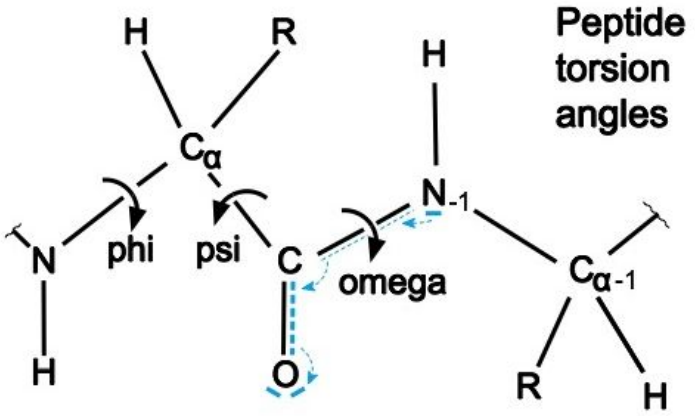
Protein Secondary Structure Prediction using Ramachandran Plot

Group :D
Shaik Mohammed Sameer(15EC10054), Atul(15EC10067), Avinab Saha(15EC10071)

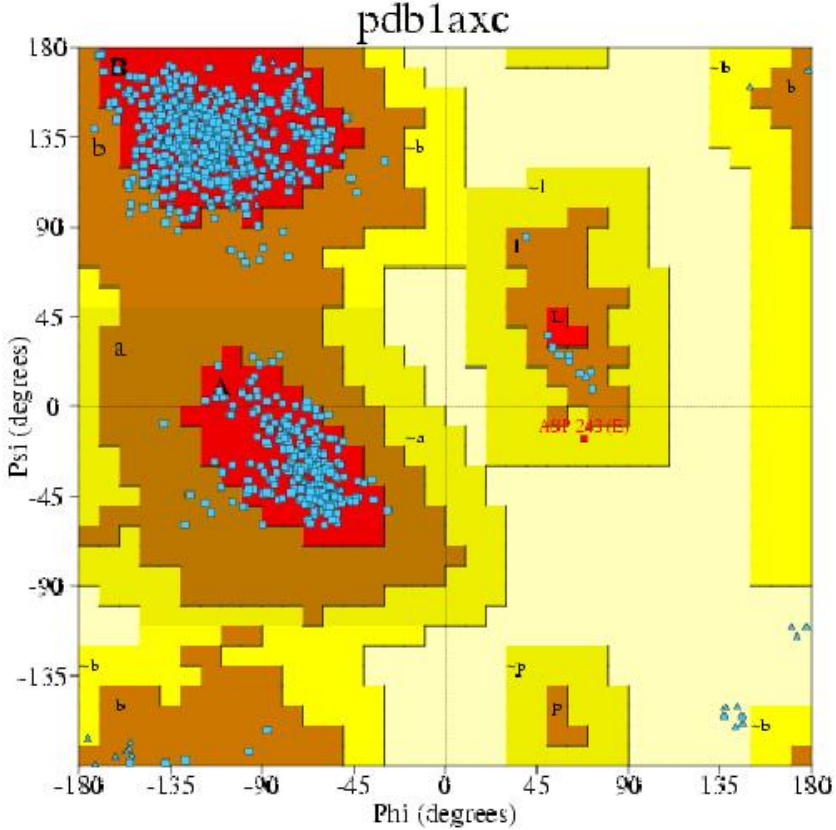
Preliminaries: Peptide Torsion Angles



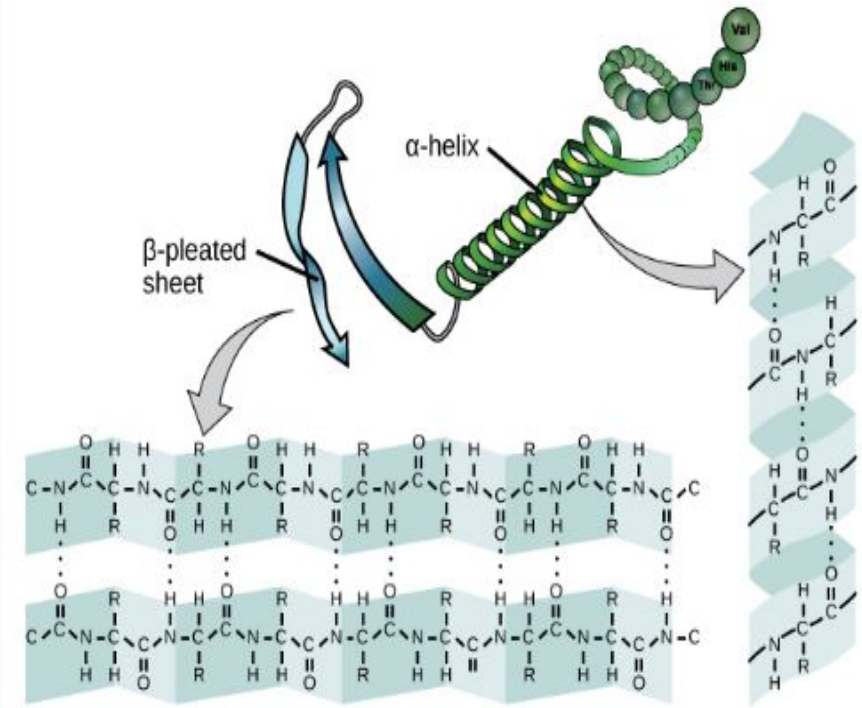
- Three Torsion angles phi (φ), psi (ψ), and omega (ω) exist in the main chain of a polypeptide
- Two torsion angles φ (N, C_α, C, N) and ψ (C_α,C,N,C_α) are on either side of the C_α atom while ω (O,C,N,H) describes the angle for the actual peptide bond
- The torsion angle ω within the protein backbone is flat and fixed to 180˚(Trans-conformation) or 0˚(Cis-conformation)

