Analysis of Clustering Fragmented Protein Bond Angles

Justin S. Diamond*
Boston University Department of Bioinformatics, Boston, MA, USA.

* Corresponding author. Email: J3tkd@bu.edu
Manuscript submitted June 10, 2019; accepted January 13, 2020.
doi: 10.17706/ijbbb.2020.10.2.74-83

Abstract: The desire for accurate protein prediction algorithms has been a hallmark of computational biology achievements. Still, better algorithms and methodologies can achieve even greater success with implication across a diverse range of biological and medicinal fields such as protein function inference. Accurate prediction methods rely heavily on sequence similarity, however structure is more evolutionary conserved, i.e. structure is an alternate characteristic for ancestral relationships between proteins. The premise of this work is that similar structural features will be clustered together, which may show a unique amino acid and secondary structure (SS) distribution, which can be, incorporated into HMMs for SS prediction and protein function inference algorithms. With structural-evolutionary relationship in mind, I propose a methodology for 'structure' based SS prediction methods using HMM and k-mean and fuzzy k -means fragmented protein clusters. When fragment distributions were incorporated into HMMs, the average accuracy increased by 1 percent but showed an increase in accuracy of up to 13 percent for particular sequences. The HMM results were not so promising, however the clustering of protein structure fragments by C-alphas bond angles shows to be a useful length-independent metric for inferring functional relationships between proteins.

Key words: Protein structure, secondary structure, protein function, k-means, UPGMA.

1. Introduction

Proteins are biological macromolecules composed of a sequential set of 20 natural amino acids that have three dimensional spatial coordinates, corresponding to their lowest energy (native) state, that are, for practical purposes, completely determined by the linear order of amino acids. Further, proteins perform biological 'functions' that are dependent upon their geometry within a biological network. Figure 1 portrays an arbitrary protein from THE-DB [1], a protein threading database.

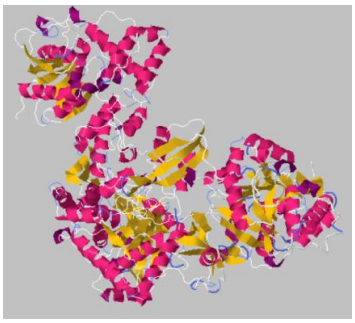


Fig. 1. A protein from THE-DB with helices in red, beta-sheets in yellow, and turns in white.

The epitome of any biological network is the degree of interaction between any numbers of proteins within that network. Interactions are characterized by binding affinities that describe how likely, and how long, two proteins are to interact with each other. The biological function of a protein within a particular pathway, such as Vitamin-D3 in Uterine Fibroid development [2], can be inferred from the degree of interaction between itself and other proteins within the network, or from structural similarity. The significance of this is multi-fold; it has implications in drug design to battle diseased states, or characterizing the underlying genetic cause of a phenotype. Protein interaction algorithms are called protein 'docking', but this class of algorithms needs determined protein structures. Now more than ever, new sequencing techniques are perpetuating the gap between the relatively large number of protein sequences to the number of determined three dimensional protein structures. It becomes tedious to experimentally determine large amounts of protein configurations with x-ray crystallography. To help fill this gap, protein prediction algorithms have been developed in which a sequence is given as input and a three-dimensional protein configuration is the output. Three classes of protein prediction algorithms exist which include ab initio, comparative, and threading modeling.

In each of these algorithms, a significant constraint that helps guide the protein conformational search space is predicted secondary structure, which is characterized by local folds within a protein. The addition of secondary structure to model tertiary structure greatly improves performance. Thus, a key aspect of correct protein prediction algorithms is correct secondary structure (SS) prediction algorithms. Current state-of-the-art SS algorithms involve either a three state model (beta-sheet, alpha-helix, turn) or eight state model (3_{10} helix , α helix , π helix , hydrogen bonded turn, extended strand in parallel and/or anti-parallel β-sheet conformation , isolated β-bridge, bend, and coil) .

Continuing, given a experimentally determined, or predicted, protein structure it is often desirable to find structurally related proteins as they are likely to pertain to similar functions such as enzyme catalysis or toxicity. Evolutionary insight can then be gathered by analyzing sets of proteins relating to a biological pathway, and other co-evolving networks, leading to better understandings of larger phenotypic phenomena.

