International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I12P213

Volume 70 Issue 12, 118-126, December 2022

Review Article

Approach and Techniques for Precise Prediction of N-Linked Glycosylation from Human Protein using Artificial Intelligence

Mubina Malik[1], Jaimin N Undavia[2]

[1,2]*Department of Computer Science & Applications, CMPICA, CHARUSAT, Charotar University of Science and Technology (CHARUSAT), CHARUSAT campus, Changa, Gujarat, India*

[1]*Corresponding Author : mubinamalik.mca@charusat.ac.in*

Received: 10 August 2022 Revised: 13 November 2022 Accepted: 26 November 2022 Published: 24 December 2022

Abstract *- Glycosylation is the most common post-translational modification of protein in all territories, which plays a significant role in biological processes. Amongst them, n-linked glycosylation is the most crucial modification, which is closely related to certain diseases such as cancer, diabetes, HIV infection, Alzheimer's disease and atherosclerosis, and liver cirrhosis. Recent advancements in biological knowledge are depicted in this article, ultimately targeting the computer science field. Machine learning and deep learning techniques are major keys to predicting various protein modifications. Through the review of several models which have been made existing for prediction and show high accuracy but result as false positives due to the poor biological knowledge, updated datasets and techniques used. Targeting precise prediction, drawbacks of the existing model and discussed parameters and techniques were emphasized to model solution in this paper. In this study, databases were combined, namely UniprotKB, dbPTM, and nGlycositeAtlas, which are experimentally verified and updated with window size 21. This window size is best for the n-linked glycosylation. After combining datasets and removing the redundancy, 11254 unique proteins and 33859 glycosites were received for further study. CD-HIT algorithm was implemented to remove the redundancy with threshold 0.9. These nearby locations for similar pattern sequences have been identified for asparagine residue for n-linked glycosylation. The protein sequence is a combination of 20 amino acids, which were required to convert into numerical form through encoding methods. Various encoding methods have conversed for n-linked glycosylation. With the biological features, amino acid encoding methods such as substitution matrices - Position Specific Scoring Matrix (PSSM) and Physicochemical properties encoding VHSE8 are the vital methods which improve the accuracy in n-linked glycosylation prediction.*

Keywords *- Artificial intelligence, Deep learning, Human protein, Machine learning, N-linked glycosylation.*

1. Introduction

Amongst Eukarya, Bacteria, and Archaea, glycosylation is the common protein post-translational modification (PTM) [1]. Protein glycosylation is the process of adding sugar molecules to a protein, lipids, and other organic molecules inside and outside the Cell. In this process, the carbohydrates attached to lipids and proteins, specifically to a residue which makes a glycosidic bond, are called glycan. It is a covalent modification which plays a vital role in immune protein localization, system responses [2], Intracellular signaling, folding, trafficking, and cell-cell interactions [3][4]. Any dysfunctional glycan can lead to diseases such as cancer, diabetes, HIV infection, Alzheimer's disease and, atherosclerosis, liver cirrhosis. In addition, glycosylation plays a key role in SARS-CoV-2 infection [5]. Many authors state that more than 50% of plasma proteins are glycosylated [6][7][8]. As glycosylation plays a vital role in biological processes within the human body, the detection of protein glycosylation is also mandatory. However, Experimental detection of glycosylation is possible; it is also challenging and requires extensive laboratory work and expense. Due to this limitation of laboratory experiments, there is a need to develop a tool which predicts glycosylation. To overcome the limitations such as accuracy, speed and cost, computational analysis of Protein glycosylation is important. In this current era, Deep learning and Machine learning, the subset of Artificial Intelligence, is booming in all areas, including healthcare and bioinformatics. [9]

Glycosylation occurs in Endoplasmic Reticulum (ER) – which helps in protein folding, and Golgi Apparatus – which informs protein where to go. Glycosylation usually occurs in the side chain of residues such as Tryptophan (Trp), Alanine (Ala), Serine (Ser), Threonine (Thr), Asparagine (Asn),

Arginine (Arg), Aspartic acid (Asp), Isoleucine (Ile), Lysine (Lys), Valine (Val), Glutamic acid (Glu), Proline (Pro), Tyrosine (Tyr), Cysteine (Cys) and Glycine (Gly) [10]. However, all residue glycosylation occurs more frequently on Ser, Thr, Asn and Trp residue [11]. Glycosylation can be classified into N-linked glycosylation, O-linked glycosylation, C-linked glycosylation, S-linked glycosylation, phoglycosylation and glypiation [12]. Three main types of glycosylation are N-linked, O-linked and C-linked. All these three types differ in terms of location in the protein chain and targeted residue within the protein sequence.

Amongst all the types, the most profound type is N-linked glycosylation. If done experimentally, the characterization process of N-linked glycosites in glycoproteins is technically tough, costly, and takes lots of time. It plays a vital role in protein folding and protein stability. So, it is closely related to the development of drug design. N-linked glycosylation occurs in both ER and the Golgi complex. According to biochemistry, an oligosaccharide (glycan) attachment to the amide nitrogen of an asparagine (Asn) residue of protein is called n-linked glycosylation. Mainly N-linked glycosylation happens in N-X-S/T (N – Asparagine, S – Serine, T – Threonine) sequence and occasionally in N-X-C (C – Cysteine) where X can be any amino acid except Proline [13].

