Pak. J. Bot., 54(3): 1147-1155, 2022. DOI: http://dx.doi.org/10.30848/PJB2022-3(24)

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 Asn^{205}, Asn^{262} 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 Ser^{74}, Ser^{219}, Ser^{292}, Thr^{330}, Ser^{320}, Thr^{333}, and Ser^{412}, 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 Ser^{320}, not present in NFS strains. *N. crassa* has four sites similar to the SFS and NFS strains with two missing sites Ser^{320} and Ser^{412}. 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 eight different 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.,* 2004). It has a conserved SET domain but devoid of pre and post SET domains and it involves in transcription activation of histone proteins (Nishioka *et al.,* 2002).

Between both types of glycosylation, N-acetylgalactosaminyl 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-Ts 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℃ 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=ATGCTCTCTCGACCCATTTCG, R1=AGGGATCGTAGACATGACACT, F2=CATCTGGAACCCTCTCTGTTT, R2=TCAACCGCCCTGGGACGCGG) 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℃ for 3 min, the second denaturation for 30 sec, annealing at 65℃ (Tm+10℃) for 45 sec and elongation at 72℃ for 60 sec. The stage 2 contained 25 cycles with denaturation at 95℃ for 30 sec, annealing at 50℃ (Tm-5℃) for 45 sec and elongation at 72℃ for 60 sec. The termination stage contained elongation at 72℃ for 5 min followed by 15 min at 4℃ (stop reaction) and final hold at 23℃ 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 EMBOSS Transeq 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: I-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. I-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 Phenylalanine (F), (Fig. 1). These all mentioned residues are highly conserved among the species of strongly similar properties.

Analysis of glycos**ylation positions:** N-glycosylation was investigated at two sites Asn^{205} and Asn^{262} in all studied strains of *S. fimicola* and *N. crassa*. O-glycosylation was found at some different positions, which are presented at Ser^{74}, Ser^{219}, Ser^{292}, Thr^{330}, and Thr^{333} in *N. crassa,* at Ser^{74}, Ser^{219}, Ser^{292}, Thr^{330}, Ser^{320}, Thr^{333}, and Ser^{412} in the SFS strains and at Ser^{74}, Ser^{219}, Ser^{292}, Thr^{330}, Ser^{320}, and Thr^{333} 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* was showed β sheets at Leu^{24}, Arg^{34}, Lys^{47}, Leu^{56}, Leu^{196}, Thr^{222}, Gln^{236}, $Asp2^{66}$, Thr^{295}, Cys^{434} (Fig. 3a), SFS strains at Leu^{24}, Arg^{34}, Lys^{47}, Leu^{56}, Leu^{196}, Thr^{222}, Gln^{236}, Asp^{266}, Thr^{295}, Val^{256}, Ala^{449} (Fig. 3b, 3c, 3d) and N5 strains were shown β sheets at Leu^{24}, Arg^{34}, Lys^{47}, Leu^{56}, Leu^{196}, Val^{215}, Val^{256}, Leu^{433}, and Arg^{436} (Fig. 3e). N6 and N7 strains were shown marked differences in the positions of β sheets from other strains, which were defined at Leu^{30}, Arg^{40}, Lys^{52}, Leu^{62}, Val^{221}, Thr^{228}, Gln^{242}, Val^{262}, Asp^{272}, Thr^{301}, and Cys^{443} (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 Ser^{74}, Ser^{219}, Ser^{292}, Ser^{320}, Ser^{412}, Thr^{330}, Thr^{333}, and Asn^{262} are exposed. Only one glycosylated site Asn^{205} was buried. Solvent accessibility of modified sites was also assessed and we found two sites (Ser^{74} and Thr^{330}) with intermediate accessibility and other sites with low accessibility except for one site Asn^{205} 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*. ISOGlyP enzymes binding values are shown for each strain; very high potential if EVP>2, high potential if EVP>1 and low potential if EVP<1. Surfaces with open accessibility are shown as exposed surfaces and surfaces with buried accessibility are represented as buried. Different servers are used for the prediction of glycosylation sites and the sites, which are most similar and have high potential, are chosen to enhance the accuracy.

