BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 10(8)

Molecular modeling and analysis of human and plant endo-β-N-acetyl- glucosaminidases for mutations effects on function

Shechinah Felice Choragudi[1], Ganesh Kumar Veeramachaneni[1], BV Raman[2] & Bondili JS[1]*

[1]Department of Biotechnology, K L University, Green fields, Vaddeswaram, Guntur, A.P-522502; [2]Department of Basic Sciences, Madanapalle Institute of Technology and sciences, Madanapalle, A.P-517325; Bondili JS – Email: jksingh@kluniversit; *Corresponding author

Received July 02, 2014; Accepted July 10, 2014; Published August 30, 2014

Abstract:
Endo- β-N-acetylgucosaminidases (ENGases) are the enzymes that catalyze both hydrolysis and transglycosylation reactions. It is of interest to study ENGases because of their ability to synthesize glycopeptides. Homology models of *Human, Arabidopsis thaliana* and *Sorghum* ENGases were developed and their active sites marked based on information available from Arthrobacter protophormiae (PDB ID: 3FHQ) ENGase. Further, these models were docked with the natural substrate GlcNAc-Asn and the inhibitor Man_3GlcNAc-thiazoline. The catalytic triad of Asn, Glu and Tyr (N171, E173 and Y205 of bacteria) were found to be conserved across the phyla. The crucial Y299F mutation showing 3 times higher transglycosylation activity than in wild type Endo-A is known. The hydrolytic activity remained unchanged in bacteria, while the transglycosylation activity increased. This Y to F change is found to be naturally evolved and should be attributing higher transglycosylation rates in human and Arabidopsis thaliana ENGases. Ligand interactions Ligplots revealed the interaction of amino acids with hydrophobic side chains and polar uncharged side chain amino acids. Thus, structure based molecular model-ligand interactions provide insights into the catalytic mechanism of ENGases and assist in the rational engineering of ENGases.

Keywords: ENGases, Transglycosylation, Modeling, Mutations

Background:
Endo- β -N-acetylgucosaminidases (ENGases) are enzymes which come under the class of EC 3.2.1.96 that hydrolyze the glycosidic bond of GlcNAc β -1,4 GlcNAc, N-linked sugar chains of glycoproteins and releases the N-glycan moiety leaving the end GlcNAc attached to the protein. These enzymes play a critical role in de-N-glycosylation event followed by degradation of misfolded glycoproteins in the eukaryotic cytosol **[1].** ENGases are found in wide variety of bacteria **[2–4],** fungi **[5],** plants **[6-8],** animals **[9]** and in humans **[10]**. These are key enzymes in the processing event of free oligosaccharides in the cytosol **[11].** The deglycosylation property of ENGases has been used for structural and functional studies of deglycosylation. Basing on the similarity in amino acid sequence, ENGases are categorized into glycoside hydrolase families GH18 and GH85. Besides hydrolysis, some of the GH85 family ENGases possesses transglycosylation activity, i.e., they transfer the released oligosaccharide moiety to a suitable acceptor other than water. The transglycosylation activity of ENGases has attracted much attention in recent years for the chemoenzymatic synthesis of oligosaccharides, glycopeptides and glycoproteins **[12, 13].**

In order to understand the structural and functional relationships of ENGases, we studied the different orthologous species like Humans, *Arabidopsis thaliana* and Sorghum in comparison with **[14]** the bacterial *'Arthrobacter protophormiae'* protein structure. Due to lack of crystal structures of plant and human ENGases, *Arthrobacter protophormiae ENGase* crystal structure **3FHQ** is considered for homology modeling. These structural comparisons lead to comprehend the dual-catalytic mechanisms like hydrolysis and transglycosylation of ENGases by docking with GlcNAc-Asn and Man_3GlcNAc-thiazoline ligands.

Methodology:
The protein sequence of Human (Uniprot ID: Q8NFI3, www.uniprot.org), Arabidopsis (NCBI ID: 79507164, www.ncbi.nlm.nih.gov) and Sorghum (Uniprot ID: C5YW98, www.uniprot.org) sequences were retrieved. Protein–protein Blast was used to identify the homologous sequence with three dimensional structures done against protein data bank (www.rcsb.org). The obtained PDB sequence was then aligned with Clustal x for multiple sequence alignment.