1.1. Introduction to Problem Statement and Research Gape Identification

As discussed above, detecting glycosylation using experimental techniques is still challenging and takes a long time and cost. Moreover, it is also quite challenging to understand glycosylation because of the various diversity of glycan attached to proteins which limits the consideration of the specific function of glycosylation. Because of the larger number of enzymatic steps involved, glycosylation is the most intricate post-translational modification. Recent advancements in Artificial Intelligence overcome the drawbacks of detecting glycosylation using experiments. Birgit Eisenhaber and Frank Eisenhaber authors have stated that "glycosylation prediction is still not satisfactory and sequence-based approach has low prediction rate because of the number of glycotransferases are not explored and even less studied" [14]. Manikandan Muthu, Sechul Chun et al. l. have highlighted available bioinformatics resources and gave the conclusion that there is a massive gap between available bioinformatics tools and real-time applications. However, many glycosylation prediction tools are developed, but not even 1% of the available tools are used for glycosylation in cancers [15]. Figure 1 shows the objective and research gap, which is considered and discussed in the article.

It was stated by many authors in 2020 and 2021 [16,17,18] that most of the prediction models have evaluated their performance at every N in protein sequence without the confirmation of N-X-S/T sequon. Also, additional features such as disordered regions and physicochemical properties can be used for accurate and better results.

Previous methods used for the prediction are accurate at a certain level but still left out with some problems. The problems are:

- Small protein sequence dataset available to train the model
- Dataset available but not verified experimentally
- Window size selection for the experiment
- Incomplete information available on amino acids for feature encoding
- No feature selection techniques were used or available

Considering these problems and the research gap, three datasets were combined, namely *UniprotKB, dbPTM, and nGlycositeAtlas*, to get the updated and experimentally verified data. Window size 21 is selected to train the model. Also, the article explored statistical and physicochemical properties to get better results.

2. Literature Review and Research Gap

To conclude, a number of papers were reviewed to predict the model for n-linked protein glycosylation and identified the methods used in the prediction model, the dataset used and limitations for accurate prediction. Based on the review, protein glycosylation prediction can be done by two approaches 1) protein sequence-based approach and 2) Protein structure-based approach. The protein sequence-based approach is also categorized in two ways: residue level and sequence level, which is represented below in Figure 2. The highlighted part in the figure depicted a suitable approach for the prediction of n-linked glycosylation. Further paragraphs describe why the selected approach is best and suitable for predicting true positive sites. Figure 2 highlights the approach used to identify the protein PTM site. For selecting the proper approach, various models and datasets have been reviewed. In the first approach, the input will be a protein sequence, and protein glycosylation is identified based on that protein sequence. For this approach, protein sequence databases are used to train the model.