Substrates	*O*-GalNAc prediction (NetOGlyc), *N*-glycosylation (NetNGlyc)	*O*-Glycosylation prediction (ISOGlyP)	Surface accessibility (NetSurfP)	Secondary structure (PolyView-2D, NetSurfP)	Relative solvent accessibility (PolyView-2D, NetSurfP, ESPript 3.0)
Residues	Positions				
Ser	74	Very high potential	Exposed	Coil	Intermediate accessibility
Ser	219	Very high potential	Exposed	Coil	Low accessibility
Ser	292	High potential	Exposed	Coil	Low accessibility
Ser	320 (only in SFS strains)	Low potential	Exposed	Helix	Low accessibility
Ser	412 (Only in SFS and NFS strains)	Low potential	Exposed	Coil	Low accessibility
Thr	330	High potential	Exposed	Coil	Intermediate accessibility
Thr	333	High potential	Exposed	Coil	Low accessibility
Asn	205	-	Buried	Coil	Buried
Asn	262	-	Exposed	Coil	Low accessibility

```
           MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILTIPAGILWTVK 60
S2         MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILTIPAGILWTVK 60
S3         MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILTIPAGILWTVK 60
N.C        MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILTIPAGILWTVK 60
N5         MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILTIPAGILWTVK 60
N6         MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILTIPAGILWTVK 60
N7         MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILTIPAGILWTVK 60
           ************************************************************

S1         HAYADALLGPALRSAQPPLSVEDTLATYNLFVKSRESGYDGQRSHIAALPASYSSSILFA 120
S2         HAYADALLGPALRSAQPPLSVEDTLATYNLFVKSRESGYDGQRSHIAALPASYSSSILFA 120
S3         HAYADALLGPALRSAQPPLSVEDTLATYNLFVKSRESGYDGQRSHIAALPASYSSSILFA 120
N.C        HAYADPLLGPALRSAQPPLSVEDTLATYILFVKSRESGYDGQRSHIAALPASYSSSILFA 120
N5         HAYADPLLGPALRSAQPPLSVEDTLATYILFVKSRESGYDGQRSHIAALPASYSSSILFA 120
N6         HAYADPLLGPALRSAQPPLSVEDTLATYILFVKSRESGYDGQRSHIAALPASYSSSILFA 120
N7         HAYADPLLGPALRSAQPPLSVEDTLATYILFVKSRESGYDGQRSHIAALPASYSSSILFA 120
           ***** ********************** **********************************

S1         EDDLEACAGTSLYTITKQLEQSIEDDHRALVVRLFVQHPDLFPLDKFTVEDYKWALCTVW 180
S2         EDDLEACAGTSLYTITKQLEQSIEDDHRALVVRLFVQHPDLFPLDKFTVEDYKWALCTVW 180
S3         EDDLEACAGTSLYTITKQLEQSIEDDHRALVVRLFVQHPDLFPLDKFTVEDYKWALCTVW 180
N.C        EDDLEACAGTSLYTITKQLEQSIEDDHRALVVRLFVQHPDLFPLDKFTVEDYKWALCTVW 180
N5         EDDLEACAGTSLYTITKQLEQSIEDDHRALVVRLFVQHPDLFPLDKFTVEDYKWALCTVW 180
N6         EDDLEACAGTSLYTITKQLEQSIEDDHRALVVRLFVQHPDLFPLDKFTVEDYKWALCTVW 180
N7         EDDLEACAGTSLYTITKQLEQSIEDDHRALVVRLFVQHPDLFPLDKFTVEDYKWALCTVW 180
           ************************************************************

S1         SRAMDFVLADGNSIRLLAPFADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFIN 240
S2         SRAMDFVLADGNSIRLLAPFADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFIN 240
S3         SRAMDFVLADGNSIRLLAPFADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFIN 240
N.C        SRAMDFVLADGNSIRLLAPFADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFIN 240
N5         SRAMDFVLADGNSIRLLAPFADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFIN 240
N6         SRAMDFVLADGNSIRLLAPFADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFIN 240
N7         SRAMDFVLADGNSIRLLAPFADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFIN 240
           ************************************************************

S1         YGPVPNSRLLRLYGFVIPGNPNDSYDLVLSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
S2         YGPVPNSRLLRLYGFVIPGNPNDSYDLVLSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
S3         YGPVPNSRLLRLYGFVIPGNPNDSYDLVLSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
N.C        YGPVPNSRLLRLYGFVIPGNPNDSYDLVLSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
N5         YGPVPNSRLLRLYGFVIPGNPNDSYDLVHSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
N6         YGPVPNSRLLRLYGFVIPGNPNDSYDLVHSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
N7         YGPVPNSRLLRLYGFVIPGNPNDSYDLVHSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
           ****************************** **********************************

S1         TDPLPKKVLRYLRIQRLDASGLAVIARQHTDATDGKISDSNEVEILRFLVESINYLLDGF 360
S2         TDPLPKKVLRYLRIQRLDASGLAVIARQHTDATDGKISDSNEVEILRFLVESINYLLDGF 360
S3         TDPLPKKVLRYLRIQRLDASGLAVIARQHTDATDGKISDSNEVEILRFLVESINYLLDGF 360
N.C        TDPLPKKVLRYLRIQRLDASGLAVIARQQTDATDGKISDSNEVEILRFLVESFNYLLDGF 360
N5         TDPLPKKVLRYLRIQRLDASGLAVIARQQTDATDGKISDSNEVEILRFLVESFNYLLDGF 360
N6         TDPLPKKVLRYLRIQRLDASGLAVIARQQTDATDGKISDSNEVEILRFLVESFNYLLDGF 360
N7         TDPLPKKVLRYLRIQRLDASGLAVIARQQTDATDGKISDSNEVEILRFLVESFNYLLDGF 360
           ******************************:*********************:********

S1         GTPLEKLEQQLAEGVYSPGGNAWAAAHVRLGEQRVLRLARKRAEELLSAVESGSGNVRNR 420
S2         GTPLEKLEQQLAEGVYSPGGNAWAAAHVRLGEQRVLRLARKRAEELLSAVESGSGNVRNR 420
S3         GTPLEKLEQQLAEGVYSPGGNAWAAAHVRLGEQRVLRLARKRAEELLSAVESGSGNVRNR 420
N.C        GTPLEKLEQQLAEGVYSPGGNAWAAAHVSLGEQRVLRLARKRAEELLSAVESGSGNVRNR 420
N5         GTPLEKLEQQLAEGVYSPGGNAWAAAHVSLGEQRVLRLARKRAEELLSAVESGSGNVRNR 420
N6         GTPLEKLEQQLAEGVYSPGGNAWAAAHVSLGEQRVLRLARKRAEELLSAVESGSGNVRNR 420
N7         GTPLEKLEQQLAEGVYSPGGNAWAAAHVSLGEQRVLRLARKRAEELLSAVESGSGNVRNR 420
           ****************************** **********************************

S1         CAICGKGSAQLSLCGRCKAVAYCGRACQVAHYKQHKARCQATASQGG*467
S2         CAICGKGSAQLSLCGRCKAVAYCGRACQVAHYKQHKARCQATASQGG*467
S3         CAICGKGSAQLSLCGRCKAVAYCGRACQVAHYKQHKARCQATASQGG*467
N.C        CANCGKGSAQLSLCGRCKAVAYCGRACQVAHYKQHKARCQATASQGG-467
N5         CAKCGKGSAQLSLCGRCKAVAYCGRACQVAHYKQHKARCQATASQGG*467
N6         CAKCGKGSAQLSLCGRCKAVAYCGRACQVAHYKQHKARCQATASQGG*467
N7         CAKCGKGSAQLSLCGRCKAVAYCGRACQVAHYKQHKARCQATASQGG*467
           ** ************************************************
```

