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

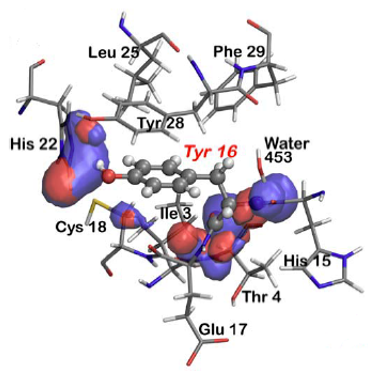
**respect to alpha helix structure using 3D modeling software**

1. **Introduction**

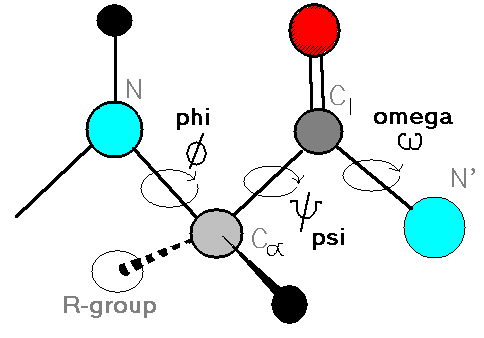
Protein modeling with computers has been a challenge area in the field of bioinformatics that dates back before the 90’s. David Searls argues that perfect protein prediction is the goal of bioinformatics itself 1. There’s a few methods of protein structure prediction, the most popular method is homology modeling. Homology modeling evaluates previously determined structures and locates close matches in primary sequence or shared domains that will most likely cause the unknown protein in question to fold similarly to a previously determined structure.

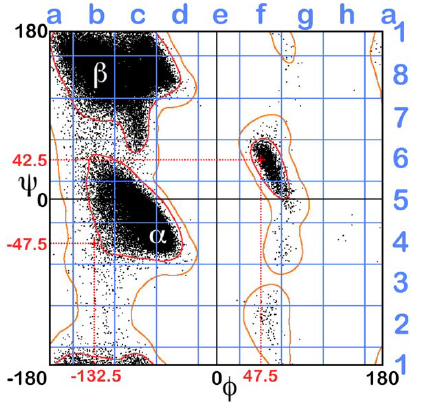
Some researchers believe a successful method that can predict protein structure with 100% accuracy would speed up research efforts exponentially 2. A functioning solution would be able to predict any protein structure based on a primary sequence. It will be a critical step in computer aided drug discovery, simulating protein-protein interactions, and so on2.

The Kellogg Lab in the VCU Department of Medicinal Chemistry, led by Dr. Glen Kellogg, my mentor, has done similar research. Their research is not limited to developing protein structure prediction software, but their work helps advance breakthroughs in computational molecular modeling, in addition to conventional x-ray crystallography and NMR.

Their primary paper that I found of interest, Ahmed et al. (2015)3, uses a different approach to the homology modeling. The foundation of their methods is program called HINT! (hydropathic interactions), developed by Drs. Glen Kellogg and Donald Abraham of the Medicinal Chemistry Department at Virginia Commonwealth University4. HINT is a computational tool that quantifies hydropathic interactions between molecules. “Hydropathy” is a collection of interactions that is comprised of four types of interactions: favorable and unfavorable hydrophobic interactions, and favorable and unfavorable polar interactions3. Ahmed et al. analyzed almost 30,000 tyrosine residues in proteins that have previously had their structures determined, then uses HINT to score the hydropathic interactions surrounding the tyrosines, this is illustrated in **Fig. 1**3where the tyrosine (highlighted in red) has it’s polar interactions highlighted by the HINT contours. The contours on the left surround the polar OH group, a favorable polar interaction could be hydrogen bonding and an unfavorable polar interaction could be base-base interactions. Favorable hydrophobic interactions could be π-stacking (when two rings stack), unfavorable hydrophobic interactions could be a hydrophobic-polar interaction. These are just a few examples of interactions from the four classes described above. The key hypothesis suggested motifs could be identified based on HINT metrics, and they found many using these methods as well and clustering methods described below.