Homology Modeling and Energy Minimization
Homology models of *Human, Arabidopsis* and *sorghum* ENGase were built with Prime (3.1) module of Schrödinger Suite (Schrödinger, LLC, New York, NY). The secondary structure of these three target sequences were predicted using the SSpro program bundled with Prime. The target *Human, Arabidopsis* and *Sorghum* ENGase and template (*Arthrobacter protophormiae* ENGase) sequences were aligned using the Clustal X method of Prime, followed by manual adjustment to avoid big gaps in the secondary structure domain. The original ligand in the template structure was removed before performing homology modeling.

Using protein preparation wizard (Schrödinger suite version 9.6) the protein was minimized. Firstly, water molecules were removed from the crystallographic structure followed by addition of hydrogen atoms. All atom charges and atom types were assigned. Finally, energy minimization and refinement of the structure was done up to 0.3 Å RMSD by applying OPLS-2005 force field. The optimized target protein was employed for docking studies. Further, the structure quality of the predicted homology models for ENGases of *Human, Arabidopsis thaliana* and *Sorghum* were evaluated by using the online tools like Rampage, PDBSum and PSVS (Protein Structural Validation Suite).

Binding Site Prediction
The binding site for bacterial ENGase was determined by using the ligand interaction diagram generated by LIGPLOT. All modeled structures were superimposed on bacterial ENGase (3FHQ) using the superimposition tool of the Schrodinger Suite. Sitemap (Schrödinger, LLC) was used to recognize the plausible active sites for the modeled proteins. Based on the scores and the larger cavity, the top ranked potential active ligand binding sites for each of the modeled protein was identified and receptor grid was generated to proceed further for molecular docking.

Preparation of Ligand Molecules
GlcNAc-Asn and $Man_3GlcNAc$-thiazoline were used as ligands in this present study. LigPrep module was used to prepare the ligand. Preparation involves the generation of tautomers, 32 stereoisomers corresponding to a pH of 7±2 and protonation states. Finally, energy minimization was done using the force field OPLS2005.