2. Related Work

2.1. Secondary Structure Prediction

Two state-of-the-art methods for obtaining accurate secondary structure predictions are using sequence profiles from psi-blast [3], [4], which tends to have better results and neural networks like DeepCNF [5]. Sequence profiles are created using psi-blast which is an iterative based blast. For instance, given an amino acid sequence, blast is ran which gives a series of sequences as an output. This series of sequences are aligned and positions are given scores based on how conserved the amino acids are in each position. The most conserved positions are the main 'profile' which is then ran through blast again. This procedure is iterated a desired number of times. This is helpful in secondary structure prediction because one can identify evolutionarily similar sequences that may share secondary structure characteristics. As well, neural networks have gained a lot of traction in a vast number of fields, including secondary structure prediction. Its usefulness is derived from the ability to recognize patterns that are difficult for humans. In DeepCNF, a window size of length 11 is used to predict the 6th position secondary structure labeling and they used 1 to 7 hidden layers. For more detailed information about the architect of DeepCNF, I suggest the reader to their paper.

Continuing literature research on the topic of Secondary Structure prediction led me to discover papers from Cheng and Sen [6] that use Fragment Database Mining (FDM) which is similar to my methodology of fragment based SS prediction but does not involve clustering or bond angles. FDM essentially breaks a

sequence into fragments and compares these sequence fragments to sequences from Protein Data Bank (PDB) to find fragment structures that are used to help infer secondary structure prediction. This methodology has accuracy ranging from 67.5 to 93.5 percent depending on the sequence similarity.

Hidden Markov Models have been used in Secondary Structure prediction previously with an initial introduction into the field during the mid 1990s. Sophie Zaloumis's paper [7] compared different HMMs [8] and showed that a normal 3-state HMM attains an accuracy of 49 percent.

2.2. Protein Clustering

The intuition comes from the evolutionary perspective that sequence-wise diverse proteins may have parts of the proteins that are structurally similar to parts (fragments) of other proteins, but this structural similarity is in correspondence with a low sequence similarity. Thus, using structural features it is possible to cluster sets of proteins to infer evolutionary or functional similarities.

This also raises the idea that evolutionarily distinct proteins could coevolve to similar conformations without sequence similarity. Bin et al. showed the ability the cluster proteins by virtual bond angles [9]. In this paper, he clusters whole proteins by the sequential C-alpha bond torsional angles between position i to i+3. However, this creates uneven dimensional vectors for proteins of differing lengths. It also fails to capture similar local folds between proteins at different indexes within proteins. By clustering fragments of proteins, I am able to compare local regions within proteins that may be similar to other local regions, it also allows for proteins to be classified by the cluster centroids of the fragments that make up a protein.

3. Methods

3.1. Secondary Structure Prediction

The challenge of secondary structure prediction is the following: an amino acid sequence must be labeled corresponding to a three or eight state system at each position, with the eight state system being an extension to the three state system, but with more strict categorizations. For simplicity, only the three state system will be considered, but this will not affect the results as the three and eight state system can be modeled in a similar fashion. To label each position I will be using a nth order Hidden Markov Model where each position is dependent upon only the n previous positions where n could be any positive integer.

3.1.1. Collecting proteins from RCSB and SS labels

Over 10,000 non-redundant protein structures from the PDB library have been collected [10].

An algorithm called STRIDE [11], which is freely available for download, considers dihedral angles between the Carbon-alpha, Nitrogen, Carbon, and Carbon-alpha of the next amino acid and hydrogen bonding patterns to characterize each amino acid into one of the eight state systems. As the eight state system is only an extension of the three state system, regrouping the labels to correspond only to helices, beta-sheets, and turns is possible.

3.1.2. Fragmenting proteins

To obtain protein fragments, I selected an arbitrary protein and an arbitrary length to cut each protein at. For instance, length of 10, corresponds a 100 amino acid protein to be cut into 10 fragments of 10 amino acids.