**Figure 1:** An interaction map generated by HINT. The red and blue contours represent favorable and unfavorable polar interactions.

I.A. Summary of *HINT* Forcefield and Basis Mapping

Ahmed et al. randomly sampled 2703 from RCSB Protein Data Bank3. Within this data set, 28,889 tyrosines were identified. The tyrosines are plotted on a Ramachandran plot based on their phi(φ)-psi(ψ) bond angles, a visualization is provided in **Fig. 2**5. The peptide bond, or omega (ω) bond, is delocalized so it has a partial double bond, which restricts its rotation, so the phi-psi bonds can form unique angles. A Ramachandran plot is a graph that contains the phi angle as the x-axis, and the psi angle as the y-axis, and effectively predicts secondary structure based on those two bond angles of a given residue6. Ahmed et al. segmented the plot into an 8 by 8 matrix where each index is 45° by 45°. Each index was termed a “chess square” and labeled a1-h8 according to their axes. Additionally, phi-psi boundaries were shifted by -20° by -25°, respectively, to highlight densely populated regions Fig. 33.

**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.

HINT’s role in quantifying data in Ahmed et al. is two-fold: It scores structures and then those scores are used to calculate interaction maps. HINT scoring takes advantage of free energy information based on experimental data from solvent partitioning and converts it into a forcefield that distinguishes hydropathic interactions. First, a box with a volume of 8712 Å3, with CA at the origin. Then, all atom-atom interactions are scored by HINT. The score between two atoms, *i* and *j* is calculated by3:

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



Where *ai* represents the partial log P*o/w*, a partition coefficient. Solvent partitioning has the same fundamental processes and atom-atom interactions as biomolecular interactions or between proteins and ligands. Solvent partition constants (LogP for water/octanol) encode thermodynamic and interaction information, it’s usually done in a wet lab to determine the hydrophobicity of a solute. The goal of HINT was to reduce bulk molecular solvent partitioning information to discrete interactions between atoms. Partition coefficients are calculated within the program or obtained from a residue-based dictionary. LogP values are calculated based on two databases in HINT that contain atomistic parameters that allow it to calculate the LogP value for each atom7,8. *Si* representssolvent-accessible surface areas; these are regions of the molecule or in this case, atom, that are accessible to the solvent, this is calculated using the GETAREA program8–10. *Tij* is a descriptor function that equates to +1 for favorable interactions and -1 for unfavorable interactions. *r* is the distance between two atoms8. *Lij*is an adaption of the Levitt implementation11 of the Lennard-Jones potential function, a function used to calculate energy potentials. When *bij* > 0, this represents favorable interactions, and when *bij* < 0, this represents unfavorable interactions3.

After all atom-atom interactions have been scored, Ahmed et al. introduces a novel application called HINT Basis Maps. HINT Basis Maps calculate the 3D interaction environment associated with a residue. Within the box that contains the side chain, grid points are computed using the following equation3:



Where ρ*xyz* represents the 3D map value at a given point, *bij* is the HINT score between atoms *i* and *j*, *xij*, *yij*, and *zij* are coordinates of the midpoint of the vector between atoms *i* and *j*, and σ is a scaling factor that controls the width of the Gaussian map peak, which in this work was set to 0.5. All interactions are summed up and separate maps are calculated for each of the four interaction classes.

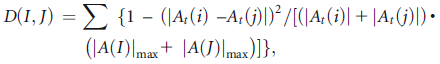
I.B. Map Similarity Metrics and Clustering Methods

Since a lot of space was vacant in a large portion of the maps, maps were scaled logarithmically on a point-by-point basis if a data point (*Gt*) was deemed high-value. *F* is a predefined value in this work at 0.5. Data point value was determined by |*Gt*|=*F* > 1.0; if so, then the scaled value is calculated by3:



Otherwise, *At* = 0

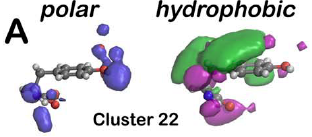
Then point by point, map-map similarity was calculated based on similarity between the two maps by a correlation coefficient-based metric given by:



Where *I* and *J* are two different maps. Summed over the set of map points *t*, where *A*(*i*)

and *A*(*j*) are corresponding point values for maps *I* and *J*, respectively. |*A*(*I*)|max is the maximum absolute value of map *I*3.

