



***NMR and molecular recognition of compounds  
related to the immune system***

**Helena Maria Nobre Coelho**

**Doctoral Thesis**

**2019**

**Doctoral Thesis Co-Tutelage between the University of the Basque Country  
UPV/EHU and the New University of Lisbon.**

Supervisors:

Dr. Jesús Jiménez Barbero (CIC bioGUNE)

Dr. Filipa Marcelo (UCIBIO)

University Tutor: Dr. Esther Lete (UPV/EHU)

This doctoral thesis was performed at Center for Cooperative Research in Biosciences (CICbioGUNE) and at Applied Molecular Biosciences Unit (UCIBIO).





*To my family*



## **Acknowledgments**

The work herein described was performed at the Center for Cooperative Research in Biosciences (CICbioGUNE) and at Applied Molecular Biosciences Unit (UCIBIO), under the supervision of Prof. Dr. Jesús Jiménez-Barbero and Dr. Filipa Marcelo.

Firstly, I would like to thank my PhD supervisors: Prof. Dr. Jesús Jiménez-Barbero and Dr. Filipa Marcelo for giving me the opportunity to participate in this research project to develop my doctoral thesis. I am extremely thankful for their continuing support and availability during these years of thesis, that helped me to grow as a person and professional.

During these years I also completed three secondment periods abroad also go my acknowledgements: 1) Center for Biological Research of the Spanish National Research Council (CIB-CSIC) (Madrid, Spain), under the supervision of Dr. Sonsoles Martín Santamaría and 2) Università degli Studi di Milano-Bicocca (UNIMIB) (Milan, Italy), under the supervision of Prof. Francesco Peri and 3) Lofarma, S.p.A, under supervision Dr. Gianni Mistrello.

In terms of collaborations, I would like to thank to all who participated to the works presented in this Thesis.

For the financial support, I am grateful to the European Union's Horizon 2020 research and innovation programme that financed this project (ETN TOLLerant project, Marie Skłodowska-Curie grant agreement No 642157) and the FCT-Portugal to IF project (IF/00780/2015) that supported the recent times.

To my laboratory colleagues from Chemical Glycobiology Lab (CIC bioGUNE) and (Bio)molecular Structure and Interactions by NMR Lab (UCIBIO).

To my family and friends.

Thank you to all! Muchas Gracias a todos! Muito Obrigada a todos!



## **Contents**

<b>Abbreviations</b>	<b>i</b>
<b>Resumen</b>	<b>iii</b>
<b>Resumo</b>	<b>ix</b>
<b>Abstract</b>	<b>xiii</b>
<b>Chapter 1 - General Introduction</b>	<b>1</b>
<b><i>1.1 General Introduction</i></b>	<b><i>1</i></b>
1.1.1 Principles of NMR Spectroscopy	1
1.1.1.1 Chemical shift	3
1.1.1.2 Spin-spin couplings	3
1.1.1.3 Spin relaxation	4
1.1.2 The Nuclear Overhauser Effect (NOE).	6
1.1.3 Diffusion ordered spectroscopy	8
1.1.4 NMR methods for the study of Protein-Ligand interactions:	8
1.1.4.1 Ligand-detected methods	10
1.1.4.2 Receptor-detected methods: Chemical Shift Perturbations	16
1.1.4.3 Specific isotopic labeling to study protein by NMR	19
1.1.5 Carbohydrates and Carbohydrate-Protein Interactions	21
1.1.5.1 Structure and Conformation	22
1.1.5.2 Carbohydrate-protein interactions	25
1.1.6 References	28
<b><i>1.2 Goals</i></b>	<b><i>33</i></b>
<b>Chapter 2 - Deciphering GalNAc O-glycosylation: From structure to function in human health &amp; disease</b>	<b>35</b>
2.1 Introduction	37
2.1.1 GalNAc-Transferases	38
2.1.1.1 Structure	40

2.1.1.2 Preferences and Specificity	43
<b>2.2 <i>GalNAc-Ts glycosylation follow an induced-fit catalytic mechanism.</i></b>	<b>47</b>
2.2.1 Introduction	49
2.2.2 Results and Discussion	51
2.2.2.1 Stability and Binding assays	51
2.2.2.2 X-Ray Crystallography & MD Simulations	53
2.2.2.3 <sup>19</sup> F labelling of the WT GalNAc-T2 and F104S mutant for NMR experiments.	57
2.2.2.4 <sup>19</sup> F-NMR experiments	60
2.2.3 Conclusions	64
2.2.4 Supporting Information	66
2.2.4.1 <sup>19</sup> F-NMR spectra	66
<b>2.3 <i>Deciphering the Mechanism of Long and Short Distance-Glycosylation of GalNAc-Ts</i></b>	<b>69</b>
2.3.1 Introduction	71
2.3.1 Results and Discussion	73
2.3.1.1 Long distance-glycosylation	73
2.3.1.1 Short distance-glycosylation	83
2.3.2 Conclusion	95
2.3.3 Supporting Information	96
<b>2.4 <i>O-glycosylation of mucin MUC1 by GalNAc-Ts</i></b>	<b>103</b>
2.4.1 Introduction	105
2.4.1 Results and Discussion	109
2.4.1.1 Design, Expression, Purification and NMR characterization of isotopic labeled MUC1 with four TR domains	109
2.4.1.2 Monitoring MUC1-4TR glycosylation by GalNAc-Ts using NMR spectroscopy. The role of the lectin domain.	110
2.4.1.3 Conformation of the glycosylated MUC1-4TR products modulates glycosylation preferences of GalNAc-Ts	137



2.4.2	Conclusions	147
2.4.3	Supporting information	151
<b>2.5</b>	<b><i>Methods</i></b>	<b>161</b>
2.5.1	NMR experiments	161
2.5.1.1	Assignment of (glyco)peptide	161
2.5.1.2	Saturation Transfer Difference (STD)	161
2.5.1.3	<sup>19</sup> F-NMR experiments	162
2.5.1.4	NMR Glycosylation Assay	163
2.5.1.5	Percentage of glycosylation of individual residues	164
2.5.1.6	Combined chemical shift perturbation (CSP)	165
2.5.2	Overexpression and Purification of <sup>15</sup> N-MUC1-4TR	165
2.5.2.1	Purification of the glycosylated products	168
2.5.3	Mass spectrometry: MALDI-TOF/TOF	168
2.5.3.1	Sample preparation	168
<b>2.6</b>	<b><i>References</i></b>	<b>170</b>
<b>Chapter 3 - The conformational and interaction features of glycolipid with TOLL-like receptors and accessory proteins</b>		<b>177</b>
<b>3.1</b>	<b><i>Introduction</i></b>	<b>179</b>
<b>3.2</b>	<b><i>The interaction of Lipid A mimics with antimicrobial peptides</i></b>	<b>183</b>
3.2.1	AMPs enhances FP7 antagonist activity in HEK-Blue hTLR4 cells	183
3.2.2	LL-37 (AMP6) enhances FP7 antagonist activity in human PBMC	187
3.2.3	NMR and TEM analysis of glycolipid/peptide interaction	188
3.2.3.1	NMR analysis from the AMP point-of-view	188
3.2.3.2	NMR analysis from the FP7 point-of-view	193
<b>3.3</b>	<b><i>Design of new TLR4 antagonists based on FP7</i></b>	<b>200</b>
<b>3.4</b>	<b><i>Characterization of natural LPSs/MD-2 interaction</i></b>	<b>203</b>
3.4.1	LPS from <i>Bradyrhizobium BTAi-1 Δshc</i>	204
3.4.1.1	NMR experiments	205

3.4.1.2 Cryo-Electron Microscopy	207
3.4.2 LPS from <i>Acetobacter pasteurianus</i>	207
3.4.2.1 NMR experiments	208
3.4.2.2 Cryo-Electron Microscopy	211
<b>3.5 Conclusions</b>	<b>212</b>
<b>3.6 Methods</b>	<b>214</b>
3.6.1 NMR experiments	214
3.6.2 Expression and Purification of MD-2 protein	215
3.6.3 Transmission Electron Microscopy	216
3.7 References	216
<b>Chapter 4 – Final Remarks</b>	<b>221</b>
<i>4.1 Scientific publications during this dissertation</i>	<i>225</i>
<i>4.2 Contribution to congresses during this dissertation</i>	<i>226</i>

## Abbreviations

<b>AMP</b>	Anti-Microbial Peptides
<b>CD14</b>	Cluster Differentiation antigen 14
<b>CHO-K1</b>	Chinese hamster ovary - K1
<b>CLD</b>	Carbohydrate lectin domain
<b>CMC</b>	Critical Micellar Concentration
<b>Cryo-TEM</b>	Cryogenic Transmission Electron Microscopy
<b>CSI</b>	Chemical shift index
<b>CSP</b>	Chemical Shift Perturbation
<b>DCs</b>	Dendritic Cells
<b>DOSY</b>	Diffusion Order Spectroscopy
<b><i>E. Coli</i></b>	<i>Escherichia coli</i>
<b>ER</b>	Endoplasmic Reticulum
<b>GalNAc-Ts</b>	GalNAc-Transferases
<b>GTs</b>	Glycosyltransferases
<b>HDL-C</b>	High-density lipoprotein cholesterol
<b>HSQC</b>	Heteronuclear Single-Quantum Coherence
<b>IRF3</b>	Interferon Regulatory Factor 3
<b>K<sub>d</sub></b>	Dissociation constant
<b>K<sub>m</sub></b>	Michaelis constant
<b>LBP</b>	Lipopolysaccharide Binding Protein
<b>LOS</b>	Lipooligosaccharide
<b>LPS</b>	Lipopolysaccharide
<b>LRR</b>	Leucine-Rich Repeat
<b>MD</b>	Molecular Dynamics
<b>MD-2</b>	Myeloid Differentiation protein
<b>MUC1</b>	Mucin-1
<b>MUC1-4TR</b>	Mucin-1 with four tandem repeats
<b>NF-κB</b>	Nuclear Factor Kappa B
<b>NMR</b>	Nuclear magnetic resonance

<b>NOE</b>	Nuclear Overhauser effect
<b>NOESY</b>	Nuclear Overhauser effect spectroscopy
<b>PAMPs</b>	Pathogen-associated molecular patterns
<b>PBS</b>	Phosphate Buffered Saline
<b>PHA</b>	Plant Lectin Phytohemagglutinin
<b>PLTP</b>	Phospholipid transfer protein
<b>PRRs</b>	Pattern Recognition Receptors
<b>RMSD</b>	Root-mean-square-deviation
<b>SAXS</b>	Small-angle X-ray scattering
<b>STD</b>	Saturation Transfer Difference
<b>TEM</b>	Transmission Electron Microscopy
<b>TEV</b>	Tobacco Etch Virus
<b>TLR4</b>	Toll-Like Receptor 4
<b>TLRs</b>	Toll-like receptors
<b>TOCSY</b>	Total correlation spectroscopy
<b>TR</b>	Tandem repeat
<b>TSP</b>	2,2,3,3-tetradeutero-3-trimethylsilylpropionic acid
<b>UDP</b>	Uridine diphosphate
<b>UDP-GalNAc</b>	Uridine diphosphate N-acetylgalactosamine
<b>V<sub>max</sub></b>	Maximum rate
<b>WT</b>	Wild-type
<b><math>\Delta\delta_{\text{comb}}</math></b>	$\Delta\delta$ $^1\text{H}/^{15}\text{N}$ combined chemical shift

## **Resumen**

Las macromoléculas biológicas: las proteínas, los ácidos nucleicos y los hidratos de carbono son los motores principales de la célula viva. La esencia de la vida está regulada por las interacciones que se dan entre estas entidades, con distintos niveles de complejidad. El conocimiento de su estructura tridimensional, su dinámica y la manera en que se relacionan entre ellas y con otras moléculas más pequeñas es imprescindible para entender en profundidad el funcionamiento de la compleja maquinaria celular.

Los hidratos de carbono son de las moléculas más variables y complejas de los sistemas biológicos. La glicosilación de proteínas y lípidos tiene la capacidad inigualable de generar una amplia gama de estructuras diferentes. El conocimiento de la comunicación intermolecular, a escala atómica y molecular, entre los hidratos de carbono y sus receptores debería llevar al entendimiento y modulación de las señales biológicas en eventos fisiológicos y patológicos.

Existen varias herramientas para obtener información a escala atómica de la comunicación de este tipo de moléculas: la espectroscopía de Resonancia Magnética Nuclear (RMN), la difracción de rayos X y la crio-microscopía electrónica (cryo-EM). También coexisten una legión de técnicas que proporcionan información estructural menos detallada, como otras espectroscopías (IR, UV, Raman), la microscopía electrónica (EM), o los métodos basados en transferencia de energía de resonancia (FRET).

En este trabajo, empleamos la espectroscopia de Resonancia Magnética Nuclear (RMN) para obtener información, a distintas escalas, sobre diferentes procesos de reconocimiento molecular entre hidratos de carbono y proteínas con interés biomédico. En particular, nos hemos centrado en cáncer (mucina-1 (MUC1)) estudiando el mecanismo de glicosilación de mucinas por GalNAc-transferasas (GalNAc-Ts) y en infecciones bacterianas (receptor Toll-like 4 (TLR4)).

## Capítulo 1

En la introducción de esta tesis he proporcionado una visión general de las bases de RMN, centrada en la descripción de los métodos de RMN que se usan para seguir eventos de reconocimiento molecular.

El reconocimiento molecular por RMN puede hacerse del punto de vista del ligando y/o del receptor. Los experimentos de RMN enfocados en la observación de las señales del ligando revelan información clave sobre el epítipo de unión del ligando. Uno de los experimentos de RMN más relevantes dentro de esta clase es el experimento de diferencia de transferencia de saturación (STD-NMR) que se basa en la transferencia de magnetización entre el receptor, que es objeto de irradiación selectiva, y aquellos ligandos que, en exceso y en régimen de intercambio rápido, se unen a él.

Entre los métodos basados en la observación de las señales del receptor, regularmente proteínas, podemos destacar los experimentos basados en la correlación heteronuclear de cuanto simple  $^1\text{H}/^{15}\text{N}$  ( $^1\text{H}/^{15}\text{N}$ -HSQC). El HSQC proporciona información a nivel atómico para definir con precisión los aminoácidos del sitio de unión.

### Objetivos

El objetivo general de la tesis es aplicar métodos de RMN basados en ligandos y receptores para estudiar la interacción de una variedad de hidratos de carbono con receptores de interés biológico y / o biomédico. Por lo tanto, un objetivo específico ha sido avanzar en la comprensión de las características estructurales que gobiernan las interacciones de los hidratos de carbono así como determinar de sus epítopos de reconocimiento.

Los objetivos específicos se han centrado en:

- Comprender el mecanismo de acción de GalNAc-Ts.
- Evaluar las interacciones de diferentes moléculas naturales y sintéticas con péptidos antimicrobianos y la proteína MD-2, en el contexto de la comprensión de la inmunidad innata relacionada con TLR4.

## Capítulo 2

La gran familia de GalNAc-Ts es responsable de una modificación postraduccional de muchas proteínas de la superficie celular. Existen cambios de expresión de GalNAc-Ts en las células en el caso de procesos tumorales. Así, el ajuste fino de la expresión de GalNAc-Ts regula la glicosilación con *O*-GalNAc en ciertas proteínas. Desde esta perspectiva, y usando una combinación de métodos de RMN y de modelado molecular, hemos analizado la especificidad y dinámica de diversas GalNAc-Ts en el proceso de *O*-glicosilación para desentrañar el mecanismo de acción de estas enzimas y definir los determinantes moleculares para el reconocimiento de sustrato.

Así, hemos descubierto que el reconocimiento de GalNAc-Ts sigue un mecanismo de ajuste inducido en el que el UDP-GalNAc es absolutamente necesario. También hemos determinado las bases moleculares de las preferencias de GalNAc-glicosilación a largo y corto rango para la GalNAc-T4.

Además, hemos demostrado que las enzimas GalNAc-T2 -T3 y -T4 tienen un proceso de glicosilación altamente ordenado, en que todas las repeticiones en tándem se glicosilan de una manera gradual. La acción combinada del dominio de la lectina y el dominio catalítico de esas enzimas es esencial para conseguir modificar todos los lugares de glicosilación de MUC1.

Estos estudios son un paso importante para conseguir el diseño racional de inhibidores que sirvan para regular la expresión de GalNAc-Ts en enfermedades, especialmente en cáncer.

### Capítulo 3

El otro apartado importante de esta tesis se ha centrado en el receptor Toll-like 4 (TLR4). TLR4 se expresa en la superficie de las células inmunitarias y reconoce específicamente endotoxinas de las bacterias, o sea, lipopolisacáridos (LPS) o lipooligosacáridos (LOS), el núcleo molecular de las paredes de las bacterias gram-negativas.

La modulación de receptores de inmunidad innata por agonistas y/o antagonistas con pequeñas moléculas sintéticas permiten controlar la actividad biológica del TLR4. Estas moléculas representan una herramienta poderosa para estudiar TLR4 y son de gran interés farmacológico como agentes antisépticos y antiinflamatorios (antagonistas) o como adyuvantes de vacunas (agonistas). El conocimiento de los aspectos moleculares del reconocimiento de LPS por CD14 (cluster de diferenciación 14) y por el complejo TLR4/MD-2 (proteína de diferenciación mieloide 2) es esencial para comprender las diferentes respuestas mediadas por TLR4 a diferentes variantes de LPS o a moléculas sintéticas.

En esta tesis nos involucramos en el estudio de estos eventos de reconocimiento molecular y caracterización de las estructuras supramoleculares de estas moléculas naturales y sintéticas mayoritariamente por métodos RMN y de microscopía electrónica. En particular, caracterizamos la interacción entre FP7 (antagonista sintético de TLR4) y péptidos antimicrobianos (AMPs) así como la agregación de diferentes glicolípidos, naturales y miméticos, por técnicas de RMN y TEM (microscopía electrónica de transmisión).



A lo largo de esta Tesis, los aspectos formativos han incluido el uso de experimentos de RMN para diseccionar los detalles moleculares de estos procesos de reconocimiento biomolecular. Así, se han empleado métodos de RMN basados en el ligando y en el receptor, en combinación con una variedad de técnicas experimentales y teóricas. Se han usado multitud de herramientas biofísicas, bioquímicas, de biología molecular, así como diversas metodologías computacionales para complementar los datos de RMN y así obtener una descripción más completa de los eventos de interacción.



## Resumo

Os hidratos de carbono estão entre as moléculas mais variadas e complexas dos sistemas biológicos. O processo de glicosilação é a modificação mais complexa de proteínas e lípidos, com uma capacidade inigualável de gerar uma ampla gama de estruturas diferentes. O reconhecimento molecular de hidratos de carbono por recetores específicos traduz-se em sinais biológicos em eventos fisiológicos e patológicos.

Neste trabalho, aplicamos a espectroscopia de Ressonância Magnética Nuclear (RMN) para obter informações estruturais de diferentes processos de reconhecimento molecular entre hidratos de carbono e diversas proteínas com relevância biomédica. Com principal destaque nos principais aspetos do mecanismo de glicosilação da proteína MUC1 pela família de enzimas GalNAc-Transferases (GalNAc-Ts). Os alvos estudados ao longo desta tese têm impacto no cancro (mucina-1 (MUC1)) e infeções bacterianas (recetor Toll-like 4 (TLR4)).

### Capítulo 1

A introdução desta tese contém uma visão geral das bases de RMN, no entanto, focou-se sobretudo em descrever as metodologias de RMN aplicadas ao estudo de eventos de reconhecimento molecular.

### Objetivo

O objetivo global desta tese foi avançar na compreensão das características estruturais que conduzem as interações de hidratos de carbono e sobretudo na determinação dos seus epítomos de reconhecimento.

Como objetivos específicos propusemo-nos a:

- Compreender o mecanismo de glicosilação orquestrado pelas GalNAc-Ts.

- Avaliar as interações de diferentes moléculas naturais e sintéticas com peptídeos antimicrobianos e a proteína MD-2, no contexto da compreensão da imunidade inata relacionada ao TLR4.

## Capítulo 2

A grande família das enzimas GalNAc-Ts é responsável por iniciar a modificação de pós-tradução de muitas proteínas da superfície celular. Alterações na expressão das GalNAc-Ts ocorrem em eventos de neoplasia celular, conduzindo a uma glicosilação aberrante das proteínas. Assim, o ajuste fino da expressão das GalNAc-Ts poderá permitir regular o processo de *O*-glicosilação. Nesta perspectiva, estudámos as especificidades e a dinâmica do mecanismo das GalNAc-Ts aquando a *O*-glicosilação por métodos de RMN, em conjunto com estudos de modelação molecular, para desvendar o mecanismo de ação de GalNAc-Ts e definir os determinantes moleculares necessários ao reconhecimento do substrato por essas enzimas.

Neste trabalho, fornecemos evidências de que o processo de reconhecimento de GalNAc-Ts segue um mecanismo de ajuste induzido para o qual o UDP-GalNAc é absolutamente necessário.

Além do mais a base molecular para as preferências de glicosilação a longa e curta distância foram identificadas. O processo de glicosilação a longa distância é orientado por um *linker* flexível que fornece capacidade de rotação ao domínio de lectina; enquanto para a glicosilação a curta distância existe no domínio catalítico um sítio de ligação que é responsável pelo reconhecimento do resíduo de GalNAc adjacente ao sítio a glicosilar.

Este estudo também fornece a evidência de que as enzimas GalNAc-T2 - T3 e -T4 tem um processo de glicosilação altamente ordenado, em que todas as repetições em tandem da proteína MUC1 são totalmente *O*-glicosiladas, de um modo gradual. Dentro deste trabalho, demonstramos evidências de que é muito importante a ação concertada entre o domínio de lectina e o domínio catalítico destas enzimas para uma catálise eficiente de todos os locais de glicosilação da proteína MUC1.

### Capítulo 3

O outro tópico desta tese é o recetor Toll-like 4 (TLR4), o qual é expresso na superfície das células imunes inatas e reconhece especificamente endotoxinas bacterianas. Em particular este recetor reconhece o lipopolissacarídeo (LPS) ou seu lipoligossacarídeo de versão incompleta (LOS), os principais componentes moleculares das paredes celulares de bactérias gram-negativas. A modulação dos recetores da imunidade inata por agonistas e/ou antagonistas, por exemplo através de pequenas moléculas sintéticas, permite controlar a atividade biológica do recetor TLR4. Sendo assim estas pequenas moléculas sintéticas representam uma ferramenta poderosa para estudar o sistema de TLR4 e são de grande interesse farmacológico como agentes antissépticos e anti-inflamatórios (antagonistas) ou como adjuvantes de vacinas (agonistas). O conhecimento dos aspetos moleculares do reconhecimento de LPS por CD14 (cluster de diferenciação 14) e pelo complexo TLR4/MD-2 (recetor Toll-like 4/proteína de diferenciação mielóide 2) é essencial para compreender as diferentes respostas mediadas pelo TLR4 a diferentes variantes de LPS / lípido A ou pequenas moléculas sintéticas.

Nesta tese dedicamo-nos ao estudo destes eventos de reconhecimento molecular e à caracterização das estruturas supramoleculares destas moléculas naturais e sintéticas maioritariamente usando técnicas de RMN e de microscopia eletrónica. Em particular, caracterizámos a interação entre FP7 (antagonista sintético de TLR4) e peptídeos antimicrobianos (AMPs) bem como

a agregação de diferentes glicolipídios, naturais e miméticos, por técnicas de RMN e TEM (microscopia eletrônica de transmissão).

Ao longo desta Tese, a RMN foi a técnica central para elucidar os detalhes moleculares, estruturais e mecanísticos, dos processos de reconhecimento biomolecular em estudo.

## **Abstract**

Glycans are among the most varied and complex molecules in biological systems. The glycosylation process is the most complex and widespread modification of proteins and lipids, with an unsurpassed capacity to generate a wide array of different structures. Recognition of glycans by specific receptors translates the glycome into unprecedented biological signals in physiological and pathological events.

In this work, we applied Nuclear Magnetic Resonance (NMR) spectroscopy to gain structural insights of different molecular recognition processes between glycans and diverse proteins with biomedical relevance. As key highlight, key aspects of the glycosylation mechanism of MUC1 by GalNAc-Transferases (GalNAc-Ts) have been unraveled, paying special attention to the site-specificity of the process. Furthermore, our targets have an important impact in disease, particularly, in cancer (mucin-1) and in bacterial infections (Toll-like receptor 4).

Throughout this Thesis, state-of-the-art NMR experiments have been applied to dissect the molecular details of these biomolecular recognition processes. Thus, ligand-based and receptor-based methods have been used, in combination with a variety of additional experimental and theoretical techniques. Thus, biophysical, biochemical, molecular biology and computational methodologies have complemented the NMR data in order to attain a complete description of the interaction events.





# *Chapter*

# **1**

*General Introduction*



## 1.1 General Introduction

Nuclear magnetic resonance spectroscopy (NMR spectroscopy), based on the discovery and development of Nuclear Magnetic Resonance by the end of 1930s-beginning of 1940s,<sup>1,2</sup> consists in a powerful and widely recognized research technique that makes use of the interaction between nuclear spins ( $I \neq 0$ ) and electromagnetic radio frequency (RF) pulses to permit the “visualization” and consequent study of molecules at atomic scale.<sup>3</sup> This development has been awarded with three Nobel prizes in physics (one in 1944 for Rabi, and two in 1952 for Bloch and Purcell)<sup>4</sup> in the last century. In this way, as the research progressed, it became possible to determine physicochemical properties of the atoms and molecules, as well as detailed information about structure, dynamics, reactions state, and chemical environment, making use of different NMR parameters and using techniques governed by quantum theory.<sup>3</sup>

Nowadays, NMR spectroscopy has seen an unprecedented growth and is now one of the most powerful and versatile tools, with increasing applications in chemistry, biology, medicine and materials science. The fundamentals and methods of NMR form a vast, complex and evolving discipline.

### 1.1.1 Principles of NMR Spectroscopy<sup>3,4</sup>

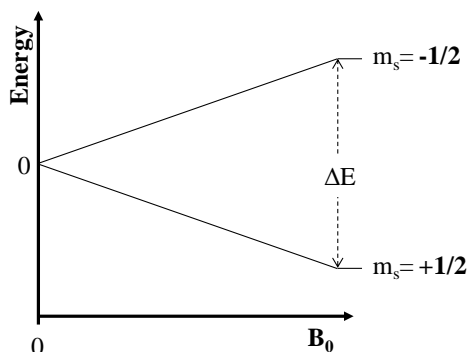
Nuclear Magnetic Resonance (NMR) spectroscopy detects nuclear-spin reorientation in an applied external magnetic field. For each kind of observed nuclei, the information provided by NMR depends on the electron environment in which the nuclei are immersed and on the positions of nuclei within molecules. Thus, NMR is a powerful tool for probing molecular structure and dynamics. The nuclei observable by NMR are those with a spin quantum number is different to zero ( $I \neq 0$ ). In an external magnetic field ( $B_0$ ), a spin will prefer an alignment along with the external field rather than oppose it. This is due to an energy difference

( $\Delta E$ ) between the two states that depends on the strength of  $B_0$  and the isotope-specific gyromagnetic ratio ( $\gamma$ ) (Figure 1.1) (equation 1).

$$\Delta E = \gamma \hbar B_0$$

**Equation 1**

Where  $\hbar$  is the reduced Planck constant.



**Figure 1.1** - Representation of the Zeeman effect in the energy states separation in NMR spectroscopy.

The radio-frequency of the electromagnetic radiation that is emitted from transitions between the two states is called the Larmor frequency, which is equal to the ratio of  $\Delta E$  and the Planck constant ( $\hbar$ ) and can be given in units of angular frequency ( $\omega_0$ ) or Hertz ( $\nu_0$ ) (Equation 2). NMR spectroscopy is the technique that allows measuring the Larmor frequency of spins in a magnetic field, thus enabling investigation of molecular properties.

$$\omega_0 = \frac{\Delta E}{\hbar} = \gamma B_0$$

**Equation 2**

To extract chemical information from a molecule, there are three fundamental concepts of solution NMR spectroscopy that are helpful, *chemical shifts* (*i.e.*, the specific Larmor frequencies of spins), *spin-spin couplings* (*i.e.*, spin interactions) and *spin relaxation* (*i.e.*, the time dependency of an NMR signal).

#### ***1.1.1.1 Chemical shift***

The magnetic field slightly differs for different spins, due to shielding effects from electron density around the specific nuclei. Therefore, spins will have shifted frequencies depending on their chemical environment (neighboring atoms and type of chemical bonding). Thus, each site experiences a slightly different magnetic field and has a different position in the NMR spectrum. Also, the chemical shifts will vary depending on the molecular orientation with respect to  $B_0$ , due to the shielding of  $B_0$  caused by anisotropic electron density of the atomic orbitals, a phenomenon called chemical shift anisotropy (CSA). In solution, this effect is averaged to zero, so that only one frequency is observed for each chemically distinct site.

#### ***1.1.1.2 Spin-spin couplings***

The NMR signal of a spin does not always appear as a single peak at a given chemical shift in an NMR spectrum. Instead, it is often observed as a split signal centered at the chemical shift. This phenomenon, the spin-spin coupling, arises from interactions between neighboring spins. The magnitude of the peak splitting is referred to as the coupling constant ( $J$ ) and is equal for both spins involved in the interaction. The most obvious spin-spin coupling in solution NMR spectroscopy is the scalar coupling, which is mediated by electron interactions through covalent bonds. The strength of the interaction is measured by the scalar coupling constant,  ${}^nJ_{IS}$ , in which  $n$  is the number of covalent bonds between the nuclei  $I$  and  $S$  and its magnitude is usually expressed in Hertz ( $Hz$ ). The important  ${}^nJ_{IS}$  have  $n$  until 4.

The  $J$ -coupling constant is also a source of information regarding the molecular shape, because of its dependence on bond geometry. The relationship can be described by Karplus-type equations (i.e., Equation 3), which differ depending on the nuclei involved in the bonds.

$$J(\theta) = A\cos^2\theta + B\cos\theta + c$$

**Equation 3**

Where  $J$  is the  $^3J$  coupling constant, and A, B, and C are constants that depend on the specific coupled nuclei.  $\theta$  is the dihedral angle.

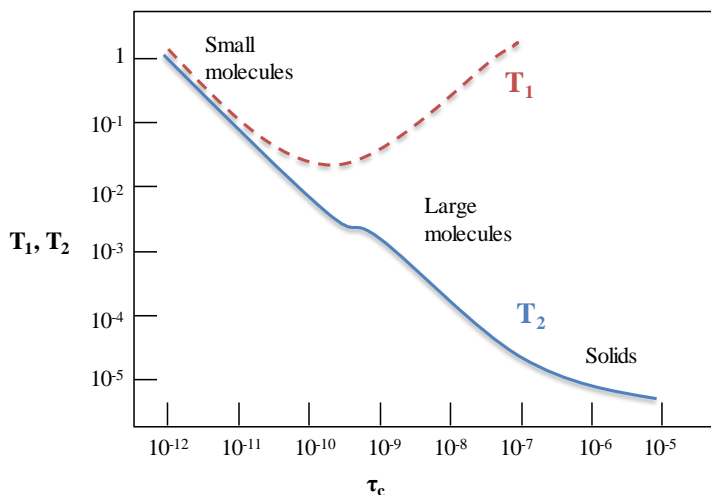
### ***1.1.1.3 Spin relaxation***

An RF pulse applied onto a sample at equilibrium causes a perturbation on the nuclear spins removing them from the thermal stationary state. As consequence of this pulse, the system will try to return to the equilibrium, losing the excess energy. However, due to the low transition energies associated with magnetic resonance, the lifetime of the excited states may be extremely long (few seconds to minutes for small molecules). These long lifetimes are fundamental for NMR spectroscopy as they result in sharp lines (as a consequence of the Heisenberg uncertainty principle).

In NMR, two relaxation parameters are defined:

- $T_1$ , the spin-lattice or longitudinal relaxation time ( $R_1$  for spin-lattice relaxation rate,  $R_1 = 1/T_1$ );
  
- $T_2$ , the spin-spin or transverse relaxation time ( $R_2$  for the spin-spin relaxation rate,  $R_2 = 1/T_2$ ).

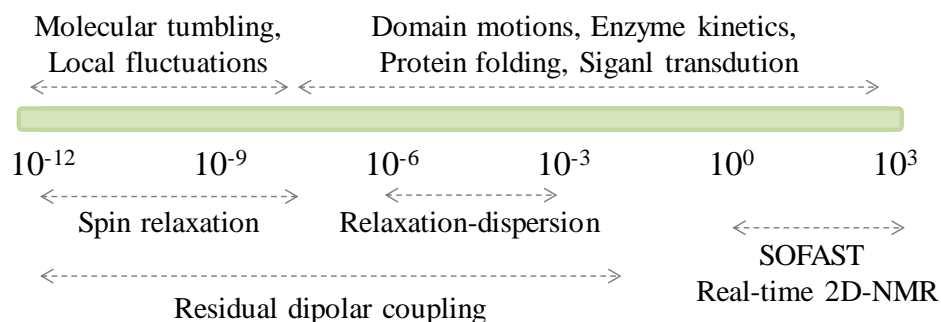
$T_1$  measures the efficiency with which the excited nuclear spins return to their ground state by exchanging energy with their surroundings.  $T_2$  is a measurement of the efficiency with which spins exchange energy with each other. The more efficient this exchange, the shorter the relaxation time. In different experiments, these relaxation times can be studied to understand protein's dynamics, transverse ( $T_2$ ) or longitudinal ( $T_1$ ), in the timescale of milliseconds (ms) to seconds (s), respectively (Figure 1.2). In solution, the resonance line widths are inversely proportional to the  $T_2$  relaxation time, which decreases with the increase in molecular size and tumbling time.



**Figure 1.2.** Behavior of  $T_1$  and  $T_2$  as a function of the correlation time for spin  $1/2$  nuclei relaxing by the Dipole-Dipole mechanism.

Dipole-dipole interaction is probably the most important mechanism of relaxation pathway for protons in molecules containing contiguous protons and for carbons with directly attached protons. This is also the source of the Nuclear Overhauser Effect (NOE). Dipolar coupling occurs when the magnetic field of one nuclear dipole affects the magnetic field at another nucleus. It depends on the distance between the nuclei and takes place through-space. Dipole-dipole relaxation is also dependent on the correlation time,  $\tau_c$ . Small molecules tumble

very fast and have short  $\tau_c$ , usually in the order of picoseconds. Large molecules, such as proteins, usually move slowly and have long  $\tau_c$ , usually in the order of nanoseconds. In summary, several dynamic processes can be studied by NMR. The different time scales of NMR observable phenomena are graphically represented in Figure 1.3.



**Figure 1.3** - Time scales of some important molecular dynamic processes and multidimensional NMR methods available to study them.

### 1.1.2 The Nuclear Overhauser Effect (NOE).

The nuclear Overhauser effect (NOE) defines the correlation between protons close in space (up to 5 - 6 Å distance) and is a consequence of dipole-dipole cross-relaxation in nuclear spin systems. Intra-proton distances can be estimated.<sup>5</sup> The NOE represents the change in intensity of a signal when the spin transitions of another nucleus cause a perturbation of its equilibrium populations. The two nuclei do not share a scalar, through bond, coupling; instead, they are sufficiently close in space to share a dipolar coupling. Thus, the NOE originates from dipolar cross-relaxation between proton pairs and depends on the proton-proton distance and on the molecular motion of the inter-proton vector (Equation 4)



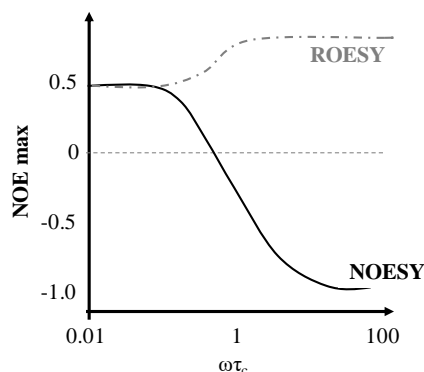
$$I_{NOE} \approx \left(\frac{1}{r^6}\right) f(\tau_c)$$

**Equation 4**

Where  $I_{NOE}$  is the NOE intensity,  $r$  is the proton-proton distance, and  $f$  is a function that depends, among others factors, on the correlation time ( $\tau_c$ ) that describes the motion of the inter-proton vector.

$\tau_c$  is the correlation time (decay time of the correlation function). When considering isotropic molecular tumbling,  $\tau_c$  is related with the time taken for the molecule to rotate by 1 radian about any axis. Therefore, rapidly tumbling molecules will have short correlation times, while slowly tumbling molecules will have long correlation times. Thus, the correlation time of a molecule is related with its molecular weight.

Cross-relaxation rates for small molecules are positive while for large molecules, as proteins, are negative (Figure 1.4). For medium size molecules, such as small peptides, cross-relaxation rates can be positive or negative, which means very small (often not measurable) NOEs. This problem can be overcome by using the ROE technique (rotating frame NOE), in which cross-relaxation occurs in the transverse instead of in the longitudinal plane (Figure 1.4).



**Figure 1.4** - Maximum NOE and ROE in function with  $\omega\tau_c$ .

### 1.1.3 Diffusion ordered spectroscopy

Diffusion ordered spectroscopy (DOSY) identifies the molecular components of a mixture and obtains, at the same time, information on their size. This information may be accessed by measuring the self-diffusion coefficient, which measures the random translational motion of molecules, driven by their internal kinetic energy. Self-diffusion coefficients are related to the structural properties of a molecule by their dependence on the physical properties of the molecule (e.g. size, charge and shape). Furthermore, self-diffusion coefficients also depend on the characteristics of the surrounding medium (e.g. temperature and viscosity).

For a spherical molecule moving in an unconstrained environment, the Stokes–Einstein law predicts a correlation between the hydrodynamic radius  $r$  and the self-diffusion coefficient ( $D$ ) (Equation 5)

$$D = \frac{kT}{6\pi\eta r}$$

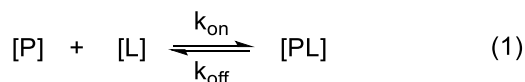
**Equation 5**

Where  $k$  is the Boltzmann constant,  $T$  is the temperature and  $\eta$  is the medium viscosity.

The diffusion of molecules is measured by evaluating the attenuation of a spin echo signal using pulsed-field gradients (PFG).<sup>6</sup>

### 1.1.4 NMR methods for the study of Protein-Ligand interactions:

The binding of small ligands to large proteins usually follows a bimolecular association reaction, which can be described by one-site binding model:



### Equation 6

The Equation 6 represents a dynamic equilibrium involving three species: the free protein P, the free ligand L, and the receptor-ligand complex PL, with  $k_{\text{on}}$  and  $k_{\text{off}}$  being the on (association) and off (dissociation) rate constants. The unimolecular rate constant  $k_{\text{off}}$  is inversely proportional to the mean lifetime  $\tau_B$  of the protein-ligand complex. The bimolecular rate constant  $k_{\text{on}}$  measures the probability of encounter between free receptor and ligand.

The binding affinity can be quantified by the temperature-dependent equilibrium dissociation constant (Equation 7):

$$K_D = \frac{[P][L]}{[PL]} = \frac{k_{\text{off}}}{k_{\text{on}}}$$

### Equation 7

Where [P], [L] and [PL] are the equilibrium concentrations of protein, ligand and the complex, respectively. [P] and [L] are also referred to as the free state and [PL] is varyingly referred to as ‘bound ligand’.  $K_D$  has the units of concentration.

When the receptor and ligand molecules are free, they retain their intrinsic NMR parameters (e.g. chemical shifts, relaxation rates, translational diffusion coefficients). However, in each other’s presence, their mutual binding affinity drives a two-state exchange process that can toggle both sets of molecules between the free and bound states (Figure 1.5). Besides providing structural information, NMR methods can supply a wide variety of transient and dynamic information on the complex. NMR methods can be divided in those detecting the ligand signals

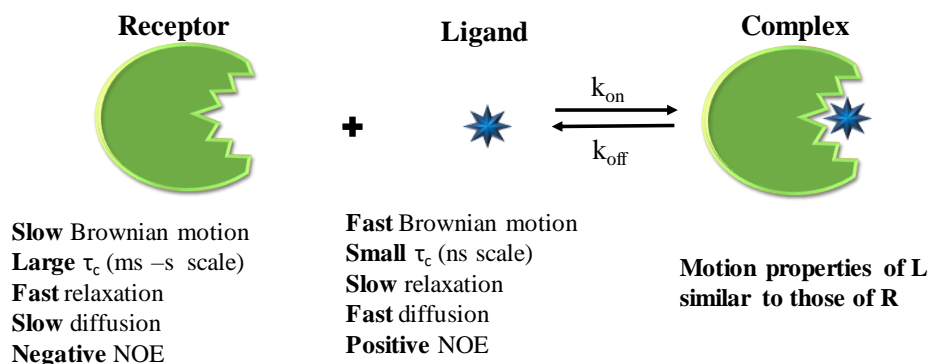
(ligand-based experiments) and those detecting the receptor signals (receptor-based experiments).<sup>7,8</sup>

#### ***1.1.4.1 Ligand-detected methods***

Generally, ligand-based NMR experiments assay for binding are employed by:

- i. Exploiting the differential mobility of the ligand in the free versus bound state: bound ligands will transiently experience the much slower rotational and translational mobility of the large receptor, leading to altered relaxation parameters and diffusion coefficients, respectively;
- ii. exploiting transfer of magnetization processes.<sup>9</sup>

Ligands, small molecules, are characterized by small relaxation rates, positive 2D-NOESY cross-peaks, and large translational diffusion coefficients,  $D$ . In contrast, bound ligands share the NMR relaxation properties of the receptor, usually large molecules, with fast relaxation; negative 2D-NOESY cross-peaks, highly efficient spin-diffusion, and smaller molecular diffusion coefficients ( $D$ ). These apparent differences are the basis of ligand-based methods, where changes in the ligand NMR parameters upon binding to the receptor can be exploited to detect and characterize the interaction (Figure 1.5).



**Figure 1.5** - Illustration of the changes in different physical properties when a small ligand interacts with a large protein.

#### 1.1.4.1.1 Line Broadening in NMR

The primary NMR methods for detection of ligand-protein interactions rely on measurement of T1/T2 relaxation rates of ligand resonances. Line broadening in the presence of the protein is often an indicator of protein–ligand interactions. The degree of line broadening depends on many factors, including the transverse relaxation rate, the exchange rate, and the fractions of the ligand in the free and bound states. Ligands that bind to a macromolecular receptor experience an enhancement of their T2 relaxation: the molecular motion of the ligand dramatically changes becoming similar to that of the receptor with the selective shortening of the T2 relaxation times, which is translated into a broadening of certain protons of the ligand, which can even disappear.<sup>10</sup>

#### 1.1.4.1.2 Saturation Transfer Difference (STD) NMR

The STD experiment is a popular and highly versatile screening method for the identification and characterization of protein-ligand interactions.<sup>7,9,11</sup> The success of this technique is a consequence of its robustness and the fact that it is focused on the signals of the ligand, without any need of processing NMR

information about the receptor, only using small amounts of non-labeled macromolecule.

The STD NMR experiment is the difference between two different  $^1\text{H}$ -NMR spectra performed on the same ligand-receptor sample and relies on the magnetization exchange from the protein-bound state to the free state of the ligand. Basically, the first one (*on-resonance* spectrum) is a spectrum in which the protein is selectively saturated, by irradiating at a region of the spectrum that contains only resonances of the receptor (from 0 ppm to -2 ppm). The saturation is efficiently propagated across the entire protein through spin-diffusion and transferred to the binding compounds via intermolecular  $^1\text{H}$ - $^1\text{H}$  cross relaxation at the ligand-receptor interface. The second spectrum (*off-resonance* spectrum) is recorded as a reference or blank spectrum. Here, the chosen saturation region is free from ligand and protein signals (e.g. 100 ppm).

The protein-to-ligand saturation transfer will affect the intensity of the ligand resonance signals in the spectrum obtained with selective receptor saturation ( $I_{SAT}$ ), and when compared to the spectrum acquired without saturation transfer ( $I_0$ ). Thus, the difference in intensity due to saturation transfer can be quantified ( $I_{STD} = I_0 - I_{SAT}$ ) and constitutes an indication of binding (Figure 1.6). STD is ideally suited to receptors with large masses (>30 kDa). Receptors with large molecular masses display large rotational correlation time,  $\tau_c$  that enhance spin diffusion and, consequently, saturation transfer within the receptor and to the ligand.

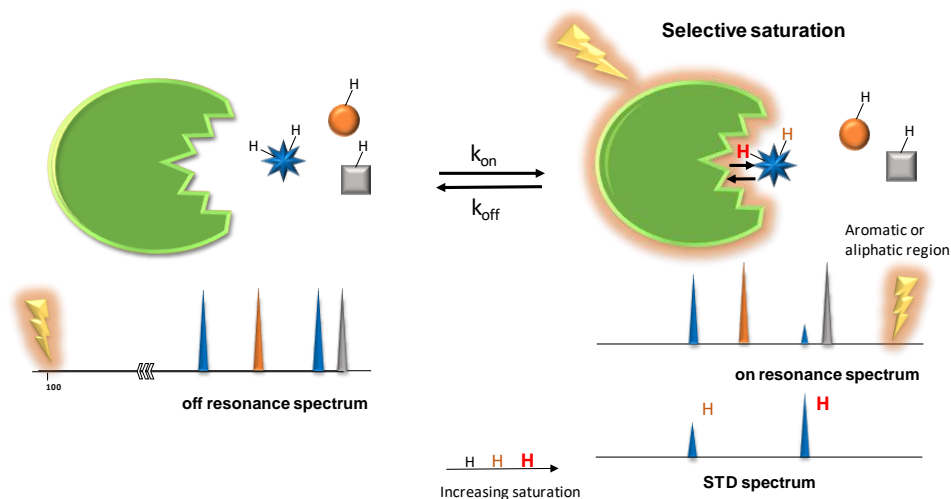


Figure 1.6 - Schematic representation of the STD experiment. When a protein becomes saturated, the intermolecular NOE is only transferred to those ligands, with shape and size that interact with the protein. The transference of saturation is detected by subtraction of on to off-resonance spectrum in order to obtain an NMR spectrum where only the signals from molecules that bind to the protein are present. Resonance signals from non-binders do not show up in the difference spectrum. Adapted from <sup>12</sup>.

Furthermore, for a given compound, the building blocks of the ligand in closest proximity to the receptor molecule receive higher amounts of magnetization, which are translated into stronger signals in the STD spectrum. Thus, the method can be also used to obtain detailed structural information on the binding epitope of the ligand (epitope mapping) (Figure 1.6).<sup>13</sup>

Regarding group epitope mapping analysis, computing methods have been developed to predict and/or interpret STD data of reversibly forming ligand-receptor complexes. In particular, CORCEMA (Complete relaxation and conformational exchange matrix) package<sup>14,15</sup> predicts the expected STD intensities for a given model of a ligand-protein complex, and compares them quantitatively with the experimental STD data. This version is very useful for rapidly determining if a model for a given ligand-protein complex is compatible with the STD-NMR data obtained in solution.

Remarkably, it is also possible to measure  $K_D$  if STD signals are recorded in the presence of a competitor inhibitor.<sup>14,16</sup> Competition experiments can be performed for designed mixtures of both ligands by evaluation of the gradual decay of the STD signals from the reference in the presence of increasing concentrations of the inhibitor. Assuming a simple bimolecular association reaction for ligand L and a competitive inhibitor I to a receptor protein, the  $K_D$  value of the ligand L can be determined from a known value of  $K_I$ . However, more complicated situations such as those arising from binding of ligands to secondary binding sites should be taken into account. STD competition experiments have been also used for the detection of high-affinity ligands.<sup>14</sup>

STD methods can be combined with any NMR pulse sequence generating a whole suite of concatenated STD NMR experiments such as STD-TOCSY or STD-HSQC.<sup>15</sup> The additional deconvolution of signals in a second dimension can be very helpful for mapping the binding epitope of ligands with complex 1D proton spectra.

The major drawbacks associated with STD experiments are related to the potential self-association of the ligands, especially with aromatic molecules, and to the possibility of non-specific binding with the receptor. These potential problems can be circumvented by employing different ligand/receptor molar ratios, different saturation frequencies, and employing competitive binders (or inhibitors) of the interaction process, if they are available.

#### **1.1.4.1.3 WaterLOGSY**

The waterLOGSY (Water-Ligand Observed via Gradient SpectroscopY) technique, like STD, relies on excitation of the receptor-ligand complex through a selective RF pulse scheme. However, waterLOGSY experiment achieves this effect indirectly, by selective perturbation of the bulk water magnetization.<sup>17</sup> The intended transfer of magnetization is therefore water  $\rightarrow$  receptor  $\rightarrow$  ligand. The magnetization of the water molecules is selectively saturated or inverted, and



during a long mixing time of up to several seconds, the magnetization is transferred, via  $^1\text{H}$ - $^1\text{H}$  cross-relaxation, to the ligand spins at the protein-ligand interface.

There are essentially two pathways by which this transfer of the magnetization occurs:

- i. Direct cross-relaxation from the water molecules tightly bound at the protein-ligand interface,
- ii. Chemical exchange between the water protons and the labile protons of amine and hydroxyl groups in the protein which in turn cross-relax with the protons of the bound ligand.

In either case, the perturbation of the bulk water magnetization is transferred to the binding compounds while residing in the receptor binding site. Distinguishing binding from nonbinding compounds in the waterLOGSY experiment is achieved by observation of the differential cross-relaxation properties of these ligands with water. Bound ligands interact directly or indirectly with inverted water spins with motional and relaxation properties of the receptor, with a slow tumbling rate and therefore, a long correlation time. This yields negative cross-relaxation rates, and inversion of the NOE signs for the bound ligand. For their part, non-interacting ligands receive the magnetization only in the free state, via water molecules involved in their solvation sphere, leading to positive cross-relaxation rates and the sign of the NOE remains unaltered. As a consequence, binders and non-binders display waterLOGSY peak intensities of opposite sign, as in *tr*-NOESY (see below).

#### **1.1.4.1.4 Methods based on transferred NOE effects**

Transferred NOE methods are based on the changes in the rotational motion properties of a ligand upon binding to a large macromolecular receptor. For small molecules, characterized by a short correlation time, NOEs are positive. However,

for large receptors, the correlation time is in the ns time scale and the associated NOEs are negative. Thus, this technique has become a classical method for studying the binding of ligands to large receptors and to deduce the conformation of the binding ligand in the protein site.<sup>9</sup>

In trNOESY, low protein:ligand ratios are typically employed (from 1:5 to 1:50, depending on the affinity and kinetic parameters). This implies that the NOEs observed for the ligand (tr-NOEs) will keep the information of its bound state, provided that the off-rate is fast in the relaxation time scale. Therefore, the signals of small ligands will experience a NOE sign change in the presence of the receptor, from positive (free state, small molecule, short correlation time) to negative (large complex, large correlation time), which will reflect the conformation of the ligand in the binding site. Technically, the tr-NOESY experiment consists of acquiring an ordinary NOESY spectrum for the ligand in presence of the protein. Normally, with lower mixing times is in the range of 50 to 100 ms, whereas for nonbinding molecules it is four- to ten-times longer.

#### ***1.1.4.2 Receptor-detected methods: Chemical Shift Perturbations & HSQC***

For small proteins (~10 kDa) and peptides structure determination, 2D experiments, as COSY (COrrrelation SpectroscopY) and TOCSY (TOtal Correlation SpectroscopY), are used to give information between protons, due to the scalar coupling through covalent bonds. NOESY experiments (Nuclear Overhauser Effect SpectroscopY) are the most important multidimensional experiments in structure determination since they correlate protons through space, allowing the determination of the distances between close protons.<sup>18</sup> These methodologies can be used to determine the structure of proteins up to 10-15 kDa.<sup>18</sup>

For large molecules, the number of hydrogen atoms is extremely high. The rotational correlation times of globular proteins and, therefore, the line widths of

the NMR resonances also increase linearly with the molecular mass. For these reasons, conventional assignment procedures based on sequential NOE correlations become very difficult, if not impossible. Heteronuclear experiments are then necessary and can solve these problems for proteins with molecular masses less than 30 kDa.<sup>18</sup> However, isotopic labeling of the samples is required with NMR active isotopes  $^{13}\text{C}$ ,  $^{15}\text{N}$ , and sometimes  $^2\text{H}$ .<sup>19</sup>

In protein NMR, the Heteronuclear Single-Quantum Coherence (HSQC) experiment<sup>19</sup> correlates the  $^{15}\text{N}$  or  $^{13}\text{C}$  nuclei with the attached  $^1\text{H}$  via the one-bond scalar coupling  $J_{N-H}$  or  $J_{C-H}$ , respectively. Thus, in the  $^1\text{H}/^{15}\text{N}$ -HSQC one signal is expected for each amino acid residue with the exception of proline, which has no amide-hydrogen due to the cyclic nature of its backbone. In this sense,  $^1\text{H}/^{15}\text{N}$ -HSQC spectrum is the ‘fingerprint’ of the protein.

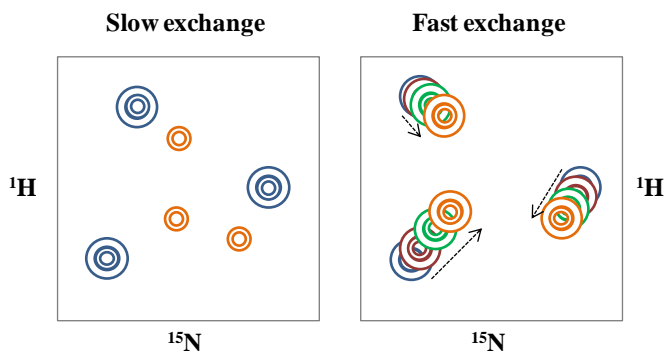
One of the most frequently used receptor-based methods relies on monitoring the perturbation of the receptor  $^1\text{H}/^{15}\text{N}$ -HSQC chemical shifts upon binding of the ligand.

#### **1.1.4.2.1 Chemical Shift Perturbations**

The chemical shift is an extremely sensitive parameter to environmental changes (pH, temperature, and folding states) and is able to reveal interaction processes. Moreover, nuclei located in the protein binding pocket usually show the largest effects. Thus, mapping chemical shift perturbations via heteronuclear shift-correlation experiments is a straightforward NMR technique to detect binding and also to define the protein residues that are involved in the interaction.

The  $^1\text{H}/^{15}\text{N}$ -HSQC spectrum of the protein, considered as the NMR fingerprint, is acquired as a reference spectrum. Then, sequential  $^1\text{H}/^{15}\text{N}$ -HSQC spectra are acquired for different amounts of ligand. The proximity of the ligand will modify the environment of the nuclei that are at the interface of the protein-

ligand complex. Therefore, the residues involved in the binding will have a different chemical shift than in the unbound form. When the exchange rate between free and bound states is fast on the chemical shift time scale, the  $^1\text{H}/^{15}\text{N}$  cross-peaks will move smoothly from their position in the free spectrum to those in the bound spectrum, with the frequency of the signal at any titration point being the weighted average of free and bound chemical shifts. When the exchange rate is slow on the chemical shift time scale, the free will signal gradually disappear and the bound signal will appear, as the intensities of the two peaks reflect the concentrations of the free and bound protein.<sup>20</sup> The chemical shift perturbations allow identifying binding and to localize the binding site on the protein structure. Mapping  $^1\text{H}/^{15}\text{N}$  chemical shift perturbations along the structure of the protein is usually applied to detect residues with larger chemical shift perturbations, which can be considered as part of the binding site (Figure 1.7).



**Figure 1.7** - Schematic representation of the chemical shift perturbation induced by the interaction of an unlabelled small ligand with a  $^{15}\text{N}$ -labelled protein as detected in the  $^1\text{H}/^{15}\text{N}$ -HSQC spectrum. Slow exchange (strong binding) after the addition of a sub-stoichiometric quantity of ligand; two signals are observed one for the free (blue) and another for the bound (red) protein. Fast exchange (weak binding), the scheme represents the superposition of  $^1\text{H}/^{15}\text{N}$ -HSQC sequence obtained for increasing quantities of ligand in the order blue, dark red, green and orange; only one signal is observed in each spectrum with the shift (arrow) depending on the structural changes induced by the ligand at the specific residue.

Fast relaxation of the protein  $^1\text{H}/^{15}\text{N}$  signals becomes an issue when dealing with high molecular weight macromolecules ( $> 35$  kDa). Fortunately, the implementation of transverse relaxation optimized spectroscopy (TROSY) HSQC variant allows targeting larger protein receptors (up to 100 kDa) by selectively recording the most slowly relaxing signal component of the  $^{15}\text{N}$ - $^1\text{H}$  correlation.<sup>21</sup>

Regarding specific labeling schemes, selective labeling methods have been also developed to improve or simplify NMR spectra of large proteins.

### ***1.1.4.3 Specific isotopic labeling to study protein by NMR***

#### **1.1.4.3.1 $^{13}\text{C}$ -methyl labeled proteins**

A frequent, although expensive, approach is to use  $^{13}\text{C}$ -methyl labeled proteins, which contain  $^{13}\text{C}$  probes in the side chain of the protein, such as alanine, valine, isoleucine, leucine and methionine. For  $^{13}\text{C}$ -methyl labeled proteins, 2D  $^1\text{H}/^{13}\text{C}$ -HMQC experiments are typically used. These experiments have some advantages in comparison to  $^1\text{H}/^{15}\text{N}$ -HSQC experiments. Specifically, methyl groups have three protons with a three-fold degeneracy and give rise to stronger signals. They also tend to resonate in a sparsely populated region of the  $^1\text{H}/^{13}\text{C}$  correlation spectrum, reducing spectral overlap.<sup>22</sup>

Several labeling schemes for amino acids, including alanine, isoleucine, leucine, methionine, and threonine, have been developed.<sup>22</sup> These labeling strategies commonly involve metabolic pathways, metabolic precursors, and the amino acid's propensity for isotopic scrambling at other sites in the protein. For example, in the case of leucine and valine, which share the same metabolic pathway, using metabolic precursors common to both amino acids, such as  $\alpha$ -ketoisovalerate, labeling techniques often result in the incorporation of isotopes into both amino acids. Isoleucine, can also be labeled using  $\alpha$ -ketobutyrate, one of its

metabolic precursor.<sup>23</sup> For alanine and methionine, residue specific labeling can be obtained by via of supplementation of minimal expression medium with the appropriate isotopically labeled amino acid.<sup>24</sup>

#### **1.1.4.3.2 <sup>19</sup>F labeled molecules**

Although classical spectroscopic methods, <sup>1</sup>H, <sup>15</sup>N, and <sup>13</sup>C, have been used to study the behavior of biomolecules in solution, <sup>19</sup>F-NMR is also becoming very popular although its implementation requires manipulation of the system. The <sup>19</sup>F nucleus displays different properties that render it ideal for NMR studies. In fact, <sup>19</sup>F is an NMR active isotope with spin ½, and it is 100% naturally abundant. It also possesses a high gyromagnetic ratio that results in excellent sensitivity (83% of <sup>1</sup>H). In addition, the fluorine chemical shifts are extremely sensitive to changes in local environment (the chemical shift is +/- 100-fold larger than that of <sup>1</sup>H).<sup>25</sup>

The <sup>19</sup>F atom has been used in biological NMR, like a molecular probe into proteins, peptides, and carbohydrates<sup>26–29</sup> to provide information on structure and dynamics,<sup>30–32</sup> protein-ligand interactions,<sup>25,33</sup> and protein (un)folding,<sup>34,35</sup> demonstrating unequivocally its versatility as probe for NMR.

In the protein field, several methods to prepare <sup>19</sup>F-modified proteins have been described.<sup>25,30–35</sup> These methods fall into three main categories:

- i. Post-translational covalent attachment of <sup>19</sup>F-containing moieties to the protein by conjugation of the <sup>19</sup>F-containing moiety to a reactive group with an -SH group on a solvent accessible cysteine.<sup>25,30,31,33</sup> The great advantage of this method is the ability to incorporate the label into proteins for which biosynthetic labeling is expensive, like a mammalian cells expression.
- ii. Specific incorporation of the type of biosynthetic amino acids modified with <sup>19</sup>F.<sup>25,33–35</sup> This approach is carried out by expression

in a defined growth media, supplemented with the  $^{19}\text{F}$ -modified amino acid

- iii. Site-specific incorporation of the  $^{19}\text{F}$ -modified amino acid using recombinant expressed orthogonal tRNA / tRNA-tRNA pairs.<sup>25,33</sup> This system is based on an extension of the genetic code to beyond the natural 20 amino acids.

### **1.1.5 Carbohydrates and Carbohydrate-Protein Interactions**

Carbohydrates (glycans, sugars, saccharides) are the most abundant biomolecules in nature and they are estimated to account for 70% of the total biomass on earth.<sup>36</sup> As a product of the photosynthesis process, carbohydrates function as energy storage in organisms and their ability to form polymers make them important for the structure of the cell. Cell membranes themselves contain sugars (as well as proteins and lipids.) These glycans play a fundamental role in a variety of biological processes, including cell adhesion and communication. Therefore, molecular recognition events between sugars and proteins constitute the first interaction process in many molecular mechanisms related to health and disease.

In the initial part of this chapter, different NMR concepts have already been introduced. Although the particular systems of interest that have been studied in this Thesis will be explicitly described below, I will briefly point out now the different types of information that can be extracted from NMR experiments in this particular topic, along with the description of the basic features that are behind the recognition of glycans by protein receptors.

In brief, key features of the structural, conformational, dynamic and interaction properties of carbohydrates will be explored in the following chapters:

- i. *Molecular shape* (structure and conformation), which can be deduced from NOEs and *J*-couplings.

- ii. *Molecular motion*, which can be investigated by NMR relaxation and chemical exchange, depending on the timescale of motion.
- iii. *Molecular interactions*, which can be assessed by chemical shift, relaxation and NOE data.

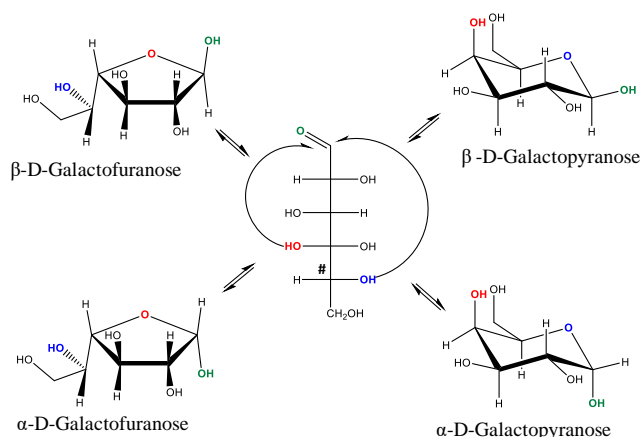
### 1.1.5.1 Structure and Conformation

Carbohydrates are extremely diverse, as consequence of the large number of different monosaccharide units that can be combined by employing different types of glycosidic linkages to form oligo- or polysaccharides.<sup>37</sup> These complex glycosylation patterns result in a large variation in shapes and geometries of carbohydrates.

