**Comparison of Exemplars of Rotamer Clusters Across the Proteinogenic Amino Acids**

Jesse Raynor

**Introduction**

Proteins are important to pharmaceutical research. They can participate in signalling pathways, thereby providing targets for drug activity, or they can function as pharmaceutical agents themselves. Problematically, our capacity to discover new proteins has far outstripped our ability to characterize their function due to technologies like high-throughput genomics. We can utilize protein similarity to putatively assign function to uncharacterized proteins, because two things which are similar physically seem likely to be similar functionally.

However, there exists a problem with current methods for determining similarity between two proteins in that they tend to rely on the identity of the amino acids which compose them. This can be reliable for long proteins possessing 40% or greater sequence identity, but the method is considered less reliable for proteins with 20%-35% sequence identity 6. There also exists the scenario where two proteins both possess a highly-conserved domain, enough such that they appear to be similar at first glance, but otherwise differ totally. In this scenario, it’s possible that they both function similarly, but it’s also quite possible that they differ totally in function.

Structural similarity is a better indicator of shared function, but it comes with problems of its own. Protein structure must be determined experimentally, typically with either nuclear magnetic resonance, which is limited to proteins below 20 kiloDaltons in weight, or X-ray crystallography, which is both difficult and time-consuming. It is possible to assign putative structure to a protein via a process called homology modelling, but the first step of this process is sequence alignment. Ergo, homology modelling fundamentally relies on sequence identity. In fact, it has been noted that if the sequences share less than 30% identity, errors tend to crop up during homology modelling. Additionally, they also require already structurally characterized proteins to match proteins of interest against. An additional method, threading, is more sophisticated than homology modelling in that it can detect structural similarity between proteins with low sequence identity by identifying which structural conformations a sequence is most compatible with. However, it too is reliant upon pre-existing structural information and leaves room for improvement 5.



Picture taken from wiseGEEK. Depicts structural diagrams of amino acids.

Proteins consist of chains of amino acids, each of which in turn consists of a main chain segment and an attached side chain. Singly-bonded atoms within the main chain and the side chains can rotate, producing a variety of conformations. These various conformations are called rotamers. One potential avenue for establishing protein similarity, investigated by Ahmed et al (2015), involves establishing similarity between proteins on the basis of the forces underlying the interaction of amino acids instead of the identity of those amino acids. An initial step in this research has involved characterizing these underlying forces in the form of rotamers and the local environments which produce them.



Picture taken from Lovell et al (2003). Depicts the psi and phi main chain dihedral angles important to the study of rotamers. The amino acid in question is a proline.

In 1962, Ramachandran et al published work on sterically forbidden and sterically allowed conformations of amino acids, conveyed as a “Ramachandran plot”. Lovell et al (2003) confirmed these results by plotting the main chain dihedral angles of 81,234 amino acids, excluding the special cases of glycine, proline, and preproline (those rotamers which precede a proline in a protein7), to form a “general case” Ramachandran plot. Ahmed et al (2015) found that tyrosine rotamers could be reliably placed into groups based on the dihedral angles of the main chain segment and the set of forces which cause the amino acid to adopt this conformation.



Picture taken from Ahmed et al (2015). Depicts the modified general case Ramachandran plot used in the paper.

To do this, they divided the general case Ramachandran plot into 64 squares and assigned tyrosines from an X-ray crystallography dataset to those squares based on the main chain segment’s conformation. Then, each tyrosine and its parent protein was aligned with the calculated tyrosine represented by the center of its respective square. From this point, the various forces which affect the tyrosine rotamer (such as hydrogen bonding) could be quantified in one of four 3D grids, each representing a class of forces in the tyrosines’ environments (favorable hydrophobic, unfavorable hydrophobic, favorable polar, and unfavorable polar). By comparing the four members of the map quartet associated with one tyrosine with the map quartet associated with another tyrosine, correlation scores were produced for each map pair and then combined, producing a single value denoting the overall similarity between two tyrosine rotamers. By sorting the tyrosine rotamers into groups based on these values, clusters of rotamers were found.

