https://doi.org/10.17952/37EPS.2024.P1278

Machine Learning Reveals Structural Properties of Monosaccharides and their Peptide Conjugates

Michele Casoria[1,2,3], Marina Macchiagodena[1,3], Carla Bazzicalupi[1,3], Claudia Andreini[1,3,4], Gianni Cardini[1,3], Anna Maria Papini[2,3], Marco Pagliai[1,3]

1 Research Unit of Computational Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy

2 Interdepartmental Research Unit of Peptide and Protein Chemistry and Biology, University of Florence, 50019 Sesto Fiorentino, Italy

3 Department of Chemistry "Ugo Schiff", University of Florence, 50019 Sesto Fiorentino, Italy

4 Magnetic Resonance Center - University of Florence, 50019 Sesto Fiorentino, Italy

e-mail: michele.casoria@unifi.it

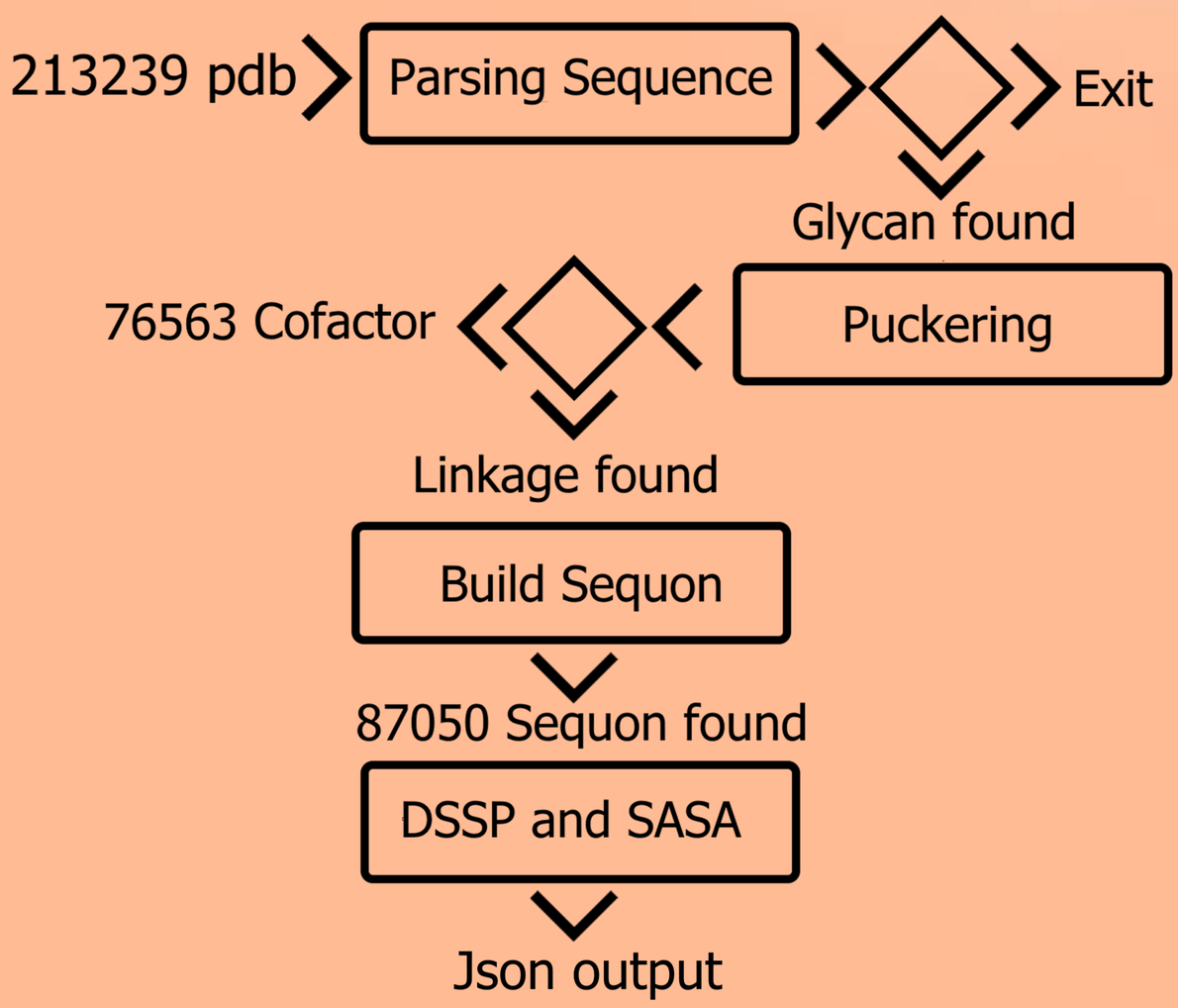
Introduction

The Protein Data Bank (PDB) is a valuable resource for investigating the three-dimensional conformations of glycoproteins, facilitating our understanding of the impact of glycosylation on proteins. It is important to note that the glycan components in these structures frequently contain inaccuracies, which can range from subtle irregularities to major errors. Projects such as PDB-REDO [1] aim to enhance precision through experimental validation. Previous studies [2] on carbohydrates report that PDB database lack precise categorization and only a limited set of structural parameters are present. For these reasons we conducted a systematic analysis of the complete PDB proteins set, aiming to categorize sugars with precision based on their glycosidic linkages

Methods

On April 2024 213239 pdb files are available on Protein Data Bank.

