TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2018

104

Matilde Teresa de las Rivas González de

Estudio d las preferencias de glicosilación de las pp-GalNAc-Ts: implicación en el desarrollo de inhibidores

Departamento

Bioquímica y Biología Molecular y Celular

Director/es

HURTADO GUERRERO, RAMÓN

ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctora

ESTUDIO D LAS PREFERENCIAS DE GLICOSILACIÓN DE LAS PP-GALNAC-TS: IMPLICACIÓN EN EL DESARROLLO DE INHIBIDORES

Autor

Matilde Teresa de las Rivas González de Garay

Director/es

HURTADO GUERRERO, RAMÓN

UNIVERSIDAD DE ZARAGOZA

Bioquímica y Biología Molecular y Celular

2018

Estudio de las preferencias de glicosilación de las pp-GalNAc-Ts:
implicación en el desarrollo de inhibidores

Memoria presentada por **Matilde Teresa de las Rivas González de Garay**,
licenciada en Química y Bioquímica, para optar al grado de doctor en Ciencias.

Zaragoza, julio de 2018

Instituto Universitario de Investigación
Biocomputación y Física de Sistemas Complejos
Universidad Zaragoza

Bioquímica y Biología Molecular y Celular
UNIVERSIDAD DE ZARAGOZA

D. RAMÓN HURTADO GUERRERO, investigador ARAID del Instituto de Biocomputación y Física de Sistemas Complejos (BIFI) de la Universidad de Zaragoza,

CERTIFICA:

Que la tesis doctoral **Estudio de las preferencias de glicosilación de las ppGalNAc-Ts: implicación en el desarrollo de inhibidores** ha sido realizada, bajo su dirección, por la licenciada MATILDE TERESA DE LAS RIVAS GONZÁLEZ DE GARAY, en el Instituto de Biocomputación y Física de Sistemas Complejos y que reúne, a su juicio, las condiciones requeridas para optar al grado de Doctor en Ciencias por la Universidad de Zaragoza en la modalidad de Tesis por compendio de publicaciones.

Zaragoza, julio de 2018.

La presente tesis doctoral, titulada **Estudio de las preferencias de glicosilación de las ppGalNAc-Ts: implicación en el desarrollo de inhibidores,** se presenta para optar al grado de Doctor en Ciencias de la Universidad de Zaragoza según la modalidad de compendio de publicaciones.

En cumplimiento de los requisitos exigidos para esta modalidad según el Programa del Doctorado en Biología Molecular y Bioquímica (RD 99/2011) y la Escuela de Doctorado de la Universidad de Zaragoza, se adjunta a continuación la referencia completa de los cinco artículos publicados durante la realización de esta tesis doctoral y que constituyen el capítulo de Resultados de la misma:

Artículo I:

Lira-Navarrete, E., **de las Rivas, M**., Compañón, I., Carmen Pallarés, M. C., Kong, Y., Iglesias-Fernández, J., Bernades, G. J. L., Peregrina, J. M., Rovira, C., Bernadó, P., Bruscolini, P., Clausen, H., Lostao, A., Corzana, F. & Hurtado-Guerrero, R. (**2015**). Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein *O*-glycosylation. *Nature Communications*, 6(6937), DOI: 10.1038/ncomms7937.

Artículo II:

de las Rivas, M., Lira-Navarrete, E., James Paul Daniel, E., Compañón, I., Coelho, H., Diniz, A., Jiménez-Barbero, J., Peregrina, J. M., Clausen, H., Corzana, F., Marcelo, F., Jiménez-Osés, F., Gerken, . A. & Hurtado-Guerrero, R. (**2017**). The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. *Nature Communications, 8(1959),* DOI: 10.1038/s41467-017-02006-0.

Artículo III:

Ghirardello, M., **de las Rivas, M.**, Lacetera, A., Delso, I., Lira-Navarrete, E,. Tejero, T., Martín-Santamaría, S., Hurtado-Guerrero, R., & Merino, P. (**2016**). Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-Polar Conjugates, *Chemistry - A European Journal*, 22, 7215-7224.

Artículo IV:

Liu, F., Xu, K., Xu, Z., **de las Rivas, M**., Wang, C., Li, X., Lu, J., Zhou, Y., Delso, I., Merino, P., Hurtado-Guerrero, R., Zhang, Y. & Wu, F. (**2017**). The small molecule luteolin inhibits *N*-acetyl-alpha-galactosaminyltransferases and reduces mucin-type O-glycosylation of amyloid precursor protein. *Journal of*

Biological Chemistry, 292(52), pp. 21304-21319.

Artículo V:

de las Rivas, M., Coelho, H., Diniz, A., Lira-Navarrete, E., Compañón, I., Jiménez-Barberó, J., Schjoldager, K., T., Bennett, E. P., Vakhrushev, S. Y., Clausen, H., Corzana, F., Marcelo, F. & Hurtado-Guerrero, R. (**2018**). Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts. *Chemistry - A European Journal.* DOI: 10.1002/chem.201800701.

Mamá, papá: gracias.

Contenido

ÍNDICE DE FIGURAS Y TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1- VISIÓN GENERAL

1. 1. LA *O*-GLICOSILACIÓN COMO MODIFICACIÓN POSTRASDUCCIONAL

Las modificaciones postraduccionales (MPTs) son las principales responsables del tamaño y de la diversidad funcional del proteoma (Schjoldager y Clausen, 2012; Figura 1).

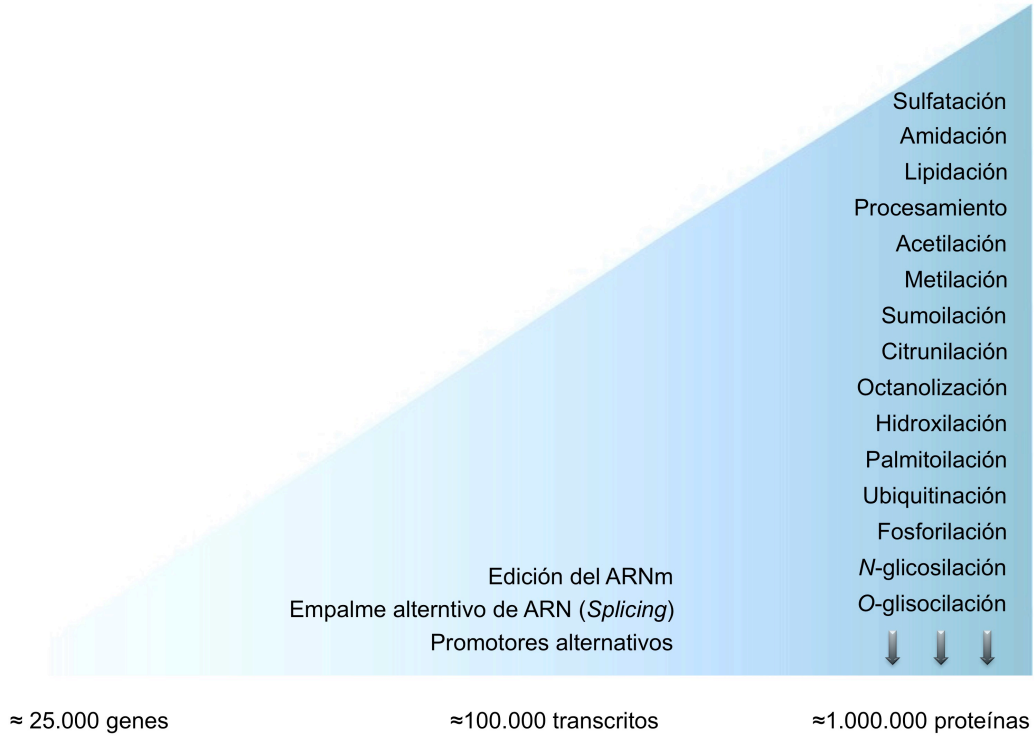


Figura 1.1. Diversificación del genoma humano causado por las modificaciones postraduccionales. La figura ilustra la expansión que se produce desde los genes hasta las proteínas, mediada por las modificaciones en el ARNm y por el empalme alternativo, pero sobre todo por las modificaciones trasduccionales, de las que sólo se listan unas pocas de las más de 200 conocidas. Elaboración propia a partir de Schjoldager y Clausen, 2012.

De las cientos de MPTs que se han detectado, muchas de ellas se han identificado como elementos clave para el mantenimiento de condiciones tanto fisiológicas como patológicas (Ohtsubo y Marth, 2010, Schjoldager y Clausen, 2012). Todas estas MPTs están orquestadas por un gran número de proteasas, quinasas y otras enzimas cuyos mecanismos de acción, a veces redundantes, están lejos de conocerse en profundidad. De entre ellas, la glicosilación —consistente en la adición de carbohidratos o glicanos sobre cadenas proteicas o lípidos— destaca por ser la MPT más compleja, regulada y abundante en el ser humano, con más del 50% del proteoma glicosilado (Apweiler *et al*., 1999).

Las formas más abundates de glicosilación son la *N*-glicosilación sobre residuos de asparagina (Paulson y Colley, 1989) y la *O*-glicosilación sobre residuos de serina o treonina, que también puede darse sobre tirosinas e

hidroxilisinas (Stanley, 2011, Steentoft *et al*., 2011). A excepción de la adición citosólica de *O*-GlcNAc (Holt et al., 1987), todos los tipos de *O*-glicosilación que se dan en humanos (adición de *O*-*N*-acetilgalactosamina, *O*-xilosa, *O*-fucosa, *O*-glucosa, *O*-manosa y *O*-galactosa; Jentoft, 1990) tienen lugar en proteínas de la vía secretora e implican dos pasos fundamentales: uno primero de iniciación, donde las glicosiltransferas encargadas reconocen el motivo en el sustrato aceptor y transfieren el primer monosacárido, y uno segundo de procesamiento, donde diversas enzimas añaden secuencialmente los carbohidratos necesarios para la construcción de las complejas glicoproteínas finales (Joshi, Natsimura, Schjoldager *et al*., 2018). Este proceso tan preciso, secuencial y orquestado que subyace tras la *O*-glicosilación proteica es uno de los más complejos de entre las MPTs, e implica la coordinación y regulación de numerosísimas enzimas (Figura 1.2), capaces de generar glicoconjugados distintos e independientes tanto a nivel estructural como funcional (Joshi, Natsimura, Schjoldager *et al*., 2018; Schjoldager y Clausen, 2012).

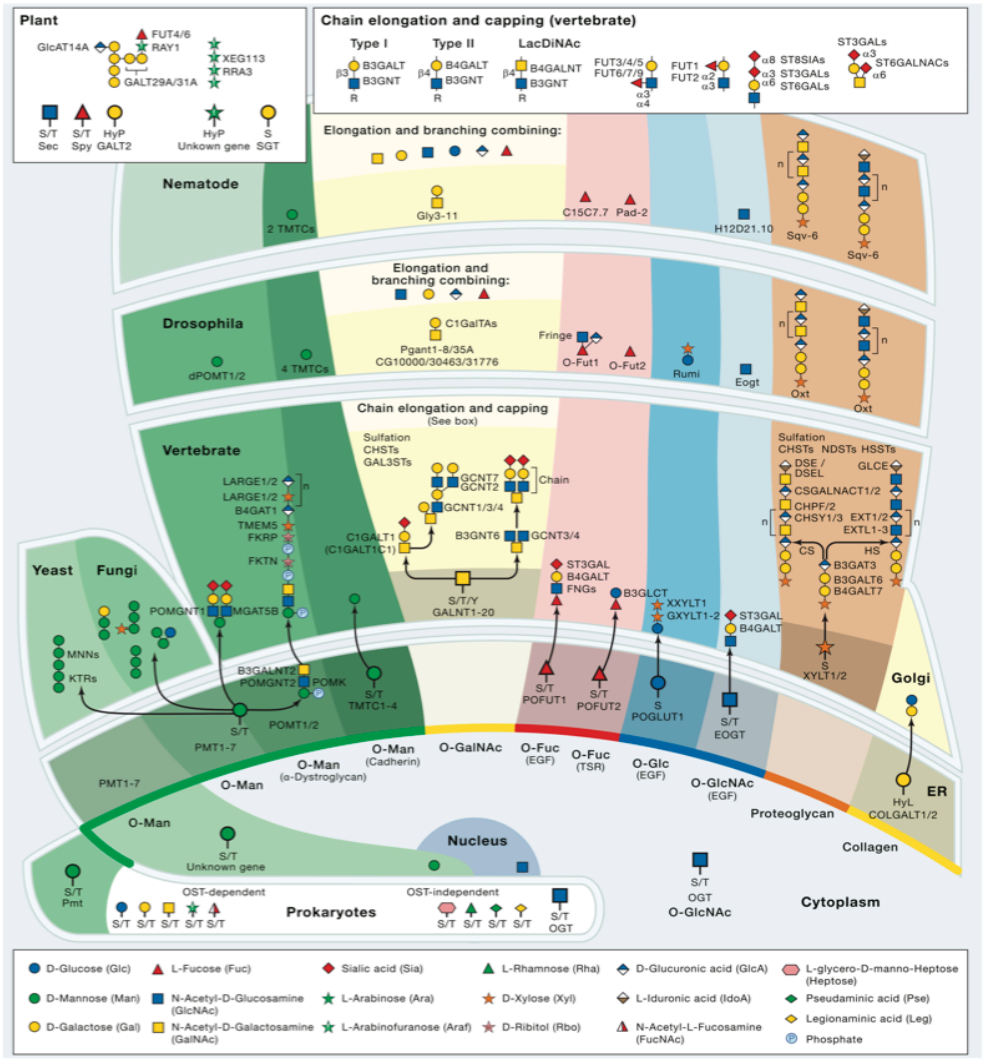


Figura 1.2. Vías de O-glicosilación en los diversos reinos. Fuente: Joshi, Nitsimura, Schjoldager *et al*., 2018.

1. 2. LA GLICOSILACIÓN DE TIPO *O*-GalNAc

Ahora bien, de entre los diferentes tipos de *O*-glicosilación mencionados, la glicosilación de tipo *O*-GalNAc —o tipo mucina— destaca sobre todas las demás por la particular complejidad de su etapa inicial, controlada por un total de 20 isoformas de las UDP-*N*-α–acetil-α-D-galactosamina *N*-acetilgalactosaminiltransferasas polipeptídicas (ppGalNAc-Ts): todas estas enzimas catalizan la primera adición de un grupo GalNAc a un residuo de serina o treonina. Esta adición se realiza de forma asombrosamente orquestada y secuencial por las diferentes isoformas que, pese a sus especificidades parcialmente redundantes, presentan selectividades contextualizadas y patrones diferenciales de actividad y expresión (Bennett *et al*., 2012; Schjoldager y Clausen, 2012; Hurtado-Guerrero, 2016; Ten Hagen *et al*., 2003). Gracias a ello son capaces, entre otras funciones, de originar los dominios densamente glicosilados conocidos como dominios de tipo mucina, con una función primordial en vías de señalización celular y en la defensa de las células de diversos epitelios (Varki *et al*., 2017; Kufe, 2009).

Si bien la deficiencia en alguna de las isoformas de las ppGalNAc-Ts produce fenotipos más sutiles que los encontrados cuando faltan las enzimas iniciadoras del resto de tipos de *O*-glicosilación, de las que no existe tanta variedad (Goreta *et al*., 2012; Joshi *et al*., 2018), se han encontrado recientemente ejemplos evidentes de cómo deficiencias puntuales en algunas de estas enzimas pueden generar patologías concretas (Topaz *et al*., 2004; Duncan *et al*., 2011; Kato *et al*., 2006; Rafaelsen *et al*., 2014; Boskovski *et al*., 2013; Khetarpal *et al*., 2016). Así mismo, numerosos estudios han señalado la clara relación existente entre la desregulación de estas enzimas y diversos procesos tumorales (Prokop y Uhlenbruck, 1969; Kufe, 2009; Thériault *et al*., 2011; Van Putten *et al*., 2017; Bafna *et al*., 2010; Fu *et al*., 2016; Haugstad *et al*., 2012; Chia *et al*., 2016, Sletmoen *et al*., 2018).

En este contexto se plantea esta tesis doctoral. En ella se recogen cinco publicaciones, todas estrechamente relacionadas entre sí y centradas en el estudio y caracterización bioquímica y estructural de diversas isoformas de estas enzimas ppGalNAc-Ts, así como de diversos mutantes de las mismas. Con ellas se pretende:

- a) Ahondar en la justificación molecular del selectivo y regulado comportamiento de algunas de las isoformas más representativas de las ppGalNAc-Ts (Lira-Navarrete, **de las Rivas** *et al*., 2015; **de las Rivas** *et al*., 2017).
- b) Llevar a cabo estudios de orientación más aplicada en los que, desde la comprensión de las bases moleculares subyacentes, se intentan desarrollar moduladores selectivos de la actividad enzimática

(Ghirardello, **de las Rivas** *et al*., 2016; Liu, Xu, Xu, **de las Rivas** *et al*., 2017) y se propone una explicación para la inactivación de una variante patogénica de la *Hs*GalNAc-T2 que conlleva problemas con el metabolismo del colesterol y cardiovasculares (**de las Rivas** *et al*., 2018).

CAPÍTULO 2- INTRODUCCIÓN

2.1. LAS MUCINAS

Las mucinas son proteínas de alto peso molecular caracterizadas por un elevado porcentaje de glicosilación: aproximadamente el 75% de su peso está constituido por carbohidratos, mientras que sólo el 25% restante corresponde a su secuencia polipeptídica (Bansil *et al*., 1995). Descritas por primera vez por Egon Eichwald y Félix Hoppe-Seyler hace más de 150 años (Eichwald, 1865), se encuentran distribuidas por el tejido epitelial de todo el organismo. Son el tipo de glicoproteínas que mayor número de *O*-glicanos contienen (también conocidos como *O*-glicanos de tipo mucina), cuya longitud puede variar desde un solo carbohidrato hasta más de 20 residuos azucarados, entre los que se pueden incluir epítopos como el grupo sanguíneo (Varki *et al*, 2017). De forma general, sus funciones principales están relacionadas con la comunicación celular y con la supervivencia de los epitelios (Kufe, 2009).

2.1.1. Estructura y función

Las mucinas contienen cientos de glicanos heterógeneos derivados de *O*-GalNAc anclados al andamiaje de la proteína o apomucina (Chugh *et al*., 2015), así como un pequeño número de *N*-glicanos. Estos *O*-glicanos se forman en el aparato de Golgi (Rottger *et al*., 1998) a partir de una primera adición de un grupo GalNAc, unido a través de un enlace *O*-glicosídico a una serina o una treonina, y mediada en mamíferos por la familia de las ppGalNAc-Ts. A partir de este primer producto GalNAcα1-*O*-Ser/Thr, al que también se le denomina antígeno Tn, los glicanos se extienden normalmente hasta formar uno de los cuatro núcleos más habituales (Figura 2.1), que posteriormente pueden seguir creciendo lineal o ramificadamente (Joshi, Nitsimura, Schjoladager *et al*., 2018; Figura 1.2).

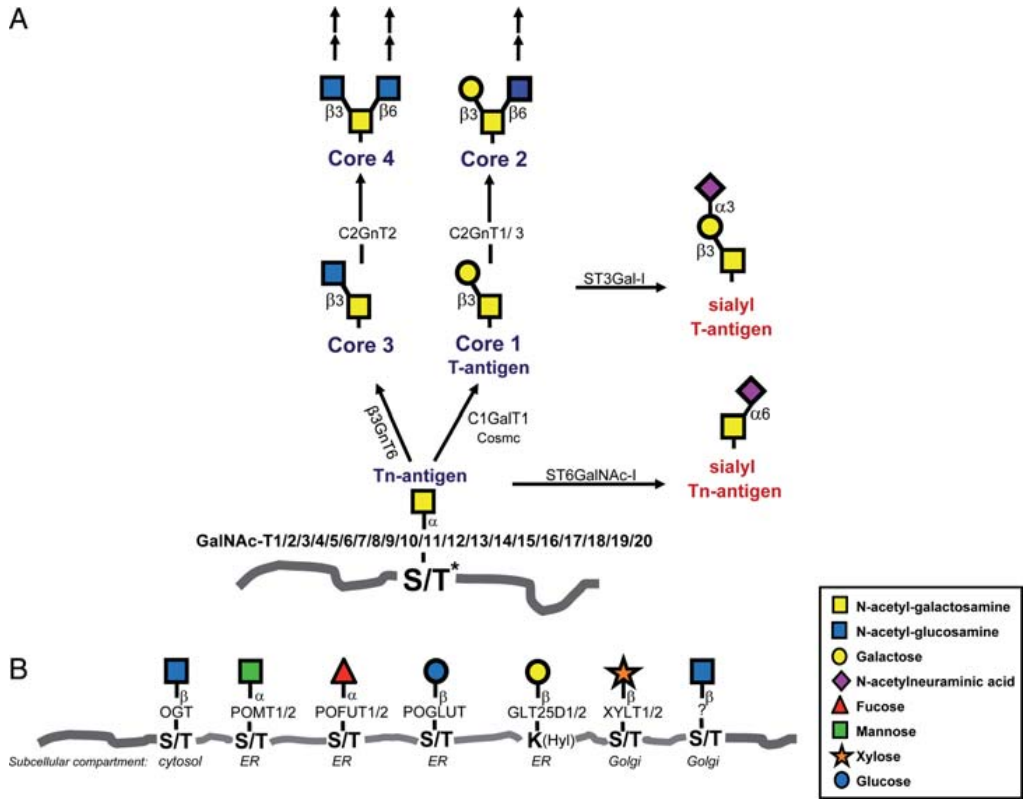


Figura 2.1. Vías de *O*-glicosilación en mamíferos. A. Rutas biosintéticas de los cuatro núcleos de *O*-glicosilación más habituales. B. Principales enzimas iniciadoras de otros tipos de glicosilación en mamíferos. **Fuente**: Benett *et al*., 2012.

Hasta el momento se han identificado 21 genes que codifican proteínas de tipo mucina (MUC). En función de su organización estructural se diferencian dos tipos fundamentales, las secretadas y las transmembrana (Rachagani et al., 2009), ambas implicadas en la homeostasis de las células epiteliales y la función barrera de las mismas (Chugh *et al*., 2015). MUC2, MUC5AC, MUC5B, MUC6, MUC7 y MUC19 son ejemplos de mucinas secretadas, que carecen de dominios transmembrana y tienen un dominio común conocido como el Factor D de von Willebrand (vWF-D), así como un dominio en el extremo C-terminal rico en cisteínas que juega un papel clave en los procesos de oligomerización que les permiten adoptar una consistencia gelificada (Pérez-Vilar y Hill, 1999, y Bansil *et al*., 2013; Figura 2.2).

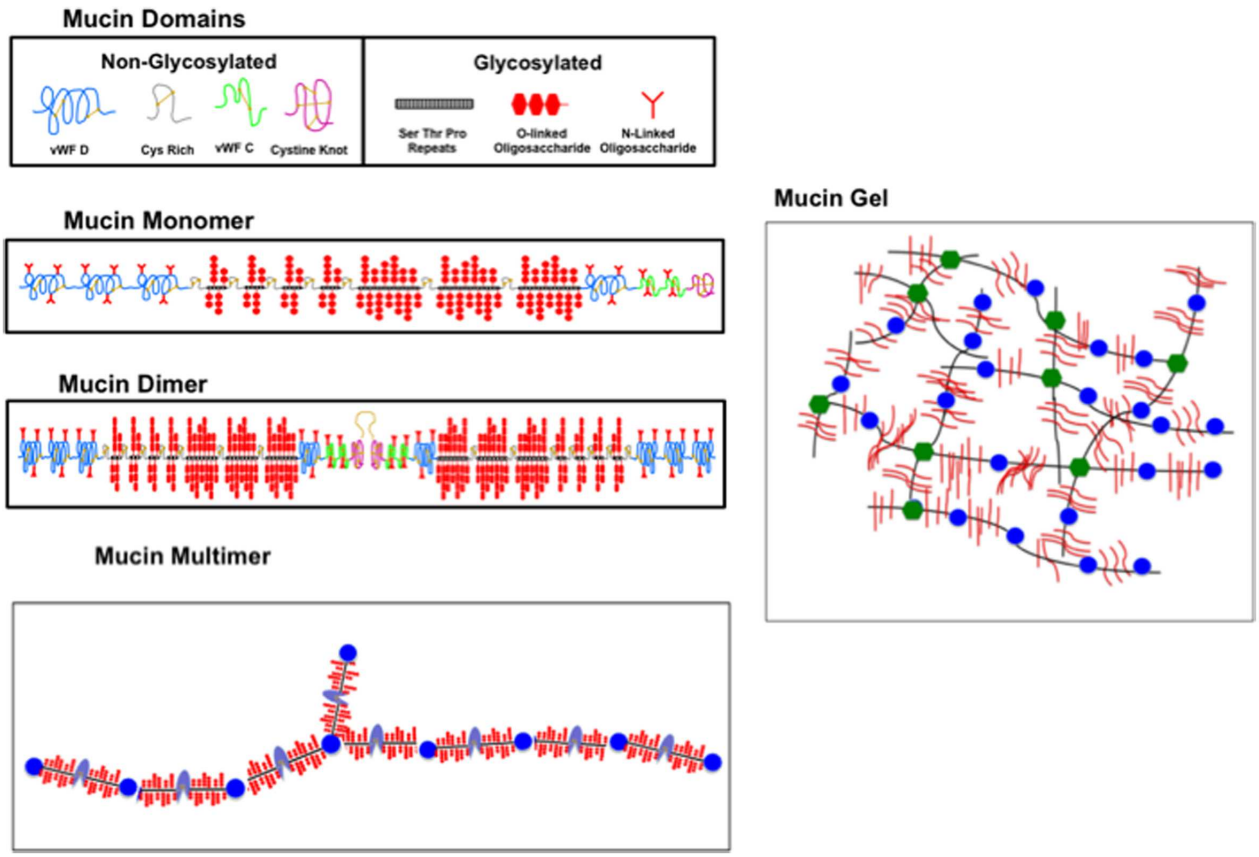


Figura 2.2. **Mucinas secretadas.** Esquema de la estructura, oligomerización y posterior gelificación de una mucina secretada. **Fuente**: Bansil *et al*., 2013.

De entre las mucinas transmembrana destacan MUC1, MUC3 MUC4, MUC12, MUC16 y MUC17, ancladas a la superficie celular a través de sus dominios transmembrana; sus alargados dominios externos híper glicosilados llegan a extenderse hasta 100 nm más allá de la superficie de la célula, lo que a su vez potencia su función defensiva y facilita su implicación en vías de señalización celular (Figura 2.3). La complejidad estructural de estas mucinas, tanto de las secretadas como de las ancladas, se debe a la gran variedad de modificaciones postraduccionales que sufren, tales como sulfatación, fosforilación y, principalmente, glicosilación (Hollingsworth y Swanson, 2004). En concreto, sus ya mencionados y numerosos glicanos de tipo *O*-GalNAc son los responsables de sus propiedades químicas, físicas y biológicas, si bien su variedad, densidad y grado de agrupación hace que sea complejo asignarles funciones específicas (Varki *et al*., 2017).

La zona extracelular de las mucinas, según Kufe (2009) y Hollingsworth y Swanson (2004) se caracteriza por la presencia de tres dominios principales:

- **Dominios PTS**: constituidos por una elevada proporción de prolinas, treoninas y serinas, que les dan el nombre. La existencia de varias repeticiones de este tipo de dominio parece justificar el elevado porcentaje de carbohidratos en estas proteínas, dado que cada una de ellas presenta numerosos sitios susceptibles de glicosilación.
- **Dominios tipo EGF**: Implicados en interacciones de tipo ligando-receptor, se desconoce su función específica en los dominios de mucina, si bien una mala glicosilación de los mismos se asocia directamente con determinados tipos de cáncer (Hollingsworth *et al.*, 2004, y Mukhopadhyay *et al*., 2013).
- **Dominios SEA**: dominios altamente glicosilados de unos 120 aminoácidos, cuyo nombre deriva de las tres proteínas en las que se encontraron, la proteína espermática, la enteroquinasa y la agrina. Si bien su función tampoco era clara, estudios recientes sugieren su papel como un punto clave de resistencia frente a tensiones mecánicas a las que las células puedan ser sometidas (Pelaseyed *et al*., 2013).