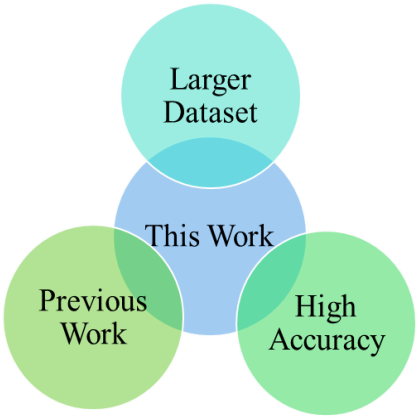

Fig. 1 Research Gap

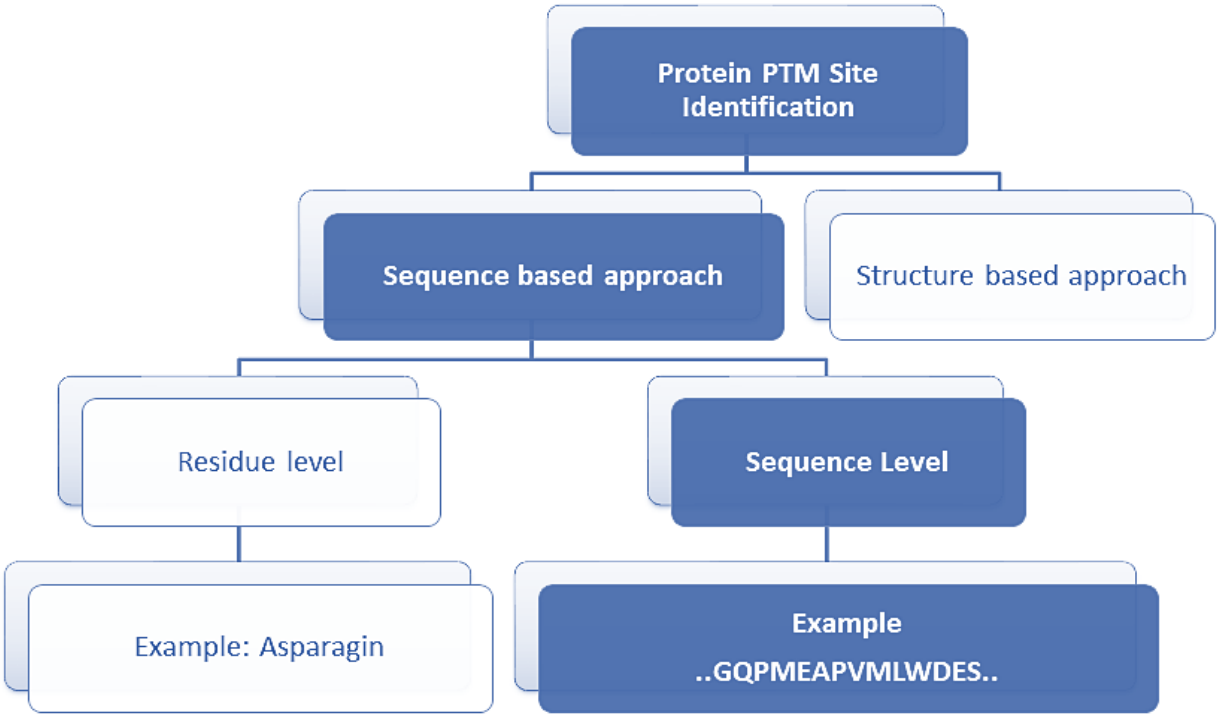

Fig. 2 Protein n-linked glycosylation prediction approach

There are various datasets available which contain protein sequence information and are categorized as primary (Swiss-Prot – protein sequence), secondary (UniProt Knowledgebase – Protein sequence with functional information), complex (UniProt - Protein sequence with a wealth of additional genomic information) [53] and specialized (Specific to disease). Moreover, other PTM databases are available, including specific post-translational modification information, e.g. glycosylation, phosphorylation etc. Here some datasets have been highlighted which are ready to use and developed from existing databases such as PTMCode [20,53], dbPTM [21], ProteomeScout [22,23], and iPTMNet [24]. These datasets are available to train the model for specific types of modification. Author Dan Ofer, Nadav Brandes et al. l. have mentioned the Natural Language Processing (NLP) Classification, which describes the local and global classification of protein sequence [25].

The residue level approach is based on the local classification where specific amino acids play an essential role; for example, Protein Post Translational Modification prediction targets specific amino acids. Global classification is based on the entire sequence, for example – Protein secondary structure prediction. In the second approach, prediction of protein, modification is carried out by the structure of the protein. Databases for protein structure include Protein Data Bank (PDB), AlphaFold DB [26] etc. are further used in computation methods such as machine learning and deep learning.

Various computational approaches for n-linked glycosylation prediction, which is based on the protein sequence as well as protein structure, were considered.

Some of the prediction approaches for n-linked glycosylation which uses sequence-based features are NetNGlyc [27], GPP [28], EnsembleGly [29], GlocoPP [30], GlycoEP [31], NGlycoGo [16], GlycoMine [32], SprintGly [33]. A few structure-based approach models have been developed: NGlycPred [54] and GlycoMine^{struct} [35]. Howbeit, Figure 2. describe the individual approach; some model uses the hybrid approach, such as N-GlycDE [17] and DeepNGlycPred [18].

As per the literature review and previous model, it is observed that solved protein structures added in Protein Data Bank (PDB) are still inadequate and do not have a solved protein structure for the query sequence. For these reasons sequence-based approach was selected, which is further classified as residue level and sequence level. Except for NetNGlyc[26], N-GlycDE[17], and DeepNGlycPred [18], almost all the listed models in the afore-mentioned paragraphs evaluated their performance using residue N without confirming N-X-[S/T] motif to identify n-linked glycosylation [13] and thus, performance is overestimated and resulted in high accuracy. To achieve comparable performance, consensus sequence N-X-[S/T] must be considered for analysis. However, the consensus sequence does not necessarily confirm n-linked glycosylation because one-third to half of the consensus sequence is hidden deep

inside the protein, which is not accessible by the glycosylation enzyme [36,37]. Thus, performance evaluation based only on the consensus sequence might result in a false positive. In addition to sequence features few predicted structural features such as secondary structure (SS) to check the helix, strand and coil in a sequence, disordered residue, and Accessible Surface Area (ASA) to check the accessibility of N residue are used to increase the accuracy for the prediction model. GlycoMinestruct [35] and SprintGly [33] use structural features, but it is evaluating performance without confirming N-X-[S/T] sequence. Recently deep learning-based approach DeepNGlycPred [18] and SVM-based approach N-GlycDE [17] are the pivotal work in the computation biology for n-linked glycosylation prediction from the protein sequence, which is based on the N-X-[S/T] consensus sequence as well as use the predicted structural features. As shown in Figure 2, for the better performance of n-linked glycosite prediction, the best approach is sequence level in which the protein sequence is fed into the model for analysis and results as n-linked glycosylated.

3. Materials and Methods

In recent years, healthcare has been booming with artificial intelligence techniques such as deep learning and machine learning. For the prediction of n-linked glycosylation, techniques can be used to solve bioinformatics problems. There are various challenges in computational biology, such as heterogeneous data, realistic interpretation of information, selection of proper architecture, hyperparameter identification, non-glycosylation sites, heterogeneous and high dimensional features, accurate performance, especially on independent datasets etc. [38,39]. The most significant challenge is dataset selection as the number of data are deposited in the databases. Still, the challenge remains to extract the proper and experimented validated data.

3.1. Protein Sequence Dataset

As mentioned above, the section discussed "PTM identification approaches" that n-linked glycosylation prediction performance is based on peptide containing N-X-[S/T] motif and experimentally validated data, so the selection of data is based on these criteria. Based on the existing model's review and online dataset available, it has been noted that dbPTM [20] and nGlycositeAtlas [40] datasets comprise updated and experimented validated data that confirm the positive sites with N-X-[S/T] peptide. Moreover, UniProtKB[41] include recently added and modified sequence information. For UniProtKB filter was applied to extract only n-linked glycosylation sites, which include 1) Reviewed and human data, 2) PTM as Glycosylation with Keyword n-linked. The n-linked glycosylation sites for human proteins have been extracted from these datasets and processed to remove the duplication. UniProtKB contains the entire protein sequence, dbPTM with 21 window sizes and nGlycositeAtlas with 41 window sizes were combined to get 11254 unique proteins and 33859 unique glycosites. Data for further study was selected after removing the duplication, shown in Table 1. These number is sufficient to train the model with experimentally verified data.

