Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2016, 8(12):52-63

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A Study on the Presence and Distribution of Carbon Percentage in and Around the Sites of Glycosylation for Eukaryotic Proteins

Parul Johri[1*], Mala Trivedi[1], Mohammed Haris Siddiqui[2] and Mrunal Gokhale[3]

[1]Amity Institute of Biotechnology, Amity University, Lucknow, Uttar Pradesh, India
[2]Department of Bioengineering, Integral University Lucknow Uttar Pradesh, India
[3]Kiran IPR Trainee, Mumbai, Maharashtra, India

ABSTRACT

Carbon is a ubiquitous element in all organic entities. Its presence in various biomolecules is of importance with regards to their structure and function. For amino acids, carbon is the only element which contributes to their hydrophobicity. Proteins necessitate presence of 31.44% carbon and approximately 27% large hydrophobic residues for structural stability. The structural accuracy of proteins is determined by hydrophobicity of its constituent residues. Similarly, post translational events also affect structural and functional aspects of proteins. Glycosylation is a universal post translational modification. It is known to influence protein folding, stability/turnover, solubility, sub-cellular localization and recognition events. Carbon being the sole contributor towards hydrophobicity, its presence is vital towards the precision of the protein function. Understanding this conception with profound study, may lead to a new dimension of protein sequence analysis. In this research we have analyzed presence and distribution of carbon percentage in and around sites of glycosylation in eukaryotic proteins. The study revealed that the site of glycosylation was flanked by sites possessing steep rise and steep dip in carbon percentage. So we may mark potential site of glycosylation by observing steep rise or dip in carbon percentage of the sequence. The site of glycosylation was seen to achieve a mean carbon percentage which was nearing mean carbon percentage for the total sequence. Thus, it was also observed that amount of carbon was being stabilized at the site of glycosylation.

Keywords: Carbon; Clustering; Eukaryotes; Glycosylation; Matlab and protein

---

INTRODUCTION

Proteins are the most abundant biological macromolecules in cells. They are present in every living system ranging from unicellular organisms to vertebrates and complex eukaryotes. They also occur in great variety; thousands of different kinds, ranging in size from relatively small peptides to huge polymers with molecular weights in millions, may be found in a single cell. They are the molecular instruments through which genetic information is expressed. Proteins make almost 50% of the dry weight of the cells and are present in profound amount, then any other bio-molecules. They have series of biological functions like DNA replication, forming cytoskeleton structures, intra and extracellular signaling and many more. Biological complications may arise due to absence or dysfunction of proteins, such as phenylketonuria (loss of phenylalanine hydroxylase). Similarly accumulation of proteins in excess amount may also be hazardous, such as Alzheimer disease (accumulation of senile plaques or neurofibrillary tangle) [1,2]. Proteins are basically organic compounds composed of amino acids arranged in linear chain and folded to a globular form. These building blocks of protein are joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. Amino acids are indeed very simple compounds. All 20 amino acids consist of combinations of only five different atoms. The atoms which can be found in these structures are; Carbon (C), Nitrogen (N), Hydrogen (H), Sulfur (S) and Oxygen (O). Amino acids are used in most body processes from

regulating the way the body works to how the brain functions. They activate and utilize vitamins and other nutrients. The liver manufactures about 80% of these amino acids, but the remaining 20% of such amino acids must be supplied directly by diet, and these amino acids are referred to as the essential amino acids. Depending on the properties of their side chains, the amino acids are classified as hydrophobic and hydrophilic. If the side chain of an amino acid is polar, then it is hydrophilic and if the side chain is non-polar it is classified as hydrophobic. The hydrophobic residues are found buried deep in the inner core of the protein whereas the hydrophilic residues are found in the outer soluble environment (can react with water). The distribution of hydrophobic residues in a protein contributes majorly towards protein folding, protein interactions, formation of core, active side formation and other biological functions [3]. The atomic composition of amino acids is responsible for the properties of hydrophobicity or hydrophilicity, which play an important role in protein interactions.

As carbon is the main element which contributes to hydrophobic interactions in proteins, the hydrophobic amino acids characteristically have greater number of carbon atoms. Each amino acid is assigned with a hydropathy index based on the hydrophobic/hydrophilic properties of its side chains. The `hydropathy index` of a protein is a number representing its hydrophilic or hydrophobic properties. The higher the index is, the more hydrophobic the amino acid (Table 1). The structure and activity of proteins are contributed by the presence of Large Hydrophobic residues (LHR) such as phenylalanine (F), isoleucine (I), leucine (L), methionine (M), and asparagines (V). It has been observed that proteins need 31.44% of total carbon content for their structure stability and activity [4].

Table 1: Hydropathy index of amino acids

Amino Acid Index	Amino Acid																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Hydropathy Index	1.8	-4.5	-3.5	-3.5	2.5	-3.5	-3.5	-0.4	-3.2	4.5	3.8	-3.9	1.9	2.8	-1.6	-0.8	-0.7	-0.9	-1.3	4.2

Glycosylation is a post translational modification. It is an enzymatic process that links saccharides to produce glycans, either free or attached to proteins or lipids. This enzymatic process produces one of four fundamental components of all cells (along with nucleic acids, proteins, and lipids) and also provides a co-translational and post-translational modification mechanism that modulates the structure and function of membrane and secreted proteins. The majority of proteins synthesized in the rough endoplasmic reticulum undergo glycosylation. It is an enzyme-directed site-specific process [5-6]. The polysaccharide chains attached to the target proteins serve various functions. It is believed to occur in more than half of the eukaryotic proteins and has been shown to affect protein folding, stability/turnover, solubility, subcellular localization and recognition events. Glycosylation also play a role in cell-cell adhesion (a mechanism employed by the immune system) as well. Experiments have shown that polysaccharides linked at the amide nitrogen of asparagines in the protein confer stability by avoiding premature degradation. It was once believed that glycoprotein were confined to eukaryotic organisms, but it is now widely accepted that post translational glycosylation is a universal modification [7-9].