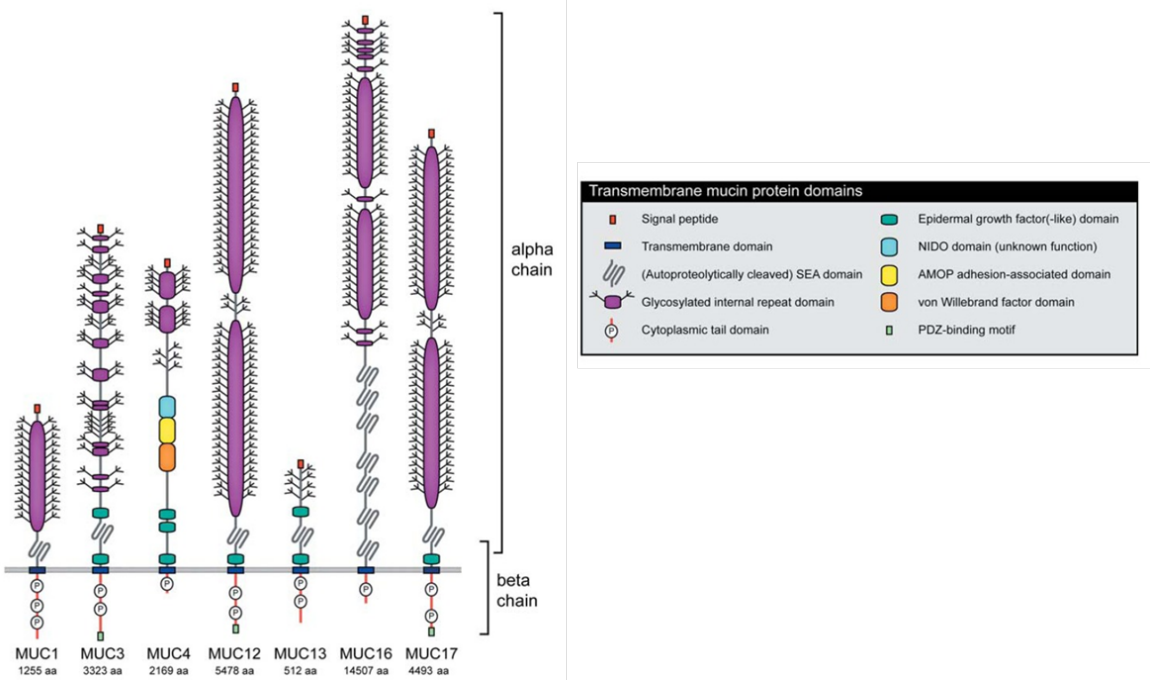


Figura 2.3. Mucinas transmembrana. Esquema de la estructura de las mucinas más representativas del subtipo transmembrana. Modificado de Putten y Strijbis, 2017.

Las mucinas recorren las superficies epiteliales del cuerpo, incluyendo los tractos gastrointestinales, genitourinarios y respiratorios, donde su función principal es la de proteger a las células de los daños físicos y químicos, así como de los agentes microbianos que puedan causar infecciones. Pueden tener

propiedades físicas antiadhesivas y repeler las interacciones con la superficie de la célula o, por el contrario, promover la adhesión a través de la unión de proteínas de unión a carbohidratos que reconocen los *O*-glicanos de su superficie. Como veremos en el siguiente apartado, deficiencias en la correcta glicosilación de todos estos dominios puede conllevar, entre otros hechos, la pérdida de estas propiedades y su consecuente patogenicidad.

2.1.2. Mucinas y patología

Se ha observado una glicosilación aberrante en los dominios de tipo mucina en numerosas enfermedades, al igual que se ha descrito su papel principal en la patogénesis del cáncer (Chugh *et al*., 2015, Van Putten y Strijbis, 2017). Durante la carcinogénesis, nuevas estructuras de glicanos emergen en numerosas glicoproteínas, entre las que se incluyen las mucinas. El origen de estos glicanos atípicos puede ser consecuencia de glicosilaciones incompletas que originan estructuras truncadas (como el antígeno Tn y los derivados sialilados sTn o sT), o la expresión aberrante de cadenas de glicanos (sLE^x, sLE^a) (Padler-Karavani, 2014; Figura 2.4).

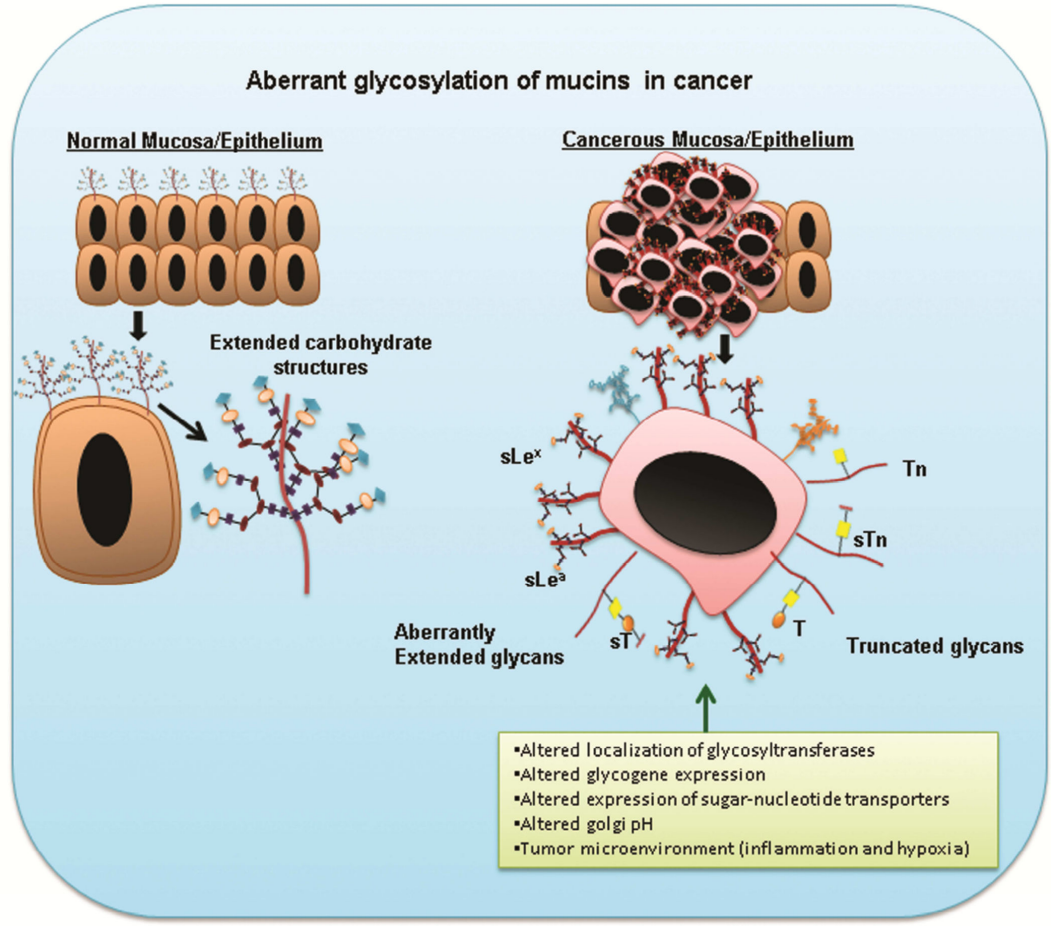


Figura 2.4. Glicosilación en cáncer. Diferencias en la glicosilación de las mucinas en condiciones fisiológicas y tumorales. **Fuente:** Chugh *et al*., 2015.

La expresión de estos antígenos truncados y el resto de glicosilaciones aberrantes en mucinas presentes en células cancerígenas (Figura 2.5) se ha llegado a observar directamente (Song et al., 2015) y está bien documentada en la literatura (Kufe, 2009, Hauselmann y Borsig, 2014; Taylor *et al*., 1999; Deng *et al.,* 2013; Chugh *et al*., 2015, entre otros), habiendo surgido por primera vez con los estudios que demostraron cómo los antígenos T eran mucho más abundantes en las zonas extracelulares de las mucinas de las células con adenocarcinoma de colon y colitis que en las células epiteliales de colon en condiciones fisiológicas (Campbell *et al*., 1995).

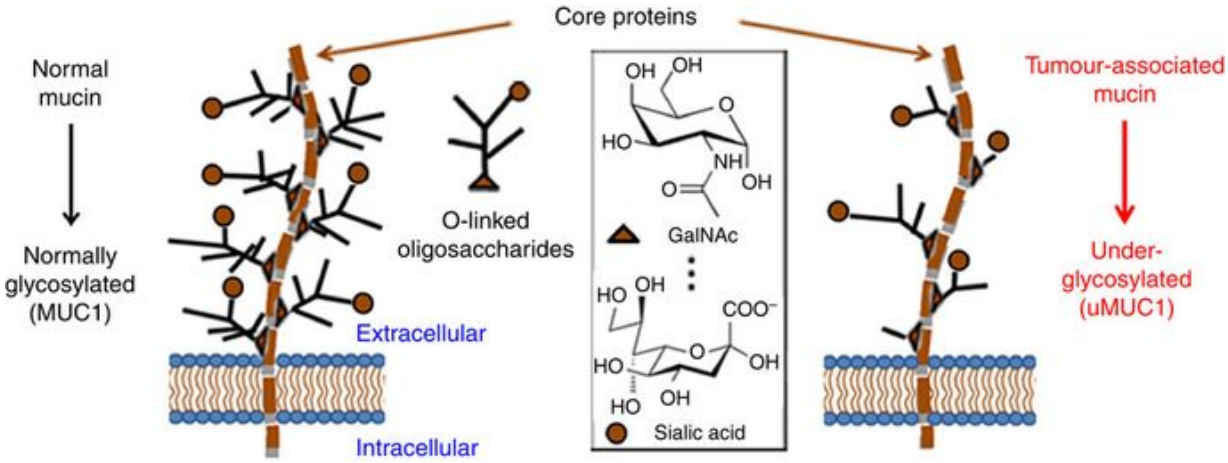


Figura 2.5. Hipoglicosilación asociada a tumores. Comparación de una mucina normal (izquierda) y una asociada a células cancerígenas (derecha), con las consecuentes deficiencias en su glicosilación. **Fuente**: Song et al., 2015.

El papel de estos glicanos aberrantes en la proliferación tumoral se ha descrito abundantemente, y parece estar muy relacionado con la adhesión de las células cancerígenas (Hauselmann y Borsig, 2014), la motilidad de las mismas y su capacidad para invadir e interaccionar de forma anormal con otras células. Son, así mismo, capaces de evadir la respuesta inmune (Madsen *et al*., 2013) y de alterar la capacidad de renovación de las células madre cancerígenas (Ponnusamy *et al*., 2011).

Así por ejemplo, de forma más concreta van Putten y Strijbis (2017) han descrito cómo la glicosilación anormal de la mucosa epitelial del colon conlleva no sólo una pérdida de sus funciones habituales como barrera defensiva, sino también un fomento de las vías de señalización celular que originan procesos inflamatorios y pueden desenvocar en la enfermedad inflamatoria del intestino. Han observado también cómo la sobreexpresión de mucinas transmembrana, asociada con diversos adenocarcinomas, deriva en una pérdida de polarización fisiológica de la membrana y el consecuente aumento de la adhesión celular directamente relacionado con la metástasis (Figura 2.6).

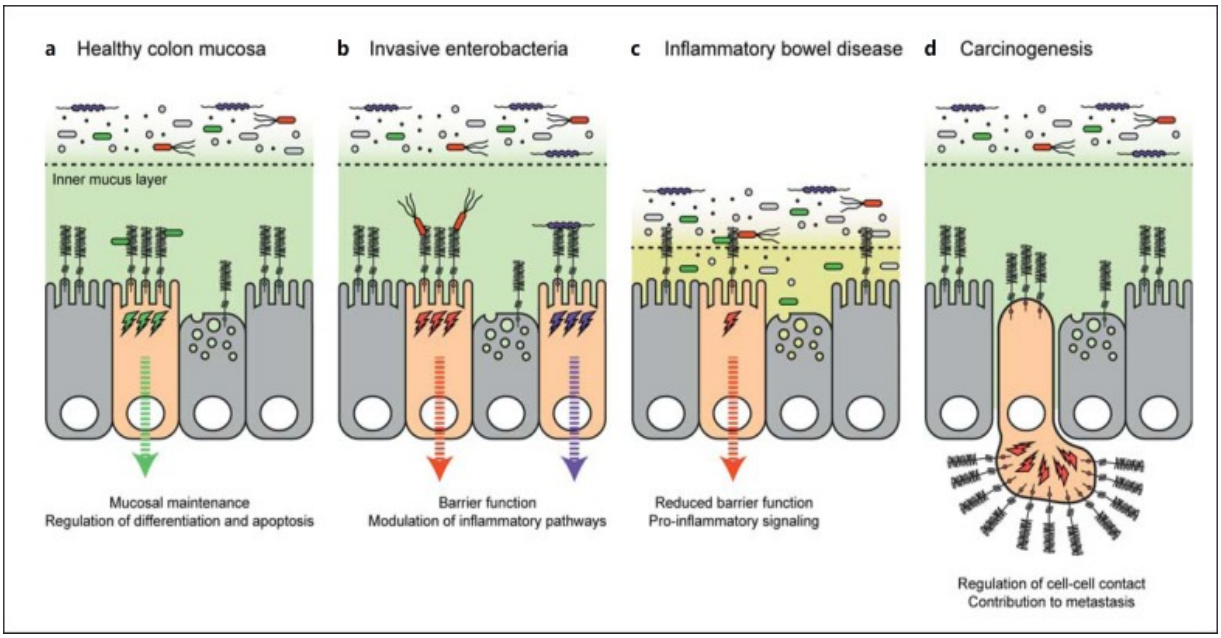


Figura 2.6. Función de las mucinas transmembrana en el colon. **a**. En condiciones fisiológicas, la mucosa de las células epiteliales del colon presenta una capa exterior formada por mucinas y encargada de la protección frente a agentes externos y una zona interior, que se mantiene en condiciones prácticamente estériles y contribuye al mantenimiento de la diferenciación y regulación de las células epiteliales. **b.** Los dominios híper glicosilados de las mucinas de la zona exterior de la mucosa de las células epiteliales del colon son los encargados de limitar las posibles invasiones de agentes patógenos. Sus dominios citosólicos actúan como señalizadores capaces de modular la respuesta inflamatoria. **c**. Deficiencias en la glicosilación de las mucinas transmembrana se asocian con la enfermedad inflamatoria intestinal, debido a la disminución de su función barrera y a un incremento de la señalización pro-inflamatoria frente a la microbiota habitual del lumen intestinal. **d.** La sobreexpresión aberrante de mucinas transmembrana es una de las características más destacadas de adenocarcinomas como el de colon, pecho, páncreas u ovario. La pérdida asociada de polaridad en la membrana celular se vincula directamente con una peor prognosis, dado que las mucinas transmembrana regulan un mayor número de interacciones célula-célula y, por tanto, contribuyen a la metástasis. **Fuente**: van Putten y Strijbis, 2017.

Otro caso de mucina directamente involucrada en la patogénesis del cáncer es MUC1, que promueve el tumor mamario a través de la activación de numerosas vías de señalización celular que fomentan la agresividad del avance tumoral, tales como la interacción con receptores tirosina-quinasa como el EGFR y el ErB2 y la estabilización de β-catenina y Erα. Puede, además, inhibir la apoptosis a través de la regulación de JNK, NF-κB y vías apoptóticas extrínsecas (Wei *et al*., 2005, Agata et al. 2008, Bafna et al., 2010 y Kufe, 2013). De forma paralela, se ha observado cómo la sobreexpresión de esta mucina transmembrana también tiene lugar en numerosas patologías gastrointestinales, tales como el cáncer pancreático (Dang *et al*., 2013 y Hinoda *et al*., 2003), y en el cáncer de ovario (Wang y Fang, 2003).

Son, como vemos, numerosos los ejemplos de mucinas implicadas en la biología patológica del cáncer, entre los que se incluyen también el de pulmón, colon, pecho, útero y páncreas, entre otros (Irimura *et al*., 1999, Matsukita *et al*., 2003). De hecho, se ha evidenciado también cómo MUC4 puede promover la metástasis pancreática y de pecho (Ponnusamy *et al*., 2011 y Chatuverdi *et al*., 2008), y cómo MUC16 juega un importante papel a la hora de proteger a los tumores de la repuesta citotóxicas de las células inmunes NK (Gubbels *et al*., 2010 y Theriault *et al*., 2011). Los ejemplos descritos en la literatura son, así, abrumadores, y de ahí que el ahondar en el conocimiento estructural y mecanístico de las enzimas implicadas en este tipo de *O*-glicosilación, así como en sus entresijos y especificidades, sea de crucial importancia para llegar a entender las vías de actuación de las mismas y desarrollar marcadores, vacunas y tratamientos específicos que nos permitan controlar su actuación aberrante y paliar, en la medida de lo posible, sus efectos patogénicos.

2.2. LA FAMILIA DE LAS ppGALNAc-Ts

Como ya se ha introducido, la *O*-glicosilación de tipo mucina es una modificación postraduccional esencial en organismos superiores (ausente en levaduras, plantas y bacterias). Aunque sólo el 1% del genoma codifica genes implicados en la biosíntesis de esta MPT (Apweiler *et al*., 1999 y Lowe y Marth, 2003), más del 80% de las proteínas que pasan a través de la vía secretora son modificadas por la maquinaria enzimática responsable de la misma (Gill *et al*., 2011, Bennet *et al*., 2012). Comienza con la adición de un azúcar *N*-acetilgalactosamina (GalNAc) al grupo hidroxilo de la cadena lateral de una serina o treonina de la proteína sustrato. Dicha adición, llevada a cabo en el aparato de Golgi en condiciones normales (Rottger *et al*., 1998, Gill *et al*., 2010; Hang & Bertozzi, 2005 y Ten Hagen *et al*., 2003) y, según se ha postulado más recientemente, en el retículo endoplasmático en condiciones tumorales (Gill *et al*., 2010, Chia *et al*., 2016), la realiza algún miembro de la gran familia de ppGalNAc-Ts, clasificadas en el grupo GT27 de la base de datos de enzimas que procesan carbohidratos, CAZy Database (http://www.cayz.org; Lombard *et al*., 2013). El producto de esta reacción, en el que el donador es el UDP-GalNAc, es el ya mencionado antígeno Tn (Hang *et al*., 2005; Gill *et al*., 2016, Chia *et al*., 2016; Figura 2.7).

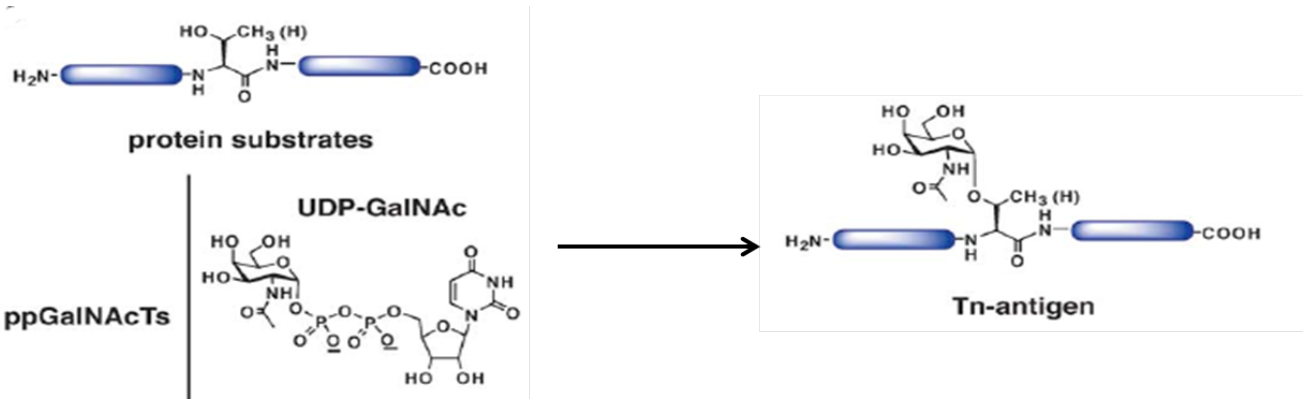


Figura 2.7. Reacción catalizada por las ppGalNAc-Ts. Fuente: Hang *et al*., 2005.

Al igual que sucede en otros organismos, en humanos se encuentran un gran número de isoformas de esta familia —20, en concreto—, si bien el número de genes que las codifican puede variar entre especies (Bennett *et al*., 2012, Ten Hagen *et al*., 2003). No obstante, la familia de las ppGalNAc-Ts se ha visto altamente conservada a lo largo de la evolución, e incluso aunque el número de miembros varíe, todos los genomas de organismos completos contienen familias de secuencias altamente homólogas, que nos indican que tanto la estructura como la función de las ppGalNAc-Ts se ha conservado (Ten Hagen *et al*., 2003, Schwientek *et al*., 2002).

Respecto a los perfiles de expresión de cada una de ellas, se ha visto cómo se encuentran de forma diferencial en función de las células y del tejido, así como muestran importantes cambios en su patrón de expresión en procesos cancerígenos (Bennet *et al*., 2012, Brooks *et al*., 2007). Se ha descrito, de hecho, cómo los genes codificantes para GalNAc-T1 y T2 (*GALNT1* y *T2*, respectivamente) son los que se expresan de forma más ubicua en el organismo (White *et al*., 1995, Bennet *et al*., 2012), mientras que hay otros que sólo lo hacen en ciertos tejidos o bajo unas condiciones fisiológicas determinadas como puede ser la diferenciación celular (Mandel *et al*., 1999, Hollingsworth *et al*., 1999, Raman *et al*., 2011, Perreni *et al*., 2010). A modo de ejemplo, cabe destacar cómo se ha detectado la presencia de un mayor número de ppGalNAc-Ts en procesos cancerígenos como los adenocarcinomas de mama y gástrico (Brooks *et al*., 2007, Gomes *et al*., 2008, Bennet *et al*, 2012). Se ha investigado acerca de cómo esta expresión anormal de diversas isoformas de ppGalNAc-Ts puede ser responsable de la iniciación de la *O*-glicosilación en lugares fisiológicamente no habituales, lo que a su vez puede generar derivados glicosilados de proteínas cuya actividad biológica se vea alterada y pueda contribuir de forma directa o indirecta a la patogénesis del cáncer (Brooks *et al*., 2007, Sletmoen *et al*., 2018).

La actividad de las GalNAcTs se identificó y caracterizó por primera vez en extractos de glándulas submaxilares bovinas y de rata, y se purificó posteriormente a partir de glándulas submaxilares ovinas. Se vio que la actividad enzimática tenía un pH de acción bastante amplio, entre 6,5 y 8,6, y que se requería de la presencia de iones metálicos como cofactores, siendo el óptimo el Mn^{2+} (Sugiura *et al*., 1982, Gerken *et al*., 1989). Gracias al empleo de diversas estrategias y fuentes, se lograron aislar y purificar dos enzimas distintas, la GalNAc-T1 bovina —de la que se encontró un ortólogo en humanos con más del 95% de coincidencia en la secuencia, lo que a su vez confirmó que el genoma humano contenía múltiples genes para ppGalNAc-Ts (White *et al*., 1995)— y la GalNAc-T2 humana. Desde entonces hasta ahora se han localizado numerosos genes de ppGalNAc-Ts en mamíferos y, cuando se inició esta tesis doctoral, se habían aislado y caracterizado prácticamente todas las isoformas, a excepción de T8, T9, T18, T19 y T20 (Li *et al*., 2010, Ten Hagen *et al*., 2003, Clausen *et al*., 1996, Hassan *et al*., 2000a y 2000b, Tian *et al*., 2007, Tabak, 2010, entre otros), además de resolverse la estructura de la GalNAc-T1 (Fritz *et al*., 2004), la GalNAc-T2 (Fritz *et al*., 2006) y la GalNAc-T10 (Kubota *et al.,* 2006).

2.2.1. Características estructurales de las ppGalNAc-Ts

Todas las ppGalNac-Ts comparten la estructura del tipo II de las glicosiltransferasas presentes en el aparato de Golgi, con una pequeña cola orientada hacia el citoplasma en el extremo N-terminal, una secuencia señal hidrofóbica de membrana, una región tipo tallo de longitud variable y un dominio catalítico en el lumen (Paulson y Colley, 1989, Wildt y Gerngross, 2005; Figura 2.8).

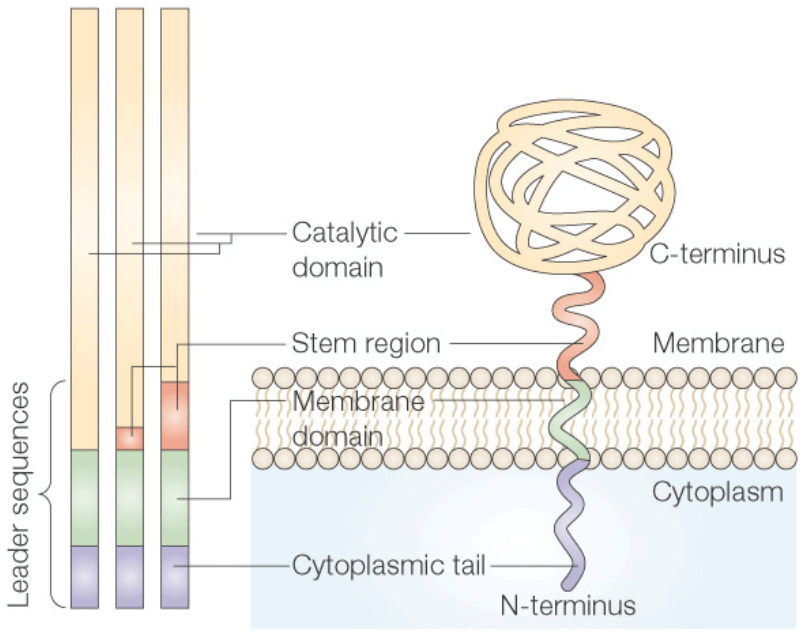


Figura 2.8. Representación esquemática de la estructura de una glicosiltransferasa de tipo II. **Fuente**: Wildt y Gerngross, 2005.

Se diferencian, no obstante, del resto de glicosiltransferasas eucariotas en que son las únicas que contienen un dominio de lectina en el extremo C-terminal de tipo ricina y que contiene alrededor de 120 aminoácidos (Imberty *et al*., 1997)

El dominio catalítico de las ppGalNAc-Ts, de aproximadamente 230 aminoácidos de longitud, contiene un motivo estructural de tipo GT-A caracterizado por dos plegamientos β-α-β de tipo Rossmann en estrecha interacción (Rao y Rossman, 1973, y Hanukoglu, 2015). Este tipo de plegamientos contiene residuos que se unen al uracilo del sustrato UDP-GalNAc (Hagen *et al*., 1999). Cerca del C-terminal del segundo plegamiento de Rossman se encuentra un dominio DxH (siendo x cualquier aminoácido) de unión a Mn^{2+}, característico de las glicosiltransferasas de tipo GT-A, conservado en todas las isoformas conocidas y que interacciona con la parte difosfato del UDP-GalNAc (Bennett *et al.*, 2012). En la parte restante del dominio catalítico, los residuos que se unen al UDP-GalNAc se sitúan o cerca o formando parte de un asa flexible (a la que también nos referiremos por su traducción inglesa *flexible loop*), que a su vez está sujeto a grandes cambios conformacionales en respuesta a la unión de UDP-GalNAc (Lira-Navarrete *et al*., 2014, y Hurtado-Guerrero, 2016). Estudios previos a esta tesis habían demostrado cómo parte de estos residuos forman un "bolsillo de prolina" que justifica la preferencia de algunas de las isoformas de las ppGalNAc-Ts por péptidos aceptores que contienen una prolina situada a 3 residuos de la Ser o la Thr que se glicosila, en la dirección del extremo C-terminal (Gerken *et al*., 2002, Gerken *et al*., 2002, Fritz *et al*., 2004, Fritz *et al*., 2006). De igual modo, la resolución de la estructura cristalográfica de GalNAc-T2 (Fritz *et al*., 2006) reveló un bolsillo adicional en el que se demostró posteriormente (Lira-Navarrete *et al*., 2014) cómo se acomoda en su interior el grupo metilo de la Thr, lo que justifica la preferencia de la enzima, también observada en algunas lectinas de plantas (Madariaga *et al*., 2014, Martínez-Sáez *et al*., 2015), por glicosilar a este residuo en vez de a la Ser.

