IN SILICO INVESTIGATION OF GLYCOSYLATION OF SET9 PROTEIN IN DIFFERENT STRAINS OF SORDARIA FIMICOLA

IQRA MOBEEN*, RABIA ARIF AND MUHAMMAD SALEEM

Molecular Genetics Research Laboratory, Department of Botany, University of the Punjab, 54590, Lahore, Pakistan
*Corresponding author’s email: iqra.phd.botany@pu.edu.pk

Abstract

Glycosylation is the most abundant and complex type of post-translational modification (PTM), which diversifies the protein functions and is affected by environmental stress. This study has investigated the polymorphism and glycosylation of the SET9 (a histone methyltransferase) protein of six strains of Sordaria fimicola and Neurospora crassa for the first time. Various bioinformatics tools are used to predict the O and N-glycosylation and those sites, which were common and have the highest EVP (enhancement value product) value, are chosen. Two N-glycosylation sites Asn206, Asn209 are found conserved in the N. crassa and all strains of S. fimicola, predicted by NetNGlyc 1.0. Seven O-glycosylation sites are reported on serine and threonine residues in SFS (South-facing slope) strains, which are presented at Ser13, Ser199, Ser202, Thr330, Ser230, Thr331, and Ser412, predicted by NetOGlyc 4.0 and ISOGlyP. While the NFS (North-facing slope) strains have five O-glycosylation sites similar to the SFS strains except for Ser202, not present in NFS strains. N. crassa has four sites similar to the SFS and NFS strains with two missing sites Ser202 and Ser412. These all sites are also presented on the 3D models of the SET9 protein. PolyView-2D, NetSurfP, and I-TASSER are used to predict the 2D and 3D structures and surface accessibility of glycosylated sites of the SET9. Polymorphism reported at five sites in the SFS strains and at one site in the NFS strains of the SET9 region, which clearly shows the effects of environmental stress on the SFS strains. The functions of the glycosylated sites of SET9 in S. fimicola are not known, but this study is evidenced by the more glycosylation sites in SFS strains than NFS strains. Therefore, the current study concludes that environmental stress is responsible for polymorphism, which alters the genes expression and this leads towards the creation of more diverse types of glycoproteins through PTMs that are specifically preferred by the organisms facing environmental stress.

Key words: Bioinformatics tools; N-glycosylation; O-glycosylation; Polymorphism; Post-translational modifications; Surface accessibility.

Introduction

Glycosylation is one of the most common and structurally diverse forms of post-translational modifications (PTMs) of proteins (Kobata, 1992; Allen & Kisailus, 1992) that takes place during or after protein synthesis (Varki, 1993). Glycosylation is a complex process in which 13 types of monosaccharides are attached with different eight types of amino acids (Spiro, 2002; Shental-Bechor & Levy, 2008) and involves many enzymes. It plays a particular role in protein secretion, localization, stability, immunogenicity and helps in mediating the communication of cell with the external environment in case of membrane glycoproteins (Lee et al., 2003).

There are predominantly two types of glycosylation in eukaryotes; O-glycosylation and N-glycosylation. N-glycosylation happens due to the attachment of glycan group Manβ1→4GlcNacβ1→4GlcNac1-N to asparagine residues having N-X-S/T motif, where X represents any amino acid except proline, known as N-glycosylation. While, O-glycosylation takes place due to the attachment of sugars to the β-hydroxyl group of serine, and threonine (Goto, 2007). Unlike N-glycosylation, O-glycosylation in fungi does not found to share a sequence motif of N-X-S/T and is diversified by the different types of glycans (Willer et al., 2003). Glycoproteins of plants and mammals are composed of GalNac, fucose, xylose, Glc, Gal, and Man amongst O-glycans (Hansen et al., 1996).

SET9 can undergoes many other types of PTMs (methylation, phosphorylation, ubiquitination, acetylation etc.), but this study is particularly focusing on the in silico study of N and O-glycosylation of SET9 protein from six strains of filamentous fungi S. fimicola. SET9 is a member of the SET-domain histone methyltransferase that can specifically methylates the lysine 4 of histone 3 (H3), (Kouskouti et al., 2003). It has a conserved SET domain but devoid of pre and post SET domains and it involves in transcription activation of histone proteins (Nishikoa et al., 2002).