3.1.3. Amino acid bond angle feature vector

Angles were obtained in two ways. The first, Fig. 2A will be referred to as the ambiguous method, which is calculated between two vectors, which is characterized by

$$v_i = p_{i+1} - p_1 \tag{1}$$

where *p* is a three dimensional coordinate corresponding to the Carbon-alpha atom of each amino acid. For instance, the first angle is taken between vi and vi+1 where the equation for angle is

$$\arccos\left(\frac{v_i \times v_{i+1}}{|v_i||v_{i+1}|}\right) \tag{2}$$

where · is the scalar product, and $||$ refers to the magnitude of the vector. The ambiguous angle can calculated between v_i and v_{i+1} only, which will create one angle corresponding to each amino acid, or more angles can be calculated between v_i and v_{i+n} where n was chosen to be between 0 and 4. This is called 'ambiguous' as it loses some information about the exact orientation from one amino acid to the next. I consider this an important angle to interpret because it has the potential to cluster fragments by similarity even if they have different secondary structures. This could find similar conformations between fragments irrespective of secondary structure, which will be interesting because fragments could have similar overall topology, but different secondary structure. The ambiguous angle has stronger correlations with protein characterization compared to secondary structure prediction. The second type of angle is the torsional angle between the carbon-alpha atoms of the ith, $i+1$, $i+2$, and $i+3$ amino acids, Figure 3B. This angle is distinct from the previous, as it is not ambiguous about direction. This angle is calculated by the normal vector of the planes created by points $\{p_i, p_{i+1}, p_{i+2}\}$ and $\{p_{i+1}, p_{i+2}, p_{i+3}\}$.

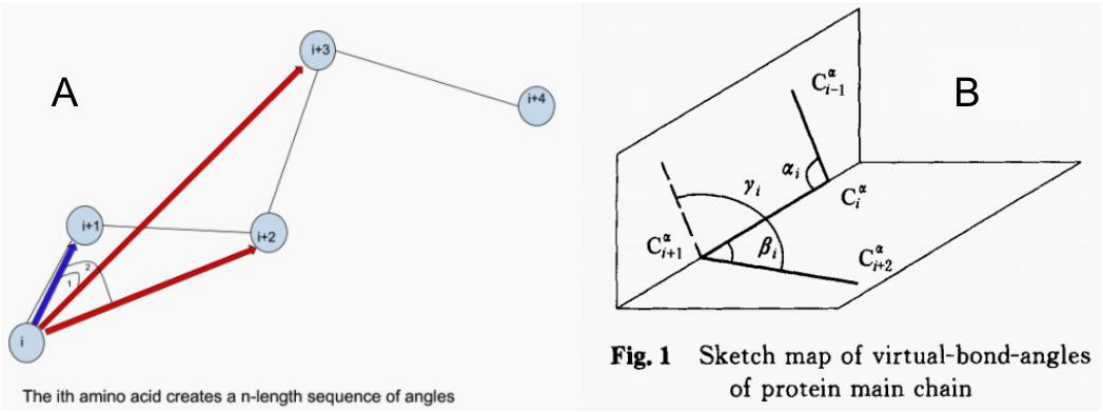


Fig. 2. Diagrams describing the ambiguous angle (3A) and the non-ambiguous angle (3B).

3.1.4. K-means and fuzzy K-means clustering

K-means is an iterative algorithm assigning labels y_i to data points x_i according to

$$y_i = \operatorname{argmin}_c(\|x_i - u_c\|^2) \tag{3}$$

where u_c is the centroid corresponding to cluster c, N is the number of data points, and C is the number of clusters. u_c can then be updated according to

$$u_c = \frac{1}{N_c}\sum_i x_i^c \tag{4}$$

with i corresponding to $x_i \in C$. This procedure is iterated until a limiting criterion, such as convergence, is met.

Fuzzy K-means is a probabilistic extension of K-means, without the theoretical guarantees of Gaussian Mixtures, which seeks to minimize

$$J_m = \sum_{i=1}^{N}\sum_{c=1}^{C} \pi_{ic} \left\| x_i - u_c \right\|^2 \quad , \; 1 \le m < inf \tag{5}.$$