En cuanto al exclusivo dominio de lectina, con plegamiento estructural de tipo ricina (Hazes, 1996), éste presenta un plegamiento de tipo β-trébol constituido por tres lóbulos homólogos, α, β y γ: tres sitios potenciales de unión a sacáridos de los que, en la mayoría de las ppGalNAc-Ts, sólo uno es funcional (Gerken *et al.,* 2013; Revoredo *et al*., 2016; Pedersen *et al*., 2011). Las características generales de este tipo de dominios son: (i) la presencia de motivos de tipo QxW y CLD de unos 40 aminoácidos en cada repetición; (ii) escasa similitud de secuencia y (iii) amplia diversidad en la especificidad de unión a sustratos monosacáridos (Hazes, 1996). En el caso de las ppGalNAc-Ts, las secuencias de los dominios están escasamente conservadas a excepción de los motivos QxW y CLD (Figura 2.9), así como seis residuos de cisteína invariables que forman puentes disulfuro entre los lóbulos de cada repetición (Gill *et al*., 2011; Ten Hagen *et al*., 2003). Han demostrado tener una alta especificidad por los glicopéptidos con GalNAc, y estudios previos a la realización de esta tesis

evidencian que una de sus funciones principales es la de modular y mejorar la eficiencia catalítica de las GalNAc-Ts frente a sustratos glicopeptídicos que ya presentan una densa *O*-glicosilación, como sucede en los dominios de tipo mucina (Ten Hagan *et al*., 2003; Bennett *et al*., 2012; Wandall *et al*., 2007; Pedersen *et al*., 2011).

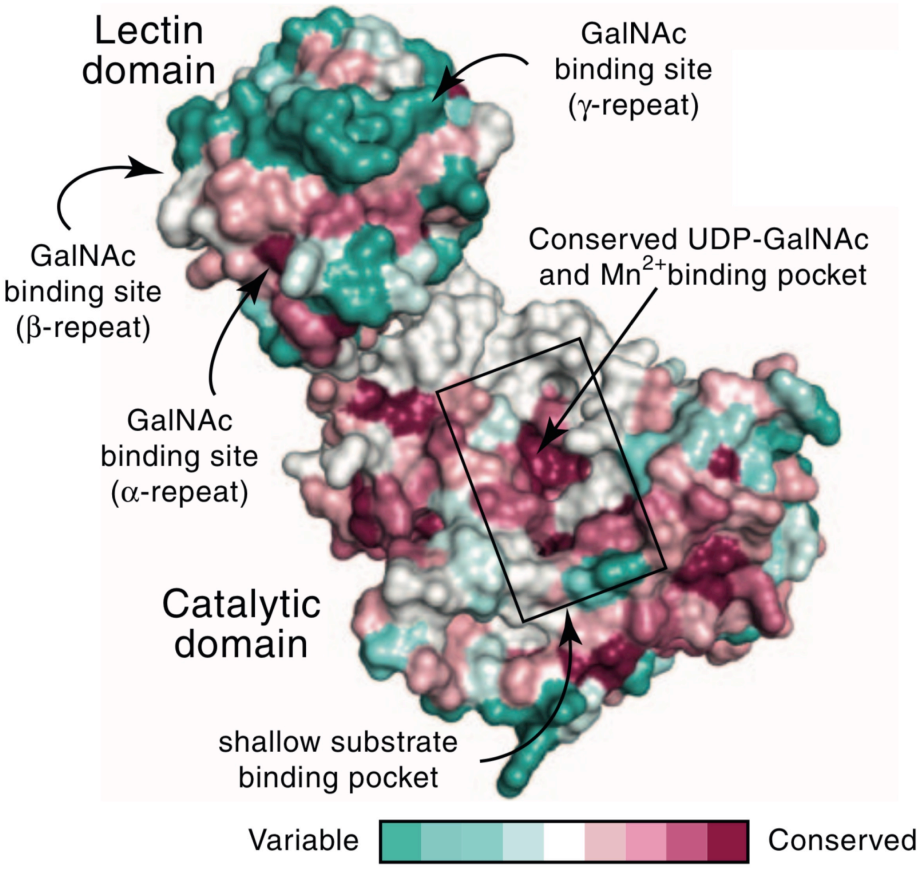


Figura 2.9. Representación de la estructura de la superficia de la *Hs*GalNAc-T2. Sobre ella se muestra el diferente grado de conservación de los residuos que integran el dominio catalítico y de lectina de las diferentes ppGalNAc-Ts. Modificado de Gill *et al*., 2011.

Es interesante mencionar por último cómo los dos dominios, catalítico y de lectina, se encuentran conectados en las ppGalNAc-Ts por un asa flexible (a la que a partir de ahora nos referiremos por su traducción en inglés *flexible linker*) de longitud variable (10-25 aminoácidos) y de elevada variabilidad y flexibilidad. Así, las estructuras resueltas hasta el comienzo de esta tesis en septiembre del 2014 habían mostrado diferencias significativas en la separación existente entre los dominios: mientras en las estructuras de *Mm*GalNAc-T1 y *Hs*GalNAc-T10 el dominio de lectina y el catalítico interaccionan a través de un área superficial considerablemente grande, donde los aminoácidos implicados varían entre las dos isoformas generando diferentes orientaciones relativas, en el complejo de *Hs*GalNAc-T2, UDP y Mn^{2+} resuelto por Fritz *et al*. (2006) se observa una separación clara entre dominios, que se incrementa en presencia del péptido no glicosilado EA2 (Figura 2.10). Éstas y otras observaciones son la base de

nuestra tesis, y parecen apuntar a que la flexibilidad del *flexible linker* podría jugar un papel clave a la hora de controlar la orientación relativa de los dominios catalítico y de lectina a través del cual dirigir parte de la especifidad de las distintas ppGalNAc-Ts. La resolución de la estructura de estas proteínas en presencia de diferentes ligandos, incluyendo un péptido glicosilado, podría dar una información crucial acerca de la flexibilidad estructural de esta familia de glicosiltransferasas.

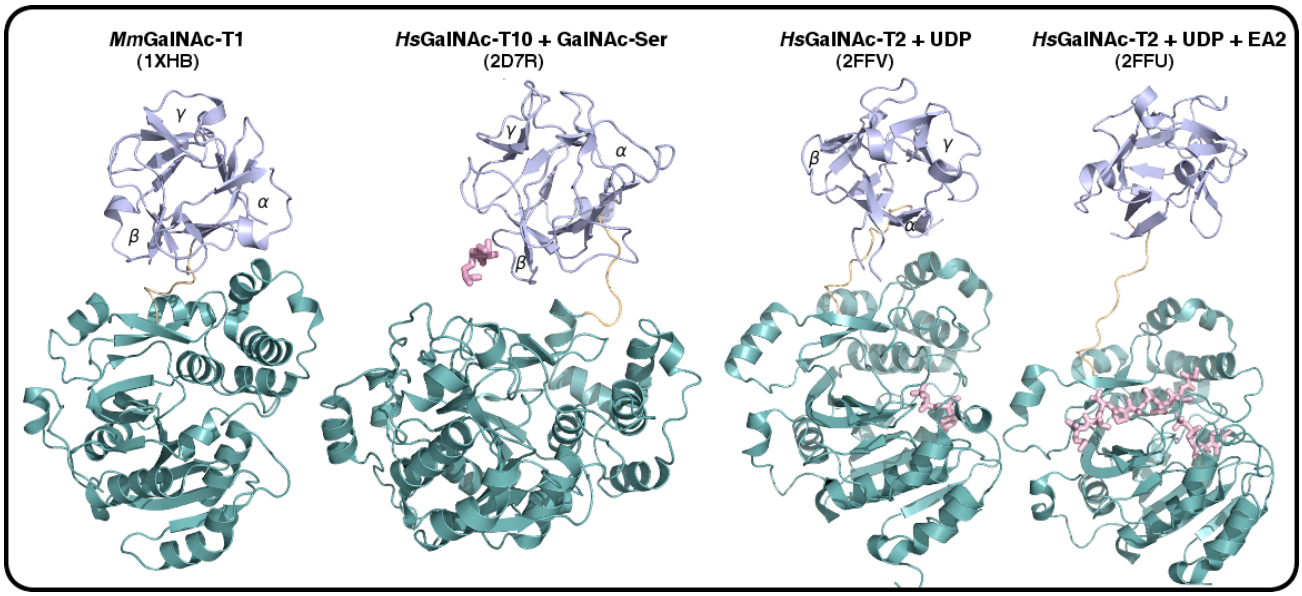


Figura 2.10. Estructuras de las GalNAc-Ts resueltas al inicio de esta tesis (septiembre del 2014). Se muestra en violeta el dominio de lectina, en verde el dominio catalítico, en naranja el linker flexible y en rosa los distintos ligandos. **Fuente:** Lira-Navarrete, 2015.

2.2.2. Características mecanísticas de las ppGalNAc-Ts

A nivel mecanístico, en función de la esteoreoquímica del enlace glicosídico del carbono anomérico del sustrato dador respecto al producto final, se puede dividir al conjunto de glicosiltransferasas en dos grandes clases, independientemente del tipo de plegamiento que adopten: de inversión o de retención (Chang *et al*., 2012, Hurtado-Guerrero y Davies, 2012, Wagner y Pesnot, 2010, Lairson *et al*., 2008). En el caso concreto de las ppGalNAc-Ts, éstas son glicosiltransferasas de retención para las que se propuso inicialmente un mecanismo de doble desplazamiento, con un intermedio en el que se observaría una unión covalente glicosil-enzima (Mosi *et al*., 2002). No obstante, los estudios mecanísticos realizados hasta el momento no han detectado acumulación alguna de dichos intermedios de reacción y, de forma complementaria, posteriores trabajos han demostrado cómo en el caso de la GalNAc-T2 tiene lugar un mecanismo de sustitución nucleófila intramolecular o SNi (Hurtado-Guerrero, 2016, Davies, 2001, Lira-Navarrete *et al*., 2014; Figura 2.11).

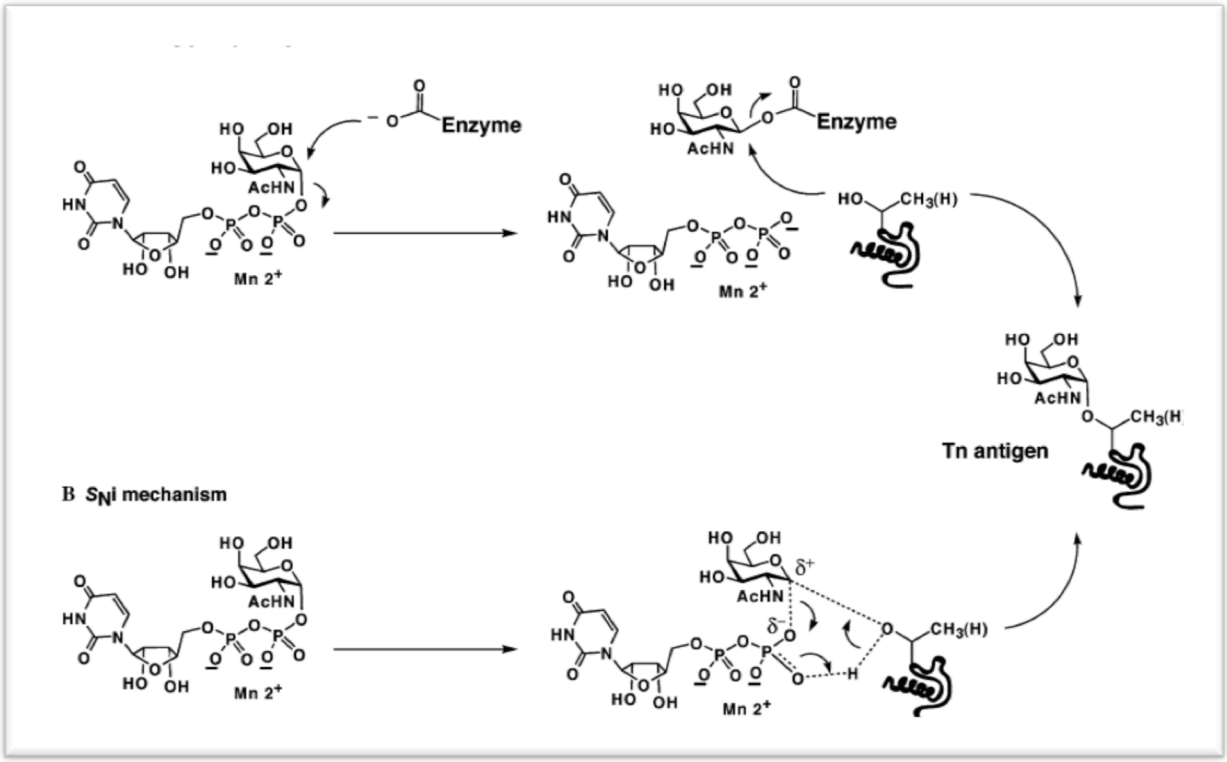


Figura 2.11. Mecanismos de retención. Mecanismos propuestos para la transferencia glicosídica de retención mediada por las ppGalNAc-Ts: formación de un intermediario covalente en A y sustitución nucleófila intramolecular en B. **Fuente:** Hang y Bertozzi, 2005.

Por otro lado, cabe destacar cómo los trabajos realizados en nuestro grupo de forma inmediatamente previa a los recogidos en esta tesis doctoral (Lira-Navarrete *et al*., 2014) demostraron que el nucleótido azucarado —y no el péptido sustrato— era el elemento determinante a la hora de que el asa flexible del dominio catalítico descrita por Fritz *et al.* (2006) para el caso de la GalNAc-T2 se estableciese en una conformación cerrada, compatible con la catálisis. En este mismo estudio, Lira-Navarrete *et al.* (2014) propusieron un mecanismo catalítico secuencial bi-bi ordenado (Figura 2.12) en el que la enzima se encuentra inicialmente en un equilibrio en el que el asa del dominio catalítico oscila entre la forma abierta y cerrada, permitiendo la unión del nucleótido glicosilado y del cofactor y la consecuente adopción de la forma cerrada y activa del asa, la cual, una vez transferido el azúcar al péptido aceptor y formado el consiguiente glicopéptido, se liberaría para permitir la salida del nucleótido.

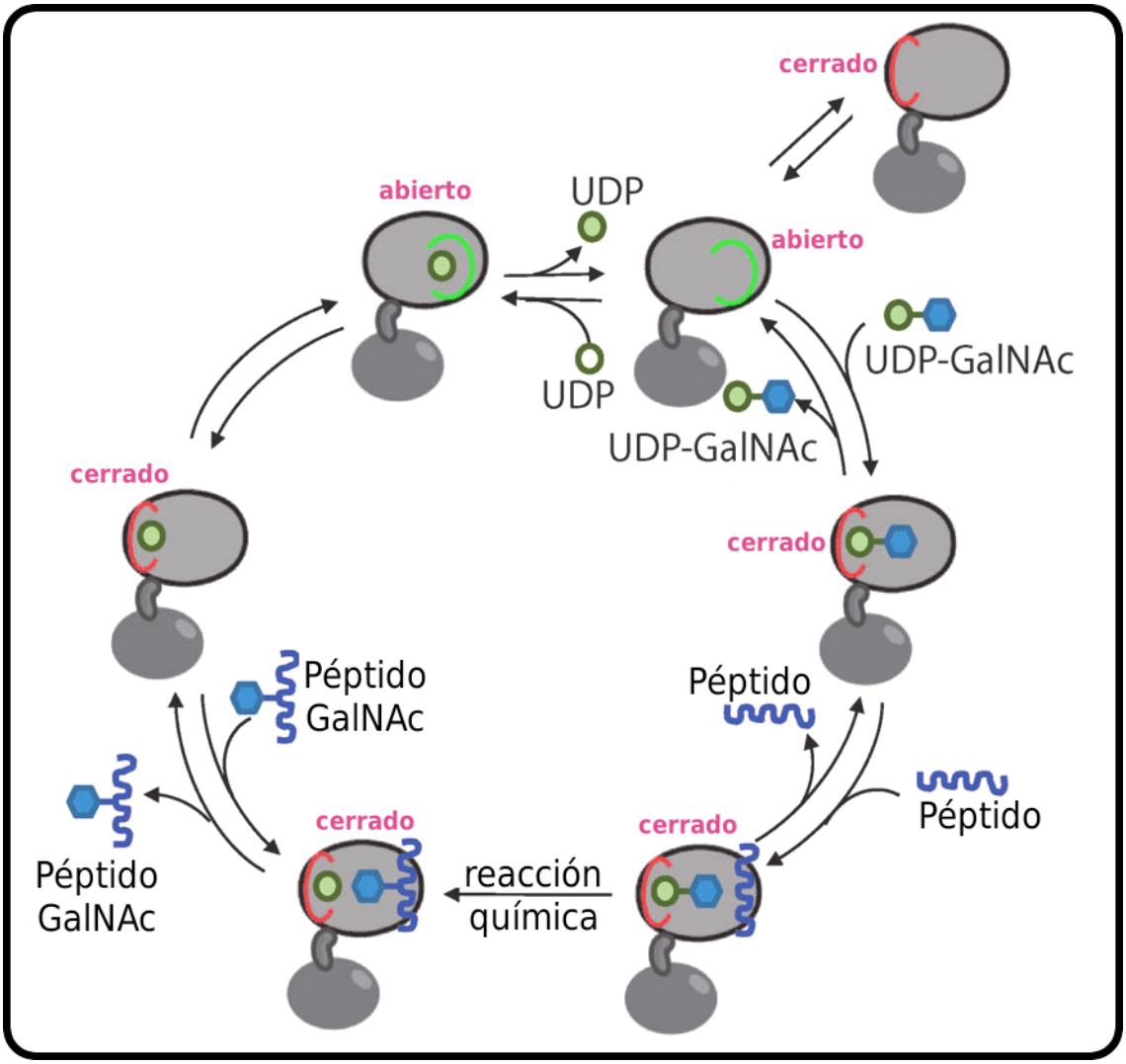


Figura 2.12. Mecanismo secuencial bi-bi ordenado. Modelo mecanístico secuencial propuesto por Lira-Navarrete *et al.* (2014), en el que el movimiento del asa flexible permite la entrada y salida del UDP-GalNAc y UDP, respectivamente. Modificado de Lira-Navarrete *et al*., 2014.

2.2.3. Especificidad de las ppGalNAc-Ts

Si comparamos esta modifación con otros procesos de glicosilación en mamíferos, como es por ejemplo el caso de la *N*-glicosilación, llevada a cabo por un solo complejo proteico (la Oligosacaril transferasa, OT) frente a una secuencia consenso determinada (NxS/T) (Apweiler *et al*., 1999), se evidencia lo incomprensible del por qué la naturaleza ha generado un número tan grande de isoformas dentro de la familia de ppGalNAc-Ts—20 en mamíferos, 9 en *C. elegans* y 12 en *Drosophila* (Bennett *et al*., 2012, Gill *et al*., 2011, Schjoldager *et al*., 2015, Hurtado-Guerrero, 2016)— para iniciar un solo camino como es el de la glicosilación de tipo *O*-GalNAc. ¿Ha sido para garantizar la eficiente glicosilación de una gran diversidad de secuencias peptídicas aceptoras o, tal vez, —sin que ambas opciones sean exluyentes— lo ha propiciado la necesidad

de una regulación sutilmente diferencial, mediada por la *O*-glicosilación, de distintas funciones proteicas implicadas en vías de comunicación celular?

Con objeto de contestar a la pregunta anterior, se han realizado numerosos estudios bioquímicos *in vitro* con enzimas GalNAc-Ts (extraídas de tejidos u obtenidas a partir de expresión recombinante) frente a bibliotecas de péptidos que imitan tándems de mucinas conocidas y derivados glicosilados de éstas (Taetert *et al*., 2001, Gerken *et al*., 2006), gracias a lo cual se han podido establecer una serie de pautas con relación a su especificidad (Festari *et al*., 2017; Revoredo *et al*., 2016; Kong et al., 2014; Gerken et al., 2013; Bennett *et al*., 2012; Gill *et al*., 2011; Pedersen *et al*., 2011; Gerken *et al*., 2011; Gerken *et al*., 2006; Gerken *et al*., 2002; Hanisch *et al*., 2001; Hassan *et al*., 2000a y 2000b, entre otros).

Gracias a estos estudios se ha diferenciado normalmente entre dos grandes tipos de "subfamilias" dentro de las 20 isoformas de las pp-GalNAc-Ts (Figura 2.13): la subfamilia I (ppGalNac-T1, T2, T3, T5, T6, T8, T9, T11, T13, T14, T15, T16, T18 y T19), con selectividad preferente hacia sustratos peptídicos, y la subfamilia II (ppGalNAc-T4, T7, T10, T12 y T17), con selectividad preferente hacia sustratos glicopeptídicos (Bennett *et al*., 2012).

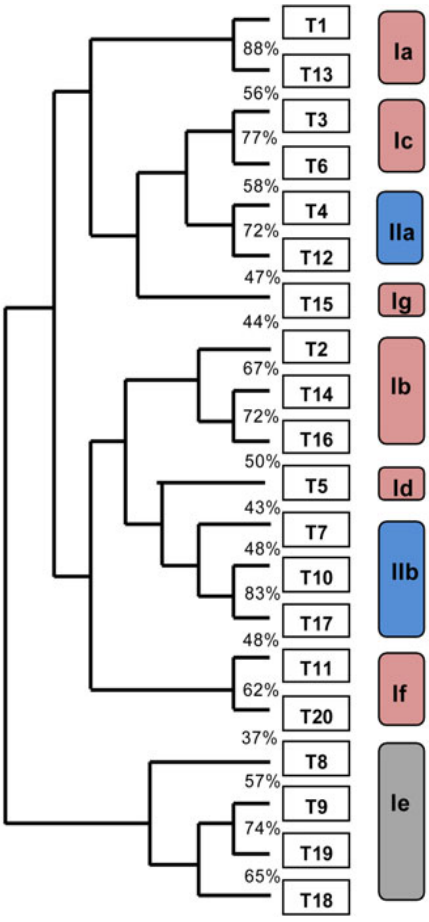


Figura 2.13. **Clasificación y árbol filogenético de las ppGalNAc-Ts.** Las isoformas con selectividad preferente hacia péptidos aparecen en rosa, las de selectividad preferente hacia glicopéptidos en azul, y las pertenecientes al subgrupo Ie aparecen en gris debido a la poca información que se tiene sobre su actividad. Modificado de Benett *et al*., 2012 y Revoredo *et al*., 2016.

De forma acorde con esta clasificación, el dominio catalítico de la enzima lleva a cabo el reconocimiento de los sustratos peptídicos, que presentan en su mayoría un motivo PxP en las posiciones de +1 a +3 respecto al sitio de adición de GalNAc, lo que a su vez está en consonancia con el "bolsillo de prolina" descrito por Fritz *et al*. (2004 y 2006) y Kubota *et al*. (2006). De igual modo, el reconocimiento de los sustratos glicopeptídicos vendría mediado por el dominio de lectina, que reconocería la presencia de un grupo GalNAc previo situado bien en el extremo N-terminal (pp-GalNAc-T3, T4, T6 y T12) o en el C-terminal (ppGalNAc-T1, T2 y T14) de la proteína sustrato, y de algún modo dirigiría a partir de ahí la nueva glicosilación hacia una serina o treonina localizada a unos 8-12 aminoácidos con respecto al sitio ya existente de glicosilación (Bennett *et al*., 2012; Hurtado-Guerrero, 2016; Kong *et al*., 2014; Kato *et al*., 2001; Hassan *et al*., 2000a y 2000b; Julenius *et al*., 2005; Raman *et al*., 2008; Taetert *et al*., 2001; Gerken, 2004; Gerken, Gilmore y Zhang, 2002; Gerken *et al*., 2002, 2006, 2011 y 2013; Hanisch *et al*., 2001; Wandall *et al*., 1997, entre otros).

No obstante, estudios más recientes (Revoredo *et al*., 2016) sugieren una clasificación algo más sutil, donde los miembros de la subfamilia Ia-d (GalNAc-T1, T2, T3, T5, T6, T13, T14 y T16; Figura 2.13) pasarían a considerarse como isoformas con selectividad preferente hacia péptidos y glicopéptidos con el GalNAc previo situado a distancias remotas del nuevo sitio aceptor, mientras que sólo los miembros Ie-f tendrían una selectividad estricta hacia sustratos peptídicos. Por otro lado, los miembros de la subfamilia IIa (GalNAc-T4 y T12), relativamente activos al enfrentarse a un número determinado de péptidos no glicosilados (Hassan *et al*., 2000b; Wandall *et al*., 2007), pasarían a considerarse como isoformas con especificidad mixta entre péptidos y glicopéptidos; sólo las isoformas de la subfamilia IIb (GalNAc-T7, T10 y T17) serían consideradas como estrictamente selectivas hacia los sustratos glicopeptídicos. Los autores de esta reclasificación proponen, así mismo, una explicación para cómo los subgrupos de ppGalNAc-Ts pueden llevar a cabo los distintos tipos de reconocimiento de sustratos glicopeptídicos (Figura 2.14).

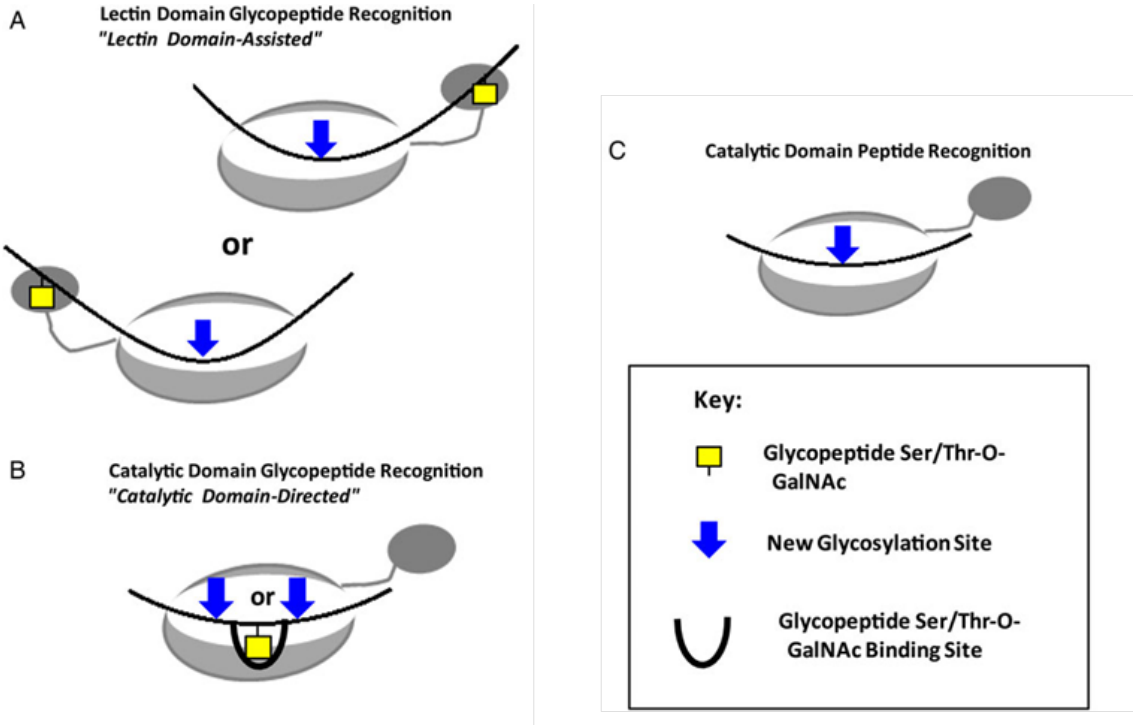


Figura 2.14. **Representación esquemática de las actividades de las ppGalNAc-Ts sobre sustratos glicopeptídicos. A**. Reconocimiento del glicopéptido asistido por el dominio de lectina (Ia-d, IIb), **B**. Reconocimiento del glicopéptido asistido por el dominio catalítico (IIa y IIb) o **C**. Reconocimiento peptídico asistido por el dominio catalítico (Ia-d y IIa). **Fuente**: Revoredo *et al*., 2016.