MATERIALS AND METHOD

In the present work we initially tried to identify the role of carbon atom in protein sequences. In order to extend this logic we retrieved glycosylated protein sequences for eukaryotes from Swiss Prot database (http://ca.expasy.org/sprot/) irrespective of the type of glycosylation (N linked and O linked). The data set is majorly classified as N linked sequences and O linked sequences. The further sub classification is based on the category of glycosylation sites present in the sequence. The sequence which possesses at least one experimentally proved glycosylation site was marked as reference. The sequence with no reference site but had potential sites for glycosylation was marked as potential. In order to complete the total carbon percentage of the sequence, we had written a PERL (v5.10.1) script using dynamic programming approach. The data was subjected to dynamic programming because the normal approach could not entertain such large such large sequences. The script basically splits the amino acids peptide chain into its constituent atomic array. A window of 10% of the total length of atomic array was then slide over, for calculating the carbon percentage of each window. The frequency of windows for each carbon percentage were identified and divided by total number of windows to provide weights. Further for the analysis, the frequencies of weighted values were plotted against the carbon percentage for each sequence. Finally, a combined analysis of all the proteins retrieved was done taking average of the weighted values and plotting graph for the same. Every sequence from the data set was processed similarly. For each graph generated, we did calculate the skewness of graph, mean frequency, carbon percentage with highest frequency, highest carbon percentage and lowest carbon percentage. This information is enlisted below (Table 2):

Table 2: Assessment of graph for each sequence from the dataset

Detailed Analysis of N - Linked (Potential) Glycosylation Sequences						
Sequence ID	**Index Number**	**Highest Carbon Content**	**Lowest Carbon Content**	**Maximum Frequency for**	**Skewness**	**Mean Frequency**
C8Z742	1	35.6617	30.147	32.3529	0.500792	32.45135
C8Z830	2	35.0378	30.4924	32.1969	0.131624	32.60909
C8ZCR2	3	33.32	30.4035	32.441	-0.346	31.98908
C8ZGM3	4	34.507	26.7605	30.9859	-0.11829	30.71211
C8ZI10	5	35.1307	30.0653	33.33	0.145325	32.91805
O76942	6	32.7239	29.616	30.5301	0.418627	30.95772
O96363	7	34.5945	30.1351	32.5675	0.479023	32.3224
P00635	8	35.0828	30.8011	32.7348	0.312506	32.61332
P14000	9	34.4457	30.3933	31.9427	0.268003	32.31426
P14217	10	34.4965	30.1096	31.6051	0.267089	32.23132
P15589	11	34.8623	30.5045	31.422	0.180043	32.31718
P22146	12	34.1523	28.1326	31.8181	0.667121	31.86643
P22282	13	35.7142	28.9115	31.6326	-0.11813	32.56905
P24031	14	34.3922	30.6629	32.5966	0.037051	32.6113
P25353	15	34.5329	29.82	32.6478	1.003844	32.65151
P25381	16	35.4166	29.8177	32.4218	0.46072	32.19942
P25574	17	34.1205	30.9446	32.0032	0.702145	32.18769
P25653	18	32.1148	29.8085	30.4177	-0.32334	30.77251
P27810	19	36.4779	30.6603	33.1761	0.610414	33.33188
P31382	20	34.9009	30.363	33.4983	-0.08928	33.06246
P32768	21	33.5881	29.7212	30.8003	2.352384	31.04555
P33727	22	34.4497	30.7416	32.177	0.274893	32.41241
P34231	23	33.7696	29.5811	30.8027	0.530328	31.5102
P36170	24	33.7748	29.5605	30.5237	1.010408	31.1136
P37370	25	32.363	29.7089	31.4212	0.580703	31.27161
P38138	26	34.3953	30.6241	32.8348	1.201029	32.73974
P38693	27	34.2975	30.5785	32.5068	0.333326	32.43656
P38694	28	34.2657	30.0699	32.4675	0.079212	31.96465
P38887	29	33.7487	32.8411	32.3987	0.601112	32.196
P39105	30	34.3117	30.0607	32.7935	0.3386227	32.08098
P40095	31	32.7272	29.2045	30.6818	0.0068787	30.88606
P40520	32	37.912	28.0219	31.3186	0.5992487	31.73736
P40533	33	35.2242	30.6068	32.9815	0.1635073	32.748772
P40583	34	34.8258	30.597	32.9601	-0.2906713	32.44499
P43600	35	36.3896	28.3667	31.5186	-0.036357	32.28878
P47055	36	36.0563	27.8873	30.7042	0.5231298	31.21982
P47078	37	38.764	27.528	32.5842	0.4734755	32.81667
P48016	38	33.9098	30.9748	31.5513	-0.0299615	32.24365
P50079	39	33.2406	29.4853	31.4325	0.6265319	31.54421
P50426	40	34.7079	30.0114	31.6151	0.3756853	32.21578
P50427	41	34.7357	30.4207	31.6073	2.0021139	31.90992
P50428	42	34.4459	30.04	31.7757	0.0710146	32.22757
P50429	43	34.3337	30.4921	32.0528	0.0509701	32.36188
P50473	44	34.0723	29.9883	31.8553	0.4633556	32.17971
P52290	45	34.3324	30.1089	32.8337	0.3903559	32.47278
P53058	46	34.9673	29.4117	32.6797	0.5138161	32.27547
P53163	47	32.8859	28.5234	30.8724	0.3029463	30.77471
P53334	48	34.657	28.3393	32.1299	0.2615403	31.3574
P53950	49	32.8409	29.0909	30.6818	1.5776942	30.94237
P54793	50	35.2173	29.5652	32.1739	-0.1438904	32.32365
Q01974	51	33.8376	30.2613	31.7744	0.9065309	31.83736
Q02783	52	33.6795	29.2284	31.1572	0.8156196	31.15089