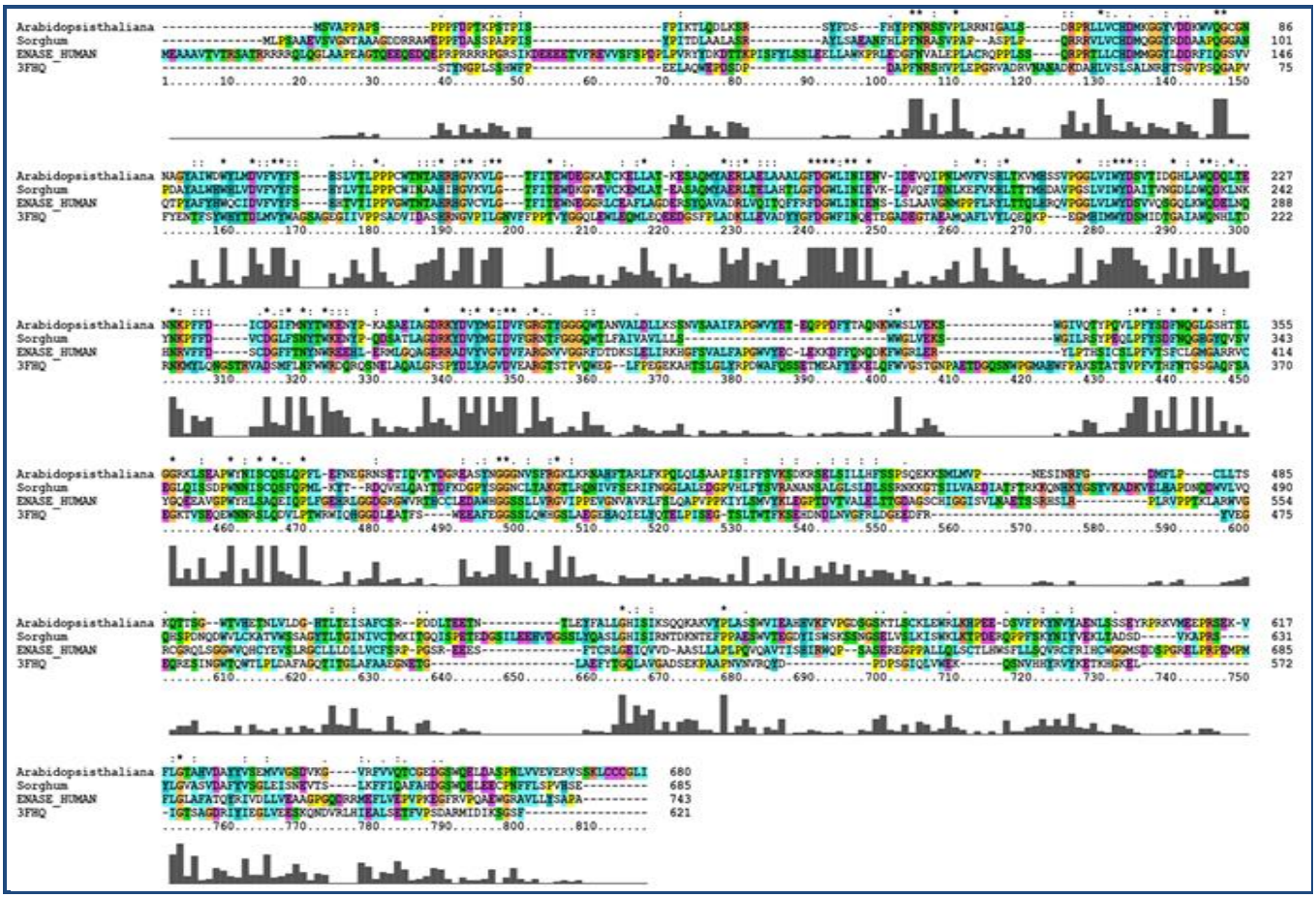

Figure 1: Depicting the Multiple Sequence Alignment (MSA) of Arthrobacter protophormiae [3FHQ], Arabidopsis thaliana, Human, Sorghum protein sequences. The MSA is developed using CLUSTAL X

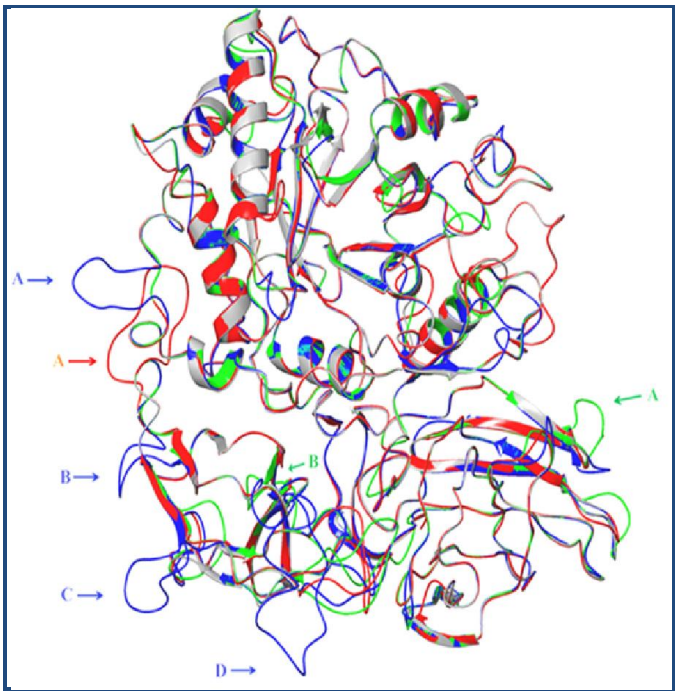

Figure 2: Shows the superimposition of the three modeled proteins Human [Green], Sorghum [Red] and Arabidopsis thaliana [Blue]. Major changes in the Human protein are depicted with green arrows: A [446-456], B [680-689], Sorghum with red arrow: A [570-581] and Arabidopsis thaliana with blue colored arrows: A [22-28], B [553-557], C [570-576] and D [584-603].