Preliminaries: Ramachandran Plot

- With the ω angles restricted, the polypeptide main chain exhibits considerable freedom to rotate around the φ & ψ torsion bonds
- In the **Ramachandran plot** shown to the right, the φ/ ψ space is visualized both the angles in range **-180° to 180°**
- The red, brown and yellow regions represent the favored, allowed, and "generously allowed" regions



Protein Secondary Structures



- The most common secondary structures in proteins are **alpha helices**, **beta sheets** and **Turns**
- Other rarely found secondary structures in natural proteins are: π helix, 3_{10} helix, polyproline helix & alpha sheets

Our Approach to predict Secondary Structures!

Workflow

Computation of RAI Index for all possible pentapeptides

Assigning secondary structure based on secondary structure stored in RAI matrix

For each central residue of pentapeptide, RAI index is obtained from RAI Index

Training Phase-

Ramachandran Adjacency Index Calculation

Ramachandran Adjacency Index (RAI)

Initialization Stage:

- A set of **PN** number of known protein sequences are considered for training.

- **RAI** matrix of dimensions *20x20x20x20x20x4* is supposed to be the array of length that stores the **average phi value**, **average psi value** and its **repetition time** (RT, stored in 'count' variable) and the assigned **secondary structure** (SS), but was instead stored in 4 arrays of dimension *20x20x20x20x20* (phi, psi, count, labels) for the sake of convenience.

- A map from Amino acid to the positions **1-20** is maintained in the vectors 'proteins' and 'initials' and is denoted by 'AA'

Ramachandran Adjacency Index (RAI)

Training Stage:

- 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

Assignment Stage:

- 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 shown in next slide

Ramachandran Map

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

- The secondary structures are assigned based on the above phi and psi values

Testing Phase

Testing Phase

Testing a new protein sequence

- Read the RAI matrix generated from the training.
- Read the unknown protein sequence as PS of length ln
- Create an empty array SS to store secondary structure of the amino acid residues
- 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 (undecided state)

Results

Results: Datasets

The algorithm was applied on 3 different classes of proteins, and then reapplied it on all the classes together. The protein databases considered are:

All alpha	Myoglobins	Train dataset size = 350	Test dataset size=94
All beta	Proteases	Train dataset size = 400	Test dataset size=79
Alpha/Beta	TransGlucosidase	Train dataset size = 650	Test dataset size=119
Mixed	All 3	Train dataset size = 1400	Test dataset size=292

- **The results obtained by our algorithm were compared with the results obtained by the famous DSSP algorithm considering them to**

Confusion Matrices: All Alpha (Myoglobins)

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

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.

Confusion Matrices: All Beta (Proteases)

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

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.

Confusion Matrices: Alpha and Beta (TransGlucosidases)

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

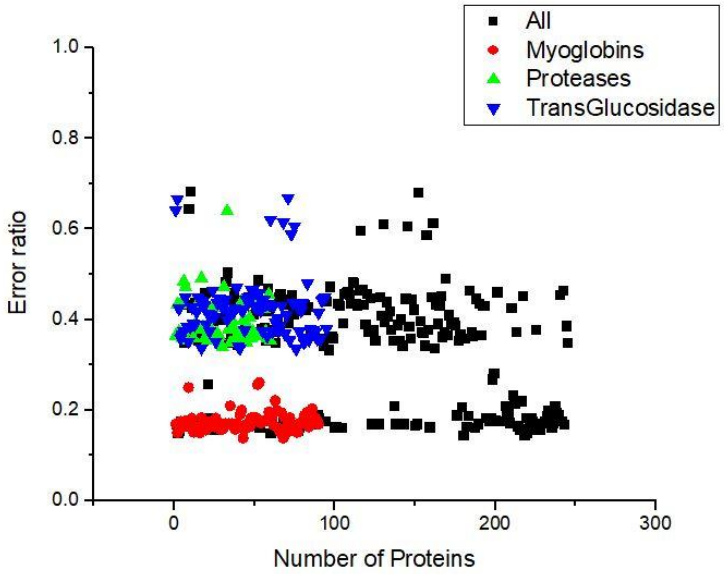
For the TransGlucosidase database considering the unassigned state X, we get an accuracy of 55.8% and an accuracy of 70.1% neglecting the unassigned states.

Confusion Matrices: Mixed (All the proteins together)

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

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

Error Ratio for different databases



We have the error ratio for different databases. For myoglobins we have a low error rate compared to other datasets. Proteases and TransGlucosidase have almost 40% of wrong predictions.

Thank You!
Questions ??