AAindex datasets are used with these sequence datasets to obtain the numeric vector representation of each amino acid's physicochemical and biochemical properties. AAindex is divided into three parts: AAindex1, AAindex2, and AAindex3 [42]. As glycosylation is related to peptide binding, the AAindex1 dataset is suitable for measuring such properties.

3.2. Window Size

The whole protein sequence cannot be processed at a time to train the model; each protein sequence is divided into sequences the program compares at a time. A window is a fixed-size fragment with no fixed position over a protein sequence. A window size W is a fixed length Protein sequence. In this sequence, the centre of this fragment is the target residue; further, the W is always an odd number generated through (W - 1)/2. It is divided by 2 because it contains the same number of residues on either end. After selecting data, it is necessary to select the optimal window size for data. There is no fixed size for the window. Window size varies from 11 to 27 for machine learning and deep learning. Some studies selected window size w through a try-and-error approach. According to past studies, an accuracy drop with W > 10 and a wider window size may result in worse performance. [43, 55] In this study, Window size 21 is selected based on two criteria:

3.2.1. Protein Structure

The development of helical structure, coils and strands are affected by amino acids, which are 9, 3 and 6 positions away, respectively, in the sequence from the targeted residue.

3.2.2. Hydrophobicity

To determine the hydrophobic regions, window size 19 or 21 is optimal [43].

Table 1. Statistics of n-linked glycosylation dataset for human protein

Dataset	# of protein	# of glycosites	Link
UniProtKB	3330	8995	https://www.uniprot.org/
dbPTM	222	481	https://awi.cuhk.edu.cn/dbPTM/index.php
nGlycositeAtlas	9260	24382	http://nglycositeatlas.biomarkercenter.org/
Unique Data	**11254**	**33859**	

```
>Cluster 82
0       21aa, >P27909_183... *
1       21aa, >P33478_183... at 100.00%
2       21aa, >P27912_183... at 100.00%
3       21aa, >P27913_183... at 90.48%
4       21aa, >P17763_183... at 90.48%
>Cluster 83
0       21aa, >P05880_334... *
1       21aa, >P31872_338... at 100.00%
2       21aa, >P05878_338... at 95.24%
3       21aa, >Q73372_340... at 90.48%
4       21aa, >P20871_336... at 95.24%
```

Fig. 3 sample result of CD-HIT algorithm with threshold 0.9, representing the protein ID with n-link glycosylation location.

3.3. Data Pre-Processing

To improve the performance of protein sequence analysis and to reduce the sequence redundancy from the selected datasets, it is necessary to cluster the protein sequence and remove the duplicate or similar identity protein sequence according to the threshold. The most widely used tool CD-HIT (Cluster Database at High Identity with Database) [45], is fast and commonly used to reduce sequence redundancy using a greedy incremental clustering algorithm based on threshold similarity. If the two protein sequences have 85% similarity over 100-residue window size, they must have at least 55 identical tripeptides and 70 dipeptides. The input file for the CD-HIT tool will be in fasta format only. NGlycoGO[16], NetNGlyc[27], N-GlycDE[17], and DeepNGlycPred [18] use CD-HIT to reduce the protein sequence with 30% similarity. Another technique is BLASTCust (Blast package for clustering the protein sequence) based on pairwise similarity and thresholds, score identity and alignment length. SPRINT-Gly [33] uses this technique with more than 30% similarity. To cluster the used protein sequence data CD-HIT algorithm was applied with a 0.9 threshold in Fasta file format. In the .fasta file, ProteinID_Location (P28841_523) were used to identify the asparagine location with protein ID. A total of 25516 clusters were identified from 33859 glycosites. As a result, it was observed that similar protein sequences have nearby glycosylation locations. It is also noticed that there is a high possibility of modification at that specific location. The sample result is highlighted below in Figure 3, where two clusters, namely clusters 82 and 83, were shown. From these sample two clusters, it is noted that in sample cluster 82, P27909 protein ID has N residue at 183 locations, and similarly, other sequences in the same cluster have similar Asn locations. This result shows a high possibility of n-glycosylation at that specific location of that specific protein sequence pattern.

3.4. Protein Sequence Encoding Methods

Amino acid encoding methods play a vital role in the success of machine learning and deep learning models. [46] 20 letters amino acids represent protein sequence; it is necessary to process the amino acid sequence into binary representation for machine learning / deep learning to predict n-linked glycosylation. Most of the models use the encoding scheme which was developed in the pre-deep learning era, such as binary encoding (one-hot encoding), substitution matrices (Block Substitution Matrix (BLOSUM)), physicochemical character-based scheme (Principal component score Vector of Hydrophobic, Steric, and Electro properties (VHSE8)) [47]. Each encoding method for n-linked glycosylation prediction has been discussed below:

3.4.1. Binary Encoding

This method represents the amino acid in the protein sequence by 0s and 1s. There are three ways to encode amino acids 20-bit, 6-bit and 5-bit. One-hot encoding is the most common binary encoding method, which is also known as orthogonal encoding. [47] Each 20 amino acid is represented in this encoding by a binary vector. First, all the 20 amino acids are sorted in alphabetical order such as [A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y] to make process simple. It can be represented i^{th} amino acid type by 20 binary bits in which i^{th} bit is set to 1, one-hot code for A is 10000000000000000000 and so on. If any unknown amino acid is identified, then one more bit is needed to add in rare cases. To simplify this 20-dimension one-hot encoding to 6-dimension conservative replacement through evolution is used, based on the Point Access Mutation (PAM) metrics [48]. In this technique, encoding is based on the six groups of amino acids: [H, R, K], [D, E, N, Q], [C], [S, T, P, A, G], [M, I, L, V], and [F, Y, W]. The third low-dimension binary encoding scheme is 5-bit encoding [49]. There are 32 (25) possible ways to represent amino acids from which all zeros, all ones and those with 1 or 4 ones (5 + 5=10) are removed and exactly 20 representations remain. Amongst these three, 5-bit encoding representation is the excellent choice for model complexity, but this method is insufficient because only the presence of the N motif does not ratify n-linked glycosylation.