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.

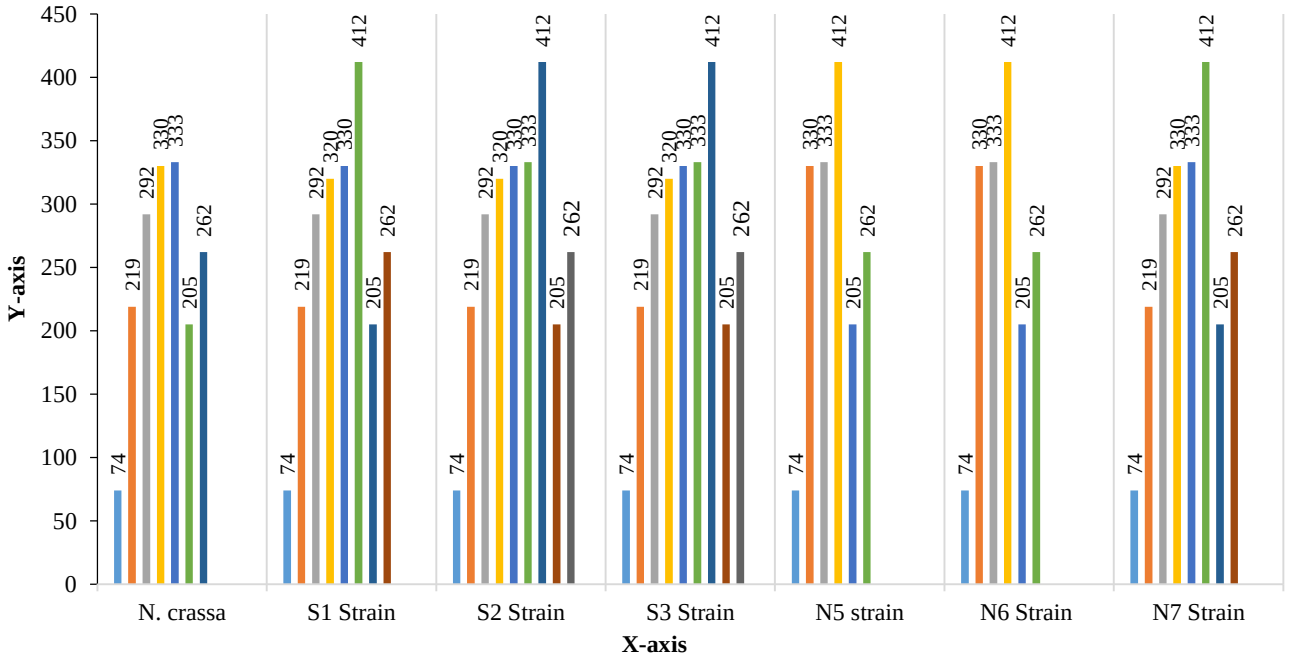


Fig. 2. This graph is showing the possible O-glycosylation and N-glycosylation sites determined at high accuracy rate for *N. crassa* and all strains of *S. fimicola.* The X-axis is showing the strains and Y-axis is representing the amino acid sequence. The numerical values shown on the top of the columns are the potential O and N-glycosylation sites. O-glycosylation positions are shown in black color and N-glycosylation positions in red colors.

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).