DETAILED ANALYSIS OF N - LINKED (Reference) GLYCOSYLATION SEQUENCES						
Sequence ID	**Index Number**	**Highest Carbon Content**	**Lowest Carbon Content**	**Maximum Frequency**	**Skewness**	**Mean**
O00391	1	34.0223	30.2325	32.8165	0.039301	32.043421
P08842	2	35.262	30.2401	32.4235	0.3595154	32.50716
P15289	3	34.8525	30.0268	31.6353	0.0121039	32.21312
P15586	4	34.7575	30.1385	31.4087	-0.0674714	32.22792
P15848	5	34.5693	30.7416	32.177	0.4927146	32.36052
P22304	6	35.0348	30.1624	32.4825	0.8441451	32.59589
P34059	7	35.1485	30.4455	32.0544	0.700144	32.55619
P51688	8	34.6835	30.3797	32.7848	-0.131493	32.5181
P51690	9	33.769	29.8474	32.6797	0.8931241	32.30305
Q3T906	10	33.2009	30.2729	31.4143	-0.1754528	31.79637
Q8C145	11	34.1687	29.5739	32.5814	1.0458064	31.88902
Q8IWU5	12	33.6909	30.1144	31.0443	0.066035	31.83778
Q8IWU6	13	34.0213	29.8932	31.032	0.8958564	31.73363
Q9VEX0	14	33.6919	30.0679	30.804	0.7999388	31.44918
Comparitive Analysis		35.262	29.5739	32.6797	0.923271	32.143743

Detailed Analysis of O - Linked (Potential) Glycosylation Sequences						
Sequence ID	**Index Number**	**Highest Carbon Content**	**Lowest Carbon Content**	**Maximum Frequency for**	**Skewness**	**Mean Frequency**
O30620	1	34.2794	28.8209	32.5327	0.8959233	32.08664
P11764	2	35.5731	27.6679	31.6205	0.2498357	31.24258
P32781	3	38.6363	26.5151	31.0606	0.4694798	31.7998
P80069	4	34.2794	28.8209	32.5327	0.8959233	32.08664
Comparative Analysis		38.6363	26.5151	31.0606	0.6693371	31.70967

Detailed Analysis of O - Linked (Reference) Glycosylation Sequences						
Sequence ID	**Index Number**	**Highest Carbon Content**	**Lowest Carbon Content**	**Maximum Frequency for**	**Skewness**	**Mean Frequency**
P02974	1	34.7826	28.0632	30.83	0.3375359	31.37345
P05431	2	35.2941	27.8431	29.8039	0.1347332	31.19574
P18774	3	34.632	27.7056	29.8701	0.1570066	30.95331
P36912	4	35.8381	29.287	33.1406	0.493275	32.61944
P36913	5	35.4	29	31.8	0.6431316	32.02278
Q05819	6	35.0162	30.1302	32.4104	0.3683436	32.46944
Q50906	7	34.4978	28.8209	32.5327	1.0459726	32.15161
Comparitive Analysis		35.8318	27.7056	29.8701	0.7242241	31.8083

Detailed Analysis of O - Linked (Reference) Glycosylation Sequences						
Sequence ID	**Index Number**	**Highest Carbon Content**	**Lowest Carbon Content**	**Maximum Frequency for**	**Skewness**	**Mean Frequency**
P47001	1	35.625	27.5	29.375	0.2337244	31.21442
Q12127	2	35.5191	27.3224	30.0546	1.0074049	30.71918
Comparitive Analysis		35.5191	27.3224	30.0546	1.7410546	30.966803

To further analyze the role of carbon content in species specific manner, we did clustering in XLMineXLMiner (Version3) is acomprehensive data mining add-in for Excel. Data mining is a discovery-driven data analysis technology used for identifying patterns and relationships in data sets. Data mining is still a nascent field, and is a convergence of fields like statistics, machine learning, and artificial intelligence. Often, there may be more than one approach to a problem. XLMiner is a toolbelt to get quickly started on data mining, offering a variety of methods to analyze your data. It has extensive coverage of statistical and machine learning techniques for classification, prediction, affinity analysis and data exploration and reduction.

The reference sequences of N linked and O linked dataset were subjected to clustering. Input provided for XLMiner was of skewness and mean frequency of these sequences. Frequency is defined as the ratio of the number of observations in a statistical category to the total number of observations. Frequency of a data set is a vital parameter in statistics because it signifies the impact of a quantity or data point in a normally distributed dataset. Since, the focus of our experiment was to analyze the significance of carbon content in proteins, the calculation of frequency is an important step in achieving potential results. Skewness is defined as the measure of asymmetry of a distribution (mathworld.wolfram.com). Skewness can come in form of “negative skewness” or “positive skewness”, depending

on whether data points are skewed to the left (negative skew) or to the right (positive skew) of the data average. The knowledge of skewness is important in terms of concluding the characteristics of a dataset. In case of carbon percentage, if the graph is right skewed, it is evident that the particular that majority of the carbon percentages lie towards the left of the mean. Similarly the analysis stands for the left skewed data.

Agglomerative hierarchical clustering was performed. The distance matrix was generated using Euclidean distance. The clustering method used was 'Complete linkage' (Figure 1).