3.4.2. Substitution Matrix

It is an evolution-based encoding method that extracts evolutionary information of residue from sequence alignment, also known as multiple sequence alignment (MSA). This method is categorized based on the position of amino acids, which are 1) position independent – PAM (Point Accepted Mutation) matrices and BLOSUM matrices. PAM focuses on the evolutionary process of protein to identify the replacement probability of single amino acid with another amino acid. BLOSUM matrices are based on the conserved region of the non-redundant protein and determine the probabilities that amino acid pairs will interchange with each other. The score is log-odds that measure pairwise identity. 2) Position depends on which encode amino acid differently at different position regardless of the amino acid type are same. Position Specific Scoring

Matrix (PSSM) [50] is the most widely used encoding representing the log-likelihoods of the occurrence probabilities of all likely molecule types in each location in sequence. Also, PSSM achieves superlative performance on the large-scale dataset. [46] The score for residue i at position j ($Score_{ij}$) will be calculated using the below formula where (f'_{ij}) is the relative frequency for the residue i at position j and (q_i) is the expected relative frequency of residue i in a random position.

$$Score_{ij} = log\left(\frac{f'_{ij}}{q_i}\right)$$

Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) is used to generate Multiple Sequence Alignment (MSA) for the specific protein sequence; after that equivalent, PSSM is calculated from the MSA, for the sequence length L, PSSM matrix is L * 20 where each row representing log-likelihoods of 20 amino acids occur at specific position [52]. Hidden Markov Models (HMM) is another position-dependent encoding in which the HHbits alignment algorithm generates a probability profile that is more sensitive than the PSI-BLAST.

3.4.3. Physicochemical Properties Encoding

Physicochemical properties are also essential to capture the environment around the targeted glycosylation residue. Amongst all the physicochemical properties, hydrophobicity plays an essential role in shaping the self-assembly of protein. Matthew J. Betts and Robert B. Russell discussed Amino Acid properties classification. [50,51] Amino Acids are categorized according to their biochemical properties into five groups, which are mentioned in Table 2. Amino acid is classified as either hydrophobic or hydrophilic and is also called polar amino acid based on the amino acid side chain contact with a polar solvent like water. [CH Asp, Ser and Thr are examples of neutral polar amino acid which does not carry any charge. Author Ghazaleh Taherzadeh et al. stated that physicochemical properties give the highest performance on human and mouse n-linked datasets, followed by evolutionary information [32].

Table 2. Amino acid classification of chemical properties

Chemical Property	Amino Acids
Polar Hydrophilic	Glycine (G), Serine (S), Threonine (T), Asparagine (N), Cysteine (C), Glutamine (Q)
Acidic	Aspartic acid (D), Glutamic acid (E)
Basic	Lysine (K), Arginine (R), Histidine (H)
Hydrophobic	Alanine (A), Valine (V), Leucine (L), Isoleucine (I), Proline (P), Methionine (M)
Aromatic	Phenylalanine (F), Tyrosine (Y), Tryptophan (W)

Other physicochemical properties include the size of the amino acid, codon diversity of amino acid etc. VHSE8 is one of the best encoding methods for physicochemical properties based on the 18 Hydrophobic, 17 Steric and 15 electronic properties, so, in total, 50 physicochemical properties of 20 amino acids. Amino Acid property can be represented by VHSE-scale as follow: VHSE1 and VHSE2 (Hydrophobic properties), VHSE3 and VHSE4 (Steric properties), VHSE5 to VHSE8 (Electronic properties).[47]

To improve the quality of the prediction model, the selection of encoding methods plays an important role. Three encoding methods were discussed in the paragraphs above, each with significance. As focused on the protein sequence data, which are n-linked glycosylated, PSSM is the most suitable encoding for multiple sequence alignment of the protein sequence. [47] Moreover, with the statistical feature PSSM, the physicochemical property VHSE8 gives the best result for predicting n-linked glycosylation for human protein sequences. Amongst all these encoding methods, it has been noted that with the occurrence of a consensus sequence for n-linked glycosylation N-X-S/T, the accessibility of N residue in that consensus sequence is also equally important.