No obstante, la falta de motivos y secuencias consenso que permitan predecir e identificar la *O*-glicosilación (Julenius *et al*., 2005; Steentoft *et al*., 2011; Schjoldager *et al*., 2015) ha supuesto siempre un gran obstáculo a la hora de mapear el conjunto del *O*-glicoproteoma. Este hecho, así como el elevado porcentaje de redundancia entre las distintas isoformas (Figura 2.15), ha dificultado durante décadas el descubrimiento de casos específicos y ejemplos en los que las ppGalNAc-Ts desempeñen una función biológica concreta, como es el caso de la GalNAc-T3 y su papel en la homeostasis del calcio y el fosfato (Topaz et al., 2004; Kato et al., 2006; Ichikawa et al., 2009), o la GalNAc-T2 y su importancia en el metabolismo lipídico (Ketharpal et al., 2016; Schjoldager et al., 2012; Teslovich et al., 2010).

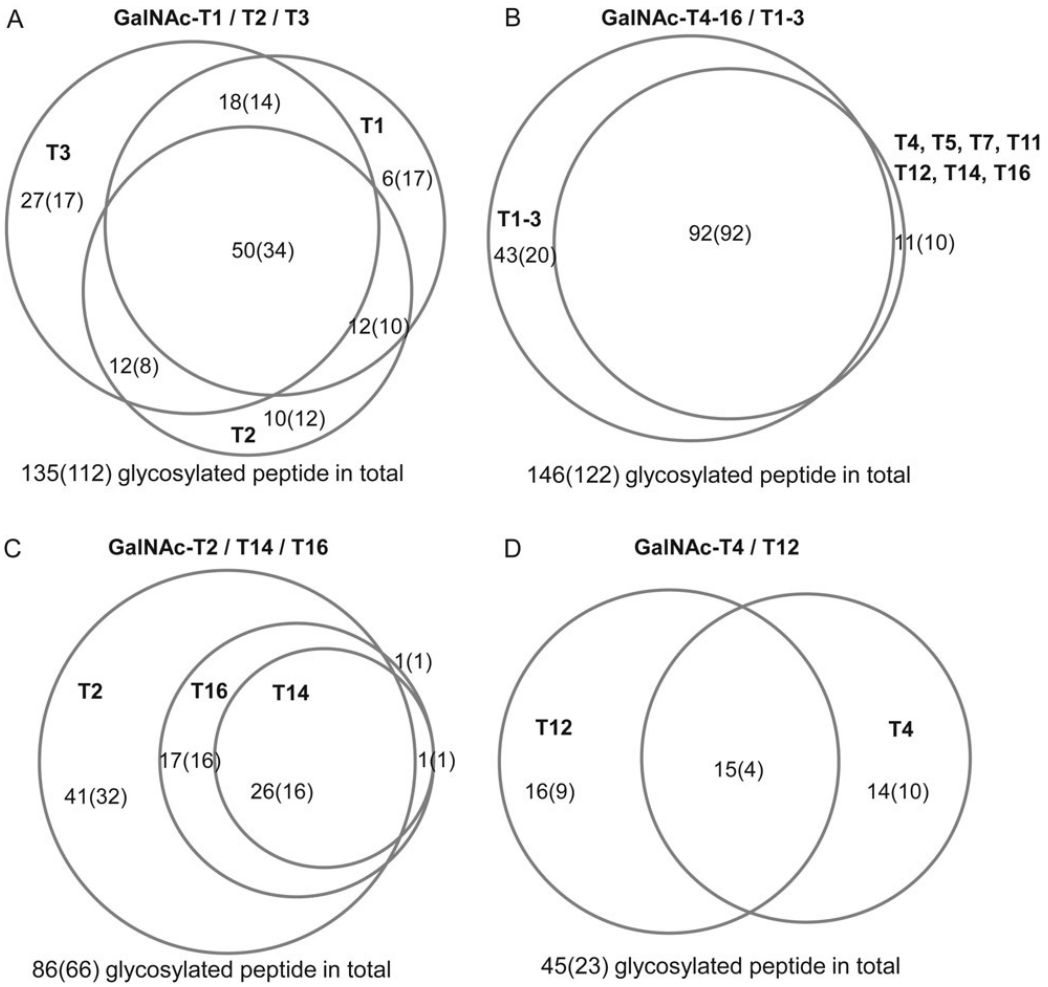


Figura 2.15. **Solapamiento entre ppGalNAc-Ts.** Representación de la selectividad y solapamiento de la actividad de las distintas isoformas de las GalNAc-Ts sobre una librería de péptidos y glicopéptidos sustratos. **A**. Solapamiento entre la actividad de GalNAc-T1, T2 y T3. **B**. Solapamiento de la actividad entre la actividad de GalNAc-T1 y T3 frente al resto de GalNAc-Ts. **C**. Solapamiento de la actividad de las isoformas de la subfamilia Ib. **D**. Solapamiento de la actividad de las isoformas de la subfamilia IIa. **Fuente**: Kong, 2015.

Afortunadamente, la evidente necesidad de desarrollar métodos para la determinación cuantitativa de los sitios de *O*-glicosilación específicos para las diferentes isoformas de las ppGalNAc-Ts y sus efectos en la función proteica empieza a solventarse gracias al desarrollo de técnicas como la estrategia *SimpleCell* (SC), implementada por el laboratorio del Profesor H. Clausen, en Dinamarca (Schjoldager *et al*., 2015; Yang *et al*., 2014; Steentoft et al., 2013; Schjoldager *et al*., 2012; Schjoldager *et al*., 2011; Steentoft *et al*., 2010; Teslovich *et al*., 2010). La base de dicha estrategia radica en la generación de líneas celulares humanas en las que el gen que sintetiza la chaperona COSMC, necesaria para la actividad de la enzima C1GalT1, está silenciado, por lo que no se puede realizar la extensión del antígeno Tn que hayan generado las isoformas de ppGalNAc-Ts presentes en cada línea celular (Schjoldager *et al*., 2015; Yang *et al*., 2014). El desarrollo de modelos celulares isogénicos de este

tipo esperan ser el fundamento para alcanzar, en un futuro próximo, un mayor conocimiento acerca de las funciones específicas de las distintas isoformas de las ppGalNAc-Ts y su implicación directa en el procesamiento, fisiológico y en condiciones de patogenicidad, de las proteínas de nuestro organismo.

Así por tanto, la necesidad de ahondar en los poco definidos factores estructurales y mecanísiticos que subyacen en los procesos de selección y en las posibles funciones específicas de las diferentes isoformas de las ppGalNAc-Ts es evidente, dado que nos permitirían no sólo justificar e incluso predecir su orquestado comportamiento (Lorenz *et al*., 2016), sino también desarrollar herramientas selectivas que puedan modular su actividad en los casos en los que sea necesario para el correcto funcionamiento del organismo (Hurtado-Guerrero, 2016).

2.2.4. Relación con patologías

Tradicionalmente, la *O*-glicosilación se ha considerado una modificación postraduccional propia de dominios de tipo mucina, densamente glicosilados (Van de Steen *et al*., 1998). No obstante, cada vez son más los estudios que evidencian cómo este tipo de *O*-glicosilación aparece también en sitios específicos de glicosilación presentes en proteínas no mucínicas, para los que la actuación de isoformas específicas de las ppGalNAc-Ts es determinante (Hanisch *et al*., 2001; Halim *et al*., 2011; Steentoft *et al*., 2011; Schjoldager *et al*., 2010; Kato *et al*., 2006; Schjoldager y Clausen, 2012). De hecho, el mayor impacto en el campo de la *O*-glicosilación puntual tuvo lugar a raíz de la vinculación de la deficiencia en GalNAc-T3 con una enfermedad humana, la calcinosis tumoral, que se originaba al no *O*-glicosilar ésta la Thr178 del factor FGF23, implicado en la regulación del fosfato (Topaz *et al*., 2004; Duncan *et al*., 2011; Kato *et al*., 2006; Rafaelsen *et al*., 2014; Folsom e Imel, 2015). A partir de ahí, se descubrieron otras implicaciones directas como son: (a) la aparición de heterotaxia (también conocida como isomerismo izquierdo) causada por la incorrecta *O*-glicosilación del receptor NOTCH1 por la isoforma GalNAc-T11 (Boskovski *et al*., 2013); (b) la incorrecta *O*-glicosilación de la proteína precursora del péptido amiloide por parte de la *Hs*GalNAc-T2, lo que conlleva la deposición de péptido β-amiloide en el parénquima cerebral y se asociada directamente con el desarrollo de la enfermedad de Alzheimer (Weidemann *et al*., 1989; Schedin-Weiss *et al*., 2014), y (c) la mala regulación del metabolismo lipídico causada por la mutación Phe104Ser en la enzima *Hs*GalNAc-T2, que la inactiva e impide la *O*-glicosilación y posterior procesamiento de diversas proteínas implicadas en la mencionada regulación (Khetarpal *et al*., 2016; Figura 2.16).

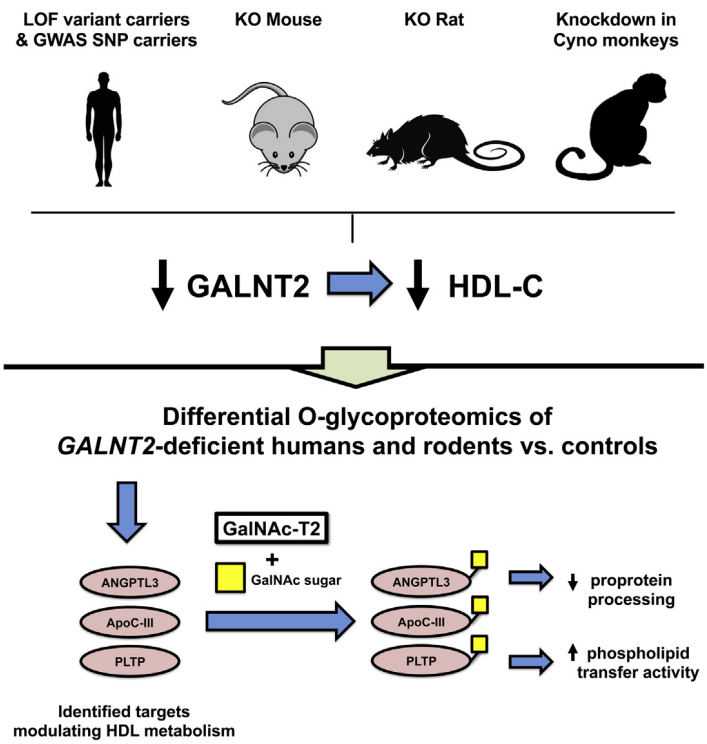


Figura 2.16. Efecto de la deficiencia de GalNAc-T2 en el metabolismo lipídico. La glicosilación deficiente impide la activación del procesamiento de las proproteínas necesarias para la correcta regulación metabólica de los lípidos. **Fuente**: Khetarpal *et al*., 2016.

Todos estos ejemplos nos dejan entrever la importancia y hasta ahora infravalorada significancia biológica de las ppGalNAc-Ts y su *O*-glicosilación, cuya relevancia metabólica parece recaer en su implicación en la regulación del procesamiento de diversas proproteínas (Schjoldager y Clausen, 2012; Schjoldager *et al*., 2011; Schjoldager *et al*., 2015; Varki *et al*., 2017). De ahí que la importancia de desentrañar los mecanismos moleculares de estas interacciones espefícas isoforma-sustrato sea cada vez mayor, al igual que lo es el comprender los motivos que inducen a la inactividad de los diversos mutantes responsables de la inactividad de las enzimas y diseñar mecanismos que palien sus efectos (Tarp y Clausen, 2008,;Supraha Goreta *et al*., 2012; Martínez-Sáez *et al*., 2016).

Mencionar, por último, cómo otra de las áreas de la medicina en las que la implicación de la *O*-glicosilación aberrante ha de ser todavía desentrañada es la de la biología del cáncer. La vinculación de una *O*-glicosilación deficiente y las células cancerígenas data de 1969, con la detección del antígeno Tn en la membrana celular de células tumorales (Prokop y Uhlenbruck, 1969), si bien se ha indagado mucho en la misma desde entonces (Taylor-Papadimitriou *et al*., 1999; Reis *et al*., 1999; Bafna *et al*., 2010; Brooks *et al*., 2007; Kufe, 2009; Thériault *et al*., 2011; Van Putten *et al*., 2017; Bafna *et al*., 2010; Fu *et al*., 2016; Haugstad *et al*., 2012, entre otros).

Entre los glicanos antigénicos específicos de células tumorales, y que aparecen en la superficie de más del 80% de las mismas, se encuentran varios de tipo *O*-GalNAc de la familia de los antígenos de Thomsen-Friedenreich (Chia *et al*., 2016; Sletmoen *et al*., 2018), como son el antígeno T nuevo (Tn), el antígeno T (T) y sus respectivos derivados sialilados (S-Tn y S-TF; Figura 2.17), los primeros de los cuales también se encuentran en otras enfermedades humanas como la nefropatía IgA (Ju *et al*., 2013).

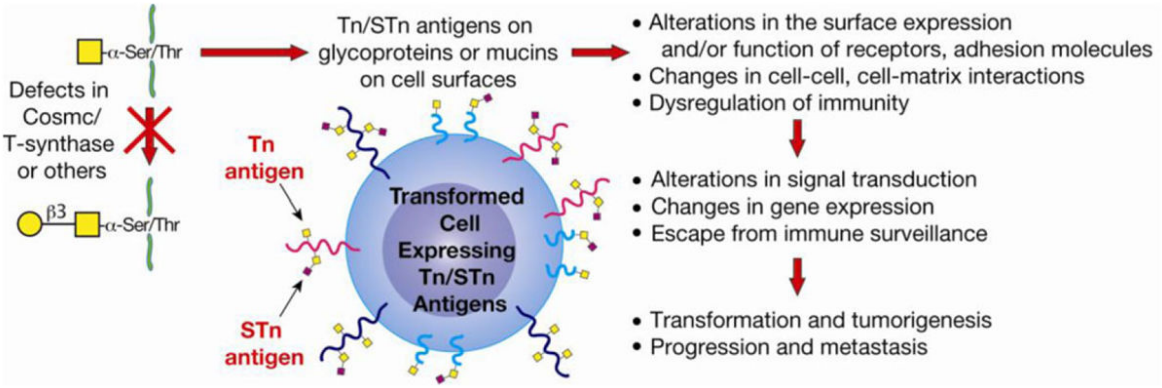


Figura 2.17. Antígenos Tn y STn en células tumorogénicas. Ambos carbohidratos antigénicos asociados a tumor se expresan solamente en células malignas, debido fundamentalmente a defectos en la chaperona COSMC. Los mecanismos por los que la presencia de los antígenos favorecen la progresión del tumor no se conocen aún en profundidad, pero la asociación con la capapidad invasiva y metastásica del tumor ha sido corroborada. **Fuente:** Ju *et al*., 2013.

Diversos trabajos han demostrado también cómo en condiciones tumorales se produce una relocación de las enzimas ppGalNAc-Ts desde el aparato de Golgi, donde se encuentran en condiciones fisiológicas (Rottger *et al*., 1998), al retículo endoplasmático (Gill *et al*., 2010; Chia *et al*., 2016; Gill *et al*., 2011), lugar en el que puede favorecerse la adición de GalNAc a sitios secundarios de glicosilación, o normalemente ocultos en el Golgi (Figura 2.18), con el consecuente incremento de la actividad de las ppGalNAc-Ts (GALA) y de la presencia de los antígenos Tn y STn.

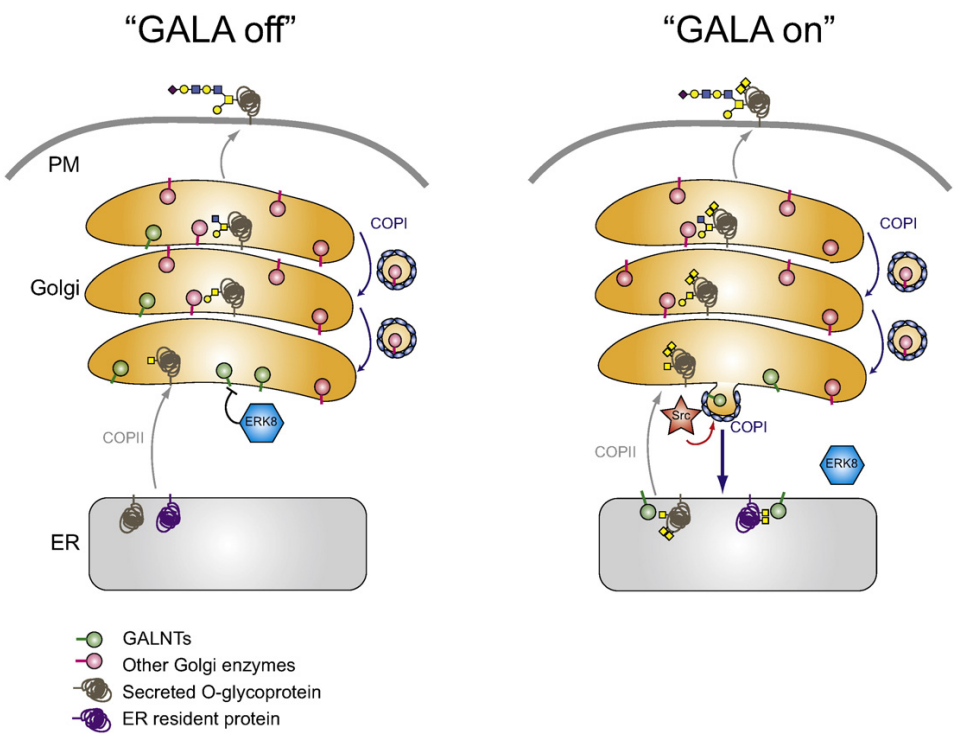


Figura 2.18. Regulación molecular de la vía de activación de las ppGalNAc-Ts (GALA). En condiciones fisiológicas (izquierda), la iniciación de la *O*-glicosilación de las proteínas tiene lugar en el aparato de Golgi, donde enseguida actúan las siguientes enzimas de la vía, como C1GalT1. En condiciones tumorogénicas (derecha) se produce el tráfico de ppGalNAc-Ts hacia el retículo endoplasmatico, donde pueden glicosilar proteínas a las que normalmente no acceden, y donde no se encuentran las siguientes enzimas de la vía de *O*-glicosilación. **Fuente**: Chia *et al*., 2016.

No obstante, las consecuencias funcionales de este cambio en el repertorio de las ppGalNAc-Ts, así como de la presencia aberrante de los antígenos Tn, T, STn y ST —que también puede generarse por la pérdida de la chaperona COSMC, responsable del plegamiento funcional de la T-sintasa (Ju *et al*., 2008; Radhakrishnan *et al*., 2014; Sletmoen *et al*., 2018; Figura 2.1)—, está lejos de entenderse por completo. Recientemente se han empezado a dilucidar las posibles implicaciones de estos antígenos en los procesos de transformación celular asociados a las condiciones tumorales (Chia *et al*., 2016; Sletmoen *et al*., 2018; Fu *et al*., 2016), sabiendo que la tumorogénesis se veía potenciada al incrementar de forma artificial la expresión de los mismos (Julien *et al*., 2001; Ozaki *et al*., 2012). A este respecto, uno de los mecanismos propuestos para la acción de los antígenos Tn y STn en el cáncer está relacionado con su capacidad para promover interacciones de tipo carbohidrato-carbohidrato (CCI) entre ellos (Sletmoen *et al*., 2018): se postula su papel como promotores secundarios de la carcinogénesis, gracias a su agregación y subsiguiente activación de receptores celulares que reconocen glicanos densamente *O*-glicosilados de tipo mucina (Haugstad *et al*., 2012; Haugstad *et al*., 2016; Brewer *et al*., 2015). Se proponen, de hecho, múltiples vías a través de las cuales los glicanos de tipo *O*-GalNAc pueden regular la formación de tumores, la invasión

celular y la progresión del cáncer, afectando para ello a propiedades de unión celular, de actividad, de niveles de expresión y de estabilidad, entre otras (Chia *et al*., 2016; Figura 2.19).

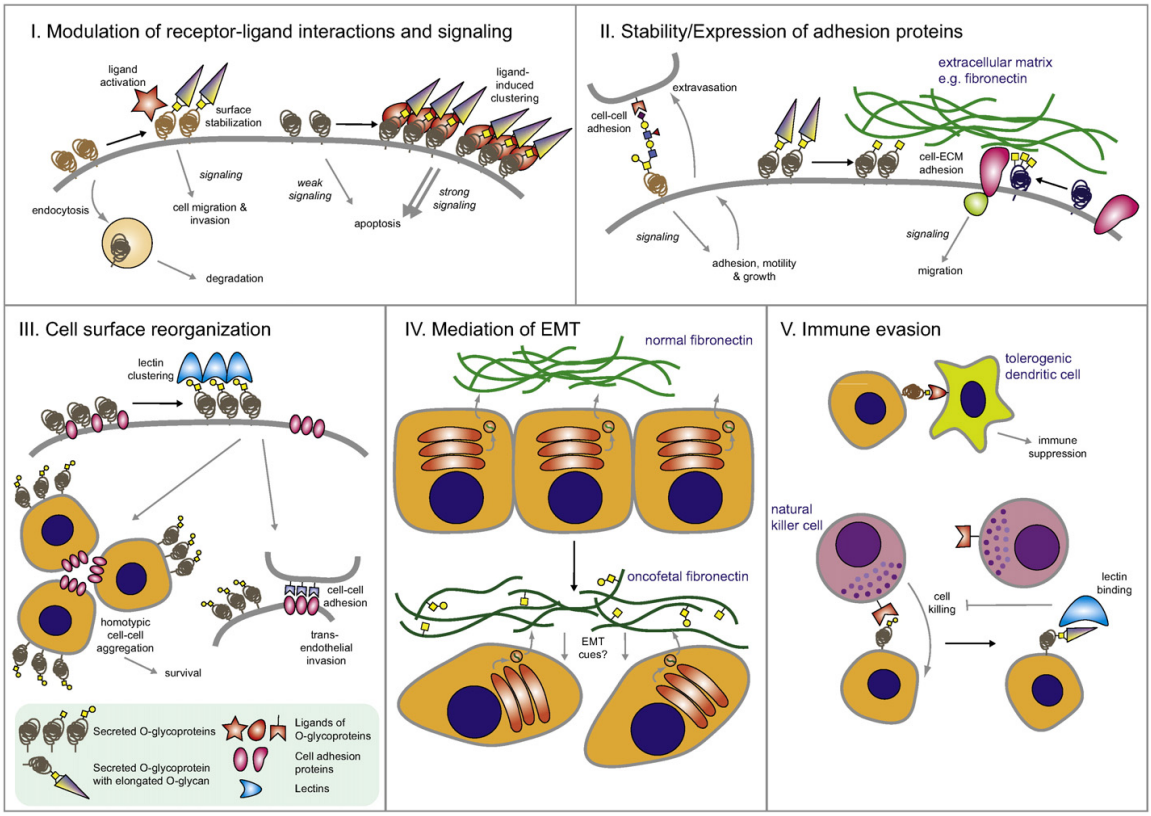


Figura 2.19. Papel de los *O*-glicanos en la invasión tumoral. Las flechas negras indican un cambio en la *O*-glicosilación, y las grises marcan las vías de señalización celular. **I.** Los *O*-glicanos pueden estabilizar la expresión de los receptores en la superficie celular bien sea reduciendo la endocitosis o promoviendo la formación de complejos. **II**. Determinados *O*-glicanos pueden condicionar las propiedades de adhesión de proteínas extracelulares de membrana, modificando las vías de señalización relacionadas. **III**. Los *O*-glicanos pueden favorecer la reoganización de la membrana, favoreciendo la adhesión entre células. **IV**. La sobreactivación de algunas ppGalNAc-Ts puede inducir la transición epitelio-mesénquima (EMT), lo que a su vez altera la morfología celular y promueve la oncogénesis. **V**. La presencia de *O*-glicanos puede impedir la actuación del sistema inmune, ya sea a través de células T tolerogénicas que reconocen el domino de lectina o mediante una unión a lectina que las enmascara frente a las *Natural Killer* (NK). Modificado de Chia *et al*., 2016.

Pese a las diversas evidencias que corroboran las implicaciones mostradas en la figura 2.19 (Wu et al., 2011; Häuselmann y Borsig, 2014; Van Putten y Strijbis, 2017; Radhakrishnan *et al*., 2014; Ponnusamy *et al*., 2011; Ozaki *et al*., 2012; Mukhopadhyay *et al*., 2013; Madsen *et al*., 2013; Park *et al*., 2011 entre otros), en la mayoría de los casos falta un entendimiento molecular acerca de cuáles son las proteínas diana implicadas e incluso acerca de qué mecanismos regula la propia glicosilación. Un mayor conocimiento acerca de las funciones específicas de las diferentes isoformas de las ppGalNAc-Ts, así como de sus

perfiles de glicosilación y mecanismos de acción específicos (objetivos generales ambos del compendio de artículos recogidos en esta tesis), permitirán —en concierto con otros estudios del tipo *SimpleCell* (Schjoldager *et al*., 2015) — desentrañar poco a poco la complicada madeja de los principios, mecanismos y funciones últimas del, hasta ahora, más desconocido paso de la *O*-GalNAc-glicosilación para, a partir de ahí, poder desarrollar moduladores, biomarcadores, vacunas , etc., que intenten paliar todos sus posibles efectos patogénicos (Tarp y Clausen, 2008; Rachagani *et al*., 2009; Martínez-Sáez *et al*., 2016).

CAPÍTULO 3- OBJETIVOS

Considerando lo expuesto anteriormente y los trabajos previos del grupo (Lira-Navarrete *et al*., 2014), el objetivo general de esta tesis y de las publicaciones que en ella se agrupan es el de ahondar en el conocimiento y justificación de las bases moleculares subyacentes en el reconocimiento mediado por las diversas isoformas de las ppGalNAc-Ts, así como de la orquestada regulación de las mismas. Para ello se plantean los siguientes objetivos específicos, asociables cada uno a las respectivas publicaciones que comprenden este compendio:

- Elucidar el papel del dominio de lectina de la *Hs*GalNAc-T2 en el perfil de glicosilación de esta enzima a través de la resolución de su estructura en complejo con péptidos glicosilados y no glicosilados (Lira-Navarrete, **de las Rivas** *et al*., 2014).
- Esclarecer el papel del *linker* flexible en el establecimiento de las preferencias de glicosilación dirigidas por el dominio de lectina de las ppGalNAc-Ts (**de las Rivas** *et al*., 2017).
- Resolver la estructura cristalográfica y caracterizar la interacción entre la *Hs*GalNAc-T2 y compuestos glicomiméticos menos polares que los de primera generación, diseñados con objeto de modular la actividad de ésta y otras glicosiltransferasas (Ghirardello, **de las Rivas** *et al*., 2016).
- Resolver la estructura cristalográfica de *Hs*GalNAc-T2 en complejo con la luteolina, una molécula de la familia de las flavinas que se postula como inhibidor selectivo de la enzima (Liu, Xu, Xu, **de las Rivas** *et al*., 2017).
- Desentrañar a nivel estructural y mecanístico las causas de la inactividad enzimática y consecuente fenotipo patogénico que conlleva la mutación Phe104Ser en la isoforma *Hs*GalNAc-T2 (**de las Rivas** *et al*., 2018).