Unlike K-means, Fuzzy K-means assigns a degree of membership of x_i with each cluster u_c by

$$\pi_{ic} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - u_c \right\|}{\left\| x_i - u_k \right\|} \right)^{\frac{2}{m-1}}} \tag{6},$$

with the cluster centroid being updated as

$$u_c = \frac{\sum_{i=1}^{N} \pi_{ic} x_i}{\sum_{i=1}^{N} \pi_{ic}} \tag{7}.$$

3.1.5. Cluster hidden Markov model for SS prediction

With STRIDE, we can label a protein at each position by its secondary structure label. As well, we can count the number of times each amino acid was emitted conditioned on the SS label, e_k, and the number of times each state transitions into each other, a_{jk}. The markov property can be extended to arbitrary conditioning on the past n states. Thus all of parameters of the HMM can be counted from the data, and the Expectation Maximization algorithm is not needed. The Viterbi algorithm will be modified into the form

$$V_j(i+1) = e_k \arg_k \max \left(V_k(i) \times a_{jk} \times Cscore_k(x_{i:i+15}) \right) \tag{8},$$

where $Cscore_k(x_{i:i+15})$ has been added to represent the probability of state k given the next 15 amino acids. To do so the frequencies of each amino acid, F_i, in $x_{i:i+15}$ is compared to the frequencies of amino acids, C_{ij}, from each cluster.

$$Cscore_k(x_{i:i+15}) = \sum_{i=1}^{i+15} abs(C_{ij} - F_i) \tag{9}.$$

Essentially, the Viterbi algorithm is weighted by the probability of the secondary structure state calculated from the distribution of amino acids and secondary structure in each cluster.

3.2. UPGMA for Protein Composition Similarity

Unweighted Pair Group Method with Arithmetic Mean will be used to quantify similarities between protein structures by calculating the distance between two fragments as the distance between fragment's cluster centroid. This will be averaged between every pair of fragments between two proteins. Important to note is that this is not a distance metric, as the distance between a protein and itself is almost never 0. This would only be true if a protein was made up of only one type of fragment type. The distance between cluster centroids is squared Euclidean distance,

$$d_{jk} = \sqrt{\sum_i (u_{ij} - u_{ik})^2} \quad (10).$$

where u_{ij} is a centroid and i is each dimension of the centroid, and j and k is any cluster. The UPGMA metric is an averaged sum over every fragment pair between two proteins,

$$D_{pr} = \frac{1}{|P||R|} \sum_p \sum_r d_{pr} \quad (11).$$

where *P* and *R* are collections of fragments that each correspond to a proteins, d_{jk} is the squared Euclidean distance metric, and | | is the number of fragments in each protein. Then a phylogenetic tree was created based on the following principles. Assign each protein into its own cluster. Identify two clusters where D_{pr} is minimum. Unify these clusters together, with the node connecting them at D_{pr} / 2. Delete the individual clusters, and add the unified cluster to the set of clusters. This procedure is iteratively done until there is only one cluster.

4. Results / Discussion

4.1. Clustered-Based HMM SS Prediction

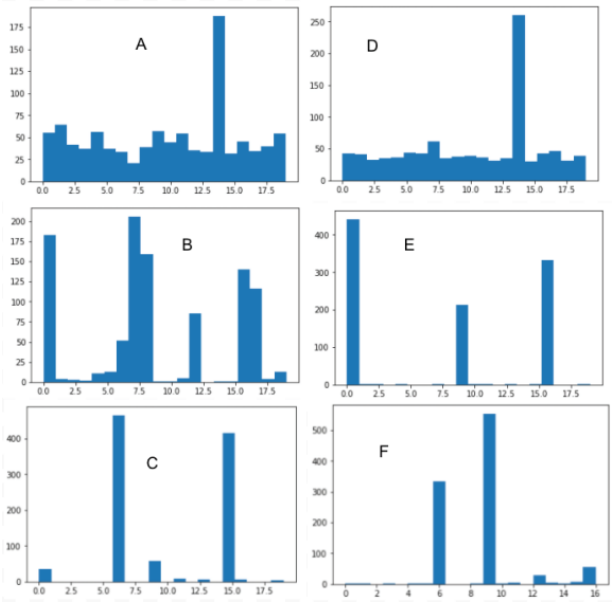


Fig. 3. A - kmeans, B - Fuzzy with m =2, C- Fuzzy with m = 4, all with ambiguous angles. D - kmeans, E - Fuzzy with m =2, F- Fuzzy with m = 4, all with torsional angles.