Monosaccharides, the simplest forms of carbohydrates, are polyhydroxylated carbon chains carrying a aldehyde functionality (aldoses) or a keto functionality (ketoses). The carbon chains contain at least three carbons (trioses), although molecules with five (pentoses) and six (hexoses) carbon atoms are the most common ones.

Monosaccharides exist in many different diastereomeric forms because of the high number of stereogenic centers. Each diastereomer has its own name (*e.g.*, glucose, galactose). Enantiomers are named according to the D and L notation, which is determined by the stereochemistry at the highest numbered stereogenic carbon in the saccharide chain (Figure 1.8). The typical cyclization process results in the formation of a new stereogenic center at the hemiacetal carbon (anomeric carbon), which now displays either  $\alpha$  or  $\beta$  configuration (Figure 1.8).



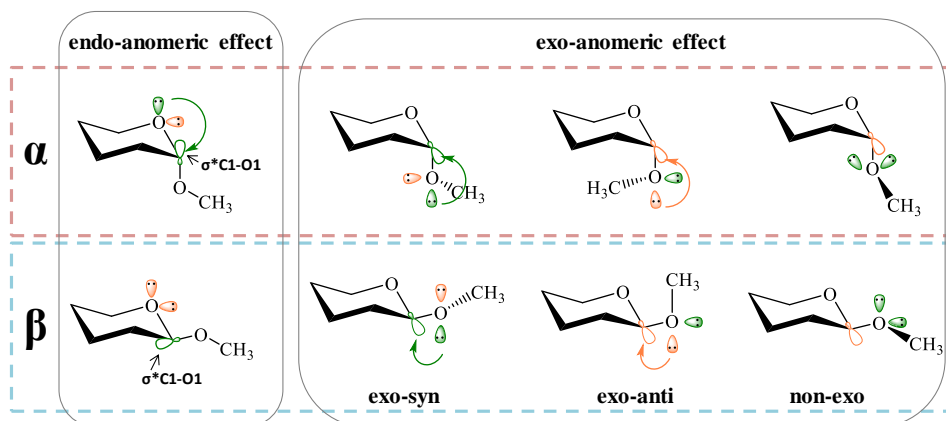


**Figure 1.8** - Cyclic forms of *D*-galactose (Gal). The hashtag marks the highest numbered stereogenic carbon that defines the enantiomeric notation.

#### 1.1.5.1.1 The anomeric effect

The anomeric effect was first described in the 1950s by J. T. Edward and R. U. Lemieux.<sup>38</sup> The anomeric effect phenomenologically describes the stabilization of the axial (*versus* equatorial) alkoxy groups at C1 of a pyranose ring. Nowadays, it is accepted that the reason is due to orbital and electronic factors produced by the simultaneous presence of two oxygens attached to one single carbon atom.

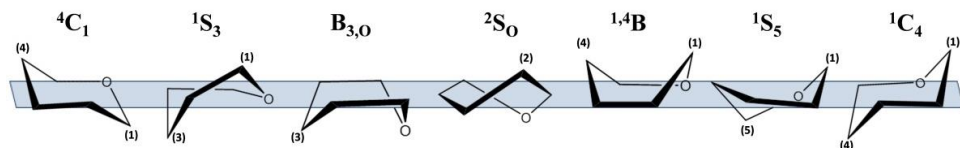
In principle, the tendency of the anomeric hydroxyl group or electronegative substituent to adopt axial orientation ( $\alpha$ -anomer) rather than equatorial ( $\beta$ -anomer) is against to what would be expected based on steric factors. However, the favorable interaction between one lone electron pair located at a molecular orbital ( $n$ ) on either glycoside oxygen atoms and the vicinal anti-bonding molecular orbital ( $\sigma^*$ ) of the contiguous C-O bond overcomes the possible steric hindrance provided by the axial orientation. The anomeric effect (*endo* and *exo*) are now reasonably explained by computational approaches, including *ab initio* calculations (Figure 1.9).<sup>39</sup>



**Figure 1.9** - Schematic representation of the lone pair- $\sigma^*$  interactions responsible for the endo- or exo-anomeric effects in  $\alpha$ - and  $\beta$ -glycosides. The endo-anomeric effect exists only in  $\alpha$ -glycosides due to the geometrical requirements that allow the molecular orbital overlapping. For the exo-anomeric effect, different conformers are represented, both for  $\alpha$ - and  $\beta$ -glycosides. However, the overlapping between the  $\sigma^*$  and the lone electron pairs is possible only for two of the three conformations. The fill color of the orbitals does not refer to the orbital phase, but it is a schematic representation for full or empty orbitals.

### 1.1.5.1.2 Conformation

The sugar ring conformation introduces another source of structural variability.<sup>40</sup> Once cyclized, the sugar ring may experience conformational flexibility, depending on the type of cycle (furanose, pyranose, septanose,...), the configuration of the stereogenic centers, and the size and chemical nature of the substituents. For pyranoses, the ring conformations may include envelope, boat and skew ring puckers, as well as alternative chairs, usually in fast equilibrium (Figure 1.10).<sup>40,41</sup>



**Figure 1.10** - Ring conformations adopted by monosaccharide rings. The energetically favored chair conformations are labeled as  ${}^4C_1$  or  ${}^1C_4$ , depending on the position of carbons 1 and 4 (above or below the sugar ring plane, as noted by the position of the numeral). The other conformations are as Skew boat (labeled as S) or boat (labeled as B).

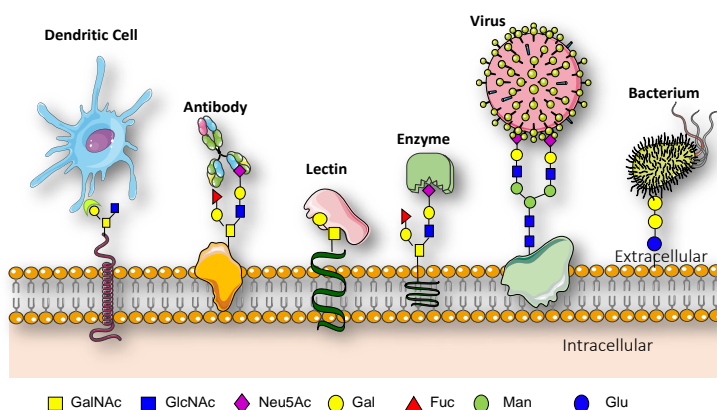
This 3D shape has special relevance for the molecular recognition of sugars by receptors, since it determines the orientation of functional groups in the three-dimensional space.<sup>42</sup> For the common D-pyranose rings, the  ${}^4C_1$  geometry is the typical chair conformation, since it minimizes steric clashes. However, monosaccharide rings may populate a number of additional, distinct 3D shapes, at relative ratios that depend on the free energy of the different conformers.<sup>43</sup>

### 1.1.5.2 Carbohydrate-protein interactions

Glycans play different roles inside the cell, as well as on the cell surface. Glycoproteins, glycolipids, and/or peptidoglycans are present at the cell surface and modulate the concomitant molecular recognition events. In this context, the knowledge of the 3D structure of these molecules at the maximum possible resolution, as well as the characterization of their recognition features by different entities, such as lectins, enzymes, antibodies, viruses, and bacteria (Figure 1.11) are essential for a thorough understanding of many vital processes and to open the possibility of their modulation.

As mentioned above, glycans play a key role in cellular transport and adhesion, cell signaling processes, cell-cell communication, immune response, hyperacute rejection of tissue transplants of nonhuman sources, fertilization, tissue

maturation, apoptosis, blood clotting, infection by bacterial and viral pathogens, tumor growth, and metastasis.<sup>44</sup> Thus, understanding the structure and function of carbohydrates is crucial to realize their function in health and disease, as well as, to develop new glycan-based therapeutics.



**Figure 1.11** - Examples of carbohydrate-protein interactions. GalNAc represents *N*-acetylgalactosamine; GlcNAc: *N*-acetylglucosamine; Neu5Ac: *N*-Acetylneuraminic Acid; Gal: Galactose; Fuc: Fucose; Man: Mannose and Glc: Glucose.<sup>44</sup>

In general, glycans bind at domains defined by shallow pockets on the hydrophilic surface of proteins. The interaction with the receptors requires the payment of an entropy penalty due to desolvation, which is usually compensated by the enthalpically favored interactions established with the receptor. Overall, sugar binding to proteins is made possible via several attractive forces including polar and hydrophobic interactions, especially CH- $\pi$  interactions, solvation effects, hydrogen bonds, van der Waals interactions, and electrostatics.<sup>45</sup>

The establishment of hydrogen bonds between sugars and their receptors constitutes perhaps the most evident protein-carbohydrate bonding potential due to the existence of numerous -OH groups in all saccharides. Furthermore, hydrogen bonding is made possible not only thanks to the hydroxyl groups, but also to the

presence of amine and carboxyl groups in many sugars. Hydroxyl groups may participate in hydrogen bonds both as donors and as acceptors, by means of their oxygen lone electron pairs. In some instances, the same –OH group simultaneously acts as donor and acceptor, a characteristic phenomenon in sugar-protein interactions, known as cooperative hydrogen bonding. Sugar hydroxyl groups establish hydrogen bond contacts with the side chains of polar residues, most often aspartic and glutamic acid, asparagine, glutamine, arginine and serine, as well as backbone amide and carbonyl groups.<sup>46,47</sup>

Apart from hydrogen bonding, non-polar interactions also contribute to the molecular recognition between biochemical species. They are of special interest when aromatic protein residues are involved in the recognition event. For instance, the non-polar CH groups of a ligand tend to pack with adjacent aromatic rings establishing CH- $\pi$  interactions. The geometry condition for CH- $\pi$  interactions requires the presence of CH groups perpendicular to the aromatic ring where the CH- $\pi$  system stabilize the complex formation by entropic and enthalpic contributions. Since the stabilizing energy of one CH- $\pi$  interaction is rather low, (CH- $\pi$  interaction ca.= 1 kcal/mol), the simultaneous CH- $\pi$  interaction of various CH groups could finally drive the molecular recognition. Indeed, CH- $\pi$  interactions have been deeply described in carbohydrate recognition.<sup>29,45,48–51</sup>

Due to the biochemical complexity of these systems, other polar and non-polar intermolecular interactions may take place in the molecular recognition event. Some of them involve third partners, such as ions or water molecules. Water mediated interactions take place when a water molecule establishes hydrogen bond simultaneously with both partners, ligand and receptor. The nature of this interaction resides in the presence of solvating water molecules in the binding pocket that enhances the ligand binding mediating as a “bridge” between the ligand and receptor.<sup>52,53</sup> Another type of interaction involves the coordination of a divalent cation bridging certain sugar hydroxyls and negatively charged aspartates or glutamates. Such interaction is displayed by proteins belonging to the C-type family of lectins, which require the presence of Ca<sup>2+</sup> ions to bind their sugar ligands.<sup>14,52</sup>

Further forces involved in the recognition of glycans by their protein receptors include electrostatic interactions between charged saccharides, such as sialic acid residues or sulfated glycosaminoglycans, and protein residues of opposite charge.<sup>54</sup>

Carbohydrate-protein interactions can be studied by several methods. High-throughput microarrays<sup>55–57</sup> can be employed to monitor the binding of a glycan to a protein. This qualitative approach is usually complemented by surface plasmon resonance (SPR) data.<sup>55</sup> Indeed, besides NMR, binding affinities can be quantified by SPR or by isothermal titration calorimetry (ITC);<sup>58</sup> These two techniques provide specific advantages. ITC is able to determine the thermodynamic parameters (enthalpy and entropy) of the interaction, whereas SPR may provide the kinetics ( $k_{\text{off}}$ ) of binding. However, neither method yields detailed structural information on the conformation of the bound partners.

From the structural point-of-view, advances in cryo-electron microscopy are providing details on different biological mechanisms involving sugars, especially during the last years. Additionally, better protocols for the structural refinement of the electron density acquired by X-ray methods for analyzing protein–sugar complexes and glycoproteins are also available, with important consequences in the glycoscience field.<sup>59,60</sup> Alternatively, NMR spectroscopy remains as one of the most rewarding techniques to explore protein–glycan interactions. Without forgetting the power of the other techniques, this Thesis is particular focused in the use of NMR approaches.<sup>61</sup>

### **1.1.6 References**

1. Purcell, E. M., Torrey, H. C. & Pound, R. V. Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Phys. Rev.* **69**, 37–38 (1946).
2. Bloch, F., Hansen, W. W. & Packard, M. Nuclear Induction. *Phys. Rev.* **69**, 127 (1946).
3. Cavanagh, J., Fairbrother, W. J., Palmer, A., Rance, M. & Skelton, N. J. *Protein*

- 
- NMR Spectroscopy*. (2007).
4. Macomber, R. S. *A Complete Introduction to NMR Spectroscopy*. (1998).
  5. Neuhaus, D. & Williamson, M. P. *The Nuclear Overhauser Effect in Structural and Conformational Analysis*. (2000).
  6. Groves, P. *et al.* Protein molecular weight standards can compensate systematic errors in diffusion-ordered spectroscopy. *Anal. Biochem.* **331**, 395–397 (2004).
  7. Calle, L. P., Cañada, F. J. & Jiménez-Barbero, J. Application of NMR methods to the study of the interaction of natural products with biomolecular receptors. *Nat. Prod. Rep.* **28**, 1118 (2011).
  8. Fernández-Alonso, M. del C. *et al.* in *New Applications of NMR in Drug Discovery and Development* 7–42 (2013).
  9. Unione, L., Galante, S., Díaz, D., Cañada, F. J. & Jiménez-Barbero, J. NMR and molecular recognition. The application of ligand-based NMR methods to monitor molecular interactions. *Med. Chem. Commun.* **5**, 1280–1289 (2014).
  10. Viegas, A. *et al.* Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - An NMR, X-ray crystallography and computational chemistry approach. *FEBS J.* **275**, 2524–2535 (2008).
  11. Bhunia, A., Bhattacharjya, S. & Chatterjee, S. Applications of saturation transfer difference NMR in biological systems. *Drug Discov. Today* **17**, 505–513 (2012).
  12. Carvalho, A. L., Santos-silva, T., Romão, M. J., Cabrita, E. J. & Marcelo, F. in *Essential Techniques for Medical and Life Scientists* 30–91 (2018).
  13. Mayer, M. & Meyer, B. Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew. Chemie - Int. Ed.* **38**, 1784–1788 (1999).
  14. Marcelo, F. *et al.* Delineating binding modes of Gal/GalNAc and structural elements of the molecular recognition of tumor-associated mucin glycopeptides by the human macrophage galactose-type lectin. *Chem. - A Eur. J.* **20**, 16147–16155 (2014).
  15. Ardá, A. *et al.* Molecular recognition of complex-type biantennary N-glycans by protein receptors: a three-dimensional view on epitope selection by NMR. *J. Am. Chem. Soc.* **135**, 2667–75 (2013).
  16. Mayer, M. & Meyer, B. Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. *J. Am. Chem. Soc.* **123**, 6108–6117 (2001).
  17. Dalvit, C. *et al.* Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water. *J. Biomol. NMR* **18**, 65–68 (2000).
  18. Braun, W., Bösch, C., Brown, L. R., Go, N. & Wüthrich, K. Combined use of proton-proton overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. Application to micelle-bound glucagon. *Biochim. Biophys. Acta* **667**, 377–396 (1981).

19. McIntosh, L. P. & Dahlquist, F. W. Biosynthetic Incorporation of  $^{15}\text{N}$  and  $^{13}\text{C}$  for Assignment and Interpretation of Nuclear Magnetic Resonance Spectra of Proteins. *Q. Rev. Biophys.* **23**, 1–38 (1990).
20. Teilum, K., Kunze, M. B. A., Erlendsson, S. & Kragelund, B. B. (S)Pinning down protein interactions by NMR. *Protein Sci.* **26**, 436–451 (2017).
21. Pervushin, K., Riek, R., Wider, G. & Wuthrich, K. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci.* **94**, 12366–12371 (1997).
22. Kerfah, R., Plevin, M. J., Sounier, R., Gans, P. & Boisbouvier, J. Methyl-specific isotopic labeling: A molecular tool box for solution NMR studies of large proteins. *Curr. Opin. Struct. Biol.* **32**, 113–122 (2015).
23. Kurauskas, V., Schanda, P. & Sounier, R. Membrane Protein Structure and Function Characterization. *Methods Mol Biol.* **1635**, 109–123 (2017).
24. Gelis, I. *et al.* Structural Basis for Signal-Sequence Recognition by the Translocase Motor SecA as Determined by NMR. *Cell* **131**, 756–769 (2007).
25. Danielson, M. A. & Falke, J. J. Use of  $^{19}\text{F}$  NMR to Probe Protein Structure and Conformational Changes. *Annu Rev Biophys Biomol Struct* **25**, 163–195 (1996).
26. Diercks, T. *et al.* Fluorinated carbohydrates as lectin ligands: versatile sensors in  $^{19}\text{F}$ -detected saturation transfer difference NMR spectroscopy. *Chem. - A Eur. J.* **15**, 5666–5668 (2009).
27. André, S. *et al.* Fluorinated carbohydrates as lectin ligands: Biorelevant sensors with capacity to monitor anomer affinity in  $^{19}\text{F}$ -NMR-based inhibitor screening. *European J. Org. Chem.* 4354–4364 (2012). doi:10.1002/ejoc.201200397
28. Unione, L. *et al.* Conformational Plasticity in Glycomimetics: Fluorocarbamethyl-L-idopyranosides Mimic the Intrinsic Dynamic Behaviour of Natural Idose Rings. *Chem. - A Eur. J.* **21**, 10513–10521 (2015).
29. Unione, L. *et al.* Fluoroacetamide Moieties as NMR Spectroscopy Probes for the Molecular Recognition of GlcNAc-Containing Sugars: Modulation of the CH– $\pi$  Stacking Interactions by Different Fluorination Patterns. *Chem. - A Eur. J.* **23**, 3957–3965 (2017).
30. Manglik, A. *et al.* Structural insights into the dynamic process of  $\beta$ 2-adrenergic receptor signaling. *Cell* **161**, 1101–1111 (2015).
31. Liu, J. J., Horst, R., Katritch, V., Stevens, R. C. & Wüthrich, K. Biased signaling pathways in  $\beta$ 2-adrenergic receptor characterized by  $^{19}\text{F}$ -NMR. *Science* **335**, 1106–1110 (2012).
32. Rydzik, A. M. *et al.* Monitoring conformational changes in the NDM-1 metallo- $\beta$ -lactamase by  $^{19}\text{F}$  NMR spectroscopy. *Angew. Chemie - Int. Ed.* **53**, 3129–3133 (2014).
33. Marsh, E. N. G. & Suzuki, Y. Using  $^{19}\text{F}$  NMR to Probe Biological Interactions of



- Proteins and Peptides. *ACS Chem. Biol.* **9**, 1242–1250 (2014).
34. Bann, J. G. & Frieden, C. Folding and domain-domain interactions of the chaperone PapD measured by 19F NMR. *Biochemistry* **43**, 13775–13786 (2004).
  35. Kitevski-Leblanc, J. L., Hoang, J., Thach, W., Larda, S. T. & Prosser, R. S. 19F NMR studies of a desolvated near-native protein folding intermediate. *Biochemistry* **52**, 5780–5789 (2013).
  36. Watt, G. D. A new future for carbohydrate fuel cells. *Renew. Energy* **72**, 99–104 (2014).
  37. Laine, R. A. Invited commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields  $1.05 \times 10$  structures for a reducing hexasaccharide: The Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* **4**, 759–767 (1994).
  38. Juaristi, E. & Cuevas, G. Recent studies of the anomeric effect. *Tetrahedron* **48**, 5019–5087 (1992).
  39. Lii, J., Chen, K., Durkin, K. A. & Allinger, N. L. Alcohols, Ethers, Carbohydrates, and Related Compounds . II . The Anomeric Effect \*. *J. Comp. Chem.comp* **24**, 1473- (2003).
  40. Angyal, S. J. The Composition and Conformation of Sugars in Solution. *Angew. Chemie Int. Ed. English* **8**, 157–166 (1969).
  41. Sattelle, B. M. & Almond, A. Is N-acetyl-d-glucosamine a rigid 4C1 chair? *Glycobiology* **21**, 1651–1662 (2011).
  42. Glaudemans, C. P. J. *et al.* Significant Conformational Changes in an Antigenic Carbohydrate Epitope upon Binding to a Monoclonal Antibody. *Biochemistry* **29**, 10906–10911 (1990).
  43. Biarnés, X. *et al.* The conformational free energy landscape of beta-D-glucopyranose. Implications for substrate preactivation in beta-glucoside hydrolases. *J. Am. Chem. Soc.* **129**, 10686–10693 (2007).
  44. Varki, A. Biological Roles of Glycans. *Glycobiology* **27**, 3–49 (2017).
  45. Ardá, A. & Jiménez-Barbero, J. The recognition of glycans by protein receptors. Insights from NMR spectroscopy. *Chem. Commun.* **54**, 4761–4769 (2018).
  46. Piotukh, K., Serra, V., Borriss, R. & Planas, A. Protein-carbohydrate interactions defining substrate specificity in Bacillus 1,3-1,4-β-D-glucan 4-glucanohydrolases as dissected by mutational analysis. *Biochemistry* **38**, 16092–16104 (1999).
  47. Gabius, H. J., André, S., Jiménez-Barbero, J., Romero, A. & Solís, D. From lectin structure to functional glycomics: Principles of the sugar code. *Trends Biochem. Sci.* **36**, 298–313 (2011).
  48. Jimenez-Moreno, E. *et al.* A thorough experimental study of CH/π interactions in water: quantitative structure– stability relationships for carbohydrate/aromatic complexes. *Chem. Sci.* **6**, 6076–6085 (2015).

49. Ramirez-Gualito, K. *et al.* Enthalpic Nature of the CH /  $\pi$  Interaction Involved in the Recognition of Carbohydrates by Aromatic Compounds , Confirmed by a Novel Interplay of NMR , Calorimetry , and Theoretical Calculations. *J. Am. Chem. Soc.* **131**, 18129–18138 (2009).
50. Asensio, J. L., Ardá, A., Cañada, F. J. & Jiménez-Barbero, J. Carbohydrate-aromatic interactions. *Acc. Chem. Res.* **46**, 946–954 (2013).
51. Fernandez-Alonso, M. del C., Cañada, F. J., Jiménez-Barbero, J. & Cuevas, G. Molecular Recognition of Saccharides by Proteins. Insights on the Origin of the Carbohydrate - Aromatic Interactions. *J. Am. Chem. Soc.* **127**, 7379–7386 (2005).
52. del Carmen Fernández-Alonso, M. *et al.* Protein-carbohydrate interactions studied by NMR: from molecular recognition to drug design. *Curr. Protein Pept. Sci.* **13**, 816–30 (2012).
53. Bermejo, I. A. *et al.* Water Sculpts the Distinctive Shapes and Dynamics of the Tumor-Associated Carbohydrate Tn Antigens: Implications for Their Molecular Recognition. *J. Am. Chem. Soc.* **140**, 9952–9960 (2018).
54. Nieto, L., Canales, Á., Giménez-Gallego, G., Nieto, P. M. & Jiménez-Barbero, J. Conformational selection of the AGA\*IA M heparin pentasaccharide when bound to the fibroblast growth factor receptor. *Chem. - A Eur. J.* **17**, 11204–11209 (2011).
55. Broecker, F. *et al.* Multivalent display of minimal *Clostridium difficile* glycan epitopes mimics antigenic properties of larger glycans. *Nat. Commun.* **7**, 1–12 (2016).
56. Coelho, H. *et al.* The Quest for Anticancer Vaccines: Deciphering the Fine-Epitope Specificity of Cancer-Related Monoclonal Antibodies by Combining Microarray Screening and Saturation Transfer Difference NMR. *J. Am. Chem. Soc.* **137**, 12438–12441 (2015).
57. Amoah, A. S. *et al.* Identification of dominant anti-glycan IgE responses in school children by glycan microarray. *J. Allergy Clin. Immunol.* **141**, 1130–1133 (2018).
58. Rodriguez, M. C. *et al.* Thermodynamic Switch in Binding of Adhesion/Growth Regulatory Human Galectin-3 to Tumor-Associated TF Antigen (CD176) and MUC1 Glycopeptides. *Biochemistry* **54**, 4462–4474 (2015).
59. Durocher, Y. & Butler, M. Expression systems for therapeutic glycoprotein production. *Curr. Opin. Biotechnol.* **20**, 700–707 (2009).
60. Sirohi, D. *et al.* The 3.8 Å resolution cryo-EM structure of Zika virus. *Science* **352**, 467–70 (2016).
61. Ardá, A. *et al.* in *Carbohydrate Chemistry: Volume 42* 47–82 (2017).

## 1.2 *Goals*

From the training perspective, the key objective of this Thesis has been to acquire knowledge on the application of NMR methods to study distinct molecular recognition events.

Therefore, ligand- and receptor-based NMR methods were used to study the interaction of a variety of ligands with diverse chemical nature with receptors of biological and/or biomedical interest. Within the explored ligands, glycans have been specially employed. Thus, the scientific objective of this thesis has been to advance in the understanding of the structural features that govern glycan interactions, including the derivation of the binding epitopes.

Within glycoscience field, the specific objectives have aimed to:

- understand of the mechanism of action of GalNAc-Ts.
  
- evaluate the interactions of different natural and synthetic molecules with antimicrobial peptides and MD-2 protein, in the context of understanding TLR4-related innate immunity.



# *Chapter*

# **2**

*Deciphering GalNAc O-glycosylation:*

*From structure to function in human health & disease*



## **2.1 Introduction**

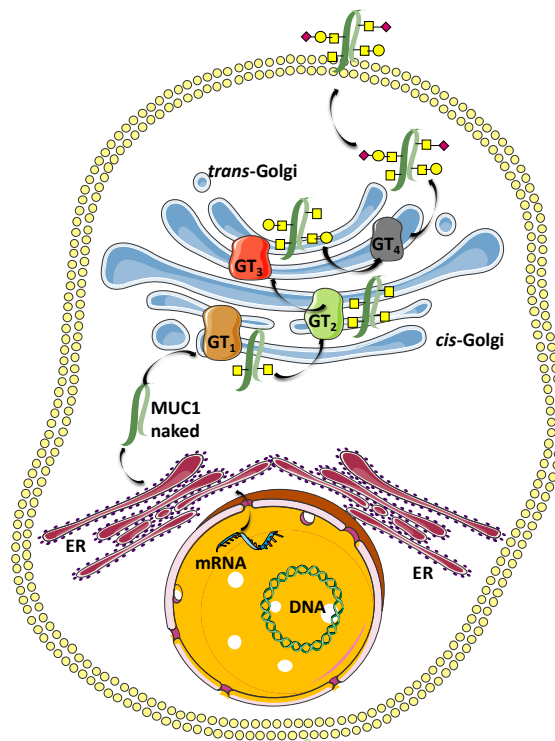
Glycosylation is the most complex and common modification of proteins and lipids, with a supreme capacity to generate a wide array of different structures.<sup>1</sup> Protein glycosylation is a complex and multifaceted post-translational modification, where several variables can be modified. They include the type of sugar-protein linkage, the sugars that are transferred and the resulting glycan structure and length. Recognition of glycans by specific receptors translates the glycome into unprecedented biological signals in diverse physiological and pathological events.<sup>2</sup> For example, human cell surface mucin-type glycoproteins play many crucial roles in biological processes and significantly influence specific cellular adhesion during differentiation, proliferation, or malignant alteration in embryogenesis, organogenesis, carcinogenesis, and cancer metastasis.<sup>3</sup>

The specific and strictly controlled glycosylation of proteins depends on the action of highly “specialized” enzymes known as glycosyltransferases (GTs) and glycosidases, which are precisely located in different organelles in cells, thus showing cell and tissue specificities.

Two of the most abundant forms of glycosylation occurring on proteins destined to be secreted or membrane-bound are dubbed *N*-linked (to asparagine) and mucin-type *O*-linked (to serine or threonine). Glycan chains of mucin glycoproteins are commonly constructed by a variety of GTs at the endoplasmic reticulum (ER) and/or the Golgi apparatus membranes through highly complicated biosynthetic pathways.<sup>4</sup> In the ER, where *N*-glycosylation is initiated, folded proteins exit using the COPII-coated protein carriers and are exported to the Golgi complex. Herein, *N*-glycans are modified and *O*-glycosylation is initiated.<sup>5</sup>

The Golgi complex is a collection of cisternae connected by tubules, which can be divided into *cis*, *medial*, and *trans* compartments.<sup>5</sup> This compartmentation permits that those proteins penetrating through the Golgi encounter the various enzymes for just a limited time. In particular, GTs are usually distributed by their

order of action: early-acting enzymes are localized in the *cis*-Golgi, while late-acting enzymes are concentrated in the *trans*-Golgi (Figure 2.1.1). It is well established that the main factors that contribute to the regulation of *O*-glycosylation in the Golgi is the competing activities of the glycosyltransferases and their localization.



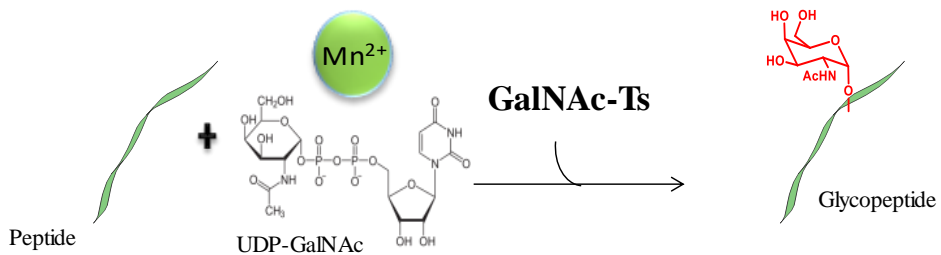
**Figure 2.1.1-** Schematic representation of the biosynthesis of mucin-type *O*-glycans.

### 2.1.1 GalNAc-Transferases

The large GalNAc-transferase (GalNAc-Ts) family is responsible to initiate the post-translational modification of *O*-GalNAc glycosylation of many cell-surface proteins.<sup>6</sup> *O*-GalNAc glycosylation is the most complex and differentially regulated type of protein glycosylation. The initial step of protein *O*-GalNAc



glycosylation pathway is mediated by a large family of GalNAc-Ts, up to 20, localized in the Golgi complex (Figure 2.1.1), which catalyze the transfer of *N*-acetylgalactosamine (GalNAc) from a sugar donor (UDP-GalNAc) to the Ser/Thr side chains on proteins/peptides (Figure 2.1.2).<sup>6</sup>



**Figure 2.1.2-** Schematic representation of the enzymatic reaction mediated by GalNAc-Ts.

Mucin type-*O*-glycosylation elongation was first thought to mainly occur in the luminal regions of the *cis*-Golgi. However, recent studies have indicated that it occurs throughout the *cis*-, *medial*-, and *trans*- Golgi regions. In fact, the localization and compartmentalization of GalNAc-Ts and other glycosyltransferases varies among these regions of the Golgi.<sup>7,8</sup> These are key factors that contribute to the regulation of *O*-glycosylation in the Golgi together with the competing activities of the GTs. Other factors include the concentrations of the metal ion Mn<sup>2+</sup>, the donor (UDP-GalNAc) and the acceptor substrates, the substrate transport rate through the Golgi, and the luminal environment of the Golgi (including pH).<sup>7-9</sup>

Mammalian GalNAc-Ts are distributed among all organs with isoforms having different expression levels in each tissue. Previous studies have shown that GalNAc-T1 and T2 mRNAs are ubiquitously (>70%) expressed among all organs.<sup>10</sup> The expression of other transferase isoforms, such as GalNAc-T5, T8, T9, T10, T13, T15, T17, T19 and T20 are less ubiquitous and more regulated to specific

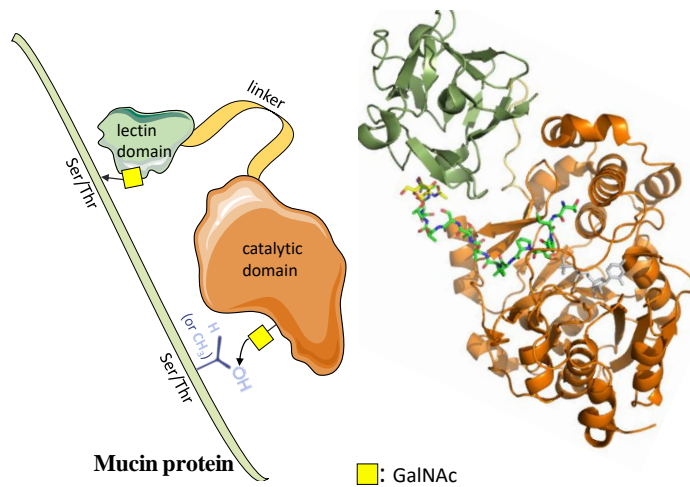
organs.<sup>6</sup> Lastly, the remaining transferases, GalNAc-T3, T4, T6, T11, T12, T16 and T18 have more broad expression levels (< 30%) spread across tissues.<sup>6</sup>

A number of distinct GalNAc-Ts have been associated with human diseases. The loss of GalNAc-T3 causes familial tumoral calcinosis due to an unregulated cleavage of FGF23.<sup>11</sup> Others include GalNAc-T2, T5 and T19, which have been linked to levels of HDL cholesterol and coronary artery disease,<sup>12,13</sup> hereditary multiple exostoses<sup>14</sup> and the Williams-Beuren Syndrome,<sup>15</sup> respectively.

GalNAc-Ts also display marked changes in their expression during malignant transformation.<sup>6,16</sup> It has been reported, using cancer cell lines, that the GalNAc-Ts can relocate to the ER, through the regulation of *Src* (a proto-oncogene).<sup>17</sup> This fact, in turn, increases the levels of the Tn-antigen ( $\alpha$ -GalNAc-*O*-Ser/Thr) and enhances tumor cell migration and invasiveness.<sup>17-19</sup> These findings indicate that the localization of the GalNAc-Ts heavily relies on the physiological state of the cell.

### **2.1.1.1 Structure**

GalNAc-Ts are type-II Golgi membrane proteins, with a extracellular domain structure that includes a short N-terminal cytoplasmic tail and a hydrophobic transmembrane domain. Structurally, the extracellular domain of GalNAc-Ts displays a unique two-domain architecture consisting of a *N*-terminal catalytic domain tethered by a short flexible linker to a *C*-terminal lectin domain (Figure 2.1.3). The catalytic domain is much larger than the lectin domain, and consists of ~230 amino acids, while the lectin domain has only ~120 amino acid residues. Both domains are linked via a short flexible linker that varies from ~10 to 25 amino acids in length among isoforms, thus creating flexibility between domains (Figure 2.1.3).



**Figure 2.1.3**-Left: Scheme of the structure of GalNAc-Ts. To yellow the catalytic domain and to green the lectin domain bound by a flexible strand, in yellow. Right: Crystal structure of GalNAc-T2 in complex with UDP and MUC5AC-13 (pdb ID: 5AJP), adapted from <sup>20</sup>.

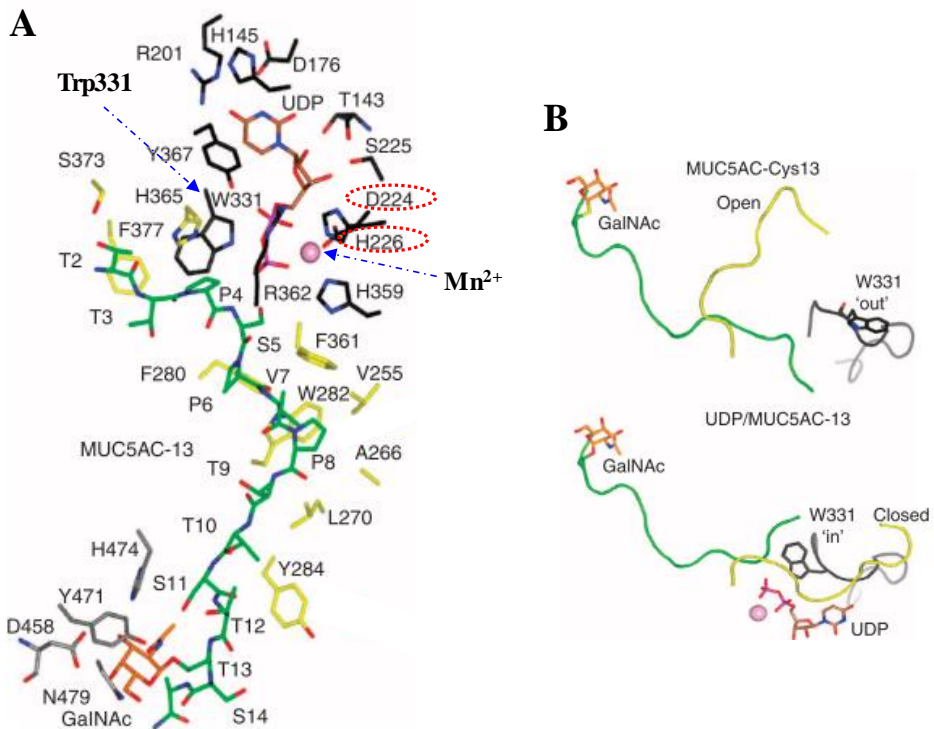
The catalytic domain is responsible for all catalytic activity: 1) the binding of the acceptor substrate and the UDP-GalNAc, and 2) the subsequent transfer of GalNAc onto the acceptor substrate. The catalytic domain is responsible to initiate glycosylation in naked acceptor substrates and to proceed short-range glycosylation in partially glycosylated substrates.<sup>21</sup>

The C-terminal lectin domain of GalNAc-Ts, which belongs to the Ricin-type lectin family, specifically recognizes GalNAc residue and is responsible to ensure long-range-glycosylation on a prior GalNAc glycosylated substrates. It consists of a folded  $\beta$ -trefoil structure built from three homologous repeat units, the  $\alpha$ ,  $\beta$  and  $\gamma$  repeats.<sup>6</sup> The  $\beta$ -trefoil repeats all share a common binding motif, known as the carbohydrate lectin domain (CLD) and QxW motifs (where x can be any amino acid) that are ~40 amino acids apart. One Asp residue of the CLD motif is critical for the lectin binding properties. Mutagenesis studies on GalNAc-Ts showed that only one or two specific subdomains ( $\alpha$ ,  $\beta$  or  $\gamma$ ) actively bind GalNAc. For example, in GalNAc-T1, the  $\alpha$  and  $\beta$  subdomains are important for GalNAc binding. For GalNAc-T2, T3 and T4 only the  $\alpha$ -subdomain recognizes the sugar,

while for GalNAc-T10, the  $\beta$ -subdomain was determined to be important for GalNAc-binding.<sup>9,22–27</sup>

One more important feature is the  $Mn^{2+}$  coordinating DxH motif at the catalytic domain (Figure 2.1.4). This is responsible for binding the phosphate moiety of the UDP-GalNAc through coordination of the  $Mn^{2+}$  ion. Previous studies showed that metal ions are required for catalysis, being  $Mn^{2+}$  the preferred divalent metal ion,<sup>28</sup> which allows the required octahedral geometry.

Another unique structural feature of the catalytic domain of GalNAc-Ts is the flexible loop that extends out of the UDP-GalNAc and substrate binding sites. This loop undergoes drastic conformational changes between the so called open (inactive) and closed (active) conformations (Figure 2.1.4). The active form appears upon binding of the UDP-GalNAc and the substrate.<sup>20,22,29</sup> Recent structural studies<sup>20</sup> have shown that, during the catalytic cycle, the flexible loop of GalNAc-T2 exists in multiple conformations; either semi-open, open and closed (Figure 2.1.4). This conformational equilibrium is coupled to the interactions of a key catalytic residue, identified as Trp<sub>331</sub>. In the active form (closed), the Trp<sub>331</sub> adopts the *in* conformation, and it is located fairly close to the UDP-GalNAc and the acceptor substrate. In contrast, when the enzyme is in its inactive (open) conformation, the Trp<sub>331</sub> displays the *out* conformation, being away from the UDP-GalNAc and the acceptor substrate (Figure 2.1.4).<sup>20</sup>



**Figure 2.1.4 - A.** Structural features of GalNAc-T2-UDP-MUC5AC-13 complex. DxH motif in red highlight and the Trp331 and Mn<sup>2+</sup> identified with blue arrow. **B.** The two conformations called open (inactive) and closed (active) and the W331 “in” and “out” was showed with GalNAc-T2-MUC5AC-Cys13, GalNAc-T2-MUC5AC-13-UDP complexes.<sup>20</sup>

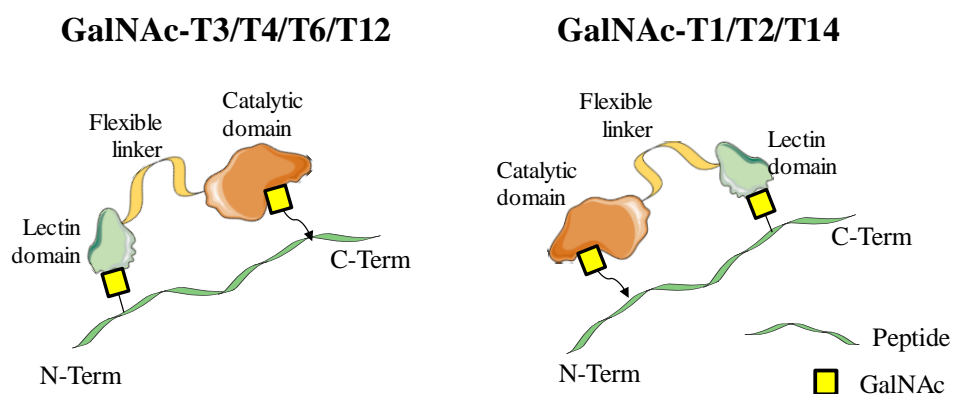
### 2.1.1.2 Preferences and Specificity

Currently there is no consensus on the specific peptide motif that undergoes O-glycosylation. To initiate the O-linked glycosylation process there are up to twenty GalNAc-Ts isoforms that can attach GalNAc, as opposed to the only one transferase enzyme that exists for N-linked glycosylation. GalNAc-Ts show distinct, although partly overlapping, kinetic properties and acceptor substrate specificities.<sup>9</sup> Thus, it is the available repertoire of GalNAc-Ts the factor that determines which proteins are O-glycosylated and where. In this context, in one hand there are individual GalNAc-Ts exclusively responsible to glycosylate a single site in proteins<sup>30-33</sup> (non-redundant glycosylation). In other hand,

glycosylation of mucins, proteins that have multiple sites of glycosylation, is covered by multiple GalNAc-Ts (redundant glycosylation).

The recognition of *O*-glycosylation sites by GalNAc-Ts depends whether these enzymes interact with naked or previous glycosylated regions of the substrate. While naked peptides appear to be exclusively recognized by the catalytic domain, glycopeptides recognition rely on the existence of a cooperative mechanism between the catalytic and lectin domains (Figure 2.1.3).<sup>34</sup> Additionally, Thr is a much better acceptor than Ser, while the preferred glycosylation sites are associated with high contents of Pro, Gly, and, obviously, Ser and Thr. Other predictive trend among most GalNAc-Ts is a common motif where a Pro residue is essential at +3 position and positive effect in positions -1, +1, +2 and +4 from the site of glycosylation (PxP motif, where x can be any amino acid).<sup>20,35,36</sup>

The lectin domain mediates the GalNAc-peptide substrate specificity thus increasing the efficiency of the enzymatic activity and modulating glycosylation site specificity in partially glycosylated acceptor substrates.<sup>23</sup> Lectin binding to the GalNAc attached at the glycopeptide directs the peptide acceptor onto the catalytic domain in a particular direction (N- and/or C-terminal) (Figure 2.1.5).<sup>21</sup> For example, GalNAc-T3/T4/T6/T12 preferentially glycosylate those C-terminal sites that are remote from the prior N-terminal GalNAc site,<sup>21,37</sup> whereas GalNAc-T1/T2/T14 exhibit the opposite preference.<sup>20,29,38</sup> Other GalNAc-Ts, such as GalNAc-T5/T13/T16 do not exhibit orientation preferences for long-range glycosylation.<sup>21</sup>



**Figure 2.1.5** – Schematic representation of preferences of two GalNAc-Ts groups.

The present chapter focuses on the application of a multidisciplinary approach combining different experimental and theoretical techniques and methods to provide new structural insights with atomic resolution to understand the mechanism of action of GalNAc-Ts.





## ***2.2 GalNAc-Ts glycosylation follow an induced-fit catalytic mechanism.***

*The work presented in this subchapter has been performed in collaboration with different research groups*

*-Matilde de las Rivas (PhD student) and Dr. Erandi Lira-Navarrete, at the laboratory of Dr. Ramon Hurtado-Guerrero (BIFI, University of Zaragoza) performed the expression and purification of GalNAc-Ts and the X-Ray crystallography experiments.*

*- Dr. Francisco Corzana at the Universidad de La Rioja was responsible for the Molecular Dynamics simulations, while Ismael Compañón (PhD student) was in charge of the synthesis of the peptide.*

*- Dr. Henrik Clausen and Dr. Sergey Y. Vakhrushev Labs at the Copenhagen Center for Glycomics, University of Copenhagen, Denmark performed the MS/MS assays.*

---

**Publication:** Matilde de las Rivas, **Helena Coelho**, Ana Diniz, Erandi Lira-Navarrete, Ismael Compañón, Jesús Jiménez-Barbero, Katrine T. Schjoldager, Eric P. Bennett, Sergey Y. Vakhrushev, Henrik Clausen, Francisco Corzana, Filipa Marcelo, Ramon Hurtado-Guerrero (2018) *Chemistry - A European Journal*, **24**, 8382 –8392. **(co-first author)**.

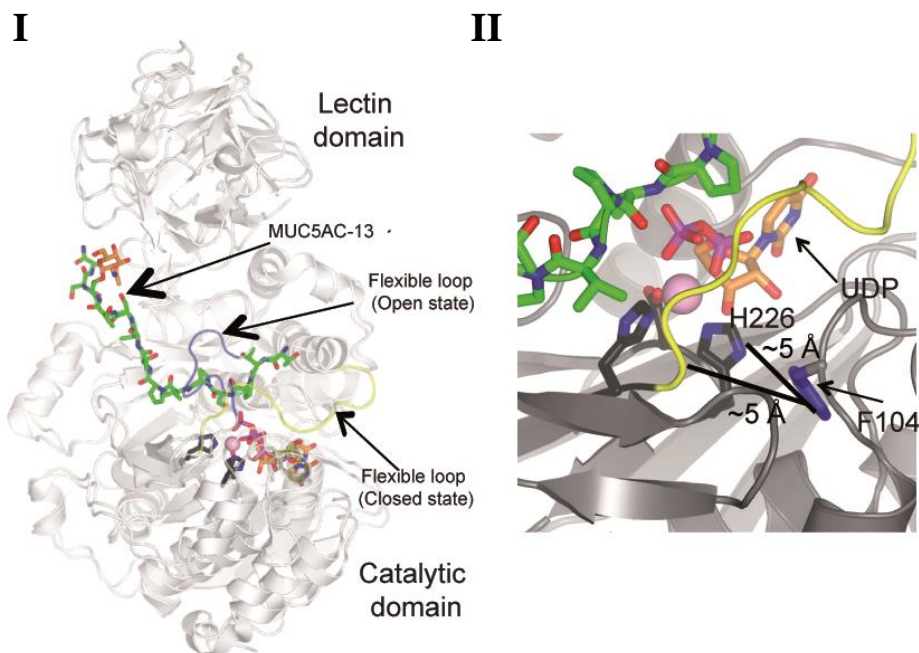


### 2.2.1 Introduction

Deficiencies and dysregulation of individual GalNAc-Ts have been found to cause diseases and predispositions. Recently, it was demonstrated that a GALNT2 mutant (F<sub>104</sub>S) leading to the inactivation of the enzyme, induces low levels of high-density lipoprotein cholesterol (HDL-C) in humans. It was also discovered that the low levels of HDL-C were likely due to the disruption of the *O*-glycosylation of phospholipid transfer protein (PLTP), which led to a decrease in its activity.<sup>32</sup>

Structurally, GalNAc-Ts display a unique characteristic among glycosyltransferases: apart of the N-terminal catalytic domain, they also possess a unique C-terminal lectin domain, being both domains connected by a short flexible linker (Figure 2.2.1). Another interesting structural feature of these enzymes is that there is a flexible loop in the catalytic domain, comprising residues Val<sub>360</sub> to Gly<sub>372</sub> in GalNAc-T2.<sup>20,29</sup> The flexible loop present in the catalytic domain may adopt different conformations during the catalytic cycle, thus dictating the catalytic activity of the enzyme, either active or inactive.<sup>20,29,39,40</sup> This unique conformational behavior of the flexible loop has been reported exclusively by using X-ray crystallography.<sup>20,29,39,40</sup> Thus, the significance of this structural element required additional experiments in solution to address the dynamic and association of this loop to GalNAc-T2 enzyme kinetics.

Structural studies on GalNAc-Ts may also serve as a platform to infer how mutations on these enzymes might lead to a loss-of-function.<sup>20,22,25,29,39,40</sup> The F<sub>104</sub>S mutant in GalNAc-T2 precisely illustrates this idea, since Phe104 is not located at the active site, but placed at a distance of ~5 Å from the flexible loop or the D<sub>224</sub>XH<sub>226</sub> motif, which in turn coordinates a manganese ion together with His<sub>359</sub> (Figure 2.2.1)<sup>20,22,25,29,39,40</sup> Thus, it is not obvious how the mutation of Phe to Ser may impair GalNAc-T2 activity. The objective this work was to understand how the mutation F<sub>104</sub>S in GalNAc-T2 leads to a loss-of-function.

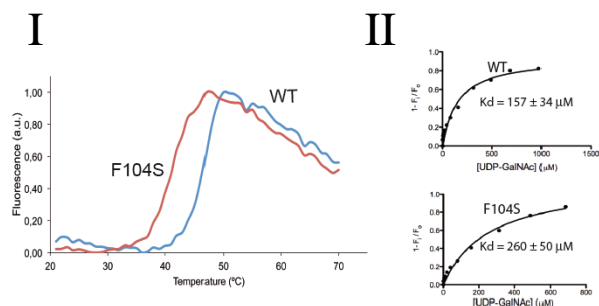


**Figure 2.2.1** - States of the flexible loop and location of Phe<sub>104</sub>. **(I)** Superimposition of the crystal structures of the complexes of GalNAc-T2 with UDP and the MUC5AC-13<sup>20</sup> glycopeptide (white color; PDB: 5AJP, with that of GalNAc-T2 with UDP (white color; PDB: 2FFV). The GalNAc residue in MUC5AC-13 is colored in orange. UDP and MUC5AC-13 are depicted as orange and green carbon atoms, respectively. The D<sub>224</sub>XH<sub>226</sub> motif/His<sub>359</sub> site is shown as black carbon atoms. The flexible loop is depicted in blue (open conformation) and yellow (closed conformation) for the PDB entries 2FFV and 5AJP, respectively. Mn<sup>+2</sup> is shown as a pink sphere. **(II)** Close-up view of the active site (PDB entry 5AJP) showing the distance of Phe<sub>104</sub> from the flexible loop and His<sub>226</sub>. The structure is shown in grey. The other structural features/amino acids are shown with the same colors above except Phe<sub>104</sub> is depicted as blue carbon atoms.

## 2.2.2 Results and Discussion

### 2.2.2.1 *Stability and Binding assays*

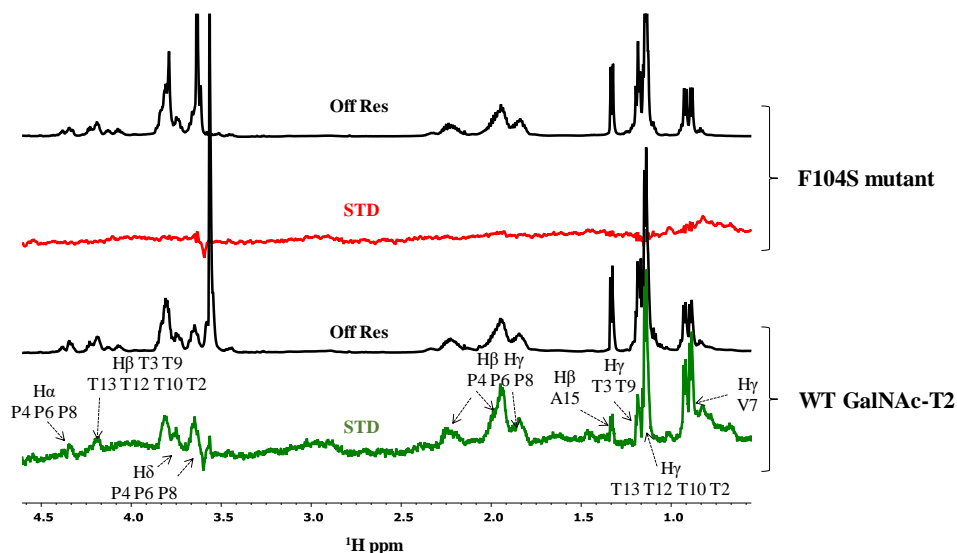
A thermal shift-assay experiment was performed to evaluate the stability of WT GalNAc-T2 and its F<sub>104</sub>S mutant. The results clearly evidenced that the WT enzyme was 5°C more than the mutant F<sub>104</sub>S (with denaturation temperature of  $47 \pm 0.09^\circ\text{C}$  for the WT enzyme  $42 \pm 0.05^\circ\text{C}$  for the mutant), implying that the mutation leads to a significant decrease on the stability of the mutant (Figure 2.2.2). In addition, the tryptophan fluorescence spectroscopy analysis of UDP-GalNAc binding revealed that both enzymes bind UDP-GalNAc in a similar manner, being the  $K_d$  of to the mutant 1.7-fold higher than that determined for the WT enzyme (Figure 2.2.2).



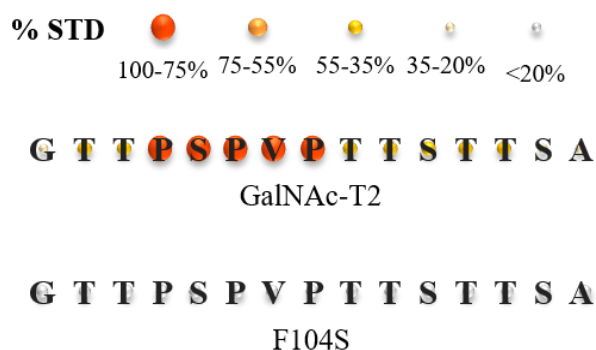
**Figure 2.2.2** - Stability and binding properties of the WT GalNAc-T2 and the F<sub>104</sub>S mutant. (I) Thermal denaturation curves of the WT GalNAc-T2 and the F<sub>104</sub>S mutant as monitored by fluorescence using 1-anilino-8-naphthalene sulfonate (ANS) as probe. (II)  $K_d$  values for UDP-GalNAc binding determined from the tryptophan fluorescence intensity as a function of the nucleotide concentration.

Fittingly, STD-NMR experiments of the peptide substrate MUC5AC only showed STD response in the presence of the WT enzyme, but not with the F<sub>104</sub>S mutant (Figure 2.2.3 and Figure 2.2.4). This result points out that the mutant is not able to properly recognize the peptide substrate, thus explaining the observed

inactivity of the mutant towards the protein substrate, PLTP.<sup>32</sup> In addition, the WT enzyme evidenced the strongest STD-NMR signals for the “PSPVPT” fragment within the peptide sequence. It is noteworthy mentioning that this Thr residue constitutes the main acceptor site on this peptide for GalNAc-T2.



**Figure 2.2.3** - STD-NMR experiments were recorded at 298K and 600MHz. (Up) STD experiments (Off resonance spectrum at the top, in black) for MUC5AC (740  $\mu$ M) in the presence of the F<sub>104</sub>S mutant (18.5  $\mu$ M), UDP (75  $\mu$ M) and MnCl<sub>2</sub> (75  $\mu$ M). (Down) STD experiments (Off resonance spectrum in black) for MUC5AC (560  $\mu$ M) in the presence of WT GalNAc-T2 (14  $\mu$ M), UDP (75  $\mu$ M) and MnCl<sub>2</sub> (75  $\mu$ M).

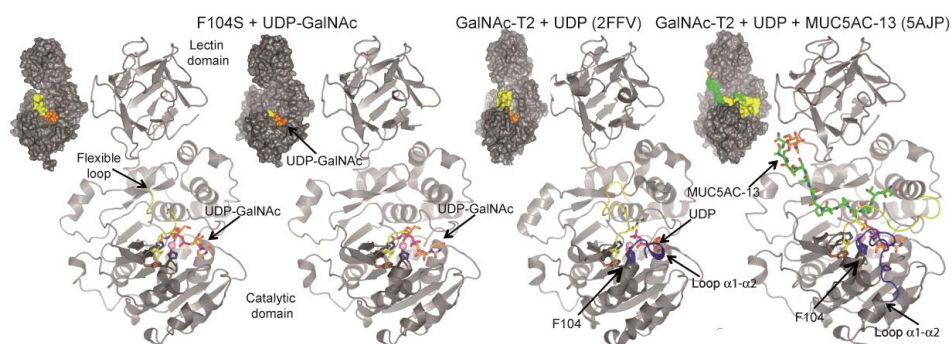


**Figure 2.2.4-** STD-NMR-derived epitope mapping of MUC5AC. The different colored spheres indicate the normalized STD signal (in percent) observed for each proton.

The STD-NMR derived epitope is in agreement with the importance of the P<sub>x</sub>P motif previously described to be essential for GalNAc-Ts recognition in a variety of protein substrates. For the interaction with the WT enzyme, Thr9 of MUC5AC was the best-recognized acceptor site, which also agrees with previous results that demonstrated that this residue is the most glycosylated one.<sup>38</sup> Hence, while there are not major differences in recognition to UDP-GalNAc between the WT enzyme and the mutant, the differences in activity are accounted for the inability of the mutant F<sub>104</sub>S to bind peptide substrates.

#### **2.2.2.2 X-Ray Crystallography & MD Simulations**

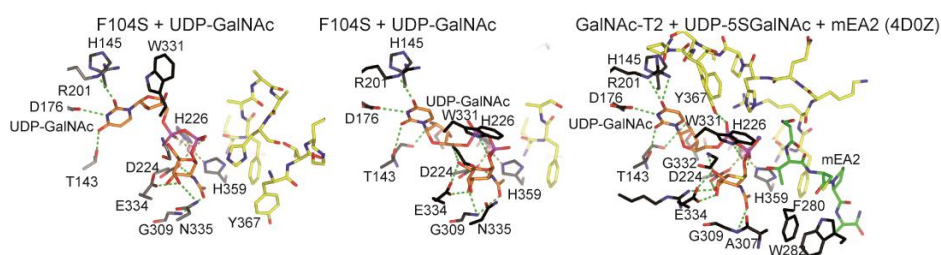
To clarify the molecular basis of why the F<sub>104</sub>S mutant does not bind to the MUC peptides, tetragonal crystals of this mutant in complex with UDP were obtained, which were subsequently soaked with UDP-GalNAc. The resulting crystal allowed us to solve the corresponding structure at 2.70 Å resolution. The crystal structure shows a compact structure with the typical GT-A fold for the catalytic domain at the N-terminus, while the lectin domain is located at the C-terminal region (Figure 2.2.5).



**Figure 2.2.5** - Crystal structure of the  $F_{104}S$  mutant. Cartoon representation of the overall mutant  $F_{104}S$  in complex with UDP-GalNAc (**left**) and the WT (**right panels**) either in complex with UDP<sup>29</sup> (PDB entry 2FFV) or UDP/MUC5AC-13<sup>20</sup> (PDB entry 5AJP). The proteins and the GalNAc moiety of MUC5AC-13 are colored in grey and orange, respectively. UDP/UDP-GalNAc (orange), MUC5AC-13 (green), the flexible loop (yellow), and the loop  $\alpha 1$ - $\alpha 2$  (blue) are also highlighted. The Phe<sub>104</sub>, Val<sub>360</sub>/Arg<sub>362</sub> and the D<sub>224</sub>X/H<sub>226</sub> motif/His<sub>359</sub> are shown as sticks in blue, yellow and black carbon atoms, respectively. Other residues interacting with Phe<sub>104</sub> are shown as brown carbon atoms. The Mn<sup>+2</sup> ion is shown as a pink sphere.

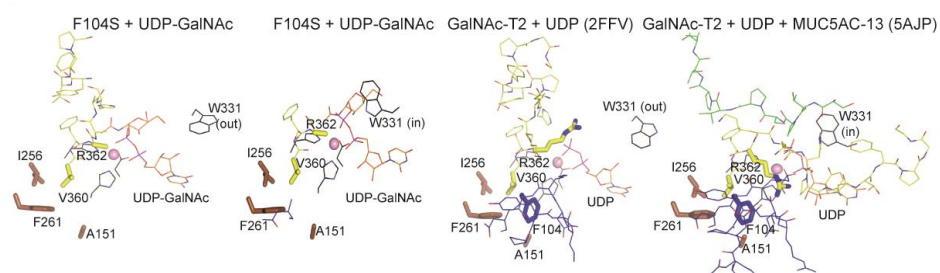
The crystal structure shows that the flexible loop in the  $F_{104}S$  mutant is disordered (Figure 2.2.5), exposing the UDP-GalNAc to the solvent. The UDP moiety in the PDB entry 2FFV is also exposed to the solvent. In this structure, the flexible loop displays an open conformation rendering the enzyme in an inactive state (Figure 2.2.5). Thus, the structure of the mutant  $F_{104}S$  with UDP-GalNAc resembles the structure of the WT enzyme in its inactive state. This conformation is completely different to that observed in the active state. In this case, the flexible loop displays a closed conformation that functions covering either UDP-GalNAc or UDP from the solvent (e.g. see PDB entries 5AJP, 4D0T and 4D0Z; Figure 2.2.5). This particular geometry is required to bind the peptide/protein substrates.<sup>20,29</sup> The closed conformation of the flexible loop is stabilized by interactions of His<sub>365</sub> and Phe<sub>369</sub> with Trp<sub>331</sub> (in-conformation) and by additional interactions with the donor substrate UDP-GalNAc (Figure 2.2.6).<sup>29</sup> Thus, it is important mentioning that the active state can only be achieved when the flexible loop adopts the close conformation and the Trp<sub>331</sub> displays the “in-conformation”.





**Figure 2.2.6** - Close-up view of the sugar nucleotide and peptide binding site of the  $F_{104S}$ -UDP-GalNAc and WT-UDP-5S-GalNAc-mEA2 complexes. Hydrogen bond interactions are shown as dotted green lines.