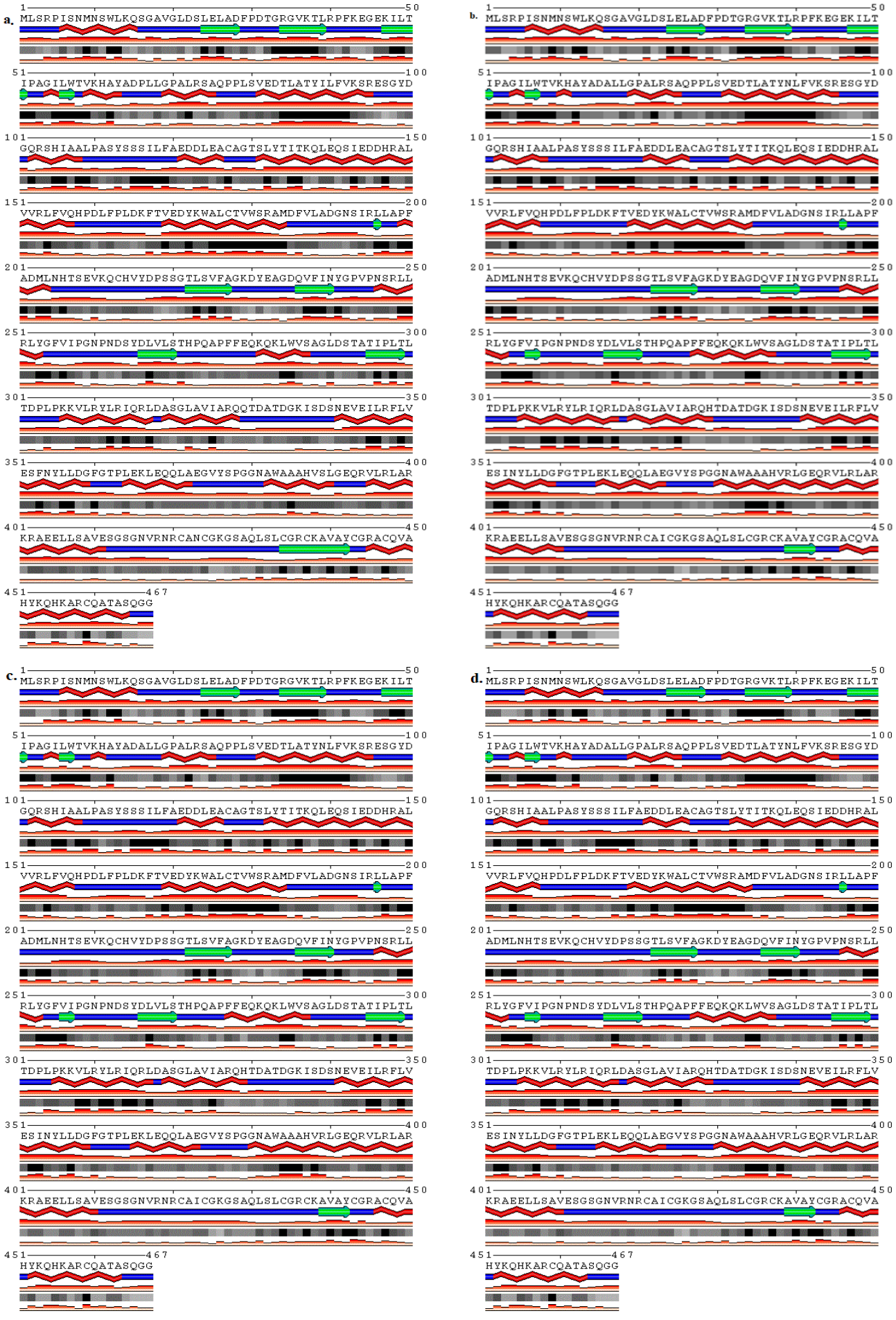
Glycosylation: Glycoprotein models were built by PyMol, which helped to label the O and N-glycosylation positions on the 3D models of SET9 protein shown in Fig. 5.

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 Asn^{184} and Asn^{200} in human PrP (prion) protein. The current study has reported two N-glycosylation positions on Asn^{205} and Asn^{262} 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 Ser^{320} 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 (Ser^{320}) in SFS strains is that it has an α-helix structure, while other glycosylated sites have coil 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*.