4. Conclusion

In conclusion, protein glycosylation is essential to the regulation of biological processes, and accurate prediction of n-linked glycosylation is mandatory to understand this biological process. The study was focused on biological and computation perspectives to identify the n-linked glycosylation for the human protein. As experimental techniques for identification are costly and time-consuming, there is the urge to develop models which can overcome these limitations. The research gap is highlighted and identified in the sequence-based prediction approach, which is more suitable in terms of accuracy as the query protein may not have a solved structure because the solved structures deposited are limited in the PDB database. Few models have been reviewed and concluded with the consensus sequence N-X-S/T not only the presence of Asn(N) residue but also accessibility of Asn(N) residue in protein sequence important in order to predict true positive sites. In addition, large and non-redundant databases have been created for n-linked glycosylation, which contains unique 11254 proteins and 33859 glysosites data from UniProtKB, nGlycositeAtlas, dbPTM. These data are updated and experimented with, and validated as well, which can improve the performance of a model to a great extent. Window size 21 was taken for protein sequence, which is optimal for n-linked glycosites prediction and the final dataset contains details - Protein ID, Location of N and Protein sequence with Window size 21. The CD-HIT algorithm was applied to remove the redundancy, through which 25516 clusters were found with a threshold value of 0.9. Upon analyzing the clusters, it was identified that more than 90% of similar protein sequences are near the glycosylation location for N residue. Also, the

listed and discussed various encoding methods for protein sequence have concluded that for n-linked glycosylation prediction, PSSM is the best technique for multiple sequence alignment as it focuses on the set of probability scores for each amino acid at each position of the alignment. With this PSSM, Physicochemical properties such as hydrophobicity are also important to improve the performance; this property can be calculated using the VHSE8 method, which is not used in the mentioned existing model. Amongst all the existing models, DeepNGlycPred and N-GlycDE evaluate performance by confirming N-X-S/T consensus sequence and accessibility of N residue using predicted structural features and PSSM encoding for multiple sequence alignment. To enhance the accuracy of the given model, physical-chemical properties such as hydrophobicity can be used.

References

[1] Kelley W. Moremen, Michael Tiemeyer, and Alison V. Nairn, "Vertebrate Protein Glycosylation: Diversity, Synthesis and Function," *Nature Reviews Molecular Cell Biology*, vol. 13, no. 7, pp. 448–462, 2012. *Crossref*, https://doi.org/10.1038/nrm3383

[2] Ząbczyńska M, and Pochec E., "The Role of Protein Glycosylation in Immune System," *Postepy Biochem*, vol. 61, no. 2, pp. 129-137, 2015.

[3] Varki A et al., editors.Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2009.

[4] John F. Rakus, and Lara K. Mahal, "New Technologies for Glycomic Analysis: Toward A Systematic Understanding of the Glycome," *Annual Review of Analytical Chemistry (Palo Alto Calif)*, pp. 367-92, 2011. *Crossref*, https://doi.org/10.1146/annurev-anchem-061010-113951

[5] Celso A Reis, Rudolf Tauber, and Véronique Blanchard, "Glycosylation is a Key in SARS-CoV-2 Infection," *Journal of Molecular Medicine*, vol. 99, no. 8, pp. 1023–1031, 2021. *Crossref*, https://doi.org/10.1007/s00109-021-02092-0

[6] Gerald W Hart, and Ronald J Copeland, "Glycomics Hits the Big Time," *Cell*, vol. 143, no. 5, pp. 672-676, 2010. *Crossref*, https://doi.org/10.1016/j.cell.2010.11.008

[7] Karin Julenius et al., "Prediction, Conservation Analysis, and Structural Characterization of Mammalian Mucin-Type O-Glycosylation Sites," *Glycobiology*, vol. 15, no. 2, pp. 153-164, 2005. *Crossref*, https://doi.org/10.1093/glycob/cwh151

[8] Radjiv Goulabchand et al., "Impact of Autoantibody Glycosylation in Autoimmune Diseases," *Autoimmunity Reviews*, vol. 13, no.7, pp. 742–750, 2014. *Crossref*, https://doi.org/10.1016/j.autrev.2014.02.005

[9] Manish Suyal, and Parul Goyal, "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms Based on Supervised Learning," *International Journal of Engineering Trends and Technology*, vol. 70, no. 7, pp. 43-48, 2022. *Crossref*, https://doi.org/10.14445/22315381/IJETT-V70I7P205

[10] Kai-Yao Huang et al., "dbPTM in 2019: Exploring Disease Association and Cross-Talk of Post-Translational Modifications," *Nucleic Acids Research*, vol. 47, no. D1, pp. D298-D308, 2019. *Crossref*, https://doi.org/10.1093/nar/gky1074

[11] Kazuaki Ohtsubo, and Jamey D Marth, "Glycosylation in Cellular Mechanisms of Health and Disease," *Cell*, vol. 126, no. 5, pp. 855-867, 2006. *Crossref*, https://doi.org/10.1016/j.cell.2006.08.019

[12] Nikolaj Blom et al., "Prediction of Post-Translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence," *Proteomics*, vol. 4, no. 6, pp. 1633-1649, 2004. *Crossref*, https://doi.org/10.1002/pmic.200300771

[13] Y Gavel, and G von Heijne, "Sequence Differences Between Glycosylated and Non-Glycosylated Asn-X-Thr/Ser Acceptor Sites: Implications for Protein Engineering," *Protein Engineering*, vol. 3, no. 5, pp. 433-442, 1990. *Crossref*, https://doi.org/10.1093/protein/3.5.433

[14] Birgit Eisenhaber, and Frank Eisenhaber, "Prediction of Post-Translational Modification of Proteins from their Amino Acid Sequence," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 609, pp. 365-384, 2010. *Crossref*, https://doi.org/10.1007/978-1-60327-241-4_21

[15] Manikandan Muthu et al., "Insights into Bioinformatic Applications for Glycosylation: Instigating an Awakening towards Applying Glycoinformatic Resources for Cancer Diagnosis and Therapy," *International Journal of Molecular Sciences*, vol. 21, no. 24, p. 9336, 2020. *Crossref*, https://doi.org/10.3390/ijms21249336