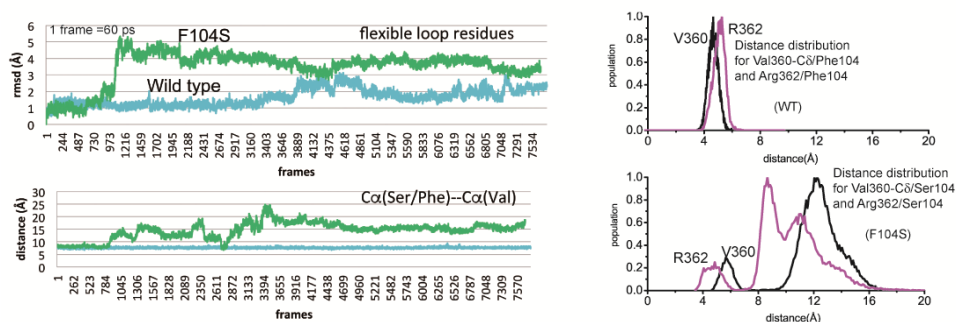
The detailed analysis of the structure of the mutant permitted to observe that there was not enough electron density to properly define the orientation of Ser<sub>104</sub> along with most of the residues of the loop between the secondary structures  $\alpha 1$  and  $\alpha 2$  (named hereafter loop  $\alpha 1$ - $\alpha 2$ ) (Figure 2.2.5). Thus, it is tempting to suggest that these disordered states of both the flexible loop and the loop  $\alpha 1$ - $\alpha 2$  in the inactive mutant might also explain why this mutant is 5°C less than the WT enzyme. On the other hand, for WT GalNAc-T2, the Phe<sub>104</sub> residue establishes hydrophobic stacking interactions with Ala<sub>151</sub>, Ile<sub>256</sub>, Val<sub>360</sub> and Phe<sub>261</sub>, probably stabilizing the loop  $\alpha 1$ - $\alpha 2$  in the WT enzyme and the flexible loop in the inactive state (Figure 2.2.7). In the active state of the enzyme, Phe<sub>104</sub> also establishes a cation- $\pi$  interaction with Arg<sub>362</sub> (Figure 2.2.7), which is located at the flexible loop. This interaction appears to be key to render the closed conformation. The mutation to Ser<sub>104</sub>, which is a polar residue, probably disrupts the interactions with the hydrophobic/aromatic residues, leading to the destabilization of the loop  $\alpha 1$ - $\alpha 2$  and the inability of reaching the required closed conformation for the flexible loop.



**Figure 2.2.7** - Close-up view of the complete sugar nucleotide and partial peptide-binding site of the different complexes. The residues interacting with  $Phe_{104}$  are magnified.  $Trp_{331}$  is shown in the “out” and “in conformations”.

To understand how the  $F_{104}S$  mutation disturbs the dynamics of the flexible loop, molecular dynamics (MD) simulations (during 200 ns) were performed in explicit water for both the WT GalNAc-T2 and the  $F_{104}S$  mutant in presence of UDP-GalNAc. The analysis of the data permitted to conclude that the flexible loop of the  $F_{104}S$  mutant exhibited significantly larger conformational changes, with root-mean-square-deviation (RMSD) ranging between 1-5.3 Å, as compared to 1-3 Å deviations for the WT analogue (Figure 2.2.8). The variations of the  $C\alpha$  distances between the  $Phe_{104}$  or  $Ser_{104}$  residues and  $Val_{360}$  were also rather different highlighting the different flexibilities between the two structures. A shorter distance of  $5.7\pm 4$  Å was found between  $Phe_{104}$  and  $Val_{360}$ , while highly variable and larger distances (between 4.5-16 Å) were found between  $Ser_{104}$  and  $Val_{360}$  (Figure 2.2.8).

Furthermore, the distance distributions between the  $Arg_{362}/Phe_{104}$  and  $Arg_{362}/Ser_{104}$   $C\alpha$  pairs were also determined (Figure 2.2.8). For the WT GalNAc-T2, the average distance of  $4.7\pm 1.4$  Å suggests that there are hydrophobic interactions between the aromatic ring of  $Phe_{104}$  and the side-chain of  $Val_{360}$ . In contrast, two populations were found for the  $F_{104}S$  mutant. A minor population, with distances ranging between 3.5-7 Å and a major population, with rather oscillating distances, between 7.5-17 Å (Figure 2.2.8). These results also point out that the flexible loop and the loop  $\alpha 1$ - $\alpha 2$  are much more structured in the WT enzyme than in the  $F_{104}S$  mutant.

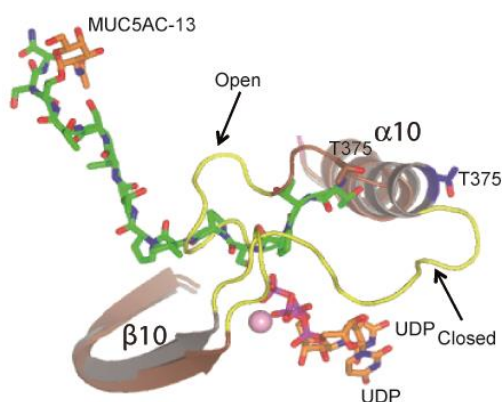


**Figure 2.2.8** - Molecular dynamics simulations on the WT GalNAc-T2 and the F<sub>104</sub>S mutant. The RMSD values calculated by 200 ns molecular dynamics simulations for the flexible loop. Distances between Ser<sub>104</sub> (C $\alpha$ ) and Val<sub>360</sub> (C $\alpha$ )/Arg<sub>362</sub> (C $\alpha$ ) compared to the corresponding distances (Phe<sub>104</sub>-Val<sub>360</sub> and Phe<sub>104</sub>-Arg<sub>362</sub>) found in the WT GalNAc-T2. The center of mass of the  $\pi$ -electron system of Phe<sub>104</sub> was used to calculate the distances.

### 2.2.2.3 <sup>19</sup>F labelling of the WT GalNAc-T2 and F<sub>104</sub>S mutant for NMR experiments.

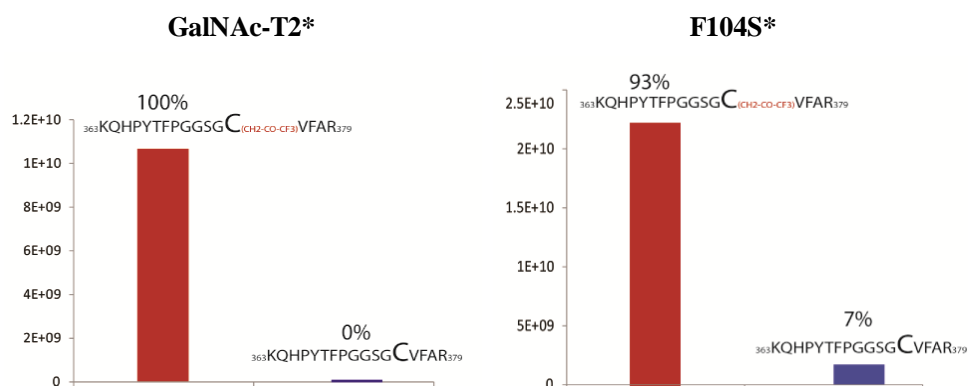
<sup>19</sup>F-NMR spectroscopy experiments were also performed on both enzymes to experimentally monitor the dynamics of the flexible loop. <sup>19</sup>F-NMR chemical shifts are highly sensitive to changes in the local conformational environment and therefore, the measurement of these parameters constitutes a well-established approach for studying protein structure and dynamics.<sup>41</sup> The inclusion of a <sup>19</sup>F label required the mutation of Thr<sub>375</sub> by a more nucleophilic Cys residue, in both WT GalNAc-T2 and the mutant F<sub>104</sub>S variants. In particular, Thr<sub>375</sub> was selected for different reasons:

- i) This residue is rather exposed and therefore amenable to modifications;
- ii) It is very close to the flexible loop (Val<sub>360</sub> to Gly<sub>372</sub>);
- iii) The X-ray crystallographic analyses indicate that Thr<sub>375</sub> is sensitive to the flexible loop dynamics and may adopt different conformations depending on the geometry of the flexible loop (open or closed) (Figure 2.2.9).



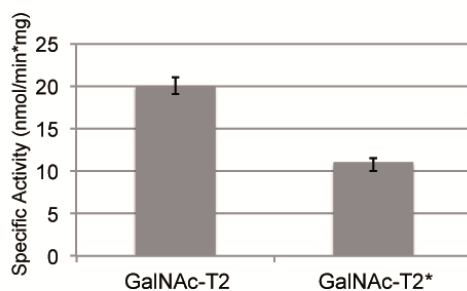
**Figure 2.2.9** – The conformation of Thr<sub>375</sub>. Close-up view of the active site of PDB entries 2FFV and 5AJP (inactive and active states, respectively). The flexible loop adopts an open (2FFV) or closed conformation (5AJP) and is colored in yellow. MUC5AC-13 is depicted with sticks using the green carbon atoms, whereas the GalNAc-moiety is in orange carbon atoms. UDP is colored in orange. The secondary structures are brown and grey for the PDB entries 2FFV and 5AJP, respectively. Thr<sub>375</sub> is depicted as sticks in brown (2FFV) and blue (5AJP) carbon atoms.

Consequently, both the WT GalNAc-T2 and the F<sub>104</sub>S mutant variants, containing the T<sub>375</sub>C mutation were produced.<sup>42</sup> The <sup>19</sup>F label could be efficiently introduced using 3-bromo-1,1,1-trifluoroacetone (BFA),<sup>41</sup> under mild conditions (5 min, phosphate buffer pH 7.0, room temperature) to give a single SCH<sub>2</sub>(CO)CF<sub>3</sub> adduct (GalNAc-T2\* and F<sub>104</sub>S\*). However, while the GalNAc-T2\* was 100% labelled, the F<sub>104</sub>S\* was not fully modified resulting in 93%, as shown by the analysis of the MS/MS data (Figure 2.2.10).



**Figure 2.2.10** - Relative quantification based on the extracted ion chromatogram between peptides with and without the  $^{19}\text{F}$  label at the position  $\text{Cys}_{375}$ .

The activities of the WT GalNAc-T2 and GalNAc-T2\* analogues were compared under the same conditions. Upon introduction of the  $\text{SCH}_2(\text{CO})\text{CF}_3$  tag, the GalNAc-T2\* variant displayed a reduction of nearly 50% in activity in comparison to the WT GalNAc-T2 analogue. Nevertheless, the new enzyme stills remained active was able to glycosylate the peptide substrates (Figure 2.2.11).

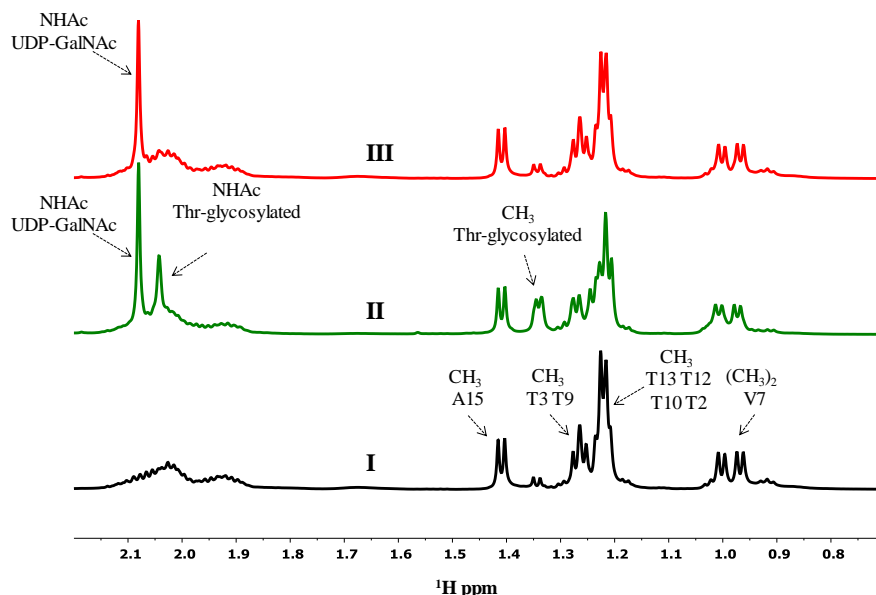


**Figure 2.2.11** - Specific activity of the WT GalNAc-T2 and GalNAc-T2\* using the MUC1 peptide as acceptor substrate. The data represent means  $\pm$ SD. for 3 independent experiments.

#### 2.2.2.4 $^{19}\text{F}$ -NMR experiments

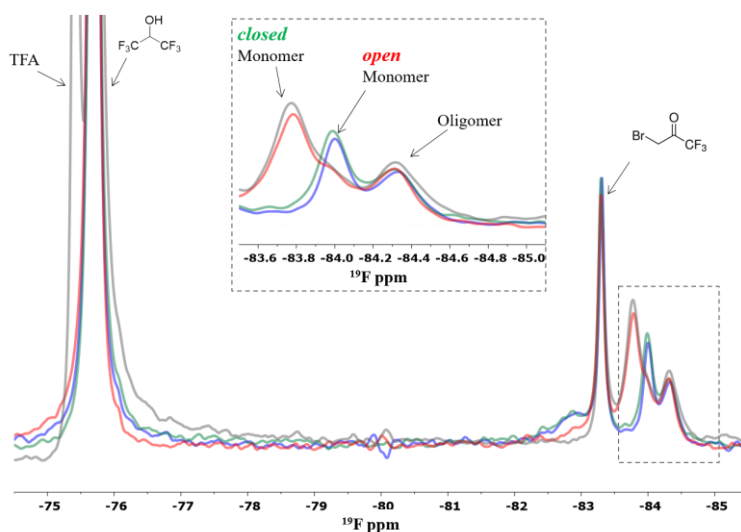
##### 2.2.2.4.1 The modified native GalNAc-T2\* enzyme

First,  $^1\text{H}$ -NMR experiments were performed to monitor the ability of both enzymes, GalNAc-T2\* and F<sub>104</sub>S\* to glycosylate the MUC5AC peptide (Figure 2.2.12). In the presence of UDP-GalNAc and MUC5AC, only GalNAc-T2\*, and not the F<sub>104</sub>S\* mutant, was able to glycosylate MUC5AC. This fact is in agreement with previous kinetic experiments that that the F<sub>104</sub>S mutant was inactive on certain peptide/protein substrates.<sup>32</sup> The appearance of new NMR signals at 2.05 ppm corresponding to the NHAc methyl group of the GalNAc-Thr motif and a dramatic alteration of the chemical shifts of the methyl protons of the Thr amino acids of the peptide ( $\delta = 1.20$ - $1.35$ ppm) clearly indicated the presence of glycosylation (Figure 2.2.12).



**Figure 2.2.12** - Glycosylation experiment of MUC5AC at 298K and 600MHz. **(I)**  $^1\text{H}$ -NMR of MUC5AC (1760  $\mu\text{M}$ ). **(II)**  $^1\text{H}$ -NMR of MUC5AC (1760  $\mu\text{M}$ ) in presence of GalNAc-T2\* (22  $\mu\text{M}$ ),  $\text{MnCl}_2$  (150  $\mu\text{M}$ ) and 5280  $\mu\text{M}$  UDP-GalNAc after 12 h **(III)**  $^1\text{H}$ -NMR of MUC5AC (1760  $\mu\text{M}$ ) in presence of F<sub>104</sub>S\* (22  $\mu\text{M}$ ),  $\text{MnCl}_2$  (150  $\mu\text{M}$ ) and 5280  $\mu\text{M}$  UDP-GalNAc after 12 h.

Then, in order to evaluate whether the  $^{19}\text{F}$ -NMR label was sensitive to the dynamics of the flexible loop upon glycosylation, different  $^{19}\text{F}$ -NMR experiments on GalNAc-T2\* were performed using a variety of conditions (Figure 2.2.13).

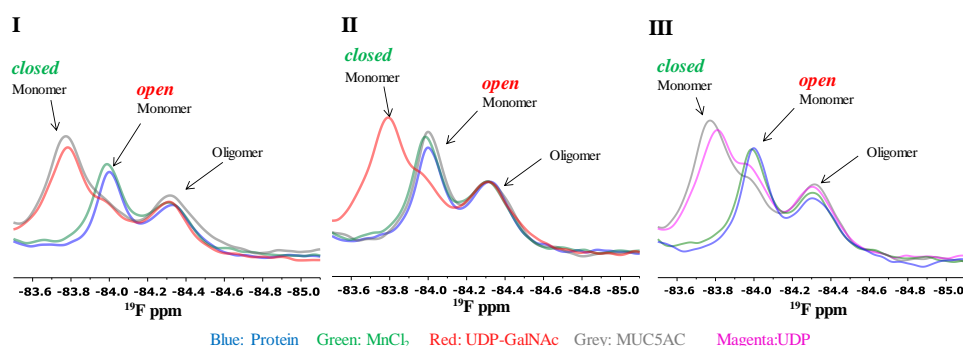


**Figure 2.2.13** -  $^{19}\text{F}$ -NMR Spectra of GalNAc-T2\* at 298K and 600MHz. Blue color corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ). Green color corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ) and  $\text{MnCl}_2$  (225  $\mu\text{M}$ ). Red color corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ),  $\text{MnCl}_2$  (225  $\mu\text{M}$ ) and UDP-GalNAc (225  $\mu\text{M}$ ). Grey color corresponds to the GalNAc-T2\* (135  $\mu\text{M}$ ),  $\text{MnCl}_2$  (225  $\mu\text{M}$ ), UDP-GalNAc (225  $\mu\text{M}$ ) and MUC5AC (150  $\mu\text{M}$ ).

The  $^{19}\text{F}$ -NMR spectra (Figure 2.2.13) also showed signals corresponding to the external reference 1,1,1,3,3,3-hexafluoro-2-propanol ( $\delta = -75.7$  ppm), added for calibration purposes. Additional signals were observed for the 3-bromo-1,1,1-trifluoroacetone reactive used for the protein labelling ( $\delta = -83.3$  ppm) as well as for the trifluoroacetic acid used in the peptide synthesis procedure ( $\delta = -75.4$  ppm). Two specific  $^{19}\text{F}$ -NMR signals at  $\delta = -84.0$  ppm and  $\delta = -84.3$  ppm were also present for GalNAc-T2\*. The  $^{19}\text{F}$  signal at  $\delta = -84.3$  ppm remained unperturbed during all the set of conditions and combinations in terms of addition of substrates. The linewidth of this  $^{19}\text{F}$ -NMR peak (161 Hz) is significantly larger than that at  $\delta = -$

84.0 ppm (111 Hz). Thus, this  $^{19}\text{F}$ -NMR resonance can potentially be attributed to the oligomeric state of GalNAc-T2\*. Which is consistent with a reduced transverse relaxation time of the oligomeric enzyme.<sup>43</sup> In fact, it has been previously demonstrated, by using SAXS experiments, that GalNAc-T2 forms oligomers that are not involved in the catalysis process.<sup>20</sup> This premise may satisfactorily explain the lack of perturbation of this NMR signal to the distinct sample conditions.

In contrast, after addition of UDP-GalNAc, the  $^{19}\text{F}$ -NMR signal at  $\delta = -84.0$  ppm vanished, and a new signal appeared at  $\delta = -83.7$  ppm (Figure 2.2.13 and Figure 2.2.14). It is remarkable that the interconversion between these two  $^{19}\text{F}$ -NMR resonances ( $-84.0$  ppm and  $-83.7$  ppm) only occurs in the presence of UDP-GalNAc. Indeed, in the absence of UDP-GalNAc, the addition of  $\text{Mn}^{+2}$  or the MUC5AC peptide (Figure 2.2.14 panel II (in grey: peptide without UDP-GalNAc)) did not induce any appreciable perturbation in the  $^{19}\text{F}$ -NMR spectrum of GalNAc-T2\*. This result strongly suggests that the  $^{19}\text{F}$ -NMR resonances at  $\delta = -84.0$  ppm and  $\delta = -83.7$  ppm correspond to the open and closed conformations of the flexible loop, respectively. These two populations are in equilibrium and in slow exchange in the NMR chemical shift timescale and therefore, are detected as two signals.



**Figure 2.2.14** -  $^{19}\text{F}$ -NMR Spectra of GalNAc-T2\* showing the open and closed loop conformations as a function of donor and acceptor substrate additions. (I) Blue color corresponds to unliganded GalNAc-T2\*. The green color corresponds to GalNAc-T2\* in the presence of  $\text{MnCl}_2$  (225  $\mu\text{M}$ ). Red color corresponds to GalNAc-T2\*,  $\text{MnCl}_2$  (225  $\mu\text{M}$ ) and UDP-GalNAc (225  $\mu\text{M}$ ). Grey color corresponds to GalNAc-T2\*,  $\text{MnCl}_2$  (225  $\mu\text{M}$ ), UDP-GalNAc (225  $\mu\text{M}$ ) and MUC5Ac (150  $\mu\text{M}$ ). (II) Blue



*color corresponds to unliganded GalNAc-T2\*. Grey color corresponds to GalNAc-T2\* and MUC5AC (150  $\mu$ M). Green color corresponds to GalNAc-T2\*, MUC5AC (150  $\mu$ M) and MnCl<sub>2</sub> (225  $\mu$ M). Red color corresponds to GalNAc-T2\* (135  $\mu$ M), MUC5Ac (150  $\mu$ M), MnCl<sub>2</sub> (225  $\mu$ M) and UDP-GalNAc (225  $\mu$ M). (III) Blue corresponds to GalNAc-T2\* (135  $\mu$ M). Green corresponds to GalNAc-T2\* (135  $\mu$ M), MnCl<sub>2</sub> (225  $\mu$ M). Magenta corresponds to GalNAc-T2\* (135  $\mu$ M), MnCl<sub>2</sub> (225  $\mu$ M) and UDP (225  $\mu$ M). Gray corresponds to GalNAc-T2\* (135  $\mu$ M), MnCl<sub>2</sub> (225  $\mu$ M), UDP (225  $\mu$ M) and MUC5AC (150  $\mu$ M). The Full spectra are present in Figure 2.2.13 and Supporting Information (Figure S1-S2).*

An additional set of experiments in the presence of UDP (Figure 2.2.14 panel III) were also recorded. The addition of UDP yielded a similar result than that described for UDP-GalNAc. However, the effect in the interconversion between the open and closed conformations of the flexible loop was not as pronounced as in the presence of UDP-GalNAc. All together, these evidences show that UDP-GalNAc is absolutely required for the interconversion between the open and closed conformations of the flexible loop and that provides a better stabilization of the closed conformation than UDP (Figure 2.2.14).

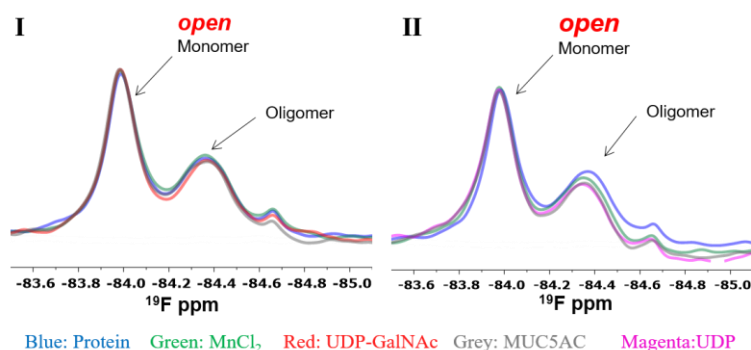
In this proposed mechanism, the unliganded GalNAc-T2 enzyme rests always in an inactive state (open flexible loop) that only moves to the active state (closed flexible loop) in the presence of UDP-GalNAc/MnCl<sub>2</sub>. In contrast to the earlier hypothesis,<sup>29</sup> these evidences precludes the potential existence of an equilibrium between the inactive and active states in the absence of UDP-GalNAc/MnCl<sub>2</sub>. Once the enzyme is in the active state, the peptide can bind and then catalysis will take place.

#### **2.2.2.4.2 The modified F<sub>104</sub>S\* mutant enzyme**

The same set of experiments was performed for F<sub>104</sub>S\* (Figure 2.2.15). In this case, and independently of the conditions used in the experiments, no alteration in the <sup>19</sup>F-NMR spectrum was observed. In particular, the addition of UDP-GalNAc

did not induce any perturbation of the  $^{19}\text{F}$ -NMR resonance signal at  $\delta = -84.0$  ppm. This result also confirms the assignment of the peak at  $\delta = -84.0$  ppm to the open conformation of the flexible loop. Furthermore, the analysis of the  $^{19}\text{F}$ -NMR experiments demonstrated that the flexible loop in  $F_{104}\text{S}^*$  is fixed in open conformation that precludes further binding to the peptides. This geometry corresponds to an inactive enzyme, which is not able to glycosylate peptides.

Overall, the  $^{19}\text{F}$ -NMR experiment provides the molecular basis to explain why the mutant  $F_{104}\text{S}$  does not recognize peptides and, therefore, it is inactive.



**Figure 2.2.15** -  $^{19}\text{F}$ -NMR Spectra of  $F_{104}\text{S}^*$  showing the open conformation even in presence of donor and acceptor substrate. (I) Blue color corresponds to unliganded  $F_{104}\text{S}^*$ . The green color corresponds to  $F_{104}\text{S}^*$  in the presence of  $\text{MnCl}_2$  ( $130 \mu\text{M}$ ). Red color corresponds to  $F_{104}\text{S}^*$ ,  $\text{MnCl}_2$  ( $130 \mu\text{M}$ ) and  $\text{UDP-GalNAc}$  ( $130 \mu\text{M}$ ). Grey color corresponds to  $F_{104}\text{S}^*$ ,  $\text{MnCl}_2$  ( $130 \mu\text{M}$ ),  $\text{UDP-GalNAc}$  ( $130 \mu\text{M}$ ) and  $\text{MUC5Ac}$  ( $100 \mu\text{M}$ ). (II) Blue color corresponds to unliganded  $F_{104}\text{S}^*$ . The green color corresponds to  $F_{104}\text{S}^*$  in the presence of  $\text{MnCl}_2$  ( $130 \mu\text{M}$ ). Magenta color corresponds to  $F_{104}\text{S}^*$ ,  $\text{MnCl}_2$  ( $130 \mu\text{M}$ ) and  $\text{UDP}$  ( $130 \mu\text{M}$ ). Grey color corresponds to  $F_{104}\text{S}^*$ ,  $\text{MnCl}_2$  ( $130 \mu\text{M}$ ),  $\text{UDP}$  ( $130 \mu\text{M}$ ) and  $\text{MUC5Ac}$  ( $100 \mu\text{M}$ ). Full spectra are present in Supporting Information (Figure S3 and S4).

### 2.2.3 Conclusions

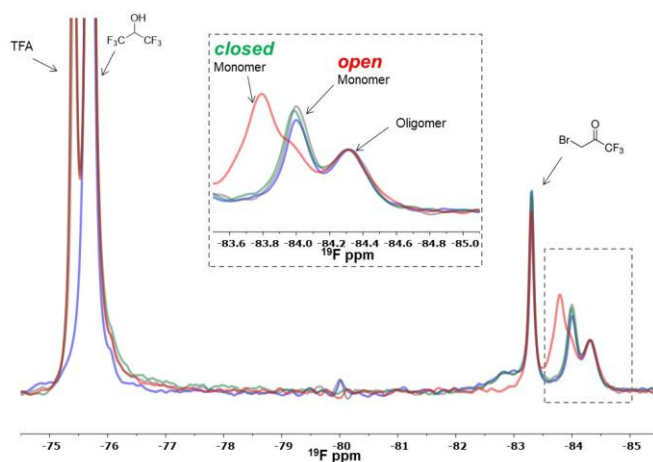
The study presented herein provides another excellent example of the effect of remote mutations in the structure and function of enzymes. Within this work, we

have provided compelling evidences that indicate that the inactive-to-active state transition of the flexible loop in the GalNAc-T2 F<sub>104</sub>S mutant is hindered. The presence of the polar residue Ser104 precludes the establishment of stabilizing interactions with a variety of residues. This causes the instability of both the flexible and the  $\alpha$ 1- $\alpha$ 2 loops, which hampers the flexible loop to adopt the closed conformation required for binding and catalysis.<sup>42</sup> The recognition process follows an induced-fit-mechanism for which UDP-GalNAc is absolutely required.

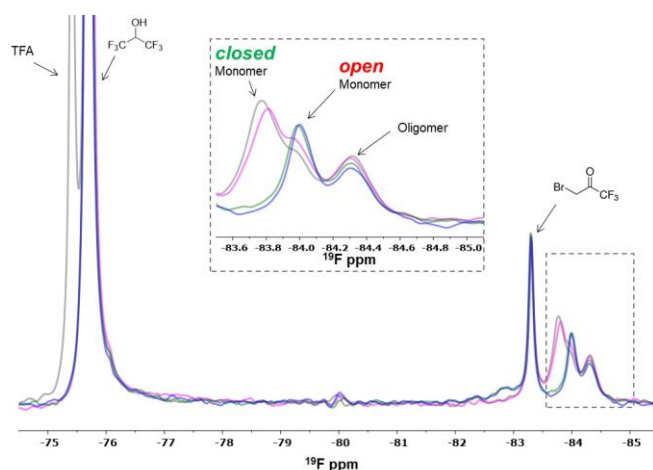
Finally, these results have exploited the knowledge obtained from a mutant associated to a pathology to get more insights into the molecular mechanism of the GalNAc-Ts. This fact also exemplifies the importance of the understanding the molecular basis of mutations associated to disease for the design and implementation of plausible therapeutic approaches.

## 2.2.4 Supporting Information

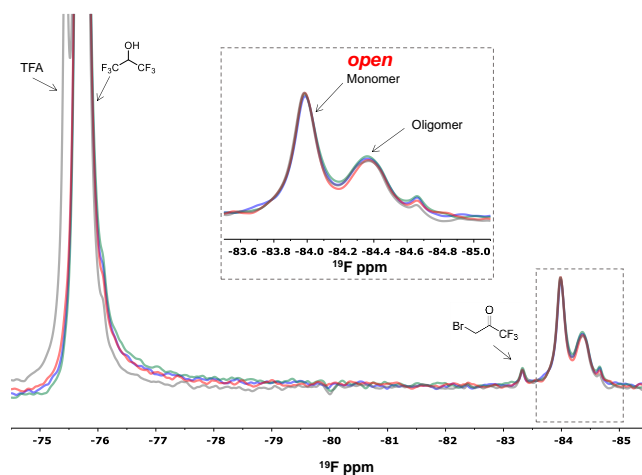
### 2.2.4.1 $^{19}\text{F}$ -NMR spectra



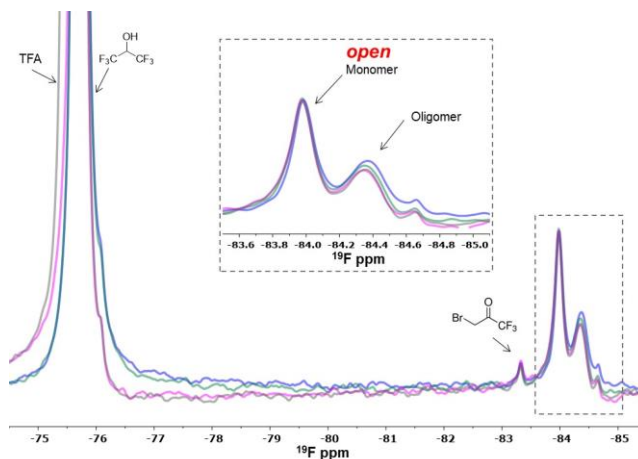
**Figure S1** -  $^{19}\text{F}$ -NMR Spectra of GalNAc-T2\* at 298K and 600MHz. Blue color corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ). Grey color corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ), MUC5AC (150  $\mu\text{M}$ ). Green color corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ), MUC5AC (150  $\mu\text{M}$ ) and  $\text{MnCl}_2$  (225  $\mu\text{M}$ ). Red color corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ), MUC5AC (150  $\mu\text{M}$ ),  $\text{MnCl}_2$  (225  $\mu\text{M}$ ) and UDP-GalNAc (225  $\mu\text{M}$ ).



**Figure S2** -  $^{19}\text{F}$ -NMR Spectra of GalNAc-T2\* at 298K and 600MHz. Blue corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ). Green corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ),  $\text{MnCl}_2$  (225  $\mu\text{M}$ ). Magenta corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ),  $\text{MnCl}_2$  (225  $\mu\text{M}$ ) and UDP (225  $\mu\text{M}$ ). Gray corresponds to GalNAc-T2\* (135  $\mu\text{M}$ ),  $\text{MnCl}_2$  (225  $\mu\text{M}$ ), UDP (225  $\mu\text{M}$ ) and MUC5AC (150  $\mu\text{M}$ ).



**Figure S3** -  $^{19}\text{F}$ -NMR Spectra of  $F_{104}\text{S}^*$  at 298K and 600MHz. Blue corresponds to  $F_{104}\text{S}^*$  ( $86\ \mu\text{M}$ ). Green corresponds to  $F_{104}\text{S}^*$  ( $86\ \mu\text{M}$ ) and  $\text{MnCl}_2$  ( $130\ \mu\text{M}$ ). Red corresponds to  $F_{104}\text{S}^*$  ( $86\ \mu\text{M}$ ),  $\text{MnCl}_2$  ( $130\ \mu\text{M}$ ) and UDP-GalNAc ( $130\ \mu\text{M}$ ). Grey corresponds to  $F_{104}\text{S}^*$  ( $86\ \mu\text{M}$ ),  $\text{MnCl}_2$  ( $130\ \mu\text{M}$ ), UDP-GalNAc ( $130\ \mu\text{M}$ ) and MUC5AC ( $100\ \mu\text{M}$ ).



**Figure S4** -  $^{19}\text{F}$ -NMR Spectra of  $F_{104}\text{S}^*$  at 298K and 600MHz. Blue corresponds to  $F_{104}\text{S}^*$  ( $86\ \mu\text{M}$ ). Green corresponds to  $F_{104}\text{S}^*$  ( $86\ \mu\text{M}$ ) and  $\text{MnCl}_2$  ( $130\ \mu\text{M}$ ) Magenta corresponds to  $F_{104}\text{S}^*$  ( $86\ \mu\text{M}$ ),  $\text{MnCl}_2$  ( $130\ \mu\text{M}$ ) and UDP ( $130\ \mu\text{M}$ ). Grey corresponds to  $F_{104}\text{S}^*$  ( $86\ \mu\text{M}$ ),  $\text{MnCl}_2$  ( $130\ \mu\text{M}$ ), UDP ( $130\ \mu\text{M}$ ) and MUC5AC ( $100\ \mu\text{M}$ ).



## 2.3 *Deciphering the Mechanism of Long and Short Distance-Glycosylation of GalNAc-Ts*

*The work presented in this subchapter has been performed in collaboration with:*

*-Matilde de las Rivas (PhD student) and Dr. Erandi Lira-Navarrete, at the laboratory of Dr. Ramon Hurtado-Guerrero (BIFI, University of Zaragoza) performed the expression and purification of GalNAc-Ts (mutants and chimeras) and the X-Ray crystallography experiments.*

*- Dr. Francisco Corzana, Dr. Gonzalo Jiménez-Osés (Universidad de La Rioja, Logroño, Spain), Dr. Carmen Rovira and Lluís Raich (PhD student) (IQTUB, Universitat de Barcelona, Barcelona, Spain) were responsible for the Molecular Dynamics simulations, while Ismael Compañón (PhD student) at the laboratory of Dr. Francisco Corzana was responsible for the synthesis of the peptides.*

*- Dr. Thomas A. Gerken and Dr. Earnest James Paul Daniel performed the kinetic studies and the Edman-based amino acid sequencing process.*

---

**Publications:** (1) - Matilde de las Rivas, Erandi Lira-Navarrete, Earnest James Paul Daniel, Ismael Compañón, **Helena Coelho**, Ana Diniz, Jesús Jiménez-Barbero, Jesús M. Peregrina, Henrik Clausen, Francisco Corzana, Filipa Marcelo, Gonzalo Jiménez-Osés, Thomas A. Gerken & Ramon Hurtado-Guerrero (2017) *Nature Communications* 8: 1959 (DOI: 10.1038/s41467-017-02006-0).

(2) - Matilde de las Rivas, Earnest James Paul Daniel, **Helena Coelho**, Erandi Lira-Navarrete, Lluís Raich, Ismael Compañón, Ana Diniz, Laura Lagartera, Jesús Jiménez-Barbero, Henrik Clausen, Carme Rovira, Filipa Marcelo, Francisco Corzana, Thomas A. Gerken, and Ramon Hurtado-Guerrero (2018) *ACS Central Science*, 4, 1274–1290 (DOI: 10.1021/acscentsci.8b00488).

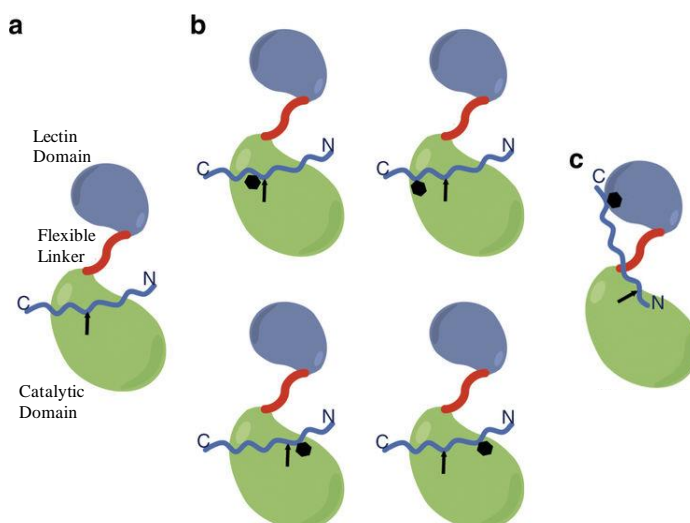




### 2.3.1 Introduction

GalNAc-Ts display distinct kinetic properties as well as different glycosylation capacities on their peptide and glycopeptides substrates. On this basis, GalNAc-Ts can be classified following three distinct glycosylation modes:

- i) Only catalytic domain-dependent glycosylation on unglycosylated peptides and glycopeptides (Figure 2.3.1a)
- ii) short distance-glycosylation<sup>21,37</sup> (Figure 2.3.2b)
- iii) lectin domain-assisted, long distance-glycosylation on glycopeptides<sup>21,37</sup> (Figure 2.3.1c).



**Figure 2.3.1** - Modes of O-glycosylation found for GalNAc-Ts. **a–c.** The panels describe the three distinct modes of O-glycosylation performed by GalNAc-Ts: the neighboring glycosylation activity by the catalytic domain in **a** and **b**, and the long-range lectin domain-mediated glycosylation activity in **c**. Peptides are indicated in blue. The black hexagon-shaped Figure denotes the position of prior GalNAc moieties at the glycopeptides. Arrows indicate the positions of the acceptor sites.

Short distance-glycosylation refers to the glycosylation on GalNAc-containing glycopeptides. In this case, the sites of glycosylation are only 1 to 3 residues away from the prior glycosite.<sup>21,37</sup> Sites adjacent and neighboring to an existing GalNAc glycosite are glycosylated in a catalytic domain-dependent manner. In particular, GalNAc-T4, together GalNAc-T7, T10 and T12 are the only GalNAc-Ts that have been shown to glycosylate contiguous or nearby sites.<sup>21</sup>

Long-range glycosylation activity refers to the glycosylation of GalNAc glycopeptides where the sites of glycosylation are 5-15 residues away from the prior glycosite. In this case, the lectin domain binds a particular GalNAc residue already present at the glycopeptide, and directs the peptide acceptor onto the catalytic domain towards the N- and/or C-terminal directions.<sup>21</sup> For example, GalNAc-T3, T4, T6, and T12 preferentially glycosylate C-terminal glycopeptides sites remote from the prior N-terminal GalNAc glycosites,<sup>21,37</sup> whereas GalNAc-T1, T2, and T14 exhibit the opposite preference.<sup>20,29,38</sup> Other GalNAc-Ts, T5, T13, or T16 do not exhibit preferences for long-range glycosylation.<sup>21</sup>

Structural studies on GalNAc-T2 bound to unglycosylated and glycopeptides have provided insights into the catalytic domain reaction, and have also demonstrated how the GalNAc-T2 lectin domain guides catalysis to N-terminal acceptor sites very distant from the prior C-terminal GalNAc glycosite.<sup>20,29</sup> On the other hand, the molecular basis of how other GalNAc-Ts such as GalNAc-T3, T4, T6, and T12 realize the opposite long range-glycosylation preferences remains unclear.<sup>21,37</sup>

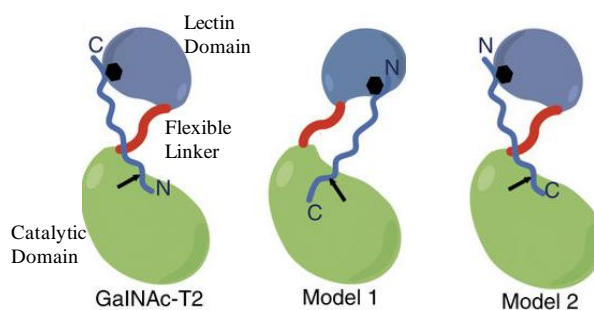
Thus, the understanding of the molecular basis of how long and short range-glycosylation proceed is of paramount importance to decipher the GalNAc-Ts preferences and how these enzymes perform their so-called “filling in” activities, this is how they complete the glycosylation process on heavily glycosylated mucin domains.

### 2.3.1 Results and Discussion

#### 2.3.1.1 *Long distance-glycosylation*

Two possible models have been proposed to explain the different preferences for long distance-glycosylation<sup>37</sup> (Figure 2.3.2). In model 1, rotation of the GalNAc-binding lectin domain is required. It is noteworthy to mention that GalNAc-Ts lectin domains contain three potential carbohydrate-binding sites dubbed  $\alpha$ ,  $\beta$  and  $\gamma$ . However, in most GalNAc-Ts, only one of them is functional.<sup>21,23,27,31</sup> In model 2, GalNAc-binding lectin domain adopts the same orientation as that found in GalNAc-T2. In contrast, the glycopeptide must adopt an inverse orientation on the catalytic domain (Figure 2.3.2).

Herein, we have adopted a multidisciplinary approach using different constructs of GalNAc-T4 and GalNAc-T2 to reveal the molecular basis of the long-distance glycosylation preferences of GalNAc-Ts.



**Figure 2.3.2** - Two possible models have been proposed to explain the different long-distance glycosylation preferences of GTs. Peptides are indicated in blue, while the black hexagon-shaped Figure denotes the position of the prior GalNAc moieties in the glycopeptides. Arrows indicate the positions of the acceptor sites.

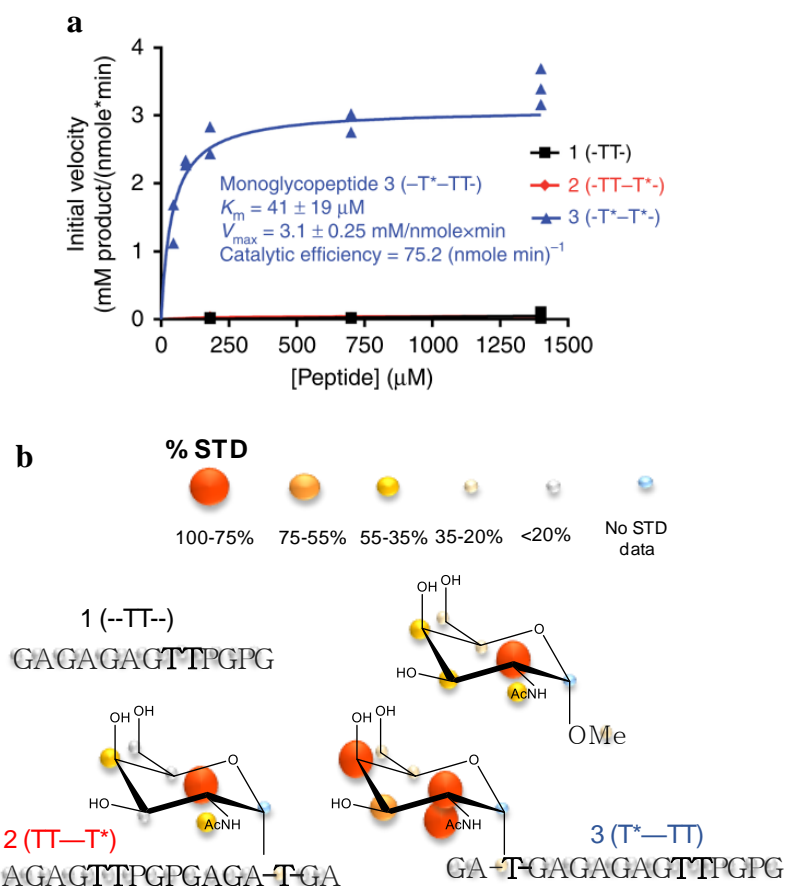
### 2.3.1.1.1 The interaction and kinetics of GalNAc-T4

Three peptides were designed and synthesized to test the long-distance glycosylation preference of GalNAc-T4 (Table 2.3.1). One of these peptides was the naked peptide, **1** (--TT--), whereas the other two were monoglyco-peptides, **2** (TT--T\*); the \* denotes the glycosylated Thr) and **3** (T\*--TT), which contained a GalNAc moiety (T\*), located either at the C- and N-terminus, respectively. All peptides show two potential Thr acceptor sites and also contain the PXP motif, which is selectively recognized by most GalNAc-Ts.<sup>20</sup>

*Table 2.3.1 - Peptide sequences used in this study. \* denotes the location of the existing GalNAc moiety.*

Peptides	Sequence
<b>1</b> (--TT--)	GAGAGAGTTPGPG
<b>2</b> (TT--T*)	AGAGTTPGPGAGAT*GA
<b>3</b> (T*--TT)	GAT*GAGAGAGTTPGPG

Based on the kinetics results (Figure 2.3.3a), GalNAc-T4 selectively glycosylates peptide **3** with high affinity,  $K_m \sim 40 \mu\text{M}$ , and catalytic efficiency  $\sim 75 \text{ nmole min}^{-1}$ . The enzyme did not show any evidence of glycosylation on peptides **1** and **2** (Figure 2.3.3a). These observations are in agreement with previous results that described that GalNAc-T4 only glycosylates C-terminal sites of monoglyco-peptides when the prior glycosite is at the N-terminus, using the lectin domain as template for the process. Thus, binding studies using saturation-transfer difference (STD) NMR experiments were performed (Figure 2.3.3b and Figure SI1).



**Figure 2.3.3** - Biophysical characterization of GalNAc-T4 against peptides **1-3**. **a**, Peptide glycosylation kinetics of GalNAc-T4 on peptides **1-3** (black, red, and blue symbols, respectively). The plot represents the average of triplicate experiments. **b**, STD-NMR-derived epitope mapping. STD spectra are present in Supporting Information (Figure S11).

The results revealed that the unglycosylated peptide was very poorly recognized by GalNAc-T4, in agreement with the kinetic assays. The recognition of glycosylated peptides by GalNAc-T4 showed weak but observable STD signals, indicating that the peptide backbone is poorly recognized by GalNAc-T4, using the methoxy group of  $\alpha$ -methyl-GalNAc as control (Figure 2.3.3b). It is noteworthy to mention that the GalNAc moiety of the mono-glycopeptides **2** and **3** showed unambiguous STD signals (Figure 2.3.3b and Figure S11). Fittingly, **2** and  $\alpha$ -methyl-GalNAc showed identical STD-derived epitope mapping. The highest STD

response corresponded to GalNAc H2, followed by H4, H3 and the N-acetyl group. In contrast, peptide **3** presents a rather different GalNAc STD pattern. In this case, H2, H4, and the N-acetyl group of the GalNAc-moiety display the highest STD response, closely followed by H3. These observations indicate that **3** has a distinct binding mode of interaction with the enzyme, different to that found for **2** and the control.

Therefore, from the kinetic and STD results, it is possible to infer that only glycopeptide **3** is displaying the proper orientation of GalNAc for lectin recognition. The interaction enables the efficient glycosylation at the C-terminal site. Hence, the location of a prior glycosite in the mono-glycopeptide with respect to the potential acceptor sites is essential for the optimal interaction that provides the correct orientation of the peptide within the enzyme, critical for getting an efficient turnover rate.

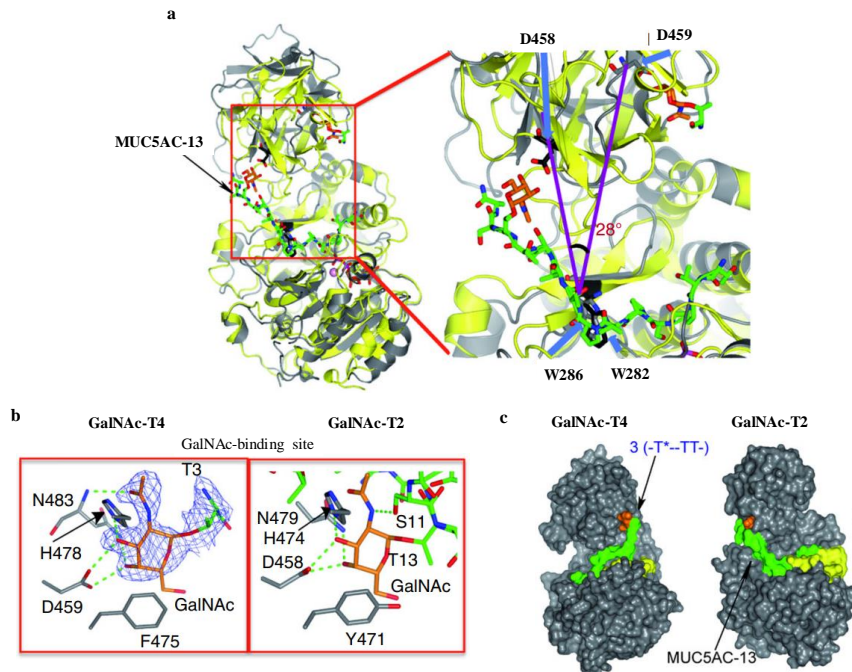
#### **2.3.1.1.2 The crystal structure of GalNAc-T4 with glycopeptide 3**

The molecular basis of GalNAc-T4 glycosylation was further explored by using X-Ray crystallography. Crystals of GalNAc-T4 were first obtained that subsequently were soaked with peptide **3**. The new crystal structure was solved at high resolution (1.90 Å), showing a compact structure with the typical GT-A fold at the N-terminus and the lectin domain at the C-terminal region.

GalNAc-T2 and T4 display a sequence identity of 40 % (Figure SI2). There are large resemblances between the two enzymes at the secondary structural level. In fact, the global superimposition renders a root-mean-square deviation (RMSD) of 2.12 Å, while the RMSD for the catalytic domains is 1.23 Å and 1.82 Å for the lectin domains (aligned C $\alpha$  atoms). Thus, a major shift takes place between the lectin domains (Figure 2.3.4a).

The GalNAc-binding site is located at the right side of the lectin domain, as in model 1. This orientation requires a rotation of  $\sim 28^\circ$  with respect to the

homologous one in GalNAc-T2 (Figure 2.3.4a). This finding explain why GalNAc-T4 achieves a distinct long distance-glycosylation on C-terminal acceptor sites. Moreover, these findings suggest that the location of the GalNAc-binding site is coupled with the long distance-glycosylation preferences of these enzymes.



**Figure 2.3.4** – **a**) (left) Superimposition of the GalNAc-T4-glycopeptide 3 (gray) and GalNAc-T2-MUC5AC-3-UDP-Mn<sup>2+</sup> (yellow) structures. The glycopeptide and the GalNAc moieties are shown in green and orange carbon atoms, respectively. The lectin  $\alpha$ -subdomain GalNAc-binding residues, Asp<sub>458</sub> and Asp<sub>459</sub>, of GalNAc-T2 and GalNAc-T4 are shown as sticks in black and gray carbon atoms, respectively. (right) Close-up view of the superimposition between GalNAc-T2 and GalNAc-T4. **b**) Close-up view of the lectin  $\alpha$ -subdomain GalNAc-binding site for both GalNAc-T4 (left) and T2 (right). The residues of both enzymes are depicted as gray carbon atoms. Hydrogen bond interactions are shown as dotted green lines. Electron density maps are FO-FC (blue) contoured at 2.2  $\sigma$  for Thr3-GalNAc. Note that both GalNAc-binding sites are depicted in the same orientation. **c**) Surface representation of GalNAc-T4 (model built with mono-glycopeptide 3 and UDP/Mn<sup>2+</sup>), and GalNAc-T2 (with UDP/Mn<sup>2+</sup>/MUC5AC-13). Both enzymes are viewed from the same orientation as in **b**. Colors for the glycopeptide and the flexible loop are the same as above.

Then, crystals of GalNAc-T4 were soaked with a high concentration of peptide **3**, UDP, and MnCl<sub>2</sub>. A well-defined density was only observed for the Thr3-GalNAc fragment of peptide **3** bound to the lectin domain (Figure 2.3.4b). There is absence of density for the rest of the peptide backbone of **3**, in agreement with its very weak affinity ( $K_d$  was estimated in the high mM range by SPR, data not shown) and with the extremely weak STD signals observed for the peptide backbone (Figure 2.3.3b).

A more exhaustive analysis of the GalNAc-binding site showed that the GalNAc-moiety is close to four conserved residues: Asp<sub>459</sub>, Asn<sub>483</sub>, Phe<sub>475</sub> and His<sub>478</sub> (equivalent residues for the  $\alpha$ - site in GalNAc-T2 are Asp<sub>458</sub>, Asn<sub>479</sub>, Tyr<sub>471</sub> and His<sub>474</sub>; see Figure 2.3.3b). Most of the interactions are due to hydrogen bonds, while CH- $\pi$  interactions also take place, involving the GalNAc moiety and Phe<sub>475</sub>/Tyr<sub>471</sub> for GalNAc-T4/T2, respectively.

### **2.3.1.1.3 Molecular dynamics simulations**

To understand how the lectin domain guides the approach of the potential acceptor sites to the catalytic domain, molecular dynamics (MD) simulations were performed on GalNAcT4 in the presence of UDP/Mn<sup>+2</sup> and **3**. The calculations showed that, although the peptide bound at the catalytic domain was highly dynamic, showing the existence of two potential acceptor Thr residues in close contact to the UDP during the simulation time (200 ns).

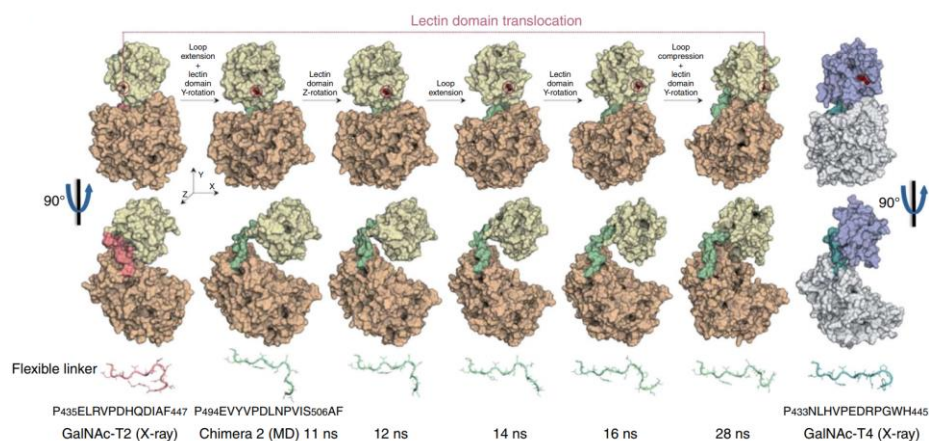
Previous studies have suggested that the flexible linker allows the existence of interdomain translational-like motion in these enzymes.<sup>20</sup> Thus, MD simulations of native GalNAc-T2/T4 in water solution were performed. These enzymes show rather different interdomain linkers. However, the MD simulations clearly showed that the lectin domains of GalNAc-T2 and GalNAc-T4 *apo* enzymes do not show significant rotation during the long 500 ns simulation time, strongly suggesting that the orientation of the GalNAc-binding sites with respect to the catalytic domain is well preserved after the GalNAc-Ts folding.



Therefore, new MD simulations were performed on chimeras 1-2 were now carried out. These constructs corresponded to GalNAc-T2 modified with different lengths of the flexible linker of GalNAc-T3 (Table 2.3.2). For chimera 2, the correct rotation of the lectin domain towards a position similar to that found for GalNAc-T4 was observed (Figure 2.3.5). In fact, the required stepwise motion was completed in only ~30 ns and involved sequential extension of the flexible linker, rotation of the lectin domain around the Z and mainly the Y axes, followed by a final compression of the linker (Figure 2.3.5). Indeed, although the GalNAc-T3 linker was initially forced to adopt the compact conformation of the GalNAc-T2 native linker, it quickly (~10 ns) recovered its native extended conformation. As a consequence, the lectin domain springs and rotates to re-assemble in a structure similar to that of the homologous GalNAc-T4 described above. Likely, this motion is smoothed by the absence of strong inter-domain interactions in the chimera, which do take in place in GalNAc-T4. In this case, a highly persistent salt bridge between Arg<sub>397</sub> and Glu<sub>487</sub> is difficult to overcome. Strikingly, the intrinsic conformational preferences of a rather small, unfolded fragment (only 14 - 16 residues) determine the relative orientation of rather large protein domains.

**Table 2.3.2-** *Sequences of the chimeras and mutant GalNAc-Ts*

<b>Transferase</b>	<b>Sequence of Flexible linker*</b>
<b>GalNAc-T2</b>	P <sub>435</sub> ELRVPDHQDIAF <sub>447</sub>
<b>GalNAc-T2-t3<sup>Flexible linker</sup> (Chimera 1)</b>	P <sub>494</sub> EVYVPDLNPVIS <sub>506</sub>
<b>GalNAc-T2-t3<sup>Flexible linker -AF</sup> (Chimera 2)</b>	P <sub>494</sub> EVYVPDLNPVIS <sub>506</sub> AF
<b>GalNAc-T2-t4<sup>Flexible linker</sup> (Chimera 3)</b>	P <sub>433</sub> NLHVPEDRPGWH <sub>445</sub>
<b>GalNAc-T2-t3<sup>Flexible linker -AF-P503A</sup> (Chimera 2_P503A)</b>	P <sub>494</sub> EVYVPDLNAVIS <sub>506</sub> AF
<b>GalNAc-T2 (double mutant, R<sub>438A</sub>-D<sub>444A</sub>)</b>	P <sub>435</sub> ELAVPDHQAI <sub>447</sub>
<b>GalNAc-T2 (triple mutant, R<sub>438A</sub>-D<sub>444A</sub>-F<sub>447A</sub>)</b>	P <sub>435</sub> ELAVPDHQAI <sub>447</sub>
<b>GalNAc-T3</b>	P <sub>494</sub> EVYVPDLNPVIS <sub>506</sub>
<b>GalNAc-T4</b>	P <sub>433</sub> NLHVPEDRPGWH <sub>445</sub>



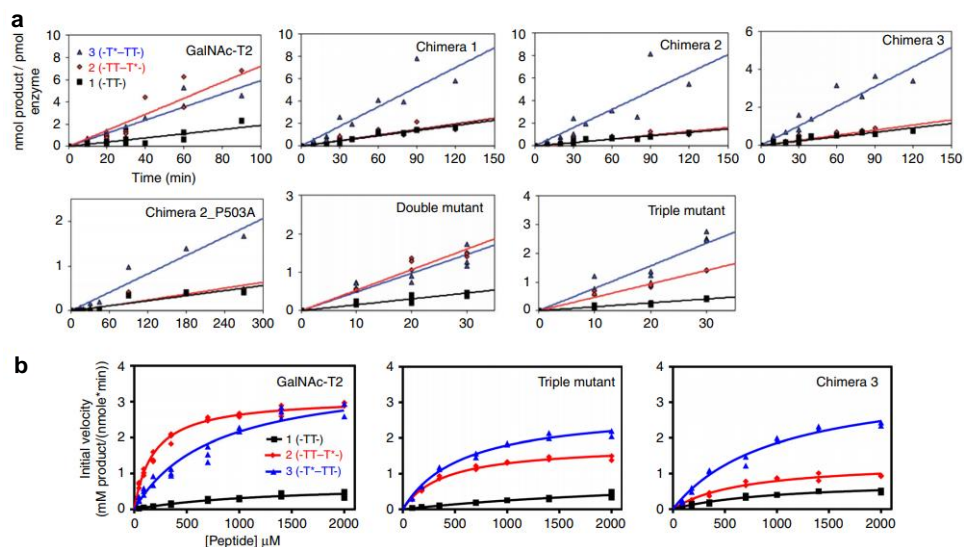
**Figure 2.3.5** - 0.5  $\mu$ s MD simulations of chimera 2. The 28 ns time-lapse snapshots show the dynamic events occurring during the lectin domain re-orientation process. The GalNAc-T2 (far left) and GalNAc-T4 (far right) crystal structures are shown as references for the initial and final states. All structures are illustrated in a surface view with two different orientations. The lectin and catalytic domain, and the flexible linker of GalNAc-T2 and GalNAc-T4 are depicted as yellow/purple blue, orange/light gray, and red/deep teal, respectively. The colors used for the chimera 2 and GalNAc-T2 are the same, except for the flexible linker of the former, which is pale green. Flexible linkers are also shown at the bottom in a cartoon- and sticks-like view.

#### 2.3.1.1.4 Kinetic characterization of the chimeras

Chimeras 1-3 were expressed and purified to validate the conclusions of the MD simulations. Chimeras 1-2 have the flexible linker of GalNAc-T3, while chimera 3 has that of GalNAc-T4 (Table 2.3.2). Although all contain most of GalNAc-T2 architecture, chimeras 1-3 rendered glycosylation preferences similar to GalNAc-T4 (Figure 2.3.6). These results support that glycosylation preferences are determined by the orientation of the functional GalNAc-binding site of the lectin domain with respect to the catalytic domain. In particular, and as suggested by MD simulations, the lectin domain of these chimeras must adopt an equivalent position to that found for GalNAc-T4 to account for their glycosylation preferences.

Notably, GalNAc-T2 is less specific than GalNAc-T4 and is able of glycosylating all peptides. However, GalNAc-T2 preferentially glycosylates mono-

glycopeptide **2**, ~2-fold and ~4-fold better than mono-glycopeptide **3** and the naked peptide **1**, respectively. On the contrary, chimeras 1-3 prefer glycosylating **3**, ~4-fold better than **2** and **1**, resembling GalNAc-T4 (Figure 2.3.6). Furthermore, the specific activity of chimeras 1-3 and GalNAc-T2, T4 towards their preferred naked peptide substrates is fairly similar, indicating that the new linkers do not affect the overall architecture of the active site. The specific Thr residues that were glycosylated were assessed by Edman amino acid sequencing. The Thr residue adjacent to the PxP motif was glycosylated by all chimeras, as expected.<sup>21</sup> Little or no glycosylation was observed at the proximal Thr, except for chimera 2\_P503A (Figure 2.3.6). Together, these results indicate that the glycosylation preferences for these chimeras can be reasonably explained by the rotation of the lectin domain of chimera 2, as predicted by the MD simulations. Hence, the flexible linker causes the rotation of the lectin domain and drives the distinct glycosylation preferences of the GalNAc-Ts.



**Figure 2.3.6 – a)** Time course plots of glycosylation by GalNAc-T2 and the GalNAc-T2 chimeras and mutants of (glyco)peptides **1–3** using a substrate concentration of 1.4 mM. **b)** Complete glycosylation kinetics (initial specific activity versus substrate concentration) for GalNAc-T2, the GalNAc-T2-triple mutant, and the GalNAc-T2 chimera 3 against **1–3**. Kinetic values are summarized in Table S11.