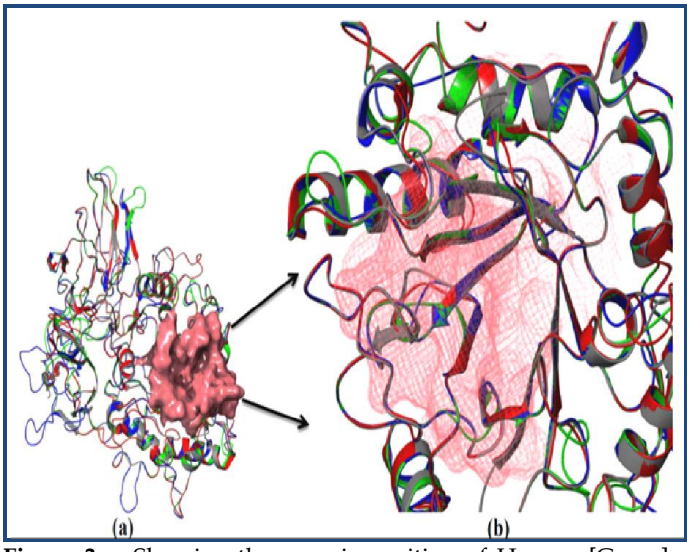

Figure 3: **a** Showing the superimposition of Human [Green], Sorghum [Red] and Arabidopsis thaliana [Blue] along with Arthrobacter protophormiae [Grey] and the active site of the GlcNAc-Asn is highlighted in pink color; **b** Zoomed version of the active site domain confirms the same homology in all the four proteins.

Molecular Docking Studies

Identified active site as predicted by sitemap in the three modeled proteins was used as a target for docking. Receptor grid generation was performed, being an essential step as the docking protocol is a grid based docking. The grid box was generated at centroid of the residues in the active site of the modeled structures continuing with the default parameters. The glide extra precision (XP) protocol (Glide Tool, Schrodinger suite) was employed for docking the prepared ligands with the modeled proteins. Ligand interaction diagram was used to understand the interactions between the ligands and modeled ENGase proteins.

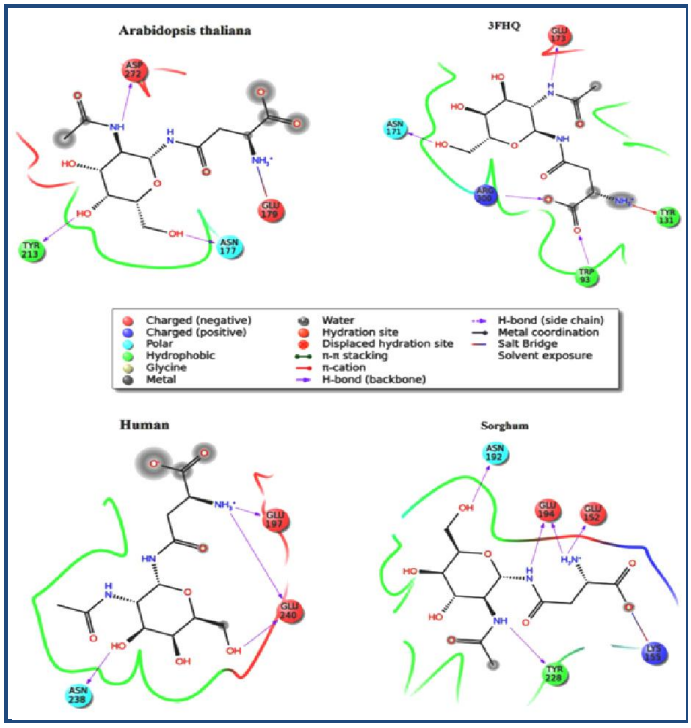

Figure 4: Lig plots highlighting the interaction of GlcNAc-Asn with active site amino acids including the catalytic triad.

Results & discussion:

Protein 3D structure prediction from amino acid sequence turns accessible if relatively similar homologous protein structure is crystallized. PSI –Blast search, an online search engine was used to identify the similar type of proteins. *Arthrobacter protophormiae* Endo-Beta-N-acetylglucosaminidase is chosen as template for modeling studies. Blast results showed 25 % similarity with *Human, Arabidopsis thaliana* and 24% with *Sorghum* protein sequences. The crystal structure 3FHQ with a resolution of 2.45 ºA was retrieved from protein data bank (www.rcsb.org). The multiple sequence alignment (MSA) of the four ENGases was done using Clustal X as shown in the **Figure 1.** The modeled 3D structures were generated using the prime module and further refined by using the protein preparation wizard. The models were evaluated by using the Procheck. Ramchandran plot for *Arabidopsis thaliana* showed 73.7% of amino acids in most favorable region, 23% in allowed regions and 3.3% in disallowed region. Human ENGase model showed 77.8% in most favorable region, 20.4% in allowed regions and only 1.8% of amino acids in disallowed regions. Sorghum ENGase protein model showed 76.7% of amino acids in most favorable regions, 20.5% in allowed regions and only 2.8% in disallowed regions. Majority of amino acids are in allowed regions of the Ramchandran plot for the three models generated and reinstate that these homology models are of good quality and dependable. Super