Between both types of glycosylation, N-acetylgalactosaminy transferase (GalNAc)-type O-glycosylation is a complex and abundant type of post-translational modification, which is highly diversified in the cells or tissues due to the presence of different types of O-glycans and also where these glycans are attached (Bennett et al., 2012). Up to 20 types of unique GalNAc-T isoenzymes have been studied which initiate and control the O-glycosylation sites. These GalNAc-TTs have substrate specificities and are differentially expressed in the cells or tissues and their assembly in a cell is supposed to orchestrate the O-glycoproteome produced (Gill et al., 2010). Site-specific O-glycosylation is required to modulate the functions of the protein (Schjoldager & Clausen, 2012) and any deficiency in the GalNAc-T isoforms can cause diseases (Fakhro et al., 2011). In the current study, ten GalNAc-Ts (T1, T2, T3, T5, T10, T11, T12, T13, T14, and T16) isoforms are studied that are involved in the O-glycosylation of SET9 protein of different strains of S. fimicola.

Glycosylation initiates in the endoplasmic reticulum (ER) in the ribosomes during protein synthesis and glycans are supplemented to the unfolded protein while it is in the translocon complex (Helenius & Aebi, 2004). Glycans are synthesized by the coordinated expressions of many genes that code for enzymes like glycosidases, glycosyltransferases and many other enzymes that are responsible for the remodeling and synthesis of glycan.
chains. These genes also code for some accessory enzymes that are involved in the transport and synthesis of nucleotide sugars. It is suggested that glycans are important for the protein in the attaining of correct folds, however, the role of glycans in attaining this function is ambiguous (Parodi, 2000; Mitra et al., 2006).

We also carried out 2D, 3D structure analysis, and investigated surface accessibility of glycosylated sites for SET9 of S. fimicola. A protein must have surface accessible regions with free backbone hydrogen bonds that can easily be accessed by enzymes involved in modification (Iakoucheva et al., 2004). A post-translational modification domain specifically approaches the surface accessible residues and cannot be able to access the highly packed and ordered protein regions due to the steric hindrance (Dunker et al., 2002). There are few studies done on the post-translational modifications of proteins in fungi and this study is carried out to bridge the knowledge gap. However, an experimental study is required to investigate the specific functions and underlying mechanisms of O and N-glycoprotein of SET9 in the S. fimicola.

Another purpose of the current study is to investigate the polymorphism in the SET9 region of SFS and NFS strains of S. fimicola due to environmental stress—a key driving force for the creation of polymorphism and its effects on the protein glycosylation. SFS strains are collected from the south-facing slope of “Evolution Canyon (EC)” have more genetic diversity due to its xeric conditions than the NFS strains isolated from the north-facing slope of EC, Israel, which has mild environmental conditions (Saleem et al., 2001; Jamil et al., 2018). The S. fimicola serves as a model organism for the genetics study and its different strains are collected from the EC because this Canyon has an evolutionary significance, and is a hotspot for genetic diversity. Moreover, it provides a dynamic microclimate for the study of the genetic diversity of different microorganisms due to its diverse environmental conditions (Nevo, 2012; Arif et al., 2017).

Materials and Methods

Sub-culturing of fungi: Sub-culturing of different strains of Sordaria fimicola (SFS strains; S1, S2, S3, NFS strains; N5, N6, N7) was carried out on potato dextrose agar (PDA) media from stock cultures under sterile conditions. The N strains were collected from the north-facing slope (NFS) and S strains were collected from the south-facing slope (SFS) of “Evolution Canyon”, Israel. The fungal cultures were incubated in a refrigerated incubator at 20°C and the growth of fungi was obtained after 9 days.

DNA extraction and amplification of SET9 region by PCR: The DNA from different strains of S. fimicola was extracted by the modified Pietro et al., (1995) method followed by separation of DNA fragments with 1% of gel in electrophoresis, stained with ethidium bromide. The PCR amplification of the SET9 region was done by using touch down PCR conditions. The two primer pairs (F1=ATGCTCTCTCGACCCATTTGCG, R1=AGGGATCTGAGACATGACACT, F2=CATTCTCGAACCTCTCTTGT, R2=TAACCGCCCTGAGGACCGG) were designed by primer3 plus tool from NCBI. The PCR reaction volume was 15µl, which contained; 10µl 2X Amp Master Mix (GeneAll), 1µl forward primer, 1µl reverse primer (100µM each), 2µl DNA sample (1 in 10 dilutions of the g-DNA stock) and 1µl ddH2O. The stage 1 included the 15 cycles with an initial denaturation at 95°C for 3 min, the second denaturation for 30 sec, annealing at 65°C (Tm+10°C) for 45 sec and elongation at 72°C for 60 sec. The stage 2 contained 25 cycles with denaturation at 95°C for 30 sec, annealing at 50°C (Tm-5°C) for 45 sec and elongation at 72°C for 60 sec. The termination stage contained elongation at 72°C for 5 min followed by 15 min at 4°C (stop reaction) and final hold at 23°C until removed from the thermal cycler. The complete amplification of SET9 regions of all strains was carried out in total 40 cycles and 120 minutes. The sequencing of DNA was done from Macrogen Korea (a sequencing company) and DNA sequences were translated into the amino acid sequences by EMBOSSTransseq server.