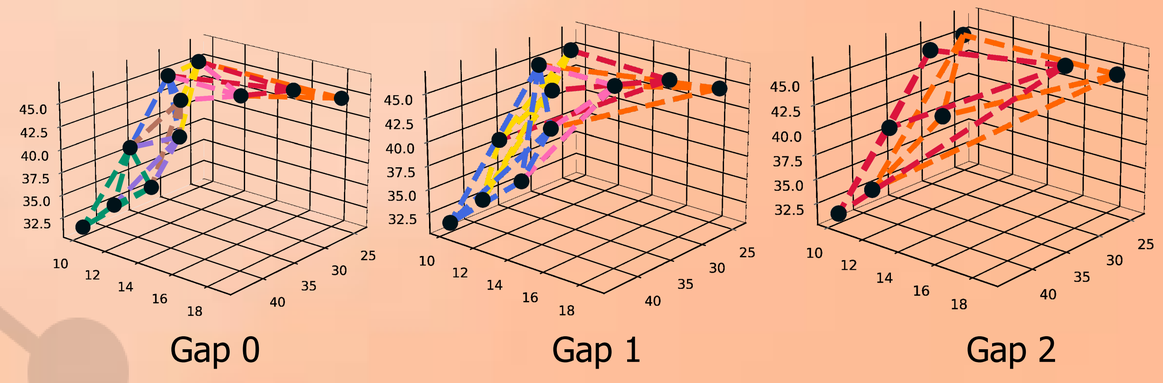
We developed a python script, which allows us to parse all files and collect the data in Json format files.
From each file resolution, taxonomy and owab are stored.
All residues within each protein chain have been analyzed, collecting relative atom coordinates and filtering the residues based on their 'resname' values.
To identify carbohydrate structures, we employed the 'HETNAME' value, specifically selecting 1724 three-letter code entries from the PDB HET Group Dictionary.
To ensure a standard procedure to determine the local structure of glycopeptides, particularly within the glycosylated regions, we deliberately selected 11-mer peptides spanning from the -5 to +5 positions with respect to the glycosylation site strategically positioned at the centre.
We computed the relative Solvent Accessible Surface Area (SASA) and secondary structures using the Shrake-Rupley and Define Secondary Structure Proteins (DSSP) algorithms respectively.



Clustering

We characterized the structures as a sequence of local conformations, which are delineated by forming eight tetrahedrons over four consecutive Cα atoms (Gap 0), five tetrahedra using alternate C_α atoms (Gap 1) and two tetrahedra with two alternate C_α atoms (Gap 2). Each tetrahedron is subsequently employed to calculate 12 geometric indices (GI parameters) as features. In total, we have 15 tetrahedra, each contributing 12 features, resulting in a collective description of 180 features for a given structure.

Tetrahedra Gap 0	Tetrahedra Gap 1	Tetrahedra Gap 2
1 2 3 4	1 3 5 7	1 4 7 10
2 3 4 5	2 4 6 8	2 5 8 11
3 4 5 6	3 5 7 9	
4 5 6 7	4 6 8 10	
5 6 7 8	5 7 9 11	
6 7 8 9		
7 8 9 10		
8 9 10 11		



Gap 0 Gap 1 Gap 2

GI Parameters

Feature	Description
1	$dist(i, i+2)$
2	$dist(i, i+3)$
3	$dist(i+1, i+3)$
4	$\sqrt[3]{volume(i, i+1, i+2, i+3)}$
5	$\text{perimeter}(i, i+1, i+2, i+3)$
6	$\sum_{i=0,\, i\neq j}^{3} variance(i)\ x\ variance(j)$
7	$area(i, i+1, i+2)$
8	$area(i+1, i+2, i+3)$
9	$area(i, i+2, i+3)$
10	$\text{perimeter}(i, i+1, i+2)$
11	$\text{perimeter}(i+1, i+2, i+3)$
12	$\text{perimeter}(i, i+2, i+3)$

These 180 features were collected into a matrix describing all the sequons, UMAP reduction to three dimensions was applied, followed by clustering of the structure using the DBSCAN algorithm, with an epsilon value of 0.4 and a minimum sample size of 4.
The resulting evaluation metrics are as follows: Silhouette Score: 0.69, Cluster Compactness: 0.04, and Cluster Separation: 12.35. Additionally, we identified 2625 clusters and 21 structures in cluster -1 as noise.

Future Developments

We are developing a database named GPS-Hub (GlycoPeptideStructure Hub), created by M. Casoria et al., which will compile comprehensive information about sequons. This database will include the biological significance of the organisms in which these peptides are found. Additionally, it will contribute to enhancing classical force fields used in molecular modeling and simulations.

Reference and Acknowledgments

1. V. Beusekom, Acta Cryst. F **2018**, 74, 463–472.
2. J. L. de Meirelles, J. Chem.Inf. Model. **2020**, 60, 684−699

European Peptide Society is kindly acknowledged for the awarded registration grant to Mr. Michele Casoria

MUR is kindly acknowledged for Michele Casoria's PhD scholarship