These rotamer clusters were formed for a single type of amino acid, tyrosine. However, it has not been examined whether rotamers of different proteinogenic amino acids will cluster together. Examining this aspect would be useful in informing further research along this line.

**Methods**

 A list of exemplars, those rotamers and their force environments, as represented by the rotamers’ map quartets, most similar to the averaged map quartet of a given rotamer cluster, will be compiled from previous results for individual amino acids. As noted in Ahmed et al (2015), average maps can lend themselves to a phenomenon called “brown mapping”, whereby they become so averaged that they begin to resemble very disparate rotamers. This can lead to misleading results, and so it is preferable to use exemplars instead of average maps.

 After doing so, the 3D maps which represent the forces around the rotamers will need to be expanded. Ahmed et al (2015) specifies that the 3D maps were sized to encompass all possible conformations of tyrosine, plus some extra buffer space. This reduces the amount of empty space in each map, thereby reducing the amount of computation needed to compare any two tyrosines. However, following this procedure for other amino acids complicates the task of comparing them. Thus, it is desirable to resize the exemplar maps by adding additional space, such that all exemplar maps are of a uniform size. Additionally, this will be done such that the exemplar rotamers and their force environments are maintained at the center of their respective maps, unmodified by any sort of scaling, rotation, or translation.

 Because the force environments for each exemplar will be already known, it will not be necessary to recalculate them. Also, because there are far fewer exemplars than the number of amino acid rotamers originally used to find them, it will not be necessary to split the Ramachandran plot into squares and assign exemplars to them.



Accordingly, the next step will be finding the degree of correlation between each of the exemplar map quartets. This will be done by first transforming the data with the piecewise function L(Gt), pictured above. In this function, Gt is the value of the real-valued data point under consideration, F is some constant real value, and At is the output value. We will follow Ahmed et al’s lead and use 0.5 for F. Employing this function on every data point in the maps was done in Ahmed et al (2015) to prevent small, highly-correlating points in the maps from skewing the results.

After transforming the data with the L(G), D(I,J) will next be employed to find the correlation score for each pair of maps between two exemplar map quartets, and the four resulting correlation scores will be combined via the function C(m,n) to find the overall similarity between the two maps 1.

To employ the correlation function D(I,J), one additional consideration needs to be made. Because the two maps under consideration, I and J, invariably contain many zero-valued points due to the effects of L(G), it is necessary to construct a Boolean mask map to prevent the maps from correlating along the zero-valued points. This mask map will be a three-dimensional array of Boolean values of equal size to the maps I and J. The values in the mask map will be set to TRUE only when At(I) >= ( 8 \* Astddev ) or At(J) >= ( 8 \* Astddev ).and will be set to FALSE otherwise. In these expressions, At(I) and At(J) are log-transformed points in the maps I and J, and Astddev is the standard deviation of the At values in the respective maps 1.



In D(I,J), At(i) and At(j) are corresponding points in maps I and J respectively, and |A(I)|max and |A(J)|max are the maximum absolute values in maps I and J respectively. As indicated by the formula, the expression performed on every corresponding point i and j in the maps I and J, and the results are summed to produce a single value denoting the similarity of the map pair 1. D(I,J) should only be calculated for points in I and J for which the corresponding point in the Boolean mask map is TRUE.



After finding each map-map correlation score for each map pair in the exemplar map quartets under consideration, they should be combined via a weighted average function, C(m,n) (where m and n are two exemplar map quartets), to produced a single similarity score denoting the overall similarity between two exemplars 1.

As C(m,n) is being calculated for every pair of exemplar quartets, the similarity scores will be stored in a matrix in which the rows represent individual exemplars and the columns represent the exemplars to which they’ve been compared. Once this is done, we will proceed to cluster the values resulting from the correlation step. In accordance with Ahmed et al (2015), we will use k-means clustering, available in R, over 1000 trials, each trial with a different set of starting points. Because k-means clustering requires a user-defined number of clusters, we will also follow Ahmed et al (2015) in this respect and employ the gap statistic, also available in R, to determine a suitable value of k 1.