CAPÍTULO 4- MATERIALES Y MÉTODOS

Todos los trabajos desarrollados en esta tesis doctoral se caracterizan por su naturaleza multidisciplinar y heterogénea, donde se integran numerosas técnicas experimentales con el objetivo común de responder de la forma más exacta y precisa posible a las preguntas biológicas y bioquímicas que los generan. Por cuestiones lógicas de tiempo y posibilidades, esta aproximación holística a los diversos estudios ha implicado la necesaria colaboración de distintos grupos y laboratorios especialistas en las diversas técnicas, gracias a cuya ayuda se han podido llevar a cabo y complementar los diferentes proyectos. A excepción del artículo IV (véase capítulo Resultados), donde el trabajo restante ya estaba finalizado en el momento de nuestra intervención, surgida como respuesta a un revisor de una revista, el total de experimentos y de técnicas de los demás proyectos han sido propuestos, diseñados y planteados en el seno del grupo de investigación del doctorando que, a partir de ahí, se ha puesto en contacto con los diversos colaboradores experimentales. En este contexto es, por tanto, de marcada importancia el señalar cuáles son las aportaciones que, dentro de los diversos trabajos, han sido desarrolladas por el doctorando y su grupo, tal y como se hace en el capítulo Resultados antes de cada uno de los Artículos y como se certifica en el apéndice correspondiente (Apéndice 10.5).

En lo que atañe al presente capítulo de materiales y métodos, en él se describirán con detalle todas las técnicas que han sido llevadas a cabo por el propio doctorando. Del resto de técnicas experimentales, a excepción de las del artículo IV, se nombrará al grupo artífice, se introducirá su importancia en el contexto del trabajo y se dará una pincelada general acerca de su fundamento, si bien no se profundizará en los detalles técnicos de su realización —disponibles, por otro lado, en la sección experimental y en el material suplementario de cada uno de los artículos—.

4.1.TÉCNICAS DE BIOLOGÍA MOLECULAR

4.1.1. Construcción del plásmido pPICZαA-galnact4 y de los plásmidos que contienen las quimeras.

4.1.1.1. Plásmido codificante para HsGalNAc-T4

La secuencia de ADN codificante para los residuos 36 al 578 de la enzima humana GalNAc-T4 (se elimina la primer parte de la secuencia aminoacídica, que corresponde a su dominio transmembrana), a la que nos referiremos como *galnact4*, fue sintetizada en el vector pUC57 por la empresa GenScript (EEUU), y se optimizaron sus frecuencias de codones para su posterior expresión en *Pichia pastoris*. Dicho ADN contenía una secuencia de reconocimiento para la enzima de corte XhoI (Roberts *et al.*, 2003) y un sitio de corte para la proteasa KEX2 de

Pichia en el extremo 5' terminal (Damasceno, 2012); en el extremo 3' contenía una secuencia codificante para una cola de 6 histidinas, un codón terminal y una secuencia de reconocimiento para la enzima de restricción SacII. Después de la doble digestión con XhoI y SacII, el constructo fue subclonado por GenScript en el vector de expresión pPICZαA (Figura 4.1), generando el plásmido de expresión pPICZαA-*galnact4*, del que se nos entregaron 4 µg liofilizados.

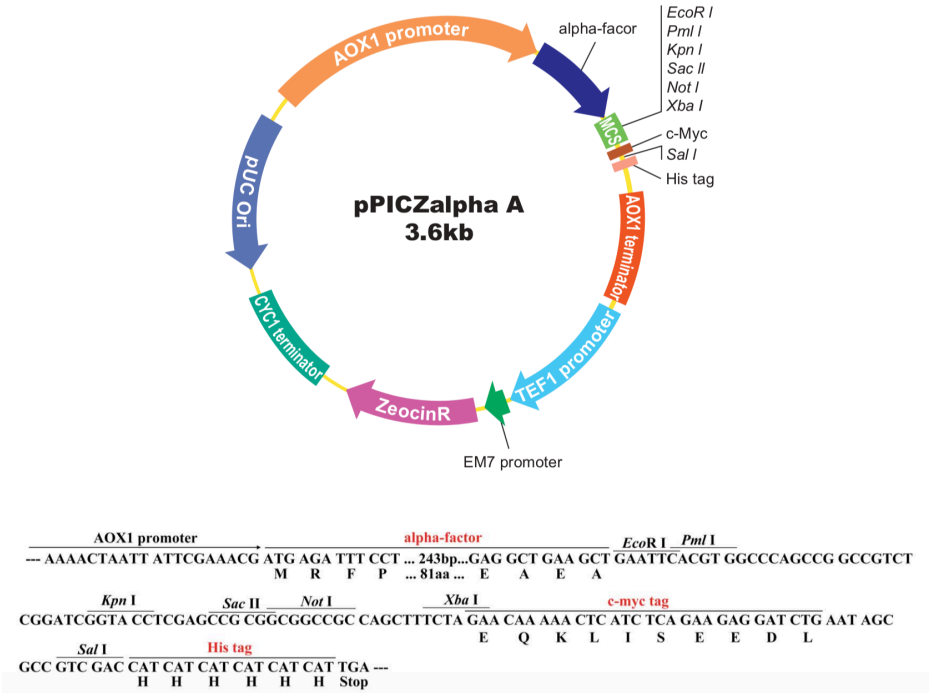


Figura 4.1. Vector pPICZαA. Empleado por GenScript (EEUU) para la posterior expresión en Pichia Pastoris. **Fuente:** https://www.genscript.com/gsfiles/vector-map/yeast/pPICZalpha-A.pdf.

4.1.1.2. Plásmidos codificantes para las quimeras

La síntesis de los plásmidos que codificaban las seis quimeras que se emplearon en el estudio del papel del *linker* flexible localizado entre los dominios catalítico y de lectina de las enzimas humanas GalNAc-T2, T3 y T4 (capítulo Resultados, Artículo II) también fue realizada por la empresa GenScrit (EEUU), partiendo para ello de las secuencias de ADN de las proteínas *Hs*GalNAc-T4, *Hs*GalNAc-T3 y la anteriormente empleada por nuestro grupo *Hs*GalNAc-T2 K75-Q571 (Lira-Navarrete *et al*., 2014).

4.1.2. Mutagénesis de sitio dirigido.

Desde el año 2015, todas las mutaciones puntuales de ADN que se diseñan en el grupo son llevadas a cabo por la empresa GenScript (EEUU). No obstante, en este trabajo hay algunas que se realizaron todavía en nuestro laboratorio,

como las reflejadas en de las Rivas *et al.*, 2015 (véase capítulo Resultados, artículo I) y algunas de las utilizadas en de las Rivas, Coelho *et al.*, 2018 (capítulo Resultados, Artículo IV). Para obtener éstas, se efectuaron diversas PCRs en las que se empleó como ADN molde el que codificaba para la proteína silvestre y se diseñaron los oligonucleótidos apropiados para la introducción de cada mutación deseada (Tabla 4.1):

Tabla 4.1. Mutantes diseñados en el laboratorio

Proteína	Mutación	Oligonucleótidos
*Hs*GalNAc-T2	W282A	W282A_F: 5'ACTGGAATTTGGTTTTCAACC**GCC**ATTATATCACCCCAGAACAAAG W282A_R: 5'CTTTGTTCTGGGGTCATATAA**TCA**GCCTTGAAAACCAAATTCCAGTC
	F361A	F361A_F: 5'CCATGCAGTAGAGTTGGTCATGTC**GCC**AGAAAAAGCACCCATACACCTTC F361A_R: 5'GGGAAGGTGATGGGTGCTGTTTTCT**GGC**GACATGACCAACTCTACTGCAT
	D458R	D458R_F: 5'GCAACAGGGTACTAATTGTTTG**AGA**ACTTTGGGTCATTTTGCTGACGG D458R_R: 5'CCGTCAGCAAAATGACCCAAAGT**TCT**CAAACAATTAGTACCCTGTTGC
	D458A	D458A_F: 5'GCAACAGGGTACTAATTGTTTG**GCT**ACTTTGGGTCATTTTGCTGACGG D458A_R: 5'CCGTCAGCAAAATGACCCAAAGT**AGC**CAAACAATTAGTACCCTGTTGC
	F104S	F104S_F: 5'TACGCAAGAAATAAG**TCT**AATCAGGTTGAAAGTGATAA F104S_R: TTATCACTTTCAACCTGATT**AGA**CTTATTTCTTGCGTA
Se indica en negrita el codón mutado		

La proporción de estos oligonucleótidos y del resto de componentes de cada PCR fueron los reflejados en la Tabla 4.2:

Tabla 4.2. Componentes de la reacción de PCR para las mutagénesis

Componente	Concentración inicial	Concentración final	Volumen (µl)
ADN Polimerasa Phusion Hot Start II High-Fidelity (Thermo Scientific)	2 U / µl	18 mU / µl	0.45
Tampón de carga de la polimerasa (Thermo Scientific)	5 x	1 x	10
ADN Molde	100 ng / µl	2 ng / µl	1
Oligonucleótidos	10 µM c / u	0.4 µM c / u	2 c / u
Mezcla de dNTP	2 mM c /u	0.2 mM c /u	5
DMSO	100 %	3 %	1.5
$MgSO_4$	25 mM	1.5 mM	3
H_2O Milli-Q			25.05
Volumen final			50

Las reacciones de amplificación del ADN se llevaron a cabo bajo las siguientes condiciones:

- Un ciclo inicial de 2 minutos a 98 ºC.
- 29 ciclos de 30 segundos a 98 ºC para provocar la desnaturalización de las hebras, 1 minuto a 55 ºC para permitir la hibridación de los oligonucleótidos y 2 minutos a 72 ºC para la extensión de la cadena.
- Un ciclo final de 10 minutos a 72 ºC.

La verificación de la correcta amplificación se realizó mediante una electroforesis en gel de agarosa al 1 % (apartado 4.1.3).

Una vez comprobada la amplificación, se digirieron las hebras parentales de las muestras de PCR añadiendo a cada una de ellas 2 µl de la enzima de restricción *DpnI* (New Englands Biolabs, 20000 U/ml), que se dejó incubar a 37 ºC durante 2 horas. Tras ello se procedió a la transformación con 6 ul de ADN de células competentes de *Escherichia coli* DH5α (apartado 4.1.4), y las bacterias transformadas se sembraron en placas de medio *Low Salt* LB agar (Apéndice 10.1) conteniendo 25 µg/ml de zeocina (Invivo Gen), que se incubaron a 37 ºC durante una noche. Al día siguiente se escogieron aleatoriamente tres colonias de la placa, cada una de las cuales se estrió en una placa nueva y se empleó como inóculo de un precultivo de 10 ml de *Low Salt* LB líquido (Apéndice 10.1) con la misma cantidad de zeocina, que se dejó incubando a 37 ºC y 250 rpm durante una noche. Los precultivos crecidos se centrifugaron al día siguiente durante 5 minutos a 3800 xg y 4 ºC (Eppendorf Centrifuge 5810R), se desechó el sobrenadante y se extrajo el ADN plasmídico de las células siguiendo para ello las instrucciones del protocolo del kit comercial Accuprep Plasmid Mini Extraction de Bionner (Figura 4.2). La elución final de cada muestra se realizó en 65 µl de H_2O Milli-Q (Millipore) a 65 ºC y el ADN resultante se cuantificó mediante absorción a 260 nm en un nanodrop NanoVue Plus (GE Healthcare Life Sciences). La secuenciación de cada gen mutado fue llevada a cabo por la empresa valenciana Sistemas Genómicos S.L. (https: //www.sistemasgenomicos.com/).

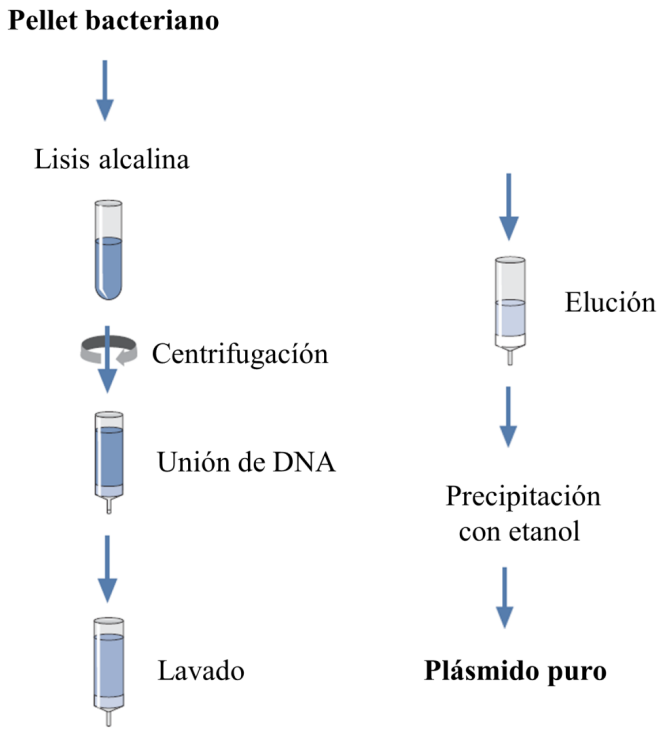


Figura 4.2. Extracción de ADN plasmídico. Esquema del proceso de extracción de ADN plasmídico mediante un kit comercial. **Fuente**: Martínez Oliván, 2014.

4.1.3. Electroforesis de ADN en gel de agarosa.

Todos los geles de agarosa para la visualización de ADN de este trabajo se realizaron, si no se especifica lo contrario, a una concentración del 1 % de agarosa en tampón TBE 1x (Apéndice 10.1). Las muestras se añadieron a los pocillos tras la adición del volumen correspondiente de tampón de carga 6x (Apéndice 10.1) y la tinción se realizó con 0,5 µl de SYBR Safe DNA gel stain (Invitrogen). El marcador de peso molecular empleado fue el 1 Kb Plus DNA Ladder (Invitrogen), del que, una vez preparado (Apéndice 10.1), se pipetearon 10 µl en un pocillo. Finalmente, las bandas de ADN se resolvieron en el gel tras la aplicación de una diferencia de potencial de 110 V durante 35 minutos, y se observaron en un equipo GeneFlash Syngere (Bio Imagin) mediante exposición a una fuente de luz azul (Safe Imager 2.0, Invitrogen).

4.1.4. Transformación de células competentes *E. coli* DH5α.

La transformación de células competentes es un proceso mediante el cual se incorpora a cepas silvestres de una determinada bacteria un plásmido que contiene el gen de interés. En nuestro caso, todos los constructos de ADN empleados en este trabajo se han transformado en cepas de *Eschericia coli* DH5α, dado que el fin último de la transformación era el de obtener ADN plasmídico en cantidad suficiente. El protocolo de transformación de las células de *E. coli* quimiocompetentes fue, en todos los casos, el siguiente:

- Adición en hielo de unos 100 ng de plásmido puro a una alícuota de 150 µl de células competentes, siempre que el ADN no proveyese del protocolo de mutágenesis dirigida anteriormente descrito (apartado 4.1.2) en cuyo caso se añaden unos 600 ng pues se asume que no todos los plásmidos habrán incorporado la mutación deseada.
- Incubación en hielo durante 20 minutos.
- Calentamiento a 42 ºC de entre 45 y 60 segundos.
- Reposo en hielo durante dos minutos.
- Adición de 500 µl de medio SOC frío (Apéndice 10.1).
- Recuperación de las células durante un periodo de 1 a 3 horas a 37 ºC con agitación de entre 200 y 250 rpm.

Una vez hecho esto, se siembran 150 µl de cada muestra en una placa de *Low Salt* LB agar (Apéndice 10.1) que contiene el antibiótico de selección correspondiente; en todos los casos que se plantean en este trabajo, dicho antibiótico es zeocina (Invivo Gen), añadido a 25 µg/ml. Ya por último, las placas se incuban durante la noche a 37 ºC.

Mencionar de forma paralela cómo, en el caso de contar con menos ADN de partida del deseado o con un menor volumen de células competentes, puede añadirse un último paso en el protocolo consistente en la centrifugación durante 5 minutos a 1500 xg de las muestras ya recuperadas (Centrífuga Beckman Coulter, Microfuge 16), para después plaquear todas las células resuspendidas en unos 100 µl de sobrenadante. De forma alternativa, puede optarse también por resuspender parte de las células directamente en unos 10 ml de medio líquido con el antibiótico correspondiente.

4.2. EXPRESIÓN DE PROTEÍNAS

Todas las proteínas recombinantes empleadas en este trabajo han sido expresadas en la cepa SMD1168 de la levadura *Pichia pastoris*; esta cepa es deficiente en Proteasa A y resulta particularmente útil para la expresión de proteínas con tendencia a degradarse (Gleeson, 1998, Ahmad *et al*., 2014), como hemos visto que sucede con las ppGalNAc-Ts. Se puede adquirir comercialmente, al igual que el resto de cepas disponibles en el laboratorio, en Thermo Fisher (http://www.thermofisher.com/order/catalog/product/C17500) y, salvo que se indique lo contrario, para su uso se siguen todas las recomendaciones del manual de usuario de *Pichia pastoris* de Invitrogen (Invitrogen, 2010).

4.2.1. Preparación de células electrocompetentes de la cepa de *P. Pastoris* SMD1168.

Para la transfección del ADN deseado en las células de *Pichia pastoris* se emplea usualmente el protocolo de electroporación; en éste, mediante la aplicación de un pulso eléctrico, se abren poros en las células a través de los que entra el ADN, cargado negativamente. No obstante, es necesario que las células se encuentren antes en unas condiciones que les permitan recuperar la ósmosis y polaridad de la membrana después de la aplicación de dicho pulso, por lo que el primer paso de la transfección ha de ser siempre la generación de células electrocompetentes.

De esta forma, el protocolo seguido es una adaptación del sugerido por Wu y Letcworth (2004), con un índice de eficiencia 20 veces superior al especificado por Invitrogen:

- Se siembran 60 µl de células SMD1168 en una placa de YPDS (Apéndice 10.1) sin antibiótico, a la par que se siembran otros 60 µl de la cepa elegida en otra placa de YPDS con 200 µg/ml de zeocina, que

hará las veces de control negativo de las placas. Se dejan crecer a 30 ºC durante 72 horas. En la placa sin antibiótico ha de observarse un césped, mientras que en la control no han debido salir colonias.

- Una vez crecida la placa, se inocula con varias colonias de la misma un precultivo de 5 ml de YPD líquido (Apéndice 10.1), que se deja creciendo a 250 rpm y 30 ºC durante unas 8 horas.
- Transcurridas las 8 horas, se transfieren 500 µl de dicho precultivo a un matraz de 500 ml que contenga 100 ml de YPD líquido y se deja creciendo durante la noche a 210 rpm y 30 ºC, hasta que haya alcanzado una densidad óptica a 600 nm (OD_{600}) que se encuentre entre 1 y 2 (aproximadamente, 15-16 horas).
- Alcanzada la densidad deseada, se centrifugan 5 minutos las células a 2000 xg y 4º C. 10^8 células se resuspenden en 1 ml de disolución estéril de composición 10 mM de acetato de litio, 0.6 M de Sorbitol, 10 mM de Tris/HCl, 10 mM de DTT, pH 7.5, considerando que, para *Pichia pastoris*, OD_{600} = 1 corresponde a 3 x 10^7 células (Wu y Letchworth, 2004).
- Una vez resuspendidas, se les deja reposar 30 minutos a temperatura ambiente (en torno a 23 ºC).
- Pasado este tiempo, se centrifugan 5 minutos a 2000 xg y 4 ºC y se resuspenden en 1.5 ml de una disolución estéril y fría de 1 M Sorbitol.
- Se repite este procedimiento de lavado tres veces.
- Finalmente, las células se resuspenden en 1 ml de Sorbitol 1 M por cada 10^{10} células y se conservan en hielo hasta su uso, que ha de tener lugar no más de dos o tres horas después.

Pese a que hay especies de levaduras que pueden conservarse a largo plazo a -80 ºC sin perder esta electrocompetencia (Suga *et al*., 2000), no se ha comprobado que suceda para el caso de *Pichia pastoris*, por lo que este proceso se lleva a cabo cada vez que se desea electroporar.

4.2.2. Electroporación de las construcciones en la cepa de *P. pastoris* SMD1168.

Una vez preparadas las células electrocompetentes, las electroporaciones de todos los plásmidos empleados en este trabajo se han realizado siguiendo el protocolo sugerido por Invitrogen.

De esta forma, en primer lugar se producen en torno a 50 µg del ADN plasmídico que se quiere electroporar, para lo cual se inoculan una media de 6 precultivos de 10 ml de *Low Salt* LB (Apéndice 10.1) y 25 µg/ml de zeocina con células DH5α transformadas con el plásmido deseado. Una vez crecidos, se centrifugan y se extrae el ADN de los mismos siguiendo el mismo protocolo que el descrito en el apartado 4.1.2. Una vez eluido el ADN, se realizan dos

digestiones de cada construcción con la enzima de restricción *SacI-HF* (New England Biolab), una de las cuales servirá de reserva. Suponiendo un rendimiento de obtención ADN razonable, los componentes de la reacción de digestión son los siguientes (Tabla 4.3):

Tabla 4.3. Componentes de la linearización de los plásmidos

Componente	Concentración inicial	Concentración final	Volumen (µl)
ADN molde	≈140 µg/ml	0,1 µg/ml (25 µg totales)	180
SacI-HF (NEB)	20000 U/ml	560 U/ml	7
Buffer de *SacI-HF*, CutSmart (NEB)	10 x	1 x	25
H_2O Milli-Q			38
Volumen final			250

Las digestiones se incuban a 37 ºC durante la noche y al día siguiente se verifica la correcta linearización de la muestra en un gel de agarosa al 1 % (apartado 4.1.3). Una vez corroborada ésta, se purifica con el kit comercial Accuprep Gel Purification de Bionner, eluyendo el ADN final en 26 µl de H_2O Milli-Q a 65 ºC. Este ADN lineal, del que en un proceso normal obtendremos entre 3 y 10 µg por muestra, puede conservarse a -20 ºC hasta el día en el que se tenga previsto preparar las células electrocompetentes y electroporar.

Llegado el día y preparadas las células como se ha descrito en el apartado anterior, se mezclan en esterilidad 80 µl de éstas con el ADN lineal y se transfieren a una cubeta de electroporación de 0.2 cm de grosor (Gene Pulser/Micropulser Electroporation Cuvettes, Bio-Rad), completamente estéril y en hielo (la noche anterior, como mínimo, han debido introducirse las cubetas en un congelador a -20 ºC). Se deja incubar la mezcla 5 minutos, tras los que se la somete a un pulso eléctrico de 2 kV en un electroporador Micropulser de Bio-Rad (son las condiciones pre-programadas para la electroporación de *Pichia pastoris* en este tipo de cubetas). Inmediatamente después del pulso, se añade 1 ml de Sorbitol 1 M a la cubeta, siempre en condiciones de esterilidad, y se trasvasa la mezcla a un falcon de 15 ml estéril, que se deja recuperando entre 1 y 3 horas a 30 ºC. Transcurrido este tiempo, se siembran 100 µl de la mezcla en una de las placas de YPDS con 200 µg/ml de zeocina (preparadas en el mismo momento que la placa control mencionada en el apartado anterior), se centrifuga el resto 5 minutos a 3800 xg y se resuspende el total del pellet restante en otros 100 µl aproximadamente, que se siembran en otra placa de las mismas características.

Estas placas, junto a las dos correspondientes al control negativo, en el que se ha realizado el mismo protocolo con células electrocompetentes que no se

han mezclado con ADN, se incuban a 30 ºC hasta la aparición de colonias (72-96 horas).

4.2.3. Prueba de expresión de las construcciones en *P. pastoris*

De las colonias crecidas se seleccionan una media de siete por construcción, que se estrían de forma diferenciada y numerada en una placa nueva de YPD (Apéndice 10.1) con 200 µg/ml de zeocina. La misma punta de pipeta con la que se ha estriado cada colonia se introduce, en esterilidad, en un falcon de 50 ml en el que previamente se ha añadido 1 ml del medio de crecimiento BMGY (Apéndice 10.1). Se incuban así todas las colonias a 30 ºC y 250 rpm hasta el día siguiente, durante un periodo aproximado de 16-20 horas.

Una vez transcurrido este periodo, las muestras se centrifugan 10 minutos a 3800 xg y 4 ºC, tras lo que, en esterilidad, se descarta el sobrenadante y se resuspenden totalmente los pellets en 1 ml del medio de expresión BMMY (Apéndice 10.1). Se incuban a 18 ºC y 250 rpm durante 72 horas, adicionándoles a las 24 y a las 48 horas un 1 % de metanol puro a cada una. Transcurrido este periodo, se centrifugan 10 minutos a 4 ºC y 3800 xg y se recuperan los sobrenadantes, que posteriormente se analizan por electroforesis desnaturalizante en SDS-PAGE (apartado 4.3.1.2). Para cada caso se selecciona la colonia que muestra una mayor expresión de la proteína deseada.

4.2.4. Conservación en glicerol de las colonias *P. pastoris*

Una vez seleccionada la colonia que mejor expresa la construcción deseada, se estría ésta de forma individual en una placa de YPD (Apéndice 10.1) con 200 µg/ml de zeocina, que se deja crecer unas 48 horas a 30 ºC. Una vez crecida y siempre en esterilidad, se resuspenden tres cuartas partes de la misma en un falcon de 15 ml en el que se han preparado 3 ml de una disolución estéril compuesta por un 70% de YPD (Apéndice 10.1) y un 30% de glicerol. Una vez resuspendida la masa celular en la disolución, se distribuye ésta en tres eppendorfs adecudamente etiquetados, que pueden almacenarse ya a -80 ºC todo el tiempo que sea necesario y sin que se llegue a apreciar pérdida en la expresión.

4.2.5. Expresión de proteínas en *P. pastoris*

Para la expresión a gran escala en *Pichia pastoris* de las distintas proteínas recombinantes que se emplean en esta tesis se ha seguido el protocolo sugerido por Invitrogen (Invitrogen, 2010), que no es sino el descrito en el apartado 4.2.3

para 1 ml pero escalado a 250 ml, 500 ml o 1 l, en función de la cantidad de proteína requerida.

Así, para la expresión del constructo deseado se siembran en primer lugar 60 µl del glicerol correspondiente (apartado 4.2.4) en una placa de YPD (Apéndice 10.1) con 200 µg/ml de zeocina, y se dejan incubar a 30 ºC durante 48-72 horas hasta la aparición de numerosas colonias. A partir de aquí, se resuspende aproximadamente una cuarta parte de la placa por cada 250 ml del medio de crecimiento BMGY (Apéndice 10.1), que se incuban en un matraz estéril de dos litros de volumen. Esta relación entre el volumen del matraz y el volumen de cultivo es importante, dado que las levaduras necesitan la mayor aireación posible para crecer de forma adecuada (Ahmad *et al*., 2014). La incubación tiene lugar durante 16-20 horas a 30 ºC y 250 rpm. Transcurrido este periodo, se centrifuga el cultivo en esterilidad, a 6000 xg y 10 ºC, durante 6 minutos. Tras desechar el sobrenadante, se resuspende vigorosamente el pellet en otros 250 ml del medio de expresión, BMMY (Apéndice 10.1), que se devuelven al matraz y se incuban a 18 ºC y 250 rpm durante 72-96 horas, añadiendo un 1% de metanol puro al cultivo cada 24 horas.

Transcurrido este tiempo, se centrifuga el sobrenadante resultante de la expresión dos veces: una primera de 5 minutos a 4 ºC y 10000 xg y una segunda de 20 minutos a 4 ºC y 10000 xg. Tras ello, en el caso de la *Hs*GalNAc-T2, sus mutantes y quimeras y la *Hs*GalNAc-T4, se añade EDTA 0.5 M de pH 7.5 hasta alcanzar una concentración final de 5 mM y β-mercaptoetanol hasta alcanzar una concentración de 10 mM (Fritz *et al*., 2006).