a.
1 MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILT 50
51 IPAGILWTVKHAYADPLLGPALRSAQPPLSVEDTLATYILFVKSRESGYD 100
101 GQRSHIAALPASYSSSILFAEDDLEACAGTSLYTITKQLEQSIEDDHRAL 150
151 VVRLFVQHPDLFPLDKFTVEDYKWALCTVWSRAMDFVLADGNSIRLLAPF 200
201 ADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFINYGPVPNSRLL 250
251 RLYGFVIPGNPNDSYDLVLSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
301 TDPLPKKVLRYLRIQRLDASGLAVIARQQTDATDGKISDSNEVEILRFLV 350
351 ESFNYLLDGFGTPLEKLEQQLAEGVYSPGGNAWAAAHVSLGEQRVLRLAR 400
401 KRAEELLSAVESGSGNVRNRCANCGKGSAQLSLCGRCKAVAYCGRACQVA 450
451 HYKQHKARCQATASQGG 467
b.
1 MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILT 50
51 IPAGILWTVKHAYADALLGPALRSAQPPLSVEDTLATYNLFVKSRESGYD 100
101 GQRSHIAALPASYSSSILFAEDDLEACAGTSLYTITKQLEQSIEDDHRAL 150
151 VVRLFVQHPDLFPLDKFTVEDYKWALCTVWSRAMDFVLADGNSIRLLAPF 200
201 ADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFINYGPVPNSRLL 250
251 RLYGFVIPGNPNDSYDLVLSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
301 TDPLPKKVLRYLRIQRLDASGLAVIARQHTDATDGKISDSNEVEILRFLV 350
351 ESINYLLDGFGTPLEKLEQQLAEGVYSPGGNAWAAAHVRLGEQRVLRLAR 400
401 KRAEELLSAVESGSGNVRNRCAICGKGSAQLSLCGRCKAVAYCGRACQVA 450
451 HYKQHKARCQATASQGG 467
c.
1 MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILT 50
51 IPAGILWTVKHAYADALLGPALRSAQPPLSVEDTLATYNLFVKSRESGYD 100
101 GQRSHIAALPASYSSSILFAEDDLEACAGTSLYTITKQLEQSIEDDHRAL 150
151 VVRLFVQHPDLFPLDKFTVEDYKWALCTVWSRAMDFVLADGNSIRLLAPF 200
201 ADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFINYGPVPNSRLL 250
251 RLYGFVIPGNPNDSYDLVLSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
301 TDPLPKKVLRYLRIQRLDASGLAVIARQHTDATDGKISDSNEVEILRFLV 350
351 ESINYLLDGFGTPLEKLEQQLAEGVYSPGGNAWAAAHVRLGEQRVLRLAR 400
401 KRAEELLSAVESGSGNVRNRCAICGKGSAQLSLCGRCKAVAYCGRACQVA 450
451 HYKQHKARCQATASQGG 467
d.
1 MLSRPISNMNSWLKQSGAVGLDSLELADFPDTGRGVKTLRPFKEGEKILT 50
51 IPAGILWTVKHAYADALLGPALRSAQPPLSVEDTLATYNLFVKSRESGYD 100
101 GQRSHIAALPASYSSSILFAEDDLEACAGTSLYTITKQLEQSIEDDHRAL 150
151 VVRLFVQHPDLFPLDKFTVEDYKWALCTVWSRAMDFVLADGNSIRLLAPF 200
201 ADMLNHTSEVKQCHVYDPSSGTLSVFAGKDYEAGDQVFINYGPVPNSRLL 250
251 RLYGFVIPGNPNDSYDLVLSTHPQAPFFEQKQKLWVSAGLDSTATIPLTL 300
301 TDPLPKKVLRYLRIQRLDASGLAVIARQHTDATDGKISDSNEVEILRFLV 350
351 ESINYLLDGFGTPLEKLEQQLAEGVYSPGGNAWAAAHVRLGEQRVLRLAR 400
401 KRAEELLSAVESGSGNVRNRCAICGKGSAQLSLCGRCKAVAYCGRACQVA 450
451 HYKQHKARCQATASQGG 467