### **2.3.1.1.5 Site-directed mutagenesis of the flexible linker**

Finally, the possible existence of particular residues within the flexible linkers that could be responsible for the lectin domain orientation was explored. A multiple alignment of the flexible linkers (Figure SI3) showed that GalNAc-T3, T4, T6, and T12, which also share the same long-distance glycosylation preferences,<sup>21</sup> contain one additional Pro residue that is not present in GalNAc-T1, T2, and T14, which display the opposite glycosylation preferences. These Pro residues (P<sub>503</sub> and P<sub>442</sub>, respectively) are located at the end of the loop in GalNAc-T3 (P<sub>494</sub>EVYVPDLNP<sub>503</sub>VIS<sub>506</sub>) and GalNAc-T4 (P<sub>433</sub>NLHVPEDRP<sub>442</sub>GWH<sub>445</sub>). There is an Asp residue at the corresponding position in GalNAc-T2 (P<sub>435</sub>ELRVDPDHQDIAF<sub>447</sub>). Since prolines are very abundant in protein turns, this additional Pro might be responsible of the different orientation found in the crystal structure of the GalNAc-T4 lectin domain and in the computational model of chimera 2. Thus, Pro<sub>503</sub> in chimera 2 was mutated to Ala to provide chimera 2\_P<sub>503</sub>A (Table 2.3.2). This chimera showed a very similar glycosylation preference as the starting chimera 2 (Figure 2.3.6), rather than the wild type GalNA-T2. This fact suggests that very complex events within the flexible linker take place that account for the suggested interplay between flexible linker motion and its coupling to the lectin domain rotation.

The interactions between different residues of the flexible linker in the compact and extended crystal structures of GalNAc-T2 were then evaluated. Two main interactions, a salt bridge between Arg<sub>438</sub> and Asp<sub>444</sub>, and a CH- $\pi$  bond between Pro<sub>440</sub> and Phe<sub>447</sub>, were observed. Their importance to maintain the compact and folded structure of the flexible linker (Figure SI4) and to fix the lectin domain orientation with respect to the catalytic domain was explored by generating the corresponding R<sub>438</sub>A-D<sub>444</sub>A double mutant and R<sub>438</sub>A-D<sub>444</sub>A-F<sub>447</sub>A triple mutant.

MD simulations on these structures showed that only in the triple mutant lectin domain rotated (160-450 ns) towards the orientation observed in chimera 2 and GalNAc-T4. These predictions were experimentally by confirmed the

corresponding kinetic analysis. Whereas the double mutant gave nearly equal mono-glycopeptide preferences as that observed for GalNAc-T2, the triple mutant provided an almost 2-fold preference for **3** over **2**, resembling the preference found for GalNAc-T4 (Figure 2.3.6, and Table SI1).

Overall, these results suggest that the entire flexible linker plays a significant role in directing the long-range glycopeptide specificity of this family of GTs, altering the relative orientations of the catalytic and lectin domains.

### **2.3.1.1 Short distance-glycosylation**

On the other hand, sites adjacent to an existing GalNAc glycosite are glycosylated just in a catalytic domain-dependent manner, with GalNAc-T4 being part of a small number of GalNAc-Ts (including GalNAc-T7, T10 and T12).<sup>21</sup> While GalNAc-T4 glycosylates acceptor sites that are located at the C-terminal direction with respect to a prior GalNAc glycosite, GalNAc-T7 and T10 do it at the N-terminal position. GalNAc-T12 places the new GalNAc unit exactly three residues away from the prior glycosite towards the N-terminal direction.<sup>21</sup>

#### **2.3.1.1.1 Kinetics of GalNAc-T4 with glycopeptide**

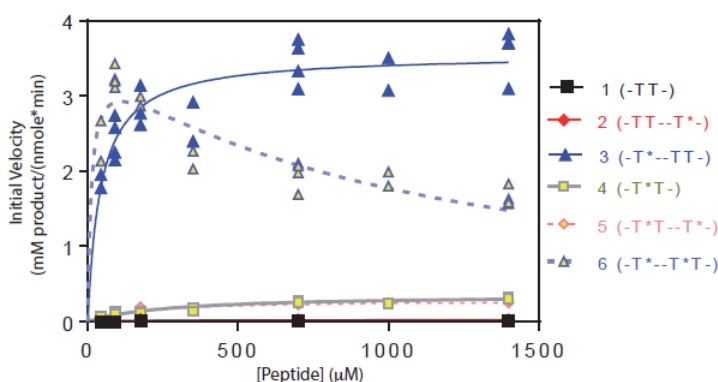
The glycopeptide library was expanded to characterize the short-distance glycosylation. Thus, mono-glycopeptide **4** (-**T\*****T**-) and di-glycopeptides **5** and **6** (denoted -**T\*****T**--**T\***- and -**T\***--**T\*****T**-, respectively) were generated (Table 2.3.3). The di-glycopeptides contain a prior C- or N-terminal glycosylation site that is designed to evaluate the effects of prior glycosylation within the same substrate. In this case, the new substrates display one potential Thr acceptor site (i.e. -**T\*****T**-) with an adjacent C-terminal PxP motif (as -PGP-).<sup>20,21</sup> According to the crystal structure of GalNAc-T4 bound to **3** (-**T\***--**TT**-) (Figure 2.3.4a and Table 2.3.1), the remote

N-terminal GalNAc motif of **3** and **6** (-T\*-T\*T-) should bind to the GalNAc binding site at the lectin domain.

**Table 2.3.3** – Peptide sequences used in this study. \* denotes the location of the GalNAc moiety.

Peptides	Sequence
<b>4</b> (-T*T-)	GAGAGAGT*TPGPG
<b>5</b> (-T*T-T*-)	AGAGT*TPGPGAGAT*GA
<b>6</b> (-T*-T*T-)	GAT*GAGAGAGT*TPGPG

The enzyme kinetics data of GalNAc-T4 are gathered in Figure 2.3.7 and Table SI2. The data for **3**, **4** and **5** could be fit to a standard Michaelis-Menten model, while those for **6** were fit to a model that included apparent substrate inhibition. As shown in Figure 2.3.7, glycopeptides **3** (-T\*-TT-) and **6** (-T\*-T\*T-) show the greatest activity, ~10-13 fold higher than those for **4** (-T\*T-) and **5** (-T\*T-T\*-). This fact is not unexpected since **1** (-TT-) and **2** (-TT-T\*-) were basically unaffected by GalNAc-T4 (Figure 2.3.4a and Table 2.3.1).



**Figure 2.3.7** – Peptide glycosylation kinetics of GalNAc-T4 against (glyco)peptides **1-6**.

For peptide **4** (-T\*T-), glycosylation takes place (Figure 2.3.7). Therefore, the neighboring glycosylated Thr must directly bind to the catalytic domain, rather than to the lectin domain, since the acceptor Thr is contiguous to the glycosylated residue. The existence of similar kinetic constants for di-glycopeptide **5** (-T\*T-T\*-) and glycopeptide **4** (-T\*T-), (Figure 2.3.7 and Table SI2) strongly suggests that the glycosylation of **5** is also promoted by binding at the catalytic domain.

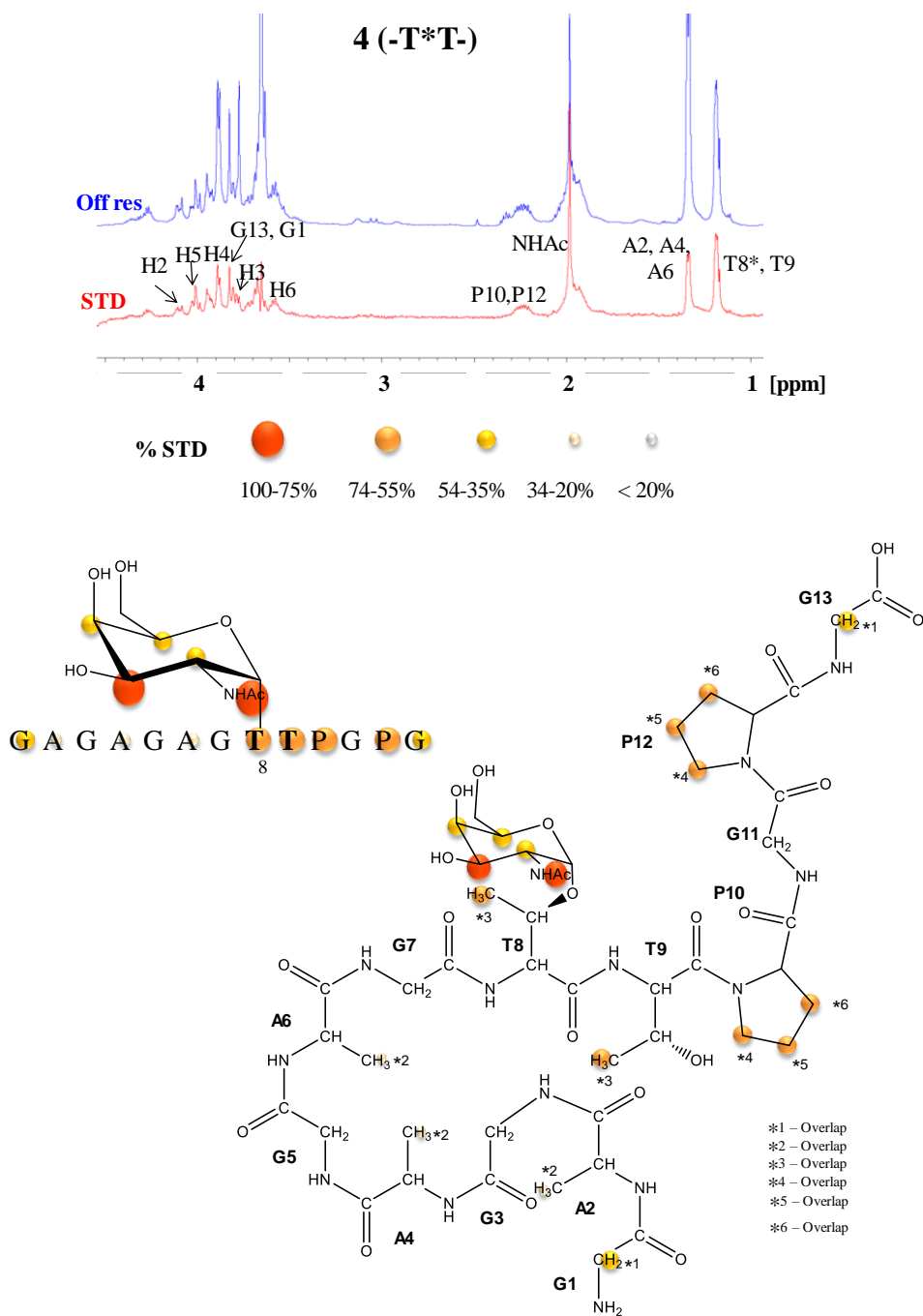
A comparison of the two mono-glycopeptides **3** (-T\*-TT-) and **4** (-T\*T-) reveals the influence of the existence of prior glycosylation at the remote N-terminal site on the enzymatic activity. Peptide **3** displays a ~10-fold higher  $V_{\max}$ , ~5-fold lower  $K_m$ , and, thus, a ~50-fold higher catalytic efficiency ( $V_{\max}/K_m$ ) compared to peptide **4**. This fact is also observed for di-glycopeptide **6** (-T\*-T\*T-), which shows both remote and neighboring glycosites at the acceptor site (Figure 2.3.7). In this case, the kinetic profile measured for **6** fits to a Michaelis Menten model with substrate inhibition (Figure 2.3.7 and Table SI2). The  $K_m$  of **6** (-T\*-T\*T-) is ~3 fold lower than that for peptide **3** (-T\*-TT-) also suggesting a synergistic effect of the two glycosites with possible separate binding events to the lectin and catalytic domains. This synergistic effect is also observed when analyzing the  $K_d$  values obtained by SPR (Figure SI5), in the presence of excess of UDP and  $MnCl_2$ . While  $K_d$  for mono-glycopeptides **4** (-T\*T-) and **3** (-T\*-TT-) are relatively weak (in the mM range), the  $K_d$  of the di-glycopeptide **6** (-T\*-T\*T-) is  $70 \pm 15 \mu M$ . Although there are certain discrepancies the  $K_m$  values from the kinetic studies (Table SI2) and the SPR-based  $K_d$  values, particularly for **3** and **4**, it should be mentioned that the kinetic studies were performed only with UDP-GalNAc. As described above, UDP-GalNAc stabilizes the flexible loop of the catalytic domain active site in the closed conformation, promoting the enzyme activity. For peptide **6**, the  $K_d$  is only ~4.5-fold higher than the  $K_m$ . We can speculate that interaction of the two GalNAc moieties of **6** with the two domains of GalNAc-T4, might further stabilize the essential closed conformation of the flexible loop.

### **2.3.1.1.2 GalNAc-T4 interactions with glycopeptide by STD-NMR**

The binding of the glycopeptides to GalNAc-T4 were also assessed using saturation-transfer difference (STD) NMR experiments. No significant STD enhancements (Figure 2.3.3b) were detected for naked peptide **1**. Differences in the relative STD intensity pattern of the GalNAc protons were deduced for glycopeptides **2** and **3**, probably reflecting their different binding mode at the lectin domain. Indeed, this fact might explain the differences in their glycosylation activities.

In contrast, glycopeptides **4** (-T\*T-), **5** (-T\*T-T\*-) and **6** (-T\*-T\*T-) (Table 2.3.3), displayed significant STD-NMR enhancements for the Thr and Pro residues at the -TTPGP- peptide sequence (Figure 2.3.8 – 2.3.10).





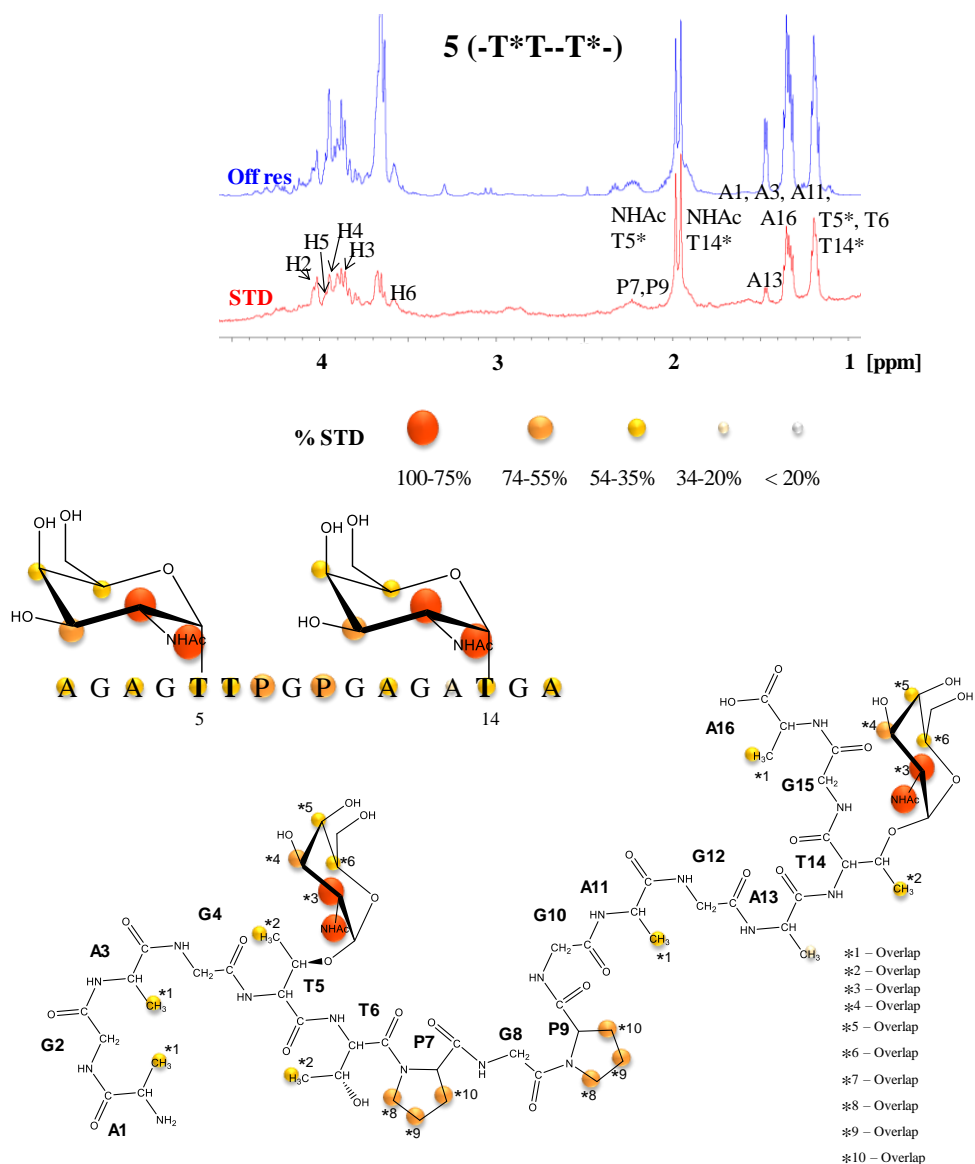
**Figure 2.3.8** – Up: STD-NMR experiments for glycopeptides **4 (-T\*T-)** at 877.5  $\mu$ M in the presence of GalNAc-T4 (13.5  $\mu$ M), UDP (75  $\mu$ M), MnCl<sub>2</sub> (75  $\mu$ M) at 298 K. The key proton resonances are marked in each STD spectrum. Down: STD-NMR-derived epitope mapping. . STD-NMR-derived epitope mapping obtained for **4 (-T\*T-)** with GalNAc-T4. The protons that could not be analyzed in the STD spectrum with accuracy were not mapped. The proton

signals that appeared at the same region of the spectrum are identified in the Figure and display \*.

The GalNAc moieties in **4**, **5** and **6** also showed clear STD-NMR signals. For **4**, **5** and **6**, and independently of the relative GalNAc position in the peptide sequence, the methyl protons of NHAc group provided high STD response (> 85%). Some differences in the relative STDs of the other GalNAc protons could be clearly deduced for **3** (Figure 2.3.3b), **4** (Figure 2.3.8), **5** (Figure 2.3.9) and **6** (Figure 2.3.10). In the case of the mono-glycopeptides **3** (-T\*-TT-) (Figure 2.3.3b) and **4** (-T\*T-) (Figure 2.3.8), the different STD patterns are consistent with the distinct binding contribution of the GalNAc residue at the lectin or catalytic domain for **3** and **4**, respectively. For glycopeptide **3** the highest STD-NMR response (>75%) is observed for the H2, H4 and NHAc while in glycopeptide **4** the H2 and H4 shows low STD effect than H3 and NHAc. These differences in STD-NMR intensities of H2, H3 and H4 protons between both glycopeptides **3** and **4** suggest that the GalNAc moiety is recognized differently which may indicate that they bind at different sites on the enzyme.

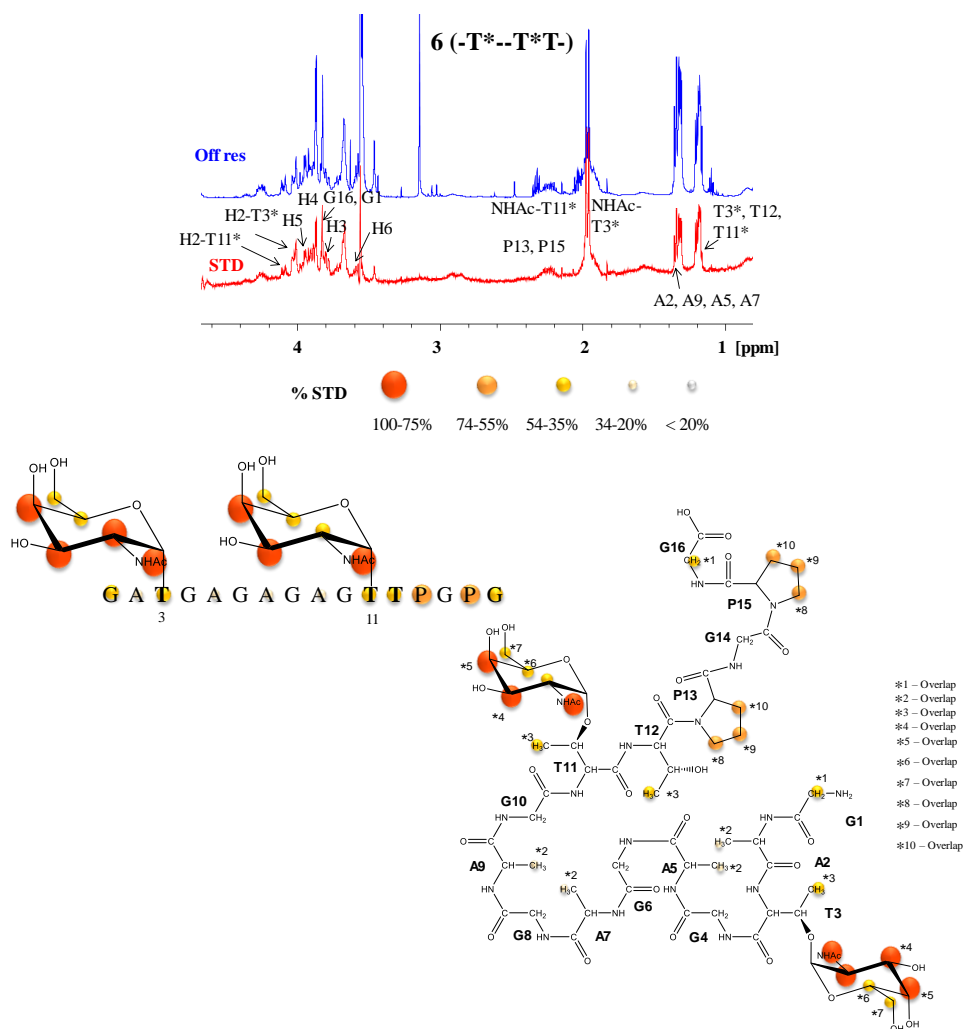
For the di-glycopeptides **5** (Figure 2.3.9) and **6** (Figure 2.3.10), it is only possible to distinguish the protons of the NHAc group for both **5** and **6** and H2 in the case of **6**. Thus, the reported STDs for the other GalNAc ring protons represent their combined enhancements. The NHAc group is strongly recognized for both glycopeptides. However, the STD response of H2 in the two GalNAc units in **6** is rather different. The H2 of GalNAc at Thr 3 (N-terminal domain) shows stronger STD intensity than that at Thr 11 (C-terminal domain), which is similar to that observed for mono-glycopeptide **4**. This fact suggests that the GalNAc moiety preferentially binds to the catalytic domain. This is further supported by the observation of STD (74%-35%) for the peptide residues of the -T\*TPGP- sequence in **4** (Figure 2.3.8), **5** (Figure 2.3.9) and **6** (Figure 2.3.10). Fittingly, such STD in the peptide backbone are absent for **2** and **3**, which lack the prior N-terminal neighboring glycosite (Figure 2.3.3b). All these observations suggest that the both the vicinal GalNAc in -T\*T-, and the -PGP- motif are required for proper peptide

binding in the absence of a remote N-terminal prior glycosite. Moreover, this binding event specifically occurs at the catalytic domain.



**Figure 2.3.9** – Up: STD-NMR experiments for glycopeptides 5 (-T\*T--T\*-) at 877.5  $\mu$ M in the presence of GalNAc-T4 (13.5  $\mu$ M), UDP (75  $\mu$ M), MnCl<sub>2</sub> (75  $\mu$ M) at 298 K. The key proton resonances are marked in each STD spectrum. Down: STD-NMR-derived epitope mapping. STD-NMR-derived epitope mapping obtained for 5 (-T\*T--T\*-) with GalNAc-

T4. The protons that could not be analyzed in the STD spectrum with accuracy were not mapped. The proton signals that appeared at the same region of the spectrum are identified in the Figure and display \*.



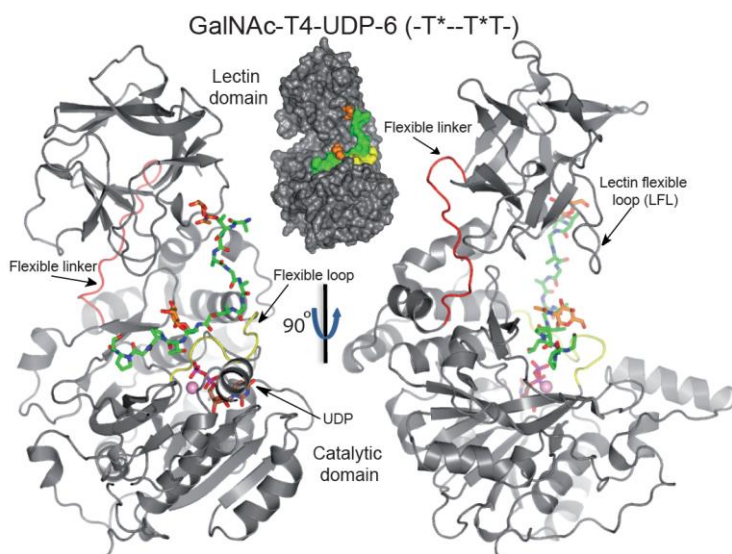
**Figure 2.3.10** – Up: STD-NMR experiments for glycopeptides **6 (-T\*-T\*T-)** at 877.5  $\mu\text{M}$  in the presence of GalNAc-T4 (13.5  $\mu\text{M}$ ), UDP (75  $\mu\text{M}$ ),  $\text{MnCl}_2$  (75  $\mu\text{M}$ ) at 298 K. The key proton resonances are marked in each STD spectrum. Down: STD-NMR-derived epitope mapping. . STD-NMR-derived epitope mapping obtained for **6 (-T\*-T\*T-)** with GalNAc-T4. The protons that could not be analyzed in the STD spectrum with accuracy were not mapped. The proton signals that appeared at the same region of the spectrum are identified in the Figure and display \*.

Strikingly, glycopeptides **3** (-T\*-TT-) and **6** (-T\*-T\*T-), showing the remote N-terminal glycosite and displaying very large catalytic efficiencies (Table SI2), display rather different STDs in their -T\*TPGP- sequences. In particular, **6** shows very high relative STD at the peptide moiety, while **3** essentially does not show any STD. We may hypothesize that this fact is consistent with a mechanism in which the remote N-terminal prior glycosite binds at the lectin domain and productively directs the peptide acceptor site onto the catalytic domain. In the presence of a N-terminal neighboring prior glycosite (-T\*T-) and a PxP motif, the substrate binding becomes further stabilized. This is also in agreement with the observed substrate (or product) inhibition deduced for **6** (-T\*-T\*T-) and with the X-ray structure of **3** (-T\*-TT-) bound to GalNAc-T4 (pdb: 5NQA). Indeed, very little electron density was found for bound the peptide that the catalytic domain of GalNAc-T4 contains an alternative GalNAc-binding site that accounts for the N-terminal short-range glycosylation preference.

#### **2.3.1.1.3 The crystal structure of GalNAc-T4 with di-glycopeptide 6**

Triclinic crystals of GalNAc-T4 were obtained that were subsequently soaked with diglycopeptide **6**, UDP, and MnCl<sub>2</sub>. The resulting crystal was solved by X-ray diffraction, providing a 3D structure of the transferase-diglycopeptide complex at 1.80 Å resolution. The structure shows a compact GT-A fold and a lectin domain located at the N- and C-terminal regions, respectively (Figure 2.3.11).

In contrast with the structure of GalNAc-T4 with peptide **3** (-T\*-TT-), where the flexible loop at the catalytic domain and the ligand were disordered, the di-glycopeptide **6** (-T\*-T\*T-) was now clearly bound to both lectin and catalytic domains. The flexible linker connecting both domains and the flexible loop of the catalytic domain (Figure 2.3.11) were also well defined. An additional loop at the lectin domain pointing towards the GalNAc binding site at the catalytic domain (residues 460-472; hereafter dubbed LFL) was also observed (Figure 2.3.11).

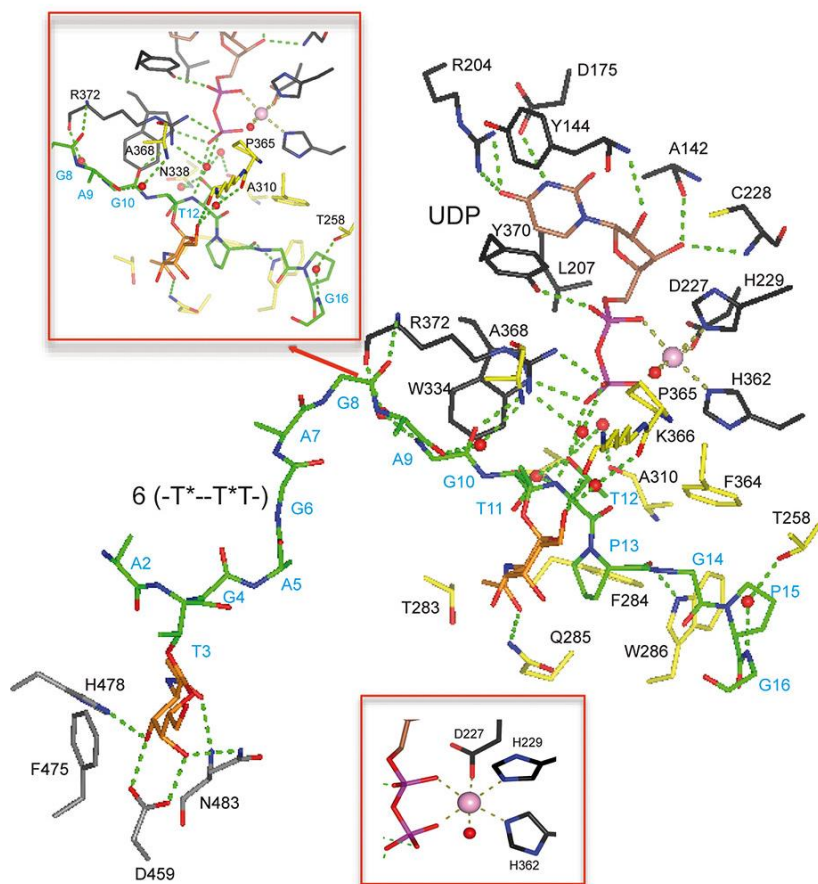


**Figure 2.3.11** - Crystal structure of GalNAc-T4 in complex with UDP-Mn<sup>2+</sup> and the peptide **6** (-T\*--T\*T-). Two different views of GalNAc-T4 in complex with **6**. The catalytic and lectin domains are colored in grey and the flexible linker and loop are depicted in red and yellow, respectively. The lectin flexible loop is indicated by a black arrow. The GalNAc moiety is shown as orange carbon atoms while the rest of the peptide is shown as green carbon atoms. The nucleotide is depicted as brown carbon atoms whereas the manganese atom is shown as a pink sphere. Surface representation of GalNAc-T4 is depicted with the same orientation as the cartoon representation of the GalNAc-T4 structure shown on the left.

The comparison of the GalNAc-T2 structure bound to MUC5AC-13 (GTTSPVPTTSTT\*SAP) and UDP/Mn<sup>2+</sup> (pdb: 5AJP)<sup>20</sup> with GalNAc-T4 bound to **6** (-T\*--T\*T-) and UDP/Mn<sup>2+</sup> was then performed. For GalNAc-T2, the closed conformation of the flexible loop is stabilized by interactions of His<sub>365</sub> and Phe<sub>369</sub> with Trp<sub>331</sub> (*in*-conformation) and by additional interactions with the donor substrate UDP-GalNAc<sup>29</sup>, as described in subchapter 2.2. The GalNAc-T4-glycopeptide **6** complex also shows the “*in*-conformation” (close) of the catalytic Trp<sub>334</sub>, accounting for the presence of an active form.<sup>20</sup>

Several direct and water-mediated interactions were observed, predominantly through hydrogen bonds, with a few hydrophobic interactions.

Direct hydrogen bonds are observed between Gly8/Ala9 and Pro13 of the peptide and the side chain NH's of Arg<sub>372</sub> and Trp<sub>286</sub>, respectively (Figure 2.3.12). A CH- $\pi$  interaction is established between Pro15 and Trp<sub>286</sub>. The Thr12 side chain hydroxyl group is hydrogen bonded to the UDP  $\beta$ -phosphate, properly placing this residue to accept the GalNAc moiety of the UDP-GalNAc. In addition, the Thr12 methyl group is located within a hydrophobic environment formed by the side chains of Thr<sub>283</sub>, Phe<sub>284</sub> and Ala<sub>310</sub> (Figure 2.3.12). This non-polar box helps place the acceptor Thr hydroxyl group in the proper position to accept the anomeric carbon of UDP-GalNAc. This fact explains why Thr residues are better acceptors than Ser moieties. Water mediated hydrogen bonds are observed through the Ala9/Gly10 and Gly16 backbones with Ala<sub>368</sub>/Arg<sub>372</sub> backbones and Thr<sub>258</sub> side chain, respectively, besides those between the Thr12 backbone with Arg<sub>372</sub> and the Thr12 side chain with the Asn<sub>338</sub> and Ala<sub>310</sub> (Figure 2.3.12).



**Figure 2.3.12** - Structural features of peptide, UDP and lectin domain-binding sites. View (central panel) of complete sugar nucleotide, peptide and lectin domain-binding sites of the GalNAc-T4-UDP-peptide **6** complex. Close-up view of peptide (upper panel) and the manganese (lower panel) binding sites. The residues forming sugar-nucleotide, peptide and lectin domain-binding sites are depicted as black, yellow and grey carbon atoms, respectively. UDP and the glycopeptide are shown as brown and green carbon atoms, respectively.  $Mn^{+2}$  and GalNAc moiety are depicted as a pink sphere and orange carbon atoms, respectively. Hydrogen bond interactions are shown as dotted green lines. Water molecules are depicted as red spheres.

This structure also supports the STDs (Figure 2.3.10) observed for the T\*TPGP sequence as well as the absence of relative STDs for the Ala methyl protons. Furthermore, the methyl group of NHAc of the GalNAc unit at Thr11 is pointing towards Thr<sub>283</sub> and Gln<sub>285</sub> of the enzyme, matching the large STD response. The structure also explains the low STD observed for the H2 proton of



the GalNAc unit at Thr11, in comparison to that measured when the GalNAc moiety binds to the lectin domain. Thus, this result also corroborates that the GalNAc residue of **4** binds to the catalytic domain.

In summary, the kinetics and STD-NMR experiments confirm the existence of a GalNAc binding site at the GalNAc-T4 catalytic domain, supporting the initial random glycopeptide studies.<sup>21</sup> Moreover, the analysis of the X-Ray structure of GalNAc-T4 bound to the di-glycopeptide **6** (-T\*-T\*T-) has revealed that the GalNAc moiety at Thr11 is recognized by the catalytic domain (Figure 2.3.12). In particular, there is a hydrogen bond between GalNAc-O6 and Lys<sub>366</sub> and another one between the GalNAc-NHAc group and Gln<sub>285</sub>. In addition, water-mediated (between GalNAc-O6 and Pro<sub>365</sub>) and hydrophobic (between GalNAc-NHAc group and Thr<sub>283</sub>) interactions also take place (Figure 2.3.12). All together, these specific interactions at the catalytic domain provide the impetus to direct the adjacent C-terminal Thr (or Ser) acceptor into the correct orientation for subsequent GalNAc transfer from the UDP-GalNAc donor: both substrate GalNAc moieties make specific and essential contacts with the enzyme, facilitating the presentation required for achieving optimal activity.

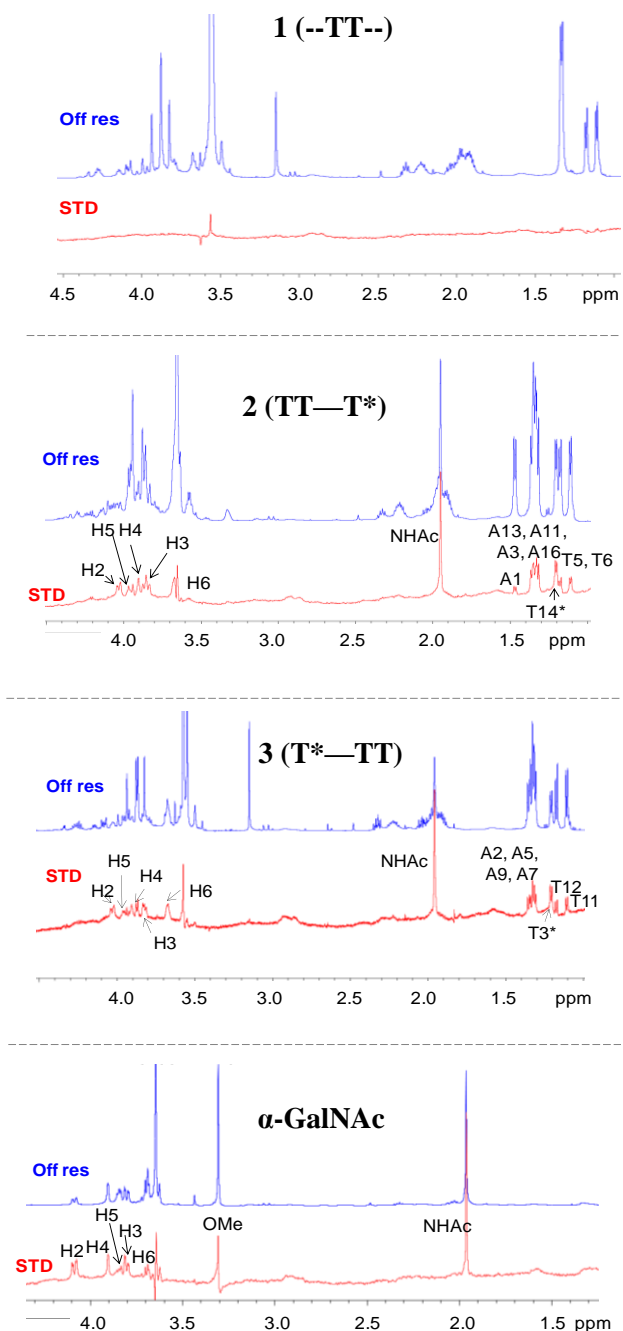
### **2.3.2 Conclusion**

In this work, we identified the molecular basis for GalNAc-T4 long and short range prior GalNAc glycosylation preferences.

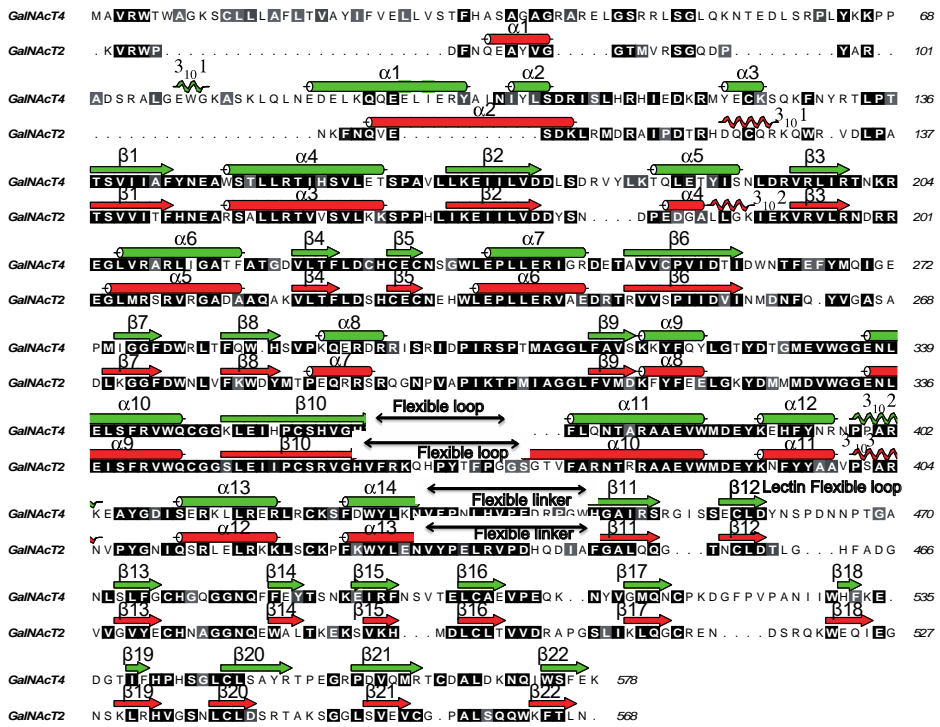
Long-range glycosylation: The distinctive long-range glycosylation preferences of GalNAc-Ts, which is based on a very small flexible linker that provides rotational capacity to the lectin domain.

Short-range glycosylation: The catalytic domain binding site residues are responsible for the binding of the neighboring GalNAc residue have been identified. Interestingly, the binding of GalNAc to the catalytic domain is mediated by few direct interactions to the enzyme suggesting weak interactions.

### 2.3.3 Supporting Information



**Figure S11** - STD-NMR Experiments. For all compounds the reference spectrum (Off res) is displayed in blue color while the STD spectrum is displayed in red. The key proton resonances are marked in each STD spectrum.

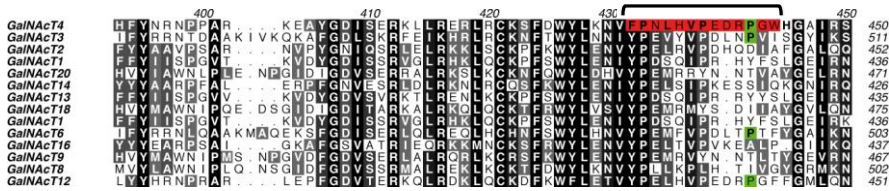


*Figure S12 - Sequence alignment of human GalNAc-T2 and GalNAc-T4. Secondary structure elements from the GalNAc-T2 and T4 structures are shown, with  $\alpha$ -helices,  $3_{10}$ -helices and  $\beta$ -strands in red and green, respectively. The regions encompassing the flexible loop and linker are indicated with double sided arrows.*

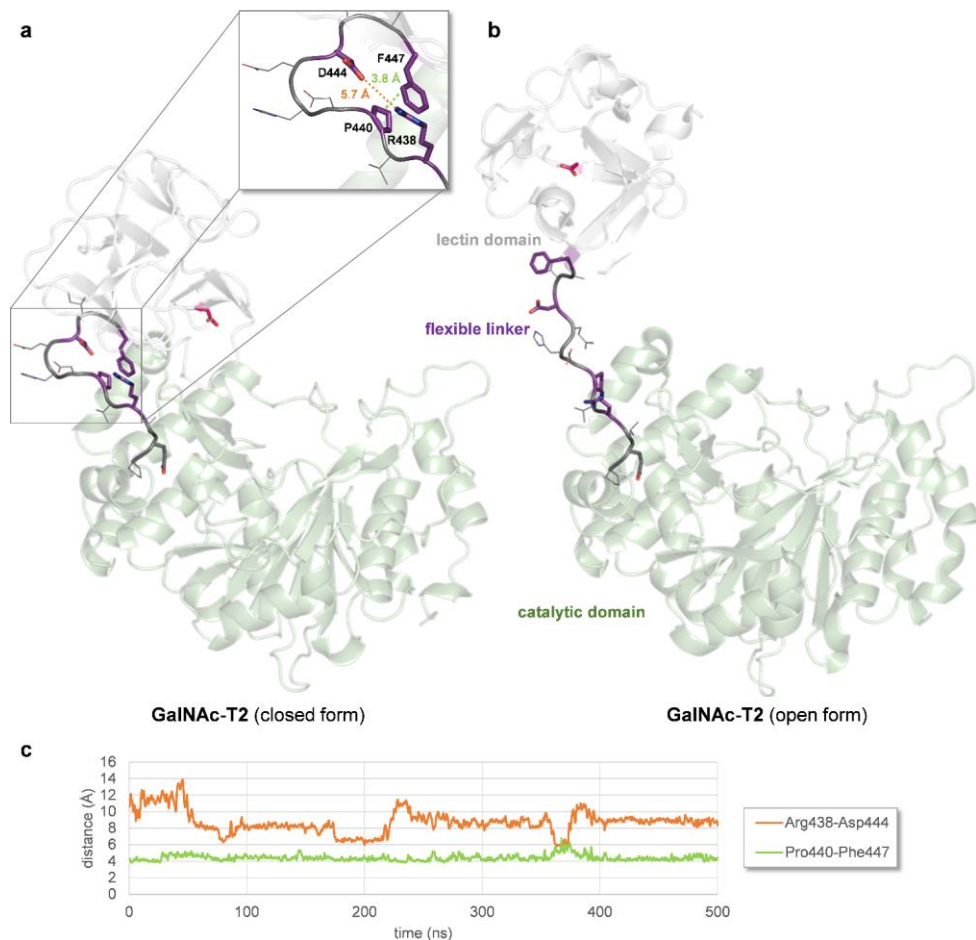
**Table S11** - Michaelis-Menten Kinetic Parameters for GalNAc-T2, GalNAc-T2-Triple Mutant, GalNAc-T2 Chimera 3, and GalNAc-T4 against (glyco)peptide substrates 1, 2 and 3. Parameters were obtained from the nonlinear fit of the data plotted in Fig. 2a and 4c using GraphPad Prism 7.03.

Transferase	Kinetic parameters	(glyco)peptide substrate		
		1(--TT--)	2(-TT—T*-)	3(-T*-TT-)
GalNAc-T2	V <sub>max</sub> <sup>1</sup>	0.7 ± 0.09	3.2 ± 0.06	4.4 ± 0.5
	K <sub>m</sub> <sup>2</sup>	1.2 ± 0.31	0.2 ± 0.01	1 ± 0.2
	Cat eff <sup>3</sup>	<b>0.57 ± 0.08</b>	<b>15.4 ± 0.81</b>	<b>4.2 ± 0.4</b>
Triple Mutant	V <sub>max</sub>	1.1 ± 0.35	1.8 ± 0.06	2.8 ± 0.1
	K <sub>m</sub>	3.3 ± 1.55	0.4 ± 0.04	0.6 ± 0.05
	Cat eff	<b>0.33 ± 0.05</b>	<b>4.5 ± 0.3</b>	<b>4.9 ± 0.3</b>
Chimera 3	V <sub>max</sub>	0.8 ± 0.1	1.35 ± 0.1	3.7 ± 0.3
	K <sub>m</sub>	1 ± 0.2	0.7 ± 0.15	1 ± 0.15
	Cat eff	<b>0.79 ± 0.09</b>	<b>1.9 ± 0.25</b>	<b>3.9 ± 0.3</b>
GalNAc-T4	V <sub>max</sub>	-	-	3.1 ± 0.25
	K <sub>m</sub>	-	-	0.041 ± 0.018
	Cat eff	-	-	<b>75.2 ± 40</b>

Note that the GalNAc-T2 kinetic data was analyzed by GraphPad Prism's 1/y weighting feature that improved the concordance of the fit for peptide 3 at high substrate concentrations, i.e its V<sub>max</sub> value. Parameters for peptides 1 and 2 were unchanged using this weighting compared to the unweighted fit. 2) V<sub>max</sub> units: mM/(nmole\*min), with standard deviation. 3) K<sub>m</sub> units: mM, with standard deviation. 4) Catalytic efficiency (V<sub>max</sub>/K<sub>m</sub>): (nmole\*min)<sup>-1</sup>. Note that the given errors were estimated from the largest percent error of the V<sub>max</sub> or K<sub>m</sub> value. 5) For GalNAc-T4 the activities of peptides 1 and 2 were too low to determine their kinetic constants.



**Figure S13** - Multiple sequence alignment for GalNAc-Ts spanning the regions around the flexible linker. The sequence highlighted in red denotes the flexible linker of GalNAc-T4. The additional Pro in GalNAc-T4/T3/T6/T12 is highlighted in green. The alignment shows the most conserved region between the flexible linkers is towards the N-termini whereas more dissimilarities are around the C-termini.



**Figure S14** - a-b, Structure and dynamics of GalNAc-T2 flexible linker (in purple) connecting the catalytic (in pale green) and lectin (in grey) domains. The X-ray structures of both the closed (PDB entry 5AJP) and open (PDB entry 2FFU) forms of the protein are used as references. c, Key polar and van der Waals interactions responsible for the folded conformation of the flexible peptide monitored throughout MD simulations in explicit water.

## Deciphering GalNAc O-glycosylation

### Deciphering the Mechanism of Long and Short Distance-Glycosylation of GalNAc-Ts

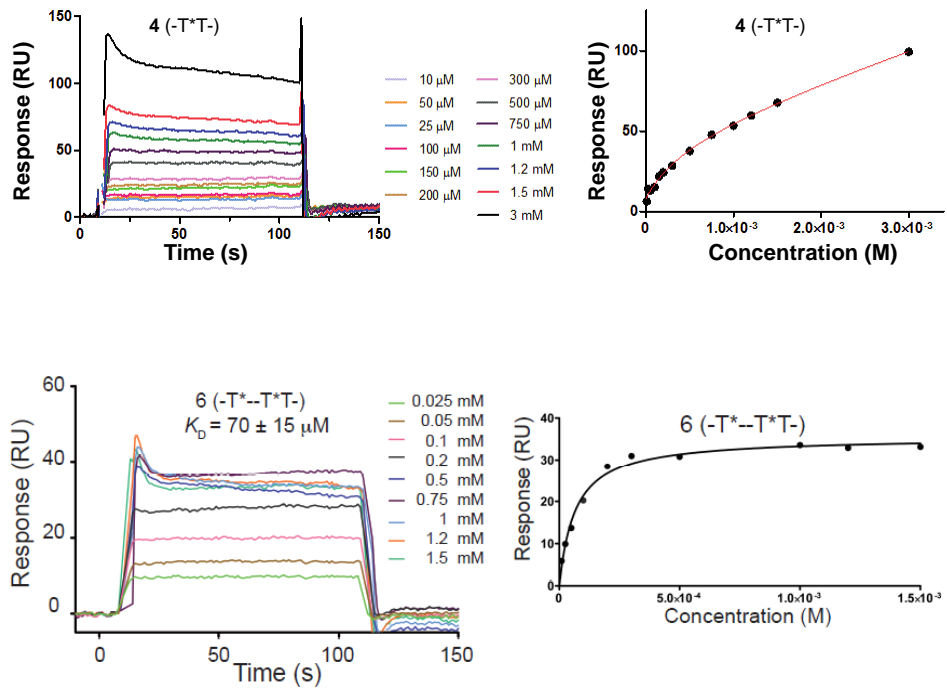
---

CH- $\pi$  interactions (in green) are tightly maintained along the whole simulation; salt-bridge (in orange) fluctuates more but stays in the attractive range most of the simulation time.

**Table SI2** - Kinetic parameters for the wild-type and the mutants.

		<b>3</b> (-T*--T-) <b>Michaelis</b> <b>Menten fit</b>	<b>4</b> (-T*T-) <b>Michaelis</b> <b>Menten fit</b>	<b>5</b> (-T*T--T*) <b>Michaelis</b> <b>Menten fit</b>	<b>6</b> (-T*--T*T-) <b>Substrate</b> <b>Inhibition fit</b>
<b>GalNAc-T4</b>	$V_{\max}$ (mM/nmol*min)	3.5 $\pm$ 0.1	0.35 $\pm$ 0.02	0.29 $\pm$ 0.02	3.6 $\pm$ 0.38
	$K_m$ ( $\mu$ M)	49	271	201	15
	$K_i$ ( $\mu$ M)	-	-	-	890
	<b>Catalytic efficiency</b> <b>(nmol*min)<sup>-1</sup></b>	<b>71.4</b>	<b>1.3</b>	<b>1.4</b>	<b>240</b>
	R <sup>2</sup> curve fit	0.82	0.89	0.90	0.80

Note that the catalytic efficiency is the  $V_{\max}/K_m$  ratio



**Figure S15 - SPR sensogram.** (Sensogram (left) and fitting (right) of SPR data for binding of the peptide **4** to GalNAc-T4. Used ligand concentrations are reported in the inset legend of the sensogram. The end-points of the various injections were plotted against protein concentration. Note that the  $K_D$  could not be determined for peptide **4** because binding saturation could not be achieved. However, the  $K_D$  was accurately determined for peptide **6**.





## **2.4      *O*-glycosylation of mucin MUC1 by GalNAc-Ts**

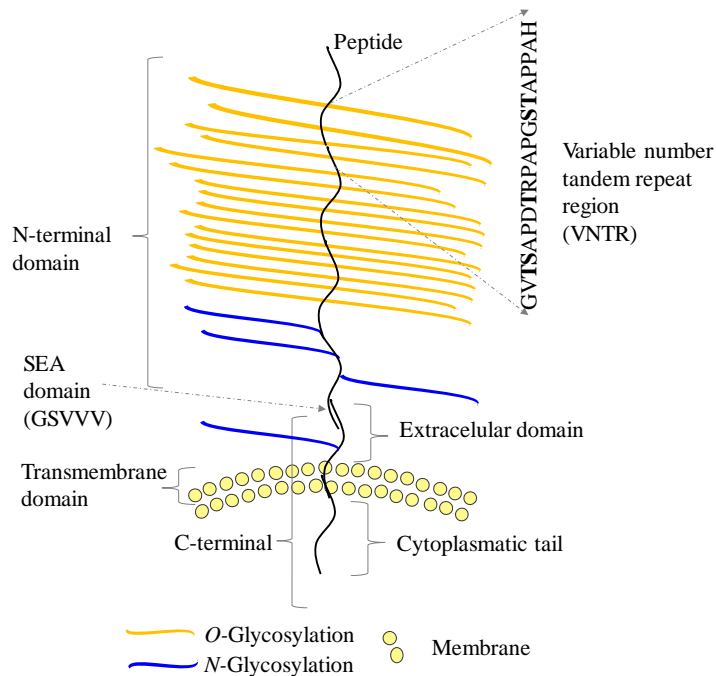
*The work presented in this subchapter has been performed in collaboration with:*

*-Dr. Ramon Hurtado-Guerrero Lab, BIFI, University of Zaragoza, Zaragoza, Spain and Matilde de las Rivas (PhD student), who were responsible for the expression and purification of GalNAc-Ts.*



### 2.4.1 Introduction

Mucin glycoproteins contain multiple sites of glycosylation decorated with complex O-GalNAc glycans with distinct core structures and are involved in fundamental biological processes in health and disease.<sup>44-46</sup> Therefore, mucin glycoproteins are one of the main protein substrates of GalNAc-Ts. In particular, MUC1 is a heavily O-glycosylated transmembrane glycoprotein of 500-1000 kDa, expressed in the apical surfaces of ductal and glandular epithelial cells.<sup>47</sup> The full-length MUC1 is composed of the VNTR and the SEA domains, both located at the N-terminal region, and the transmembrane domain, the cytoplasmic tail and some residues of the extracellular domain, which are in the C-terminal region (Figure 2.4.1).



**Figure 2.4.1** - Schematic representation of the structure of MUC1. The N-terminal is constituted by the VNTR domain and by the SEA domain (sea urchin sperm protein Enterokinase domain), while the C-terminal is composed by the transmembrane domain, cytoplasmic domain and some residues of the extracellular domain.

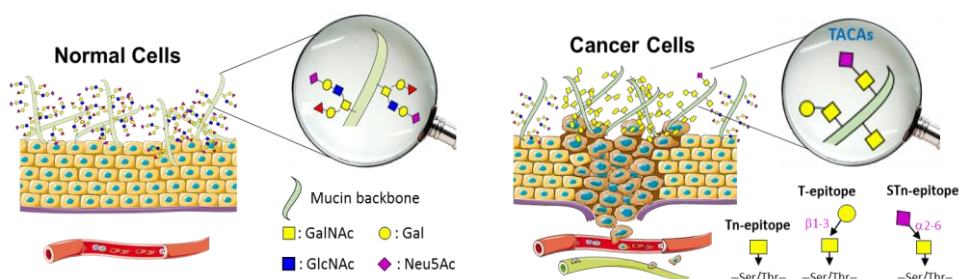
The C-terminal domain (*N*-glycosylated (23-25 kDa)) is composed by an extracellular domain constituted by 58 residues, a single hydrophobic spanning transmembrane domain (28 amino acids) and a cytoplasmatic tail of 72 amino acids with various tyrosine, serine and threonine phosphorylation sites that can bind to various signaling motifs (kinases and growth factor receptors).<sup>48</sup> The later interactions affect and regulate several cancer processes, like the proliferation, the apoptosis and the transcription of various genes.<sup>49–51</sup> MUC1 has been shown to be a docking protein for signaling molecules such as  $\beta$ -canetin (a regulator of transcription), FGFR3 (Fibroblast growth factor receptor 3) and other proteins involved in cancer regulation via the phosphorylation of its cytoplasmic tail.<sup>49,52</sup>

The N-terminal domain of MUC1 contains a variable number of 20-120 tandem repeats (VNTR). Each tandem repeat (TR) encodes a polymorphic sequence of 20 amino acids, with five potential sites for *O*-glycosylation in Ser and Thr residues (Figure 2.4.2).



**Figure 2.4.2** – Sequence of one tandem repeat of MUC1. The larger letters in bold are the five sites for *O*-glycosylation.

It has been proposed that the aberrant glycosylation of the N-terminal domain of MUC1 glycoproteins is a universal feature of cancer (Figure 2.4.3).<sup>17,45,53–56</sup>



**Figure 2.4.3** – MUC1 expressed at the cell surface. Left: MUC1 in normal cells, Right:

*MUC1 in cancer cells. Complex glycans are expressed in normal cells. Short glycan epitopes are exclusively overexpressed in cancer cells.*

In this context, the VNTR domains of MUC1 protein core are one of the major sites for GalNAc-Ts O-glycosylation and alterations in expression, location and function of GalNAc-Ts also occur during malignant transformation.<sup>57</sup> Therefore, it is vitally needed to understand GalNAc-Ts binding specificities and their mechanism of action in mucin templates, like MUC1, to further develop potential glycan-based therapeutic approaches for cancer.

Specific transfer of the GalNAc unit to the five potential positions of the MUC1 sequence (Figure 2.4.2) was studied *in vitro* by Edman degradation and mass spectrometric analysis.<sup>27,34,58</sup> GalNAc-T1 and GalNAc-T3 prefer to glycosylate Thr3 in the VTSA region but they are also capable to glycosylate the Ser14 and Thr15 residues of the GSTA region, (Figure 2.4.2).<sup>34,59</sup> On the other hand, GalNAc-T2 transfers sugar faster to Thr15 at the GSTA region, while Thr3 in the VTSA region and Ser14 in the GSTA region are glycosylated with lower efficiency.<sup>34</sup> The remaining sites, Ser4 at the VTSA region and Thr8 at the PDTR region are exclusively catalyzed by other enzyme, GalNAc-T4.<sup>27,33,60,61</sup> Furthermore, GalNAc-T4 appears to exhibit a strict dependence on the prior glycosylation of the MUC1 substrate.<sup>27,33,62</sup> Hence, the concerted action of GalNAc-T1, T2, T3 and T4 enzymes yields the fully glycosylation of MUC1 TR domains. Nevertheless, the complete glycosylation state may or may not be achieved in a disease context and can be dependent of various parameters, such as the repertoire and cellular activity of the GalNAc-Ts, the location of these enzymes in the cell compartments (Golgi and/or endoplasmic reticulum) and the accessibility of the donor substrate.<sup>19</sup>

Our current knowledge about GalNAc-Ts family and their glycosylation mechanism arise in great extension from kinetic studies in tandem with mass spectrometry analysis.<sup>21,23,26,38,63</sup> In addition, from a structural perspective new molecular details to explain the mechanism of action of GalNAc-T2/T4 by

employing X-Ray and NMR techniques have recently been reported<sup>42,64,65</sup> (described in the subchapters 2.2 and 2.3). However, most of these studies have been performed on short peptide acceptor substrates ignoring the real complexity of the multiple domains of mucins. In fact, three main aspects of this complex process are poorly understood:

1) how distinct GalNAc-Ts cooperate in an orchestrated manner the catalysis of a substrate with multiple sites of glycosylation such as the ones present in mucins;

2) how the interplay between the lectin and the catalytic domain dictates the efficiency and the preference order of glycosylation of GalNAc-Ts during the glycosylation process of mucins;

3) whether the TR repeats are independently *O*-glycosylated of each other or whether the order of the *O*-glycosylation of the acceptor sites takes place in a step-wise manner in the context of multiple domains of mucins.

To address these questions, NMR methods have been herein employed to follow the *O*-glycosylation process of the MUC1 acceptor substrate. Thus, a MUC1 construct encoding multiple TR domains was designed (MUC1-4TR) and heteronuclear NMR experiments were carried out. By simply monitoring the chemical shift perturbations of the amide signals in <sup>1</sup>H/<sup>15</sup>N-HSQC spectra, the mechanism of action of different GalNAc-Ts (GalNAc-T2/T3/T4) have been dissected within the context of a mucin multivalent analogue. In addition, special emphasis has been focused on the understanding of the role of the lectin domain to guide GalNAc-Ts catalysis. For that purpose, the lectin binding ability of GalNAc-T2/T3/T4 has been disrupted, mutating a critical residue at the lectin domain binding site.

Furthermore, we study how conformation of MUC1 changes during glycosylation and how these modifications influence the glycosylation process by GalNAc-Ts, namely GalNAc-T4.

## 2.4.1 Results and Discussion

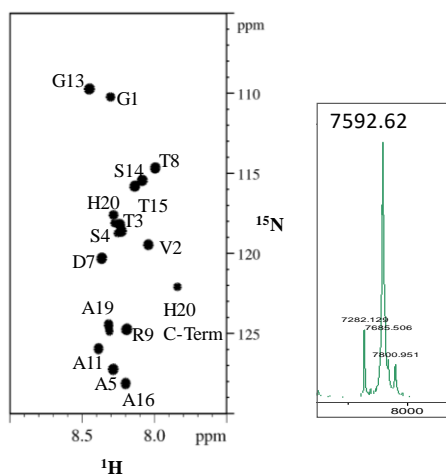
### 2.4.1.1 *Design, Expression, Purification and NMR characterization of isotopic labeled MUC1 with four TR domains*

A DNA construct encoding a KSI protein, a histidine tag, a Tobacco Etch Virus (TEV) cleavage site and four TR domains of MUC1 (MUC1-4TR) was synthesized and sub-cloned into the bacterial expression vector pHTP-KSI by NZYTech (pHTP-KSI-MUC1-4TR). The sequence of the KSI-MUC1-4TR construct is presented in Figure 2.4.4.

**MGHTPEHITAVVQRFVAALNAGDLGDGIVALFADDATVEDPVGSEPRSGTAAI  
REFYANSLKLPLAVELTQEVRAVANEAAFAFTVSFEYQGRKTVVAPIDHFRF  
NGAGKVVSIRALFGEKNIHACQAMGSSHHHHHSSGPQQLRENLYFQGV  
SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTA  
PPAHGVTSAPDTRPAPGSTAPPAH**

*Figure 2.4.4 - KSI-MUC1-4TR sequence. This sequence consists of KSI-tag (blue), a His-tag (green), a TEV (Tobacco Etch Virus) site (red) and the amino acid sequence of MUC1-4TR (underlined).*

The expression vector encodes for a hexahistidine tag and a TEV recognition site to facilitate the MUC1-4TR purification. The expression and purification of MUC1-4TR is described in detail on Material and Methods. The success of the expression and purification was confirmed by mass spectrometry (MALDI-TOF) with the presence of the peaks with 7592.62 m/z for <sup>15</sup>N-MUC1-4TR (Figure 2.4.5). The yield of purified MUC1-4TR was approximately 9 mg L<sup>-1</sup> of culture. The yield of expression/purification was estimated by <sup>1</sup>H-NMR analysis and using 2,2,3,3-tetradeutero-3-trimethylsilylpropionic acid (TSP) as a chemical shift reference ( $\delta$  TSP = 0 ppm). The <sup>1</sup>H/<sup>15</sup>N-HSQC-NMR resonances of MUC1-4TR at pH 6.3 and 283K were completely assigned through standard 2D-TOCSY, 3D-TOCSY <sup>15</sup>N-edited, 2D-NOESY and 3D-NOESY <sup>15</sup>N-edited (Figure 2.4.5 and Table S4).