4.3. PURIFICACIÓN DE PROTEÍNAS

4.3.1. Purificación de *Hs*GalNAcT2, los mutantes de ésta y *Hs*GalNAcT4

4.3.1.1. Concentración y diálisis

El sobrenadante del cultivo de *Pichia pastoris*, ya centrifugado (apartado 4.2.4), se filtra con un equipo de filtración al vacío a través de filtros de membrana (Merck-Millipore; Figura 4.3) de 0.45 µm de poro en primer lugar y de 0.22 µm en segundo. Una vez filtrado, se concentra el sobrenadante con la ayuda de un concentrador Pellicon XL Biomax 10 kDa (Merck-Millipore) hasta un volumen final de 20-40 ml (Wu *et al*., 2006; Figura 4.3).

Figura 4.3. Equipo de filtración al vacío y concentrador. A la izquierda, sistema de filtración al vacío. A la derecha, concentrador Pellicon XL Biomax.

Al sobrenadante concentrado se le realizan tres diálisis a 4 ºC durante in mínimo de dos horas cada una. El tampón empleado para estas proteínas es 25 mM Tris/HCl, 5 mM EDTA, 10 mM β-mercaptoetanol, pH 7.5.

4.3.1.2. Electroforesis desnaturalizante y tinción

La electroforesis desnaturalizante en geles de poliacrilamida es una técnica clásicamente empleada para la separación y visualización de proteínas durante y tras la purificación de las mismas (Fishman y Moyer, 1942). El fundamento de la misma reside en la separación en función de su peso molecular de una muestra proteica cargada negativamente en un gel de acrilamida, al someterla a una diferencia de potencial eléctrico.

Para su realización, las muestras se desnaturalizan térmicamente (5 minutos a 100 ºC) en presencia de una disolución de carga (Apéndice 10.1) que, entre otros componentes, contiene dodecil sulfato sódico (SDS), un detergente aniónico que se une inespecíficamente a las proteínas y les confiere una relación tamaño/carga constante. Una vez cargadas las muestras en el gel, con una fuente de alimentación se les aplica un campo eléctrico a voltaje constante de 120 V durante unos 90 minutos. Al hacer esto, las proteínas migran por la red de acrilamida a una velocidad inversamente proporcional al logaritmo de su peso molecular y se pueden comparar con un patrón estándar (marcador de peso molecular, en nuestro caso EZ-Run prestained recombinant protein ladder, de Thermo Fisher; Figura 4.4).

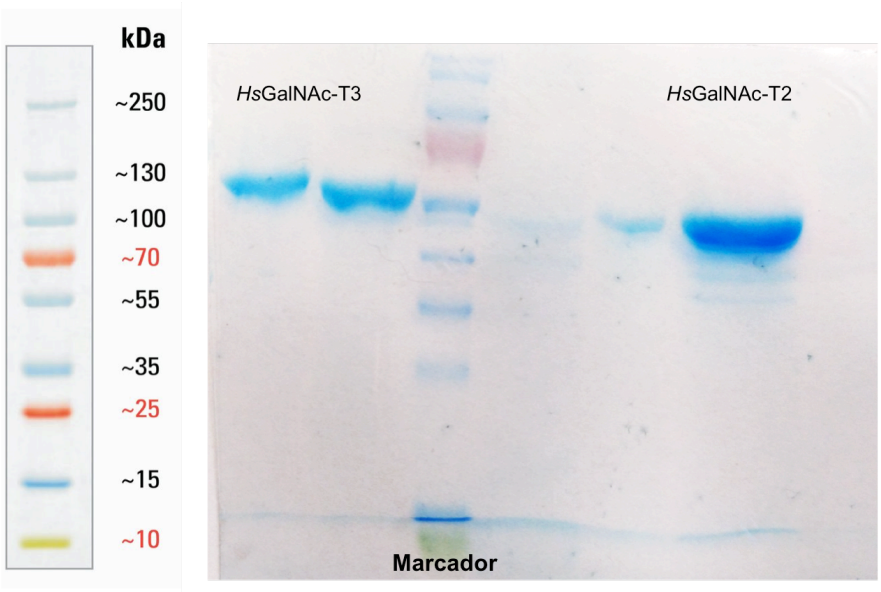


Figura 4.4. Ejemplo de electroforesis desnaturalizante. A la izquierda se observa el patrón del marcador de peso molecular EZ-Run de Thermo Fisher y a la derecha un ejemplo de gel desnaturalizante. En él se observan dos muestras de *Hs*GalNAc-T3, con y sin *N*-glicosilaciones, seguidas del marcador y de tres muestras correspondiente al primer paso de purificación de *Hs*GalNAc-T2.

En las distintas purificaciones descritas en esta tesis se han empleado geles de acrilamida convencionales, conteniendo un 10% o un 12% de ésta. La visualización de las bandas proteicas se ha llevado a cabo por tinción con coomassie y posterior destinción con una disolución de metanol y ácido acético (Apéndice 10.1), tras la que sólo quedan en el gel las bandas correspondientes a la proteína presente en el mismo (Figura 4.4).

4.3.1.3. Cromatografía de afinidad con HiTrap Blue

La cromatografía de afinidad es una de las técnicas más comúnmente empleadas para la purificación de proteínas. A rasgos generales, se basa en la generación de interacciones bioespecíficas entre la proteína presente en la muestra y la matriz cromatográfica de la columna, que permiten separar a la primera de gran parte de las impurezas que la puedan acompañar (Cuatrecasas, 1970; Figura 4.5).

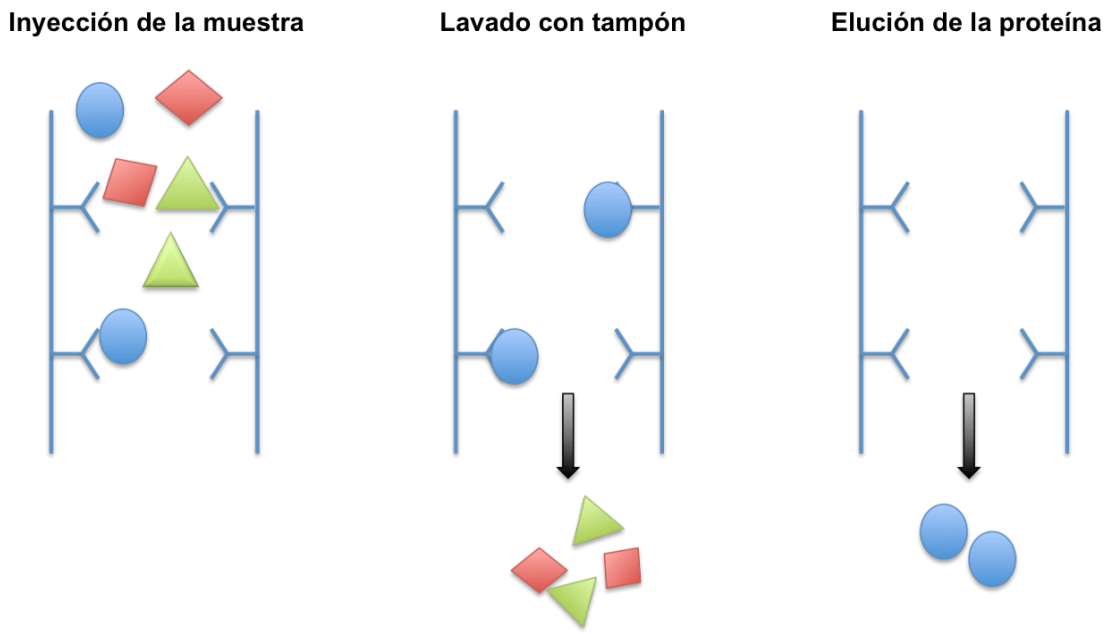


Figura 4.5. Esquema de una purificación por cromatografía de afinidad. De izquierda a derecha, en la muestra original se encuentra la proteína de interés, que va uniéndose de forma específica a la matriz de la columna. En la fase de lavado las proteínas unidas inespecíficamente se liberan, permitiendo la recuperación de las proteínas deseadas en la fase de elución.

Respecto a la purificación tanto de la *Hs*GalNAc-T2, como de sus diversos mutantes y de la *Hs*GalNAc-T4, el paso inicial de las mismas consiste en una cromatografía de afinidad a la columna HiTrap Blue (GE Healthcare). Esta columna está pre-empaquetada con el colorante Cibacron Blue F3GA, que mimetiza nucleótidos y permite que algunas proteínas de unión a estos queden retenidas en su interior (Wilson, 1976). Dado que tiene un volumen total de 5 ml, se realiza en primer lugar un lavado a 5 ml/min con dos o tres volúmenes de H_2O Milli-Q, seguidos de un equilibrado con el tampón A, que es el mismo en el que habíamos dializado la proteína previamente (25 mM Tris/HCl, 5 mM EDTA, 10 mM β-mercaptoetanol, pH 7.5). Una vez equilibrada, se realiza la inyección de la proteína a un flujo de 1 ml/min, para así favorecer la unión de ésta a la matriz de Cibacron. Tras la carga se lavan las posibles proteínas que hubiesen podido quedar unidas a la columna con baja especificidad pasando 2 o 3 volúmenes de tampón A, y finalmente se eluye la proteína de interés con un cambio al 100% de tampón B (25 mM Tris/HCl, 5 mM EDTA,1 M NaCl, 10 mM β-mercaptoetanol, pH 7.5).

Todo este proceso se lleva a cabo en un equipo de cromatografía líquida a baja presión (FPLC) tipo ÄKTA (GE Healthcare), a 6 ºC y manteniendo la presión en el lecho de la columna por debajo de 0.15 MPa. Tras la elución de la proteína, la columna se regenera con 20 ml de H_2O Milli-Q seguidos por 10 ml de cloruro de guanidinio 6 M y otros 50 ml de H_2O Milli-Q. Finalmente, se pasan 20 ml de una disolución de conservación compuesta por 20 mM de KH_2PO_4 pH 6 y 20% de etanol, en la que se guarda la columna a 4 ºC (Figura 4.6).

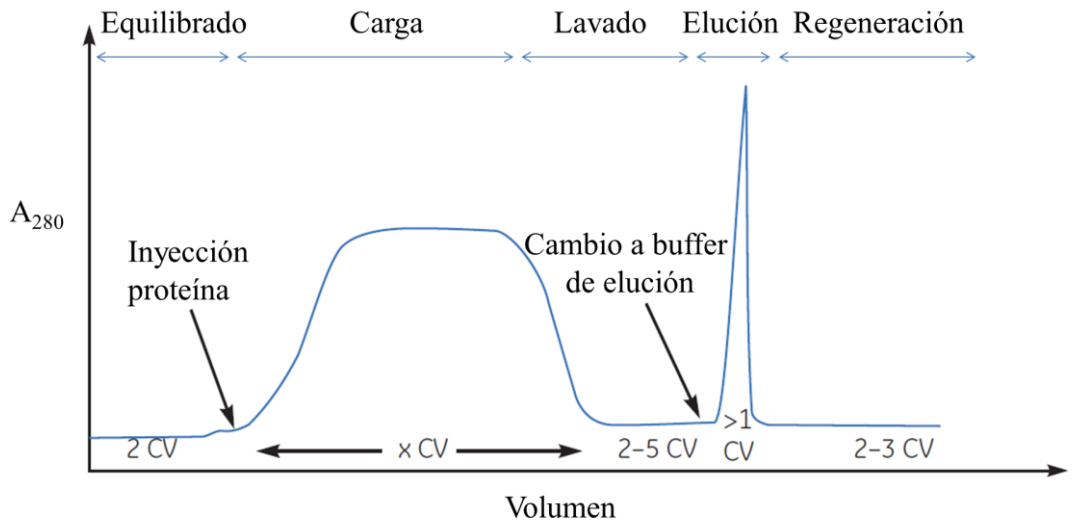


Figura 4.6. Esquema del registro en un ÄKTA de una cromatografía de afinidad típica. El proceso se sigue mediante la absorbancia a 280 nm. En primer lugar, se equilibra la columna con dos volúmenes de columna (CV) del tampón deseado. Una vez hecho esto se carga la mezcla de proteínas, de las que la mayor parte no se unen a la columna, y posteriormente se procede al lavado que arrastra las uniones inespecíficas hasta que no se observa absorbancia a 280 nm. Por último, la proteína objetivo se eluye con el tampón de elución correspondiente, tras lo que se limpia y regenera la columna.

4.3.1.4. Cromatografía de exclusión molecular

La cromatografía de exclusión molecular se realiza con columnas de matriz porosa y canales de tamaño selectivo que hacen que las muestras queden retenidas a distintos tiempos en función de su peso molecular, lo que a su vez permite el aislar la proteína de restos de impurezas y de posibles agregados, la presencia de los cuales dificulta enormemente la posterior cristalización de la proteína (van der Laan *et al.*, 1989). En nuestro caso, se emplea la columna HiLoad 26/60 Superdex 75, con un rango de separación óptimo entre 3 y 70 kDa que abarca todas las proteínas con las que trabajamos.

De esta forma, la muestra de *Hs*GalNAc-T2, el mutante de la misma que corresponda o *Hs*GalNAc-T4, proveniente en cualquier caso de la elución de la HiTrap Blue y una vez visualizada en el correspondiente gel de acrilamida (apartado 4.3.1.2), se separa de los restos de impurezas y agregados mediante este tipo de cromatografía. Se introduce la muestra a un flujo de 1.5 ml/min en la columna previamente equilibrada en tampón 25 mM Tris/HCl, 5 mM EDTA, 150 mM NaCl, 10 mM β-mercaptoetanol, pH 7.5, y se recogen las alícuotas correspondientes al volumen de elución de la proteína monomérica, verificando su pureza mediante electroforesis en gel de acrilamida.

4.3.1.5. Intercambio de tampón

Un paso habitual en muchas purificaciones es el de intercambio de tampón. En el caso de *Hs*GalNAc-T2, sus mutantes y *Hs*GalNAc-T4, se lleva a cabo cuando el objetivo no es el de cristalizar la proteína y no se realiza el último paso de exclusión molecular. Si bien el intercambio se puede hacer mediante diálisis o centricón (apartado 4.3.1.6), es muy común el empleo de un dispositivo de *desalting* cuando los volúmenes oscilan entre 10 y 40 ml, que es lo corriente tras un paso cualquiera de purificación. En ese caso se emplea una columna HiPrep 26/10 Desalting (GE Healthcare), la cual, mediante un proceso de exclusión molecular, permite separar a través de una matriz porosa las moléculas grandes (en nuestro caso, la proteína) de las pequeñas (los componentes del tampón en el que la muestra se encuentra disuelta), que quedan retenidas más tiempo.

Así, se procede al equilibrado de la columna con 3 volúmenes de H_2O Milli-Q y otros tres del tampón en el que se quiere dejar la muestra. Se inyecta un volumen no superior a 15 ml de la muestra a un flujo de 4 ml/min y se recogen alícuotas hasta que acaba de eluir la proteína (comienza a hacerlo pasados unos 20 ml). En cuanto esto sucede se puede apreciar el cambio de conductividad asociado a la salida del tampón inicial en el que la misma estaba disuelta (Figura 4.7). Si es necesario, se repite la operación con el volumen restante de muestra una vez que la conductividad vuelve a los valores iniciales. Finalizado el proceso y con la proteína disuelta en el nuevo tampón, si bien algo más diluida (1.3 veces más, aproximadamente), se procede a lavar la columna de Desalting con 3 volúmenes de H_2O Milli-Q y dos de etanol al 20%.

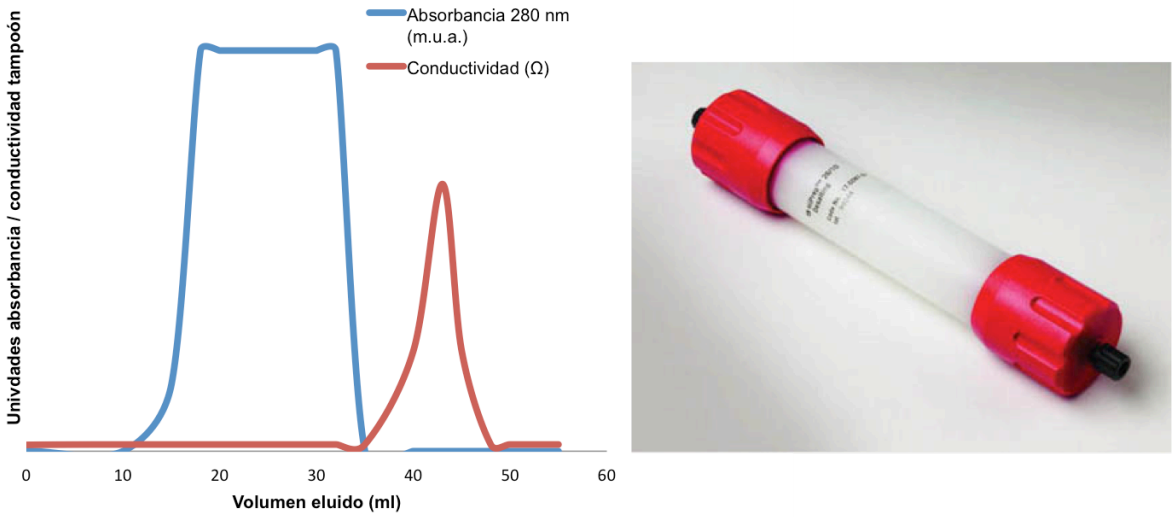


Figura 4.7. Intercambio de tampón mediante Desalting. A la izquierda, perfil de absorbancia y conductividad de un intercambio de tampón típico con columna de Desalting HiPrep 26/10. A la derecha, fotografía de la columna.

4.3.1.6. Concentración y conservación

Las alícuotas obtenidas tras la exclusión molecular o el intercambio de tampón (típicamente 15-20 ml) se concentran por ultrafiltración en centricones de corte de poro de 30 kDa (Amicon Ultra-15 centrifugal Filter de Merck Millipore; Figura 4.8) con los que también se procede a cambiar el tampón en el que se encuentra la muestra, realizando tres lavados con tampón 25 mM Tris/HCl, 0.5 mM EDTA, pH 7.5. Una vez alcanzado un volumen final cercano a 300 µl, se le añade TCEP.HCl 455 mM hasta alcanzar una concentración final de 1 mM. En estas condiciones se determina la concentración final de la alícuota (apartado 4.6.1) y, en el momento oportuno, se congela en N_2 líquido y se guarda a -80 ºC.

Figura 4.8. Centricón. Amicon Ultra-15 centrifugal Filter de Merck Millipore.

4.3.2. Purificación de las quimeras

Las quimeras que se lograron expresar y, consecuentemente, se emplearon en el Artículo II (véase Tabla 2 en de las Rivas *et al*., 2017), se purifican de forma idéntica a la descrita para *Hs*GalNAc-T2, *Hs*GalNAc-T4 y mutantes (apartado 4.3.1). La variación en la secuencia del *flexible linker* entre los dominios catalítico y de lectina no altera su comportamiento a lo largo de los diferentes pasos de purificación y siguen interaccionando con la matriz de la columna HiTrap Blue. No sucede lo mismo con su expresión que, pese a ser suficiente, es considerablemente menor que en el caso de las proteínas silvestres.

4.4. SÍNTESIS DE PÉPTIDOS, GLICOPÉPTIDOS Y GLICOMIMÉTICOS

Los diferentes péptidos y glicopéptidos empleados en los artículos I, II y V (véase capítulo Resultados) han sido diseñados de forma racional, tal y como se explica en el desarrollo y en la información suplementaria correspondientes, a

partir de trabajos previos en los que se estudiaba el perfil de glicosilación de la ppGalNAc-T2 (Raman *et al*., 2008).

Si bien el diseño se ha realizado en nuestro grupo, la síntesis orgánica y purificación de los mismos ha sido realizada por el grupo del Doctor Francisco Corzana, en la Universidad de La Rioja, siguiendo la metodología que se detalla en la sección experimental de los respectivos artículos.

Respecto al diseño de los compuestos glicomiméticos empleados en el artículo III (véase capítulo Resultados), estos fueron diseñados y sintetizados por el grupo del Doctor Pedro Merino, de la Universidad de Zaragoza, siguiendo para ello el protocolo descrito en el artículo correspondiente.

4.4.1. Preparación de las disoluciones de péptidos, glicopéptidos y glicomiméticos

Todos los péptidos sintetizados en el laboratorio del Doctor Corzana, así como los compuestos glicomiméticos del laboratorio del Doctor Merino, se reciben puros y liofilizados. Debido a su carácter generalmente muy ácido, incompatible con el rango de pH en el que las diversas ppGalNAc-Ts son estables, se disuelven en una disolución 25 mM Tris/HCl, pH 7.5 y se ajusta el pH con NaOH 2 M y papel indicador de pH (Fisher) hasta dejarlos a una concentración final de 100 mM de compuesto o péptido y un pH entre 7 y 8.

4.5. TÉCNICAS BIOFÍSICAS DE ANÁLISIS ESTRUCTURAL

4.5.1. Cristalización de proteínas

La cristalización de proteínas es el proceso mediante el cual las moléculas de proteína homogénea, en unas condiciones determinadas, se ordenan de forma repetitiva y periódica a lo largo de las tres dimensiones espaciales, generando lo que conocemos como un cristal. La difracción por Rayos X de estos cristales permite, después del procesamiento matemático adecuado, la resolución de su estructura tridimensional con una resolución inferior a 2 Å (Figura 4.10). Hoy en día, y pese a las dificultades experimentales que conlleva la obtención de cristales de proteínas, sigue siendo la técnica más empleada para la caracterización estructural de proteínas, con más del 86% de las estructuras depositadas en el Protein Data Bank (PDB) obtenidas gracias a ella (Dessau y Modis, 2011).

Las diferentes condiciones en las que se han logrado cristalizar cada una de las proteínas o/y complejos proteicos obtenidos a lo largo de los cinco trabajos reunidos en esta tesis doctoral se describen de forma pormenorizada en el apartado experimental de cada uno de los artículos (véase capítulo Resultados).

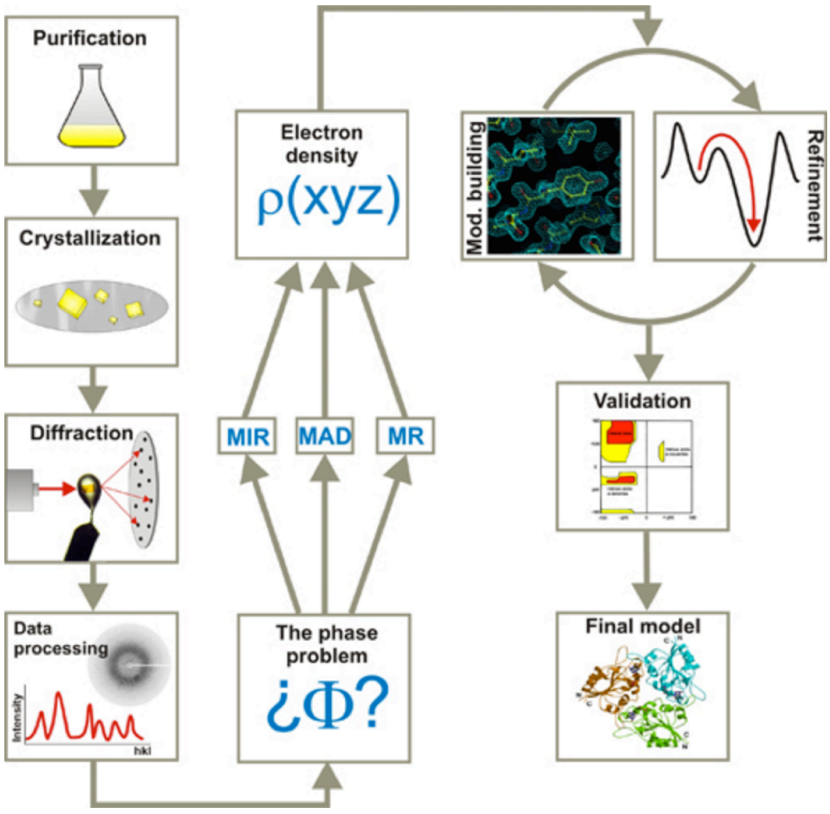


Figura 4.10. Esquema del proceso de la cristalografía de proteínas. A partir de la proteína purificada se obtienen cristales en una condición determinada y se difractan en una fuente de Rayos X. Los datos de difracción obtenidos, una vez procesados, aportan información acerca de las coordenadas de posición de los diferentes átomos, si bien la posición absoluta de las mismas no puede asignarse hasta que se resuelve el problema del faseado, que otorgaría una información comparable a la de un origen de coordenadas. En nuestro caso, este problema se soluciona normalmente por reemplazamiento homólogo con una estructura previamente resuelta de elevado % de identidad, como puede ser otra ppGalNAc-T. Tras ello, se obtiene un mapa de la densidad electrónica de la molécula, que se refina con los datos experimentales hasta validar su viabilidad estructural y obtener una representación final de la estructura de la proteína. **Fuente**: Párraga *et al*., 2010.

4.5.1.1. Cocristalización

El procedimiento habitual para la búsqueda de las condiciones iniciales de cristalización de cada proteína cristalizada en esta tesis doctoral ha sido el de la obtención de cristales por difusión de vapor en gota sentada (Davies y Segal, 1976). Los experimentos se realizan en placas de 96 pocillos, a cada uno de los cuales se añaden 60 µl de una de las 96 condiciones de los cribados

comerciales JCSG HTS, Classic I, Basic, PEG-SALT (Jena Bioscience), $AmSO_4$ suite (QUIAGEN), Proplex, Midas y Morpheus (Molecular Dimensions). De estos 60 µl, se mezclan 0.5 µl de cada condición comercial (o líquido madre) con 0.5 µl de la proteína, ya sea sola o en presencia de sus ligandos. La concentración inicial de proteína oscila en torno a los 5-10 mg/ml, si bien se ajusta de forma experimental en cada caso verificando que la proporción de gotas precipitadas al hacer el primer cribado se encuentre en torno al 40%. La concentración de los péptidos sustratos y de UDP y Mn^{2+} ha de estar siempre muy en exceso respecto a la de la proteína para forzar su posible entrada dentro de la misma. Tal y como se detalla en las correspondientes secciones experimentales de los artículos, las condiciones habituales son 5 mM para los péptidos y el UDP y 2 mM para el Mn^{2+}. Una vez puestas las gotas, las placas se almacenan a 18 ºC en una granja cristalográfica termostatizada (Formulatrix) que va fotografiando las gotas a intervalos cada vez más espaciados durante un total de tres semanas.

Una vez determinada cada condición inicial de cristalización, se diseñan cribados que modifican parámetros de la misma tales como la concentración de agente precipitante, el pH de la disolución tampón o la concentración de los posibles aditivos. Identificadas las condiciones óptimas de cristalización dentro de esta mejora, puede recurrirse también a poner una placa en la que a cada gota se le añaden 0.25 µl de uno de los compuestos que componen el cribado de aditivos de Hampton Research. El objetivo final es el de obtener cristales de buen tamaño, resitentes y bien formados (Figura 4.11), que hayan integrado el ligando en su estructura.

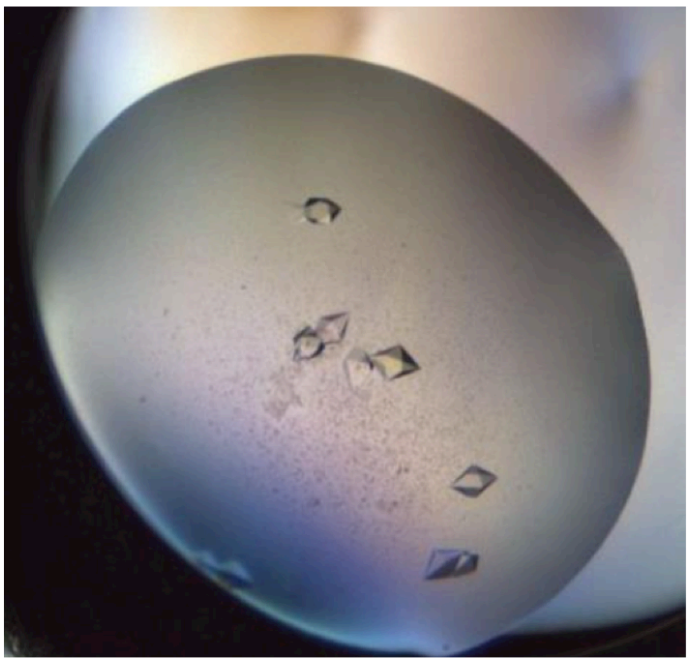


Figura 4.11 Ejemplo de cristales de proteína. Cristales de la HsGalNAc-T2 cocristalizados con el péptido MUC5AC-13, UDP y Mn^{2+}.