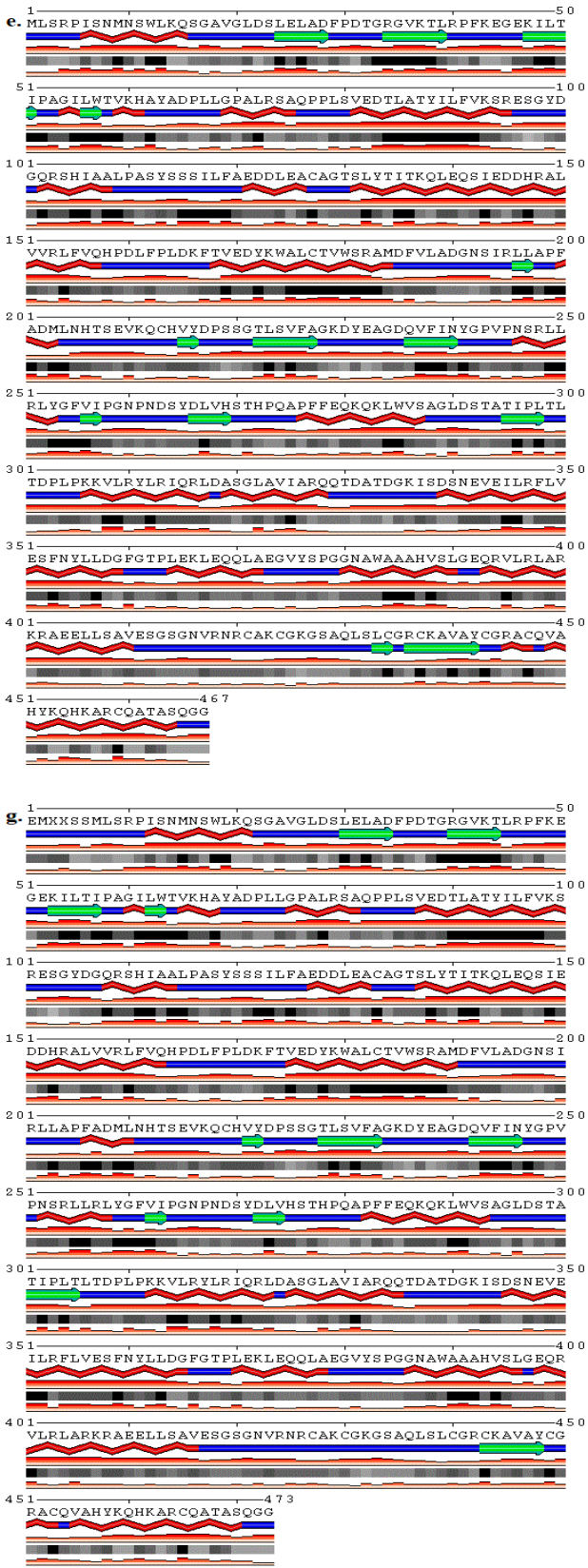


Fig. 3. Prediction of secondary structure of SET9 protein by Polyview-2D SABLE. The first line (scale) is showing the amino acid numbers, the second line is showing the amino acid sequence and the third line is showing α-helix motif (red zigzag), coil structure (blue) and β-sheet motif (green arrow). The fourth row is representing the confidence level of prediction. The fifth and sixth lines are showing the solvent accessibility. Secondary structures of (a) *N. crassa* (b) S1 (c) S2 (d) S3 (e) N5 (f) N6 (g) N7.

Two sites Ser^{74} and Ser^{219} have a very high O-glycosylation potential with EVP value above 06, 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). Uslupehlivan *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. Ahmad

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.

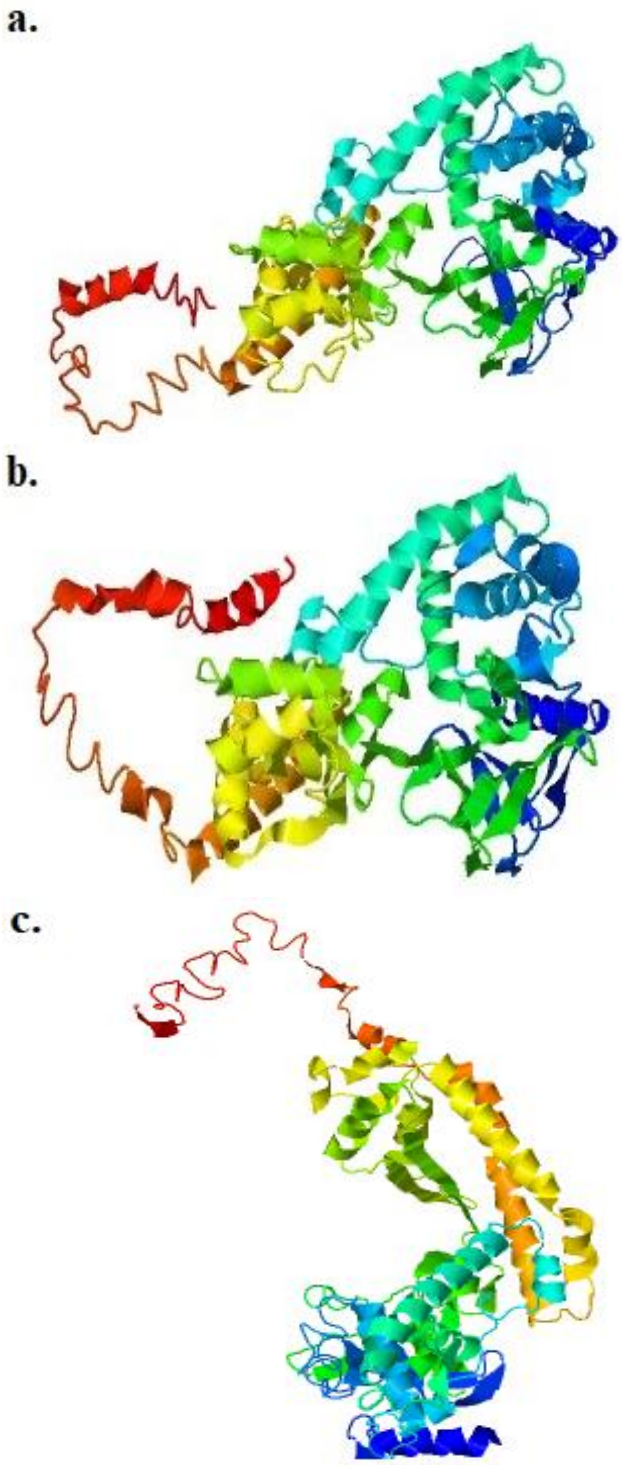


Fig. 4. 3D structures of SET9 protein of (a) *N. crassa* (b) S1, S2, S3 and (c) N5, N6, N7 strains of *S. fimicola.* 3D structures of SFS and NFS strains are shown once for each due to the similarity in their structures. Motifs shown in red color indicate α-helix, yellow indicate β-sheet and motifs shown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