All nth order HMM, with n ranging from 0 to 9, were evaluated for accuracy of SS prediction. The minimum accuracy for the non-clustered HMM was 49.26 for the zeroth order markov chain, and the maximum was 49.36 for the fourth order. For the following results, unless otherwise stated, the order of the HMM is four as it was the max in the initial test. My basic HMM had an average accuracy of 49.36 percent among 200 proteins in the test set, which is comparable to the literature I found [7]. When implementing CHMM using ambiguous angles the average accuracy actually decreased to 47.1 percent using clusters generated from K-means, and 48.3 percent for clusters generated from fuzzy K-means with m = 2. However, the fragment-based model reaches accuracy of up to 73 percent, which shows its potential (with a low of

23). The average accuracy using the fragment clusters generated from fuzzy k-means with m = 4 was 49.6 which shows a .3 increase from the normal model. Using torsional angles, the K-means achieved an accuracy of 50.5 percent, the fuzzy K-means with m = 2 achieved an accuracy of 48.3 with a high of 78 percent, and 47.8 with m =4. The only case in which the accuracy increased beyond one percent was with K-means clustering with torsional angles.

To cluster protein fragments by a length of 10, the average length of secondary structure label durations, was chosen. There were distinct differences between K-means and fuzzy K-means. The measure of fuzziness is characterized by the variable m, where when m=1, the two models should be equivalent. As m increased, there tended towards a few larger clusters and more smaller clusters. K-means, and fuzzy K-means with m=2, had a more uniform distribution of amino acids and SS labels, One possible reasoning for fuzzy K-means to lose discrepancy between the fragments with higher m is that the angles remained, for the most part, within 40-70 degrees using the ambiguous angles, which means that there may not be completely distinct differences between fragment clusters.

4.2. Protein Composition Similarity

Characterizing proteins based on the fragments that make up a particular protein can be compared to Li [9], that showed classifying proteins by C-alpha amino acid torsional angles is feasible. Yet, he did not show how well the classification method is justified. He clustered proteins by a sequence of angles that corresponds to the whole structure. This method assumes that similar proteins must have similar angle sequences throughout the whole sequence. Whole protein clustering does not catch similarities between proteins that have similar local fragments permuted in different ways intra-proteins. Fragment based clustering should be able to identify similar fragments within proteins by clustering them together.

Table 1. Three Clusters of 6-8 Proteins Taken Randomly from UPGMA Hierarchical Clustering

Name	Organism	Function
1a04A2	E. coli	Signal transduction, nitrate response
1a2xA1	E. coli	Skeletal Muscle function
1aabA	E. coli	DNA binding to Rat HMGA
1a5zA2	E. coli	Oxidoreductase
1aa7A2	Influenza Virus	RNA-matrix binding
1a41A2	Virus	Topoisomerase catalytic fragment
1a0pA1	E. coli	Site specific recombinase

Name	Organism	Function
1a6eA	Thermoplasm acidophilum	chaperone
1a4zA	Bus Taurus	Oxidoreductase
1ac5A	Saccharomyces Cerevisiae	Carboxypeptidase
1a4sA	Gadus Morhua	Oxidoreductase
1a3wA	Saccharomyces Cerevisiae	Pyruvate Kinase
1a6eB (second domain)	Thermoplasm acidophilum	chaperone

Name	Organism	Function
1a6bA	Virus	Leukemia virus
1a1tA	Virus	HIV-1 nucleocapsid bound to RNA element
1aafA	Virus	Nucleocapsid protein from HIV-1
1abtA	Bungarus multicinctus	toxin
1ae2A	E. Coli	DNA binding
1a5jA	E. Coli	DNA binding

The three protein clusters chosen, Table 1, do show some obvious similarities. Such as the leftmost table identifying mostly E. coli proteins. Here 4 of the 7 proteins are DNA/RNA related. The middle table shows somewhat more variability but was able to cluster together two oxidoreductases, and it also clusters together the two subunits of 1a6e. The rightmost table shows the most similarity having three viral proteins clustered together, and two DNA binding proteins clustered together. In comparison to the fuzzy k-means counterpart, the overall topology of diagrams remained similar. For instance, in both diagrams 1a29A2 and 1a79A1 are paired close together, whereas 1a79A2 is placed on the opposite side of the diagrams.