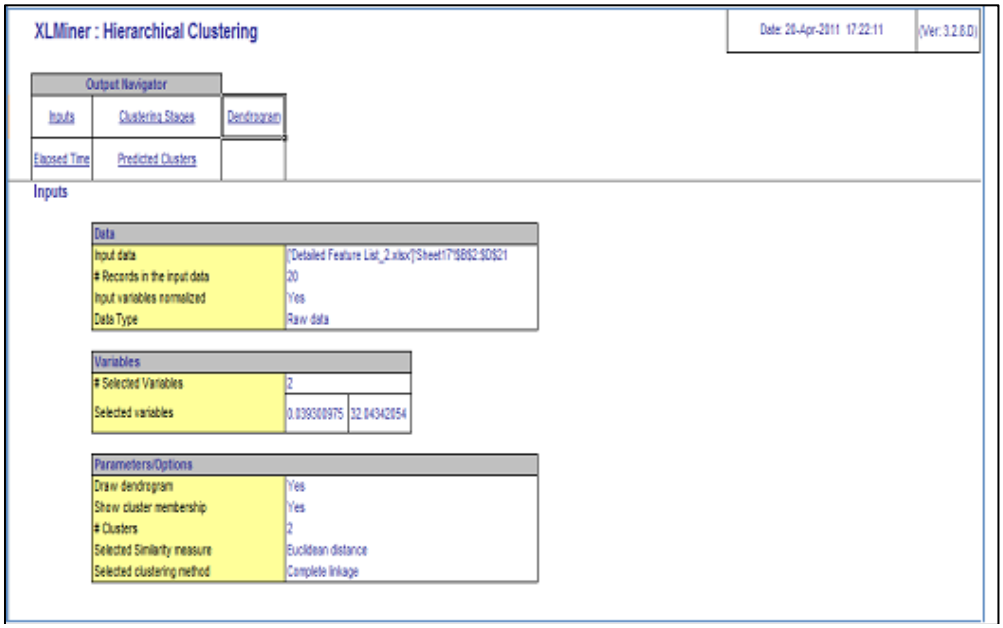

Figure 1: Data Analysis Uisng XLMiner Software

The next step of analysis was done in MATLAB (Version 7.11.0.584(R2010b); License Number: 161051). The input provided was mean frequency and skewness of N linked reference sequences. MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. It can be used in a wide range of applications, including signal and image processing, communications, control design, test and measurement, financial modeling and analysis, and computational biology. MATLAB provides a number of features for documenting and sharing work. It can be integrated with other languages and applications.

Agglomerative hierarchical clustering for the data was performed in MATLAB. Procedure for MATLAB:

- The input for MATLAB was provided from the excel data sheets. The parameters input were skewness and mean frequency.
- This data was assigned to a variable.
- The next step was to calculate the distance matrix. It was calculated using Euclidean distance. Function pdist() was used for the same.
- After generation of distance matrix, the creation of agglomerative hierarchical clustering tree using function linkage () was performed. The linkage generated provides link between objects and clusters.
- This linkage generated was provided as input for generation of dendrogram using dendrogram() function.
- The last step was the calculation of co-phenetic correlation coefficient (c). The co- phenetic correlation for a cluster tree is defined as the linear correlation coefficient between cophenetic distances obtained from the tree, and the original distances used to construct the tree. Thus, it is a measure of how faithfully the tree represents the dissimilarities among observation.

The results for the first phase of analysis were emphasizing on the carbon percentage in protein atomic array. The next phase was designed to deal with specific sites of glycosylation. Here we report the pattern of carbon present in

and around the sites of glycosylation. Glycosylation sites present in each sequence were marked and the distribution of carbon percentage, along with the region flanking the site was observed for each sequence.
This phase is specifically designed for glycosylated sequences. A PERL script was written for the same based on dynamic programming approach. The text file containing the sequence data must contain the locations for sites of glycosylation. The script splits the protein amino acid peptide chain into its constituent atomic array. Here, we consider the atomic content only for side chain of amino acids because the backbone for all the amino acids is constant. So we do not consider the atomic content for backbone. A window taking 2% of the total length of atomic array, with an increment value of 10 atoms, was then slide over for calculation of counts of carbon, nitrogen and sulfur. The carbon percentage was calculated for each window. For analysis purpose, carbon percentage was plot against window number on a graph. Every sequence was analyzed in the same manner and individual graph was plotted for each sequence.
The analysis for sequence ID O00754 (N – linked potential) from *Homo sapiens* is provided. Here we have marked the sites of glycosylation occurring throughout the sequence. We do calculate the mean and median for sites of same length upstream and downstream of the site of glycosylation.

ALGORITHM

The first aim of this project was to study prominence of carbon content in protein sequences. The second aim, after analysis of carbon content and its distribution in protein sequence, was to study the prominence of carbon content at site of glycosylation for sequences undergoing post translational event of glycosylation. The arrangement of carbon content in and around the site of glycosylation was to be analyzed. To understand this it was first necessary to understand if carbon distribution did have any particular pattern or prominent arrangement in protein sequences. With this in view we divided the project in two phases. The first phase was designed to analyze protein atomic carbon content and its distribution. The results generated by this method were expected to through light on the pattern and percentage of carbon atoms in the whole protein sequence. The second phase was designed specifically for protein sequences undergoing glycosylation as post translational modification. The results of this phase were expected to reveal the arrangement of carbon content in and around all the glycosylation sites present in the sequence. Both the programs were written using PERL language (v5.10.1).

Approximately 160 sequences were analyzed using these programs. The sequences for eukaryotes undergoing glycosylation were selected from Swiss prot database and were downloaded in FASTA format. The text files of these were used as input and the sequences were subjected to carbon amount calculation using dynamic programming. Dynamic programming is a method used in solving complex problems by breaking them down into simpler steps. This approach is optimized to solve complicated big problems by breaking them down into smaller sub problems in a recursive fashion. Another feature of problems which are solved using dynamic programming is overlapping sub problems. This means that the space of sub problems should solve the same sub problem over and over, rather than generating new sub problems. It is an approach which provides relative results to a problem in an absolute manner rather than the direct approach. Dynamic programming was used as it is a more absolute approach as compared to the direct approach. The results generated through this method could be subjected to various statistical formulae and analysis to get relative results when considering large datasets. The dataset used to conduct all the experiments involved over 160 protein sequences, hence the program used required relative and absolute approaches in order to yield results which could be used to generate a commendable hypothesis. Hence dynamic programming was deemed more suitable.