Alignment of protein sequences: Protein sequences of SET9 of all strains of S. fimicola were aligned with reference N. crassa by clustal omega online tool to underpin the polymorphism.

Prediction of potential glycosylation sites: The N-glycosylation and O-glycosylation sites of SET9 were predicted from NetOGlyc 4.0 and NetNGlyc 1.0 servers, respectively. The threshold level for both N and O-glycosylation potential sites were chosen as 0.5. ISOglyP server was used to investigate the enzyme binding activity of potential O-glycosylation sites found by NetOGlyc. ISOglyP found potential positions of GalNAc-Ts (transferases) (T1, T2, T3, T5, T10, T11, T12, T13, T14, and T16) isoforms. ISOglyP calculated EVP (enhancement value product) values for each GalNAc-T isoforms, which showed the glycosylation rate.

Analysis of solvent accessibility and structural information of predicted glycosylation sites: NetSurfP v1.1 server was used to assess the surface accessibility of predicted glycosylation sites of serine, tyrosine, and asparagine. The secondary structure prediction and solvent accessibility of all strains were carried out using PolyView-2D SABLE, and ESPript 3.0. The most potential glycosylation sites were chosen by comparing and testing the most similar and highest scores for these sites.

Prediction of 3D structures of protein, visualization, and Glycoprotein building: TASSER server was used to predict the 3D structures of SET9 of all strains and N. crassa and structures were visualized by molecular visualization system PyMol. TASSER (Iterative Threading Assembly Refinement) uses the hierarchical approach to predict the 3D structures and functions of the protein. The confidence of each model is quantitatively measured by the C-score that is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. PyMol was used to mention the O-glycosylation and N-glycosylation positions with their respective amino acid residues on the 3D models of SET9.
Results

Analysis of polymorphism in the SET9 region:
Polymorphism was observed at six sites in the amino acid sequence of the SET9 region of *S. fimicola* with respect to the *N. crassa*. SFS strains have more polymorphism than NFS strains, five polymorphic sites were observed in the SFS strains and only one polymorphic site at 423RD position in NFS strains was observed, which is a unique for both SFS and NFS strains. At this position, *N. crassa* has asparagine (N), while the SFS and NFS strains have isoleucine (I) and lysine (K), respectively (Fig. 1).

A unique substitution was observed in the SET9 region at 423RD site where Asparagine (N) has been substituted with Isoleucine (I) and Lysine (K) in the SFS and NFS strains, respectively. There are two novel sites (329 and 353), which are depicting the conservation among the species of strongly similar properties, represented by (-). At 329 position, SFS strains have Histidine (H) while NFS strains and *N. crassa* has Glutamine (Q). At 353 position, SFS strains have Isoleucine (I) whereas NFS strains and *N. crassa* have Phenyalanine (F), (Fig. 1). These all mentioned residues are highly conserved among the species of strongly similar properties.

Analysis of glycosylation positions: N-glycosylation was investigated at two sites Asn329 and Asn353 in all studied strains of *S. fimicola* and *N. crassa*. O-glycosylation was found at some different positions, which are presented at Ser196, Ser219, Ser262, Thr330, and Thr333 in *N. crassa*, at Ser197, Ser262, Thr330, Ser320, Thr333, and Ser347 in the SFS strains and at Ser197, Ser219, Ser262, Thr330, Ser320, and Thr333 in the NFS strains of *S. fimicola* (Table 1). Different servers for the prediction of glycosylation were used and the most similar positions with the highest EVP values and glycosylation potential were picked to ensure accuracy. Those sites that have more potential are more likely to be glycosylated and vice versa.

Secondary structure prediction: A comparison of the secondary structure of *N. crassa* was made with different strains of *S. fimicola*, which showed some differences in the positions of coils, α and β sheets. *N. crassa* showed β sheets at Leu24, Arg34, Lys47, Leu56, Leu196, Val256, Ala49 (Fig. 3b, 3c, 3d) and N5 strains were shown β sheets at Leu24, Arg34, Lys47, Leu56, Thr222, Gln356, Asp366, Thr395, Val89, and Ala49 (Fig. 3b, 3c, 3d) and N5 strains were shown β sheets at Leu24, Arg34, Lys47, Leu56, Val215, Val256, Leu435, and Arg436 (Fig. 3e). N6 and N7 strains were shown marked differences in the positions of β sheets from other strains, which were defined at Leu28, Arg32, Lys57, Leu62, Val221, Thr238, Gln382, Val402, Asp272, Thr301, and Cys425 (Fig. 3f and 3g). The positions of coils and α sheet structures are shown in Fig. 3a-g.