**Figure 2.4.5** – MUC1-4TR template. Left:  $^1\text{H}/^{15}\text{N}$ -HSQC of the MUC1-4TR with corresponding NMR assignment. Right: MALDI-TOF spectrum of the MUC1-4TR.

The assignment of MUC1-4TR is in agreement with previously reported data for a similar MUC1-5TR construct<sup>59</sup> and confirms that MUC1-4TR construct was successfully expressed and isotopically labeled with high degree of purity.

#### **2.4.1.2 Monitoring MUC1-4TR glycosylation by GalNAc-Ts using NMR spectroscopy. The role of the lectin domain.**

The glycosylation of the multivalent MUC1 with multiple TR domains was investigated by employing GalNAc-T2, T3 and T4 isoforms. These enzymes were selected based on their well-established preferences towards MUC1 peptide substrates.<sup>27,33,34</sup>

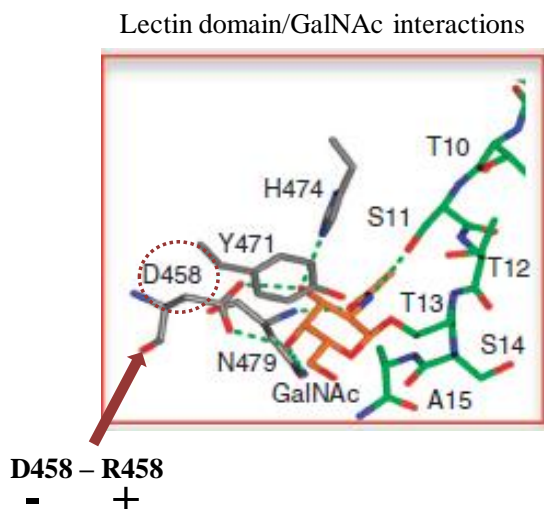
Before monitoring the *O*-glycosylation process of MUC1-4TR by GalNAc-Ts, several control experiments were initially carried out a)  $^1\text{H}/^{15}\text{N}$ -HSQC of MUC1-4TR in presence of UDP-GalNAc and  $\text{MnCl}_2$  and in absence of GalNAc-Ts; b)  $^1\text{H}/^{15}\text{N}$ -HSQC of MUC1-4TR in presence of GalNAc-Ts and  $\text{MnCl}_2$  and in absence of UDP-GalNAc. For both combinations, the spectra remained



unperturbed. Indeed, the presence of distinct components on the sample does not induce any chemical shift or line width perturbation on the  $^1\text{H}/^{15}\text{N}$ -HSQC of MUC1-4TR.

The GalNAc transfer to mucins induces a perturbation on the structure of the amino acid directly glycosylated, in this case the Ser and Thr amino acids, as well as in the neighboring amino acids of the glycosylation site. Therefore,  $^1\text{H}/^{15}\text{N}$ -HSQC experiments on isotopically labeled MUC1-4TR can simultaneously monitor the catalysis of GalNAc-Ts within the context of a MUC1 multivalent analogue with five *O*-glycosylation sites and multiple tandem repeat domains.

To investigate the role of the lectin domain during MUC1-4TR glycosylation process, critical mutations in the  $\alpha$ -repeat of the lectin domain of GalNAc-T2, T3 and T4 were used.<sup>3,23,26</sup> For all the isoforms, the mutated residue was the conserved Asp in GalNAc-T2/T3 and T4 (D<sub>458</sub> in GalNAc-T2, D<sub>517</sub> in GalNAc-T3 and D<sub>459</sub> in GalNAc-T4). This residue is critical for GalNAc binding, since it establishes H-bonds with OH-3 and OH-4 of GalNAc (Figure 2.4.6). Figure 2.4.6 has been adapted from the X-Ray structure of the complex between GalNAc-T2 and MUC5AC-13 and displays the key interactions between the Asp residue and GalNAc.<sup>20</sup>



**Figure 2.4.6** - View of lectin domain-binding site of the GalNAc-T2 bound to the glycopeptide MUC5AC-13. The residues of the lectin domain-binding sites are depicted as grey carbon atoms, the peptide is displayed in green and the GalNAc moiety is depicted with orange carbon atoms. Hydrogen bond interactions are shown as dotted green lines. Adapted from <sup>20</sup>.

The substitution of Asp (negative charge) to Arg or His amino acids (positive charge) in the lectin domain of GalNAc-T2/T3 and T4 strongly affects the recognition of GalNAc by the lectin. Indeed, previous studies have demonstrated that the GalNAc-T2 D<sub>458</sub>R, GalNAc-T3 D<sub>517</sub>H and GalNAc-T4 D<sub>459</sub>H mutants lost the capacity for GalNAc binding.<sup>3,23,26</sup> Hence, by using these mutants, the role of the lectin domain in the glycosylation process of a multivalent MUC1 substrate was investigated.

#### 2.4.1.2.1 GalNAc-T2 WT vs GalNAc-T2 D<sub>458</sub>R

GalNAc-T2 WT catalyzes the glycosylation of three out of the five glycosylation sites in the MUC1 sequence, namely Thr3, Ser14 and Thr15 (Figure 2.4.7 panel I).<sup>63</sup> To ascertain the role of the lectin domain during glycosylation of MUC1-4TR, distinct <sup>1</sup>H/<sup>15</sup>N-HSQC of MUC1-4TR were recorded in the presence

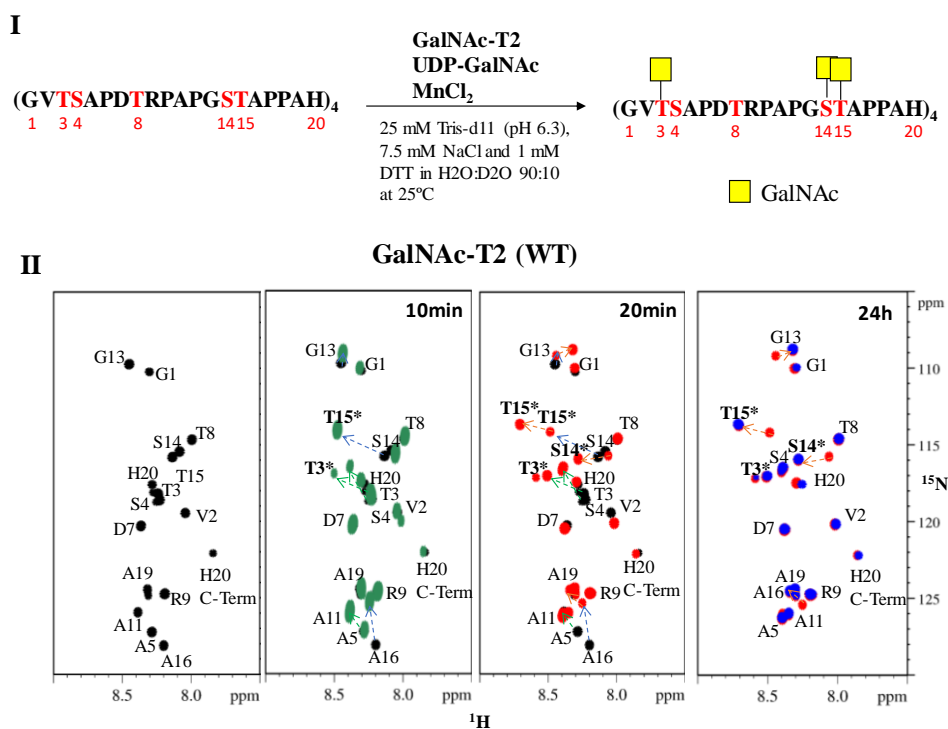
of GalNAc-T2 WT and the GalNAc-T2 D<sub>458</sub>R mutant. The <sup>1</sup>H/<sup>15</sup>N-HSQC spectra were recorded over time and after adding an excess or restricted amount of UDP-GalNAc.

Figure 2.4.7 - Panel II shows the <sup>1</sup>H/<sup>15</sup>N-HSQC of <sup>15</sup>N-labeled MUC1-4TR in the presence of GalNAc-T2 WT and MnCl<sub>2</sub> after addition of UDP-GalNAc (excess) at 0, 10 min, 20 min and 24 h. The shifts observed in the spectra are due to GalNAc transfer to the Ser and Thr residues of MUC1-4TR by GalNAc-T2. At 10 min, the observed shifts correspond to complete glycosylation of Thr15 (blue arrows) along with partial glycosylation of Thr3 (green arrows). Glycosylation of Thr15 induces strong chemical shift perturbation (CSP)<sup>i</sup> of Thr15 (0.42 ppm). However, also the vicinal amino acids suffer CSP<sup>ii</sup>. Ala16 experiences significant CSP (0.60 ppm), while Ser14 and Gly13 experience small ones (0.05 ppm and 0.11 ppm, respectively). Glycosylation of Thr3 prompts its corresponding CSP (0.44 ppm) along with that of Ser4 (0.37 ppm), Ala5 (0.29 ppm) and Val2 (0.06 ppm). The significant CSP of the neighboring residues to the glycosylation site indicates an alteration of the chemical and structural environment of MUC1 TR domain due to the presence of GalNAc units.

---

<sup>i</sup> The values of chemical shift presented are the combined chemical shift calculated to each backbone NH groups of MUC1-4TR.<sup>79</sup>

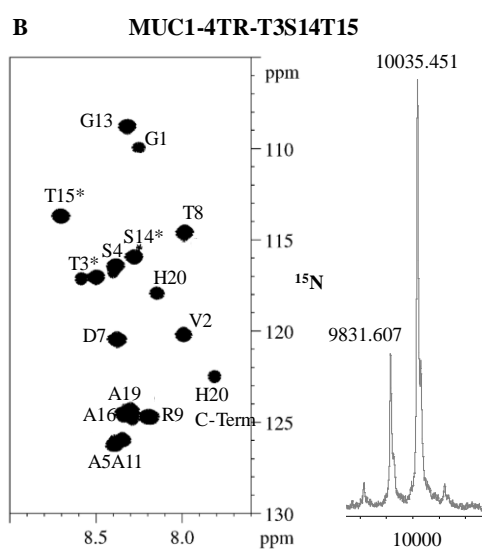
<sup>ii</sup> See on conformational studies, section 2.4.2.1.



**Figure 2.4.7 – I:** Scheme of the MUC1-4TR glycosylation event by GalNAc-T2 under the NMR experimental conditions. The amino acid sequence of MUC1-4TR displays in red the glycosylation sites; **II** - Glycosylation of MUC1-4TR by GalNAc-T2 (WT) in the presence of an excess of UDP-GalNAc, over time. HSQC of MUC1-4TR in the presence of GalNAc-T2 and MnCl<sub>2</sub> before (black) and after addition of UDP-GalNAc (in excess) at 10 min (green), 20 min (red) and 24 h (blue). The arrows indicate the observed CSP that occur as result of GalNAc additions to the MUC1 at Thr15 (blue arrows), Thr3 (green arrows) and Ser14 (orange arrows). The \* labeling in the amino acid on the spectra indicates the glycosylation site.

After 20 min, the glycosylation at Thr3 is completed and a new peak arise for Ser14 which corresponds to the third glycosylation event catalyzed by GalNAc-T2 (orange arrows). After 20 min, only 75 % of Ser14 was glycosylated, which matches with the glycosylation of three tandems out of the four present in MUC1-4TR. The complete glycosylation of Ser14 (the rest 25 %) was only reached at the end of 24 h. Glycosylation of Ser14 induces its corresponding CSP (0.16 ppm) and that at the neighbor glycosylated Thr15\* (\* indicates GalNAc, 0.19 ppm). Furthermore, CSP were also detected for Ala16 (0.19 ppm) and Gly13 (0.12 ppm).

After 24h, MUC1-4TR contains 12 GalNAc residues as confirmed by mass spectrometry (Figure 2.4.8). The full NMR assignment of the product allowed to confirm that Thr15, Thr3 and Ser14 are glycosylated, while Ser4 and Thr8 remain non-glycosylated (Figure 2.4.8, Table S6 on the supporting information). Indeed, even longer reaction times and further additions of GalNAc-T2 did not yield glycosylation at Ser4 or Thr8.

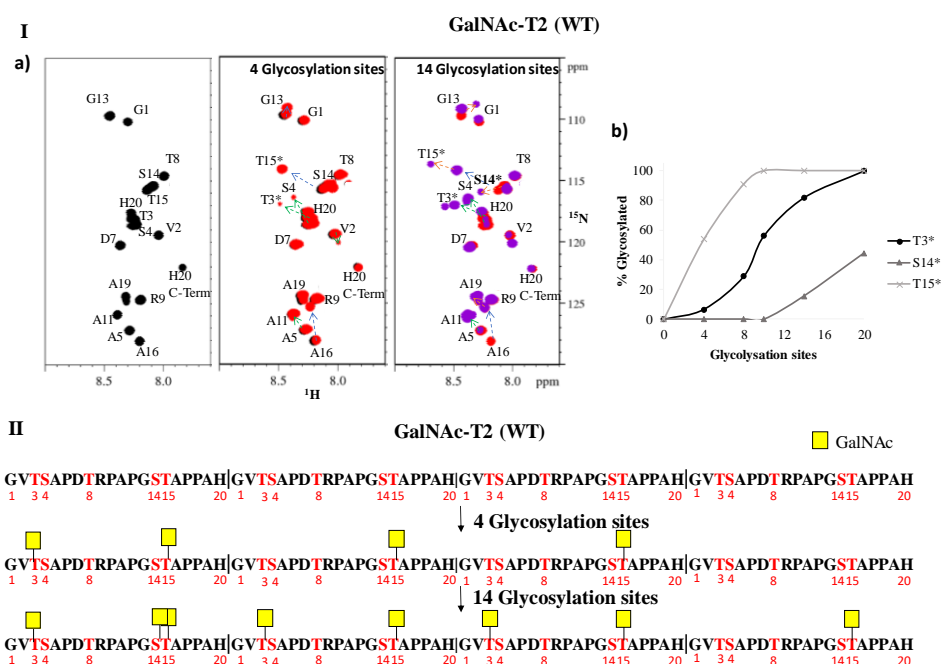


**Figure 2.4.8** – The purified MUC1-4TR-T3S14T15 product. **Left:**  $^1\text{H}/^{15}\text{N}$ -HSQC with the corresponding NMR assignment. The \* labeling in the amino acid on the spectra indicates glycosylation. **Right:** MALDI-TOF spectrum.

The *O*-glycosylation process of MUC1-4TR by GalNAc-T2 WT was also exhaustively monitored. For that purpose, restricted amount of the donor substrate (UDP-GalNAc) were sequentially added to a  $^{15}\text{N}$ -labeled MUC1-4TR sample in the presence of  $\text{MnCl}_2$  and GalNAc-T2 and  $^1\text{H}/^{15}\text{N}$ -HSQC experiments were carried out after each addition (Figure 2.4.9).

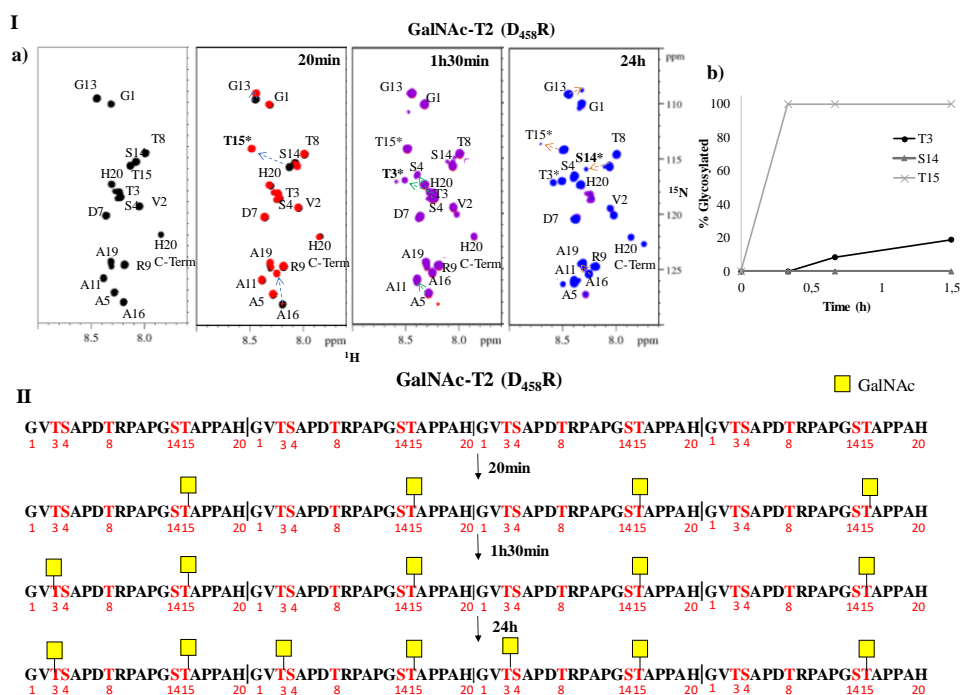
The first  $^1\text{H}/^{15}\text{N}$ -HSQC was acquired with the minimum UDP-GalNAc concentration required to glycosylate 4 glycosylation sites out of the 20 available

in MUC1-4TR. The analysis (Figure 2.4.9, Panel I-a, 4 sites) evidenced that GalNAc-T2 prefers to glycosylate Thr15 (blue arrows). However, before finishing glycosylation of all the Thr15 residues, GalNAc-T2 starts the addition of GalNAc at Thr3 (green arrows). Figure 2.4.9 – Panel I-b noticeably shows that, at this point (4 sites), 54 % of the Thr15 residues are glycosylated (two TR domains) and only 7 % of Thr3 contains GalNAc. Thus, it is expected that glycosylation of Thr3 takes place in a TR domain with a GalNAc-containing Thr15 moiety (Figure 2.4.9- Panel I-a 4 sites). As previously mentioned, GalNAc-T2 uses the lectin domain to assist the long-range glycosylation of glycopeptides and thus it is capable to glycosylate the N-terminal site (Thr3), which is 12 amino acids far from the prior glycosylated site at the C-terminus (Thr15).<sup>23</sup> After adding the proper amount of UDP-GalNAc that may allow the glycosylation of the 20 glycosylation sites (Figure 2.4.9 Panel I-b), Thr15 and Thr3 sites are fully glycosylated (100 %, 4 TR domains), while Ser14 presents a glycosylation around 45 %. From Figure 2.4.9-Panel I-a/b (UDP-GalNAc for 14 sites), it is also possible to deduce that glycosylation in Ser14 occurs after full glycosylation of Thr15 (100 %, 4 TR domains) and when glycosylation of Thr3 is almost finished (83 %, 3 TR domains). Hence, it is very likely that glycosylation in Ser14 is initiated in a TR where both Thr3 and Thr15 are already glycosylated (Figure 2.4.9 – Panel II 14 sites).



**Figure 2.4.9 - I** - Glycosylation of MUC1-4TR by GalNAc-T2 (WT) following addition of small amounts of UDP-GalNAc (step-by-step). **a)**  $^1\text{H}/^{15}\text{N}$ -HSQC of MUC1-4TR in the presence of GalNAc-T2 and  $\text{MnCl}_2$  before (black) and in the presence of UDP-GalNAc for 4 glycosylation sites (red) and 14 glycosylation sites (purple). The arrows indicate CSP that occur as result of GalNAc additions to the MUC1 at Thr15 (blue arrows), Thr3 (green arrows) and Ser14 (orange arrows). The \* labeling in the amino acid on the spectra indicates glycosylation. **b)** Relative glycosylation of MUC1-4TR by GalNAc-T2 (WT) with control addition of UDP-GalNAc (step-by-step). **II** - Schematic representation of the products obtained in presence of GalNAc-T2 (WT) after the step-by-step experiment. The glycosylation sites are displayed in red.

As mentioned above, glycosylation of MUC1-4TR was also investigated by using the D<sub>458</sub>R mutant. For that purpose, distinct  $^1\text{H}/^{15}\text{N}$ -HSQC of MUC1-4TR in the presence of GalNAc-T2 D<sub>458</sub>R and  $\text{MnCl}_2$  and after addition of UDP-GalNAc (excess) were recorded. In particular, Figure 2.4.10 – Panel I-a shows the  $^1\text{H}/^{15}\text{N}$ -HSQC spectra at 0, 20 min, 1 h 30 min and 24 h.



**Figure 2.4.10 – I -** Glycosylation of MUC1-4TR by the mutant GalNAc-T2 (D<sub>458</sub>R) in presence of an excess of UDP-GalNAc, over time. **a)** <sup>1</sup>H/<sup>15</sup>N-HSQC spectrum of the MUC1-4TR in presence of the mutant GalNAc-T2 (D<sub>458</sub>R) before addition of UDP-GalNAc (black) and after addition of an excess of UDP-GalNAc at 20 min (red), 1 h 30 min (purple) and 24 h (blue). Arrows indicate peak shifts that occur as result of GalNAc additions to the MUC1 at Thr15 (blue arrows), at Thr3 (green arrows) and at Ser14 (orange arrows). The \* labeling in the amino acid on the spectra indicates glycosylation. **b)** Relative glycosylation of MUC1-4TR by GalNAc-T2 (D<sub>458</sub>R) over time. **II -** Schematic representation of the products obtained in presence of the mutant GalNAc-T2 (D<sub>458</sub>R). The amino acid sequence of the MUC1-4TR displays in red the glycosylation sites.

After 20 min, new HSQC peaks arise only for Thr15, Ala16, Ser14 and Gly13 residues, supporting that Thr15 is glycosylated (Figure 2.4.10 Panel I-a, 20 min). In contrast to GalNAc-T2 WT, full glycosylation at Thr15 occurs after 20 min (100 %, 4 TR domains) with no glycosylation at Thr3 (Figure 2.4.10 Panel I-b, 20 min). No CSP of Thr3 residues and neighboring amino acids are perceived after 20min. After 1 h 30 min, glycosylation of Thr3 around 20 % takes place (Figure 2.4.10 Panel I-a/b, 1h 30 min). Thus, after finishing glycosylation of all Thr15 sites the mutant initiates the glycosylation at Thr3 (Figure 2.4.10 Panel II).

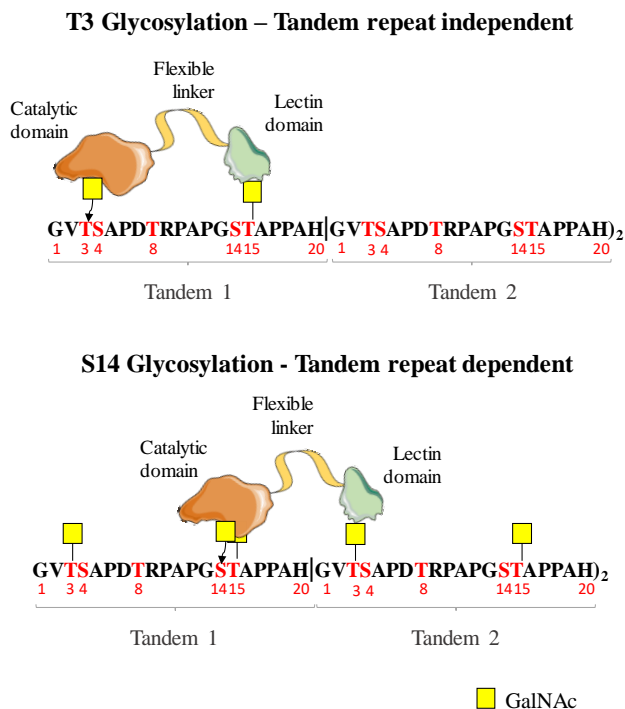


After 24 h, (Figure 2.4.10 Panel I-a) glycosylation at Thr3 was 80 % (Figure 2.4.10 Panel II) with a very small fraction at Ser14 (5 %). These results show that the D<sub>458</sub>R mutant is able to glycosylate Thr3 and Ser14 sites, although the catalytic efficiency of the enzyme is severely affected (for the WT enzyme and after 20 min of incubation with the substrates, both Thr15 and Th3 are fully O-glycosylated while Ser14 is 75% O-glycosylated (Table 2.4.1).

**Table 2.4.1** – Percentage of glycosylation / min at Thr3, Ser14 and Thr15 by GalNAc-T2 WT and GalNAc-T2 D<sub>458</sub>R.

	<i>GalNAc-T2 (WT)</i> (%/min)	<i>GalNAc-T2 (D<sub>458</sub>R)</i> (%/min)
<b>T3</b>	4,6	0,2
<b>S14</b>	3,5	0
<b>T15</b>	5	5

As expected, glycosylation at Thr15 is not affected by the mutation, while the kinetics of glycosylation of Thr3 and Ser14 are strongly modulated by the lectin binding function. However, the D<sub>458</sub>R mutant is still capable to finish the glycosylation at Thr3. In contrast, glycosylation at Ser14 is severely compromised. GalNAc-T2 WT glycosylates Thr15 depending exclusively of the catalytic domain and glycosylates Thr3 and Ser14 with assistance of the lectin domain. For glycosylation of Thr3, the lectin binds the GalNAc moiety located at Thr15 of the same TR domain (distant 12 amino acids of the glycosylation site, TR independent), while for Ser14 glycosylation, the lectin binds Thr3 of the downstream TR domain (distant 9 amino acids of the glycosylation, TR dependent) (Figure 2.4.11).



**Figure 2.4.11** – Schematic representation of the glycosylation processes at Thr3 and Ser14 in MUC1 by GalNAc-T2.

Therefore, glycosylation at Ser14 depends on the glycosylation at Thr3 of the following TR domain. Indeed, for GalNAc-T2 WT, a Ser14 site corresponding to one TR domain of MUC1-4TR is difficult to glycosylate (after 20 min, 75 % of Ser14 are glycosylated, 3 TR domains). An HSQC recorded after 16 h shows that glycosylation at Ser14 only progresses up to 85 %. Fully glycosylation of all Ser14 (100 %, 4 TR domains) is only achieved after 24 h. This result points out that the last Ser14 to be glycosylated is located at the C-terminal TR domain of MUC1-4TR, and without assistance of the lectin domain. Indeed, a mass around 9831 m/z was detected corresponding to a product of MUC1-4TR with 11 GalNAc units attached (Figure 2.4.8).

The difficulty to glycosylate Ser14 by GalNAc-T2 has previously been reported.<sup>34</sup> The presence of a GalNAc moiety at the adjacent Thr15 residue strongly

influences the catalytic efficiency of the enzyme. Molecular interactions between GalNAc at Thr15 and the catalytic domain cannot be ruled out, prompting the inhibition of the catalytic center by the glycosylated product.

#### **2.4.1.2.1.1 Summary:**

— **Thr15** is the **first glycosylation site** catalyzed by GalNAc-T2. This glycosylation of Thr15 is **not affected with the D<sub>458</sub>R mutation**. The GSTA region is the natural preference for the catalytic domain at MUC1. The **GalNAc-T2 displays preference to glycosylate at the GSTA region** in MUC1-4TR.

— Glycosylation of **Thr3** is the **second glycosylation site** catalyzed by GalNAc-T2. The glycosylation of this site is **assisted by the lectin domain** and is **TR independent**. The lectin binds to a GalNAc moiety located at the Thr15 of the same TR domain.

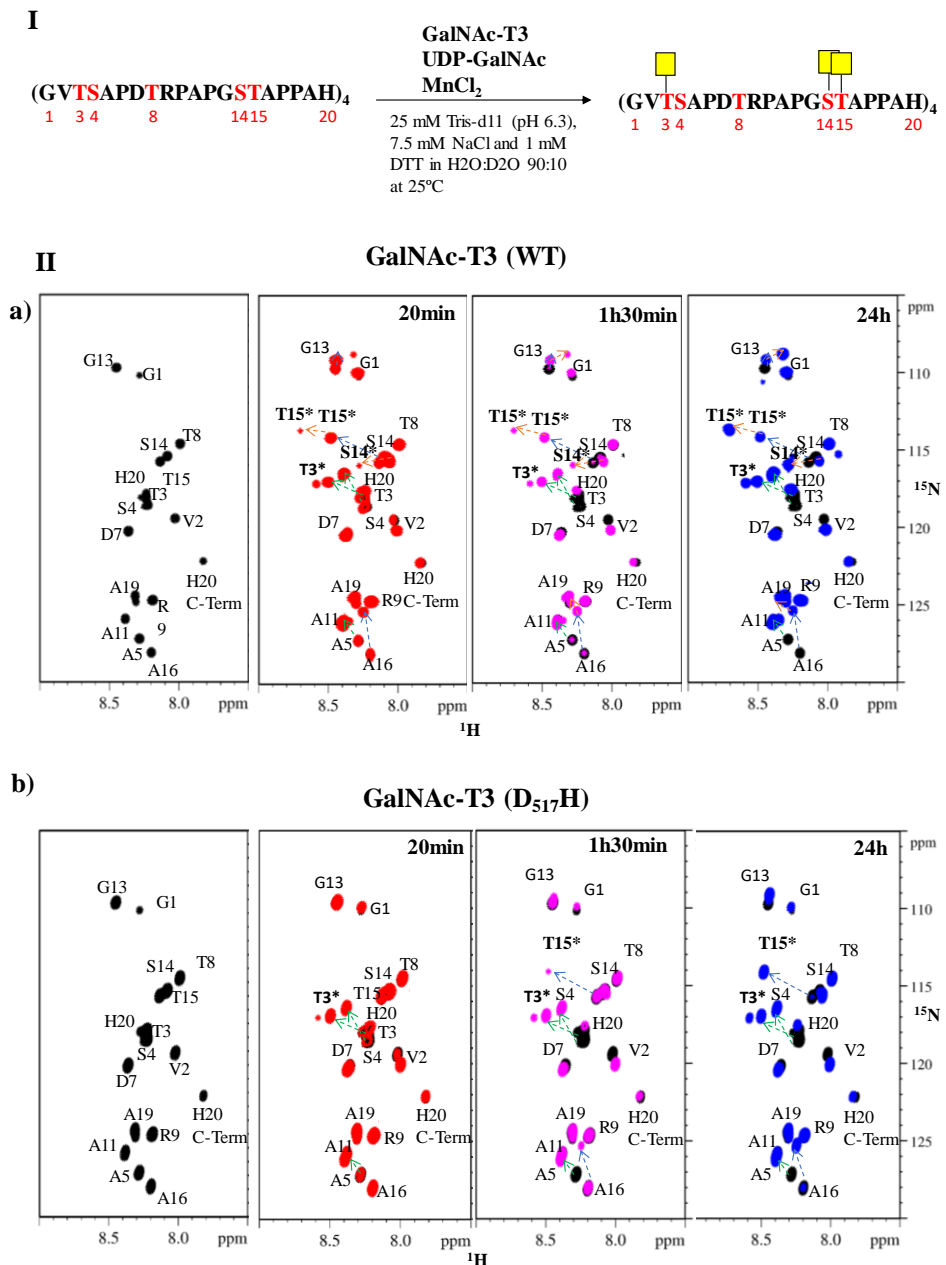
— Finally, the glycosylation of **Ser14** is **assisted by the lectin domain** and is **TR dependent**. Herein, the lectin binds to a GalNAc moiety located at Thr3 of the downstream TR domain.

#### **2.4.1.2.2 GalNAc-T3 WT vs GalNAc-T3 D<sub>517</sub>H**

The employed methodology to monitor the *O*-glycosylation process of MUC1-4TR by GalNAc-T2 and their mutant was also followed to investigate the glycosylation process of MUC1-4TR by GalNAc-T3 WT and its corresponding D<sub>517</sub>H mutant. GalNAc-T3 is able to initiate glycosylation of MUC1, modifying the same three sites as GalNAc-T2 (Figure 2.4.12 - Panel I).<sup>34</sup> Nevertheless, the order of glycosylation preference is different than that for GalNAc-T2. GalNAc-T3 glycosylates Thr3 first and then, Thr15.<sup>34</sup> GalNAc-T3 preferentially glycosylates

peptides with a Val residue at -1.<sup>66</sup> As GalNAc-T2, GalNAc-T3 ends with the transfer of GalNAc to Ser14.

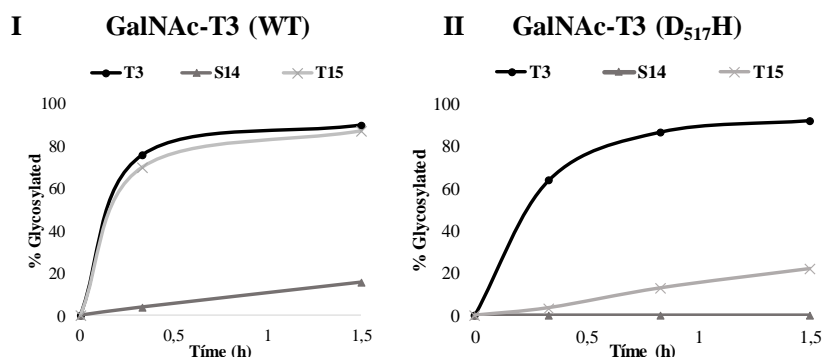
Figure 2.4.12 - Panel II-a and II-b shows the HSQC of <sup>15</sup>N-labeled MUC1-4TR in the presence of MnCl<sub>2</sub> and GalNAc-T3 WT and GalNAc-T3 D<sub>517</sub>H, respectively, after addition of UDP-GalNAc (excess) at 0, 20 min, 1 h 30 min, and 24 h. After 20 min of the addition of UDP-GalNAc, the CSPs correspond to the partial glycosylation at Thr3 (blue arrows) and Thr15 (green arrows) with glycosylation percentages around 76 % and 70 %, respectively (Figure 2.4.12 Panel II-a). In addition, a very small fraction of Ser14 is glycosylated (less than 5 %) (Figure 2.4.13 panel I). For the D<sub>517</sub>H mutant, the HSQC after 20 min only shows CSP for Thr3, Val2, Ser4 and Ala5 compatible with glycosylation at Thr3 (Figure 2.4.12 Panel II-b). At this point, 65 % of Thr3 of MUC1-4TR, which corresponds to 2-3 TR domains, are glycosylated. No glycosylation at Thr15 was detected. Indeed, for the mutant, just after 1 h 30 min, after the conclusion of glycosylation at Thr3, the transfer of GalNAc at Thr15 is initiated (around 25 %, one TR domain). After the same period, in the presence of GalNAc-T3 WT, glycosylation at Thr3 and Thr15 are ca. 100 %. These observations are rather similar to those perceived for the analogue mutant GalNAc-T2 D<sub>458</sub>R.



**Figure 2.4.12 – I:** Scheme of MUC1-4TR glycosylation by GalNAc-T3 under the NMR experimental conditions. the glycosylation sites are displayed in red; **II:** Glycosylation of MUC1-4TR construct by GalNAc-T3 (WT) and mutant D<sub>517</sub>H, in presence of excess of UDP-GalNAc. **II-a):** HSQC of MUC1-4TR in the presence of GalNAc-T3 (WT) before (black) and after addition of UDP-GalNAc (in excess) at 20 min (red), 1 h 30 min (pink) and 24 h (blue). Arrows indicate CSPs that occur as result of GalNAc additions at Thr3 (green arrows), Thr15 (blue arrows) and Ser14 (orange arrows). The \* labeling in the amino acid on the spectra indicates glycosylation. **II-b)** HSQC of the MUC1-4TR in presence of the GalNAc-T3 D<sub>517</sub>H mutant before (black) and after addition of an excess of UDP-GalNAc at

20 min (red), 1 h 30 min (pink) and 24 h (blue). Arrows indicate CSP that occur as result of GalNAc additions to the MUC1 at Thr3 (green arrows), Thr15 (blue arrows) and Ser14 (orange arrows).

Figure 2.4.13 Panel I and II show the plots of the percentage of glycosylation in function of time for GalNAc-T3 and the D<sub>517</sub>H mutant, respectively. Glycosylation at Thr15 and Ser14 by GalNAc-T3 is strongly affected by the mutation in the lectin domain.



**Figure 2.4.13-** Relative glycosylation of MUC1-4TR over time. **Panel I.-** In the presence of GalNAc-T3 WT. **Panel II.-** In the presence of the GalNAc-T3 D<sub>517</sub>H mutant.

After 24 h, for GalNAc-T3 WT, the fraction of glycosylation at Ser14 is ca. 70 % (Figure 2.4.13 Panel II-a), contrasting with that observed for the D<sub>517</sub>H mutant, where no glycosylation at Ser14 is detected (Figure 2.4.13 Panel II-b). For GalNAc-T3 WT, long time reaction (72 h) yielded glycosylation at Ser14 close to 88 %, in contrast with the mutant, which is unable to transfer GalNAc at Ser14.

Table 2.4.2 shows the percentage of glycosylation between the GalNAc-T3 WT and the D<sub>517</sub>H mutant. Thus, while glycosylation at Thr3 by GalNAc-T3 is only dependent on the natural affinity of the catalytic domain to the GVTS region of

MUC1, remote glycosylation at Thr15 and Ser14 is assisted by the lectin domain. In contrast to Thr3 residue, the percentage of glycosylation/min of Thr15 and Ser14 amino acids is severally decreased by the mutation in the lectin domain (Table 2.4.2).

**Table 2.4.2** – Percentage of glycosylation / min at Thr3, Ser14 and Thr15 by GalNAc-T3 WT and GalNAc-T3 D<sub>517H</sub>.

	<i>GalNAc-T3 (WT)</i>	<i>GalNAc-T3 (D<sub>517H</sub>)</i>
	( <i>%/min</i> )	( <i>%/min</i> )
<b><i>T3</i></b>	3.8	3.2
<b><i>S14</i></b>	0.2	0
<b><i>T15</i></b>	3.5	0.2

Therefore, to accomplish long-range glycosylation, GalNAc-T3 uses the lectin domain to bind the GalNAc moiety located at Thr3 and directs the transfer of GalNAc to Thr15 and Ser14 at the same TR domain. This result is in agreement with the preference of GalNAc-T3 to glycosylate C-terminal sites remote from the prior N-terminal GalNAc glycosites in glycopeptides.<sup>21,37</sup> This behavior is opposite to that observed for GalNAc-T2 and -T1 isoforms.<sup>37,38</sup> The directionality of the lectin-mediated long-range glycosylation is modulated by the linker that connects the lectin and catalytic domain in GalNAc-Ts, as reported on subchapter 2.3 of this thesis.<sup>64</sup>

As for GalNAc-T2, interactions of the GalNAc moiety at the adjacent Thr15 with the catalytic center of GalNAc-T3 are expected. Comparing GalNAc-T2 with GalNAc-T3, it is evident that GalNAc-T3 presents lower efficiency to glycosylate Ser14 than GalNAc-T2. The percentage of glycosylation / min for this residue is 3.5 for GalNAc-T2 and 0.2 for GalNAc-T3. These differences in the ability to incorporate GalNAc at Ser14 have already been described by Wandall and co-workers<sup>34</sup>. Kinetics studies assisted by mass spectrometry have demonstrated that

GalNAc-T2 glycosylates Ser14 faster than GalNAc-T1 and -T3 isoforms. The differences in the directionality of the lectin-mediated long-range glycosylation cannot explain the difference in the kinetics of glycosylation of Ser14 by the different GalNAc-Ts, since GalNAc-T1 and -T2 display the same behavior.<sup>20,29,38</sup>

#### **2.4.1.2.2.1 Summary:**

— Glycosylation at **Thr3 is the first event processed by GalNAc-T3**. This glycosylation is **not affected with the D<sub>459</sub>H mutation**. The GVTS region is the natural preference for the GalNAc-T3 catalytic domain at MUC1. The **GalNAc-T3 displays preference to glycosylate the GVTS region** in MUC1-4TR

— **Thr15 is the second site glycosylated** by GalNAc-T3. It is **assisted by the lectin domain** and is **TR independent**. The lectin binds the GalNAc moiety located at Thr3 of the same TR domain.

— The last site glycosylated by GalNAc-T3 is **Ser14**. It is **assisted by the lectin domain** and is **TR independent**. The lectin binds the GalNAc moiety located at Thr3 of the same TR domain, opposite to GalNAc-T2.

— **GalNAc-T3 has lower efficiency to glycosylate Ser14 than GalNAc-T2**.

#### **2.4.1.2.3 GalNAc-T4 WT vs GalNAc-T4 D<sub>459</sub>H**

GalNAc-T4 is the unique GalNAc-T isoform able to glycosylate the latest two sites of MUC1 sequence, the Ser residue at GVT\*SA (\* indicates that the residue is *O*-glycosylated) and the Thr moiety at the immunogenic PDTRP region.<sup>27,33,60,61</sup> Therefore, *O*-glycosylation of these two latest sites was also investigated by NMR. The purified glycosylated product MUC1-4TR-T3S14T15, obtained after enzymatic glycosylation by GalNAc-T2/T3 (see description in

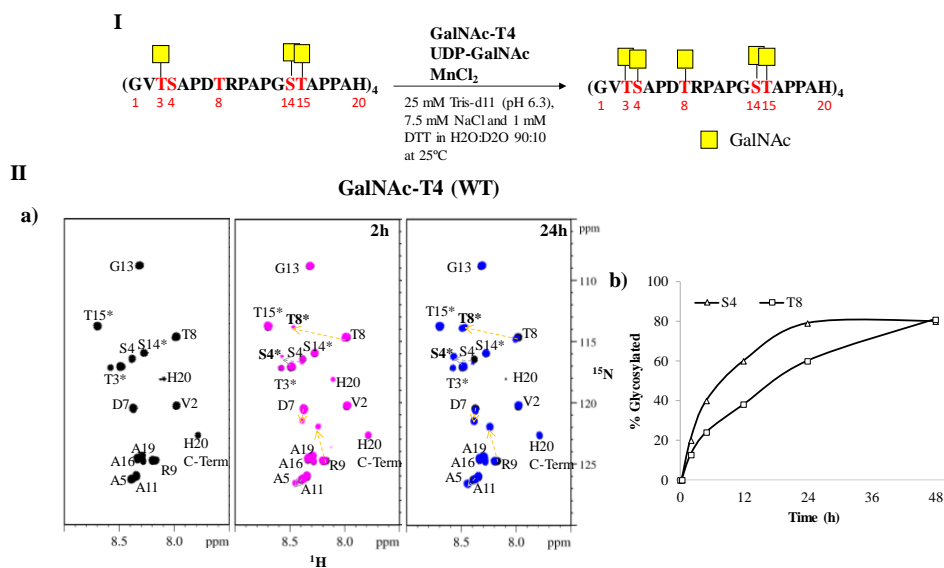


material and methods), was used as starting material (Figure 2.4.14 – Panel I). From analysis of the HSQC spectra (Figure 2.4.14 Panel II) the detected CSP arise from glycosylation at Ser4 and at Thr8. Glycosylation at Thr8 induces its strong CSP<sup>iii</sup> (0.39 ppm),<sup>iv</sup> together with CSP at Arg9 (0.62 ppm) and Asp7 (0.23 ppm). Interestingly, glycosylation at Ser4 only provides a moderate CSP at the corresponding Ser4 (0.13 ppm) and at Ala5 (0.09 ppm). Additionally, Figure 2.4.14 Panel II-a shows that GalNAc-T4 slightly prefers to glycosylate Ser4 than Thr8. After 5 h, glycosylation at Ser4 and Thr8 reaches 40 % and 24 %, respectively (Figure 2.4.14 Panel II-b). The mass spectrometry analysis shows that the product after GalNAc-T4 catalysis only contains nineteen GalNAc moieties instead of the expected twenty GalNAc residues (Figure 2.4.15). Assignment of the HSQC of the purified product after glycosylation confirms that Ser4 at the N-terminus is in fact non-glycosylated suggesting that the glycosylation of this acceptor site is lectin assisted (Figure 2.4.15). In fact, glycosylation at Ser4 by GalNAc-T4 would likely require the binding of the lectin domain to a GalNAc moiety present at the preceding MUC1 TR domain. In this context, the lectin could bind the GalNAc moiety at either Ser14 or Thr15, distant 10 or 9 amino acids, respectively, to the Ser4 glycosylation site of the subsequent TR domain.

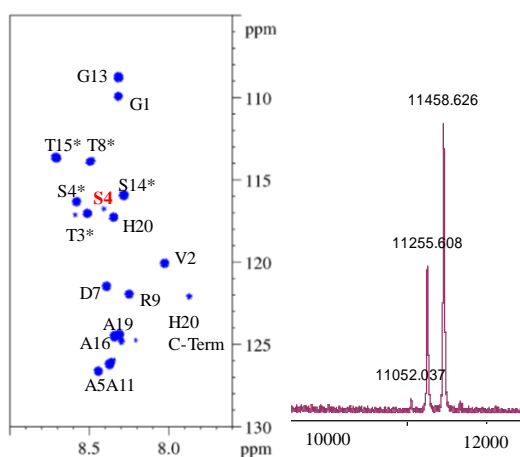
---

<sup>iii</sup> The values of chemical shift presented are the combined chemical shift calculated to each backbone NH groups of MUC1-4TR.<sup>79</sup>

<sup>iv</sup> See on conformational studies, section 2.4.2.1.

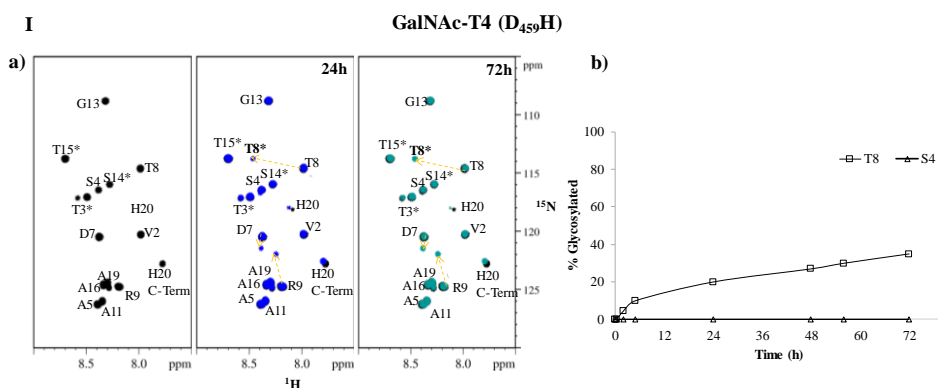


**Figure 2.4.14 – I:** The glycosylation pattern of the initially glycosylated MUC1-4TR-T3S14T15 provided by GalNAc-T4 under the NMR experimental conditions. the glycosylation sites are displayed in red; **II – a)** Glycosylation of MUC1-4TR-T3S14T15 by GalNAc-T4 WT in the presence of an excess of UDP-GalNAc. HSQC of the MUC1-4TR-T3S14T15 product in the presence of GalNAc-T4 and  $MnCl_2$  before (black) and after addition of UDP-GalNAc (in excess) at 2 h (magenta) and 24 h (blue). The arrows indicate CSP that occur as result of GalNAc additions at Ser4 (gray arrows) and at Thr8 (yellow arrows). The \* labeling in the amino acid on the spectra indicates glycosylation. **b)** Relative glycosylation of MUC1-4TR-T3S14T15 by GalNAc-T4 (WT) along the reaction time.



**Figure 2.4.15 -** Data of purified MUC1-4TR-T3S4T8S14T15 product. **Left:** HSQC with the corresponding NMR assignment. The \* labeling in the amino acid on the spectra indicates glycosylation. **Right:** MALDI-TOF spectrum.

It has been previously demonstrated that GalNAc-T4 displays a strict dependence on the prior glycosylation of MUC1 substrate.<sup>33,61</sup> Therefore, to address whether Ser4 and Thr8 are likely to be glycosylated in a lectin domain dependent manner, the D<sub>459</sub>H mutant was utilized. From HSQC analysis (Figure 2.4.16 Panel D), it is clear that mutation at the lectin domain affects glycosylation at Ser4 and Thr8 in different manners. Hence, our results indicate that the abolishment of the lectin domain function entirely precludes glycosylation at Ser4, supporting the hypothesis that the lectin domain of GalNAc-T4 binds the previous TR domain to direct the catalysis at Ser4 at the succeeding TR domain. Besides to be severely affected, glycosylation at Thr8 is carried out by the D<sub>459</sub>H mutant and after 24 h, a 20 % of the Thr8 residues are glycosylated, reaching a maximum of 35 % after 72 h. Table 2.4.3 compares the percentage of glycosylation / min of Ser4 and Thr8 for GalNAc-T4 and its respective D<sub>459</sub>H mutant. The results show that not only GalNAc-T4 prefers to glycosylate Ser4 over Thr8 but also this process relies on the lectin domain for an efficient glycosylation. In addition, the mutant is not capable of glycosylating Ser4 and partly glycosylates Thr8. Our data agrees well with previous cellular studies using CHO-K1 cells overexpressing GalNAc-T4 and the D<sub>459</sub>H mutant, respectively.<sup>60</sup> These results showed an increase in the glycan occupancy of Ser4 and Thr8 in the former cell line while Ser4 glycosylation was decreased ~50 % and Thr8 glycosylation was not affected in the latter cell line.<sup>60</sup> Note also that the glycosylation rate of Ser4 is lower than the observed for Thr3 and Thr15 achieved by GalNAc-T2 and T3, respectively. However, Ser4 glycosylation rate is similar to the one observed for Ser14 by both GalNAc-T2 or T3. Both acceptor sites are glycosylated in a lectin domain dependent manner and are contiguous to a prior glycosite that likely impedes an optimal glycosylation (see subchapter 2.3).<sup>65</sup>



**Figure 2.4.16 – I a)** Glycosylation of MUC1-4TR-T3S14T15 product by GalNAc-T4 (D<sub>459</sub>H) in the presence of an excess of UDP-GalNAc. HSQC of the MUC1-4TR-T3S14T15 product in the presence of GalNAc-T4 (D<sub>459</sub>H) before (black) and after addition of UDP-GalNAc (in excess) at 24 h (blue) and 72 h (green). The arrows indicate CSP that occur as result of GalNAc additions to the MUC1 at Thr8 (yellow arrows). The \* labeling in the amino acid on the spectra indicates glycosylation. **b)** Relative glycosylation of MUC1-4TR by GalNAc-T4 D<sub>459</sub>H over time.

**Table 2.4.3 – Percentage of glycosylation / min at Ser4 and Thr8 by GalNAc-T4 WT and GalNAc-T4 D<sub>459</sub>H.**

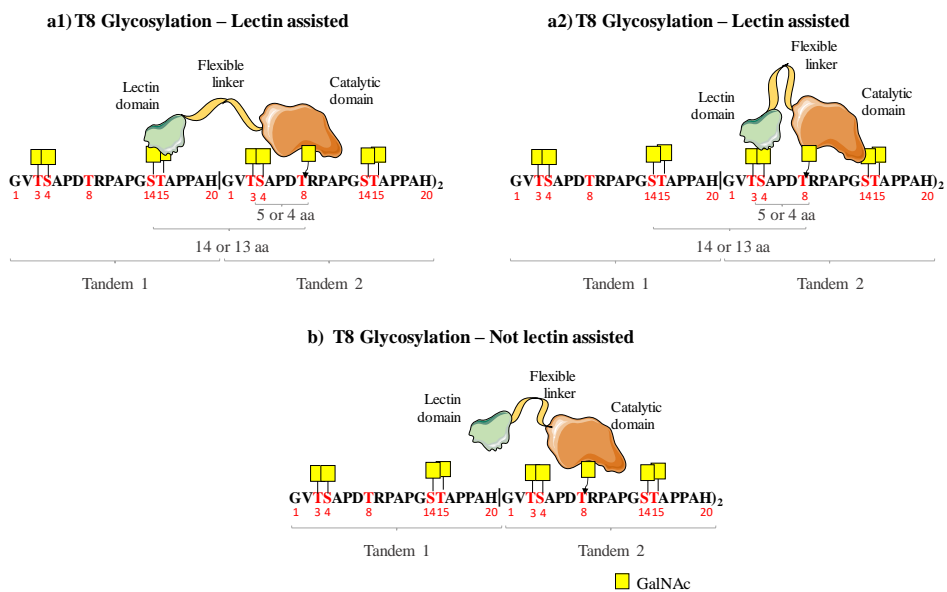
	GalNAc-T4 (WT) (%/min)	GalNAc-T4 (D <sub>459</sub> H) (%/min)
S4	0.13	0
T8	0.08	0.03

The effect of the mutation on glycosylation at Thr8 can be explained by two different scenarios (Figure 2.4.17):

a) **lectin assisted**, where the lectin domain would likely bind to either the GalNAc moiety bound to Ser14 or Thr15 of the previous TR domain (14 or 13 amino acids away from Thr8), since the binding of the lectin domain at Thr3 or Ser4 within the same TR domain is highly unlikely because of their short distance to Thr8 (5 and 4 amino acids away from Thr8). However, the catalytic domain of

GalNAc-T4 is still capable to partially glycosylate the Thr8 located at the PDTRP region.

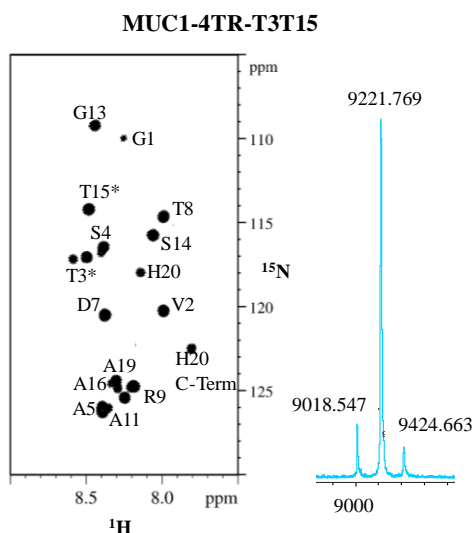
b) **not lectin assisted**, where the lectin domain would not bind to the GalNAc located at the preceding MUC1 TR domain and glycosylation at Thr8 would be entirely dependent on the catalytic domain. Thus, it is likely that glycosylation at Thr8 indirectly depends on the previously glycosylation at Ser4. In other words, in this scenario the simultaneous presence of GalNAc at Thr3, Ser4, Ser14, and Thr15 of MUC1 TR is essential to induce the conformation change of MUC1 required to favor glycosylation at Thr8 within the PDTRP region.



**Figure 2.4.17** – Schematic representation of the two possible scenarios to explain glycosylation at Thr8 by GalNAc-T4 **Panel a1)** Lectin assisted via GalNAc binding located Ser14/Thr15. **Panel a2)** Lectin assisted via GalNAc binding located at Thr3/Ser4 **Panel b)** Not lectin assisted.

The potential of GalNAc-T4 to glycosylate the Ser residue at GST\*A has also been described (Ser14 in our MUC1 template).<sup>60</sup> To investigate the glycosylation preference of GalNAc-T4 and its D<sub>459</sub>H mutant at Ser14, a MUC1-

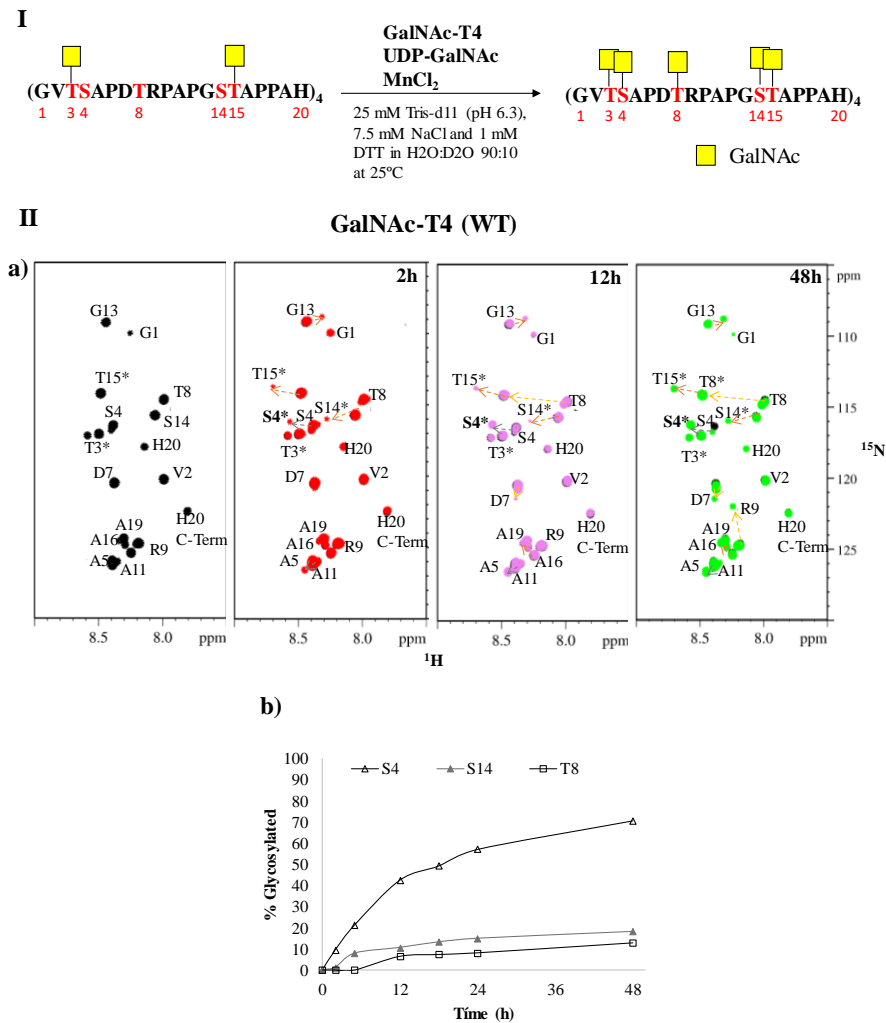
4TR peptide already glycosylated at Thr3 and Thr15 sites (MUC1-4TR-T3T15) was used as acceptor substrate. In particular, MUC1-4TR-T3T15 was obtained by glycosylation of MUC1-4TR by the GalNAc-T2 D<sub>458</sub>R mutant (see materials and methods). Figure 2.4.18 shows the HSQC and the corresponding mass of the purified MUC1-4TR-T3T15.



**Figure 2.4.18** - Data of the purified MUC1-4TR-T3T15 product. **Left:** HSQC with corresponding NMR assignment. The \* labeling in the amino acid indicates glycosylation. **Right:** MALDI-TOF spectrum.

Glycosylation of MUC1-4TR-T3T15 by GalNAc-T4 WT and its D<sub>459</sub>H mutant (Figure 2.4.19 Panel I) was monitored by HSQC-NMR experiments. The CSP are compatible with glycosylation at Ser4, along with reactions at Ser14 and Thr8 (Figure 2.4.19 Panel II-A). Furthermore, the results clearly pinpoint that Ser4 at the GVT\*SA region is preferentially glycosylated by GalNAc-T4 rather than Ser14 at GST\*AP and then Thr8 at PDTRP (Figure 2.4.19 Panel II-B). Noteworthy, after 2h, besides glycosylation at Ser4 (9.5 %), a small fraction of glycosylation at Ser14 is also detected (1 %) (Figure 2.4.19 Panel II 2h). Thus, GalNAc-T4 initiates glycosylation of Ser14 after glycosylating Ser4 and before glycosylation of Thr8

starts. In fact, the  $^1\text{H}/^{15}\text{N}$ -HSQC spectrum after 12 h shows a small percentage of glycosylation at Thr8 (6.4 %). After 48h, glycosylation at Ser4 reaches 70%, while at Ser14 and Thr8 are only 18 % and 13%, respectively, fairly similar. On the other hand, after 48 h, when MUC1-4TR-T3S14T15 is used as acceptor substrate, glycosylation at Ser4 and Thr8 reaches 79 % and 81 %, respectively. Thus, these results clearly show the relevance of the complementarity and hierarchy of GalNAc-Ts during the *O*-glycosylation process of MUC1.

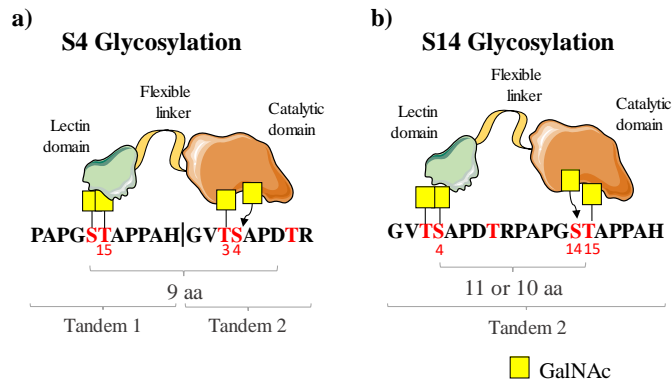


**Figure 2.4.19 – I:** Scheme of the MUC1-4TR-T3T15 glycosylation product synthesized by GalNAc-T4 (WT) under the NMR experimental conditions. The glycosylation

sites are displayed in red; **II – a**) The glycosylation process of MUC1-4TR-T3T15 by GalNAc-T4 (WT) in the presence of excess of UDP-GalNAc.  $^1\text{H}/^{15}\text{N}$ -HSQC spectrum of the MUC1-4TR-T3T15 glycosylation product in the presence of GalNAc-T4 (WT) (black) and after addition of UDP-GalNAc 2 h (red), 12 h (pink) and 48 h (green). The arrows indicate the CSP that occur at Ser4 (gray arrows), Ser14 (orange arrows) and Thr8 (yellow arrows). The \* labeling indicates the glycosylation site. **b**) The time course of relative glycosylation at the different residues.

Long-range glycosylation at Ser4 is guided by the lectin domain of GalNAc-T4. The results previously described for MUC1-4TR-T3S14T15 showed that the lectin domain could bind the GalNAc moiety either located at Ser14 or Thr15 of the upstream TR to drive the glycosylation of Ser4 residue of the downstream and contiguous TR domain. However, the percentage of glycosylation/min at Ser4 using either MUC1-4TR-T3S14T15 or MUC1-4TR-T3T15 as acceptor substrates are similar (around 0.1 %/min) (Table 2.4.4), which strongly suggests that the lectin domain of GalNAc-T4 should bind the GalNAc moiety at Thr15 of the prior TR domain (Ser4 is 9 amino acids away from the GalNAc moiety at Thr15 of the precedent TR domain, Figure 2.4.20 a). Glycosylation event at Ser14 is also probably assisted by the lectin domain (Ser14 is 11 or 10 amino acids away from the GalNAc moieties at Thr3 or Ser4 in the same TR domain, Figure 2.4.20 b). However, GalNAc-T4 prefers to glycosylate Ser4 over Ser14. This is likely explained by the different peptide sequence around the acceptor site and the different location of the prior GalNAc moiety with respect to the acceptor site (GVT\*SA vs GST\*AP, where the Ser to be glycosylated is underlined (Figure 2.4.20).