Phase 1: Calculation of carbon content in protein using dynamic programming
The phase one program was aimed at evaluation of atomic carbon content in the protein sequence. This program can be used on any type of proteins written in FASTA format.
The algorithm for the first program is as follows:

- The program starts by accepting the text file of sequence provided by the user. This sequence is in FASTA format.
- The next step is parsing of this sequence. Here, the amino acid sequence of protein is extracted from rest of the annotations.
- The length of amino acid sequence is calculated.

- Every amino acid is replaced with the corresponding atomic sequence. The amino acid sequence is converted into its atomic array sequence.
- The formation of polypeptides is due to peptide bond formation. This bond is formed by removal of water molecule between two adjacent amino acids. Hence, one water molecule is removed from between two adjacent amino acids atomic content. This leads to the formation of the actual atomic level polypeptide sequence.
- The next step is the analysis of the atomic content of the sequence and calculation of its length.
- As said earlier, dynamic programming does break down a problem into smaller sub problems. The sub problems are of equal magnitude. Following this, a window of 10% of the total length of atomic array is slide over, for calculating the carbon count of each window. The window frame moves with a distance of one atom at a time.
- The carbon count is used for the calculation of carbon percentage in each window.
- The frequency of windows for each carbon percentage were identified and divided by total number of windows to provide weights.
- Further for the analysis, the frequencies of weighted values were plotted against the carbon percentage for each sequence.

Second phase does deal with the sequences which are of specific nature. These proteins do under go glycosylation as post translational modification. Our interest is in analyzing the sites of glycosylation and the region flanking them. The second program written does analysis of these sites.
The algorithm for the second program is as follows:

- The sequences used as input for this program are glycosylated and the locations for sites for glycosylation are provided in Swiss prot database.
- The text file with FASTA sequence is to be input in the program. But before inputting the sequence the user has to add the sites of glycosylation in the file as comment at the beginning.
- Parsing of sequence takes place.
- The program does calculate amino acid sequence length.
- Further, the amino acids are substituted with their atomic content.
- Here, we consider only the side chain atoms of the amino acids and not the whole atomic content with backbone and side chain. The reason to take only side chain atoms is because the glycosylation event does take place on the side chain only. The peptide bond formation is hence not taken into account.
- The length of the atomic sequences is calculated.
- Dynamic programming approach is used for this program. Window of length 2% of the total length of sequence is slide over the entire sequence. It moves with an increment of 10 atoms.
- The glycosylation sites provided at the beginning of the sequences are noted and the amino acids are found. Further the atom to which the glycosylation takes place is marked.
- The windows, in which the glycosylated atom is present, are marked.
- We further calculate carbon count of each window.
- Using the carbon count, carbon percentage for each window is calculated.

The next step was to analyze carbon percentage. For this, the carbon percentage was plotted in graph against the window number. Various aspects of carbon percentage like mean, median, highest and lowest points of carbon percentage were calculated. The windows containing the glycosylation site were encircled in the graph. The carbon percentage is plotted against window number. The sites of glycosylation are encircled in the graph.

RESULTS

The entire project was being divided into two phases. In the first half, we tried to find out importance of carbon element in eukaryotic glycosylated protein sequences. The carbon content for N linked glycosylated proteins was calculated using dynamic programming code. The frequency of carbon percentage for all the N linked glycosylated proteins was calculated, tabulated (Table 1) and subjected to frequency distribution plot (Figure 2).

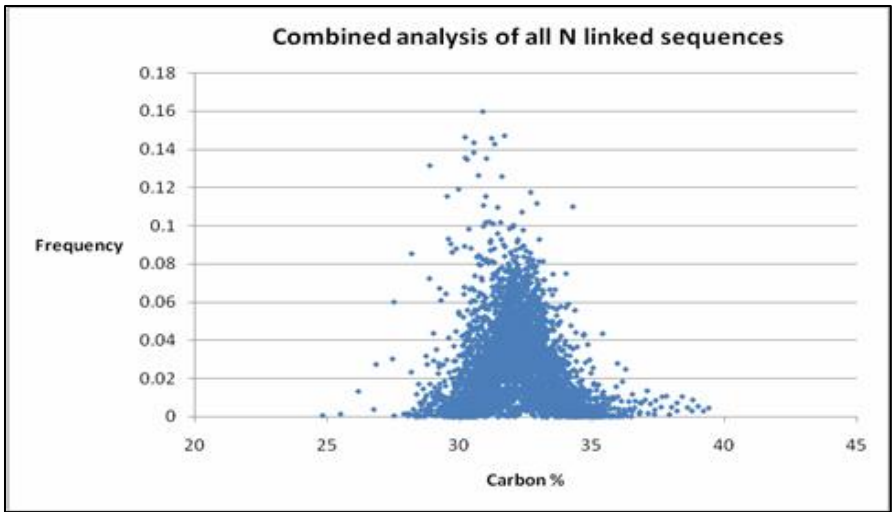

Figure 2: Combined analysis of all N linked sequences

The plot turned to be a bell shaped curve ranging from 24.83% to 39%. The maximum frequency of carbon percentage was observed for 30.87%. The mean frequency calculated was 32.05 and the skewness for the plot was 1.2586.

The same procedure was subjected to O linked glycosylated proteins (Figure 3).