Assessment of Surface and solvent accessibility:
Surface accessibility of glycosylated sites was determined which showed that eight sites Ser196, Ser219, Ser262, Ser320, Ser347, Thr330, Thr333, and Asn353 are exposed. Only one glycosylated site Asn205 was buried. Solvent accessibility of modified sites was also assessed and we found two sites (Ser196 and Thr330) with intermediate accessibility and other sites with low accessibility except for one site Asn205 that was found buried (Table 1).

### Table 1. Prediction of potential O and N-glycosylation sites, surface accessibility, and secondary structure analysis for *N. crassa*, SFS and NFS strains of *S. fimicola*.

<table>
<thead>
<tr>
<th>Substrates</th>
<th>Positions</th>
<th>Glycosylation (NetNGlyc)</th>
<th>Glycosylation (NetOGlyc)</th>
<th>Surface accessibility (NetSurfP)</th>
<th>Secondary structure (PolyView-2D, ESPript 3.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O-Glycosylation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residues</td>
<td>Ser</td>
<td>Ser</td>
<td>Ser</td>
<td>Ser</td>
<td>Ser</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>219</td>
<td>320</td>
<td>320</td>
<td>330</td>
<td>333</td>
</tr>
<tr>
<td>Very high potential</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
</tr>
<tr>
<td>High potential</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
</tr>
<tr>
<td>Low potential</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
</tr>
<tr>
<td>Intermediate accessibility</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
</tr>
<tr>
<td>Low accessibility</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
<td>Exposed</td>
</tr>
<tr>
<td>Buried</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Exposed</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IN SILICO GLYCOSYLATION OF SET9 PROTEIN
Fig. 1. Multiple sequence alignment of amino acid sequence of SET9 regions of six strains of *S. fimicola* with respect to the reference *N. crassa* by Clustal Omega. Amino acid residues (spaced sites, without asterisks) highlighted in yellow color are showing polymorphic sites, sites with asterisks (*) are fully conserved sites (in green color), sites with symbol (:) are depicting conservation between groups of strongly similar properties (in red color) and asterisks represented in blue color at the end of amino acid sequence are stop codons.
S2 Strain is sent in owing the strains and Y320. N6 Strain is favored by the S1 Strain N7 Strain on, localization, and stability similar properties. Arif conservation among the species of strongly has also been found in this study, which shows the organisms facing environmental stress (Chistiakov 2006; Bhargava & Fuentes, 2010). Modifications which are specifica

Discussion

We found five polymorphic sites in the amino acid sequence of SET9 of six strains S. fimicola with respect to the model organism N. crassa (Fig. 1). Four polymorphic sites are observed in the SFS strains as compared to the NFS strains for which only one sites has been observed. This is due to environmental stress and harsh conditions of EC, which SFS strains are facing and produce more polymorphism in the exonic regions to ensure their survival. Such environmental stress is responsible for base substitutions and mutation which alter the genes expression and this lead towards the creation of more diverse types of frontier molecules or proteins and glycoproteins through post-translational modifications which are specifically favored by the organisms facing environmental stress (Chistiakov et al., 2006; Bhargava & Fuentes, 2010).

Two novel sites at 329 and 353 amino acid positions has also been found in this study, which shows the conservation among the species of strongly similar properties. Arif et al., (2019) also studied such positions in COX1 protein of parental strains of S. fimicola. Glycosylation is an important and most abundant form of post-translational modification in eukaryotes that is involved in protein secretion, localization, and stability (Deshpande et al., 2008). N-glycosylation initiates during protein folding in endoplasmic reticulum and O-glycosylation takes place in the Golgi (Cumming, 2003). To the best of our knowledge, glycosylation of the SET9 protein of S. fimicola is first time studied. Uslupehlivan et al., (2018) studied two N-glycosylation sites Asn184 and Asn200 in human PrP (prion) protein. The current study has reported two N-glycosylation positions on Asn205 and Asn262 found to be conserved among N. crassa and in all strains of S. fimicola (Table 1 and Fig. 2).