imposition of the modeled proteins along with bacterial counterpart showed helix broken to helix-loop-helix confirmation at 680 to 689 amino acids and extended loops with 446-456 amino acids in Human protein. Four extended loops were also seen in *Arabidopsis thaliana* structure ranging 22-28, 553-557, 570-576, 584-603 amino acids. 570-581 amino acids extended loop was also seen in Sorghum as depicted in **Figure 2.** The Glide XP based molecular docking was done with all the three modeled proteins with GlcNAc-Asn and Man_3GlcNAc-thiazoline as ligands. All the four ENGases were docked with GlcNAc-Asn and Man_3GlcNAc-thiazoline inside their binding pockets generated by the sitemap. The binding cavity site position is common for all the four ENGases but they differ in amino acids residues number as depicted in **Figure 3**. Based on the nature of amino acids, all the four cavities of the proteins contain majority of aromatic and acidic type of amino acids in their cavity. This suggests that cavities of all the three modeled modeled structures are in line with crystallized protein and could be attributed with similar functionality.

When compared with the bacterial (3FHQ) catalytic triad, N171, E173 and Y205 are found to be conserved in Human and plant ENGase proteins as shown in Table 1. Critical catalytic residues N171 (Bacteria), N238 (Human), N177 (*A.thaliana*) and N192 (Sorghum) were showing back bone hydrogen bonding with GlcNAc-Asn as represented in **Figure 4**. E179 (*A.thaliana*) showed a salt bridge while E173 (Bacteria), E240 (Human) and E194 (Sorghum) were hydrogen bonded with GlcNAc-Asn. Y213 (*A.thaliana*) and Y228 (Sorghum) were also found connected with back bone hydrogen bonds to GlcNAc-Asn. Tryptophan (W) at 216 and 244 positions of bacterial enzyme were reported to be crucial for transglycosylation activity and are speculated as gate keepers for the active site during transglycosylation **[14].** W (Trp) 216 of Bacteria is found to be conserved in Humans and Plant ENGases where as W (Trp) at 244 is replaced with N (Asn) at 305 positions in humans and T (Thr) in plants at 244 and 259 positions as tabulated in **Table 1 (see supplementary material).**

Considering the previous mutagenic studies of Endo M **[15]** and the structures of binary complexes of Endo A and mutational studies, Y299 F showed 3 times higher transglycosylation activity than wild type Endo A. The hydrolytic activity also remained unchanged **[14].** Interestingly, Y299 of bacteria is not found to be conserved and is replaced with F(PHE) at 361 in humans and at 300 in *A.thaliana.* The same tyrosine is found to be completely deleted in Sorghum protein sequence. As predicted, the loss of hydroxyl group in the Y to F mutation might help in the release of the product faster and there by resulting in increased transglycosylation rate in human and plant ENGases. This could also be considered as a natural selection criteria of evolution for elevated rates of enzyme activity in higher phyla. Further, the docking results of Man_3GlcNAc-thiazoline with *Arabidopsis thaliana* model showed hydrogen bonding with Tyr 213, Glu 179, Asp 139, Ser 105 amino acids present in the binding pocket. Bacterial ENGase (3FHQ) with Man_3GlcNAc-thiazoline complex was maintained through Tyr 299 and Gly 95. Binding between Human ENGase model and Man_3GlcNAc-thiazoline was maintained through Asp 233, Glu 240, Asn 308, Lys 284 Trp 285 and Asp 287 amino acids. Man_3GlcNAc-thiazoline bound to Lys 242, Glu 265, Trp 263, Lys 264, Trp 239 and Thr 262 amino acids of Sorghum model.