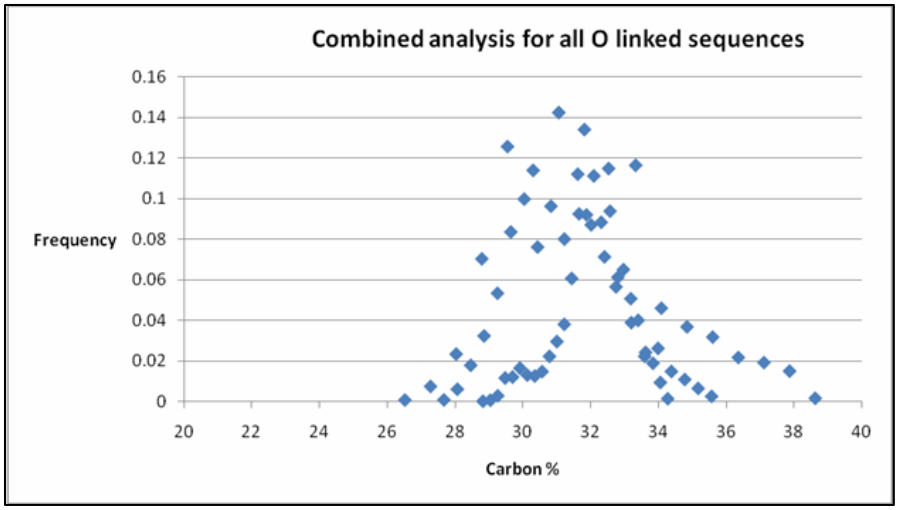

Figure 3: Combined analysis of all O linked sequences

The plot did provide curve in a range of 26.51% to 38.63%. Maximum frequency of carbon percentage was observed for 31.06%. Mean frequency was calculated as 31.71 and skewness of graph was 0.6693.

After the analysis of N linked and O linked glycosylated proteins, we plotted all the sequences togrther in one single plot (Figure 4).

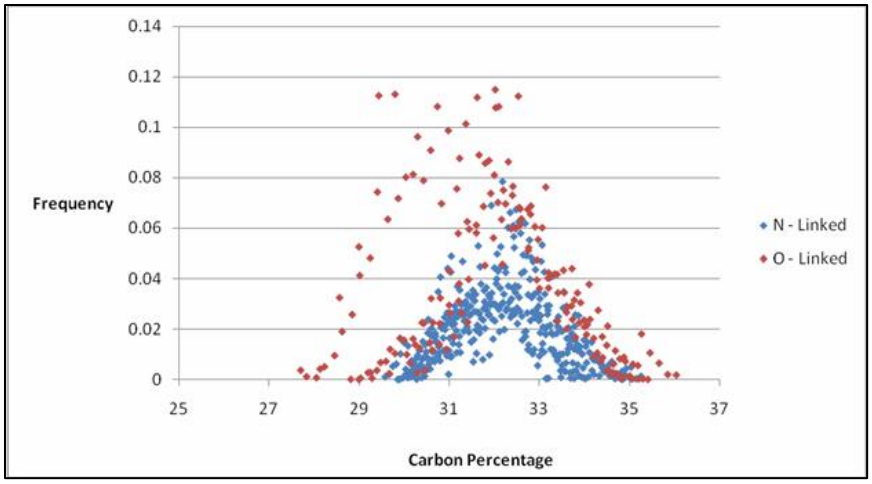

Figure 4: Comparative analysis between all N linked and O linked sequences

This plot significantly seperated the N linked protein sequences from O linked proteins sequences. This showed different carbon precentage for different types of glycosylation. Further to extend our analysis, we tried to cluster protein sequences based on carbon percentage. A distance matrix was created taking two parameters in account – skewness and mean frequency of the frequency distribution plot for N linked glycosylated sequences (Figure 5).

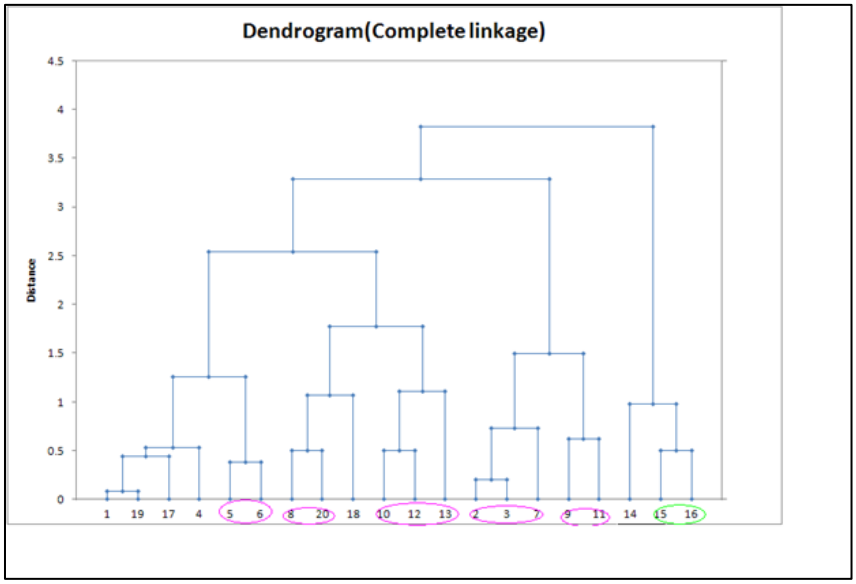

Figure 5: Dendrogram for all N linked and O linked reference sequences by XLMiner

The above dendrogram was analyzed and it was observed that the clusters formed were evidently categorized species wise (Table 3).

Table 3: Clusters formed in dendrogram from XLMiner

Nodes	Organisms
5-6	Homo sapiens
10-12-13	
2-3-7	
8-20	Homo sapiens and Pedobacter herpavirus
9-11	Homo sapiens and Mus musculus
15-16	Neisseria gonorrhoeae and Niesseria meningitides

By study of this dendrogram it was evident that the cluster formation was according to the species. Further to back up the analysis, we performed analysis of only N linked glycosylation sequences in MATLAB software (Figure 5).