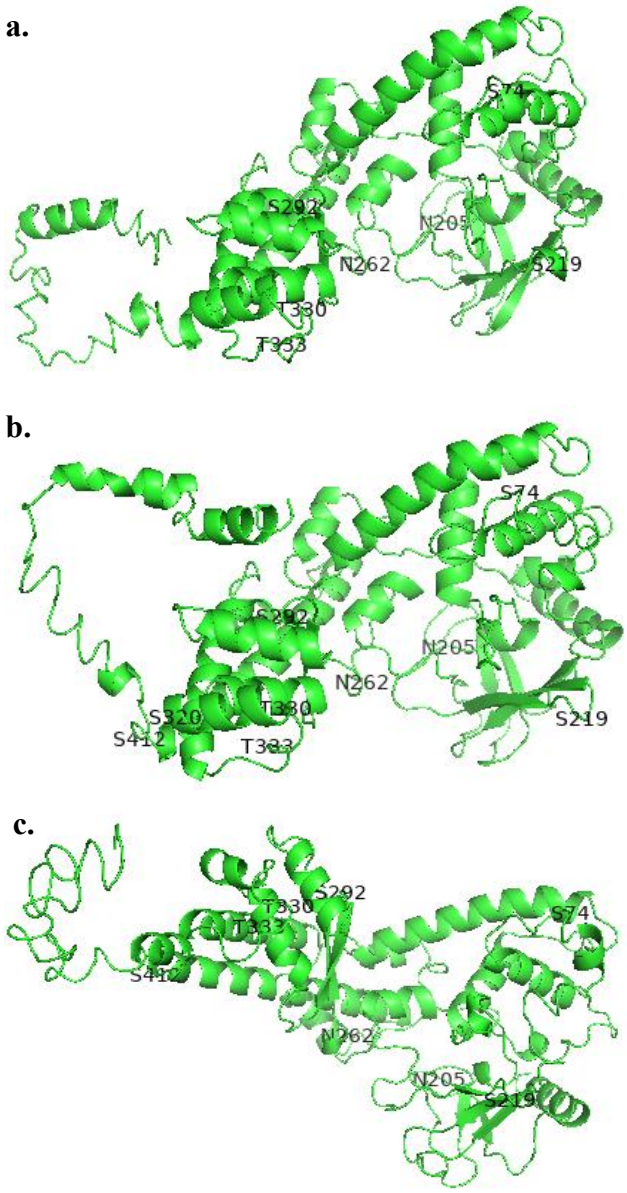


Fig. 5. SET9 Protein 3D structures with O-glycosylation and N-glycosylation residues labeled with PyMol. Glycoprotein of *N. crassa* (a), SFS strains (b), and NFS strains (c) of *S. fimicola*.

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*.

References

Ahmad, S., M.M. Gromiha and A. Sarai. 2003. RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, 19(14): 1849-1851.

Allen, H.J. and E.C. Kisailus (Eds.) 1992. *Glycoconjugates: Composition, Structure and Function*. Marcel Dekker, New York.

Arif, R., F. Akram, T. Jamil, H. Mukhtar, S.F. Lee and M. Saleem 2017. Genetic variation and its reflection on posttranslational modifications in frequency clock and mating type a-1 proteins in *Sordaria fimicola*. *Bio. Med. Res. Int.*, 2017: 1-10.

Arif, R., S.H. Bukhari, M. Ishfaq, M.G. Shahid, S.F. Lee and M. Saleem. 2019. Genetic variation and post-translational modifications of cytochrome c oxidase-1 (COX1) in different strains of *Sordaria fimicola*. *Int. J. Agric. Biol*., 21: 1055-1062.

Bennett E.P., U. Mandel, H. Clausen, T.A. Gerken, T.A. Fritz and L.A. Tabak. 2012. Control of mucin-type O-glycosylation: A classification of the polypeptide GalNAc-transferase gene family. *Glycobiol.*, 22 (6): 736-756.

Bhargava, A. and F.F. Fuentes. 2010. Mutational Dynamics of Microsatellites. *Mol. Biotech.*, 44 (3): 250-266.

Chistiakov, D., A.B. Hellemans and F.A.M. Volckaert. 2006. Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. *Aquaculture*, 255(1-4): 1-29.

Cumming, D.A. 2003. Pathways and functions of mammalian protein glycosylation. In: (Ed.): Makrides, S.C. *Gene Transfer and Expression in Mammalian Cells. Elsevier Science*, pp. 433-455.

Deshpande, N., M.R. Wilkins, N. Packer and H. Nevalainen. 2008. Protein glycosylation pathways in filamentous fungi. *Glycobiol.*, 18(8): 626-637.

Dunker, A.K., C.J. Brown, J.D. Lawson, L.M. Iakoucheva and Z. Obradovic. 2002. Intrinsic disorder and protein function. *Biochem.*, 41(21): 6573-6582.