Conclusion:

ENGases are inherently a class of hydrolases; the hydrolytic activity in general is relatively higher in comparison to their transglycosylation activities with natural N-glycan's as substrates. The major concern for the chemo-enzymatic approach is the product of hydrolysis, where the product thus formed tends to be the substrate for the enzyme. Lack of crystallized structures was a challenge for ENGase engineering and in the understanding of the structural basis for hydrolysis and transglycosylation activities. This study provides homology models of human and plant ENGases and depicts the mutations for enhancing the transglycosylation activities. The human and plant enzymes are excellent source for engineering it by selectively modifying the hydrolytic catalytic triad.

Acknowledgment:

The authors thank the management of K L University for providing the infrastructure of **Schrödinger suite**. The authors also acknowledge the financial support of DST SR/FT/LS-033/2009 for the execution of the project.

References:

[1] Spiro RG *et al. Cell Mol Life Sci*. 2004 **61**: 1025 [PMID: 15112051]
[2] Bourgerie S *et al. J Bacteriol*. 1994 **176**: 6170 [PMID: 7928985]
[3] Takegawa K *et al. Appl Environ Microbiol*. 1989 **55**: 3107 [PMID: 16348072]
[4] Fujita K *et al. Biosci Biotechnol Biochem* 2004 **68**: 1059 [PMID: 15170110]
[5] Kadowaki S *et al. Agric Biol Chem* 1990 **54**: 97 [PMID: 1368528]
[6] Li SC *et al. Biochim Biophys Acta*. 1981 **660**: 278 [PMID: 6793075]
[7] Kimura Y *et al. Biochim Biophys Acta*. 1998 **1381**: 27 [PMID: 9659369]
[8] Kimura Y *et al. Biosci Biotechnol Biochem* 1998. **62**: 253 [PMID: 9532781]
[9] Overdijk B *et al. Biochim Biophys Acta*. 1981 **659**: 255 [PMID: 7260095]
[10] Ito K *et al. J Biol Chem* 1993 **268**: 16074 [PMID: 8340428]
[11] Suzuki T *et al. J Biol chem*. 1998 **273**: 10083 [PMID: 9553052]
[12] Wang LX *et al. Carbohydr Res*. 2008 **343**: 1509 [PMID: 18405887]
[13] Wang LX *et al. Trends Glycosci Glycotechnol*. 2011 **23**: 33
[14] Yin J *et al. PLoS One*. 2009 **4**: e4658 [PMID: 19252736]
[15] Umekawa M *et al. J Biol Chem*. 2008 **283**: 4469 [PMID: 18096701]

Edited by P Kangueane

Citation: **Choragudi *et al.*** Bioinformation 10(8): 507-511 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: Mutations and their respective positions are mentioned. Serial No 1-3 [bold] depicts the conserved catalytic triad of the bacterial and their respective positions in *Human, Arabidopsis thaliana, Sorghum* protein sequences. Serial No. 4-9 represents the amino acids which are interacting with the substrate. Serial No.8 and 9[bold] represents the major mutational changes responsible for the increased transglycosylation activity.

No	**Bacteria *Arthrobacter protophormiae* (621 amino acids)**		**Human (743 amino acids)**		***Arabidopsis thaliana* (680 amino acids)**		**Sorghum (685 amino acids)**	
	Amino acid	**Position**	**Position**	**Amino acid**	**Amino acid**	**Position**	**Amino acid**	**Position**
1	**N (ASN)**	**171**	**N (ASN)**	**238**	**N (ASN)**	**177**	**N (ASN)**	**192**
2	**E (GLU)**	**173**	**E (GLU)**	**240**	**E (GLU)**	**179**	**E (GLU)**	**194**
3	**Y (TYR)**	**205**	**Y (TYR)**	**274**	**Y (TYR)**	**213**	**Y (TYR)**	**228**
4	W (TRP)	93	F (PHE)	164	F (PHE)	104	F (PHE)	119
5	F (PHE)	125	DELETED	-	DELETED	-	DELETED	-
6	W (TRP)	216	W (TRP)	282	W (TRP)	221	W (TRP)	236
7	F (PHE)	243	Y (TYR)	304	Y (TYR)	243	Y (TYR)	258
8	**W (TRP)**	**244**	**N (ASN)**	**305**	**T (THR)**	**244**	**T (THR)**	**259**
9	**Y (TYR)**	**299**	**F (PHE)**	**361**	**F (PHE)**	**300**	**DELETED**	**-**