a group of interactions that is comprised of four types of interactions: favorable and unfavorable hydrophobic interactions, and favorable and unfavorable polar interactions. His paper analyses almost 30,000 tyrosine residues in previously discovered proteins, then uses HINT to score the hydropathic interactions surrounding the tyrosines illustrated in **Fig. 1** (Ahmed et al., 2015). The key hypothesis was to analyze these residues to see if there are any recurring motifs, and they found many using HINT and clustering methods.



Tyrosines are plotted on a Ramachandran plot based on their phi-psi bond angles, this is illustrated in **Fig. 2** (“Peptide Bonds and Protein Backbones,” n.d.). A Ramachandran plot is a graph that contains the phi angle as the x-axis, and the psi angle as the y-axis, and is effectively used to predict secondary structure based on those two bond angles of a given residue (Ramachandran, Ramakrishnan, & Sasisekharan, 1963). Once the plot is complete, the plot is divided into 64 chest squares based on the phi-psi angles Kellogg’s paper only looks at the a1 chest square. The Ramachandran plot used in his paper is **Fig. 3** (Ahmed et al., 2015). My experiment uses the data from the d4 square. Once tyrosines have been sorted based on their phi-psi angles, they are clustered into hypothetical “motifs” based on their similarity metrics. The a1 chest square produced 14 unique clusters which are analyzed in the paper. In summary, his research proposes that for the given phi-psi tyrosine angles in that chest square, there are only 14 possible conformations that tyrosine can take on. This is significant because despite the fact that there may be near infinite possibilities of primary sequence, this suggests there is a *limited* number of possibilities a secondary structure may take on. And that is something worth pursuing.

**Figure 2.** Locations of the phi-psi bonds in a peptide, note that the “R-group” represents the side chain, and the omega bond is the peptide bond.

**Figure 1.** An example of HINT scoring from the same Kellogg article. Blue and red contours represent polar interactions. And green and purple contours represent hydrophobic interactions.

In my experiment, I would like to look at Phenylalanine and compare it to the Tyrosine results found in Kellogg’s paper. I have chosen to examine Phenylalanine because it is very similar to tyrosine. The goal of my experiment to build on Kellogg’s research and possibly pave the way to the next big paper.

This proposal looks at Tyrosine’s and Phenylalanine’s in the d4 ramachandran chest square. Doing Kellogg’s experiment with Phenylalanine would likely produce novel results. Yet would be too ambitious to say I could handle, and it wouldn’t be original, invalidating the goal of this semester long project. The d4 Ramachandran plot predicts a right handed alpha helix as the secondary structure with the given phi-psi angles. So my experiment effectively focuses in on tyrosines and phenylalanines in right handed alpha helices. I chose the d4 square because out of all of Kellogg’s data, the d4 square had the most tyrosines fall into it, likewise it has the most phenylalanines in it too. So in theory it gives me the most data available to work with, and when it comes to statistical analyses, a bigger sample helps predict a more accurate population parameter.

**Figure 3.** The Ramachandran plot from Dr. Kellogg’s paper. My experiment uses this data from the d4 chest square.

Nothing has been previously done to analyze these two amino acids in right handed alpha helices, however, I found papers describing a general composition of amino acids in alpha helices as well as their general locations (Pace & Scholtz, 1998; Ulmschneider & Sansom, 2001). [Implementation to be added].

My key hypothesis questions whether among the limited set of “motifs” formed by tyrosine, are there any shared with phenylalanine? The other papers support the fact that they may share structural properties [again, implementation will be added soon], as well as Kellogg’s own data set cited later in this article. Not to mention the obvious similarities between the two amino acids. What will I find by comparing these two amino acids? What major differences will we find? Will I find anything unique? Since nothing like this has been done before, I would hope to find something unique and novel. And hopefully will be able to give us a better understanding of secondary structure afterwards.

1. **The Experiment**

The goal of this experiment is to use the techniques outlined in Kellogg’s and on a smaller scale and a novel purpose to compare two entities that have not been compared before. I will be focusing in on rotamers that fall in the d4 Ramachandran plot chess square. Rotamers are effectively structural conformations that the side chain can take on. In short, I would like to HINT score all phenylalanines that fall into that same chest square. Then, calculate maps based on the HINT score. Afterwards, use the map-map similarity metric from Kellogg’s paper to score the relationship between tyrosine vs. phenylalanine maps to see if there are any similarities. According to the data set of proteins used from Kellogg’s paper, there are a total of 5376 tyrosines that fall into the d4 chest square. And there are 6364 phenylalanines that fall into the d4 chest square. This is significant because d4 contains the highest numbers of residue count per chest square. Additionally, they are similar in both data set size and characteristics. Chest square d4 is not the most populated square for every residue, but it is for both tyrosine and phenylalanine, which I believe supports my hypothesis that we may find similar clustering patterns between these two residues.

