Computational Biophysics Term Project (CS61060) Group: D

Secondary Structure Assignment of Proteins based on Ramachandran Plot

Prepared By: *Shaik Mohammed Sameer*(15EC10054), *Atul*(15EC10067), *Avinab Saha*(15EC10071)

Introduction

Protein structure is the three dimensional arrangement of the atoms in its amino acid sequence. Protein structure has repetitive local segments like a-helix, b-sheet etc which are further arranged three dimensionally. The classification of the structure into these repetitive segments is known as the secondary structure of the protein. The secondary structures are mainly influenced by the hydrogen bonds present in the biopolymer, as observed in an atomic-resolution structure under experimental conditions. Different algorithms to predict the secondary structure have been proposed. Its prediction from amino-acid sequence can provide information to build the initial model for molecular simulation.

Theoretically, if a protein sequence is given, then its secondary structures can be determined by various simulation methods. However, a major drawback of this approach is strong bias by the similarity of the protein sequence content. If a similarity between the target sequence and the template sequence is detected, structural similarity can be assumed but cannot be confirmed. So, more information is needed to predict the structure. several methods have been proposed by various research groups. The Directory of Secondary Structure of Proteins (DSSP) by Kabsch and Sander makes the secondary structure assignment solely on the basis of backbone-backbone hydrogen bonds. The Ramachandran method assigns the secondary structure based on the phi and psi values. The secondary STRuctural IDEntification method (STRIDE) by Fisherman and Argos, uses an empirically derived hydrogen bond energy and phi, psi torsion angle criteria to assign secondary structure. Each method takes different approach in predicting the secondary structure, but every method lacks in one way or the other, giving a nominal accuracy.

A different method to construct a protein secondary structure without any additional information of the new protein except its amino acid sequence is presented in this work. The ramachandran method globally assigns the same structure based on phi and psi angles irrespective of the neighbouring amino acids, but practically, the neighbouring amino acids play a significant role. A modified and reverse approach is to train a neighbourhood dependent ramachandran assignment of structures. We denominate this trained matrix as **Ramachandran Adjacency Index (RAI)** for protein secondary structure prediction. Using this, we have a set of virtual phi and psi values without actually finding them, which can then be used to predict the secondary structure based on the ramachandran plot. This is further discussed in detail in the upcoming sections.

Protein Secondary Structures and Ramachandran Plot

As defined earlier, protein secondary structure is the three-dimensional form of local segments of proteins. The two most common secondary structural elements are **alpha helices** and **beta sheets**, though **beta turns** and **omega loops** occur as well. Secondary structure elements typically spontaneously form as an intermediate before the protein folds into its three-dimensional tertiary structure.

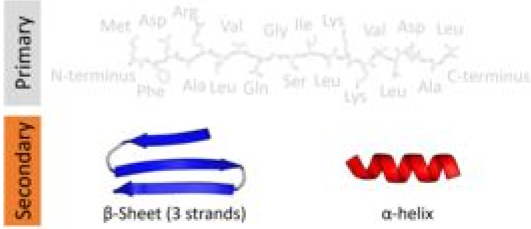
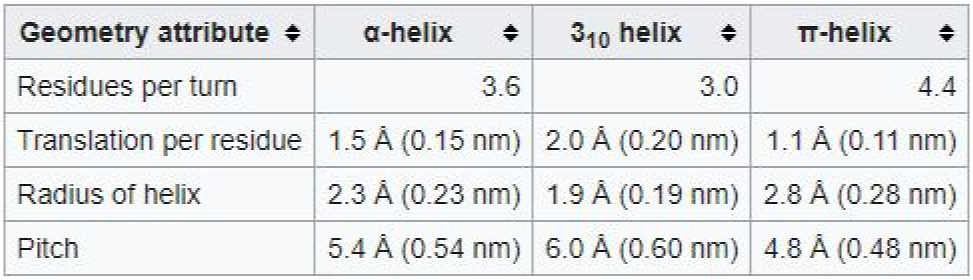


Fig1: Primary And Secondary Protein Structures PDB: 1AXC

The most common secondary structures are **alpha helices** and **beta sheets**. Other helices, such as the **3_{10}** helix and **π** helix, are calculated to have energetically favorable hydrogen-bonding patterns but are rarely observed in natural proteins except at the ends of α helices due to unfavorable backbone packing in the center of the helix.