4.5.1.2. Inmersión de ligandos (*soaking*)

En las ocasiones en las que las condiciones de cocristalización no han logrado que entre el ligando en la proteína o si éste, como sucede en el caso de de la luteolina (véase artículo IV), no es totalmente soluble en el tampón empleado, se recurre a la técnica de *soaking*. En ésta, los cristales previamente obtenidos de la proteína se traspasan de su gota original a una nueva gota en la que se encuentra el compuesto que se quiere introducir a gran concentración (de 20 a 50 mM, según la solubilidad, el aguante del cristal y la disponiblidad del compuesto), disuelto en el líquido madre de la condición del cristal. Una vez introducido el cristal dentro de la nueva gota, se va observando su evolución durante el tiempo de incubación y se saca de la misma cuando se ve dañado o, si aguanta sin resquebrajarse, transcurridas varias horas e incluso días.

Todos las operaciones que se les hacen a los cristales se realizan empleando los llamados crioloops (Hampton), que no son sino pequeños lazos de material transparente a la radiación de Rayos X y de tamaño del orden de micras con los que los cristales se cogen de la gota mientras se observan aumentados a través de un microscopio (Figura 4.12).

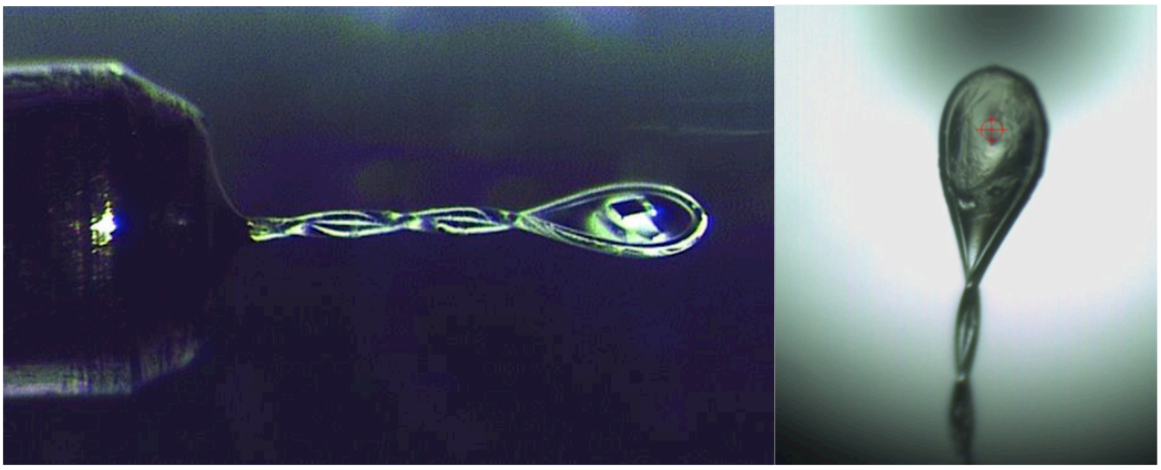


Figura 4.12. Montaje de un cristal en un loop. A la izquierda, imagen de un cristal en el centro de un *loop*, unido a su base (**Fuente**: Mac Sweeney y D'Arcy, 2003). A la derecha, imagen de un cristal de *Hs*GalNAc-T4 tomada durante una difracción remota en el sincrotrón Diamond Light Source (DLS), en marzo de 2017. El punto rojo es el lugar en el que incide el láser de Rayos X cuando se le da la orden, y el cristal ocupa todo el centro del *loop*.

4.5.1.3. Toma de datos de los cristales

Los cristales, una vez formados y recolectados, se almacenan en N_2 líquido, mismo medio en el que posteriormente serán sometidos al haz de rayos X en la

línea de rayos X del sincrotrón correspondiente. Para evitar que la temperatura de dicho N_2 líquido (78 K) conlleve la congelación del agua interna de los cristales y cause la rotura o desestructuración de los mismos, antes de almacenarlos en los viales de transporte se sumergen en una solución crioprotectora. Las soluciones crioprotectoras de los diferentes cristales se componen o bien, en el caso de que el precipitante de la condición sea un compuesto de tipo PEG, de una disolución de la misma condición que aquélla en la que ha crecido el cristal pero añadiendo un 25% de un alcohol de bajo peso molecular como el etilenglicol o el glicerol, o bien con sulfato de litio saturado si el precipitante del líquido madre es una sal.

Los datos de los cristales resueltos en esta tesis doctoral se han tomado o bien en el sincrotrón ALBA, en Barcelona, o bien en el sincrotrón Diamond Light Source (DLS), en Oxford. Ambos presentan una características similares, si bien la potencia y capacidad del procesamiento del segundo es mucho mayor, por lo que los datos se pueden tomar a mayor velocidad (Figura 4.13).

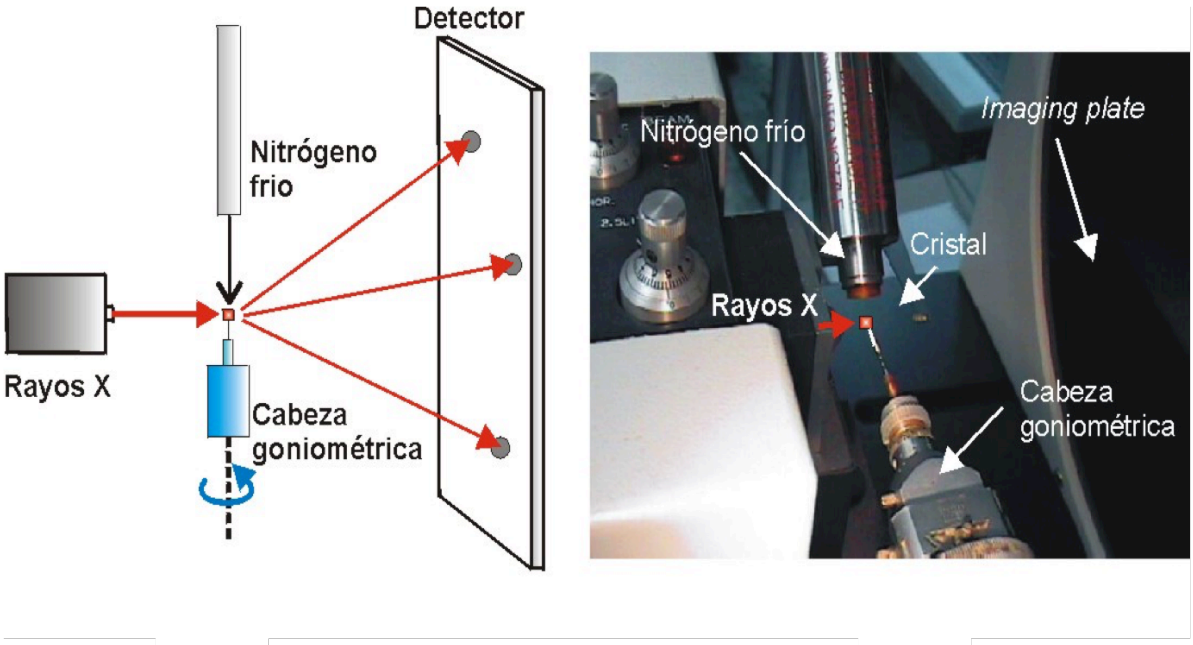


Figura 4.13. Toma de datos de difracción de rayos X. A la izquierda, esquema de la cabeza del láser de rayos X, la cabeza goniométrica en la que se coloca el loop con el crsistal montado, en línea con el haz de N_2 líquido, y el detector que recoge la señal difractada. A la derecha, imagen real de la línea BL-13 XALOC, en el sincrotrón ALBA, en Barcelona.

4.5.1.4. Resolución de las estructuras cristalográficas

Todos los cristales obtenidos para la consecución de los artículos recogidos en esta tesis doctoral se han procesado con el programa XDS (Kabsch, 2010) y el software CCP4 (Winn *et al.*, 2011). Tal y como se indica en la sección experimental de los artículos correspondientes (capítulo Resultados), las estructuras se resolvieron por reemplazamiento molecular usando Phaser (Winn

et al., 2011) y las siguientes entradas de PDB de estructuras de la ppGalNAc-T2 previamente resueltas: 2FFV en los Artículos I, III y IV, 5AJP en el Artículo II y 4D0T en el Artículo V (Fritz et al., 2006; Lira-Navarrete *et al*., 2014). La asignación de los aminoácidos correspondientes al mapa de densidad electrónica se realizó mediante ciclos de *building* en Coot (Emsley y Cowtan, 2004), seguidos de refinamiento en REFMAC5 (Murshudov *et al*., 2011). La validación de los modelos finales se llevó a cabo con el programa PROCHECK (Laskowski *et al*., 1993). Finalmente, las coordenadas y factores estructurales de cada estructura se depositaron en el World Wide Protein Data Bank (wwPDB) con el número de entrada que se indica en cada uno de los correspondientes artículos.

4.5.2. Dispersión de Rayos X de ángulo pequeño (SAXS)

La técnica de SAXS consiste, al igual que la cristalografía por difracción de rayos X, en el análisis de la dispersión de rayos X que produce una muestra. La diferencia entre ambas es que en el SAXS se registra la dispersión elástica de los rayos X en un rango angular muy bajo (entre 0.1 y 10 º), que aporta información acerca no ya de la posición de los átomos sino de la forma, densidad, nivel de agrupación y del tamaño de las macromoléculas (Glatter y Kratky, 1982). Esto permite, como sucede en el caso de la *Hs*GalNAc-T2 (Artículo I), obtener valiosa información complementaria acerca del comportamiento de la proteína en disolución y de su estado de oligomerización en presencia o ausencia de diferentes ligandos.

Las mediciones experimentales fueron realizadas por un miembro de nuestro grupo de investigación en el sincrotrón DESY, en Hamburgo. Cada medida se llevó a cabo en tampón 25 mM Tris/HCl, 10 mM $MnCl_2$, pH 7.5, en un volumen final de 50 µl. Se realizaron medidas de HsGalNAc-T2 en presencia y ausencia de diferentes ligandos, siempre en concentración 10 veces superior a la de la proteína (Tabla 4.5).

Tabla 4.5. Combinaciones de experimentos de SAXS con *Hs*GalNAc-T2

*Hs*GalNAc-T2
*Hs*GalNAc-T2 + UDP-GalNAc
*Hs*GalNAc-T2 + UDP
*Hs*GalNAc-T2 + EA2
*Hs*GalNAc-T2 + UDP + EA2
*Hs*GalNAc-T2 + MUC5AC-13
*Hs*GalNAc-T2 + UDP + MUC5AC-13
*Hs*GalNAc-T2 + MUC5AC
*Hs*GalNAc-T2 + UDP + MUC5AC

Todas estas combinaciones se realizaron a tres concentraciones diferentes de *Hs*GalNAc-T2: 10 mg/ml, 5mg/ml y 2.5 mg/ml.

Los resultados de los experimentos fueron analizados por el grupo de investigación del Doctor Pau Bernadó del Centro de Bioquímica Estructural de la Universidad de Montpellier, Francia.

4.5.3. Microscopía de fuerza atómica (AFM)

La microscopía de fuerza atómica es una técnica que, mediante un microscopio de fuerza atómica, permite obtener un mapa topográfico de la estructura molecular de una muestra (Doerr, 2015). Dado que se realiza en condiciones fisiológicas, resulta otra herramienta tremendamente útil para complementar la información estructural obtenida a partir de técnicas como la cirstalografía de rayos X. En proteínas como la *Hs*GalNAc-T2, donde la variación estructural de la macromolécula varía tanto en función de la presencia o ausencia de sus ligandos, resulta una técnica particularmente ilustrativa y visual (Artículo I).

Todos los experimentos y datos obtenidos mediante esta técnica fueron realizados por María del Carmen Pallarés y Carlos Marcuello, investigadores del grupo de la Doctora Ana Isabel Gracia Lostao, del Laboratorio de Microscopías Avanzadas del Instituto de Nanociencia de Aragón.

4.6. OTRAS TÉCNICAS BIOFÍSICAS PARA EL ANÁLISIS Y CARACTERIZACIÓN DE PROTEÍNAS

4.6.1. Absorción UV-Visible y cuantificación de proteínas

Las proteínas presentan un espectro de absorción (UV) dominado por dos grandes picos, uno en torno a los 225 nm y otro alrededor de los 280 nm (Creighton, 1997). El pico a 225 nm se corresponde con la absorción del enlace peptídico, mientras que el de 280 nm se debe a la absorción de los enlaces tipo sp^2 presentes en los anillos aromáticos de los aminoácidos tirusoina y, fundamentalmente, triptófano (Creighton, 1997).

A lo largo de esta tesis doctoral se ha empleado esta técnica de forma casi exclusiva para la cuantificación de proteínas. Otro uso común de la técnica es el de la detección de agregación, que conlleva una dispersión de luz apreciable como absorción por debajo de 450 nm y que aumenta conforme disminuye la longitud de onda de emisión (Creigton, 1997).

Para cuantificar las proteínas obtenidas en esta tesis se ha determinado su absorción UV-Visible a 280 nm y su correlación con el coeficiente de extinción téorico (Apéndice 10.2) , fáclmente obtenible en varios servidores informáticos (ProtParam; Noble y Bailey, 2009). Este coeficiente de absorción se relaciona con la absorbancia a 280 nm según la ecuación de Beer-Lambert:

$$A = \varepsilon \cdot C \cdot l$$

donde A es la absorbancia medida experimentalmente a una longitud de onda, ε es el coeficiente de extinción molar de la proteína ($M^{-1}cm^{-1}$) a esa misma longitud de onda, C es la concentración y l es la longitud del paso óptico de la muestra. En nuestro caso, las variaciones en el coeficiente de extinción a 280 nm entre proteínas silvestres y mutadas suele ser muy pequeña, dado que como mucho afecta a uno de los aminoácidos implicados en la absorción (Artículo I, mutante Trp282Ala de la *Hs*GalNAc-T2) y se trata de proteínas de considerable número de residuos (Apéndice 10.2). Es por ello por lo que muchas veces se emplea el de las primeras al realizar cálculos no analíticos.

A nivel experimental, la medida de absorbancia a 280 nm se ha realizado en un espectrofotómetro Cary (*Varian*; Figura 4.14), en cubetas de cuarzo de 1 cm de paso óptico. Se realizaba primero una medida de referencia con 120 µl de tampón y después se medía la muestra concentrada realizando una dilución 1:120 de la misma como primera aproximación, adecuada para concentraciones de proteína en el rango micromolar.

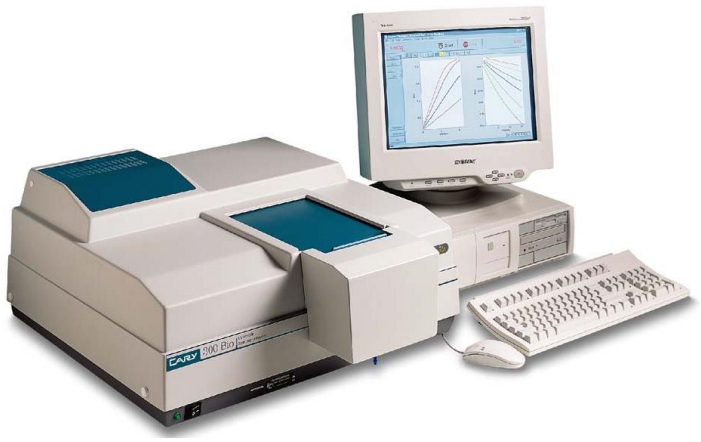


Figura 4.14. Espectrofotómetro UV-Vis. Espectrofotómetro Cary (Varian) utilizado para la obtención de espectros UV-Vis a lo largo de esta tesis.

4.6.2. Determinación de constantes de disociación por espectroscopía de fluorescencia

La emisión de fluorescencia es un fenómeno producido cuando moléculas que previamente han absorbido energía mediante su exposición a luz de una determinada longitud de onda emiten luz a una longitud de onda mayor. El espectro de emisión de fluorescencia de las proteínas está dominado por la contribución de los triptófanos, cuyo máximo de absorción se sitúa en torno a 280 nm y su máximo de emisión en torno a 350 nm (Creighton, 1997). Tanto la intensidad de la emisión como la longitud a la que se sitúa el máximo se ven afectadas por el entorno cercano de dichos triptófanos, de forma que los cambios en la estructura de una proteína son fácilmente detectables mediante esta técnica (Creighton, 1997; Sancho, 2013). Variaciones en la fluorescencia del triptófano son, así, una herramienta directa para la determinación de constantes de disociación entre proteínas y sus ligandos, que al interaccionar con el dominio catalítico de éstas modifican la fluorescencia de los triptófanos presentes según la ecuación siguiente:

$$\frac{1 - F_1}{F_0} = \frac{f_a \cdot [Q]}{K_d + [Q]}$$

donde F_0 es la fuorescencia de la enzima en ausencia de ligando, F_1 es la fluorescencia de la enzima en presencia de una concentración dada de ligando [Q], f_a es el grado fraccionario de fluorescencia y K_d es la constante de disociación proteína-ligando (La *et al*., 2008).

Los diversos experimentos, tanto en el caso de la *Hs*GalNAc-T2 como en el de su mutante Phe104Ser, se han llevado a cabo en un espectrofluorímetro Cary Eclipse (*Varian*). Todas las medidas se realizan a una concentración total de proteína de 1 µM, disuelta en 1 ml de tampón 25 mM Tris/HCl, 400 µM UDP, 1 mM $MnCl_2$, pH 7.5, excepto en el caso de los inhibidores glicomiméticos (Artículo III), donde se eliminan el UDP y el $MnCl_2$ del tampón. Los diferentes péptidos y compuestos, disueltos en 25 mM Tris/HCl, pH 7.5, se añaden en concentraciones crecientes, desde 1 µM hasta 300 µM. Las reacciones se incuban 30 segundos a 25 ºC antes de cada medida, que se realizan excitando a 280 nm y registrando la emisión entre 300 nm y 400 nm.

Las constantes de disociación se determinan ajustando los datos obtenidos para el máximo de emisión de fluorescencia a la ecuación anterior (Creighton, 1997), para lo que se emplea el programa Prism (GraphPad) y se considera un modelo de unión 1:1, en el que una molécula de ligando se une a una molécula de proteína.

4.6.3. Determinación de la temperatura de fusión (Tm)

Hablamos de temperatura de fusión de una proteína para referirnos a la temperatura a la que el 50% de moléculas de proteína presentes en la disolución se encuentran desplegadas, esto es, se han desnaturalizado. La determinación de la misma es un reflejo directo de la estabilidad de la proteína, así como sus posibles variaciones en presencia de distintas condiciones o de determinados ligandos. En el caso de esta tesis, es una técnica útil y sencilla para comprobar si un mutante es particularmente menos estable que la correspondiente proteína silvestre (Artículo V).

Para la determinación experimental de la Tm se realiza un ensayo termofluorescente en el fluorímetro FluoDIA T7.0 (Photal Otsuka Electronics). En placas de fondo cónico de 96 pocillos se depositan 100 µl de una disolución de la proteína correspondiente (*Hs*GalNAc-T2 o su mutante Phe104Ser) a concentración 2 µM y disuelta en tampón 25 mM TRIS/HCl, 1 mM TCEP.HCl, pH 7.5. Se le añade también el fluoróforo 1-anilino-8-naftaleno sulfonato (ANS) hasta alcanzar una concentración final por pocillo de 100 µM. La presencia de este fluoróforo posibilita el seguimiento de la desnaturalización de la proteína conforme la temperatura aumenta, dado que también lo hace la superficie expuesta al solvente con el desplegamiento y, por tanto, su capacidad de unión (Lo *et al*., 2004; Semisotnov *et al*., 1991; Figura 4.15).

A nivel experimental, se realiza la excitación de la mezcla a 390 nm y se recoge la emisión a 500 nm (longitudes de onda de excitación y emisión del ANS; Lo *et al*., 2004). Las medidas se realizan desde los 21 ºC hasta los 70 ºC, aumentando 0.5 ºC por minuto y recogiendo variaciones en la fluorescencia cada 0.1 ºC. Las muestras se realizan por triplicado y, para evitar la evaporación de la mezcla con el aumento de la temperatura, se le añaden a cada pocillo 50 µl de aceite mineral.

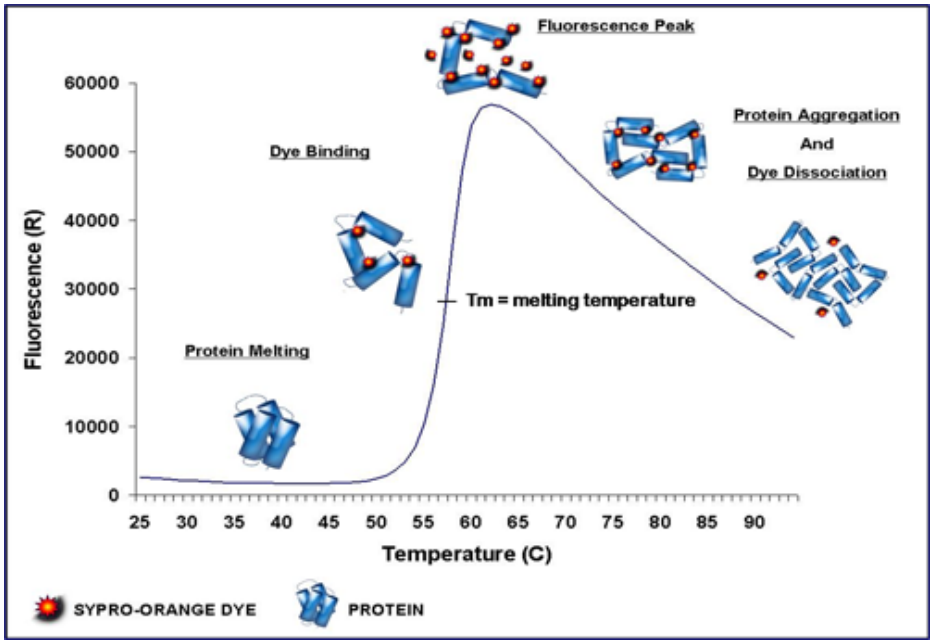


Figura 4.15. Esquema de ensayo de desnaturalización térmica seguido por fluorescencia. El funcionamiento del fluorófro Spyro-Orange es extrapolable al del ANS. **Fuente**: http://www.bio.anl.gov/molecular_and_systems_biology/Sensor/sensor_images/assay.png

Los datos recogidos se ajustan con el programa Origin (OriginLab) a la ecuación del modelo de Boltzmann:

$$I = \frac{(A + B) \cdot (B - A)}{1 + e^{Tm - T / slope}}$$

Donde I es la intensidad de la fluorescencia a la temperatura T, A y B son las intensidades de fluorescencia pretransicional y postransicional y *slope* es el factor de la pendiente.

4.6.4. Marcaje de los mutantes de *Hs*GalNAc-T2 con 3-bromo-1, 1, 1-trifluoroacetona

El marcaje de los mutantes T375C y F104S-T375C de la *Hs*GalNAc-T2 con el fluoróforo 3-bromo-1,1,1-trifluoroacetona (BFA) se realiza siguiendo exactamente el protocolo descrito por Rydzik *et al.* (2014). Tal y como éste indica, previa reducción de las proteínas mediante 10 minutos de incubación en hielo con 1 mM de TCEP.HCl, se incuba una mezcla de las mismas con BFA en un exceso molar 14:1 durante 5 minutos a temperatura ambiente y tampón 20 mM Na_2HPO_4, 200 mM NaCl, pH 7. Transcurrido este tiempo, se elimina el exceso de fluoróforo intercambiando el tampón gracias a una columna de desalting desechable de 10 ml (GE Healthcare). La proteína marcada se preserva en 25 mM Tris/HCl, pH 7.5.

La correcta incorporación del fluoróforo fue verificada por el grupo colaborador del Profesor Henrik Clausen, en la Universidad de Copenhage, por espectrometría de masas. El protocolo detallado se especifica en la sección experimental del artículo correspondiente (Artículo V).

4.6.5. Análisis de la actividad enzimática mediante radioactividad

La actividad enzimática de diversas ppGalNAc-Ts, así como la de sus mutantes y derivados, ha sido corroborada por el grupo de investigación del Profesor Henrik Clausen, en la Universidad de Copenhague. Para ello, han realizado la incubación de cada enzima con uno de sus sustratos peptídicos, $MnCl_2$ y un exceso del reactivo radioactivo UDP-[^{14}C]GalNAc. Retirado el exceso de reactivo, se detectan los grupos [^{14}C]GalNAc unidos al péptido sustrato mediante un contador de centelleo y se compara la actividad frente al control.

4.6.6. Análisis de la actividad enzimática mediante espectrometría de masas

Otra ténica empleada en los trabajos recogidos en esta tesis para la determinación de la actividad enzimática de las diversas isoformas ha sido el de la espectrometría de masas. También llevado a cabo por el laboratorio del Profesor Henrik Clausen en la Universidad de Copenhage, consiste en incubar la proteína con los reactivos necesarios y determinar el número de grupos GalNAc incorporados al péptido aceptor. El equipo utilizado para ello es un aparato de desorción/ionización mediante láser asistida por matriz, acoplado a un analizador de tiempo de vuelo y a un espectrómetro de masas (MALDI-TOF-MS). La unidad de enzima, tanto en este caso como en el del contador de centelleo (apartado 4.6.5), la definen como la cantidad necesaria de enzima para transferir 1 µmol de GalNAc en 1 minuto en las condiciones de reacción estándares.

Los motivos que conducen al empleo de una u otra técnica (espectrometría o radioactividad) han dependido tan sólo de la disponibilidad y accesibilidad de los equipos por parte del personal del grupo de investigación del Profesor Clausen, sin que haya variaciones apreciables entre ambos en lo que al uso concreto que les hemos dado en nuestro trabajo se refiere.

4.6.7. Resonancia magnética nuclear (RMN)

En esta tesis doctoral sólo se han empleado espectros monodimensionales ^{1}H de RMN para corroborar la unión de la luteolina a la enzima *Hs*GalNAc-T2 (Artículo IV).

El fenómeno de RMN en proteínas se observa mediante la cuantificación de las transiciones que se producen entre dos niveles próximos de energía, presentes en los núcleos atómicos con spin ½, cuando absorben energía a una frecuencia determinada en presencia de un campo magnético (Ferentz y Wagner, 2000). En proteínas convencionales, los únicos núcleos con spin ½ son los de sus átomos de Hidrógeno. En estas condiciones, pueden obtenerse espectros monodimensionales ^{1}H que aportan información sobre el plegamiento general de una proteína y que presentan unas señales a su vez caracterizadas por su valor de intesidad (I) y su desplazamiento químico (δ). El valor de este δ depende de las propiedades de cada núcleo atómico y del entorno de éste, por lo que permite detectar las variaciones causadas por la presencia o ausencia de interacción con un determinado ligando.

Estos experimentos de RMN han sido llevados a cabo por el Doctor Ignacio Delso, responsable del Servicio de Resonancia Magnética Nuclear de la Universidad de Zaragoza.

4.6.7.1. Resonancia magnética de diferencia de transferencia de saturación (STD)

La espectroscopía RMN de diferencia de transferencia de saturación (STD) ha sido una de las técnicas complementarias más empleadas en alguno de los trabajos que conforman esta tesis (Artículo II, Artículo V). Esta variante de RMN permite detectar y caracterizar interacciones receptor-ligando en disolución, observando para ello las señales de la molécula pequeña (ligando), que presenta propiedades espectroscópicas apropiadas para estudios a alta resolución (Angulo y Nieto, 2011).