Fakhro, K.A., M. Choi, S.M. Ware, J.W. Belmont, J.A. Towbin, R.P. Lifton, M.K. Khokha and M. Brueckner. 2011. Rare copy number variations in congenital heart disease patients identify unique genes in left-right patterning. *Proc. Natl. Acad. Sci. USA*, 108: 2915-2920.

Gill, D.J., J. Chia, J. Senewiratne and F. Bard. 2010. Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. *J. Cell Biol.*, 189: 843-858.

Goto, M. 2007. Protein O-glycosylation in fungi: diverse structures and multiple functions. *Biosci. Biotechnol. Biochem.*, 71(6): 1415-1427.

Hansen, J. E., O. Lund, J. O. Nielsen, J. E. Hansen and S. Brunak. 1996. O-GLYCBASE: a revised database of O-glycosylated proteins. *Nucl. Acids Res.*, 24: 248-252.

Helenius, A. and M. Aebi. 2004. Roles of N-linked glycans in the endoplasmic reticulum. *Ann. Rev. Biochem.*, 73:1019-1049.

Iakoucheva, L.M., P. Radivojac, C.J. Brown, T.R. O'Connor, J. G. Sikes, Z. Obradovic and A.K. Dunker. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucl. Acids Res.*, 32(3): 1037-1049.

Jamil, T., N. Sami, R. Arif, Q. Rashid and M. Saleem. 2018. H3/H4 Histone genes variations and its effect on post-translational modifications in various strains of *Sordaria fimicola*. *Int. J. Agri. Biol.*, 20: 1021-1026.

Kobata, A. 1992. Structures and functions of the sugar chains of glycoproteins. *Eur. J. Biochem*., 209: 483.

Kouskouti, A., E. Scheer, A. Staub, L. Tora and I. Talianidis. 2004. Gene-specific modulation of TAF10 function by SET9-mediated methylation. *Mol. Cell*, 14(2): 175-182.

Lee, J., J.S. Park, J.Y. Moon, K.Y. Kim and H.M. Moon. 2003. The influence of glycosylation on secretion, stability, and immunogenicity of recombinant HBV pre-S antigen synthesized in *Saccharomyces cerevisiae. Biochem. Biophys. Res. Commun.*, 303(2): 427-432.

Mitra, N., S. Sinha, T.N. Ramya and A. Surolia. 2006. N-linked oligosaccharides as outfitters for glycoprotein folding, form and function. *Trends Biochem. Sci.*, 31: 156-163.

Nevo, E. 2012. "Evolution Canyon," a potential microscale monitor of global warming across life. *Proc. Natl. Acad. Sci.*, 109: 2960-2965.

Nishioka, K., S. Chuikov, K. Sarma, H. Erdjument-Bromage, C.D. Allis, P. Tempst and D. Reinberg. 2002. Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. *Genes Dev.*, 16(4): 479-489.

Pang, C.N.I., A. Hayen and M.R. Wilkins. 2007. Surface accessibility of protein post-translational modifications. *J. Prot. Res.*, 6(5): 1833-1845.

Parodi, A.J. 2000. Role of N-oligosaccharide endoplasmic reticulum processing reactions in glycoprotein folding and degradation. *Biochem. J.*, 348(Pt1): 1-13.

Pietro, S., T.M. Fulton, J. Chunwongesm and S.D. Tanksley. 1995. Extraction of high quality DNA for genome sequencing. *Mol. Biol. Rep.*, 13: 207.

Rana, J., R. Arif, M. Ishfaq, S.F. Lee and M. Saleem. 2018. Amplification and posttranslational modification study of manganese superoxide dismutase gene in *Sordaria fimicola*. *Int. J. Agri. Biol.*, 20(9): 2059-2063.

Rini, J.M. and H. Leffler. 2010. Carbohydrate recognition and signaling. In: *Handbook of Cell Signaling*. Academic Press, pp. 85-91.

Saleem, M., B.C. Lamb and E. Nevo. 2001. Inherited differences in crossing over and gene conversion frequencies between wild strains of *Sordaria fimicola* from "Evolution Canyon". *Genetics*, 159: 1573-1593.

Schjoldager, K.T. and H. Clausen. 2012. Site-specific protein O-glycosylation modulates proprotein processing-deciphering specific functions of the large polypeptide GalNAc-transferase gene family. *Biochim. Biophys. Acta.*, 1820: 2079-2094.

Shental-Bechor, D. and Y. Levy. 2008. Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. *Proc. Natl. Acad. Sci.*, 105(24): 8256-8261.

Spiro, R.G. 2002. Protein glycosylation: Nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiol.*, 12: 43R-56R.

Uslupehlivan, M., R. Deveci and C. Ün. 2018. *In silico* investigation of the prion protein glycosylation profiles in relation to scrapie disease resistance in domestic *sheep (Ovis aries). Mol. Cell. Probes*, 42: 1-9.

Varki, A. 1993. Biological roles of oligosaccharides: All of the theories are correct. *Glycobiol.*, 3: 97-130.

Willer, T., M. C. Valero, W. Tanner, J. Cruces and S. Strahl. 2003. O-mannosyl glycans: from yeast to novel associations with human disease. *Curr. Opin. Struct. Biol.*, 13: 621-630.

(Received for publication 3 January 2020)