Geometry attribute ◆	α-helix ◆	3_{10} helix ◆	π-helix ◆
Residues per turn	3.6	3.0	4.4
Translation per residue	1.5 Å (0.15 nm)	2.0 Å (0.20 nm)	1.1 Å (0.11 nm)
Radius of helix	2.3 Å (0.23 nm)	1.9 Å (0.19 nm)	2.8 Å (0.28 nm)
Pitch	5.4 Å (0.54 nm)	6.0 Å (0.60 nm)	4.8 Å (0.48 nm)

Table 1: Structural features of the three major forms of protein helices

Other extended structures such as the **polyproline helix** and **alpha sheet** are rare in native state proteins but are often hypothesized as important protein folding intermediates. Tight turns and loose, flexible loops link the more regular secondary structure elements. The random coil is not a true secondary structure but is the class of conformations that indicate an absence of regular secondary structure.

Amino acids vary in their ability to form the various secondary structure elements. Proline and glycine are sometimes known as **helix breakers** because they disrupt the regularity of the α helical backbone conformation; however, both have unusual conformational abilities and are commonly found in turns. Amino acids that prefer to adopt helical conformations in proteins include methionine, alanine, leucine, glutamate, and lysine (**MALEK** in amino-acid 1-letter codes); by contrast, the large aromatic residues (tryptophan, tyrosine, and phenylalanine) and C_β-branched amino acids (isoleucine, valine, and threonine) prefer to adopt **β-strand** conformations. However, these preferences are not strong enough to produce a reliable method of predicting secondary structure from sequence alone.

Gopalasamudram Narayana Ramachandran developed a method to predict the secondary structures in 1968. He classified the secondary structures into alpha helix or beta sheet based on phi and psi torsion angles. The ramachandran plot (as shown below) shows that, for a particular range of phi and psi values, an amino acid has high chances of being alpha helix or beta sheet.

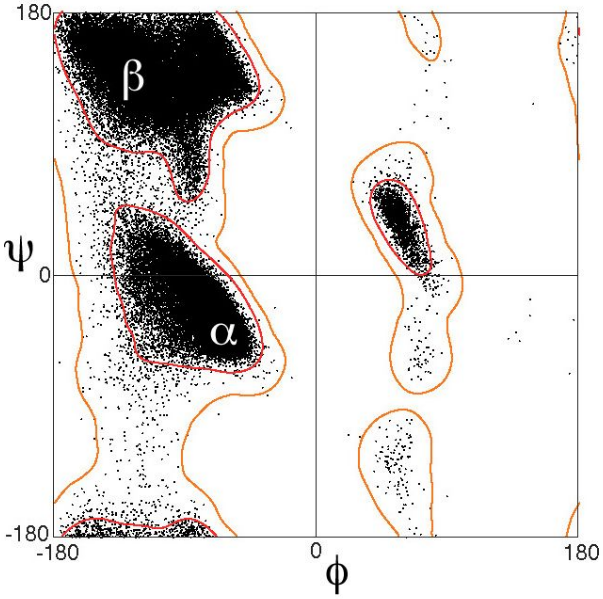


Fig2: Ramachandran plot representing the dependence of secondary structure of proteins on phi and psi torsional angles

Related Works

The accuracy of secondary structure prediction has improved significantly by aligning protein sequences. Recently, a two-stage neural network has been used to predict protein secondary structure based on the *position-specific scoring matrices* generated by **PSI-BLAST**. And by adding neural network units that detect periodization in the input sequence, the use of tertiary structural class causes a marked increase in accuracy. The best case for prediction is **79%** for the class of all-alpha proteins. A method has also been developed to check against from an updated protein database. This method correctly predicts **69%** of amino acids for a *three-state description* of the secondary structure, including *helix, 3-sheet,* and *coiled* in the whole database. Meanwhile, the effect of training a neural network with different types of multiple sequence alignment profiles derived from the same sequences is shown to provide a range of accuracy from **70.5%** to **76.4%.** Another method is using evolutionary information contained in multiple sequence alignments as input to neural networks. It has a sustained overall accuracy of **71.6%** in multiple cross-validation tests on 126 unique protein chains. The above methods are composed of several neural network approaches and taking into account for sequence similarity. In addition, there is a notable algorithm called **PHD** based on a neural network learning process system. This is the first method to break the **65%** boundary on Q3 accuracies of secondary structure prediction methods. Besides, a method has been tested on 59 proteins in the database and yields **72%** success in class prediction, **61.3%** of residues correctly predicted also for three states.