The k-means clustering algorithm was used as the clustering method. K-means is a statistical tool that forms clusters in data sets based on proximity. The data from the pairwise map correlation coefficient was used in this method3,12.

Ahmed et al. continues to describe the observations of a1 chess square, produced 14 unique clusters which are analyzed in the paper. These clusters, in a way, are comparable to secondary structure motifs due to their recurrence in these proteins. An example of a cluster from the a1 chess square is illustrated in **Fig. 4**3. In summary, the findings suggest that for the given phi-psi tyrosine angles in that chess square, there are 14 identified conformations that tyrosine can take on. This is significant because even though there may be near infinite possibilities of primary sequence, this suggests there is a *limited* number of possibilities a secondary structure may take on.

I.C. Phenylalanine v. Tyrosine

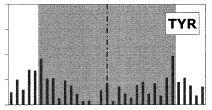
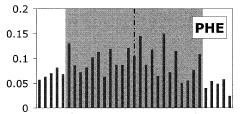
**Figure 4**. This is the first of the 14 clusters found in chess square a1.

This proposal seeks to acquire Phenylalanine Hint basis map data, then perform pairwise comparison with Tyrosine data, limited to the d4 chess square. Using Ahmed et. al.’s experiment with Phenylalanine would produce novel results. The d4 Ramachandran plot predicts a right handed alpha helix as the secondary structure with the given phi-psi angles6. Therefore, 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 too3,13. 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.

My key hypothesis questions whether among the limited set of clusters formed by tyrosine, are there any shared with phenylalanine? There are similarities between the two amino acids, which leads me to believe there may. Since nothing like this has been done before, my goal is to find something unique. And hopefully will be able to give a better understanding of secondary structure afterwards.

**Table 1** (top): graphically plots the 10 most populated chess squares for tyrosine. Adapted from Ahmed et al. (2015) **Table 2** (bottom): plots the 10 most populated chess squares for phenylalanine. Adapted from Kellogg unpublished observation.

1. **The Experiment**

The goal of this experiment is to use the techniques outlined in Ahmed et al., on a smaller scale, and with a novel purpose to compare two entities that have not been compared before. I chose phenylalanine to compare to tyrosine due to it’s similar structure. My initial fascination came from imagining the similar folding patterns phenylalanine could share with tyrosine simply because they share a phenyl group. Yet at the same time, I also question what differences will appear because of phenylalanine not having the polar OH that tyrosine has. In the d4 chess square, there are a total of 5376 tyrosines as well as 6364 phenylalanines **Tables 1**3and **2**13. This is significant because d4 contains the highest numbers of residue count per chess square. Additionally, they are similar in both data set size and characteristics. Chess 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. Another reason to pursue this study comes from the nature of the d4 chess square. The chess square forms into right handed helices. Ulmshneider et al. (2000) analyzed the amino acid distributions in transmembrane alpha helices of 29 proteins and found that the alpha helix composition had about 500 phenylalanines and 325 tyrosines14. And when they calculated average amino acid distributions among alpha helices in the lipid bilayer, tyrosines and phenylalanines had pronounced frequency peaks at both interfacial regions, the distributions are included in **Fig 5**14. And on the extracellular side, the peaks happen to appear at same frequency. Additionally, aromatic residues are believed to have preference to interfacial regions by anchoring the helix into the membrane by interactions between rings and lipid head groups15. To summarize, I believe this is ample proof to back up my thesis that there may be shared structures between phenylalanine and tyrosine in alpha helices.