[16] Ching-Hsuan Chien et al., "N-GlycoGo: Predicting Protein N-Glycosylation Sites on Imbalanced Data Sets by Using Heterogeneous and Comprehensive Strategy," *IEEE Access*, vol. 8, pp. 165944-165950, 2020. *Crossref*, https://doi.org/10.1109/ACCESS.2020.3022629

[17] Thejkiran Pitti et al., "N-Glyde: A Two-Stage N-Linked Glycosylation Site Prediction Incorporating Gapped Dipeptides and Pattern-Based Encoding," *Scientific Reports*, vol. 9, no. 1, p. 15975, 2019. *Crossref*, https://doi.org/10.1038/s41598-019-52341-z

[18] Subash C. Pakhrin et al., "DeepNGlyPred: A Deep Neural Network-Based Approach for Human N-Linked Glycosylation Site Prediction," *Molecules*, vol. 26, no. 23, pp. 7314, 2021. *Crossref*, https://doi.org/10.3390/molecules26237314

[19] Tian Jipeng, Suma P, and Dr. T.C.Manjunath, "AI, ML and the Eye Disease Detection," *SSRG International Journal of Computer Science and Engineering*, vol. 7, no. 4, pp. 1-3, 2020. *Crossref*, https://doi.org/10.14445/23488387/IJCSE-V7I4P101

[20] Pablo Minguez et al., "PTMcode: A Database of Known and Predicted Functional Associations Between Post-Translational Modifications in Proteins," *Nucleic Acids Research*, vol. 41, pp. 306-311, 2013. *Crossref*, https://doi.org/10.1093/nar/gks1230

[21] Zhongyan Li et al., "dbptm in 2022: An Updated Database for Exploring Regulatory Networks And Functional Associations of Protein Post-Translational Modifications," *Nucleic Acids Research*, vol. 50, no. D1, pp. 471–479, 2022. *Crossref*, https://doi.org/10.1093/nar/gkab1017

[22] Bingjie Xue et al., "KinPred: A Unified and Sustainable Approach for Harnessing Proteome-Level Human Kinase-Substrate Predictions," *PLoS Computational Biology*, vol. 17, no. 2, 2021. *Crossref*, https://doi.org/10.1371/journal.pcbi.1008681

[23] Alex S Holehouse, and Kristen M Naegle, "Reproducible Analysis of Post-Translational Modifications in Proteomes--Application to Human Mutations," *PLoS One*, vol. 10, no. 12, 2015. *Crossref*, https://doi.org/10.1371/journal.pone.0144692

[24] Sachin Gavali et al., "RESTful API for iPTMnet: A Resource for Protein Post-Translational Modification Network Discovery," *Database: The journal of Biological Databases and Curatio*, vol. 2020, 2020. *Crossref*, https://doi.org/10.1093/database/baz157

[25] Dan Ofer, Nadav Brandes, and Michal Linial., "The Language of Proteins: NLP, Machine Learning & Protein Sequences," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1750-1758, 2021. *Crossref*, https://doi.org/10.1016/j.csbj.2021.03.022

[26] Mihaly Varadi et al., "AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models," *Nucleic Acids Research*, vol. 50, no. D1, pp. 439–444, 2022. *Crossref*, https://doi.org/10.1093/nar/gkab1061

[27] Gupta R, and Brunak S., "Prediction of Glycosylation Across the Human Proteome and the Correlation to Protein Function," *Pacific Symposium on Biocomputing, Pacific Symposium on Biocomputing*, pp. 310-322, 2002.

[28] Stephen E Hamby, and Jonathan D Hirst, "Prediction of Glycosylation Sites Using Random Forests," *BMC Bioinformatics*, vol. 9, p. 500, 2008. *Crossref*, https://doi.org/10.1186/1471-2105-9-500

[29] Cornelia Caragea et al., "Glycosylation Site Prediction Using Ensembles of Support Vector Machine Classifiers," *BMC Bioinformatics*, vol. 8, pp. 438, 2007. *Crossref*, https://doi.org/10.1186/1471-2105-8-438

[30] Chauhan JS et al., "GlycoPP: A Web Server for Prediction of N- and O-Glycosites in Prokaryotic Protein Sequences," *PLoS One*, vol. 7, no. 7, 2012.

[31] Jagat Singh Chauhan, Alka Rao, and Gajendra P. S. Raghava, "In Silico Platform for the Prediction of N-, O- and C-Glycosites in Eukaryotic Protein Sequences," *Plos One*, vol. 8, 2013. *Crossref*, https://doi.org/10.1371/journal.pone.0067008

[32] Fuyi Li et al., "Glycomine: A Machine Learning-Based Approach for Predicting N-, C- and O-Linked Glycosylation in the Human Proteome," *Bioinformatics*, vol. 31, no. 9, pp. 1411–1419, 2015. *Crossref*, https://doi.org/10.1093/bioinformatics/btu852

[33] Ghazaleh Taherzadeh et al., "SPRINT-Gly: Predicting N- and O-Linked Glycosylation Sites of Human and Mouse Proteins by Using Sequence and Predicted Structural Properties," *Bioinformatics*, vol. 35, no. 20, pp. 4140-4146, 2019. *Crossref*, https://doi.org/10.1093/bioinformatics/btz215

[34] Kolapo Adetomiwa, "Adoption And Utilization of Artificial Intelligence (Ai) In Poultry Production: Evidence From Smart Agricultural Practices in Nigeria," *SSRG International Journal of Agriculture & Environmental Science*, vol. 7, no. 3, pp. 46-54, 2020. *Crossref*, https://doi.org/10.14445/23942568/IJAES-V7I3P106