The first step would be to evaluate HINT scoring of the phenylalanine residues. This will be done by scoring all the atom-atom interactions between the molecules. Then I will use the same algorithm used in Kellogg’s paper, the HINT basis map, which takes all the atom-atom interactions for the vicinity of a given residue and computes it into a list which is then converted into a 3D object that displays interaction type and strength just as mentioned in the introduction. This will be calculated for all four interactions that make up HINT.

The method that I’m describing is the same one used in Kellogg’s paper. In his paper he scored all tyrosine residues the same way and calculated HINT basis maps for each tyrosine. The following was also done in his experiment and is what I plan on doing with the phenylalanine residues, with the exception of the comparison at the end. To quantitate the similarity of the maps, and effectively compare two residues, a composite metric was made based on the four maps made in the HINT basis mapping. Clustering will be done with the k-means method as described above [Citation and description to be added]. The difference is that the clustering will be performed comparing phenylalanine’s against tyrosines.

1. **Discussion**

In the best case scenario, the clusters would suggest that some phenylalanine and tyrosine rotamers form the same “motifs” in right handed alpha helices. This can then give reason to investigate further and possibly look at another chest square to see if there are other similarities in “motif” formation between these two residues. In the big picture, this would be good news because it would take vast universe of unexplained knowledge regarding how molecules work and shrink it down a tiny, tiny, bit [I think this sounds a bit dramatic and I’m thinking of rewording it]. Alternatively, if we find no cluster correlation between phenylalanine and tyrosine, we can heed this and focus our resources on something else. More importantly, we can build on the data generated for the d4 square and finish the rest of the Ramachandran plot of perform another clustering analysis of just phenylalanine.

Limitations and obstacles that I can foresee before beginning this experiment would be the following. It would probably be a lot of data to sift through.

[Final words and concluding sentence will be added soon]

**References**

Ahmed, M. H., Koparde, V. N., Safo, M. K., Neel Scarsdale, J., & Kellogg, G. E. (2015). 3D interaction homology: The structurally known rotamers of tyrosine derive from a surprisingly limited set of information-rich hydropathic interaction environments described by maps. *Proteins: Structure, Function and Bioinformatics*, *83*(6), 1118–1136. https://doi.org/10.1002/prot.24813

Berg, J., Tymoczko, J., & Stryer, L. (2002). Protein Structure and Function. In *Biochemistry* (5th ed.). New York: W. H. Freeman. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK21177/

Crick, F. (1958). On Protein Synthesis. *The Symposia of the Society for Experimental Biology*, 138–166. https://doi.org/10.1038/227561a0

Huang, P.-S., Boyken, S. E., & Baker, D. (2016). The coming of age of de novo protein design. *Nature*, *537*(7620), 320–327. https://doi.org/10.1038/nature19946

N/A. (2017). Refined Search on PDB. Retrieved from https://www.rcsb.org/pdb/results/results.do?tabtoshow=Current&qrid=5451C702

Pace, C. N., & Scholtz, J. M. (1998). A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical Journal*, *75*(1), 422–427. https://doi.org/10.1016/S0006-3495(98)77529-0

Peptide Bonds and Protein Backbones. (n.d.). Retrieved from http://www.biochem.ucl.ac.uk/~martin/c40/peptide.html

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*(1), 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6

Salzberg, S. L., Searls, D. B., & Kasif, S. (1998). Grand Challenges in Computational Biology. In *Computational Methods in Molecular Biology* (pp. 3–10). Elsevier Science B.V. Retrieved from https://app.knovel.com/web/view/khtml/show.v/rcid:kpCMMB0001/cid:kt00BYFA5K/viewerType:khtml/root\_slug:1-grand-challenges-in-computational-biology/url\_slug:grand-challenges-in-computational?b-toc-cid=kpCMMB0001&b-toc-root-slug=&b-toc-url-slug=grand-challe

Ulmschneider, M. B., & Sansom, M. S. P. (2001). Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta*, *1512*(1), 1–14. https://doi.org/10.1016/S0005-2736(01)00299-1