**Figure 5a** (left): displays the distribution of phenylalanine in transmembrane alpha helices. The shaded region is the length of the bilayer with the central dotted line being the center of the bilayer. The left shaded boundary is the cytoplasmic interface and the right boundary is the extracellular interface. **Figure 5b** (right): displays the distribution of phenylalanine. Not much of this data overlaps, but the peaks on the right boundary show that these amino acids appear at the same frequency at the cytoplasmic interface

Since the data is already compiled for the tyrosine data from Ahmed et al., I would use the methods described in the introduction to HINT score and map for all phenylalanines in chess square d4. Rather determine correlation by plotting phenylalanine v phenylalanine, I will plot phenylalanine against tyrosine using the same map-map correlation coefficient metric as described above. Then I’ll use the k-means algorithm to cluster and determine if any homologous structures exist between tyrosine and phenylalanine.

1. Discussion

In the best-case scenario, there would be clusters that suggest some phenylalanines and tyrosines form the similar structures in right handed alpha helices. This is the best-case scenario because it would suggest that not only are folding patterns conserved for specific amino acids, but also, that some are conserved *among* amino acids. This can then give reason to investigate further and possibly look at another chess square to see if there are other similarities in formation between these two residues, or even open a window to compare to a different amino acid. In the big picture, this will narrow down the potential structures that have been identified, and can ultimately aid in developing a more effective homology modeling prediction algorithm. Alternatively, if I find no cluster correlation between phenylalanine and tyrosine, this could suggest that phenylalanine only has unique structures. Chess square d4 is an ideal data set to test this novel comparison since research (mentioned above) has shown that despite missing an OH group, phenylalanine has similar propensities to appear in the same positions as tyrosine due to environmental demands. I believe that if something were to be found between these two amino acids, it would be in this chess square. Moving forward, whether best or worst results are obtained, this data can be built and finish the rest of the Ramachandran plot of perform another clustering analysis of just phenylalanine to identify it’s own unique clusters.

References

1. Salzberg SL, Searls DB, Kasif S. Grand Challenges in Computational Biology. In: *Computational Methods in Molecular Biology*. Elsevier Science B.V.; 1998:3-10. 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.

2. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature*. 2016;537(7620):320-327. doi:10.1038/nature19946.

3. Ahmed MH, Koparde VN, Safo MK, Neel Scarsdale J, Kellogg GE. 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 Struct Funct Bioinforma*. 2015;83(6):1118-1136. doi:10.1002/prot.24813.

4. hint! (Hydropathic Interactions).

5. Peptide Bonds and Protein Backbones. http://www.biochem.ucl.ac.uk/~martin/c40/peptide.html.

6. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol*. 1963;7(1):95-99. doi:10.1016/S0022-2836(63)80023-6.

7. Hansch C. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. New York: Wiley; 1979.

8. Kellogg GE. hint! User’s Guide. http://www.edusoft-lc.com/hint/manuals/300/. Published 2008.

9. Fraczkiewicz R, Braun W, Braun W. Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules. *J Comput Chem*. 1998;19(3):319-333. https://s3.amazonaws.com/objects.readcube.com/articles/downloaded/wiley/81e1d84099e780096ffcff22ecc55887507b9a0482c37fa1e2713561d54fc070.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIS5LBPCM5JPOCDGQ%2F20171210%2Fus-east-1%2Fs3%2Faws4\_request&. Accessed December 9, 2017.

10. Negi S, Zhu H, Fraczkiewicz R, Braun W. Calculation of Solvent Accessible Surface Areas, Atomic Solvation Energies and Their Gradients for Macromolecules. http://curie.utmb.edu/getarea.html. Published 2015. Accessed December 9, 2017.

11. Levitt M. Molecular dynamics of native protein: I. Computer simulation of trajectories. *J Mol Biol*. 1983;168:595-620.

12. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Source J R Stat Soc Ser C (Applied Stat*. 1979;28(1):100-108. http://www.jstor.org/stable/2346830. Accessed December 10, 2017.

13. Kellogg GE. Unpublished Observation.

14. Ulmschneider MB, Sansom MSP. Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta*. 2001;1512(1):1-14. doi:10.1016/S0005-2736(01)00299-1.

15. Deisenhofer J, H M. The photosynthetic reaction center from the purple bacterium Rhodopseudomonas viridis. *Science (80- )*. 1989;245(1463-1473).