[35] Fuyi Li et al., "GlycoMine(struct): A New Bioinformatics Tool for Highly Accurate Mapping of the Human N-Linked and O-Linked Glycoproteomes by Incorporating Structural Features," *Scientific Reports*, vol. 6, 2016. *Crossref*, https://doi.org/10.1038/srep34595

[36] Benjamin Luke Schulz, "*Beyond the Sequon: Sites of N-Glycosylation*," Glycosylation, Petrescu, S., Ed., InTech: Rijeka, Croatia, pp. 21–40, 2012. *Crossref*, https://doi.org/10.5772/50260

[37] Mihai Nita-Lazar et al., "The N-X-S/T Consensus Sequence is Required But not Sufficient for Bacterial N-Linked Protein Glycosylation," *Glycobiology*, vol. 15, no. 4, pp. 361–367, 2005. *Crossref*, https://doi.org/10.1093/glycob/cwi019

[38] Mubina Malik, and Jaimin N Undavia, "Trials, Skills, and Future Standpoints of AI-Based Research in Bioinformatics," *International Journal of Recent Technology and Engineering*, vol. 9, no. 1, pp. 968–972, 2020. *Crossref*, https://doi.org/10.35940/ijrte.A1920.059120

[39] Alhasan Alkuhlani et al., "Intelligent Techniques Analysis for Glycosylation Site Prediction," *Current Bioinformatics*, vol. 16, no. 6, pp. 774-788, 2021. *Crossref*, https://doi.org/10.2174/1574893615666210108094847

[40] Shisheng Sun et al., "N-GlycositeAtlas: A Database Resource for Mass Spectrometry-Based Human N-Linked Glycoprotein and Glycosylation Site Mapping," *Clinical Proteomics*, vol. 16, no. 35, pp. 1-11, 2019. *Crossref*, https://doi.org/10.1186/s12014-019-9254-0

[41] The UniProt Consortium, "UniProt: The Universal Protein Knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, 2021. *Crossref*, https://doi.org/10.1093/nar/gkaa1100

[42] Shuichi Kawashima, and Minoru Kanehisa, "Aaindex: Amino Acid Index Database," *Nucleic Acids Research*, vol. 27, no. 1, pp. 368-369, 1999. *Crossref*, https://doi.org/10.1093/nar/27.1.368

[43] Ke Chen, Lukasz Kurgan, and Jishou Ruan, "Optimization of the Sliding Window Size for Protein Structure Prediction," *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pp. 1-7, 2006. *Crossref*, https://doi.org/10.1109/CIBCB.2006.330959

[44] Vedant Bhatt, and Mohammad Makki, "Artificial Intelligence for Curing Skin Disorders," *SSRG International Journal of Computer Science and Engineering*, vol. 5, no. 10, pp. 7-9, 2018. *Crossref*, https://doi.org/10.14445/23488387/IJCSE-V5I10P103

[45] Limin Fu et al., "CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data," *Bioinformatics*, vol. 28, no. 23, pp. 3150-3152, 2012. *Crossref*, https://doi.org/10.1093/bioinformatics/bts565

[46] Xiaoyang Jing et al., "Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 1918–1931, 2020. *Crossref*, https://doi.org/10.1109/TCBB.2019.2911677

[47] Hesham ElAbd et al., "Amino Acid Encoding for Deep Learning Applications," *BMC Bioinformatics*, vol. 21, no. 235, pp. 1-14, 2020. *Crossref*, https://doi.org/10.1186/s12859-020-03546-x

[48] J. T. L. Wang et al., "New Techniques for Extracting Features from Protein Sequences," *IBM Systems Journal*, vol. 40, no. 2, pp. 426–441, 2001. *Crossref*, https://doi.org/10.1147/sj.402.0426

[49] Gilbert White, and William Seffens, "Using a Neural Network to Back Translate Amino Acid Sequences," *Electronic Journal of Biotechnoloy*, vol. 1, no. 3, pp. 17–18, 1998.

[50] Michael Beckstette et al., "Fast Index Based Algorithms and Software for Matching Position-Specific Scoring Matrices," *BMC Bioinformatics*, vol. 7, no. 389, 2006. *Crossref*, https://doi.org/10.1186/1471-2105-7-389

[51] Matthew J. Betts, and Robert B. Russell, "Amino Acid Properties and Consequences of Substitutions," *Bioinformatics for Geneticists*, vol. 317, no. 289, 2003. *Crossref*, https://doi.org/10.1002/0470867302.ch14

[52] Stephen F. Altschul et al., "Gapped BLAST And PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997. *Crossref*, https://doi.org/10.1093/nar/25.17.3389

[53] Pablo Minguez et al., "PTMcode v2: A Resource for Functional Associations of Post-Translational Modifications within and Between Proteins," *Nucleic Acids Research*, vol. 43, pp. 494-502, 2015. *Crossref*, https://doi.org/10.1093/nar/gku1081

[54] Gwo-Yu Chuang et al., "Computational Prediction of N-Linked Glycosylation Incorporating Structural Properties and Patterns," *Bioinformatics*, vol. 28, no, 17, pp. 2249–2255, 2012. *Crossref*, https://doi.org/10.1093/bioinformatics/bts426

[55] Ying Xu et al., "Phoscontext2vec: A Distributed Representation of Residue-Level Sequence Contexts and its Application to General and Kinase-Specific Phosphorylation Site Prediction," *Scientific Reports*, vol. 8, p. 8240, 2018. *Crossref*, https://doi.org/10.1038/s41598-018-26392-7