The present study investigated the O-glycosylation at two threonine and four serine residues of all strains. A unique polymorphic site Ser230 was found to be present in the SFS strains (not in NFS strains) that might be involved in performing some important functions in these strains and this site may have an evolutionary potential (Table 1 and Fig. 2). This is because SFS, strains are facing environmental stress that is a key driving evolutionary force for species survival and adaptation (Saleem et al., 2001). Another unique feature for this site (Ser230) in SFS strains is that it has an α-helix structure, while other glycosylated sites havecoil structures (Table 1). On the other hand, Rini & Leffler (2010) have studied that the O-glycosylation of threonine and serine residues of nuclear and cytoplasmic proteins are involved in the signal transduction in the multicellular organisms. It can be argued that the O-glycosylation of SET9 (a cytoplasmic protein) might be able to perform a similar function in S. fimicola.

Prediction of 3D structures of SET9 Protein: The 3D structures of each strain were built by I-TASSER and differences were found at loop regions of SET9 of N. crassa, SFS and NFS strains. 3D structures of SFS strains were altogether similar and 3D structures of NFS strains were shown strong similarity with each other (Fig. 4).

Glycosylation: Glycoprotein models were built by PyMol, which helped to label the O and N-glycosylation positions on the 3D models of SET9 protein shown in Fig. 5.
Two sites Ser$_{74}$ and Ser$_{219}$ have a very high O-glycosylation potential with EVP value above 0.6, which revealed that these are the potential sites for O-glycosylation. Goto (2007) has reported that O-glycosylation in filamentous fungi plays an important role in the maintenance of fungal morphology, development of hyphae, and differentiation. Like other filamentous fungi, it can be assumed that O-glycosylation might be involved in performing similar functions in _S. fimicola_.

We did not find any study, which has reported the O and N-glycosylation of SET9 protein in any fungus or any other eukaryotic organism. Some recent studies have been carried out on the post-translational modifications of various other proteins in different strains of _S. fimicola_. Arif _et al._ (2019) have studied glycosylation of COX1 protein in the parental strains of _S. fimicola_ and reported two O-glycosylation positions (Thr$_{126}$ and Thr$_{127}$). Another study by Jamil _et al._ (2018) has shown the nine serine and five threonine O-glycosylation modifications for histone-3 protein in the SFS and NFS strains of _S. fimicola_. Rana _et al._ (2018) have studied O-glycosylation on six serine and four threonine residues of manganese superoxide dismutase of different strains of _S. fimicola_.

Post-translational modifications do not take place at surface accessible regions of protein all the time, but these regions are more likely to be modified and have more interaction with enzymes involved in modifications (Iakoucheva _et al._, 2004; Dunker _et al._, 2002). In this study, surface accessibility of glycosylated sites (modified residues) has shown that all sites are exposed and only one site Asn$_{205}$ is found buried (Table 1). These results support the view of Iakoucheva _et al._ (2004) and Dunker _et al._ (2002). Uslupehliyan _et al._ (2018) used the NetSurfP server to predict the surface accessibility of glycosylated sites of prion protein of sheep and they found that all glycosylated residues are exposed. Ahmadv
et al., (2003) predicted surface accessibility of various proteins by RVP-net. Likewise, Pang et al., (2007) have studied the surface accessibility of modified sites of various proteins and described that surface accessibility is important for protein-protein interactivity. It is not known yet that either surface accessible regions of SET9 are involved in protein-protein interaction or not.

![3D structures of SET9 protein](image)

**Fig. 4.** 3D structures of SET9 protein of (a) *N. crassa* (b) S1, S2, S3 and (c) N5, N6, N7 strains of *S. fimicola*. 3D structures of SFS and NFS strains are shown once for each due to the similarity in their structures. Motifs shown in red color indicate α-helix, yellow indicate β-sheet and motifs shown in green represent coil structure.

![3D structures of SFS and NFS strains](image)

**Fig. 5.** SET9 Protein 3D structures with O-glycosylation and N-glycosylation residues labeled with PyMol. Glycoprotein of *N. crassa* (a), SFS strains (b), and NFS strains (c) of *S. fimicola*.

**Conclusion**

In this study, we have first time demonstrated the O and N-glycosylation of SET9 protein in different polymorphic strains of *S. fimicola* and *N. crassa*. In addition to this, different bioinformatics tools were used to generate 2D and 3D structures of SET9. N-glycosylation positions were found to be conserved among *N. crassa* and all strains, while O-glycosylation sites showed some variations among SFS and NFS. More polymorphic positions/sites were observed in the SFS strains due to the stressful environment of EC, they are facing. Likewise, more glycosylation sites have been observed in the SFS strains, which evidenced the consequences of environmental stress. There are few studies on the post-translational modifications of some proteins of *S. fimicola* by different bioinformatics tools. Experimental studies are required to underpin the specific functions and underlying mechanisms related to the glycosylation of SET9 protein in *S. fimicola*. 
Reference