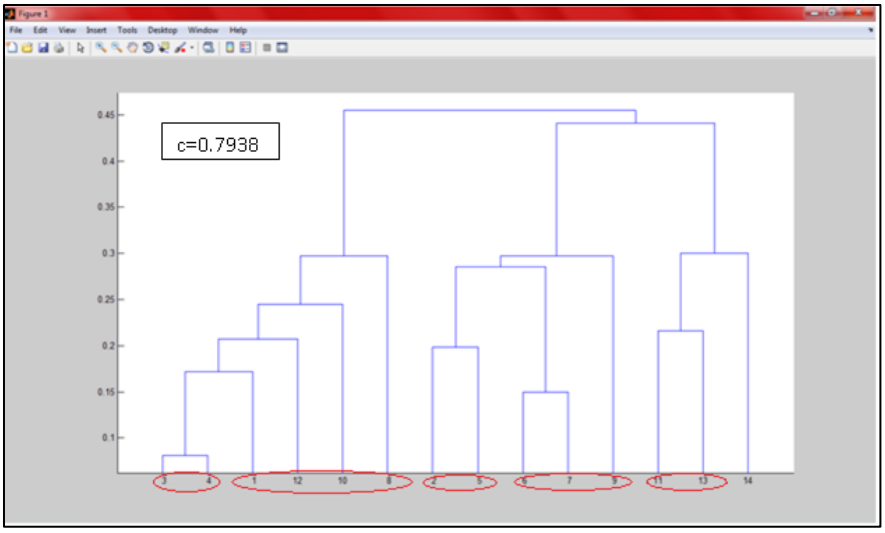

Figure 5: Dendrogram for all N linked reference sequences in MATLAB

The analysis of the dendrogram we concluded that the clustering was more advanced. It was now based on the function and subcellular location of the proteins. The cophenetic correlation coefficient for the above dendrogram was c=0.7938. Here the presence of the last (14^{th}) node was as an outlier. After masking that node the value for c=0.7953 (Figure 6).

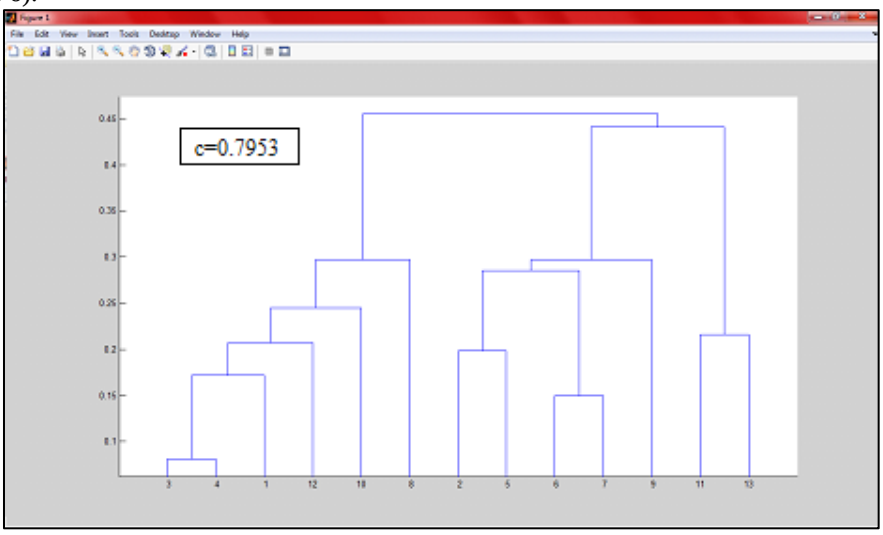

Figure 6: Dendrogram after masking the 14^{th} node

The clusters formed above we analyzed according to their function and location in cell (Table 4).

After examining the carbon content analysis in different types of glycosylated proteins in eukaryotes, we further proceed to analyze the role of carbon at particular glycosylation site. For this study we again subject the glycosylated protein sequence to a dynamic programming code, which focused only the side chains of the peptide and not the back bone. From the dynamic programming approach we obtained data regarding carbon percentage in particular site of glycosylation. This data was plot on a graph (Figure 7).

Table 4: Clusters formed in dendrogram from MATLAB.

Nodes	Function	Location
3	Hydrolases	Lysosome
4	Hydrolases	Lysosome
1	Catalyzes oxidation	Golgi apparatus membrane
12	Arylsulfatase activity	Golgi apparatus membrane
10	Catalyze formation of mannose	Golgi apparatus membrane
8	NADH Dehydrogenase	Mitochondria
2	Conversion of precursor to estrogen	Endoplasmic reticulum membrane
5	Hydrolysis	Lysosome
6	Degradation	Lysosome
7	Hydrolysis	Lysosome
9	NADH Dehydrogenase	Mitochondrial inner membrane
11	Transporter protein	Cell membrane
13	Arylsulfatase activity	Endoplasmic reticulum