Different methods use different classification of secondary structures. As mentioned earlier, Ramachandran plot considers only alpha helices and beta sheets. DSSP and STRIDE algorithms mentioned earlier, classifies the protein sequence into 7 types. It differentiates between different helices and sheets and further considers the building blocks of these structures as separate type if occurred deserted.

Our Method

Since the neighbouring amino acids play a significant role on phi and psi values, we can train a model with a set of sequences with known sequences and torsional angles, and use their result to predict the secondary structure of the new protein. Only the effect due to the 2 amino acids on either sides is considered to keep the computational complexity low. The algorithm for training goes as follows:

- A set of PN number of known protein sequences are considered for training.
- An empty array (RAI) of dimensions 20x20x20x20x20x4 is considered where array of length 4 stores the average phi value, average psi value and its repetition time (RT) and the assigned secondary structure (SS).
- A map from Amino acid to the positions 1-20 is defined as function AA.
- For p in 1 to PN do
    - ln =length of the protein
    - For i=3 to (ln-2) do
        - RAI [AA(i-2)] [AA(i-1)] [AA(i)] [AA(i+1)] [AA(i+2)] [1] = (existing value*RT + phi(AA(i)))/(RT+1)
        - RAI [AA(i-2)] [AA(i-1)] [AA(i)] [AA(i+1)] [AA(i+2)] [2] = (existing value*RT + psi(AA(i)))/(RT+1)
        - RAI [AA(i-2)] [AA(i-1)] [AA(i)] [AA(i+1)] [AA(i+2)] [3] = existing value +1
- For i1, i2, i3, i4, i5 in 1 to 20 do
    - Assign H or S or E or X to RAI[i1][i2][i3][i4][i5][3] based on the phi and psi values

This approach generates **φ** and **ψ** angles for a given protein sequence which are averaged values for all the occurrences of all the pentapeptides in the training data set. From the raw data obtained from the pdb files, the 'phi' and 'psi' angles of the train database were computed by simple mathematical calculations involving the relative coordinates of the atoms in the main chain of the amino acids. The labels are assigned based on the following table.

Secondary Structure	**φ range**	**ψ range**
α-helix (H)	$-100^\circ \leq \varphi \leq -40^\circ$	$-60^\circ \leq \psi \leq 30^\circ$
B-sheets (E)	$-150^\circ \leq \varphi \leq -50^\circ$	$100^\circ \leq \psi \leq 180^\circ$
Coils (S)	**All other accessible regions**	
Unassigned (X)	**-**	**-**

Table 2: The region represented for each secondary structure

The influence of protein nearest neighbors for the angles on **φ** and **ψ** is evaluated as interrelated, so we make an assumption that the **φ** and **ψ** angles of **nth** amino acid are influenced by the pentapeptide as *(n − 2), (n − 1), (n +1), and (n +2)*. We could also use tripeptide or hepta peptides and their corresponding **φ** and **ψ** angles to predict the unknown protein structure. Besides, it is apparent that the formation of a **helix** or **sheet** including the amino acid at position (n) involves amino acids at positions (n−4), (n−3), (n+3), and (n+4). Consequently, by considering the problem of computational complexity and data set size, the pentapeptide relation is chosen as our rule.

The algorithm for testing goes as follows:

```
● Read the RAI matrix generated from the training.
● Read the unknown protein sequence as PS.
● Create an empty array SS = secondary structure of the amino acid residues
● ln=length of PS
● For i=3 to ln-2 do
    ○ Query PS(i-2,i-1,i,i+1,i+2) with RAI matrix
    ○ If Yes
        ■ If RT of RAI !=0 do
            ● Assign the structure from RAI to SS(i)
        ■ Else
            ● Query PS(i-1,i,i+1) with reduced RAI matrix
            ● If Yes
                ○ If RT of RAI !=0 do
                    ■ Assign the structure from RAI to SS(i)
                ○ Else
                    ■ Assign X
```

The possible pentapeptides are 20x20x20x20x20, which is a huge size, and there is a high probability that some of them never occur in the training dataset. As mentioned in the algorithm, when the count of a matched pentapeptide is zero, we use a reduced RAI matrix. Reduced RAI matrix is a tripeptide based matrix taking the average of all the pentapeptides which contain the target tripeptide. If even the reduced RAI matrix has repeat count 0, then X is assigned.