**Figure 2.4.20** – Schematic representation to illustrate the different locations of the existing GalNAc respect to the glycosylation site. **Panel a)** Glycosylation of Ser4 within the GVT\*SA sequence. **Panel b)** Glycosylation of Ser14 within the GST\*AP sequence.

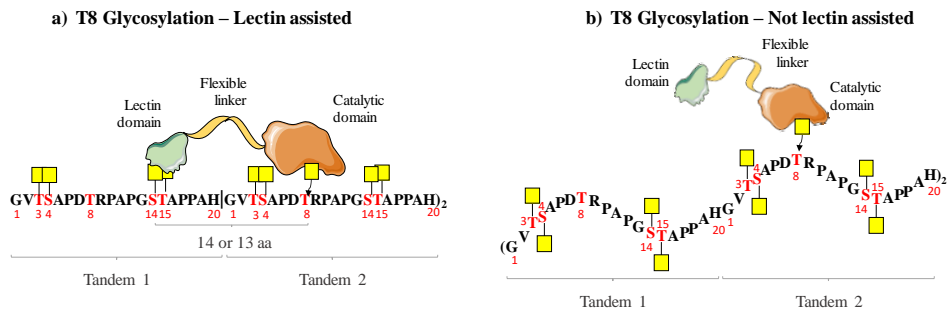
The glycosylation process of MUC1-4TR-T3T15 was also investigated in the presence of the D<sub>459</sub>H mutant. In the case of the mutant Ser4, Ser14 and Thr8 were not glycosylated (Table 2.4.4).

**Table 2.4.4** – Percentage of glycosylation / min at Ser4, Thr8 and Ser14 by GalNAc-T4 WT and GalNAc-T4 D<sub>459</sub>H.

	<i>GalNAc-T4 (WT)</i> (%/min)	<i>GalNAc-T4 (D<sub>459</sub>H)</i> (%/min)
<b>S4</b>	0.08	0
<b>T8</b>	0.01	0
<b>S14</b>	0.02	0

The absence of glycosylation at S14 and S4 was expected. However, the absence of glycosylation at Thr8 was more surprising and implies two potential hypotheses (Figure 2.4.21): a) glycosylation of Thr8 is lectin assisted and lectin binds to the GalNAc located at S14 of the preceding TR domain (in this scenario the glycosylation of Ser14 by GalNAc-T2/T3 is crucial to the further glycosylation

of Thr8 by GalNAc-T4) or b) the concomitant glycosylation of Ser4 and Ser14 modulates MUC1 conformation allowing to increase the glycosylation at Thr8 by GalNAc-T4.



**Figure 2.4.21** – Schematic representation of the two possible scenarios to explain glycosylation at Thr8 by GalNAc-T4 **Panel a)** Lectin assisted. Lectin binds to the GalNAc located at Ser14. **Panel b)** Not lectin assisted and the conformation of glycosylated MUC1 is important for glycosylation at Thr8.

#### 2.4.1.2.3.1 Summary:

— **Ser4 is the first residue** glycosylated by GalNAc-T4. Glycosylation of Ser4 is **assisted by the lectin domain** and is **TR dependent**. The lectin binds the GalNAc moiety located at Thr15 of the upstream TR domain.

— **GalNAc-T4 is also capable to glycosylate Ser14 with assistance of the lectin domain**. However, this glycosylation is **TR independent**. GalNAc-T4 glycosylates Ser14 using the GalNAc moiety located either at Thr3 or at Ser4 of the same TR domain.

— **The glycosylation of Thr8** may be **either lectin-assisted or entirely dependent on the catalytic domain**. We cannot exclude any of the two hypotheses. In the last case, the process would depend on the density of GalNAc moieties and thus, on the concomitant glycosylation at Thr3, Ser4, Ser14 and Thr15.

— The lectin domain of GalNAc-T4 binds contiguous GalNAc moieties.

#### ***2.4.1.3 Conformation of the glycosylated MUC1-4TR products modulates glycosylation preferences of GalNAc-Ts***

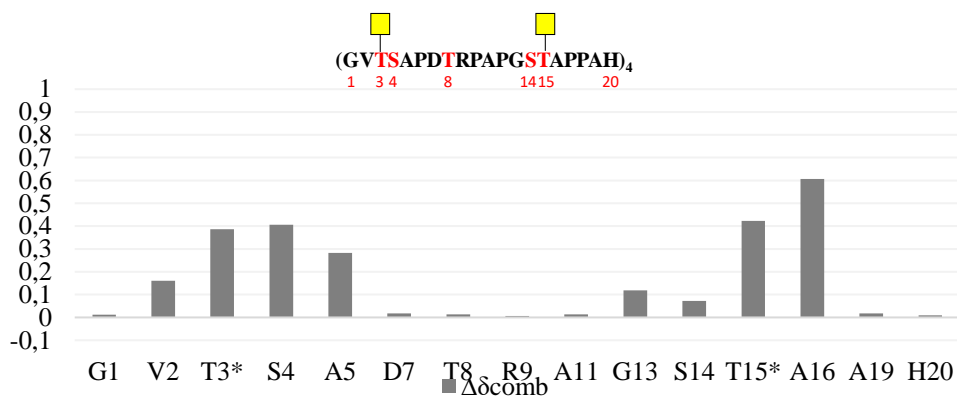
It is established that *O*-GalNAc glycosylation induces changes in the protein backbone conformations,<sup>67–70</sup> with well-defined and distinct preferences for Ser and Thr residues.<sup>71,72</sup>

The conformational analysis of the MUC1-4TR glycosylation products was carried out to evaluate the alterations on the native MUC1-4TR conformation upon *O*-GalNAc glycosylation.

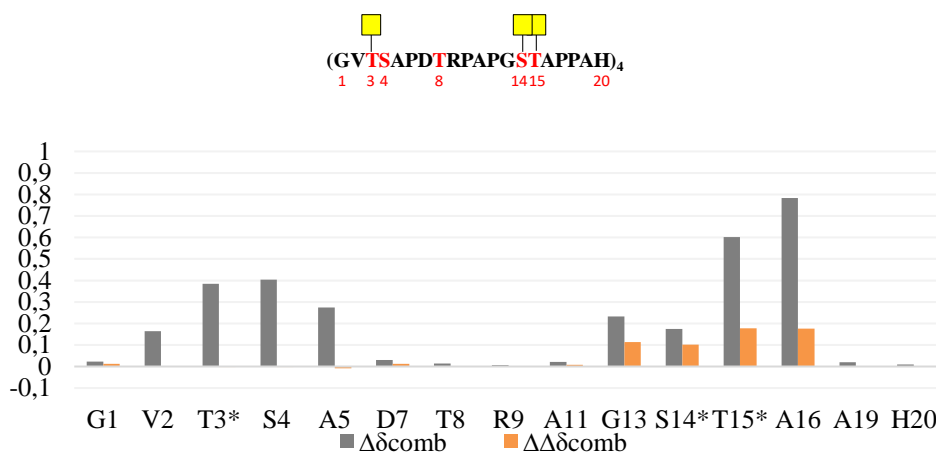
The glycosylation-induced chemical shifts were deduced<sup>v</sup> for each residue of the glycosylated MUC1-4TR-T3T15 (Figure 2.4.22), MUC1-4TR-T3S14T15 (Figure 2.4.23) and MUC1-4TR-T3S4T8S14T15 (Figure 2.4.24), using the unglycosylated MUC1-4TR as reference. In some of the cases, for comparison reasons, the data acquired for the precedent glycosylated product were also used. Figures 2.4.22 – 2.4.24 clearly illustrate that the shifts occur both at the glycosylated- and neighboring peptide residues. Furthermore, the plots illustrate the downfield shift tendency observed upon *O*-glycosylation with respect to the reference (with exception of A5 orange bar in Figure 2.4.24).

---

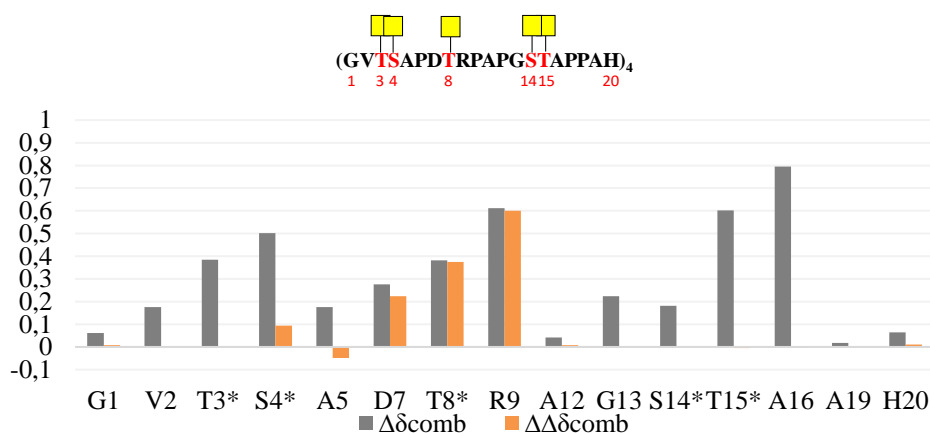
<sup>v</sup> The chemical shift presented are the combined chemical shift ( $\Delta\delta_{\text{comb}}$ ) calculated to each backbone NH groups of MUC1-4TR.<sup>79</sup>



**Figure 2.4.22** - Plot of the  $\Delta\delta_{comb}$  values for backbone NH groups of MUC1-4TR-T3T15 relative to naked MUC1-4TR. The \* indicates the glycosylation site.



**Figure 2.4.23**- Plot of the  $\Delta\delta_{comb}$  values for backbone NH groups of MUC1-4TR-T3S14T15 relative to naked MUC1-4TR (Gray bars). The \* indicates the glycosylation site.  $\Delta\Delta\delta_{comb}$  is  $\Delta\delta_{comb_{T3S14T15}} - \Delta\delta_{comb_{T3T15}}$  (Orange bars).

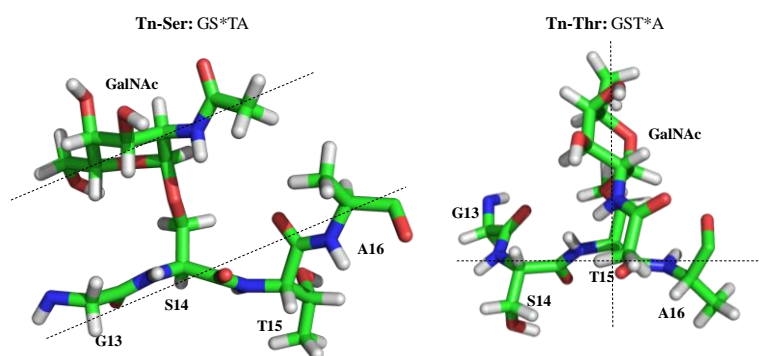


**Figure 2.4.24** - Plots of the  $\Delta\delta_{comb}$  values for backbone NH groups of MUC1-4TR-T3S4T8S14T15 relative to naked MUC1-4TR (Gray bars). The \* indicates the glycosylation site.  $\Delta\Delta\delta_{comb}$  is  $\Delta\delta_{comb_{T3S4T8S14T15}} - \Delta\delta_{comb_{T3S14T15}}$  (Orange bars).

The  $^1\text{H}$  and  $^{15}\text{N}$  NMR chemical shifts of the unglycosylated and glycosylated MUC1-4TR products are indicative of a random coil-like structure (Tables S4 - S7 in supporting information).<sup>73</sup> However, the downfield tendency upon *O*-GalNAc glycosylation is indicative of a deviation of the random coil to an extended conformation.

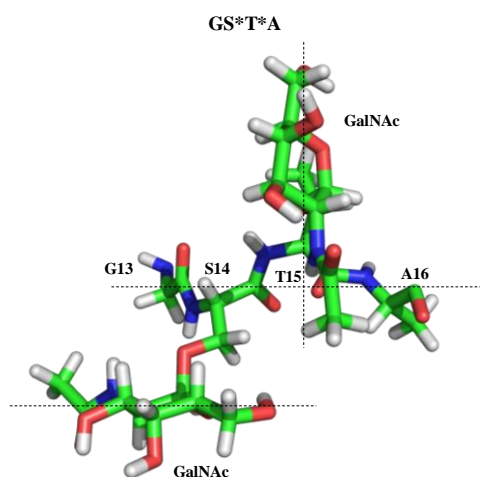
The analysis also indicates that glycosylation at Thr3, Thr15 and Thr8 induces larger chemical shift perturbations in the NH groups of the neighboring peptide residues than glycosylation at Ser4 and Ser14. Between the Thr units, glycosylation at Thr15 at GSTA region shows the larger variation in the chemical shift of the neighboring amino acids followed by glycosylation at Thr8 at the PDTR region. Another interestingly observation is that the highest chemical shift is usually experienced by the amino acid located at site +1 relative to the glycosylation site. For example, glycosylation at Thr15 (Figure 2.4.22) promotes a shift of 0.60 ppm in Ala16 and of only 0.07 ppm in Ser14 (site -1 relative to the glycosylation site). The glycosylated Thr15 shifts 0.42 ppm. The same trend was observed for glycosylation at Thr3 and Thr8. The impact on the chemical shift of the NH of the amino acid located at the site +1 respective to the glycosylation site is more prominent when the glycosylated residue is a threonine than serine.

The differences in the chemical shift perturbation induced by *O*-GalNAc glycosylation in Thr or Ser can be correlated with the differences around the glycosidic linkages of GalNAc-Thr and GalNAc-Ser. Indeed, previous studies have already demonstrated that GalNAc-Thr glycopeptides are rather rigid in solution, with a *O*-glycosidic linkage in eclipsed conformation ( $\Phi=80^\circ$  and  $\psi=120^\circ$ ), while GalNAc-Ser analogues are more flexible and adopt the *exo*-anomeric/*syn* conformation ( $\Phi=80^\circ$  and  $\psi=180^\circ$ ).<sup>71,72,74</sup> In the eclipsed geometry (GalNAc-Thr), the GalNAc moiety displays an almost perpendicular arrangement with respect to the amino acid. In contrast, in the GalNAc-Ser glycopeptides the GalNAc adopts a parallel disposition.<sup>71,72,74</sup> The conformational analysis of GalNAc-Ser and GalNAc-Thr in MUC1-derived short glycopeptides<sup>75</sup> clearly pinpoint those conformers around the glycosidic sugar-peptide linkage (Figure 2.4.25). A characteristic NOE between the GalNAc moiety and the Thr residue is indicative of the perpendicular arrangement in the case of GalNAc-Thr glycopeptides. Thus, in the particular case of MUC1-4TR, the perpendicular disposition of the GalNAc moiety attached to all Thr residues (Thr3, Thr8 and Thr15) is supported by a NOE contact between the NH of the NHAc group of GalNAc and the backbone NH of the threonine (Table S1-S3). This NOE is not present in the GalNAc-Ser motifs (Ser4 and Ser14) indicating that a parallel arrangement between the GalNAc moiety and the peptide should take place.



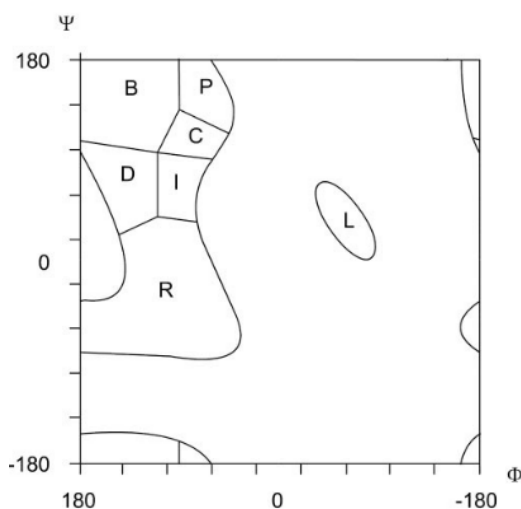
**Figure 2.4.25** – Conformations of the GalNAc-Ser and GalNAc-Thr moieties in the GSTA region of MUC1.

An identical conformational behavior for GalNAc-Ser (parallel) and GalNAc-Thr (perpendicular) takes place for the contiguous GalNAc O-glycosylation structures. Strong NOE contacts between the NH of the NHAc of GalNAc and the NH of threonine<sup>72,74</sup> were detected in products MUC1-4TR-T3S14T15 (Table S2) and MUC1-4TR-T3S4T8S14T15 (Table S3), corroborating this 3D disposition (Figure 2.4.26).



**Figure 2.4.26** – Conformation of the di-glycosylated GS\*T\*A motif of MUC1 (\* indicates the site of glycosylation).

Deviations from random-coil chemical shift values of H $\alpha$  are also indicative of tendencies for a given secondary structure.<sup>76</sup> The chemical shift index (CSI H $\alpha$ ) values (observed – random coil) for all residues of the unglycosylated MUC1-4TR and glycosylated MUC1-4TR products are indicative of the presence of extended-like conformations (Figure 2.4.28 – 2.4.30 top panel).<sup>73,76</sup> Indeed, the conformation of the unglycosylated MUC1 peptide thumbs in two major areas of the Ramachandran plot (Figure 2.4.27) both of them compatible with extended-like conformations: polyproline II-like and the  $\beta$ -strand conformations.



**Figure 2.4.27** – Conformational clusters on the Ramachandran plot.  $\psi$  and  $\Phi$  torsional angles are in degrees. Adapted from <sup>77</sup>.

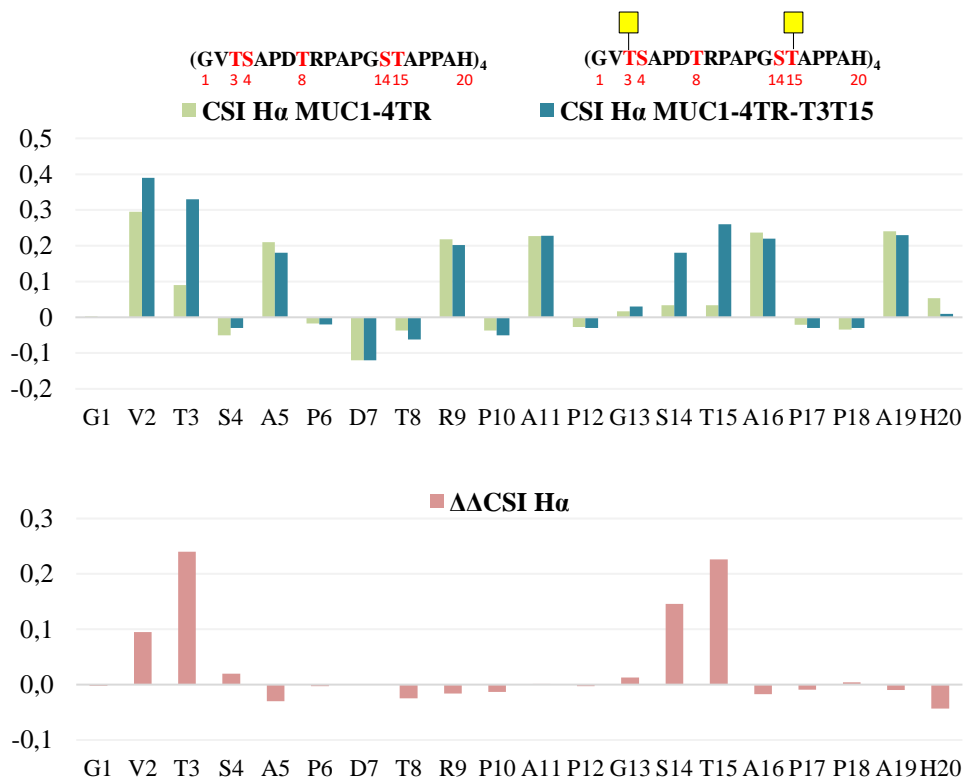
Association of the catalytic affinities of different GalNAc-Ts with the major conformations displayed by the three distinct MUC1 regions (GVTS, DTRP, GSTA) was previously reported.<sup>77</sup> By using MUC1 short peptides, Kinarsky and co-workers, demonstrated that the most populated area on the Ramachandran plot (Figure 2.4.27) for the non-glycosylated GVTS motif was the B area ( $\beta$ -strand conformation), while the DTRP region matches in D area (inverse  $\gamma$ -turn conformation) and GSTA in the P area (polyprolines II-like conformation).<sup>77</sup> GalNAc-T2 initiates MUC1 glycosylation at Thr15 of the GSTA region, indicating that the catalytic domain of GalNAc-T2 has a preference to polyprolines II-like conformations. In contrast, GalNAc-T3 initiates MUC1 glycosylation at Thr3 of GVTS region, showing that the enzyme prefers to glycosylate peptides higher populated in  $\beta$ -strand conformations. GalNAc-T4 is the single GalNAc-T able to glycosylate the DTRP region therefore showing a marked preference to glycosylate peptide holding inverse  $\gamma$ -turn conformations.

After glycosylation, the  $^1\text{H}\alpha$  chemical shift of the glycosylated residue and the residues around always experience a considerable downfield shift, pointing out



that the *O*-GalNAc glycosylation increases the population of extended-like conformations in peptides (polyprolines II-like and  $\beta$ -strand conformations).

The Figure 2.4.28 – 2.4.30 analyze the effect of glycosylation on the CSI  $^1\text{H}\alpha$  of MUC1-4TR products. To distinguish the contribution of GalNAc glycosylation on the MUC1 conformation the difference between the CSI  $^1\text{H}\alpha$  of a MUC1-4TR product versus the CSI  $^1\text{H}\alpha$  of the previous MUC-4TR structure was carried out (Figure 2.4.28 – 2.4.30 bottom panel).

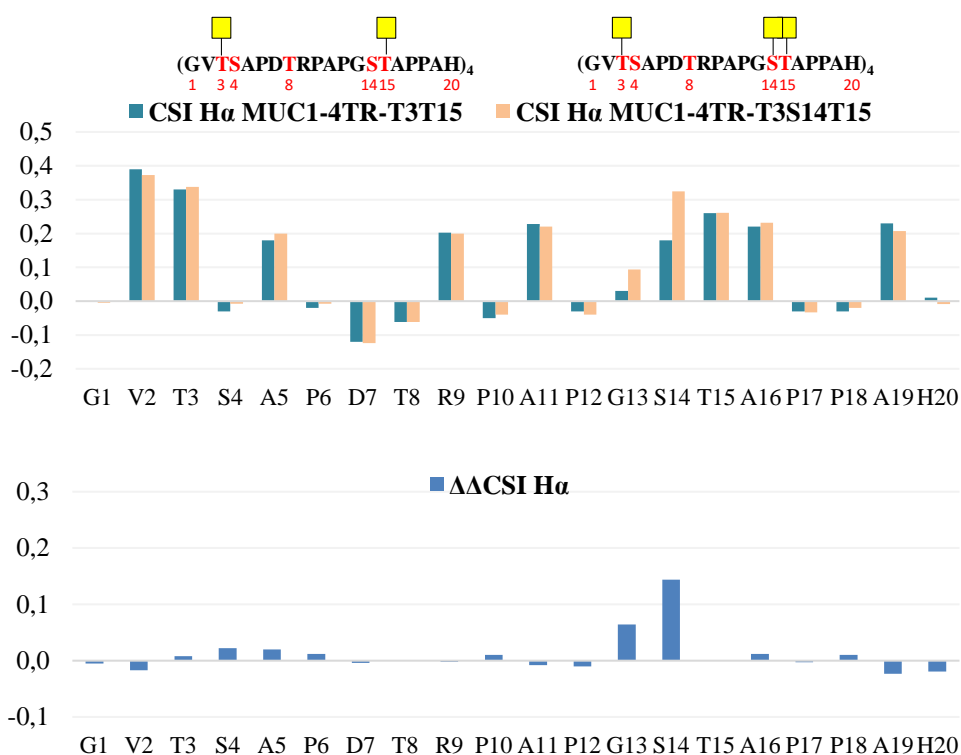


**Figure 2.4.28** – Effect of glycosylation at Thr3 and Thr15 on CSI  $^1\text{H}\alpha$  of MUC1-4TR. **Top Panel.** CSI  $^1\text{H}\alpha$  for the unglycosylated MUC1-4TR (green bars) and the MUC1-4TR-T3T15 product (dark blue bars). **Bottom Panel** Difference between CSI  $^1\text{H}\alpha$  of MUC1-4TR-T3T15 and MUC1-4TR.

Glycosylation at Thr3 and Thr15 affect two other glycosylation sites, Ser4 and Ser14, respectively. Thus, the precedent glycosylation at Thr3 and Thr15

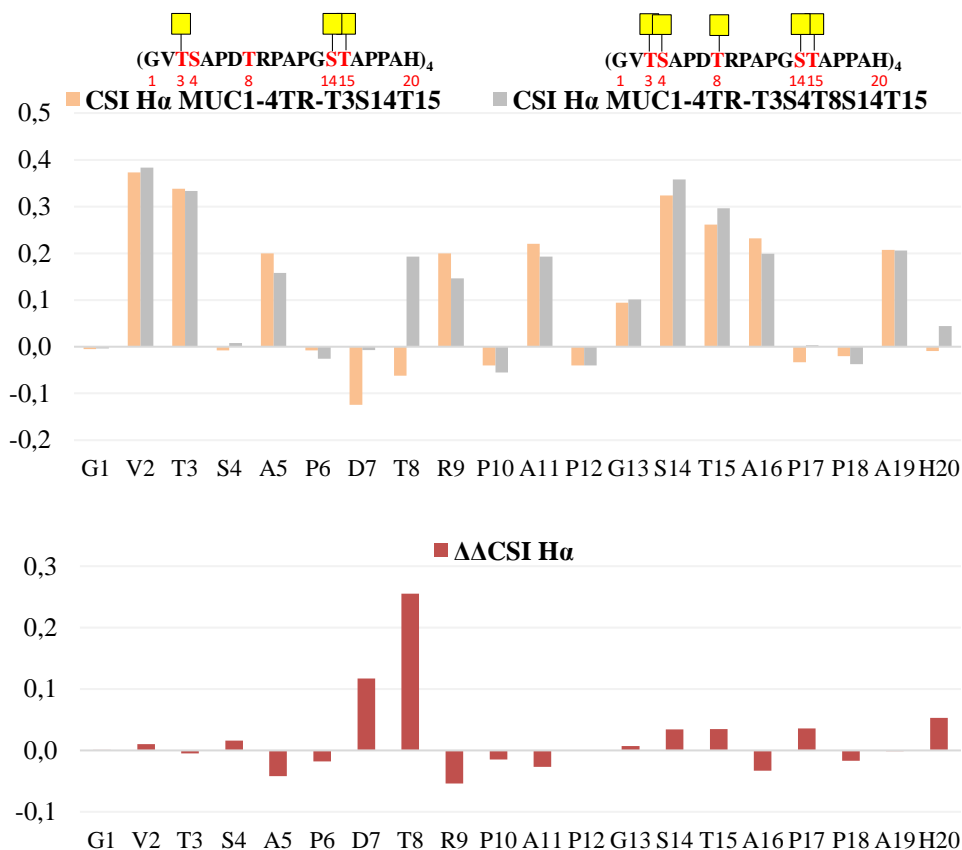
residues clear modulates the succeeding glycosylation at Ser4 and Ser14, respectively. Noteworthy, the structural perturbation at Ser14 induced by glycosylation at Thr15 is larger than that observed in the case of Ser4 (induced by glycosylation at Thr3). The orientations of the side chains of the flanking amino acids to Thr and Ser residues are affected by glycosylation. In addition, the orientation of the side chains into opposite directions is typical for extended-like conformations. In this context, a virtual torsional  $\theta (i, i+1)$  between vectors  $C\beta-C\alpha(i) \dots C\alpha-C\beta (i+1)$ , demonstrated that glycosylation at Thr15 modifies the torsional angle ca.  $80^\circ$  (from  $85^\circ$  in the unglycosylated to  $165^\circ$  in the glycosylated residue), while glycosylation at Thr3 only changes ca.  $45^\circ$  (from  $120^\circ$  to  $165^\circ$ ).<sup>77</sup> This observation explains why glycosylation at Thr15 strongly affects the conformation of the flanking Ser14.

Furthermore, glycosylation at Thr15 (GSTA region) induces a reduction of the population at B region ( $\beta$ -strand conformation) in the Ramachandran plot and an increase of the population in the D area (inverse  $\gamma$ -turn conformation).<sup>77</sup> Ser14 is the last residue that is glycosylated by GalNAc-T2/T3. Additionally, in section 2.4.2.2 we showed that the transfer of GalNAc to Ser14 by GalNAc-T2 and GalNAc-T3 is lectin assisted. However, the catalysis is less efficient in GalNAc-T3 compared to GalNAc-T2. This observation suggests that GalNAc-T3 glycosylates less favorably acceptor sites located in inverse  $\gamma$ -turn conformation.



**Figure 2.4.29** – Effect of glycosylation at S14 on CSI <sup>1</sup>H $\alpha$  of MUC1-4TR. **Top Panel.** CSI <sup>1</sup>H $\alpha$  for the MUC1-4TR-T3T15 (dark blue bars) and MUC1-4TR-T3S14T15 products (orange bars). **Bottom Panel.** Difference between CSI <sup>1</sup>H $\alpha$  of MUC1-4TR-T3S14T15 and MUC1-4TR-T3T15 products.

Ser4 at the GVT\*S region (\* indicates glycosylation site) and Thr8 at PDTR region are both glycosylated by GalNAc-T4. GalNAc-T4 preferentially glycosylate peptides that adopt inverse  $\gamma$ -turn conformations,<sup>77</sup> as that present in the PDTR motif. Also, glycosylation at Thr3 induces a change in the  $\psi$  and  $\Phi$  angles of GVTS region from  $\beta$ -strand to inverse  $\gamma$ -turn conformations. The prevalence of inverse  $\gamma$ -turn conformers in both of these regions can be associated with the catalytic preferences of GalNAc-T4 towards the MUC1 sequence.



**Figure 2.4.30** - Effect of glycosylation at Ser4 and Thr8 on CSI <sup>1</sup>H $\alpha$  of MUC1-4TR. **Top Panel.** CSI <sup>1</sup>H $\alpha$  for the MUC1-4TR-T3S14T15 (orange bars) and MUC1-4TR-T3S4T8S14T15 products (gray bars). **Bottom Panel.** Difference between CSI <sup>1</sup>H $\alpha$  of MUC1-4TR-T3S4T8S14T15 product and MUC1-4TR-T3S14T15.

The analysis of Figures 2.4.28 – 2.4.30 highlight that glycosylation at Thr residues has a major impact on the structural environment around the glycosylation site than at Ser. Specifically, our CSI data indicates that the glycosylation at Ser4 has a negligible effect on the MUC1 conformation. Additionally, a reduced number of sugar-peptide NOEs was observed upon GalNAc addition at Ser residues in comparison to those detected after GalNAc introduction at Thr (Table S1-S3). This result is in agreement with the observation that glycans attached to serine residues are likely to hold more rotational flexibility.<sup>71,72,74</sup> These observations together with the obtained NOE data (Table S1-S3) pinpoint that GalNAc-Ser barely modulates

the peptide conformation around the glycosylation site, when compared with GalNAc-Thr.

This last conclusion rules out the hypothesis that concomitant glycosylation at Ser4 and Ser14 modulates MUC1 conformation allowing to increase the glycosylation event at Thr8 by GalNAc-T4. Hence, glycosylation at Thr8 by GalNAc-T4 should be assisted by the lectin domain, with the lectin recognizing the GalNAc attached to Ser14 of the preceding TR domain.

#### **2.4.1.3.1.1 Summary**

- The ***O*-GalNAc glycosylation at Thr moieties induces higher chemical shift perturbations than at Ser units**. Thus, **glycosylation at Thr generate larger structural variations**.
  
- The **effect of the glycosylation at the two serines (Ser4 and Ser14) is minimal and does not affect the conformation of APDTR region**.

#### **2.4.2 Conclusions**

The study presented herein provides the evidence that GalNAc-Ts glycosylate in a highly-ordered form the MUC1 substrate. Besides that, all the TR domains of MUC1 are fully *O*-glycosylated in a stepwise manner. Within this work, we have provided compelling evidences that the interplay between the lectin and catalytic domain of these enzymes, for the catalysis of all glycosylation sites of MUC1, is crucial.

GalNAc-T2 and GalNAc-T3 are responsible to glycosylate the same three glycosylation sites at MUC1 (Thr3, Ser14 and Thr15). However, each enzyme has

own preferences and directionality. GalNAc-T2 prefers GSTA sequence (Thr15) at a naked MUC1 and, in contrast GalNAc-T3 has preference for GVTS region (Thr3). These preferences of the catalytic domain should be correlated with the conformation of each region at MUC1. The other two sites of glycosylation (Thr3 in case of GalNAc-T2, Thr15 in case of GalNAc-T3 and Ser14 in both of cases) are glycosylated using the lectin domain to guide the glycosylation process. When the lectin domain is disrupted, considering the mutants GalNAc-T2 D<sub>458</sub>R and GalNAc-T3 D<sub>517</sub>H, the glycosylation of these sites was highly affected. Additionally, in the case of GalNAc-T3 D<sub>517</sub>H, the mutation disallows the glycosylation of Ser14.

GalNAc-T4 is the only GalNAc-T isoform able to glycosylate the last two sites of MUC1 sequence, the Ser residue at GVT\*SA (\* indicates that the residue is *O*-glycosylated) (Ser4) and the Thr moiety at the immunogenic PDTRP region (Thr8). Within this work, we observed that Ser4 is assisted by lectin domain because the mutant GalNAc-T4 D<sub>459</sub>H is not able to glycosylate this glycosylation site. The lectin domain of GalNAc-T4 binds at GalNAc of the previous glycosylated Thr15 of upstream TR to guide the catalytic glycosylation at Ser4.

It was more challenging to discriminate if the glycosylation of Thr8 is to be glycosylated in a lectin assisted manner or is entirely dependent of the catalytic domain. However, based on our results seems that the glycosylation of Thr8 is lectin assisted, using the *O*-GalNAc at Ser14 of the upstream TR. This suggestion is based considering that the glycosylation of Thr8 is affected in the case of the mutant GalNAc-T4 D<sub>459</sub>H, but still capable to glycosylate this position (Thr8) in the tri-glycosylated MUC1 (Thr3, Ser14, Thr15). Though, when the substrate is a di-glycosylated MUC1 (Thr3, Thr15), the mutant D<sub>459</sub>H loses completely the capacity to glycosylate Thr8. We also observed, that the effect of the glycosylation Ser4 and Ser14 in the conformation of MUC1 is minimal, and these glycosylations do not affect the APDTR region conformation at all. Leaving only the hypotheses that Thr8 is lectin assisted and is necessary a concomitant glycosylation at Thr3, Ser4, Ser14 and Thr15.

GalNAc-T4 is also capable to glycosylate Ser14 with assistance of the lectin domain. GalNAc-T4 glycosylates Ser14 likely using GalNAc moieties located at Thr3 or Ser4 of the same TR domain. Thus, the lectin domain of GalNAc-T4 binds contiguous GalNAc moieties.





### 2.4.3 Supporting information

*Table S1 – Carbohydrate-peptide NOEs of the MUC1-4TR-T3T15. The \* labeling in the amino acid indicates glycosylation.*

H8 GalNAc-T3*	H $\alpha$ S4	Medium
H8 GalNAc-T3*	H $\alpha$ T3*	Medium
H8 GalNAc-T3*	H $\alpha$ A5	Medium
H8 GalNAc-T3*	H2 GalNAc-T3*	Weak
H8 GalNAc-T3*	H $\beta$ A5	Strong
NH GalNAc-T3*	NH T3*	Strong
H5 GalNAc-T15*	H $\gamma$ T15*	Strong
H8 GalNAc-T15*	H $\alpha$ T15*	Medium
H8 GalNAc-T15*	H $\alpha$ S14	Medium
H3 GalNAc-T15*	H $\gamma$ T15*	Weak
H6 GalNAc-T15*	H $\beta$ A16	Strong
H8 GalNAc-T15*	H $\beta$ A16	Weak
H2 GalNAc-T15*	H $\beta$ A16	Medium
H6 GalNAc-T15*	H $\gamma$ T15*	Medium
H2 GalNAc-T15*	H $\beta$ T15*	Medium
NH GalNAc-T15*	NH T15	Strong

*Table S2 – Carbohydrate-peptide NOEs of the MUC1-4TR-T3S14T15. The \* labeling in the amino acid indicates glycosylation.*

H8 GalNAc-T3*	H $\alpha$ T3*	Medium
NH GalNAc-T3*	H $\alpha$ S4	Weak
NH GalNAc-T3*	H $\beta$ T3*	Weak
H2 GalNAc-T3*	H $\beta$ T3*	Weak
H8 GalNAc-T3*	H $\alpha$ S4	Medium
H8 GalNAc-T3*	H $\beta$ A5	Medium
NH GalNAc-T3*	NH T3	Strong
H5 GalNAc-S14*	H $\beta$ T15	Strong
H8 GalNAc-S14*	H $\alpha$ S14*	Weak
H8 GalNAc-T15*	H $\beta$ A16	Medium
H2 GalNAc-T15*	H $\beta$ T15*	Medium
H6 GalNAc-T15*	H $\beta$ A16	Strong
H8 GalNAc-T15*	H $\alpha$ T15*	Medium
H5 GalNAc-T15*	H $\gamma$ T15*	Strong
H3 GalNAc-T15*	H $\gamma$ T15*	Medium
H6 GalNAc-T15*	H $\gamma$ T15*	Medium
NH GalNAc-T15*	NH T15	Strong

**Table S3** – Carbohydrate-peptide NOEs of the MUC1-4TR-T3S14T15. The \* labeling in the amino acid indicates glycosylation.

H8 GalNAc-T3*	H $\alpha$ A5	Medium
H3 GalNAc-T3*	H $\gamma$ V2	Medium
NH GalNAc-T3*	H $\beta$ T3*	Medium
NH GalNAc-T3*	NH T3	Strong
H8 GalNAc-T3*	H $\alpha$ A5	Medium
H8 GalNAc-T3*	H $\alpha$ S4	Medium
H8 GalNAc-S4*	NH S4*	Medium
H2 GalNAc-T8*	H $\gamma$ T8*	Medium
H4 GalNAc-T8*	H $\gamma$ T8*	Medium
H5 GalNAc-T8*	H $\gamma$ T8*	Strong
H8 GalNAc-T8*	H $\gamma$ T8*	Strong
NH GalNAc-T8*	NH T8*	Medium
NH GalNAc-S14*	H $\beta$ T15*	Medium
H8 GalNAc-S14*	H $\alpha$ T15*	Weak
NH GalNAc-T15*	H $\alpha$ A16	Weak
H8 GalNAc-T15*	H $\beta$ A16	Medium
NH GalNAc-T15*	NH T15	Strong
H8 GalNAc-T15*	H $\alpha$ T15*	Medium
H3 GalNAc-T15*	H $\beta$ A16	Strong
H5 GalNAc-T15*	H $\gamma$ T15*	Strong
H3 GalNAc-T15*	H $\gamma$ T15*	Medium

**Table S4** -  $^1\text{H-NMR}$  assignments of the  $^{15}\text{N-MUC1-4TR}$ .

<b>Residue</b>	<b>Atom</b>	<b>ppm</b>				
G1	N	110.42			QG	1.98
	QA	3.97	A11	QB	1.37	
	H	8.52		N	126.24	
				HA	4.58	
		H		8.63		
V2	QG2	0.95	P12	HB	4.24	
	N	119.76		HA	4.41	
	HB	2.13		QB	2.30	
	HA	4.25		QG	2.00	
	H	8.26		HD3	3.81	
T3	QG2	1.20		HD2	3.66	
	N	118.71	G13	N	110.04	
	HA	4.44		QA	3.99	
	H	8.46		H	8.67	
S4	N	118.97	S14	N	115.56	
	QB	3.85		QB	3.91	
	HA	4.45		HA	4.53	
	H	8.45		H	8.26	
A5	QB	1.37	T15	QG2	1.21	
	N	127.51		N	116.14	
	HA	4.62		HB	4.23	
	H	8.51		HA	4.38	
P6	HA	4.42		H	8.35	
	QB	2.32	A16	QB	1.35	
	QG	1.99		N	128.52	
	HD2	3.66		HA	4.59	
	HD3	3.82		H	8.42	
		P17		HB	4.24	
D7	N		120.77	HA	4.42	
	HB3		2.68	QG	2.03	
	HB2		2.75	QB	2.22	
	HA		4.64	HD3	3.83	
	H	8.59	HD2	3.65		
T8	QG2	1.19	P18	HA	4.41	
	N	115.09		QG	2.03	
	HA	4.31		QB	2.27	
	H	8.21		HD2	3.65	
R9	QG	1.71		HD3	3.82	
	QB	1.80	A19	QB	1.38	
	N	125.02		N	124.77	
	QD	3.23		HA	4.59	
	HA	4.60	H	8.55		
	H	8.38	H20	N	117.66	
HA	4.40	QB		3.22		
HD3	3.82	HA		4.68		
P10	HD2	3.61				
	QB	2.28				

*Deciphering GalNAc O-glycosylation*  
*O-glycosylation of mucin MUC1 by GalNAc-Ts*

---

	H	8.55
H20 C-term	N	122.17
	H	8.06
	QB	3.16
	HA	4.59

**Figure S5** -  $^1\text{H}$ -NMR assignments of the  $^{15}\text{N}$ -MUC1-4TR-T3T15. The \* labeling in the amino acid indicates glycosylation.

Residue	Atom	ppm
G1	H	8.45
	N	110.33
	QA	3.97
V2	QG2	0.99
	HA	4.34
	HB	2.12
	H	8.21
T3*	N	120.36
	H	8.80
	N	117.73
	QG2	1.28
	HA	4.68
	HB	4.33
	H3-GalNAc	3.99
	H2-GalNAc	4.20
	H5-GalNAc	4.10
	H6-GalNAc	3.67
	H1-GalNAc	4.89
H4-GalNAc	4.03	
H8-GalNAc	2.03	
NH-GalNAc	7.65	
S4	H	8.61
	N	116.46
	QB	3.85
	HA	4.47
A5	H	8.62
	N	126.47
	QB	1.41
P6	HA	4.53
	HG2	2.03
	HD2	3.84
	HA	4.42
	HD3	3.65
	QB	2.30
D7	HG3	1.90
	HB2	2.68
	HA	4.64
	HB3	2.76
	H	8.60
T8	N	120.93
	H	8.20
	N	115.08
	HA	4.29
R9	QG2	1.22
	H	8.39
	N	124.96
	HB3	1.72
	HA	4.58
	HD3	3.21
	HB2	1.82
	QG	1.36
P10	HG2	2.05
	HA	4.39
	QB	2.28
	HD2	3.83
	HD3	3.67
	HG3	1.83
A11	H	8.63
	N	126.32
	HA	4.58
	QB	1.39
P12	HG2	2.04
	HA	4.41
	QB	2.29
	HD2	3.81
G13	HG3	1.92
	H	8.67
	N	109.64
S14	QA	4.00
	H	8.25
	N	115.99
	QB	3.96
T15*	HA	4.68
	H	8.76
	N	114.41
	QG2	1.27
	HA	4.61
	HB	4.36
	H3-GalNAc	3.89
	H2-GalNAc	4.09
	H5-GalNAc	3.98
	H8-GalNAc	2.05
H1-GalNAc	4.93	
A16	H4-GalNAc	3.94
	H6-GalNAc	3.89
	NH-GalNAc	7.88
	H	8.46
P17	N	125.37
	QB	1.34
	HA	4.57
	HG2	2.02
	HA	4.41
	QB	2.32

*Deciphering GalNAc O-glycosylation  
O-glycosylation of mucin MUC1 by GalNAc-Ts*

---

	HG3	1.92
P18	HG2	1.92
	HA	4.41
	QB	2.33
	HG3	1.92
A19	H	8.54
	N	124.71
	QB	1.34
	HA	4.58
H20	H	8.39
	N	118.14
	QB	3.19
	HA	4.64
H20 C-term	HB3	3.12
	HA	4.45
	HB2	3.21
	H	8.02
	N	122.31

**Figure S6** -  $^1\text{H}$ -NMR assignments of the  $^{15}\text{N}$ -MUC1-4TR-T3S14T15. The \* labeling in the amino acid indicates glycosylation and # labeling indicates the amino acid perturbed with the neighboring glycosylation.

Residue	Proton	ppm
G1	H	8.443
	N	110.311
	QA	3.965
V2	H	8.195
	N	120.683
	HA	4.323
	HB	2.107
	QG2	0.998
T3*	H	8.796
	N	117.735
	HA	4.688
	HB	4.320
	QG2	1.252
	H2-GalNAc	4.090
	H1-GalNAc	4.977
	H4-GalNAc	3.969
	H3-GalNAc	3.915
	HN-GalNAc	7.652
	H8-GalNAc	2.044
H6-GalNAc	3.782	
H5-GalNAc	4.024	
S4	H	8.610
	N	116.534
	HA	4.492
	HB3	3.871
	HB2	3.805
A5	H	8.620
	N	126.387
	HA	4.550
P6	QB	1.413
	HA	4.432
	QB	2.317
	HD3	3.840
	HD2	3.680
	HG3	2.054
	HG2	1.974
D7	H	8.596
	N	120.901
	HA	4.636
	HB3	2.747
	HB2	2.675
T8	H	8.200
	N	115.115
	QG2	1.202
R9	HA	4.288
	H	8.386
	N	125.077
	HA	4.58
	HD3	3.255
	QB	1.818
P10	QG	1.709
	HD2	3.207
	HA	4.410
	HG3	2.034
	HG2	1.901
	QB	2.302
A11	HD3	3.845
	HD2	3.662
	H	8.577
	N	126.278
P12	HA	4.571
	QB	1.401
	HA	4.40
	HG3	2.040
	HG2	1.821
G13	QB	2.278
	HD3	3.839
	HD2	3.630
	H	8.534
S14*	N	109.165
	HA3	4.064
	HA2	3.977
	H	8.505
	N	116.234
	HA	4.824
	HB3	4.073
	HB2	3.877
	H2-GalNAc	4.134
	H1-GalNAc	4.884
	HN-GalNAc	7.841
H8-GalNAc	2.034	
H3-GalNAc	3.895	
H6-GalNAc	3.762	
H5-GalNAc	4.024	
H4-GalNAc	3.974	
T15*	H	8.992
	N	113.750
	HA	4.611
	HB	4.295

*Deciphering GalNAc O-glycosylation  
O-glycosylation of mucin MUC1 by GalNAc-Ts*

	QG2	1.252
	HN-GalNAc	7.787
	H2-GalNAc	4.102
	H8-GalNAc	2.032
	H3-GalNAc	3.868
	H6-GalNAc	3.727
	H5-GalNAc	4.006
	H4-GalNAc	3.977
	H1-GalNAc	5.008
A16	QB	1.383
	HA	4.582
	H	8.543
	N	124.368
P17	HA	4.407
	HG3	2.038
	HG2	1.950
	QB	2.293
	HD3	3.824
	HD2	3.680
P18	QB	2.358
	HD3	3.824
	HD2	3.632
	HG3	2.047
	HG2	1.921
	HA	4.420
A19	H	8.524
	N	124.641
	QB	1.356
	HA	4.557
H20	H	8.353
	N	118.226
	QB	3.156
	HA	4.621
H20 C-Term	H	8.019
	N	122.703
	HA	4.463
	HB3	3.230
	HB2	3.106



**Figure S7** -  $^1\text{H-NMR}$  assignments of the  $^{15}\text{N-MUC1-4TR-T3S4T8S14T15}$ . The \* labeling in the amino acid indicates glycosylation and # labeling indicates the amino acid perturbed with the neighboring glycosylation.

Residue	Proton	ppm
G1	H	8.472
	N	110.209
	QA	3.966
V2	H	8.206
	N	120.637
	HA	4.333
	HB	2.112
	QG1	0.992
T3*	H	8.876
	N	117.646
	HA	4.683
	HB	4.351
	QG2	1.324
	HN-GalNAc	7.707
	H4	3.894
	H1	4.936
	H5	4.089
	H3	3.977
	H2	4.215
	H6	3.672
H8	2.049	
T3* N-term	H	8.798
	N	117.680
	HA	4.683
	QG2	1.294
	HB	4.313
	HN-GalNAc	7.657
	H4-GalNAc	3.903
	H5-GalNAc	4.105
	H3-GalNAc	3.980
	H2-GalNAc	4.203
	H6-GalNAc	3.670
	H8-GalNAc	2.043
S4*	H	8.629
	N	116.779
	HA	4.508
	H	8.798
	N	116.258
	H3-GalNAc	4.114
	H2-GalNAc	4.214
	H8-GalNAc	2.077
	HN-GalNAc	7.846
	H1-GalNAc	4.882
	H6-GalNAc	3.641
	S4	HA
HB3		3.931
HB2		3.791
A5	H	8.671
	N	126.794
	HA	4.508
	QB	1.429
P6	HA	4.414
	HG2	1.889
	HG3	2.040
	QB	2.298
D7#	QD	3.839
	H	8.608
	N	121.842
	HA	4.753
D7	HB3	2.811
	HB2	2.654
	H	8.594
	N	121.079
T8*	HA	4.648
	HB3	2.794
	HB2	2.689
	H	8.756
	N	114.351
	HA	4.543
	HB	3.756
	H2-GalNAc	4.170
	H4-GalNAc	3.942
	H3-GalNAc	3.984
	H6-GalNAc	3.670
	H5-GalNAc	4.090
HN-GalNAc	7.885	
H1-GalNAc	4.850	
H8-GalNAc	2.043	
T8	H	8.220
	N	115.218
	HA	4.333
	QG2	1.202
R9#	H	8.446
	N	121.946
	HA	4.526
	QD	3.245
	HB2	1.863
	QG	1.418
HB3	1.746	

*Deciphering GalNAc O-glycosylation  
O-glycosylation of mucin MUC1 by GalNAc-Ts*

R9	H	8.414
	N	125.079
	HA	4.613
	QD	3.284
	QB	1.814
	QG	1.381
P10	HA	4.385
	HG2	1.889
	HG3	2.030
	QB	2.285
	QD	3.846
A11	H	8.606
	N	126.471
	HA	4.543
	QB	1.394
P12	HA	4.401
	HG2	1.840
	HG3	2.038
	QB	2.278
	QD	3.853
G13	H	8.532
	N	109.155
	HA3	4.071
	HA2	3.948
S14*	H	8.508
	N	116.270
	HB3	4.071
	HB2	3.913
	HA	4.858
	HN-GalNAc	7.784
	H4-GalNAc	3.870
	H1-GalNAc	4.834
	H5-GalNAc	4.099
	H3-GalNAc	3.993
	H2-GalNAc	4.206
H6-GalNAc	3.688	
H8-GalNAc	2.032	
T15*	H	9.005
	N	113.673
	HB	4.313
	QG2	1.259
	HA	4.646
	HN-GalNAc	7.836
	H4-GalNAc	3.898
	H1-GalNAc	4.891
	H3-GalNAc	4.109
	H2-GalNAc	4.195
	H5-GalNAc	3.989
H6-GalNAc	3.679	
H8-GalNAc	2.072	

A16	H	8.545
	N	124.478
	QB	1.342
	HA	4.549
A19	H	8.532
	N	124.655
	QB	1.352
P17	HA	4.556
	HA	4.443
	HG3	2.051
	HG2	1.955
P18	QB	2.313
	QD	3.839
	HA	4.403
	HG3	2.042
H20	HG2	1.895
	QB	2.301
	QD	3.828
	H	8.432
H20 C-term	N	117.979
	HA	4.674
	HB3	3.221
	HB2	3.170
	H	8.061
H20 C-term	N	122.519
	HA	4.473
	HB3	3.231
	HB2	3.126

## **2.5 Methods**

### **2.5.1 NMR experiments**

#### **2.5.1.1 *Assignment of (glyco)peptide***

All NMR experiments were recorded on a Bruker Avance 600 MHz spectrometer equipped with a triple channel cryoprobe head. The  $^1\text{H}$  NMR resonances of the peptides were completely assigned through standard 2D-TOCSY (30 and 80 ms mixing time) and 2D-NOESY experiments (400 ms mixing time). Solution conditions used for the NMR characterization studies were 1-3 mM (glyco)peptide, 25 mM perdeuterated Tris- $\text{d}_{11}$  in 90:10  $\text{H}_2\text{O}/\text{D}_2\text{O}$ , 7.5mM NaCl and 1mM DTT, uncorrected pH 7.4. The assignments were accomplished either at 278K or 298K. The resonance of 2,2,3,3-tetradeutero-3-trimethylsilylpropionic acid (TSP) was used as a chemical shift reference in the  $^1\text{H}$  NMR experiments ( $\delta$  TSP = 0 ppm). Peak lists for the 2D-TOCSY and 2D-NOESY spectra were generated by interactive peak picking using the Computer Aided Resonance Assignment (CARA) software<sup>78</sup>.

#### **2.5.1.2 *Saturation Transfer Difference (STD)***

Samples for STD-experiments were prepared in perdeuterated 25 mM Tris- $\text{d}_{11}$  in deuterated water, 7.5 mM NaCl and 1 mM DTT, uncorrected pH 7.4. STD-NMR experiments were performed at 298 K in the presence of UDP,  $\text{MnCl}_2$  with peptide (or GalNAc-O-Me) and GalNAc-Ts, the concentrations of each intervenient are described in the legend of the respective experiment.

The STD-NMR spectra were acquired on a Bruker Avance 600 MHz spectrometer equipped with a triple channel cryoprobe head with 1920 transients in a matrix with 64k data points in  $t_2$  in a spectral window of 12335.53 Hz centered at 2819.65 Hz. An excitation sculpting module was employed to suppress the water

proton signals. Selective saturation of the protein resonances (on resonance spectrum) was performed by irradiating at  $-1$  ppm using a series of Eburp2.1000-shaped  $90^\circ$  pulses (50 ms, 1 ms delay between pulses) for a total saturation time of 2.0 s. For the reference spectrum (off resonance), the samples were irradiated at 100 ppm. Proper control experiments were performed with the ligands in the presence and absence of the protein to optimize the frequency for protein saturation ( $-1$  ppm) and to ensure that the ligand signals were not affected. However, all glycopeptides, when irradiated at  $-1$  ppm in the absence of protein, showed residual saturation on the aliphatic methyl groups in the STD-NMR spectra. This nonspecific saturation was considered, by subtraction, when quantifying the STD-NMR data in the presence of transferase. A blank STD experiment with only the protein was also recorded. The subtraction of this protein STD spectrum allowed eliminating the signal background of the protein. The STD-NMR total intensities were normalized with respect to the highest STD-NMR response. The STD response of each amino acid corresponds to the average of STD percentages of all amino acid proton resonances that were measured with sufficient accuracy. The signal of the anomeric proton as well as the  $H_\alpha$  protons of the Ala aminoacids of glycopeptides could not be analyzed in the STD-NMR spectra due to their close proximity to the HDO resonance. Some resonances from Gly and Pro overlapped and could not be discriminated.

### **2.5.1.3 $^{19}\text{F}$ -NMR experiments**

$^{19}\text{F}$ -NMR experiments were carried out to monitor the  $^{19}\text{F}$  reporter  $(\text{CF}_3)_2\text{COCH}_2$ - selectively introduced at the SH- group of the mutants T<sub>375</sub>C from both constructs of the WT GalNAc-T2 and the F<sub>104</sub>S mutant. The  $^{19}\text{F}$  experiments of GalNAcT2- $^{19}\text{F}$  and F<sub>104</sub>S- $^{19}\text{F}$  constructs were recorded at distinct conditions in absence and presence of  $\text{MnCl}_2$ , UDP or UDP-GalNAc and peptide. The UDP or UDP-GalNAc and  $\text{MnCl}_2$  were added in an excess of 1.5 times of the protein concentration while peptide was in a 1.1:1 peptide:protein molar ratio. Higher concentrations of UDP or UDP-GalNAc (500  $\mu\text{M}$ ),  $\text{MnCl}_2$  (500 $\mu\text{M}$ ) and peptide

(protein:peptide molar ratio 1:2.2) did not induce higher  $^{19}\text{F}$  chemical shift perturbations. The  $^{19}\text{F}$ -NMR experiments were recorded on a Bruker Avance III 600 MHz spectrometer equipped with a  $^{19}\text{F},^1\text{H}$  SEF dual probe optimized for direct  $^{19}\text{F}$  detection. The  $^{19}\text{F}$ -NMR experiments were acquired at 298 K using a acquisition of 640 scans in a matrix with 8k data points in a spectral window of 11295.2 Hz centered at  $(\text{CF}_3)_2\text{CHOH}$  -45174.9 Hz. The signal of 1,1,1,3,3,3-hexafluoro-2-propanol was used as a chemical shift reference ( $\delta = -75.7$  ppm). Samples for  $^{19}\text{F}$ -NMR were prepared in 25 mM Tris-d11 in 90:10  $\text{H}_2\text{O}/\text{D}_2\text{O}$ , uncorrected pH 7.5.

Trace elements of the 3-Bromo-1,1,1-trifluoroacetone reactive used for protein labelling ( $\delta \text{CF}_3\text{COCH}_2\text{Br} = -83.3$  ppm) were also present. In addition, the trifluoroacetic acid (TFA) used in the synthesis of peptides can be seen in  $^{19}\text{F}$ -NMR spectra in the presence of the peptide ( $\delta \text{TFA} = -75.4$  ppm).

In addition, to monitor the glycosylation event of MUC5Ac in presence of GalNAcT2- $^{19}\text{F}$  and F<sub>104</sub>S- $^{19}\text{F}$ , two  $^1\text{H}$ -NMR spectra were recorded before and after addition of UDP-GalNAc. All conditions contained  $\text{MnCl}_2$ .

#### **2.5.1.4 NMR Glycosylation Assay**

Glycosylation of  $^{15}\text{N}$ -MUC1-4TR by recombinant GalNAc-Ts was monitored by collecting a series of  $^1\text{H}/^{15}\text{N}$ -HSQCs experiments using a 600-MHz Bruker Avance III spectrometer equipped with a 5mm inverse detection triple-resonance z-gradient cryogenic probe. All NMR experiments were performed at 298 K. The spectra were acquired with  $2048 \times 128$  points and 4 scans. The spectral widths were 9615.4 Hz for  $^1\text{H}$  and 1946.2 Hz for  $^{15}\text{N}$ . The central frequency for  $^1\text{H}$  was set on the solvent signal and on the center of the amide region for  $^{15}\text{N}$ . The time between each experiment was approximately 10 min. Spectra were processed with the Bruker TopSpin 3.5 software. The volumes of the peaks in the  $^1\text{H}/^{15}\text{N}$ -HSQC spectra corresponding to the glycosylated and non-glycosylated forms of MUC1-

4TR were quantified by using the Mnova v.9.0 software. The ratio of the peak volumes for glycosylated residues to the total volume of peaks for both glycosylated and non-glycosylated forms was plotted versus time or sites.

All samples were prepared in 200  $\mu$ L buffer (25 mM Tris-d<sub>11</sub>, 7.5 mM NaCl, 1 mM DTT, 150  $\mu$ M MnCl<sub>2</sub>, in H<sub>2</sub>O:D<sub>2</sub>O 90:10) at pH 6.3. The relationships between <sup>15</sup>N-MUC1-4TR and GalNAc-Ts were kept between WT enzymes and the corresponding mutants. Depending on the experiments, different amounts of UDP-GalNAc were used. For the experiments where an excess of UDP-GalNAc was added, the ratio between UDP-GalNAc and each glycosylation site was set to 40/1 (UDP-GalNAc/MUC1-4TR). In the step-by-step experiments, controlled amounts of UDP-GalNAc were added to glycosylate 4, 8, 10, 14 and 20 sites of MUC1-4TR using the required amount of enzyme to glycosylate all accessible acceptor sites. In most of the cases, 24 h were enough to achieve the exhaustive glycosylation of MUC1-4TR. The “endpoint” products of MUC1-4TR after glycosylation by GalNAc-T2/T3 and GalNAc-T4 were purified using HPLC as described below. Glycosylated MUC1-4TR, with GalNAc attached only to T3 and T15 sites, and MUC1-4TR-T3T15 were obtained using the D<sub>458</sub>R mutant and stopping the GalNAc transfer by adding EDTA (0.5 mM).

#### ***2.5.1.5 Percentage of glycosylation of individual residues***

These values were determined by the calculating the peak volumes for glycosylated and non-glycosylated residues in <sup>1</sup>H/<sup>15</sup>N-HSQC spectra. Relative glycosylation values were defined as the ratio of the volume of the glycosylated peak to the total volume of glycosylated and non-glycosylated peaks.

### 2.5.1.6 Combined chemical shift perturbation (CSP)

The  $\Delta\delta_{comb}$  was calculated for each *backbone NH groups of MUC1-4TR* using the equation 2.5.1. A column plot was created with the  $\Delta\delta_{comb}$  values. Once again, the cut-off was calculated to identify the amino acids with higher perturbation after glycosylation.

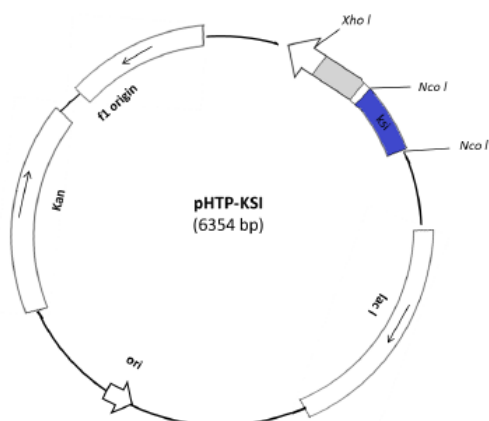
$$\Delta\delta_{comb} = \sqrt{((\Delta\delta H)^2 + (\frac{\omega H}{\omega N} \Delta\delta N)^2)} \quad \text{Equation 2.5.1}$$

where  $\Delta\delta H$  is the  $^1\text{H}$  chemical shift difference for the signal with and without glycosylation,  $\Delta\delta N$  is the  $^{15}\text{N}$  chemical shift difference for the signal with and without glycosylation, and the  $\omega H/\omega N$  is the quotient between the gyromagnetic constant of the proton and the gyromagnetic constant of the nitrogen atom.

With the values of  $\Delta\delta_{comb}$  obtained for each amino acid after glycosylation, a column plot was performed.

### 2.5.2 Overexpression and Purification of $^{15}\text{N}$ -MUC1-4TR

A construct containing 4 repeats of 20 amino acids of the conserved sequence GVTSAPDTRPAPGSTAPPAH (MUC1-4TR) of the N-terminal domain of MUC1 was synthesized by NZYtech. MUC1-4TR was sub-cloned in the expression vector pHTP-KSI, produced by NZYTech, to yield MUC1-4TR construct (Figure 2.5.1). The MUC1-4TR construct concomitantly contains a fusion protein KSI tag, a His tag and a TEV protease recognition sequence (Glu-Asn-Leu-Tyr-Phe-Gln↓Gly). The arrow indicates the cut site of TEV enzyme.