This also brings an important question to light because the two domains for 1a79 were separated whereas one may expect domains of the same protein to be clustered together. Intuitively, this could be justified by saying that distinct domains within a protein can be functionally different, as they must function together to achieve the protein's main function. For instance, one could imagine that an important part of a protein is to maintain stability while it binds to something. This could require two domains, one, which

functions to keep the stability of the protein, and the other, which has the function of binding.

Overall, the results suggest that this methodology could be used to infer some functional similarity between proteins, but it can also be used to infer structural similarities in cases where traditional methods like RMSD are not as effective either because two proteins are of varying sizes, or perhaps the overall topology of a two proteins are distinct, but the fragments which make them up are similar. The biggest downfall of this methodology is that it contains subjectivity on exactly how to validate the similarity of two proteins functionally and structurally. This method is sequence independent, so it believe it could be helpful in aiding sequence dependent methodologies in recognizing similar proteins. As a baseline, Blosum62 [12] was used to cluster the same set of protein sequences. Many protein pairs, in close proximity, were maintained showing this method can detect close homology. However, my method should be able to find more distant evolutionary connectedness. (Appendix)

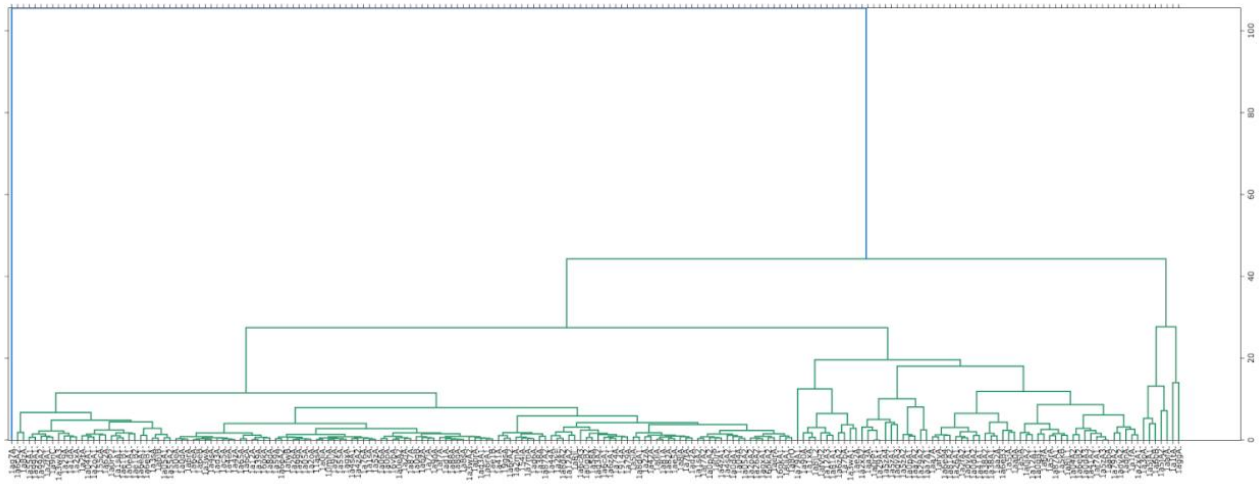
5. Conclusion

Clustering fragmented proteins by amino acid angles shows a useful measure in protein informatics as it can be used jointly with sequence dependent methods. Future work could be implementing these clustering techniques with psi-blast to identify sequences for SS prediction to slightly improve quality. As well, functional inference from compositional similarity could be used to help classify newly discovered protein structures, or find evolutionarily related proteins. Further, these angle vectors could be used as protein representations for deep learning methods.