Test results:

Now, moving on to the training dataset consideration. DSSP, STRIDE and other algorithms also work on trained data bases. They consider all the proteins available in the PDB bank. But, instead of training on all the protein database, we propose a different approach of adaptively choosing database as per the target protein. From the structural classification of proteins (SCOP), we know that homologous proteins have similar structural and functional features. So, for a protein of class A, prediction based on the training database of class A proteins can be better than prediction based on protein database in general. For the same, we applied our algorithm on 3 different classes of proteins, and then reapplied it on all the classes together. The protein databases considered are:

1) All alpha : Myoglobins :Train dataset size = 350 ;Test dataset size=94 proteins

2) All beta : Proteases :Train dataset size = 400 ;Test dataset size=79 proteins

2) Alpha/Beta : TransGlucosidases :Train dataset size = 650 ;Test dataset size=119 proteins

4) Mixed : All 3 :Train dataset size = 1400 ;Test dataset size=292 proteins

For evaluating the results of our work, we also ran the DSSP algorithm on the datasets, the output of which was assumed to be the ground truth for our performance analysis. One major thing to be noted here is that, the DSSP algorithm doesn't assign any structure in some cases where it is unable to classify. So we first compared our results with DSSP considering the unassigned resides and then also compared them, neglecting the unassigned residues.

First the residue based error was measured using confusion matrix and accuracy. Later, protein wise errors were measured to check the consistency over various proteins.

Given below are the confusion matrices for all the four datasets mentioned above.

A) For the Myoglobin database considering the unassigned state X, we get an accuracy of 82.5% and an accuracy of 91% neglecting the unassigned states.

This work\DSSP	H	S	E	X
H	11230	894	0	75
S	168	621	0	702
E	0	104	8	633
X	4	6	0	350

Table 3a: Confusion matrix for Myoglobin database

B) For the Protease database considering the unassigned state X, we get an accuracy of 59.7% and an accuracy of 73.1% neglecting the unassigned states.

This work\DSSP	H	S	E	X
H	1370	1351	95	161
S	159	1560	914	826
E	287	710	6630	1971
X	0	2	81	157

Table 3b: Confusion matrix for Protease database

C) For the TransGlucosidase database considering the unassigned state X, we get an accuracy of 55.8% and an accuracy of 70.1% neglecting the unassigned states.

This work\DSSP	H	S	E	X
H	23568	8260	1722	2153
S	1633	7430	3595	5685
E	494	3668	16174	10259
X	19	6	49	306

Table 3c: Confusion matrix for TransGlucosidase database

D) For the combined database considering the unassigned state X, we get an accuracy of 59.8% and an accuracy of 74% neglecting the unassigned states.

This work\DSSP	H	S	E	X
H	36149	10500	1857	2499
S	1992	9617	4492	7139
E	768	4481	22789	12827
X	23	14	130	813

Table 3d: Confusion matrix considering all the proteins as database

- On averaging the test accuracy for the three databases including the unassigned residues is equal to (94*82.5+59.7*79+119*55.8)/(94+79+119)=65.45%