**Sequence:**

**MGHTPEHITAVVQRFVAALNAGDLGDGIVALFADDATVEDPVGSEPRSGTAAI  
REFYANSLKLPLAVELTQEVRAVANEAAFAFTVSFEYQGRKTVVAPIDHFRF  
NGAGKVV SIRALFGEKNIHACQAMGSSHHHHHSSGPQQLRENL~~YFQ~~GV~~T~~  
SAPDTRPAPGSTAPPAHGVT SAPDTRPAPGSTAPPAHGVT SAPDTRPAPGSTA  
PPAHGVT SAPDTRPAPGSTAPPAH**

**Figure 2.5.1** - *pHTP-KSI* expression vector and *KSI-MUC1-4TR* sequence. This sequence consists of *KSI*-tag (blue), a *His*-tag (green), a *TEV* site (*Tobacco Etch Virus*) (red) and the amino acid sequence of *MUC1-4TR* (underlined).

One single colony of the transformed *E. coli* cells (BL21 DE3) was inoculated and incubated in 10 mL of sterile LB medium with 50 µg/mL of kanamycin and left to grow overnight at 37 °C at 220 rpm (pre-inoculum). The culture of pre-inoculum (10 mL) was then inoculated in 2 L of M9 minimal media with 50 µg/mL of kanamycin and let to grow. Protein expression was induced at  $OD_{600nm} = 0.6$  with a final concentration of 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), overnight at 25°C. To harvest cell was centrifuged at 6000 rpm for 15 minutes, at 4 °C in an Avanti J-26 XPI Beckman Coulter with the rotor JA-10. The harvested cells of 1 L of culture were re-suspended (40 mL) in 10 mM phosphate buffer (pH 7.4) with 150 mM of NaCl and 1 mM of β-mercaptoethanol, ruptured by sonication with 10 cycles of 1 min with 80% of



amplitude, in a Hielscher Ultrasound Technology (UP100H). After sonication, the cell lysate was centrifuged at 10000 rpm for 15 minutes, the pellet was recovered and solubilized in 50 mL of solubilization buffer (phosphate buffer (10 mM), 150 mM NaCl, 8M Urea,  $\beta$ -mercaptoethanol (1 mM) (pH=7.4)) overnight at room temperature with agitation. Before the purification, the solution was centrifuged at 10000 rpm, at 4 °C for 15 minutes. The supernatant was loaded onto a Ni-NTA-agarose column (QIAGEN) previously equilibrated with the solubilization buffer with 10 mM Imidazole. The resin was washed with a buffer containing 10 mM phosphate at pH 7.4, 150 mM NaCl and 50 mM imidazole, during the wash MUC1-KSI refolds. MUC1-KSI was eluted with a gradient with buffer containing 10 mM phosphate at pH 7.4, 150 mM NaCl and 1 M imidazole (during 70 mL), with the flow 3 mL/min. To remove the imidazole, a desalting step (5 HiTrap™ desalting columns) against phosphate buffer (10 mM) with NaCl 150 mM and  $\beta$ -mercaptoethanol 1 mM was performed.

To obtain MUC1-4TR without KSI (fusion protein) and His-tag, the MUC1-KSI was digested with TEV protease (Tobacco Etch Virus), overnight at room temperature, in the same buffer, after desalting chromatography with 500  $\mu$ M of EDTA (Ethylenediaminetetraacetic acid) and TEV protease (1:100 (w/w)).

Since after digestion, no aromatic amino acids were present, the last step to obtain pure MUC1-4TR employed a reversed-phase chromatography (AKTA Prime Plus GE HealthCare), using a C18 column (Purospher® STAR RP-18 end capped 5 $\mu$ m column (HPLC-Cartridge) with the detector placed at 220 nm, 230 nm and 254 nm. To purify MUC1-4TR were used as buffer (A) water/TFA (99.9:0.1, v/v), and (B) acetonitrile. The column and system were first equilibrated in 90% of buffer A and 10% of buffer B, then the sample was injected, and the elution recorded using a gradient from 10-40% of acetonitrile (B), with a flow of 2 mL/min for 40 minutes. All peaks with absorbance at 220 nm were collected. At the end the column was washed with 100% of acetonitrile.

Pure MUC1-4TR was lyophilized and stored at 4 °C. The purity of the protein was confirmed by <sup>1</sup>H/<sup>15</sup>N-HSQC experiments. The yield of expression/purification

was estimated by  $^1\text{H}$ -NMR analysis and using 2,2,3,3-tetradeutero-3-trimethylsilylpropionic acid (TSP) as a chemical shift reference ( $\delta$  TSP = 0 ppm).

### ***2.5.2.1 Purification of the glycosylated products***

To purify MUC1-4TR glycosylated products, different buffers (A, water/TFA 99.9:0.1, v/v) and (B, acetonitrile) were used. The column and system were first equilibrated in 90% of buffer A and 10% of buffer B (20 min), then the sample was injected, and the elution recorded using a gradient from 10-30% of acetonitrile (B), with a flow of 2 mL/min for 70 minutes. All peaks with absorbance at 220 nm were collected and analyzed by  $^1\text{H}/^{15}\text{N}$ -HSQCs and MALDI-TOF/TOF. At the end, the column was washed with 100% of acetonitrile.

### **2.5.3 Mass spectrometry: MALDI-TOF/TOF**

An Autoflex III MALDI-TOF/TOF spectrometer (Bruker Daltonics) was used in linear mode with the following settings: 5000-40000 Th window, linear positive mode, ion source 1: 20 kV, ion source 2: 18.5 kV, lens: 9 kV, pulsed ion extraction of 120 ns, high gating ion suppression up to 1000 Mr. Mass calibration was performed externally with protein 1 standard calibration mixture (Bruker Daltonics) in the same range as the samples. Data acquisition was performed using FlexControl 3.0 software (Bruker Daltonics), and peak peaking and subsequent spectra analysis was performed using FlexAnalysis 3.0 software (Bruker Daltonics).

#### ***2.5.3.1 Sample preparation***

Dried droplet preparations were spotted following the standard Dried Droplet application method (0.5  $\mu\text{L}$  sample + 0.5  $\mu\text{L}$  matrix (sinapinic acid, 10 mg/ml in

[70:30] Acetonitrile:Trifluoroacetic acid 0.1%)) onto a GroundSteel massive 384 target (Bruker Daltonics). The preparations were desalted using ZipTip® C4 micro-columns (Millipore) (2 µL sample) with elution using 0.5µL SA (sinapinic acid, 10 mg/ml in 70:30 acetonitrile:trifluoroacetic acid 0.1%) matrix onto a GroundSteel massive 384 target (Bruker Daltonics).

## 2.6 References

1. Weinbaum, S., Tarbell, J. M. & Damiano, E. R. The Structure and Function of the Endothelial Glycocalyx Layer. *Annu. Rev. Biomed. Eng.* **9**, 121–167 (2007).
2. Bertozzi, C. R. & Kiessling, L. L. Chemical glycobiology. *Science* **291**, 2357–2364 (2001).
3. Yoshimura, Y. *et al.* Elucidation of the sugar recognition ability of the lectin domain of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase 3 by using unnatural glycopeptide substrates. *Glycobiology* **22**, 429–438 (2012).
4. Ten Hagen, K. G., Fritz, T. A. & Tabak, L. A. All in the family: The UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases. *Glycobiology* **13**, 1–16 (2003).
5. Papanikou, E. & Glick, B. S. Golgi compartmentation and identity. *Curr. Opin. Cell Biol.* **29**, 74–81 (2014).
6. Bennett, E. P. *et al.* Control of mucin-type O-glycosylation: A classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* **22**, 736–756 (2012).
7. Röttger, S. *et al.* Localization of three human polypeptide GalNAc-transferases in HeLa cells suggests initiation of O-linked glycosylation throughout the Golgi apparatus. *J. Cell Sci.* **111**, 45–60 (1998).
8. Stanley, P. Golgi glycosylation. *Cold Spring Harb. Perspect. Biol.* **3**, 1–13 (2011).
9. Gill, D. J., Clausen, H. & Bard, F. Location, location, location: New insights into O-GalNAc protein glycosylation. *Trends Cell Biol.* **21**, 149–158 (2011).
10. White, T. *et al.* Purification and cDNA cloning of a human UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase. *J. Biol. Chem.* **270**, 24156–24165 (1995).
11. Topaz, O. *et al.* Mutations in GALNT3, encoding a protein involved in O-linked glycosylation, cause familial tumoral calcinosis. *Nat. Genet.* **36**, 579–581 (2004).
12. Holleboom, A. G. *et al.* Heterozygosity for a Loss-of-Function Mutation in GALNT2 Improves Plasma Triglyceride Clearance in Man. *Cell Metab.* **14**, 811–818 (2011).
13. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of Coronary Artery Disease. *Nat Genet* **40**, 161–169 (2008).
14. Simmons, A. D. *et al.* A direct interaction between EXT proteins and glycosyltransferases is defective in hereditary multiple exostoses. *Hum. Mol. Genet.* **8**, 2155–2164 (1999).
15. Nakayama, Y. *et al.* A putative polypeptide N-acetylgalactosaminyltransferase/Williams-Beuren syndrome chromosome region 17 (WBSCR17) regulates lamellipodium formation and macropinocytosis. *J. Biol. Chem.* **287**, 32222–32235 (2012).
16. Mandel, U. *et al.* Expression of polypeptide GalNAc-transferases in stratified

- epithelia and squamous cell carcinomas: Immunohistological evaluation using monoclonal antibodies to three members of the GalNAc-transferase family. *Glycobiology* **9**, 43–52 (1999).
17. Gill, D. J. *et al.* Initiation of GalNAc-type O-glycosylation in the endoplasmic reticulum promotes cancer cell invasiveness. *Proc. Natl. Acad. Sci.* **110**, E3152–E3161 (2013).
  18. Cazet, A., Julien, S., Bobowski, M., Burchell, J. & Delannoy, P. Tumour-associated carbohydrate antigens in breast cancer Tumour-associated carbohydrate antigens in breast cancer. *Breast Cancer Res.* **12**: 204 (2010).
  19. Bard, F. & Chia, J. Cracking the Glycome Encoder: Signaling, Trafficking, and Glycosylation. *Trends Cell Biol.* **26**, 379–388 (2016).
  20. Lira-Navarrete, E. *et al.* Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein O-glycosylation. *Nat. Commun.* **6**:6937 (2015).
  21. Revoredo, L. *et al.* Mucin-type o-glycosylation is controlled by short- And long-range glycopeptide substrate recognition that varies among members of the polypeptide GalNAc transferase family. *Glycobiology* **26**, 360–376 (2016).
  22. Fritz, T. A., Raman, J. & Tabak, L. A. Dynamic association between the catalytic and lectin domains of human UDP-GalNAc:polypeptide  $\alpha$ -N-acetylgalactosaminyltransferase-2. *J. Biol. Chem.* **281**, 8613–8619 (2006).
  23. Wandall, H. H. *et al.* The lectin domains of polypeptide GalNAc-transferases exhibit carbohydrate-binding specificity for GalNAc: Lectin binding to GalNAc-glycopeptide substrates is required for high density GalNAc-O-glycosylation. *Glycobiology* **17**, 374–387 (2007).
  24. Tenno, M., Kézdy, F. J., Elhammer, Å. P. & Kurosaka, A. Function of the lectin domain of polypeptide N-acetylgalactosaminyltransferase 1. *Biochem. Biophys. Res. Commun.* **298**, 755–759 (2002).
  25. Fritz, T. A., Hurley, J. H., Trinh, L.-B., Shiloach, J. & Tabak, L. A. The beginnings of mucin biosynthesis: The crystal structure of UDP-GalNAc:polypeptide -N-acetylgalactosaminyltransferase-T1. *Proc. Natl. Acad. Sci.* **101**, 15307–15312 (2004).
  26. Pedersen, J. W. *et al.* Lectin domains of polypeptide GalNAc transferases exhibit glycopeptide binding specificity. *J. Biol. Chem.* **286**, 32684–32696 (2011).
  27. Hassan, H. *et al.* The lectin domain of UDP-N-acetyl-D-galactosamine: Polypeptide N-acetylgalactosaminyltransferase-T4 directs its glycopeptide specificities. *J. Biol. Chem.* **275**, 38197–38205 (2000).
  28. Hagen, F. K., Hazes, B., Raffo, R., DeSa, D. & Tabak, L. a. Structure-Function Analysis of the UDP-N-acetyl-D-galactosamine: Polypeptide N-acetylgalactosaminyltransferase. *J. Bol. Chem.* **274**, 6797–6803 (1999).
  29. Lira-Navarrete, E. *et al.* Substrate-guided front-face reaction revealed by combined structural snapshots and metadynamics for the polypeptide N-

- acetylgalactosaminyltransferase 2. *Angew. Chemie - Int. Ed.* **53**, 8206–8210 (2014).
30. Chefetz, I. *et al.* GALNT3, a gene associated with hyperphosphatemic familial tumoral calcinosis, is transcriptionally regulated by extracellular phosphate and modulates matrix metalloproteinase activity. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1792**, 61–67 (2009).
31. Kato, K. *et al.* Polypeptide GalNAc-transferase T3 and familial tumoral calcinosis: Secretion of fibroblast growth factor 23 requires O-glycosylation. *J. Biol. Chem.* **281**, 18370–18377 (2006).
32. Schjoldager, K. T., Christoffersen, C., Leguern, E., Clausen, H. & Rader, D. J. Clinical and Translational Report Loss of Function of GALNT2 Lowers High-Density Lipoproteins in Humans , Nonhuman Primates , and Clinical and Translational Report Loss of Function of GALNT2 Lowers High-Density Lipoproteins in Humans , Nonhuman Primates ., *Cell Metab.* **24**, 234–245 (2016).
33. Bennett, E. P. *et al.* Cloning of a human UDP-N-acetyl-alpha-D-galactosamine : polypeptide N-acetylgalactosaminyltransferase that complements other GalNAc-transferases in complete O-glycosylation of the MUC1 tandem repeat. *J. Biol. Chem.* **273**, 30472–30481 (1998).
34. Wandall, H. H. *et al.* Substrate specificities of three members of the human UDP-N-acetyl-alpha-D-galactosamine:Polypeptide N-acetylgalactosaminyltransferase family, GalNAc-T1, -T2, and -T3. *J. Biol. Chem.* **272**, 23503–23514 (1997).
35. Gerken, T. A. *et al.* Emerging paradigms for the initiation of mucin-type protein O-glycosylation by the polypeptide GalNAc transferase family of glycosyltransferases. *J. Biol. Chem.* **286**, 14493–14507 (2011).
36. Kong, Y. *et al.* Probing polypeptide GalNAc-transferase isoform substrate specificities by in vitro analysis. *Glycobiology* **25**, 55–65 (2015).
37. Gerken, T. A. *et al.* The lectin domain of the polypeptide GalNAc transferase family of glycosyltransferases (ppGalNAc Ts) acts as a switch directing glycopeptide substrate glycosylation in an N- or C-terminal direction, further controlling mucin type O-Glycosylation. *J. Biol. Chem.* **288**, 19900–19914 (2013).
38. Raman, J. *et al.* The catalytic and lectin domains of UDP-GalNAc:polypeptide  $\alpha$ -N-acetylgalactosaminyltransferase function in concert to direct glycosylation site selection. *J. Biol. Chem.* **283**, 22942–22951 (2008).
39. Hurtado-Guerrero, R. Recent structural and mechanistic insights into post-translational enzymatic glycosylation. *Biochem. Soc. Trans.* **44**, 61–67 (2016).
40. Ghirardello, M. *et al.* Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-Polar Conjugates. *Chem. - A Eur. J.* **22**, 7215–7224 (2016).
41. Rydzik, A. M. *et al.* Monitoring conformational changes in the NDM-1 metallo- $\beta$ -lactamase by 19F NMR spectroscopy. *Angew. Chemie - Int. Ed.* **53**, 3129–3133 (2014).
42. De Las Rivas, M. *et al.* Structural analysis of a GalNAc-T2 mutant reveals an

- induced-fit catalytic mechanism for GalNAc-Ts. *Chem. Eur. J.* **24**, 8382-8392 (2018).
43. Aramini, J. M. *et al.* 19F NMR Reveals multiple conformations at the dimer interface of the nonstructural protein 1 effector domain from influenza A virus. *Structure* **22**, 515–525 (2014).
44. Hollingsworth, M. A. & Swanson, B. J. Mucins in cancer: protection and control of the cell surface. *Nat. Rev. Cancer* **4**, 45–60 (2004).
45. Finn, O. J. *et al.* MUC-1 epithelial tumor mucin-based immunity and cancer vaccines. *Immunol. Rev.* **145**, 61–89 (1995).
46. Bergstrom, K. S. B. & Xia, L. Mucin-type O-glycans and their roles in intestinal homeostasis. *Glycobiology* **23**, 1026–1037 (2013).
47. Song, W. *et al.* MUC1 glycopeptide epitopes predicted by computational glycomics. *Int. J. Oncol.* **41**, 1977–1984 (2012).
48. Nath, S. & Mukherjee, P. MUC1: A multifaceted oncoprotein with a key role in cancer progression. *Trends Mol. Med.* **20**, 332–342 (2014).
49. Bafna, S., Kaur, S. & Batra, S. K. Membrane-bound mucins: the mechanistic basis for alterations in the growth and survival of cancer cells. *Oncogene* **29**, 2893–2904 (2010).
50. Lakshmanan, I. *et al.* Mucins in lung cancer: Diagnostic, prognostic, and therapeutic implications. *J. Thorac. Oncol.* **10**, 19–27 (2015).
51. Madariaga, D. *et al.* Detection of tumor-associated glycopeptides by Lectins: The peptide context modulates carbohydrate recognition. *ACS Chem. Biol.* **10**, 747–756 (2015).
52. Duraisamy, S., Kufe, T., Ramasamy, S. & Kufe, D. Evolution of the human MUC1 oncoprotein. *Int. J. Oncol.* **31**, 671–677 (2007).
53. Hollingsworth, M. A. & Swanson, B. J. Mucins in cancer: protection and control of the cell surface. *Nat. Rev. Cancer* **4**, 45–60 (2004).
54. Springer, G. T and Tn, general carcinoma autoantigens. *Science (80-. )*. **224**, 1198–1206 (1984).
55. Julien, S., Videira, P. A. & Delannoy, P. Sialyl-tn in cancer: (how) did we miss the target? *Biomolecules* **2**, 435–66 (2012).
56. Saitoh, O., Gallagher, R. E. & Fukuda, M. Expression of Aberrant O-Glycans Attached to Leukosialin in Differentiationdeficient HL-60 Cells. *Cancer Res.* **51**, 2854–2862 (1991).
57. Pinho, S. S. & Reis, C. A. Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev Cancer* **15**, 540–555 (2015).
58. Stadie, T. R. E., Chai, W., Lawson, A. M., Byfield, P. G. H. & Hanisch, F. -G. Studies on the Order and Site Specificity of GalNAc Transfer to MUC1 Tandem Repeats by UDP-GalNAc: Polypeptide N -Acetylgalactosaminyltransferase from

- Milk or Mammary Carcinoma Cells. *Eur. J. Biochem.* **229**, 140–147 (1995).
59. Brox, R. D. *et al.* Nuclear Magnetic Resonance-Based Dissection of a Glycosyltransferase Specificity for the Mucin MUC1 Tandem Repeat. *Biochemistry* **42**, 13817–13825 (2003).
60. Olson, F. J., Bäckström, M., Karlsson, H., Burchell, J. & Hansson, G. C. A MUC1 tandem repeat reporter protein produced in CHO-K1 cells has sialylated core 1 O-glycans and becomes more densely glycosylated if coexpressed with polypeptide-GalNAc-T4 transferase. *Glycobiology* **15**, 177–191 (2005).
61. Hanisch, F. G., Reis, C. A., Clausen, H. & Paulsen, H. Evidence for glycosylation-dependent activities of polypeptide N-acetylgalactosaminyltransferases rGalNAc-T2 and -T4 on mucin glycopeptides. *Glycobiology* **11**, 731–740 (2001).
62. Bennett, E. P., Hassan, H., Hollingsworth, M. a & Clausen, H. A novel human UDP-N-acetyl- D -galactosamine : polypeptide N-acetylgalactosaminyltransferase , GalNAc-T7 , with speci c city for partial GalNAc-glycosylated acceptor substrates. *FEBS Lett.* **460**, 226–230 (1999).
63. Tetaert, D., Richet, C., Gagnon, J., Boersma, A. & Degand, P. Studies of acceptor site specificities for three members of UDP-GalNAc:N-acetylgalactosaminyltransferases by using a synthetic peptide mimicking the tandem repeat of MUC5AC. *Carbohydr. Res.* **333**, 165–171 (2001).
64. De Las Rivas, M. *et al.* The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. *Nat. Commun.* **8**: 1959 (2017).
65. De Las Rivas, M. *et al.* Structural and Mechanistic Insights into the Catalytic-Domain-Mediated Short-Range Glycosylation Preferences of GalNAc-T4. *ACS Cent. Sci.* **4**, 1274–1290 (2018).
66. Schjoldager, K. T. B. G. & Clausen, H. Site-specific protein O-glycosylation modulates proprotein processing - Deciphering specific functions of the large polypeptide GalNAc-transferase gene family. *Biochim. Biophys. Acta - Gen. Subj.* **1820**, 2079–2094 (2012).
67. Meyer, B. & Möller, H. Conformation of glycopeptides and glycoproteins. *Top. Curr. Chem.* **267**, 187–251 (2007).
68. Coltart, D. M. *et al.* Principles of mucin architecture: Structural studies on synthetic glycopeptides bearing clustered mono-, di-, tri-, and hexasaccharide glycodomains. *J. Am. Chem. Soc.* **124**, 9833–9844 (2002).
69. Barchi, J. J. Mucin-type glycopeptide structure in solution: Past, present, and future. *Biopolymers* **99**, 713–723 (2013).
70. Barb, A. W., Borgert, A. J., Liu, M., Barany, G. & Live, D. *Intramolecular glycan-protein interactions in glycoproteins. Methods in Enzymology* **478**, (Elsevier Inc, 2010).
71. Corzana, F. *et al.* Serine versus threonine glycosylation: The methyl group causes a drastic alteration on the carbohydrate orientation and on the surrounding water



- shell. *J. Am. Chem. Soc.* **129**, 9458–9467 (2007).
72. Madariaga, D. *et al.* Serine versus threonine glycosylation with  $\alpha$ -o-GalNAc: Unexpected selectivity in their molecular recognition with lectins. *Chem. - A Eur. J.* **20**, 12616–12627 (2014).
73. Wishart, D. S., Sykes, B. D. & Richards, F. M. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* **222**, 311–333 (1991).
74. Bermejo, I. A. *et al.* Water Sculpts the Distinctive Shapes and Dynamics of the Tumor-Associated Carbohydrate Tn Antigens: Implications for Their Molecular Recognition. *J. Am. Chem. Soc.* **140**, 9952–9960 (2018).
75. Coelho, H. *et al.* The Quest for Anticancer Vaccines: Deciphering the Fine-Epitope Specificity of Cancer-Related Monoclonal Antibodies by Combining Microarray Screening and Saturation Transfer Difference NMR. *J. Am. Chem. Soc.* **137**, 12438–12441 (2015).
76. Wishart, D. S., Sykes, B. D. & Richards, F. M. The Chemical Shift Index: A Fast and Simple Method for the Assignment of Protein Secondary Structure through NMR Spectroscopy. *Biochemistry* **31**, 1647–1651 (1992).
77. Kinarsky, L. *et al.* Conformational studies on the MUC1 tandem repeat glycopeptides: Implication for the enzymatic O-glycosylation of the mucin protein core. *Glycobiology* **13**, 929–939 (2003).
78. Keller, R. *The computer aided resonance assignment tutorial*. Goldau, Switzerland: Cantina Verlag (2004).
79. Schumann, F. H. *et al.* Combined chemical shift changes and amino acid specific chemical shift mapping of protein-protein interactions. *J. Biomol. NMR* **39**, 275–289 (2007).



# Chapter

# 3

## *The conformational and interaction features of glycolipid with TOLL-like receptors and accessory proteins.*

*The work presented in this chapter has been performed within the framework of the TOLLERANT EU project, in close collaboration with Dr. Francesco Peri, University of Milano-Bicocca, Milano, Italy. Alberto Minotti and Florent Cochet (PhD students) were responsible for the synthetic aspects, while Fabio Facchini and Stefania Enza Sestito (PhD students) performed the in vitro assay with cells. Dr. David Andreu (Pompeu Fabra University, Barcelona Biomedical Research Park, Spain) provided the antimicrobial peptides. Furthermore, in collaboration with Dr. Alba Silipo, University of Naples Federico II, Napoli, Italy, Mateusz Pallach (PhD student) was responsible for the extraction, isolation and characterization of the natural LPSs.*

---

**Publications:** (1) - Florent Cochet, Fabio A. Facchini, Lenny Zaffaroni, Jean-Marc Billod, **Helena Coelho**, Aurora Holgado, Harald Braun, Rudi Beyaert, Roman Jerala, Jesus Jimenez-Barbero, Sonsoles Martin Santamaria, Francesco Peri (2019) Novel carboxylate-based glycolipids: TLR4 antagonism, MD-2 binding and self-assembly properties *Scientific Reports* 9:919 (DOI: 10.1038/s41598-018-37421-w)

(2) - Fabio Alessandro Facchini, **Helena Coelho**, Stefania Enza Sestito, Sandra Delgado, Alberto Minotti, David Andreu, Jesús Jiménez-Barbero, Francesco Peri (2017) Co-administration of Antimicrobial Peptides (AMPs) Enhances Toll-like Receptor 4 (TLR4) Antagonist Activity of a Synthetic Glycolipid. *ChemMedChem*, 13, 280-287. (DOI: 10.1002/cmdc.201700694).



### **3.1 Introduction**

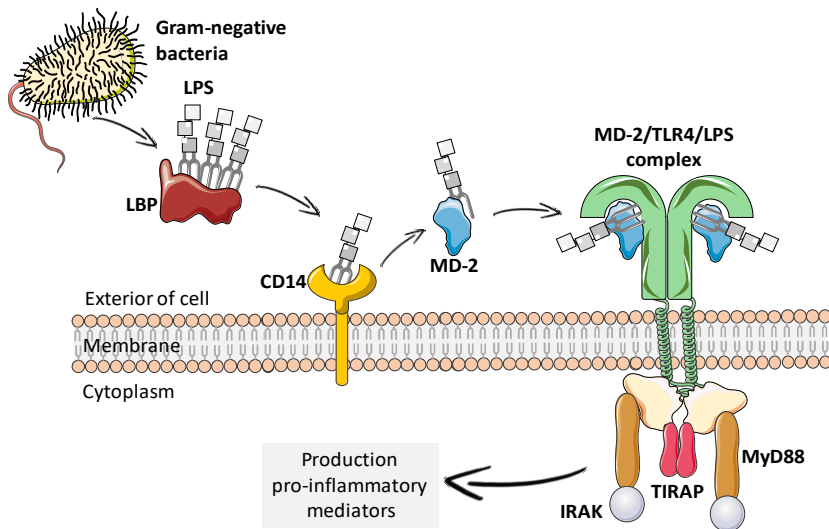
The immune system can be described as a set of molecules, cells and tissues responsible for conferring protection against external aggressions. In particular, against antigens that are strange to the body, whether a microorganism or macromolecule, and various internal injuries, or modified tumor cells. Thus, immunity can be defined as a set of defense mechanisms that our body has to protect itself from attacks and to maintain the immune homeostasis. The immune system under normal conditions does not show any kind of response to host cells (immunological tolerance).<sup>1-3</sup>

There are two types of immune responses that complement each other. Innate immune response is the first line of attack to pathogens, recognizing only molecular patterns stored in microorganisms, the so-called Pathogen-Associated Molecular Patterns (PAMPs) and the danger-associated molecular patterns (DAMPs), derived from host cells with changes in homeostasis. The cells that are part of this kind of immunity are phagocytic cells (neutrophils, macrophages), dendritic cells (DCs), natural killer (NK) cells, and biological barriers (as epithelia).<sup>4</sup> The adaptive immune system acts latter on the immune response. In contrast to innate immune responses, adaptive ones are stimulated by an immunogen (substance that is capable of trigger an immune response). In this case, the response magnitude and efficacy is usually increased after each exposure, including the development of antibodies memory. Therefore, innate immunity initiates the responses to directly kill pathogens, and, if not enough to clear it, the antigen-specific adaptive immune responses are triggered to provide a second layer of protection.<sup>4</sup>

Toll-like receptors (TLRs) are mainly expressed on innate immune cells and play an important role in the cellular pathogen recognition and consequent immune responses. TLRs belong to the family of pattern recognition receptors (PRRs) and participate in immune surveillance by detecting PAMPs.<sup>5,6</sup> This recognition triggers the activation of phagocytic cells that can lead to the internalization of the microorganism, as well as to cytokines and inflammatory mediators release,

resulting in an inflammatory process.<sup>1,3</sup> Structurally, the TLRs have a single transmembrane helix that connects the extracellular ligand-binding domain to the intracellular signaling domain.

In humans, TLRs comprise a gene family of 10 receptors with various functions, including defense against a variety of diseases such as infections, cancers, autoimmune, inflammation, etc. Among the TLR family members, Toll-Like Receptor 4 (TLR4) was the first receptor to be identified and characterized in humans.<sup>7</sup> TLR4 is expressed at the surface of innate immune cells (macrophages, DCs) and specifically recognizes bacterial endotoxins, i.e., lipopolysaccharides (LPS) or its truncated version lipooligosaccharides (LOS), the main molecular components of gram-negative bacteria cell walls.<sup>8</sup> LPS (or endotoxin) was confirmed as the first agonist ligand of TLR4. LPS is provided by the cell wall of Gram-negative bacteria and recognized by a cascade of receptors, including the cluster differentiation antigen 14 (CD14), TLR4, and the myeloid differentiation protein (MD-2). The recognition of LPS starts with the binding of LPS aggregates to the LPS binding protein (LBP), followed by the transfer of LPS monomers to CD14. After that, CD14 presents LPS monomers to MD-2, triggering the formation of the TLR4-MD-2 complex. The process concludes with the formation of the activated (TLR4.MD-2.LPS)<sub>2</sub> dimer, which has a key role in starting the inflammatory cascade (Figure 3.1). In fact, the receptor dimerization induces the recruitment of adapter proteins to the intracellular TIR domain of TLR4, prompting the activation of NF-κB (Nuclear Factor Kappa B) and Interferon Regulatory Factor 3 (IRF3) and the triggering of cytokines secretion.<sup>9-11</sup>



**Figure 3.1-** Scheme of Toll-Like Receptor 4-associated proteins that are involved in LPS sensing.

Modulation of innate immunity receptors by agonists and antagonists with synthetic small molecules able to modulate this activity represents a powerful tool to study the TLR4 receptor system. These molecules display pharmacological interest as antiseptics and anti-inflammatory agents (antagonists) or as vaccine adjuvants (agonists).<sup>12,13</sup>

A classification for TLR4 modulators is based on the effect they have on the TLR4 pathway. These modulators can act as antagonists, when they are able to inhibit the LPS-triggered TLR4 activation, or as agonists, when they stimulate the activation of TLR4 pathway. To modulate the activation of TLR4 pathway, these molecules can interact with one or more receptors or co-receptors of the pathway, interfering at different stages of the signaling.

For TLR4 antagonists, the greatest effort has been made to treat septic shock and potentially all the pathologies due to the deregulated activation of TLR4 pathway, including asthma, cardiovascular disorders, diabetes, obesity, metabolic syndromes, autoimmune disorders, neuroinflammatory disorders, neuropathic pain,

central nervous system disorders such as amyotrophic lateral sclerosis and Alzheimer disease, psychiatric diseases, skin inflammations (dermatitis), psoriasis, and some tumors.<sup>13</sup> A second motivation for the discovery of TLR4 ligands has been the development of new agonists that could be employed as vaccine adjuvants: molecules that elicit a controlled immune response.<sup>14,15</sup>

LPS and derivatives, LOS or mimics, are amphiphilic molecules with low sub-micromolar/nanomolar values of critical micellar concentration (CMC) in aqueous solutions, thus showing a propensity to aggregate in the concentration range relevant for biological responses. CMC values for *E. coli* LPS of 1.3 - 1.6  $\mu\text{M}$  have been measured by fluorescence correlation microscopy.<sup>16</sup> Nevertheless, as for any amphiphilic system, monomers are also present in dynamic equilibrium. The key question of what is the biologically active unit of endotoxins, whether large or small aggregates or monomers, has been amply debated in the literature.<sup>17-19</sup> The first steps of LPS recognition, as binding to LBP and CD14, are almost certainly influenced by the size and 3D shape of LPS aggregates. However, the binding of single LPS molecules to receptors (CD14 and MD-2) is the basis of sensitive and selective of endotoxin molecular recognition by the (TLR4/MD-2/lipid A)<sub>2</sub> activated complex.<sup>20-23</sup>

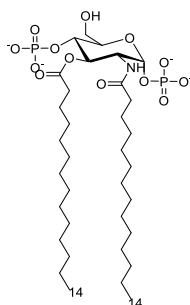
This work aims to study the interaction of a synthetic small molecule with antagonist of TLR4 properties with antimicrobial peptides and the characterization of the supramolecular structures formed in water solution of natural LPS or mimetic of Lipid A (synthetic molecules) by NMR and Electron Microscopy.



### ***3.2 The interaction of Lipid A mimics with antimicrobial peptides***

Antimicrobial peptides (AMP) are known to bind to and neutralize LPS, interacting with endotoxin. However, it was observed that neutralization of LPS by AMP is associated with a drastic change of the LPS aggregate-type, from a cubic to a multilamellar shape, increasing the size of the aggregate and thus leading to the inhibition of the binding of endotoxin to LBP and other mammalian proteins.<sup>24</sup>

Based on this premises, we herein focused on investigating the possibility that AMPs could interact with a simple LPS glycomimetic, the anionic monosaccharide FP7 (Figure 3.2), which has been described as a TLR4 antagonist,<sup>25</sup> and could modulate its antagonist activity on TLR4.

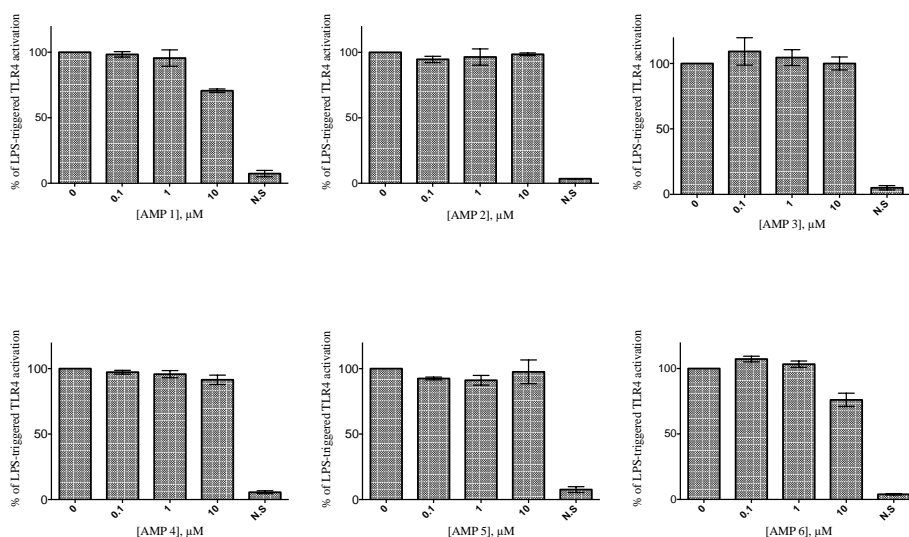


**Figure 3.2** - Chemical structure of FP7, a synthetic diacylated glucopyranose derivative active as TLR4 antagonist.

#### **3.2.1 AMPs enhances FP7 antagonist activity in HEK-Blue hTLR4 cells**

The effect of FP7 and antimicrobial peptides (AMPs) co-administration was first investigated in HEK-Blue hTLR4 cells, i.e., HEK293 cells stably transfected

with human TLR4, MD-2, and CD14 genes. Furthermore, HEK-Blue hTLR4 cells display a secreted embryonic alkaline phosphatase (SEAP), produced upon activation of NF- $\kappa$ B, as reporter gene. Thus, the LPS binding event causes TLR4 dimerization, myddosome formation, and NF- $\kappa$ B activation, leading to SEAP production and secretion. FP7 (Figure 3.2) has been reported previously as a TLR4 antagonist.<sup>25</sup> These results were further confirmed by inhibiting the LPS-stimulated TLR4 activation, in a dose-dependent manner, providing a  $IC_{50}$  of 2.5  $\mu$ M. Remarkably, no antagonist effects were detected for the AMPs administered alone on the same cell line, at the concentration range used (Figure 3.3).



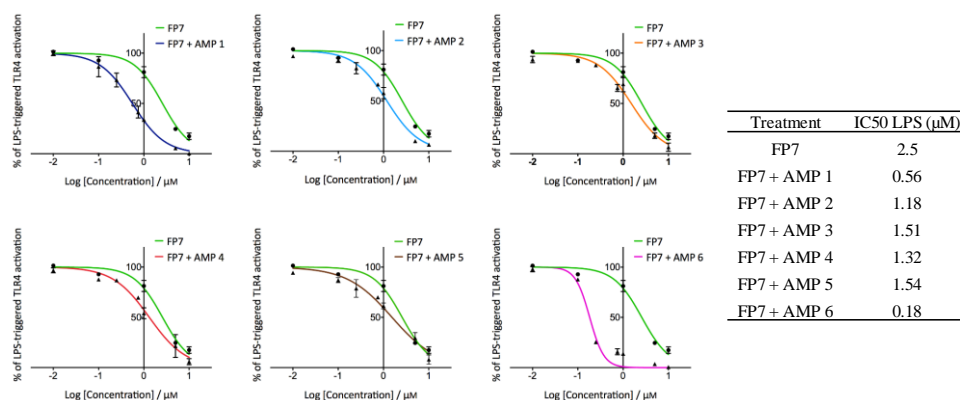
**Figure 3.3** - Effect of AMPs 1-6 administration on LPS-stimulated TLR4 signal in HEK-Blue hTLR4 cells. HEK-Blue hTLR4 cells were pre-treated with the indicated concentrations of AMPs 1-6 and stimulated with LPS (100 ng/mL) after 30 minutes. Data were normalized to stimulation with LPS alone. Data represent the mean of percentage  $\pm$  SEM of at least 3 independent experiments. N.S. No stimulated with LPS.

However, when FP7 was co-administered (1:1 stoichiometric ratio) with AMPs (Table 3.1), different results were obtained. The AMPs **2-5** poorly improved FP7 antagonist activity ( $IC_{50}$  around 1.1 - 1.5  $\mu$ M), while AMPs **1** and **6** exhibited

stronger activity ( $IC_{50}$  0.56  $\mu$ M and 0.18  $\mu$ M respectively) (Figure 3.4). On the other hand, AMP 1 and AMP 6 efficiently improved the TLR4 antagonist activity of FP7 in HEK-Blue hTLR4 cells, in a 0 to 10  $\mu$ M concentration range.

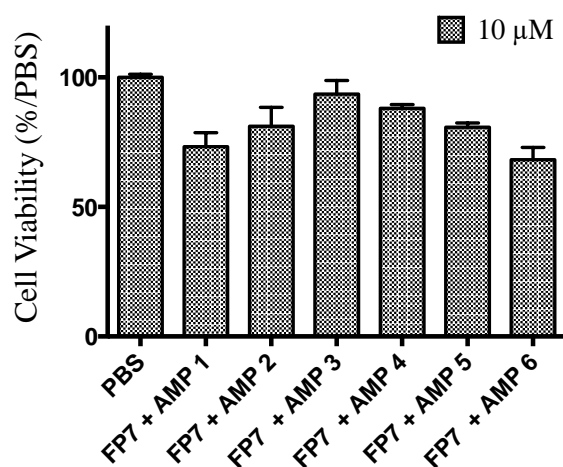
**Table 3.1** - Sequences of the anti-microbial peptides.

AMP	Name	Sequence	
1	CA(1-8)M(1-8)	KWKLFKKIGIGAVLKVLTTGLPALIS-amide	26
2	CA(1-7)M(2-9)	KWKLFKKIGAVLKVL-amide	27
3	[K <sup>6</sup> (Me <sub>3</sub> )]CA(1-7)M(2-9)	KWKLFK(Me <sub>3</sub> )KIGAVLKVL-amide	28
4	N <sup>o</sup> -Oct-CA(1-7)M(2-9)	Octanoyl-KWKLFKKIGAVLKVL-amide	29
5	CA(1-7)M(5-9)	KWKLFKKVLKVL-amide	27
6	LL-37	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLPRTES	30



**Figure 3.4** - Dose-dependent inhibition of LPS-stimulated TLR4 signal in HEK-Blue hTLR4 cells by FP7/AMPs co-administrations. Activities of FP7/AMPs administrations on LPS-stimulated TLR4 signal in HEK-Blue hTLR4 cells.

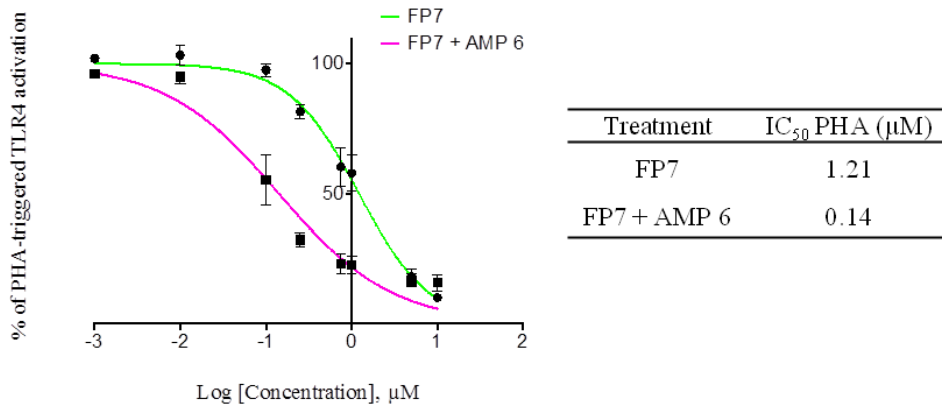
The cytotoxic effects of the co-administration were also tested to exclude the possibility that the activity increase was related with cytotoxic activity. The toxicity of all co-administrations was evaluated by a MTT test, which permitted showing the absence of significant toxicity even at the highest concentration tested (10  $\mu$ M) (Figure 3.5).



**Figure 3.5** - MTT assay of FP7/AMPs co-administrations in HEK-Blue hTLR4 cells. Cells were treated with the six co-administrations used in the other assays; the bars represent the cell viability estimated by using 10  $\mu$ M of compounds, equivalent to the maximum concentration used previously. Data are normalized with PBS and represent the mean of percentage  $\pm$  SEM of at least 3 independent experiments.

To exclude the hypothesis that the effects of AMPs in the activity of FP7 could be due to a neutralizing effect on LPS, additional experiments were performed. The addition of phytohemagglutinin (PHA), in a dose-dependent manner, stimulated TLR4 activation in HEK-Blue hTLR4 cells, when co-administered with FP7 and LL37. PHA-L induces the activation of TLR4 as agonist, but with a lower potency than LPS.<sup>31</sup> The treatment with FP7 alone inhibited TLR4 activation in a dose-dependent way, as expected. In addition, the co-administration with LL37 showed enhancement of FP7 antagonism activity

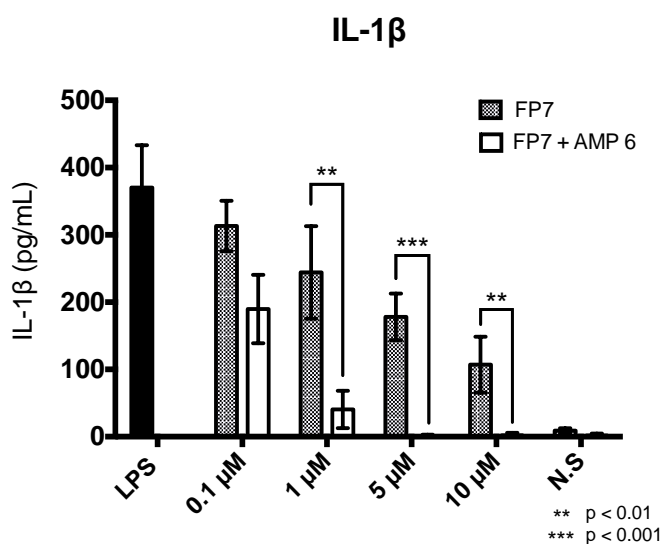
(Figure 3.6), suggesting that this effect is, at least in part, independent from an LPS neutralizing effect.



**Figure 3.6** - Dose-dependent inhibition of PHA-stimulated TLR4 activation by FP7 and FP7/AMP6. Cells were treated with increasing concentrations of compounds and stimulated with PHA-P (5 μg/mL). The results represent normalized data with positive control (PHA-P alone). Concentration-effect data were fitted to a sigmoidal 4 parameter logistic equation to determine IC<sub>50</sub> values and represent the mean of percentage ± SEM of at least 3 independent experiments. The IC<sub>50</sub> values are shown Table.

### 3.2.2 LL-37 (AMP6) enhances FP7 antagonist activity in human PBMC

HEK-Blue hTLR4 cell is a model. Thus, the capacity of LL-37 to improve FP7 antagonist activity in human monocytes was also investigated. Human Peripheral Blood Mononuclear Cells (h)PBMCs were isolated from buffy coats, pre-incubated with increasing concentrations (0.1-10 μM) of FP7 or FP7/LL37 (1/1) and stimulated with LPS (100 ng/mL). As expected, FP7 induced the reduction in the production of IL-1β. The addition of LL37 to FP7 produced a much more powerful inhibitory response, decreasing the production of IL-1β already at the lowest dose of 1 μM. (Figure 3.7).



**Figure 3.7.** LL-37 (AMP6) potentiation of FP7 antagonist activity in human PBMCs. PBMCs isolated from buffy coats were preincubated with FP7 or FP7/AMP6 mix for 30 minutes and then stimulated with LPS (100 ng/mL). IL-1 $\beta$  production was quantified after one night's incubation. Data represent the mean  $\pm$  SEM of at least three independent experiments.

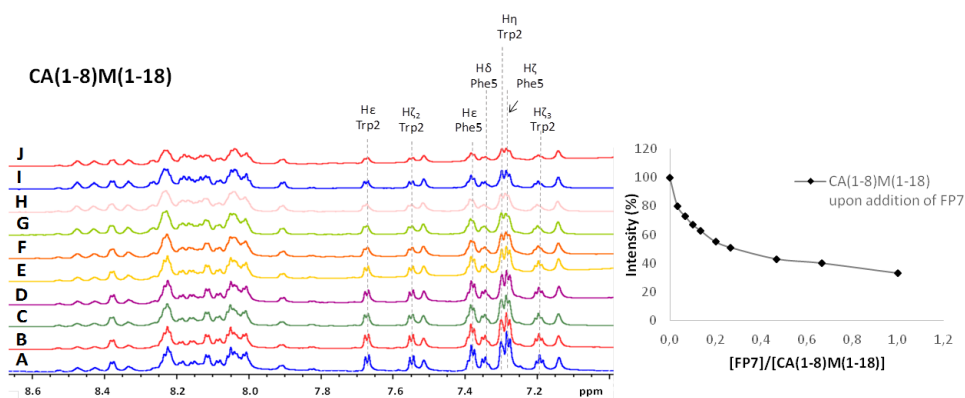
### **3.2.3 NMR and TEM analysis of glycolipid/peptide interaction**

The characterization of the interaction between the synthetic TLR4 antagonist FP7, and these two synthetic anti-microbial peptides, CA(1-8)M(1-18) and LL37 was studied by NMR experiments. Nuclear magnetic resonance (NMR) spectroscopy has become a powerful tool to investigate structural features, dynamics and interactions of biomolecules in solution, at atomic resolution.

#### **3.2.3.1 *NMR analysis from the AMP point-of-view***

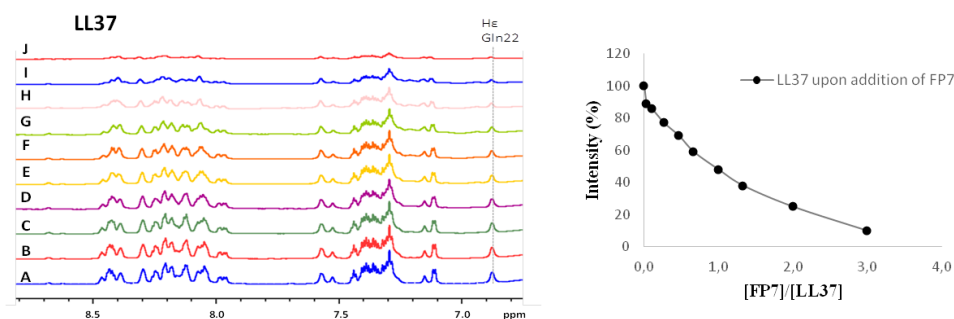
The effect of binding can be successfully traced by the perturbations observed on characteristic NMR parameters (e.g. chemical shifts, line widths, and

signal intensities) of either partner. The titration of CA(1-8)M(1-18) (**AMP1**) peptide with FP7 (Figure 3.8) permitted to observe clear perturbation of the signals of the lateral chains of the hydrophobic aminoacids upon addition of the FP7 glycolipid. The experimentally observed reductions in intensity (Figure 3.8), due to specific line broadening of these signals, probably arise from the changes in the transverse relaxation times of these signals, a clear indication of the existence of a binding event. Additional variations in the amide region were also observed (Figure 3.8). This dramatic change likely arises from the existence of interaction between CA(1-8)M(1-18) peptide and the FP7 glycolipid.



**Figure 3.8- Left** <sup>1</sup>H-NMR of CA(1-8)M(1-18) (300 μM) with upon addition of FP7. **A:** CA(1-8)M(1-18) alone; **B:** FP7 (10 μM); **C:** FP7 (20 μM); **D:** FP7 (30 μM); **E:** FP7 (40 μM); **F:** FP7 (60 μM); **G:** FP7 (80 μM); **H:** FP7 (140 μM); **I:** FP7 (200 μM); **J:** FP7 (300 μM); **I-Right** Intensity (%) of the Hε- Trp2 of CA(1-8)M(1-18) as a function of the [FP7]/[CA(1-8)M(1-18)] molar ratio. The samples have 10 % DMSO in PBS 100 mM pH=5.5 in H<sub>2</sub>O/D<sub>2</sub>O 90:10 at 310K, 64 scans.

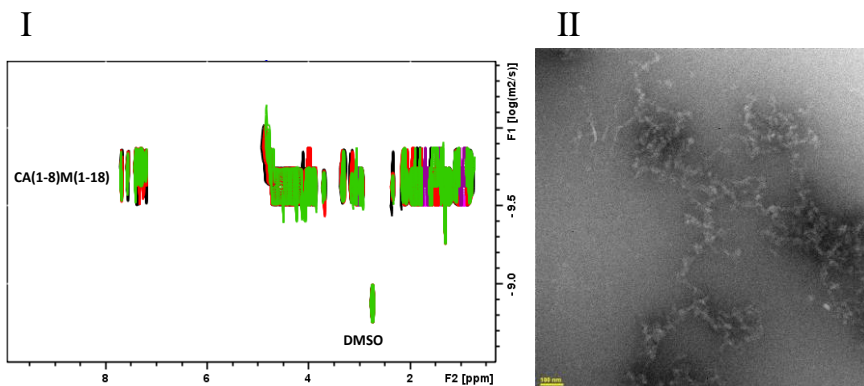
The interaction of the second anti-microbial peptide, LL37 (**AMP6**) with FP7 was also investigated by using <sup>1</sup>H-NMR (Figure 3.9). Once more, the existence of broadening of the peptide NMR resonance signals upon addition of FP7 permitted to deduce the existence of a molecular recognition phenomenon.



**Figure 3.9 - Left)** <sup>1</sup>H-NMR of LL37 (300 μM) with upon addition of FP7. **A:** LL37 alone; **B:** FP7 (10 μM); **C:** FP7 (30 μM); **D:** FP7 (80 μM); **E:** FP7 (140 μM); **F:** FP7 (200 μM); **G:** FP7 (300 μM); **H:** FP7 (400 μM); **I:** FP7 (600 μM); **J:** FP7 (900 μM); **Right)** Intensity (%) of the Hε-Gln22 of LL37 peptide as a function of [FP7]/[LL37]. The samples have 10 % DMSO in PBS 100 mM pH=5.5 in H<sub>2</sub>O/D<sub>2</sub>O 90:10 at 310K, 64 scans.

The DOSY (Diffusion Order Spectroscopy) experiments on the CA(1-8)M(1-18) peptide showed a strikingly low diffusion coefficient, far from its small/medium molecular weight, strongly suggesting that the peptide is aggregated (Figure 3.9 -I and Table 3.2). This experimental evidence was also deduced by transmission electron microscopy (TEM) negative staining analysis (Figure 3.9 - II). Indeed, filament-like shapes were observed for the peptide alone.



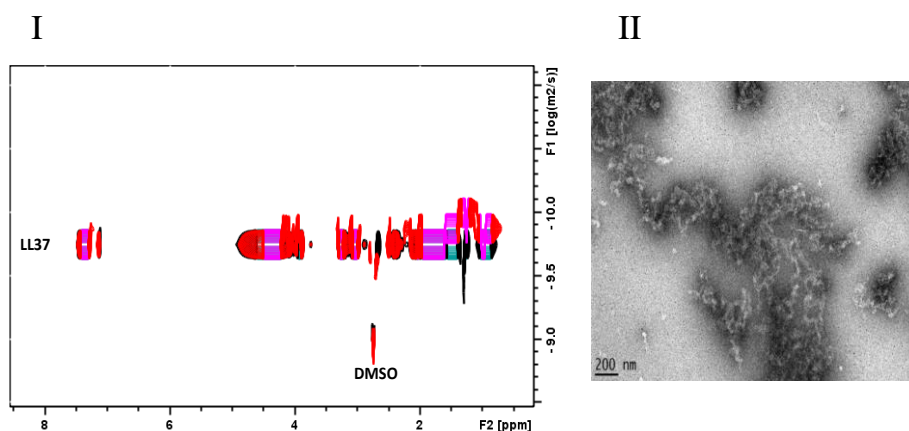


**Figure 3.9 – I)** DOSY spectrum Black: CA(1-8)M(1-18) (300µM) Red: : CA(1-8)M(1-18) (300µM) FP7 (80µM). Green : CA(1-8)M(1-18) (300µM) FP7 (200µM). **II)** Transmission Electron Microscopy - Negative Staining Analysis. CA(1-8)M(1-18) peptide at 2.5mg/ml. nominal magnification of 30,000 X (0.36nm/pixel).

**Table 3.2 - CA(1-8)M(1-18) peptide (300µM) diffusion data values.**

	$D/m^2s^{-1}$
CA(1-8)M(1-18)	$2.29 \times 10^{-10}$
CA(1-8)M(1-18) + FP7	$2.29 \times 10^{-10}$

However, the diffusion co-efficient remains unchanged in the presence of FP7 (Figure 3.9 and Table 3.2). The interaction of FP7 with CA(1-8)M(1-18) peptide (in excess), does not show a large effect in the average size of the CA(1-8)M(1-18) aggregates. As for CA(1-8)M(1-18), DOSY experiments were recorded for LL37 (Figure 3.10 -I and Table 3.3) upon addition of FP7 Lipid. The same result was observed in this case. The diffusion co-efficient is not affected in the presence of FP7 (Figure 3.10 -I and Table 3.3).



**Figure 3.10 – I)** DOSY spectrum. **Black:** LL37 (300 $\mu$ M) **Red:** LL37 (300 $\mu$ M) FP7 (200 $\mu$ M). **II)** Transmission Electron Microscopy. Negative Staining Analysis. Left: LL37 peptide at 45mg/ml. nominal magnification of 30,000 X (0.36nm/pixel).

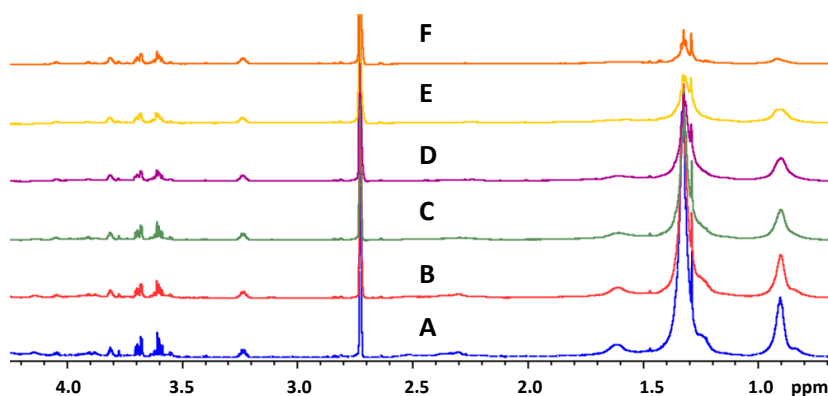
**Table 3.3 - LL37 peptide (300 $\mu$ M) diffusion data values.**

	<b>D/m<sup>2</sup>s<sup>-1</sup></b>
<b>LL37</b>	1.78x10 <sup>-10</sup>
<b>L37 + FP7</b>	1.78 x10 <sup>-10</sup>

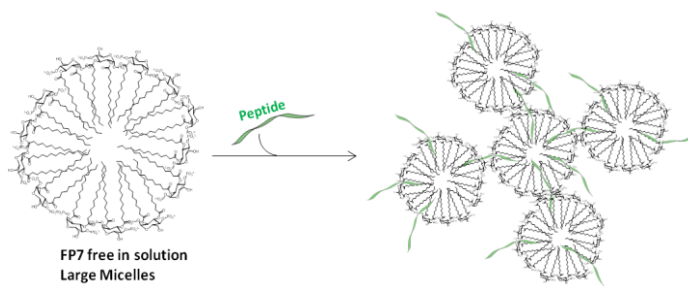
However, as also deduced for **AMP1**, the DOSY of LL37 showed a small diffusion co-efficient, far from that expected for the small/medium molecular weight. Again, this fact strongly suggests that there is aggregation of the peptide. This hypothesis was confirmed with the corresponding TEM analysis (Figure 3.10 -II).

### 3.2.3.2 NMR analysis from the FP7 point-of-view

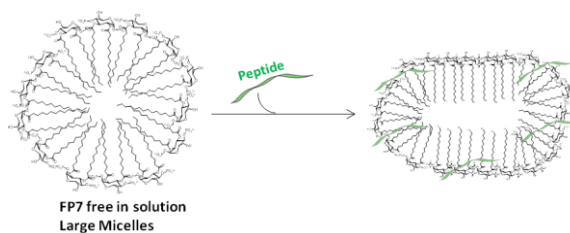
The process was also monitored by looking at NMR signals of the  $^1\text{H}$  signals of the FP7 glycolipid upon addition of the peptide (Figure 3.11). In this case, the dramatic reduction of the intensity of the NMR signals of the aliphatic chains permitted to deduce that the interaction with CA(1-8)M(1-18) peptide involved the lipid chains (Figure 3.11). Three alternative hypotheses could be invoked to explain these data. The peptide could act as linker between different FP7 aggregates (Figure 3.12) and/or deform the FP7 micelle (Figure 3.13) or the peptide could participate in the formation of large aggregates (Figure 3.14), behaving as a large molecule, as earlier described by us for MD2.<sup>25</sup>



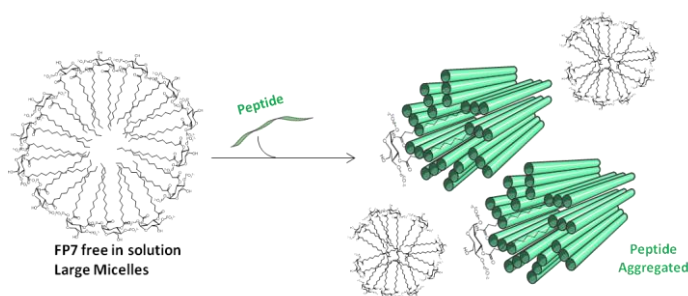
**Figure 3.11-**  $^1\text{H}$ -NMR of FP7 (500  $\mu\text{M}$ ) with upon addition of CA(1-8)M(1-18) peptide. **A:** FP7 alone **B:** CA(1-8)M(1-18) (10  $\mu\text{M}$ ) **C:** CA(1-8)M(1-18) (30  $\mu\text{M}$ ) **D:** CA(1-8)M(1-18) (50  $\mu\text{M}$ ) **E:** CA(1-8)M(1-18) (90  $\mu\text{M}$ ) **F:** [CA(1-8)M(1-18) (170  $\mu\text{M}$ )]. The sample has 10% DMSO in Buffer PBS 100mM pH=5.5 in H<sub>2</sub>O/D<sub>2</sub>O 90:10 at 310K, 560 scans, DS=4.



**Figure 3.12** - Cartoon showing the possibility that the AMP act as linker between different FP7 aggregates

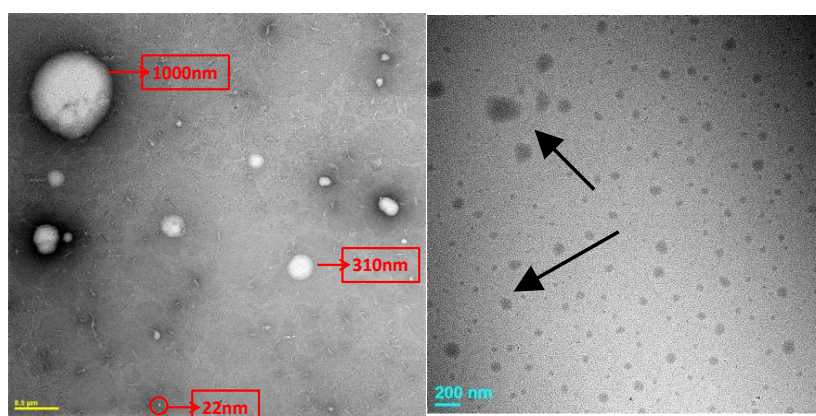


**Figure 3.13** - Cartoon showing the possibility that the AMP changes the shape of FP7 aggregates



**Figure 3.14** - Cartoon showing the possibility that the AMP aggregation and the aggregate behaves as a large molecule.

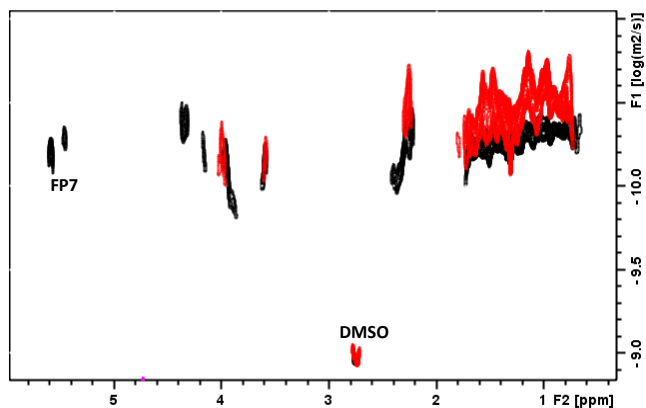
Indeed, FP7, being an amphipathic molecule, forms aggregates or micelles in water solution, as previously demonstrated by DOSY.<sup>25</sup> Thus, the existence of FP7 aggregation and formation of micelles of very diverse size was demonstrated. However, they show the same morphology when analyzed by negative staining TEM (Figure 3.15, Left) or by cryogenic transmission electron microscopy (Cryo-TEM) (Figure 3.15, Right).