Last, we should validate the results of the clustering procedure. This can be done using a variation of the Peeples’ R protocol, detailed in Ahmed et al (2015). The first step of this procedure is to generate 250 new matrices from the similarity matrix by randomly rearranging similarity scores in their respective rows. This has the effect of preserving the mean and standard deviation associated with each exemplar while scrambling the underlying relationships between exemplars. After doing so, we can apply our k-means clustering with the gap statistic method. With the output of the k-means clustering in hand, we then find the sum of squares error for both the randomized data and the actual data and plot each against the number of clusters. If successful, the plot of sum of squares error vs k should tend to fall more slowly than it should for the actual data, in which the error should fall relatively rapidly with an increasing value of k 1.

**Discussion**

 In Ahmed et al (2015), the the composite similarity scores of tyrosine clusters in square A1 averaged 0.95763 with a standard deviation of 0.01490. Meanwhile, the intercluster similarity scores (or “dissimilarity” scores) between those clusters averaged 0.8486 with a standard deviation of 0.0421. It is hoped that an experiment in clustering rotamers of disparate identity would produce clusters with similarity scores matching the former and intercluster similarity scores matching the latter. Since the data isn’t segmented by the Ramachandran plot squares, as it is in Ahmed et al (2015), there also exists the possibility that this method might find that multiple clusters of the same type actually occupy the same space and are similar enough that they can be combined.

 In the event that the exemplars do not cluster well, alternative explanations need to be considered, particularly with regard to the implications for the relationship between individual amino acids and the protein structures they form. In particular, it might be worth considering whether rotamers are highly dependent upon their neighbors and, if so, what sets of neighboring amino acids tend to be associated with any given rotamer cluster. This might entail constructing protein contact maps for each of the proteins in the original X-ray crystallography dataset, identifying to which rotamer cluster each amino acid in the dataset belongs, and employing association rules-based machine learning to discover which rotamers and, by extension, which amino acids each rotamer tends to be associated with.

 The approach described in the Methods section does have a few apparent shortcomings. The most problematic is that there are amino acids which have not yet been studied with the method described in Ahmed et al (2015). Consequently, there’s a not insignificant amount of legwork left to be done with regards to finding rotamer clusters among those amino acids. The other biggest problem is the lack of X-ray crystallography and NMR data for membrane proteins, which comprise ~25% of total proteins, but as of 2008, have less than 150 structures determined 4. The environment in which membrane proteins exist has been found to affect the distribution of rotamers throughout membrane proteins in a way that differs from soluble proteins 8. However, without sufficient data on membrane proteins, it’s hard to account for these differences with this method without resorting to artificial methods.

 One of the major challenges of biology is the task of discovering how to determine the tertiary, three-dimensional structure of a protein from its primary structure, the amino acid sequence. An approach which looks “beneath the hood” seems to be needed for this, and this experiment will help direct research in that vein.

**Bibliography**

1. Ahmed, M. et al (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 83: 1118-1136.

2. Ramachandran, G.N., Ramakrishnan, C. & Sasisekharan, V. (1962). Stereochemistry of

Polypeptide Chain Configurations. J. Mol. Biol. 7: 95-99.

3. Lovell, S. et al (2003). Structure Validation by C-alpha Geometry: Phi, Psi, and C-beta

Deviation. Proteins 50: 437-450.

4. Carpenter, E.P., Beis, K., Cameron, A.D. & Iwata. S. (2008). Overcoming the challenges of

membrane protein crystallography. Current Opinion in Structural Biology 18: 581-586.

5. Forster, M.J. (2002). Molecular modelling in structural biology. Micron 33: 365-384.

6. Chen, J., Guo, M., Wang, X. & Liu, B. (2016). A comprehensive review and comparison of

different computational methods for protein remote homology detection. Briefings in

Bioinformatics.

7. Ho, B.K. & Brasseur, R. (2005). The Ramachandran plots of glycine and pre-proline. BMC

Structural Biology 5:14.

8. Chamberlain, A.K. & Bowie J.U. (2004). Analysis of Side-Chain Rotamers in Transmembrane

Proteins. Biophysical Journal 87: 3460-3469.