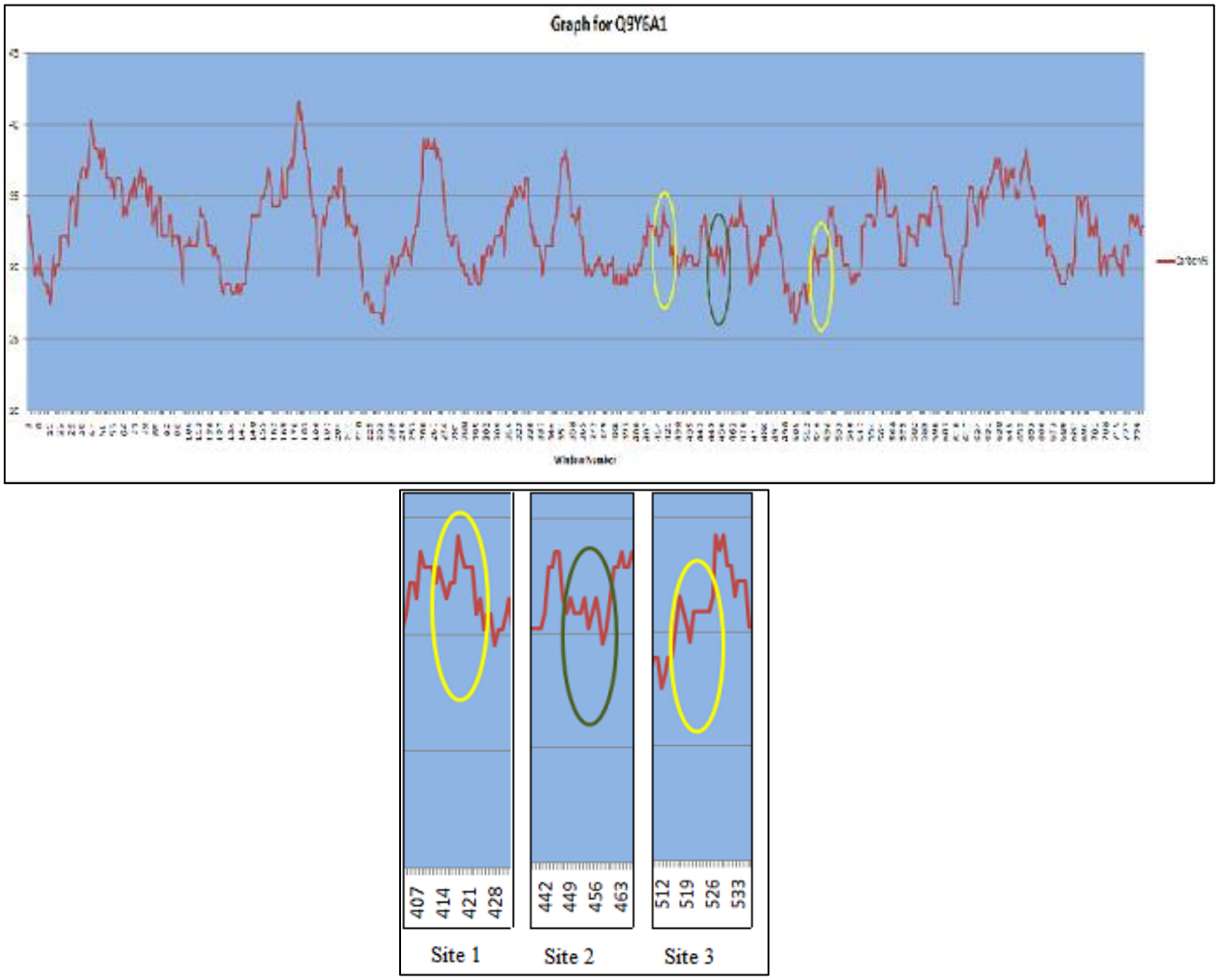

Figure 7: Site wise analysis for sequence Q9Y6A1

We subjected the same analysis to all reference glycosylated proteins. From the study of carbon percentage in each site of glycosylation, it was evident that the sites flanking the site of glycosylation site had steep rise or steep dip in carbon amount. So we concluded that the site of glycosylation does lie between very high and very low amount of glycosylation.

CONCLUSIONS

The placement of hydrophobic residues in protein sequence is characteristic to the protein function and structure. The post translational event of glycosylation is a commonly occurring phenomenon in all eukaryotes. The glycans are attached to a specific site on the protein chain. The amino acid sequence flanking the glycosylation site may have influence by modulating the access of oligosaccharide transferase to its active site, or on affinity of such interactions that may interfere glycosylation. Carbon, being the sole contributor to of hydrophobicity, is considered as one of the core factor in protein stability. Considering these facts it may be said that, the glycoprotein formation process is dependent on the carbon content present in and around the site of glycosylation. This concept may lead to greater understanding about the relation between the carbon content of protein and its significance with its function.

ACKNOWLEDGEMENT

Authors are grateful to Dr. A. K. Chauhan, Founder President & Mr. Aseem Chauhan, Chancellor Amity University Haryana & Chairperson AMITY Lucknow for providing necessary facilities and support. We also extend our gratitude to Maj. Gen. K.K Ohri, AVSM (Retd.), Pro Vice Chancellor, Amity University, Uttar Pradesh Lucknow Campus for constant support and encouragements.

REFERENCES

[1] C Chothia; AM Lesk, *EMBO J*, **1986**, 5(4), 823-836.
[2] P Johri. *Int J Comput Bioinfo In Silico Model*, **2013**, 2(4), 173-179.
[3] P Johri; M Gokhale. *J Comput Biol*, **2013**, 2(1), 1-6.
[4] P Anandagopu; S Suhanya; V Jayaraj; E Rajasekaran. *Bioinformation*, **2008**, 2(7), 304-307.
[5] V Jayaraj; M Vijayasarathy; R Geerthana; R Senthil; E Rajasekaran; SP Sarabjot; SP Simarjot; H Panchal; G Sathyamurthy; N Ramachandra Swamy; U B Sharma; A Patel; S Khatri; V Rishi; S Kr Gupta; MK Verma; GACI Omar; Stefan BALEV; S Paul; M K Rajak; G Sahoo. *Int J Computl Intell in Bioinfo*, **2009**, 2(2), 99-102.
[6] A Varki. Essentials of Glycobiology, 2nd Edition, Cold Spring Harbor Laboratories Press, New York, US, **2008**,
[7] E Maverakis; K Kim; M Shimoda; M Gershwin; F Patel; R Wilken; S Raychaudhuri; L Renee Ruhaak; C B Lebrilla. *J Autoimmun*, **2015**, 57(6), 1-13.
[8] WG Flynne. Biotechnology and Bioengineering, 1st Edition, Nova Science Publishers, New York, US, **2010.**
[9] Johri P; Tivedi M; Siddhique H; A Singh. *J Chem Pharm Res*, **2016,** 8(5), 204 -207.
[10] S Sachan; P Johri; M Tivedi; H Siddhique; A Singh. Study of carbon distribution at glycosylation sites in Arabidopsis thaliana. *J Chem Pharm Res*, **2016,** 8(5), 204 -207.