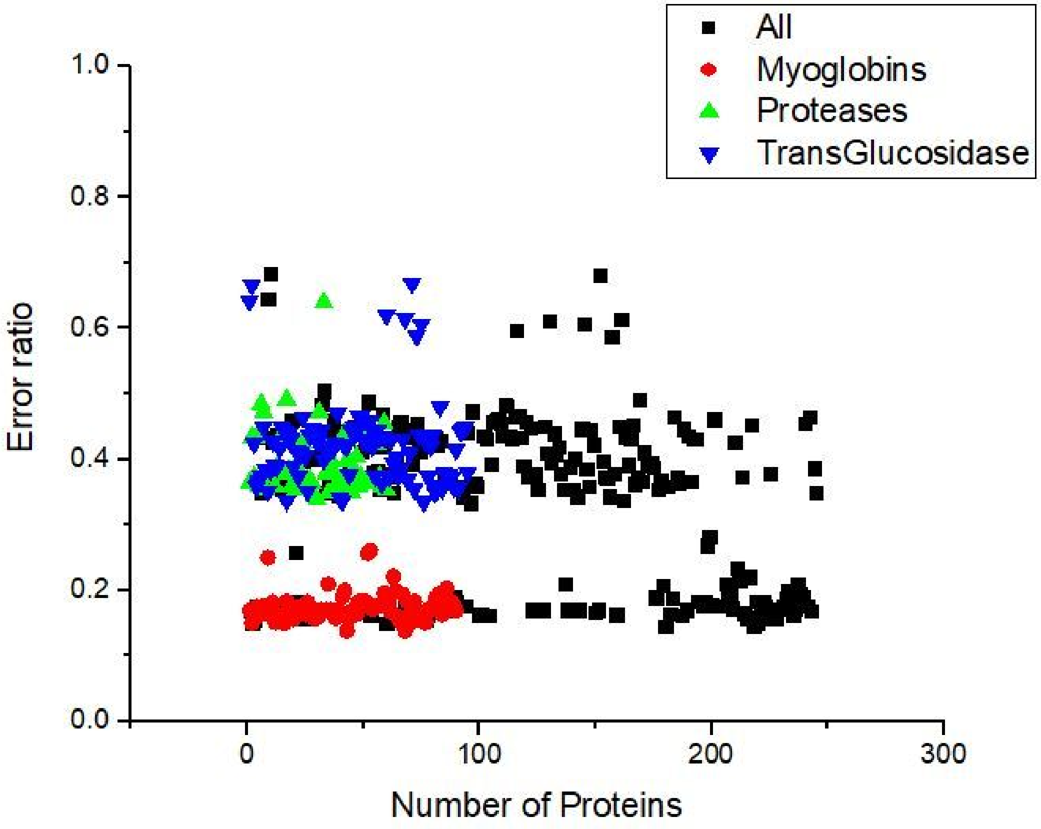
  But if we combine the training database, we get an accuracy of 59.8%

- On averaging the test accuracy for the three databases neglecting unassigned residues is equal to (94*91+73.1*79+119*70.1)/(94+79+119)=77.64%

  But if we combine the training database, we get an accuracy of 74%

So, using a segregated training data set is better than using a common training database for different kinds of proteins.

In figure 3, we have the error ratio for different databases. For myoglobins we have a low error rate compared to other datasets. Proteases and TransGlucosidases have almost 40% of wrong predictions.



Conclusion:

Ramachandran plot can be used to assign secondary structures to protein sequences from the torsional angles (phi and psi) between Nitrogen & alpha-Carbon and alpha-Carbon & the next Carbon atom respectively, but a universal boundary on phi and psi cannot be set for all amino acids. It largely depends upon the amino acids and their neighborhood. Because of which, the nearest 4 neighbors (2 before and 2 after the amino acid in consideration) are observed and the average value of the torsional angles is computed. This gives the deserved attention to the neighboring amino acids and was observed to produce good results of over 70% on average.

Prepared By: *Shaik Mohammed Sameer*(15EC10054), *Atul*(15EC10067), *Avinab Saha*(15EC10071)

References:

1. Kabsch W, Sander C (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers*. 22 (12): 2577–637. doi:10.1002/bip.360221211. PMID 6667333.
2. Frishman D, Argos P. Knowledge-Based Protein Secondary Structure Assignment Proteins: Structure, Function, and Genetics 23:566-579 (1995)
3. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963;7:95–9
4. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins Scott A. Hollingsworth and P. Andrew Karplus, doi: 10.1515/BMC.2010.022
5. Protein Secondary Structure Prediction Based on Ramachandran Maps Yen-Ru ChenSheng-Lung PengYu-Wei Tsay, Part of the Lecture Notes in Computer Science book series (LNCS, volume 5226)
6. PDB website : https://www.rcsb.org/
7. Wikipedia: https://www.wikipedia.org/

x