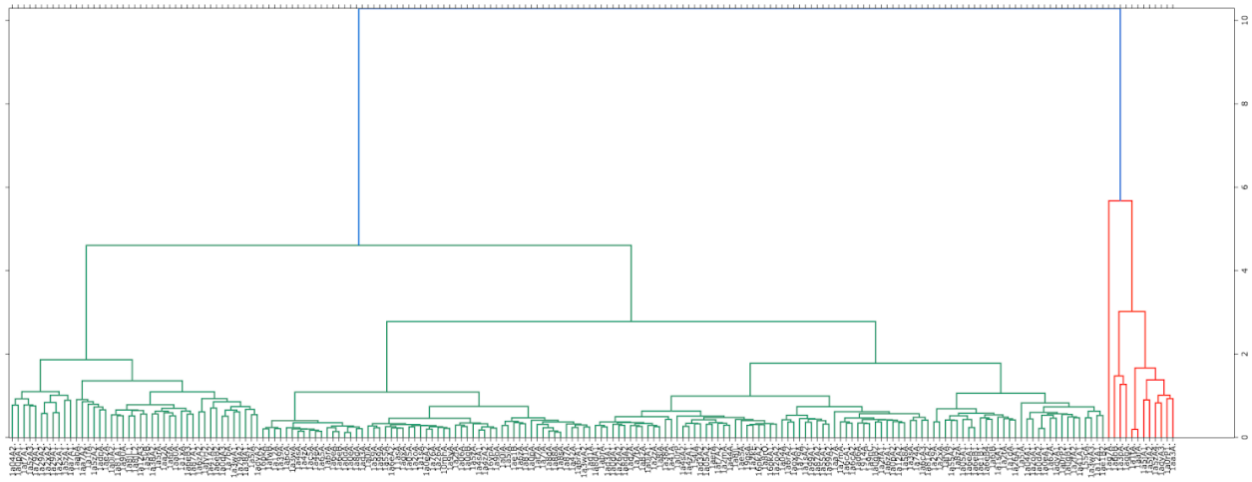
Fragment based clustering did prove to have some slight extra predictive power, however it was not significant in the present methodology. In addition, other features of proteins could be incorporated that may have more predictive powers such as distances between amino acids, side chain orientation, surface area, and amino acid volumes. To extend the protein classification methodology we would need to find a quantitative measure that can accurately describe the conformational similarity between two proteins. However, this remains a difficult problem because comparing the geometry of proteins fragments becomes an abstract challenge. Functional similarity is somewhat easier to conceptualize because either two proteins have similar functions (i.e. kinase, DNA binding) or they do not. Trying to judge the conformational similarity is not so distinct. Future work could include finding a quantitative measurement of conformational similarity between protein structures, and how conformational similarity relates to functional or evolutionary relationships.

Appendix

My Method:



Blosum62 Baseline:



Conflict of Interest

There are no conflicts to report.

Author Contribution

All work performed by Justin Diamond.

References

[1] Justin, S. D., & Yang, Z. (2018). THE-DB: A threading model database for comparative protein structure analysis of the *E. coli* K12 and human proteomes. *Database*, 2018.

[2] Brakta, S., Diamond, J. S., Al-Hendy, A., Diamond, M. P., & Halder, S. K. (2015). The role of vitamin D in uterine fibroid biology. *Fertility and Sterility*, *104(3)*, 698–706.

[3] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol. 292*, 195-202.

[4] McGuffin, L. J., Bryson, K, & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics, 16(4)*, 404-5.

[5] Wang, S., Peng, J., Ma, J., & Xu, J. (2006). Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports 6, 18962*.

[6] Sen, T. Z. A Consensus data mining secondary structure prediction by combining GOR V and fragment database mining. *Protein Sci., 15(11), 2499-506.*

[7] Zaloumis, S. (2005). The application of hidden Markov models to protein secondary structure prediction.

[8] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77(2),* 257–286.

[9] Li, B., He, & H., Li, Y., *et al*. (2005). *J Cent. South Univ. Technol*., *12*, 465.

[10] Berman, H. M., Westbrook, & J., Feng, Z., *et al*. (2000). The protein data bank. *Nucleic Acids Res*., *28(1)*, 235–242.

[11] Heinig, M., & Frishman, D. (2004). STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucl. Acids Res., 32,* W500-2.

[12] Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*., *89(22)*, 10915–10919.

Justin Diamond studies atomic systems at the intersection of physics, biology, and machine learning. As a graduate of Michigan State University's bachelors program of human biology and Boston University's masters program of bioinformatics, accompanied by machine learning research experience at the University of Michigan and Toyota Technological Institute of Chicago, Justin joined Dr. Alexander Tkatchenko's Theoretical Chemical Physics laboratory at the University of Luxembourg, as a PhD student, to research fundamental questions related to biologically significant systems.