**Figure 3.15** - Transmission Electron Microscopy - Negative Staining Analysis. Left: FP-7 at 2.5mg/ml. nominal magnification of 10,000 X (1.1nm/pixel) Right: Transmission Electron Microscopy - cryo-TEM of FP7 alone (2.5 mg/mL) with a nominal magnification of 30000X (0.36 nm/pixel). The samples have 10% DMSO in PBS 100 mM pH=5.5.

In contrast to the case with the antimicrobial peptides, the DOSY experiment of FP7 in the presence of CA(1-8)M(1-18) peptide (Figure 3.16) induced clear perturbations in the diffusion behavior of FP7. Under substoichiometric ratios ( $[CA(1-8)M(1-18)]/[FP7]=0.06$ ) of the peptide, an increase in the diffusion coefficient was evident. This effect could be, in principle, due to changes either in the size (Figure 3.12) or shape (Figure 3.12) of the lipid. Thus, TEM with Negative Staining Analysis (Figure 3.17) and Cryo-TEM (Figure 3.18) were employed to obtain the required morphological information. The presence of the CA(1-8)M(1-18) peptide induced formation of aggregates between different FP7 micelles, thus

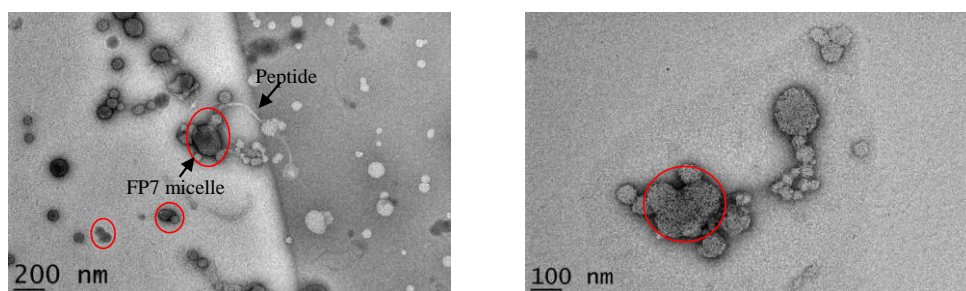
supporting the hypothesis outlined in Figure 3.12. The peptide is linking various FP7 micelles, displaying peanut-shaped structures. This fact suggests the presence of fusion events (indicated in red in the Figure 3.17 and Figure 3.18).



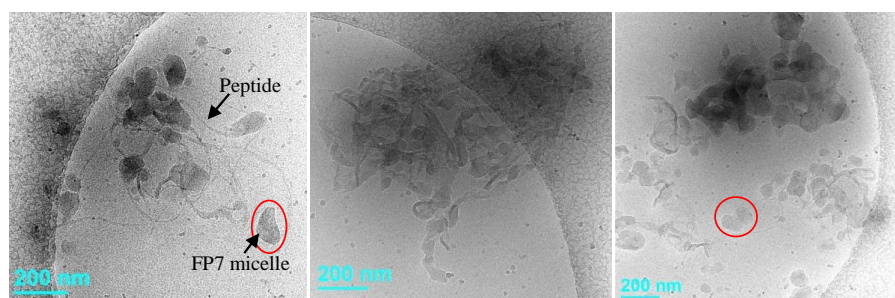
**Figure 3.16** - DOSY spectrum. *Black*: FP7 (500 $\mu$ M) *Red*: FP7 (500 $\mu$ M) and CA(1-8)M(1-18) (30 $\mu$ M)

**Table 3.4** - FP7 Lipid (500 $\mu$ M) diffusion data values.

	<b>D/m<sup>2</sup>s<sup>-1</sup></b>
<b>FP7</b>	5.01x10 <sup>-11</sup>
<b>FP7 + CA(1-8)M(1-18)</b>	3.16 x10 <sup>-11</sup>

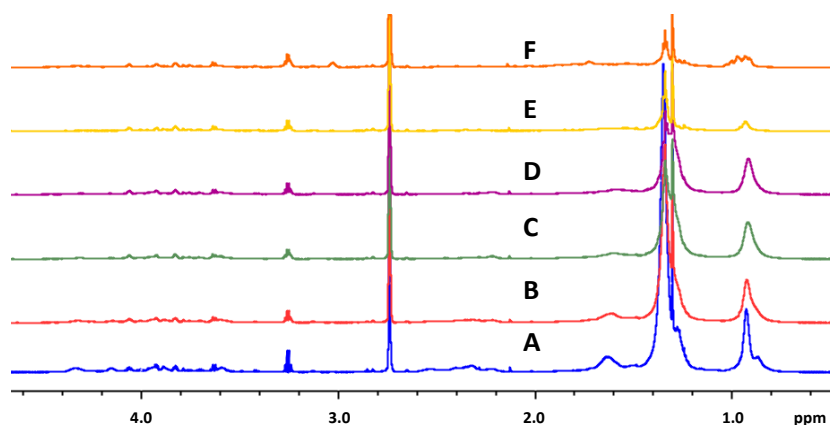


**Figure 3.17** - Transmission Electron Microscopy - Negative Staining Analysis. Left: FP7 Lipid (320  $\mu$ M) with CA(1-8)M(1-18) peptide (80  $\mu$ M) in H<sub>2</sub>O and DMSO 10%, nominal magnification of 10000 X (0.36 nm/pixel) Right : FP7 Lipid (320  $\mu$ M) with CA(1-8)M(1-18) peptide (80  $\mu$ M) in PBS 100 mM pH=5.5 with DMSO 10%, nominal magnification of 20000 X (0.36 nm/pixel)



**Figure 3.18** - Transmission Electron Microscopy - Cryo-TEM FP7 Lipid (320  $\mu$ M) with CA(1-8)M(1-18) peptide (80  $\mu$ M) in PBS 100 mM pH=5.5 with DMSO 10%, nominal magnification of 30000 X (0.36 nm/pixel)

As well as for the CA(1-8)M(1-18) peptide, the <sup>1</sup>H-NMR of FP7 upon addition of LL37 (Figure 3.19) showed that the aliphatic chains are clearly involved in the binding event, since they broaden significantly in the presence of the peptide.

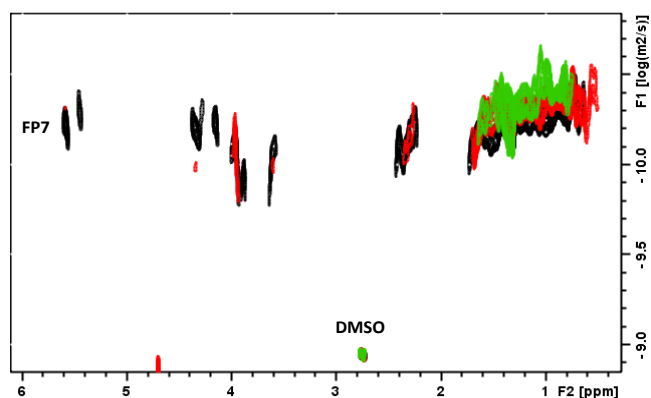


**Figure 3.19** -  $^1\text{H-NMR}$  of FP7 ( $500\ \mu\text{M}$ ) with upon addition of LL37 peptide. **A:** FP7 alone. **B:** LL37 ( $10\ \mu\text{M}$ ) **C:** LL37 ( $20\ \mu\text{M}$ ) **D:** LL37 ( $30\ \mu\text{M}$ ) **E:** LL37 ( $50\ \mu\text{M}$ ) **F:** LL37 ( $90\ \mu\text{M}$ ). The sample has 10% DMSO in Buffer PBS 100mM pH=5.5 in  $\text{H}_2\text{O}/\text{D}_2\text{O}$  90:10 at 310K, 320 scans,  $DS=4$ .

DOSY experiments showed that, upon addition of LL37 peptide (Figure 3.20 and Table 3.5), there were clear perturbations of the diffusion of FP7. The peptide causes a decrease in the diffusion coefficient, although the observed perturbation is smaller compared to that in the presence of CA(1-8)M(1-18) peptide.

The TEM analysis showed that the effect of LL37 is different to that of CA(1-8)M(1-18) in the presence of FP7 micelles. In this case, a dramatic change in the shape, from spheres to cylinders, was induced when LL37 was present. Long entangled cylindrical micelles are now displayed in the cryo-TEM image (Figure 3.21).

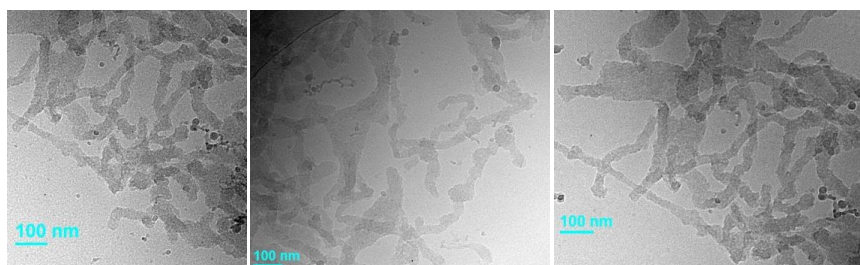




**Figure 3.20** - DOSY spectrum. **Black:** FP7 (500µM) **Red:** FP7 (500µM) and LL377 (10µM) **Green** FP7 (500µM) and LL377 (50µM)

**Table 3.5** - FP7 Lipid (500µM) diffusion data values.

	$D/m^2s^{-1}$
<b>FP7</b>	$6.31 \times 10^{-11}$
<b>FP7 + [LL377]=10µM</b>	$5.0 \times 10^{-11}$
<b>FP7 + [LL377]=50µM</b>	$3.98 \times 10^{-11}$



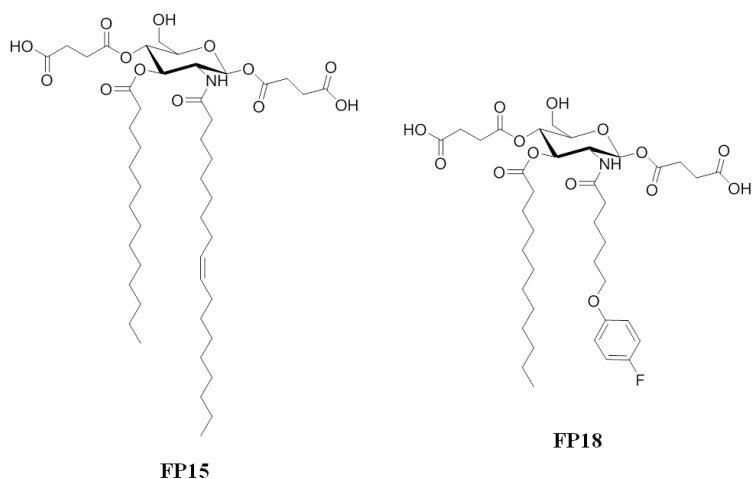
**Figure 3.21** - Transmission Electron Microscopy. Cryo-TEM. FP7 Lipid (588 µM) with LL37 peptide (400 µM) in PBS 100 mM pH=5.5 with DMSO 10%, nominal magnification of 30000 X ((0.36 nm /pixel).

NMR and TEM experiments clearly showed an effect of either CA(1-8)M(1-18) or LL37 on FP7 aggregation state. NMR shows addition of either AMP to FP7 to cause the formation of larger aggregates, as revealed by the reduction in the intensities of FP7 aliphatic chain <sup>1</sup>H NMR signals and the decrease of FP7 diffusion coefficient in DOSY. Cryo-TEM images confirm these data and clearly show that, upon peptide addition, FP7 micellar aggregates undergo a change in size and 3D shape from spherical to rod-like cylindrical.

In the end, although the NMR experiments have been performed at a concentration two orders of magnitude higher than that at which FP7 displays biological activity, they provide a valuable indication on the ability of these antimicrobial peptides to affect FP7 aggregation state in aqueous environment.

### ***3.3 Design of new TLR4 antagonists based on FP7***

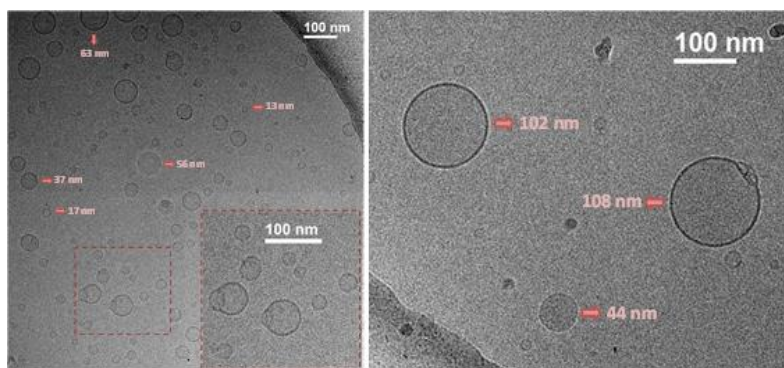
FP7 monosaccharide structure is a proper structural scaffold for designing new active antagonists on TLR4.<sup>32</sup> Substitution of phosphates with carboxylates was proposed to provide new molecules that could maintain their activity on TLR4. Thus, the newly proposed FP15 and FP18 molecules (Figure 3.22) display two carboxylic groups at C1 and C4, which are obviously in the form of  $\beta$ -carboxylated anions at neutral pH and therefore could mimic the two negative phosphates of the FP7 molecule.



**Figure 3.22** - Chemical structure of FP15 and FP18 compounds.

The powerful antagonist Eritoran contains saturated and unsaturated lipophilic chains attached at the C3 and C2 positions of the glucosamine moieties. Thus, FP15 was designed to display aliphatic chains at C3 (C14) and oleic chains at C2 (C18, cis-9). On the other hand, FP18 displays a myristic moiety at C3 (C14) and an aliphatic chain of 5 carbons at C2, finalizing with a <sup>19</sup>F-containing *para*-phenyl group.

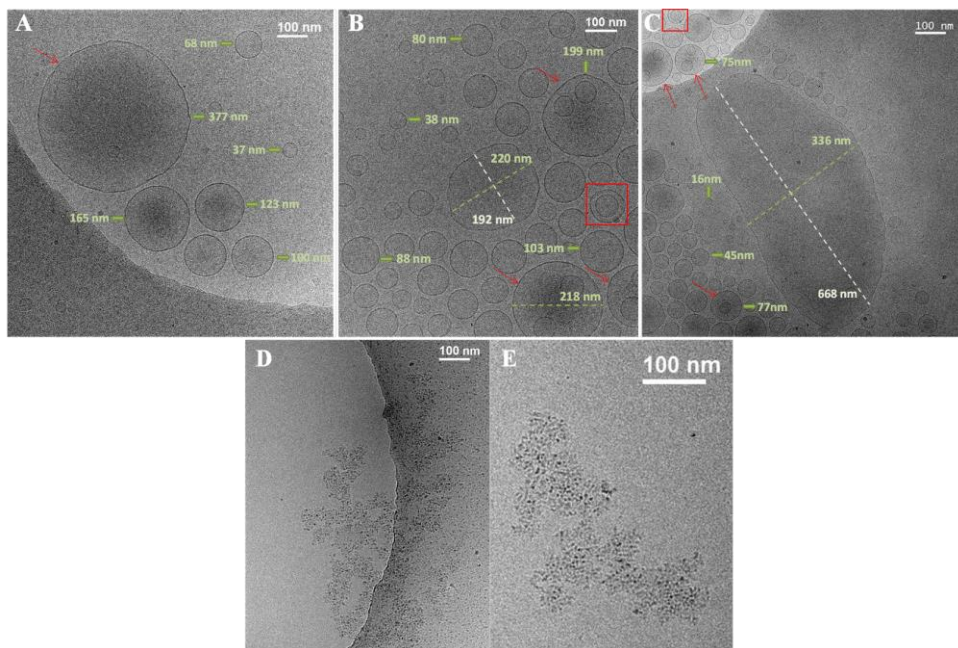
As already described above, FP7 aggregates in aqueous media. Thus, to investigate the behavior of FP15 and FP18 in solution, Cryogenic Transmission Electron Microscopy (Cryo-TEM) methods were employed (Figure 3.23). The detailed inspection of the cryo-TEM data for FP15 allowed showing that this glycolipid displays a different aggregation behavior in solution to that of FP7.



**Figure 3.23** - Cryogenic Transmission Electron Microscopy (Cryo-TEM) of FP15 (7 mg/mL) nominal magnification of 40,000 X (0.26 nm/pixel).

FP15 mainly generates circular and homogeneous small unilamellar vesicles (SUV), although with rather different size distributions, from 10 to 110 nm. Moreover, it was possible to detect the presence of fusion events (as highlighted in the zoomed picture in Figure 6, dark red squares) as well as the existence of open bilayers. Indeed, the use of vitrified samples allowed trapping the potentially un structures associated with the formed intermediates in the solubilization process of the vesicles.

According to the Cryo-TEM data, FP18 forms more heterogeneous structures than FP15. In this case, (Figure 3.23) it was possible to observe the presence of circular Small Unilamellar Vesicles (SUV, a vesicle of a single bilayer and small size) with different sizes, between 16 – 110 nm, as well as circular and amorphous Large Unilamellar Vesicles (LUV) (vesicle of a single bilayer of larger size) with sizes ranging from 200 – 700 nm. In addition, FP18 forms Multivesicular Vesicles (MVV, vesicles of the bilayer that are not arranged in a concentric manner, red arrows in Figure 3.23),<sup>33</sup> as well as Multilamellar Vesicles (MLV, vesicles made of concentric layers of the bilayer,<sup>33</sup> red squares in Figure 3.23).



**Figure 3.23** – Cryogenic Transmission Electron Microscopy - Cryo-TEM of FP15 (7 mg/mL) nominal magnification of 40000 X (0.36 nm/pixel).

Besides vesicles, in the Cryo-TEM of FP18, it was possible to see long and numerous entangled structures (Figure 3.23 down), which are probably formed by wormlike micelles.<sup>34</sup>

Therefore, these experimental data show that these compounds are able to form vesicles/liposomes displaying a double layer. This fact can be a good starting point to use these them for drug delivery issues.

### **3.4 Characterization of natural LPSs/MD-2 interaction**

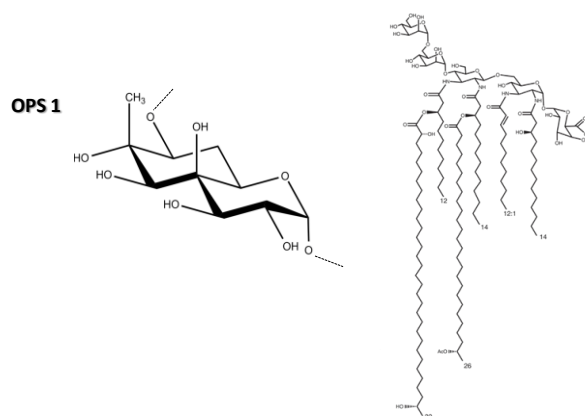
Natural LipoPolySaccharides (LPSs), as amphiphilic molecules, form aggregates in aqueous environments above a critical concentration (CMC). At higher LPS concentrations, the amphiphilic aggregates form nearly spherical

supramolecular aggregate structures, which can be multilamellar or non-lamellar, depending on the physicochemical environment.<sup>35</sup> The aggregation process of LPS from *Escherichia coli* serotype 026:B6 has been described by Santos *et al.*,<sup>36</sup> with an apparent critical micellar concentration (CMCa) value of 14 µg/mL.

### 3.4.1 LPS from *Bradyrhizobium BTAi-1 Δshc*

Rhizobia are Gram-negative bacteria that can establish a symbiotic relationship with legumes.<sup>37</sup> Their abilities to fix nitrogen from natural sources hold promise as a viable alternative to the use of industrial N-fertilizers in agriculture.<sup>38</sup> Furthermore, the lipid A of rhizobial LPS, characterized by a peculiar lipid and sugar composition and by the presence of very long-chain fatty acids (VLCFA), is important for their adaptation to the intracellular life.<sup>39</sup>

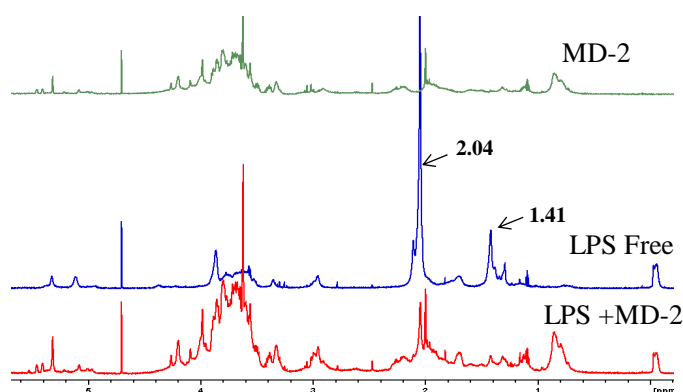
In this chapter, in the framework of our EU project TOLLERANT, we have focused our study on the mutant LPS from *Bradyrhizobium BTAi-1 Δshc*, where the hopanoid biosynthesis pathway have been removed (Figure 3.24).



**Figure 3.24** -. Chemical structure of the LPS components from *Bradyrhizobium BTAi-1 Δshc*.

### 3.4.1.1 NMR experiments

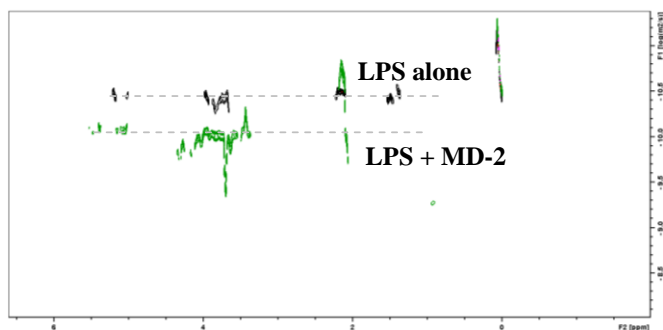
The  $^1\text{H}$ -NMR of LPS from *Bradyrhizobium BTAi-1  $\Delta$ shc* in presence of the accessory MD-2 protein (Figure 3.25) permitted to observe clear perturbation of the signals at 2.04 and 1.41 ppm. However, other  $^1\text{H}$  NMR signals from the could not be analyzed due to extensive overlapping with the signals from hMD-2. The experimentally observed reductions of intensity (Figure 3.25), due to specific line broadening of these signals, probably arise from the changes in the transverse relaxation times of these signals, an indication of the existence of a binding event.



**Figure 3.25** -  $^1\text{H}$ -NMR spectra. **Green:**  $^1\text{H}$ -NMR of 60 $\mu\text{M}$  of hMD-2, **Blue:**  $^1\text{H}$ -NMR of 0.9 mg/mL of LPS from *Bradyrhizobium BTAi-1  $\Delta$ shc*; **Red:**  $^1\text{H}$ -NMR of 0.9 mg/mL of LPS from *Bradyrhizobium BTAi-1  $\Delta$ shc* in presence of 60 $\mu\text{M}$  of MD-2; The spectra were acquired in deuterated phosphate buffer at pH 7.5, 298 K, 64 scans.

We have previously demonstrated, using DOSY measurements, that FP7 forms aggregates or micelles in water solutions, with a small diffusion coefficient value. Fittingly, the interaction of this molecule with MD-2 makes the diffusion coefficient significantly faster, indicating that the initial aggregate is indeed disrupted.<sup>25</sup> The same behavior was observed herein for the LPS from *Bradyrhizobium BTAi-1  $\Delta$ shc* (Figure 3.26 and Table 3.6). The diffusion coefficient increases one order of magnitude, strongly suggesting that the presence of MD-2

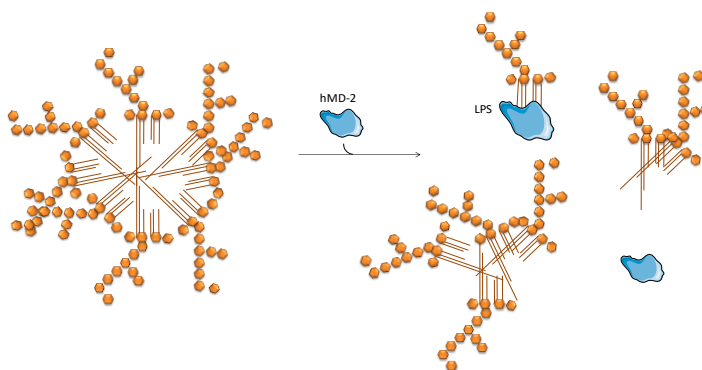
perturbs the aggregation process of the LPS, providing a supramolecular entity of smaller size (Figure 3.27).



**Figure 3.26** - DOSY spectrum **Black**: DOSY of 0.9 mg/mL of LPS from *Bradyrhizobium* BTAi-1  $\Delta$ shc **Green** : DOSY of 0.9 mg/mL of LPS from *Bradyrhizobium* BTAi-1  $\Delta$ shc in presence of 60 $\mu$ M of MD-2; The spectra were acquired in deuterated phosphate buffer at pH 7.5, 298 K, 64 scans, 32 points and  $d_{20} = 500$  ms and  $p_{30} = 5$  ms.

**Table 3.6.** Diffusion coefficient value

	$D/m^2s^{-1}$
0.9 mg/mL LPS from <i>Bradyrhizobium</i> BTAi-1 $\Delta$ shc	$5.781 \times 10^{-12}$
0.9 mg/mL LPS from <i>Bradyrhizobium</i> BTAi-1 $\Delta$ shc + MD-2 60 $\mu$ M	$2.09 \times 10^{-11}$



**Figure 3.27** - Schematic representation of the putative processes that take place before and after addition of MD-2, based on the DOSY experimental data.



A similar behavior has been previously described by Tobias *et. al.*<sup>40</sup> for the interaction of the LPS from *Salmonella minnesota Re595* with the LPS-binding Protein (LBP) and the bactericidal/permeability-increasing protein (BPI) by sedimentation, light scattering, and fluorescence analyses. Shortly, they also reported that the presence LBP promoted the disaggregation of the LPS supramolecular structure.

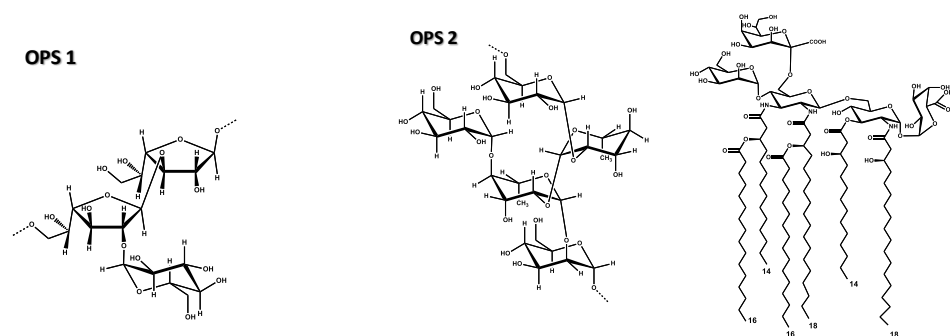
#### **3.4.1.2 Cryo-Electron Microscopy**

It is important to note that the attempt of Cryo-EM with *Salmonella minnesota Re595* LPS, it was impossible to freeze the samples.

#### **3.4.2 LPS from *Acetobacter pasteurianus***

*Acetobacter pasteurianus* is an aerobic Gram-negative bacterium used in the production process of a Japanese black vinegar called *kurozu*. Consumption of this vinegar is believed to carry health benefits as recently recommended as a health drink in Japan.<sup>41</sup> Since oral administration of LPS has been reported to modulate immune responses,<sup>42</sup> the LPS-like components in *kurozu* might be responsible for the beneficial health effects of *kurozu*. However, recent studies with the separated LPS from *A. pasteurianus* showed that it only displays weak TLR4-stimulating activity in comparison with that of the *E. coli* analogue.<sup>43</sup>

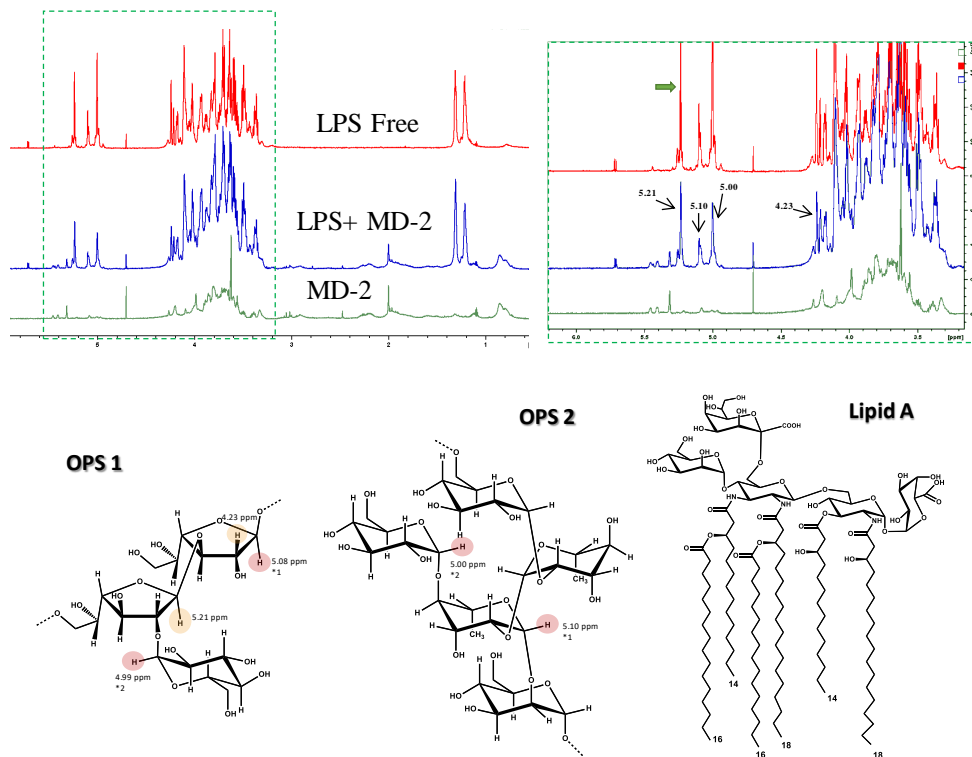
After extraction and isolation, the LPS from *A. pasteurianus*, the molecule was further characterized using extensive analytical methods, including chemical modifications, MS, MALDI, MS/MS, NMR. Finally, the chemical structure of the different components of the molecule were fully characterized (Figure 3.28).



**Figure 3.28** - Structure of the different LPS components of the LPS from *A. pasteurianus*.

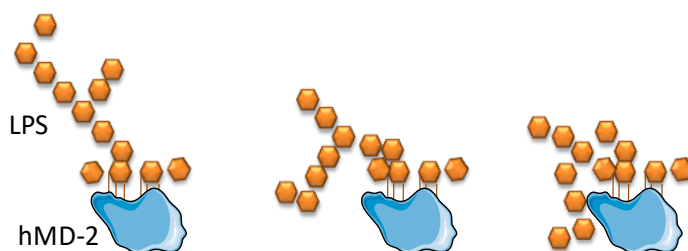
### 3.4.2.1 NMR experiments

The presence of MD-2 provided alterations in the  $^1\text{H}$ -NMR signals of the LPS from *A. pasteurianus* (Figure 3.29), especially at the anomeric proton signals of the OPS part. (Figure 3.29). Unfortunately, it was not possible to obtain non-ambiguous information about the acyl chains, due to the presence of MD-2 signals in this region of the spectrum.



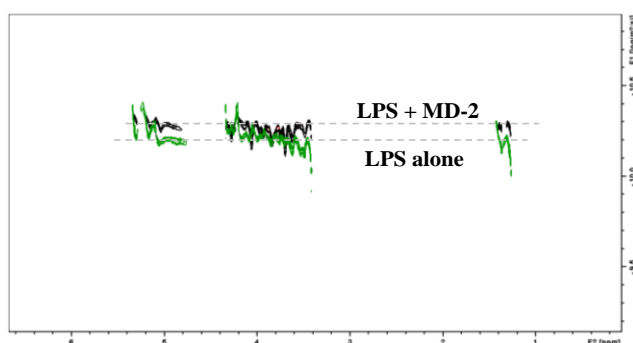
**Figure 3.29 - Up:**  $^1\text{H-NMR}$  spectra. **Green:**  $^1\text{H-NMR}$  of  $60\mu\text{M}$  of MD-2, **Blue:**  $^1\text{H-NMR}$  of  $0.9\text{ mg/mL}$  of LPS from *A. pasteurianus*; **Red:**  $^1\text{H-NMR}$  of  $0.9\text{ mg/mL}$  of LPS from *A. pasteurianus* in presence of  $60\mu\text{M}$  of MD-2; The spectra were acquired in deuterated phosphate buffer at pH 7.5, 298 K, 64 scans. **Down:** “Epitope mapping” where red represents the protons with more attenuation and yellow less perturbation.

From a structural perspective, the OPS domain is rather distant from lipid A fatty acid chains, which are known to act as the binding site for MD-2. However, we cannot discard the hypothesis that the observed perturbation at the anomeric signals is due the recognition of MD-2. The surface of MD-2 has hydrophobic and hydrophilic regions,<sup>44</sup> that might be involved with additional transient interactions with the OPS, as schematized in Figure 3.30.



**Figure 3.30** - Schematic representation of the intrinsic flexibility of the OPS together with the hypothetical contact-points with the MD-2 protein.

The effect of MD-2 at *A. pasteurianus* LPS by DOSY experiments was different to that observed for the LPS from *Bradyrhizobium BTAi-1 Δshc*. In this case, the addition of MD-2 also induced clear perturbations of the diffusion coefficient of the LPS. However, now the diffusion coefficient decreased one order of magnitude, in contrast with the observations for the LPS from *Bradyrhizobium BTAi-1 Δshc*. This behavior is associated with an increase in size or a change in the shape of the supramolecular structure. (Figure 3.31 and Table 3.7).



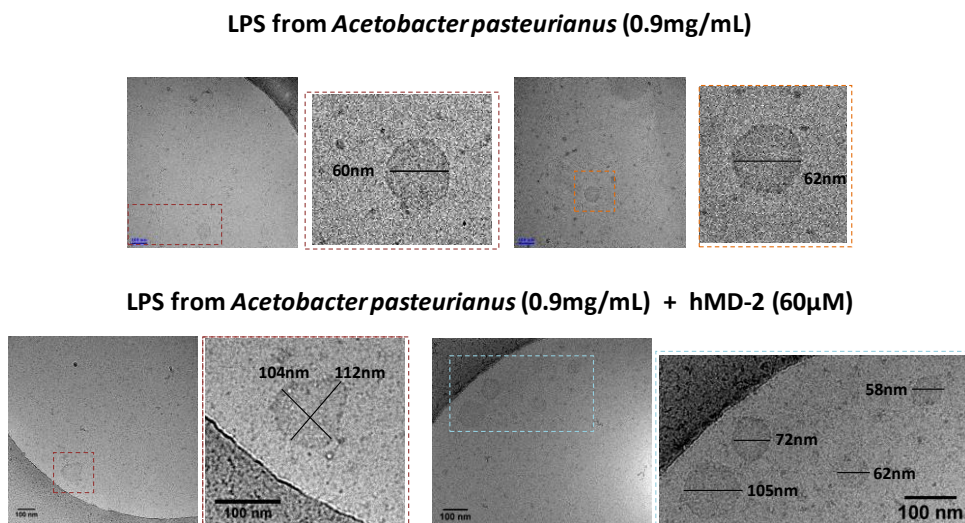
**Figure 3.31** - DOSY spectrum **Black**: DOSY of 0.9 mg/mL of LPS from *A. pasteurianus* **Green** : DOSY of 0.9 mg/mL of *A. pasteurianus* in presence of 60 $\mu$ M of MD-2; The spectra were acquired in deuterated phosphate buffer at pH 7.5, 298 K, 64 scans, 32 points and  $d_{20} = 300$  ms and  $p_{30} = 4$  ms.

**Table 3.7 - Diffusion coefficient value**

	$D/m^2s^{-1}$
0.9 mg/mL LPS from <i>A. pasteurianus</i>	$1.064 \times 10^{-11}$
0.9 mg/mL LPS from <i>A. pasteurianus</i> + MD-2 60uM	$5.175 \times 10^{-12}$

### 3.4.2.2 Cryo-Electron Microscopy

The LPS from *A. pasteurianus*, being an amphipathic molecule, forms aggregates or micelles in water solution (Figure 3.32 up). The LPS from *A. pasteurianus* generates mainly circular and homogeneous micelles with a highly conserved size distribution close to 60 nm. Upon addition of MD-2, some micelles increased the size, while others lost their circular form. In fact, the results obtained with Cryo-TEM are in agreement with the reduction in the diffusion coefficient value observed in the DOSY experiments described above.



**Figure 3.32 - Cryogenic Transmission Electron Microscopy (Cryo-TEM) of LPS from *A. pasteurianus* (0.9mg/mL) nominal magnification of 40,000 X (0.26 nm/pixel).**

Therefore, according to the NMR data, both LPSs (from *Bradyrhizobium BTAi-1 Δshc* and *A. pasteurianus*) interact with MD-2 protein. However, the DOSY experiments also showed that the two LPS molecules show a different behavior in the presence of MD-2. This different performance could be due to the different structural features of the two molecules. The LPS from *Bradyrhizobium BTAi-1 Δshc* possesses two very long-chain fatty acids (VLCFA), namely 32:0 (2,31-2OH) and 26:0 (25-OAc) (rest 2 times 14:0, 12:0, 12:1), whereas the LPS from *A. pasteurianus* does not (18:0 (3-OH), 16:0, 14:0(3-OH), 14:0). This LPS from *Bradyrhizobium BTAi-1 Δshc* showed the smaller diffusion coefficient ( $5.781 \times 10^{-12} \text{ m}^2/\text{s}$ ), indicating the presence of much larger aggregates than in the case of the LPS from *A. pasteurianus* ( $1.064 \times 10^{-11} \text{ m}^2/\text{s}$ ). Only in this case, MD-2 is able to disrupt the supramolecular aggregates.

### **3.5 Conclusions**

In this chapter, we have described the characterization of different features of the interaction of glycolipids with TOLL-accessory proteins.

The interaction of the FP7 glycolipid, a TLR4 antagonist, with known antimicrobial peptides (CA(1-8)M(1-18)<sup>26</sup> and LL37<sup>30</sup>) showed that this co-administration enhances the antagonist potency observed for isolated FP7. Thus, in pathologies where inflammation is exacerbated by bacterial infection, the combination of anti-TLR small molecules with AMPs, as discussed here for the TLR4 antagonist FP7 and peptides CA(1-8)M(1-18) and LL-37, may become a valuable and innovative therapeutic approach.

The characterization of the aggregation of the different glycolipids, natural or mimetic, showed that each molecule displays very different types of aggregation, which are strongly related to the different chemical features of the molecules.

Finally, the interaction of natural LPS (from *Bradyrhizobium BTAi-1 Δshc* and *A. pasteurianus*) with MD-2 showed that both LPSs interact with the protein.

However, the DOSY experiments and Cryo-EM showed that these two LPS present a different behavior in presence of MD-2. The LPS from *Bradyrhizobium BTAi-1*  $\Delta shc$  in presence of the protein underwent disaggregation. In contrast, this is not the case for the LPS from *A. pasteurianus*.

## **3.6 Methods**

### **3.6.1 NMR experiments**

All NMR experiments were recorded on a Bruker Avance III 800 MHz spectrometer equipped with a TCI cryoprobe and Bruker Avance III 600 MHz spectrometer equipped with a TBI probe.

The  $^1\text{H}$  NMR resonances of the peptides (CA(1-8)M(1-18) and LL37) were characterized through 2D-TOCSY (75 ms mixing time) and 2D-NOESY experiments (300 ms mixing time). The concentration of the compounds was set to 500  $\mu\text{M}$  (LL37) and 300  $\mu\text{M}$  (CA(1-8)M(1-18)) in perdeuterated PBS 100  $\mu\text{M}$  in  $\text{H}_2\text{O}/\text{D}_2\text{O}$  90:10 with 10% DMSO, uncorrected pH meter reading 5.5. The peptide characterization was accomplished either at 293 K. The resonance of 2,2,3,3-tetradeutero-3-trimethylsilylpropionic acid (TSP) was used as a chemical shift reference in the  $^1\text{H}$  NMR experiments ( $\delta$  TSP = 0 ppm). Peak lists for the 2D-TOCSY and 2D-NOESY spectra were generated by interactive peak picking using the CARA software.<sup>45</sup>

The DOSY spectra of FP7 were recorded at 310 K with the `tdDOSYccbp.2D` pulse sequence by acquisition of 256 scans, with a diffusion time of 300 ms, a gradient length of 2 ms, and a gradient ramp from 5 % to 95 % in 16 linear steps. Additions of the peptides to the solution were then performed and new DOSY spectra recorded up to a molar ratio of  $[\text{CA(1-8)M(1-18)}]/[\text{FP7}]=0.06$  and  $[\text{LL37}]/[\text{FP7}]=0.1$ .

The DOSY spectra of the isolated CA(1-8)M(1-18) and LL37 peptides as blank were recorded at 310 K by acquisition of 128 scans, with a diffusion time of 250 ms, a gradient length of 1.5 ms, and a gradient ramp from 5 % to 95 % in 16 linear steps. Additions of the FP7 to the solution were then performed and new DOSY spectra recorded up to a molar ratio of  $[\text{FP7}]/[\text{CA(1-8)M(1-18)}]=0.667$  and



[FP7]/[LL37]=0.667. FP7 samples were prepared by diluting the stock solution of FP7 (50 mM in DMSO) with the PBS buffer 100 mM pH=5.5, with a final 10 % DMSO ratio. Peptide samples were prepared by dissolving the solid molecules in DMSO (20 mM stock solution).

The  $^1\text{H}$  NMR experiments were recorded with an concentration of the compounds was set to 0.9 mg/mL (LPS from *Acetobacter pasteurianus*) and 1 mg/mL (LPS from *Bradyrhizobium BTAi-1  $\Delta$ shc*) in perdeuterated PBS 50 mM in D<sub>2</sub>O, uncorrected pH meter reading 7.5, with 64 scans.

DOSY spectra of LPS were recorded at 298 K with the ledbpgppr2s pulse sequence by acquisition of 64 scans, with a diffusion time of 500 ms (for LPS from *Acetobacter pasteurianus*) or 300 ms (for LPS from *Acetobacter pasteurianus*), a gradient length of 4 ms (in both cases), and a gradient ramp from 5% to 95 % in 32 linear steps. Was used the same conditions for acquisition of the sample in presence of MD-2.

### **3.6.2 Expression and Purification of MD-2 protein**

MD-2 was expressed in *Pichia pastoris*, in the laboratory of Prof. Peri by Lenny Zaffaroni (University of Milano-Bicocca, Milano, Italy) as previously described.<sup>46</sup> The MD-2 was purified by affinity chromatography (TALON Cobalt). A 0.5 M solution of Tris HCl pH 7.5 and 1.5 M NaCl was added to the medium to a final concentration of 50 mM Tris HCl and 150 mM NaCl. The *Pichia* medium containing MD-2 was incubated with 5 mL of TALON Cobalt resin (Clontech) previously equilibrated with the 50 mM of Tris buffer at pH 7.5, 150 mM NaCl overnight at 4 °C with slow agitation. The resin was washed with 50 mL of 50 mM of Tris buffer at pH 7.5, 150 mM NaCl. MD-2 and eluted with 1 M imidazole in 1.5 mL fractions, which were analyzed for protein concentration and by SDS-PAGE. The buffer of fractions containing MD-2 were was exchange by means of several cycles of dilution-ultrafiltration (Vivaspin Turbo 15, 3.000 Dalton

molecular weight cut-off, Sartorius Stedim Biotech, Germany) to 50 mM PBS in deuterated water, at pH 7.5. The purity of the protein was confirmed with SDS-PAGE electrophoresis and the yield of the purification was approximately 3 mg per litter of initial culture.

### **3.6.3 Transmission Electron Microscopy**

The samples of FP7 and the antimicrobial peptides were prepared with 90 % of PBS 100mM in H<sub>2</sub>O and 10 % of DMSO, and 16% DMSO in the cases of FP15 and FP18. Negative staining samples were applied to glow-discharged carbon-coated copper grids and stained with 2 % (w/v) NANOVAN. Digital micrographs were taken at room temperature in low dose mode radiation on a Jeol transmission electron microscope operated at 100 kV and equipped with an orius camera. For cryo-microscopy studies the samples were vitrified on Quantifoil 2/2 grids, using vitrobot (FEI) and were analyzed at nitrogen liquid temperature with a TEM operated at 200 kV in low-dose conditions. Micrographs were taken at low radiation dose on a JEM-2200FS/CR transmission electron microscope JEOL, Japan) operated at 200 kV and equipped with an UltraScan4000 SP (4008x4008 pixels) cooled slow-scan CCD camera (GATAN, UK). The samples of FP7 and the antimicrobial peptides were prepared with 90 % of PBS 100mM in H<sub>2</sub>O and 10 % of DMSO, and 16% DMSO in the cases of FP15 and FP18. And the samples of natural LPSs were prepared with PB (50mM) pH 7.5.

## **3.7 *References***

1. Mogensen, T. H. Pathogen recognition and inflammatory signaling in innate immune defenses. *Clin. Microbiol. Rev.* **22**, 240–273 (2009).
2. Akira, S. Innate immunity and adjuvants Shizuo. *Phil. Trans. R. Soc. B* **366**, 2748–2755 (2011).
3. Hoving, J. C., Wilson, G. J. & Brown, G. D. Signalling C-type lectin receptors, microbial recognition and immunity. *Cell. Microbiol.* **16**, 185–194 (2014).

4. Suresh, R. & Mosser, D. M. Pattern recognition receptors in innate immunity, host defense, and immunopathology. *Adv. Physiol. Educ.* **37**, 284–291 (2013).
5. Akira, S. & Takeda, K. Toll-like receptor signalling. *Nat. Rev. Immunol.* **4**, 499–511 (2004).
6. Miyake, K. Innate immune sensing of pathogens and danger signals by cell surface Toll-like receptors. *Semin. Immunol.* **19**, 3–10 (2007).
7. Medzhitov, R. Toll-like receptors and innate immunity. *Nat. Rev. Immunol.* **1**, 135–145 (2001).
8. Poltorak, A. *et al.* Defective LPS Signaling in C3H / HeJ and C57BL / 10ScCr Mice : Mutations in Tlr4 Gene. *Science*. **282**, 2085–2088 (1998).
9. Zanoni, I. *et al.* CD14 controls the LPS-induced endocytosis of toll-like receptor 4. *Cell* **147**, 868–880 (2011).
10. Park, B. S. & Lee, J.-O. Recognition of lipopolysaccharide pattern by TLR4 complexes. *Exp. & Mol. Med.* **45**, e66 (2013).
11. Yang, J. *et al.* Cellular uptake of exogenous calcineurin B is dependent on TLR4/MD2/CD14 complexes, and CnB is an endogenous ligand of TLR4. *Sci. Rep.* **6**, 24346 (2016).
12. Peri, F. & Piazza, M. Therapeutic targeting of innate immunity with Toll-like receptor 4 (TLR4) antagonists. *Biotechnol. Adv.* **30**, 251–260 (2012).
13. Peri, F. & Calabrese, V. Toll-like receptor 4 (TLR4) modulation by synthetic and natural compounds: an update. *J. Med. Chem.* **57**, 3612–3622 (2014).
14. Kanzler, H., Barrat, F. J., Hessel, E. M. & Coffman, R. L. Therapeutic targeting of innate immunity with Toll-like receptor agonists and antagonists. *Nat. Med.* **13**, 552–559 (2007).
15. Mata-Haro, V. *et al.* The vaccine adjuvant monophosphoryl lipid A as a TRIF-biased agonist of TLR4. *Science*. **316**, 1628–1632 (2007).
16. Yu, L., Tan, M., Ho, B., Ding, J. L. & Wohland, T. Determination of critical micelle concentrations and aggregation numbers by fluorescence correlation spectroscopy : Aggregation of a lipopolysaccharide. *Anal. Chim. Acta* **556**, 216–225 (2006).
17. Schromm, A. B. *et al.* Physicochemical and Biological Analysis of Synthetic Bacterial Lipopeptides. *J. Biol. Chem.* **282**, 11030–11037 (2007).
18. Gutsmann, T., Schromm, A. B. & Brandenburg, K. The physicochemistry of endotoxins in relation to bioactivity. *Int. J. Med. Microbiol.* **297**, 341–352 (2007).
19. Mueller, M. *et al.* Aggregates Are the Biologically Active Units of Endotoxin \*. *J. Biol. Chem.* **279**, 26307–26313 (2004).
20. Gioannini, T. L. *et al.* Isolation of an endotoxin – MD-2 complex that produces Toll-like receptor 4-dependent cell activation at picomolar concentrations. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4186–4191 (2004).
21. Park, B. S. *et al.* The structural basis of lipopolysaccharide recognition by the

- TLR4-MD-2 complex. *Nature* **458**, 1191–1195 (2009).
22. Teghanemt, A., Widstrom, R. L., Giannini, T. L. & Weiss, J. P. Isolation of monomeric and dimeric secreted MD-2. Endotoxin.sCD14 and Toll-like receptor 4 ectodomain selectively react with the monomeric form of secreted MD-2. *J. Biol. Chem.* **283**, 21881–21889 (2008).
  23. Jerala, R. Structural biology of the LPS recognition. *Int. J. Med. Microbiol.* **297**, 353–363 (2007).
  24. Kim, H. M. *et al.* Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran. *Cell* **130**, 906–917 (2007).
  25. Cighetti, R. *et al.* Modulation of CD14 and TLR4-MD-2 activities by a synthetic lipid A mimetic. *ChemBioChem* **15**, 250–258 (2014).
  26. Wade, D. *et al.* All-D amino acid-containing channel-forming antibiotic peptides. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4761–4765 (1990).
  27. Andreu, D. *et al.* Shortened cecropin A-melittin hybrids Significant size reduction retains potent antibiotic activity. *FEBS Lett.* **296**, 190–194 (1992).
  28. Fernández-Reyes, M. *et al.* Lysine N-trimethylation, a tool for improving the selectivity of antimicrobial peptides. *J. Med. Chem.* **53**, 5587–5596 (2010).
  29. Chicharro, C., Granata, C., Lozano, R., Andreu, D. & Rivas, L. N-Terminal Fatty Acid Substitution Increases the Leishmanicidal Activity of CA(1-7)M(2-9), a Cecropin-Melittin Hybrid Peptide. *Antimicrob. Agents Chemother.* **45**, 2441–2449 (2001).
  30. Gudmundsson, G. H. *et al.* Structure of the gene for porcine peptide antibiotic PR-39, a cathelin gene family member: comparative mapping of the locus for the human peptide antibiotic FALL-39. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 7085–7089 (1995).
  31. Unitt, J. & Hornigold, D. Plant lectins are novel Toll-like receptor agonists. *Biochem. Pharmacol.* **81**, 1324–1328 (2011).
  32. Perrin-Cocon, L. *et al.* TLR4 antagonist FP7 inhibits LPS-induced cytokine production and glycolytic reprogramming in dendritic cells, and protects mice from lethal influenza infection. *Sci. Rep.* **7**, 1–13 (2017).
  33. Talsma, H., Jousma, H., Nicolay, K. & Crommelin, D. J. A. Multilamellar or multivesicular vesicles? *Int. J. Pharm.* **37**, 171–173 (1987).
  34. Kumar, S., Awang, M. B., Abbas, G. & Kalwar, S. A. Wormlike Micellar Solution: Alternate of Polymeric Mobility Control Agent for Chemical EOR. *J. Appl. Sci.* **14**, 1023–1029 (2014).
  35. García-Verdugo, I., Sánchez-Barbero, F., Soldau, K., Tobias, P. S. & Casals, C. Interaction of SP-A (surfactant protein A) with bacterial rough lipopolysaccharide (Re-LPS), and effects of SP-A on the binding of Re-LPS to CD14 and LPS-binding protein. *Biochem. J.* **391**, 115–124 (2005).
  36. Santos, N. C., Silva, A. C., Castanho, M. A. R. B., Martins-Silva, J. & Saldanha, C. Evaluation of Lipopolysaccharide aggregation by light scattering spectroscopy.

- ChemBioChem* **4**, 96–100 (2003).
37. Oldroyd, G. E. D., Murray, J. D., Poole, P. S. & Downie, J. A. The Rules of Engagement in the Legume-Rhizobial Symbiosis. *Annu. Rev. Genet.* **45**, 119–144 (2011).
  38. Olivares, J., Bedmar, E. J. & Sanjuán, J. Biological Nitrogen Fixation in the context of Global Change. *Mol. Plant Microbe Interact.* **26**, 486–494 (2013).
  39. Brown, D. B., Huang, Y. C., Kannenberg, E. L., Sherrier, D. J. & Carlson, R. W. An acpXL mutant of *Rhizobium leguminosarum* bv. phaseoli lacks 27-hydroxyoctacosanoic acid in its lipid a and is developmentally delayed during symbiotic infection of the determinate nodulating host plant *Phaseolus vulgaris*. *J. Bacteriol.* **193**, 4766–4778 (2011).
  40. Tobias, P. S. *et al.* CARBOHYDRATES , LIPIDS , AND OTHER NATURAL PRODUCTS : Lipopolysaccharide ( LPS ) -binding Proteins BPI and LBP Form Different Types of Complexes with LPS Lipopolysaccharide ( LPS ) -binding Proteins BPI and LBP Form Different Types of Complexes with LPS \*. *J. Biol. Chem.* **272**, 1–5 (1997).
  41. Hashimoto, M. *et al.* Characterization of outer membrane vesicles of *Acetobacter pasteurianus* NBRC3283. *J. Biosci. Bioeng.* **125**, 425–431 (2018).
  42. Inagawa, H., Kohchi, C. & Soma, G. I. Oral administration of lipopolysaccharides for the prevention of various diseases: Benefit and usefulness. *Anticancer Res.* **31**, 2431–2436 (2011).
  43. Hashimoto, M. *et al.* Characterization of a novel D-Glycero-D-talo-oct-2-ulosonic acid-substituted lipid A moiety in the lipopolysaccharide produced by the acetic acid bacterium *Acetobacter pasteurianus* NBRC 3283. *J. Biol. Chem.* **291**, 21184–21194 (2016).
  44. Ohto, U., Fukase, K., Miyake, K. & Satow, Y. Crystal Structures of Human MD-2 and Its Complex with Antiendotoxic Lipid IVa. *Science.* **316**, 1632–1634 (2007).
  45. Keller, R. *The computer aided resonance assignment tutorial.* Goldau, Switzerland: Cantina Verlag (2004).
  46. Facchini, F. A. *et al.* Structure–Activity Relationship in Monosaccharide-Based Toll-Like Receptor 4 (TLR4) Antagonists. *J. Med. Chem.* **61**, 2895–2909 (2018).



# *Chapter*

# **4**

*Final Remarks*





## **Final Remarks**

Besides the specific conclusions detailed in each of the chapters, the global conclusions are highlighted here:

Within this work, we have provided experimental evidences that the catalytic events mediated by GalNAc-Ts follow induced-fit-mechanism, for which the presence UDP-GalNAc is essential. The long-range glycosylation preference of GalNAc-Ts is based on the existence of a very small flexible linker that provides rotational capacity to the lectin domain. On the other hand, for the short-range glycosylation, there is in the catalytic domain a binding site that is responsible for the binding of the neighboring GalNAc residue. All the GalNAc-Ts (GalNAc-T2, -T3 and -T4) studied herein glycosylate in a highly-ordered form the MUC1 substrate, and the MUC1 TR domains are fully *O*-glycosylated in a stepwise manner. The interplay between the lectin and the catalytic domains is essential for the catalysis of all the glycosylation sites of MUC1.

Additionally, we have also been studying different molecular recognition features related to innate immunity processes. We have characterized the supramolecular structures of a variety of natural and synthetic molecules, TLR4-agonists or -antagonists, by using a combination of NMR and biophysical methods. In particular, the co-administration of the so-called FP7 glycolipid, a TLR4 antagonist, with antimicrobial peptides (CA(1-8)M(1-18) and LL37) enhances the antagonist potency observed for isolated FP7. Moreover, the characterization of the aggregation of different glycolipids have shown the possibility of existence of very different types of aggregation, which are strongly related to the different chemical features of the molecules.



## 4.1 Scientific publications during this dissertation

7. Cochet, F., Facchini, F. A., Zaffaroni, L., Billod, J. M., **Coelho, H.**, Holgado, A., Braun, H., Beyaert, R., Jerala, R., Jimenez-Barbero, J., Martin Santamaria, S., Peri, F. Novel carboxylate-based glycolipids: TLR4 antagonism, MD-2 binding and self-assembly properties *Scientific Reports* 9:919 (DOI: 10.1038/s41598-018-37421-w)
6. Rivas, M.; Daniel, E. J. P.; **Coelho, H.**; Lira-Navarrete, E.; Raich, L.; Compañón, I.; Diniz, A.; Lagartera, L.; Jiménez-Barbero, J.; Clausen, H.; Rovira, C.; Marcelo, F.; Corzana, F.; Gerken, T. A.; Hurtado-Guerrero, R.; *ACS Central Science*, **2018**, *4*, 1274–1290. (DOI: 10.1021/acscentsci.8b00488)
5. Rivas, M.; **Coelho, H.**; Diniz, A.; Lira-Navarrete, E.; Compañón, I.; Jiménez-Barbero, J.; Schjoldager, K. T.; Bennett, E. P.; Vakhrushev, S. Y.; Clausen, H.; Corzana, F.; Marcelo, F.; Hurtado-Guerrero R., Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts. *Chemistry - A European Journal*, **2018**, *24*, 8382-8392. (DOI: 10.1002/chem.201800701) (co-first author)
4. Facchini, F. A.; **Coelho, H.**; Sestito, S. E.; Delgado, S.; Minotti, A.; Andreu, D.; Jiménez-Barbero, J.; Peri, F., Co-administration of Antimicrobial Peptides (AMPs) Enhances Toll-like Receptor 4 (TLR4) Antagonist Activity of a Synthetic Glycolipid. *ChemMedChem* **2017**, *13*, 280-287. (DOI: 10.1002/cmdc.201700694)
3. Rivas, M.; Lira-Navarrete, E.; Daniel, E. J. P.; Compañón, I.; **Coelho, H.**; Diniz, A.; Jiménez-Barbero, J.; Peregrina, J. M.; Clausen, H.; Corzana, F.; Marcelo, F.; Jiménez-Osés, G.; Gerken, T. A.; Hurtado-Guerrero R. The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. *Nature Communications* **2016**, *8*:1959 (DOI: 10.1038/s41467-017-02006-0)
2. Unione, L.; Gimeno, A.; Valverde, P.; Calloni, I.; **Coelho, H.**; Mirabella, S.; Poveda, A.; Arda, A.; Jiménez-Barbero, J. Glycans in Infectious Diseases. A molecular recognition perspective. *Curr Med Chem.* **2017**, *24*, 4057-4080. (Review) (DOI: 10.2174/0929867324666170217093702)
1. Ardá, A.; **Coelho, H.**; Fernández de Toro, B.; Galante, S.; Gimeno, A.; Poveda, A.; Sastre, J.; Unione, L.; Valverde, P.; Cañada, F. J.; Jiménez-Barbero, J. Recent advances in the application of NMR methods to uncover the conformation and recognition features of glycans. *Carbohydr. Chem.* **2016**, *42*, 47–82. (Book Chapter) (DOI: 10.1039/9781782626657-00047)

## ***4.2 Contribution to congresses during this dissertation***

### ***Poster awards***

1. “<sup>19</sup>F-NMR Spectroscopy shows that GalNAc-Ts glycosylation mechanism follows an induced-fit mechanism” VIII Ibero-American NMR Meeting (Lisbon, Portugal, June 26-29, **2018**).

### ***Flash Presentation***

1. “Co-administration of antimicrobial peptides helps in the activity of FP7 glycolipid (TLR4 antagonist)” 29th International Carbohydrate Symposium (ICS2018) (Lisbon, Portugal, July 14 -19, **2018**) (Presented by Helena Coelho)

### ***Poster communications***

11. “Co-administration of antimicrobial peptides helps in the activity of FP7 glycolipid (TLR4 antagonist)” 29th International Carbohydrate Symposium (ICS2018) (Lisbon, Portugal, July 14 -19, **2018**). [Author] (Presented by Helena Coelho)
10. “<sup>19</sup>F-NMR Spectroscopy shows that GalNAc-Ts glycosylation mechanism follows an induced-fit mechanism” 1<sup>st</sup> Meeting of Young Biophysicists (Oeiras, Portugal, July 6, **2018**). [Author] (Presented by Helena Coelho)
9. “<sup>19</sup>F-NMR Spectroscopy shows that GalNAc-Ts glycosylation mechanism follows an induced-fit mechanism” VIII Ibero-American NMR Meeting (Lisbon, Portugal, June 26-29, **2018**). [Author] (Presented by Helena Coelho)
8. “The effect of co-administration of antimicrobial peptides in the activity of synthetic glycolipid FP7 (TLR4 antagonist)” 16th Iberian Peptide Meeting (16EPI) / 4th Chemical Biology Group Meeting (4GEQB), (Barcelona, Spain, February 5-7, **2018**) [Author] (Presented by Helena Coelho)
7. “Elucidation of lectin specificity for MUC1 tumor-associated carbohydrate antigens by NMR and Molecular Modeling” 19th European Carbohydrate Symposium – EUROCARB (Barcelona, Spain, July 2nd -6th **2017**) [Author] (Presented by Helena Coelho)
6. “Elucidation of the key structural elements of MUC1-based glycan antigens for recognition by monoclonal antibodies.” NMR: a tool for biology Xth (Paris, France, January 30th - February 1st, 2017) [Author] (Presented by Helena Coelho)

5. “Deciphering the secrets of GalNAc O-glycosylation: From structure to function in human health & disease” 6th EuCheMS Chemistry Congress (Seville, Spain, September 11-15, **2016**) [**Author**] (Presented by Helena Coelho)
4. “Molecular recognition of MUC1 tumor-associated carbohydrate antigens by human macrophage galactose-type lectin.” I Doctoral Conference of the UPV / EHU (Bilbao, Spain, July 11-12, **2016**). [**Author**] (Presented by Helena Coelho)
3. “Molecular recognition of MUC1 tumor-associated carbohydrate antigens by human macrophage galactose-type lectin.” 8th GERMN / 5th Iberian NMR Meeting (Valencia, Spain, June 27-29, **2016**). [**Author**] (Presented by Helena Coelho)
2. “Molecular Recognition of Tumor-Associated MUC1 Peptides by Monoclonal Antibodies. A Microarray and Saturation Transfer Difference (STD) NMR Spectroscopy Approach.” 1st joint French-Portuguese NMR conference / XXVII GERM conference and the 3rd Portuguese RNRMN meeting (Lisbon, Portugal, April 19-23, **2016**). [**Author**] (Presented by Helena Coelho)
1. “Molecular Recognition Studies of Cancer-Related Monoclonal Antibodies by Microarrays and STD-NMR.” III Biennial Meeting of the Chemical Biology Group / XII Carbohydrate Symposium (Madrid, Spain March 14-16, **2016**). [**Author**] (Presented by Helena Coelho)