La base de un experimento de STD es la siguiente (Angulo y Nieto, 2011; Figura 4.16):

- En primer lugar, se realiza un espectro monodimensional ^{1}H estándar de la mezcla de proteína y un exceso de ligando a una frecuencia de irradiación alejada de la frecuencia de resonancia del ligando o de la proteína. Dicho espectro, llamado *off-resonance*, mostrará las intensidades de referencia del ligando (las de la macromolécula no se ven).
- Se realiza un segundo experimento en el que la proteína se satura de forma selectiva irradiando, por ejemplo, en la frecuencia de los carbonos alifáticos. La saturación selectiva se transfiere al conjunto de

la proteína por el denominado efecto NOE intramolecular (NOEs), que se ve favorecido por el gran tamaño de las macromoléculas. De igual modo, esta saturación se transfiere desde la proteína hasta aquellas zonas del ligando que estén directamente unidas a la misma. Todas estas señales saturadas no aparecerán en el espectro.

- La resta de las señales del segundo espectro respecto a las del primero es lo que nos proporciona el espectro STD, que solamente contiene las señales asignables a los protones del ligando que están en contacto directo con la superficie de la proteína. El valor de sus intensidades serán un reflejo de la proximidad del protón a dicha superficie.

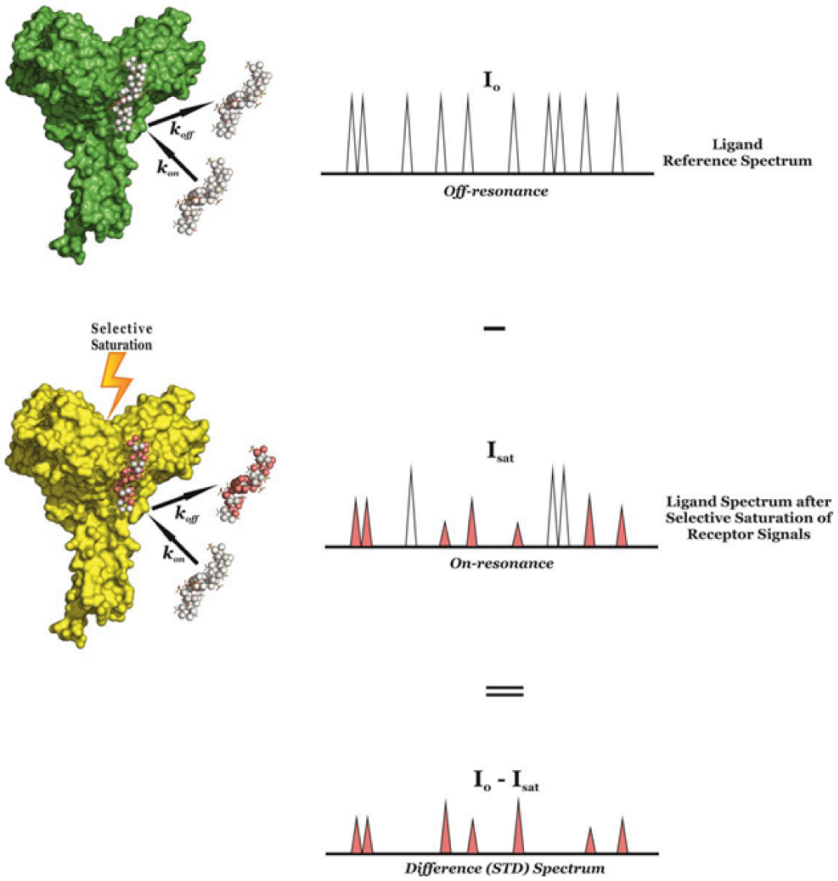


Figura 4.16. Esquema de un experimento de RMN STD. Se muestra la proteína en representación de superficie y los protones no intercambiables del ligando como esferas. En el espectro monodimensional ^{1}H superior se observan las intensidades de referencia del sistema en equilibrio. En el espectro intermedio, la proteína se ha saturado selectivamente, y transmite vía difusión de spin (NOEs) dicha saturación a los protones que interaccionan directamente con su superficie (color naranja). El último espectro es la resta de los dos primeros, propiamente el espectro diferencial, y sólo muestra las señales correspondientes a los protones que han interaccionado directamente, con mayor o menos intensidad en función de cuál haya sido su proximidad a la superficie. **Fuente:** http://www.glycopedia.eu/Saturation-Transfer-Difference

Este tipo de experimentos nos ha permitido, así, caracterizar las diversas interacciones proteína-ligando en disolución, algo de suma importancia en aquellos casos en los que no se ha podido cristalizar el complejo. Incluso en los casos en los que sí se ha resuelto la estructura cristalina del compuesto, técnicas como el STD nos garantizan que las interacciones observadas tienen también lugar en condiciones fisiológicas y no se ven alteradas por las condiciones de la cristalización.

Todos los experimentos de STD realizados en esta tesis doctoral se han llevado a cabo en el grupo de investigación de la Doctora Filipa Marcelo, en la Universidad de Cohimbra, Portugal. Los detalles técnicos de los mismos aparecen debidamente pormernorizados en las secciones experimentales y suplementarias de los artículos correspondientes.

4.6.8. Ensayos de transferencia y determinación de parámetros cinéticos de la reacción

La determinación de los parámetros cinéticos (K_M y $v_{máx}$) de las reacciones de transferencia del grupo GalNAc por parte de las diversas isoformas de las ppGalNAc-Ts y sus variantes y mutantes ha sido una técnica de enorme utilidad a la hora de caracterizar y comparar las interacciones entre las quimeras, así como analizar la variación de sus preferencias de glicosilación respecto a las proteínas silvestres (Artículo II).

Todos los ensayos cinéticos fueron realizados por el doctorando Earnest James Paul Daniel, del grupo de investigación del Profesor Thomas Gerken, de la Universidad Case Western, en Ohio. Los detalles técnicos de la metodología se describen con profusión en el apartado del artículo correspondiente.

4.6.9. Determinación de constantes de disociación por resonancia de plasmón superficial (SPR)

La resonancia de plasmón superficial (SPR) es una técnica biofísica que permite medir tanto la cinética como la termodinámica de la unión entre dos o más moléculas, una de las cuales se encuentra inmovilizada en la superficie de un chip de material inerte, generalmente de oro (Myszka, 2000). En esta técnica, un haz de luz polarizada incide sobre la superficie del film metálico provocando una interacción con los electrones de dicha superficie, que absorben la energía y se transforman en plasmones superficiales. El fenómeno de resonancia entre los fotones del haz y los plasmones tiene lugar cuando el momento de ambos es el mismo, lo que a su vez depende de la intensidad y del ángulo de la luz incidente. El momento de los plasmones, a su vez, depende de las condiciones del medio,

y es por ello que cuando éstas varían es necesario que el ángulo de la luz incidente también lo haga para que se vuelva a alcanzar la resonancia. Es este cambio de ángulo la magnitud que miden la mayoría de equipos de SPR (Myszka, 2000). Esta variación es proporcional al espesor de la superficie de la lámina y a la masa adherida sobre ella, lo que hace que se pueda monitorizar, en tiempo real, la masa adherida a la superficie del chip en cada instante de la realización del experimento (Myszka, 2000). Para facilitar la interpretación de resultados, normalmente las señales de SPR se representan en unidades arbitrarias de resonancia (RU) frente al tiempo.

En el caso de esta tesis doctoral, se realizaron experimentos de SPR para intentar determinar la Kd entre la *Hs*GalNAc-T4 y el glicopéptido 3 (Artículo II), pero la unión fue tan pobre que no pudo llegar a saturarse el chip y no se obtuvo ningún valor. Se recurrió a esta técnica debido a que tiene una mayor sensibilidad que otras a las que podíamos recurrir en nuestro laboratorio, y era algo necesario debido a la mala unión que, acertadamente, se preveía. No obstante, es ésta una técnica tremendamente útil para la obtención de Kd, entre otros motivos por la poca muestra que precisa, por lo que se ha seguido recurriendo a ella en otros proyectos complementarios a los recogidos en esta tesis doctoral.

Para la realización de estos experimentos de SPR se contrató el servicio del Instituto de Química Médica del Centro Superior de Investigaciones Científicas de Madrid.

4.6.10. Determinación del sitio de glicosilación

La determinación del sitio en el que se produce la glicosilación ha sido un aspecto clave a la hora de estudiar el papel del *flexible linker* en las preferencias de glicosilación de las distintas isoformas de ppGalNAc-Ts. A nivel experimental ha sido llevado a cabo por el grupo del Profesor Clausen, en la Universidad de Copenhagen.

4.7. TÉCNICAS COMPUTACIONALES

A lo largo de los cinco artículos que conforman esta tesis doctoral han sido numerosas las técnicas computacionales que nos han permitido extrapolar, interpretar, previsualizar y, en general, complementar y enriquecer los resultados experimentales obtenidos. De entre ellas, de forma muy resumida y superficial, destacan: (a) la dinámica molecular, constitente en simulaciones informáticas en las que, partiendo de estructuras de las proteínas de interés, se definen matemáticamente parámetros estéricos y de interacción con el medio a partir de

los cuales el programa calcula y predice la posición y movimiento de los átomos y sus posiciones respectivas a lo largo de un tiempo determinado; (b) el *coarse-grained model* (modelo de puntos gruesos; Rinker *et al*., 2012), donde la simulación parte de estructuras o modelos con menor resolución, lo que permite realizar dinámicas con moléculas más grandes y durante más tiempo, y (c) el *docking* o acoplamiento molecular (Lengauer y Rarey, 1996), método de modelado informático con el que se intenta predecir la conformación preferida por una molécula al formar un complejo con otra. Este último se emplea mucho para predecir la interacción de proteínas con ligandos y otros compuestos pequeños, como sucede en el caso de los inhibidores glicomiméticos (Artículo III), mientras que las simulaciones clásicas de dinámica molecular y el corase-grained estudian la evolución del conjunto de sistemas más grandes y complejos.

Todos estos experimentos de bioinformática han sido desarrollados por colaboradores bien del grupo de Química Biológica Computacional, en el Centro de Investigaciones Biológicas en Madrid, o del Departamento de Física Teórica de la Universidad de Zaragoza o del grupo del Doctor Francisco Corzana, en la Universidad de La Rioja.

CAPÍTULO 5- RESULTADOS

ARTÍCULO I

Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein *O*-glycosylation.

El presente artículo fue publicado en la revista *Nature Communications* el año 2015, primer año del periodo de tesis de la doctoranda. Fue realizado en coautoría con: (a) el grupo del Doctor Francisco Corzana, del Departamento de Química Orgánica de la Universidad de La Rioja; (b) el Instituto de Nanociencia de Aragón, de la Universidad de Zaragoza; (c) el grupo de investigación del Profesor Henrik Clausen, en la Universidad de Copenhage; (d) el Departamento de Química Orgánica y Computacional de la Universidad de Barcelona; (e) el Centro de Bioquímica Estructural de la Universidad de Montepellier, Francia y (f) el Departamento de Física Teórica de la Universidad de Zaragoza.

De los coautores del artículo, el único que no está en posesión del Grado de Doctor a fecha de defensa de la presente tesis doctoral es Ismael Compañón, del Departamento de Química Orgánica de la Universidad de la Rioja. Se adjunta su renuncia a presentar este artículo en la modalidad de compendio de pubilcaciones en el Apéndice 10.6 de la presente tesis.

Los resultados de esta publicación fueron previamente defendidos por un miembro anterior de nuestro grupo, la Doctora Erandi Lira Navarrete, que presentó su tesis doctoral por la modalidad clásica unos meses antes de la publicación del artículo (Lira-Navarrete, 2015).

La contribución del doctorando en esta publicación ha sido compartida con la Doctora Lira Navarrete y ha consistido en lo siguiente:

- Clonaje de las diversas construcciones y mutantes.
- Expresión y purificación de las proteínas implicadas.
- Cristalización de las proteínas.
- Determinación y cálculo de las constantes de disociación proteína-ligando.

Se adjunta a continuación la referencia completa del artículo:

Lira-Navarrete, E., **de las Rivas, M**., Compañón, I., Carmen Pallarés, M. C., Kong, Y., Iglesias-Fernández, J., Bernades, G. J. L., Peregrina, J. M., Rovira, C., Bernadó, P., Bruscolini, P., Clausen, H., Lostao, A., Corzana, F. & Hurtado-Guerrero, R. (**2015**). Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein *O*-glycosylation. *Nature Communications*, 6(6937), DOI: 10.1038/ncomms7937.

nature COMMUNICATIONS

ARTICLE

Received 15 Dec 2014 | Accepted 16 Mar 2015 | Published 5 May 2015

DOI: 10.1038/ncomms7937 OPEN

Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein *O*-glycosylation

Erandi Lira-Navarrete[1], Matilde de las Rivas[1], Ismael Compañón[2], María Carmen Pallarés[3], Yun Kong[4], Javier Iglesias-Fernández[5,†], Gonçalo J. L. Bernardes[6,7], Jesús M. Peregrina[2], Carme Rovira[5,8], Pau Bernadó[9], Pierpaolo Bruscolini[1,10], Henrik Clausen[4], Anabel Lostao[3,11], Francisco Corzana[2] & Ramon Hurtado-Guerrero[1,11]

Protein *O*-glycosylation is controlled by polypeptide GalNAc-transferases (GalNAc-Ts) that uniquely feature both a catalytic and lectin domain. The underlying molecular basis of how the lectin domains of GalNAc-Ts contribute to glycopeptide specificity and catalysis remains unclear. Here we present the first crystal structures of complexes of GalNAc-T2 with glycopeptides that together with enhanced sampling molecular dynamics simulations demonstrate a cooperative mechanism by which the lectin domain enables free acceptor sites binding of glycopeptides into the catalytic domain. Atomic force microscopy and small-angle X-ray scattering experiments further reveal a dynamic conformational landscape of GalNAc-T2 and a prominent role of compact structures that are both required for efficient catalysis. Our model indicates that the activity profile of GalNAc-T2 is dictated by conformational heterogeneity and relies on a flexible linker located between the catalytic and the lectin domains. Our results also shed light on how GalNAc-Ts generate dense decoration of proteins with *O*-glycans.

[1] BIFI, University of Zaragoza, BIFI-IQFR (CSIC) Joint Unit, Mariano Esquillor s/n, Campus Rio Ebro, Edificio I + D, Zaragoza 50018, Spain. [2] Departamento de Química, Universidad de La Rioja, Centro de Investigación en Síntesis Química, E-26006 Logroño, Spain. [3] LMA, INA, Universidad de Zaragoza, 50018 Zaragoza, Spain. [4] Copenhagen Center for Glycomics, Department of Cellular and Molecular Medicine, School of Dentistry, University of Copenhagen, Copenhagen DK-2200, Denmark. [5] Departament de Química Orgànica i IQTCUB, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain. [6] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. [7] Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Av Prof Egas Moniz, 1649-028 Lisboa, Portugal. [8] ICREA, Passeig Lluís Companys 23, 08020 Barcelona, Spain. [9] Centre de Biochimie Structurale, INSERM U1054, CNRS UMR 5048, Université Montpellier 1 and 2, 29 rue de Navacelles, 34090 Montpellier, France. [10] Departamento de Física Teórica, Universidad de Zaragoza, Zaragoza 50009, Spain. [11] Fundación ARAID, 50018 Zaragoza, Spain. †Present address: Department of Chemistry, Britannia House, 7 Trinity Street, King's College London, London SE1 1DB, UK. Correspondence and requests for materials should be addressed to R.H.-G. (email: rhurtado@bifi.es).

Mucin-type (GalNAc-type) *O*-glycosylation is by far the most differentially and complex regulated type of protein glycosylation, and likely the most abundant with over 80% of all proteins passing through the secretory pathway that is predicted to be *O*-glycosylated[1]. In metazoans, this post-translational modification is initiated by a large family (20 in humans) of polypeptide *N*-acetylgalactosaminyl-transferases (GalNAc-Ts), which transfer a GalNAc residue from uridine diphosphate *N*-acetylgalactosamine (UDP-GalNAc) to Ser/Thr side chains in the presence of manganese[2]. These isoenzymes are grouped in CAZy database as family 27 (ref. 3) and further classified into several subfamilies based on the enzyme protein sequences and the genomic structures of the encoding genes[2]. They are also classified into two major classes based on their primary acceptor substrate preferences for peptides and partially glycosylated GalNAc-glycopeptides. The acceptor substrate specificities of these isoenzymes are distinct but partly overlapping, and so far no clear global consensus motifs or isoform-specific motifs that govern protein *O*-glycosylation have emerged, although improved algorithms for predictions have been proposed[1,4]. These isoenzymes have different cell and tissue expression patterns[2] and they play important roles in health and disease[5,6]. The immediate product of the GalNAc-Ts is also known as the Tn (GalNAcα1-*O*-Ser/Thr) antigen (if sialylated the STn (NeuAcα2-3GalNAcα1-*O*-Ser/Thr) antigen), and while this structure in normal cells is masked by elongation of the glycan structures by a number of other glycosyltransferases, expression of Tn (and STn) is a hallmark of cancer cells[7,8]. Expression of these immature truncated *O*-glycans is strongly correlated with poor prognosis and low overall survival and truncated *O*-glycans serve as targets for immunotherapies[9]. Mechanisms leading to expression of Tn and STn *O*-glycans in cancers include somatic mutations and/or epigenetic silencing of the private chaperone Cosmc responsible for *O*-glycan elongation[10] as well as relocation of GalNAc-Ts from Golgi to endoplasmic reticulum[11]. More recently, we have shown that the truncated *O*-glycophenotype appear to directly induce oncogenic features with enhanced growth and invasion[12].

GalNAc-T isoforms contain an *N*-terminal catalytic domain adopting a GT-A fold and a unique C-terminal lectin domain, classified as a carbohydrate-binding module (CBM) 13 in the CAZy database[3], with a β-trefoil fold that are connected through a short flexible linker[13–16]. A key structural feature of these enzymes in catalysis is the existence of a flexible loop formed by residues Arg362 to Ser373 (residue numbers in GalNcAc-T2) that adopts different conformations during the catalytic cycle and renders the enzyme catalytically inactive or active[16].

Another level of complexity in this family of enzymes comes from the lectin domains that play a very important role in the *O*-glycosylation process by modulating the specificity and the use of partially glycosylated GalNAc-glycopeptide substrates[17,18]. Lectin domains or CBMs also exist in some glycosyl hydrolases (GHs). For these enzymes, CBMs are important for targeting the catalytic GHs onto their substrates and enhancing their hydrolytic activity, and even can potentiate the action of a cognate catalytic module towards polysaccharides in intact cell walls through the recognition of nonsubstrate polysaccharides[19,20]. We have hypothesized that the lectin domains of GalNAc-Ts are required to accommodate efficient glycosylation of diverse protein sequences with densely located acceptor sites in a process where the lectin domains bind to initially attached GalNAc residues and promote catalytic efficiency and further incorporation to less efficient acceptor sites[2,21]. The importance of the lectin domains is exemplified by the rare disease familial tumoral calcinosis associated with hyperphosphatemia, which is caused by deficiency in GalNAc-T3 or FGF23 (ref. 22). Thus, loss of GalNAc-T3, which mediates the *O*-glycosylation of Thr178 in a proprotein convertase-processing site (RHTR$^{179}\downarrow$) of FGF23, results in inactivation of FGF23 and cause hyperphosphatemia[22]. Notably, glycosylation of Thr178 in FGF23 by GalNAc-T3 requires the lectin domain and a GalNAc *O*-glycan positioned *N*-terminal to this site[5]. However, the mechanism for how these unique lectins modulate *O*-glycosylation is essentially unknown. In this regard, recent studies have provided evidence that the *N*/*C*-terminal orientation of GalNAc residues and available proximal acceptor sites is an important factor in the overall catalytic activity and specificity of GalNAc-Ts, and moreover, that the preferences for this orientation differs between isoforms[23]. This adds another level of differential regulation and complexity to the *O*-glycosylation process and help to explain how large mucins with thousands of glycosites are densely decorated with *O*-glycans with high fidelity. Interestingly, studies with random glycopeptide libraries suggest that there is a distance preference for the effect of prior GalNAc residues, in which the optimal distance from the prior site of glycosylation corresponds to residues that are located approximately ten residues apart for GalNAc-T1, T2 and T13, and eight or nine residues for T3 (ref. 23). This distance preference suggests a cooperative binding of GalNAc-Ts to GalNAc-glycopeptide substrates through both the catalytic and lectin domains[23].

Here we present the first crystal structures of the inactivated and activated forms of the GalNAc-T2 isoform in binary and ternary complexes with different GalNAc-glycopeptides enabling us to study in detail the lectin-mediated modulatory mechanism of catalysis. The crystal structures, in combination with metadynamics simulations, provides a rational explanation of why this isoform favours glycosylation of acceptor sites positioned at the *N*-terminal region and ten residues apart from the prior GalNAc moieties. We also present single-molecule atomic force microscopy (AFM) and small-angle X-ray scattering (SAXS) data indicating that GalNAc-Ts sample a highly complex conformational landscape formed by ensembles of compact and extended structures. Analysis of SAXS shows that alterations of the equilibrium towards the compact structures in the presence of acceptor substrates appear to be required for lectin-mediated catalysis. In addition, by using a coarse-grained model, we demonstrate that GalNAc-T2 structural heterogeneity, as well as its activity profile, can be simply related to the mechanical properties of the flexible linker.

Results

Architecture of the GalNAc-glycopeptides complexes. To understand how lectin domains direct *O*-glycosylation, we obtained tetragonal crystals of GalNAc-T2 in complex with three GalNAc-glycopeptides (Table 1 and Supplementary Table 1). The resulting crystals allowed us to solve the structure at high resolution (from 1.48 to 1.67 Å) and easily interpret the density maps (Table 1). Despite the co-crystallization experiments were performed with GalNAc-T2, UDP and Mn^{+2}, we obtained binary complexes with glycopeptides MUC5AC-Cys13 and MUC5AC-3-13 (the binary complexes are referred to the enzyme complexed to the different glycopeptides MUC5AC-Cys13 and MUC5AC-3-13, respectively), and a ternary complex with MUC5AC-13 (the ternary complex is referred to the enzyme bound to UDP and the glycopeptide MUC5AC-13; Fig. 1a,b).

Within the asymmetric unit (AU) of $I4_1$ crystals, one molecule of GalNAc-T2 is arranged as a dimer with another molecule from the neighbouring AU (Fig. 1a). These compact dimeric forms are consistent with previous structures of GalNAc-T2 in complexes with UDP and Mn^{+2} (protein data bank (PDB) entry 2FFV[14]), and UDP or UDP-GalNAc/Mn^{+2} and peptides (PDB entries

Table 1 | Data collection and refinement statistics*.

	GalNAc-T2 -MUC5AC-Cys13	GalNAc-T2-MUC5AC-3,13	GalNAc-T2-MUC5AC-13-UDP-Mn^{+2}
Data collection			
Space group	I 4_1	I 4_1	I 4_1
Cell dimensions			
a, *b*, *c* (Å)	87.35, 87.35, 178.43	87.2, 87.2, 179.07	87.27, 87.27, 178.59
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution (Å)	20–1.67 (1.76–1.67)	20–1.48 (1.56–1.48)	20–1.65 (1.74–1.65)
R_{merge}†	0.103 (0.653)	0.058 (0.535)	0.043 (0.601)
$I/\sigma(I)$	11.3 (3.0)	18.6 (2.3)	22.1 (2.9)
Completeness (%)	99.8 (100)	99.6 (97.5)	99.8 (99.2)
Multiplicity	8.9 (9.2)	8.6 (4.2)	6.7 (5.8)
Refinement			
Resolution (Å)	20–1.67 (1.76–1.67)	20–1.48 (1.56–1.48)	20–1.65 (1.74–1.65)
No. of reflections (test)	76,947 (2,308)	110,205 (3,306)	79,709 (2,391)
R_{work}/R_{free}‡	0.196/0.232	0.149/0.178	0.156/0.185
No. of atoms			
Protein	4,023	4,040	3,991
Glycopeptide	114	129	107
Sulfate	60	155	110
Ethylenglycol	56	—	—
UDP	—	—	25
Mn^{+2}	—	—	1
Water	556	800	614
B-factors (Å^2)			
Protein	22.41	19.03	26.71
Glycopeptide	33.81	19.46	19.29
Sulfate	56.01	61.48	69.44
Ethylenglycol	38.75	—	—
UDP	—	—	21.88
Mn^{+2}	—	—	19.29
Water	33.51	35.06	39.02
R.m.s deviations			
Bond lengths (Å)	0.019	0.012	0.013
Bond angles (°)	2.027	1.613	1.627
PDB ID	5ajn	5ajo	5ajp

UDP, uridine diphosphate.
*Highest resolution shell is shown in parenthesis.
†$R_{merge} = \Sigma hkl \Sigma i\ |I(hkl)i - [I(hkl)]| / \Sigma hkl\ \Sigma i\ I(hkl)$.
‡$R_{work} = \Sigma hkl |F(hkl)_o - [F(hkl)_c]| / \Sigma hkl\ F(hkl)_o$; R_{free} was calculated as R_{work}, where $F(hkl)_o$ values were taken from 3% of data not included in the refinement.

4D0T, 4D0Z and 4D11 (ref. 16)). However, members of this family of enzymes are proposed to be monomers in solution[16,24,25] (discussed below). The crystal structures show that the typical GT-A fold is located in the *N*-terminal region and the lectin domain is located in the *C*-terminal region (Fig. 1a). The density for all the glycopeptides and UDP/Mn^{+2} was well defined in most monomers (Fig. 1b). These results are further supported by tryptophan fluorescence spectroscopy experiments showing that peptides/glycopeptides can bind fairly well to GalNAc-T2 in the absence or presence of UDP/Mn^{+2} (Supplementary Table 2). The dissociation constants (K_ds) of the peptides were in the low μM range and very similar among them except for the naked peptide MUC5AC, in which the K_d in the absence of UDP/Mn^{+2} is fourfold better than in the presence of the nucleotide (Supplementary Table 2).

The structures also feature a flexible loop that can oscillate between closed and open conformations (Fig. 1c,d) and is associated with the active and inactive states of the enzyme, as described in previous studies[14,16]. We also reported that the UDP-GalNAc moiety was a key factor to maintain the flexible loop in a closed conformation regardless of whether the peptide was present or not (Fig. 1c and PDB entry 4D0T)[16]. On the contrary, previous structures containing UDP (Fig. 1c and PDB entries 2FFU[14], 4D11 (ref. 16) and 2FFV[14]) including our complex of GalNAc-T2-UDP-MUC5AC-13 (Fig. 1c,d) exhibit a mix of loop conformations (semi-open, open and closed) along the catalytic cycle (Fig. 1c)[16]. Furthermore, the flexible loop dynamics are coupled to the key catalytic residue Trp331 mobility that can adopt 'in' (inside of the active site) and 'out' (outside of the active site) conformations, which in turn are associated to the active or inactive states of GalNAc-T2, respectively (Fig. 1d). Interestingly, MUC5AC-13 was bound to an active form of GalNAc-T2, whereas MUC5AC-Cys13 and MUC5AC-3-13 were unpredictably trapped bound to the inactive form of GalNAc-T2 (Fig. 1b–d). In these cases, the flexible loop was found in an open conformation, different from the one previously described for the GalNAc-T2-UDP complex (PDB entry 2FFV[14] and root-mean-square deviation (RMSD) of 4.68 Å for aligned Cα atoms corresponding to the flexible loop; Fig. 1c,d), which further demonstrates the loop versatility.

Peptide and lectin domain-binding sites. As shown in the crystal structures (Fig. 1), the GalNAc-T2-binding site is large and set up by three regions: the sugar nucleotide, the peptide and the lectin domain-binding sites (Fig. 2a). In the GalNAc-T2-MUC5AC-13-UDP complex, GalNAc-T2 is in an active state with the flexible loop in a closed conformation with UDP in the sugar nucleotide-binding site, and MUC5AC-13 acts as a bridge between the catalytic unit and the lectin

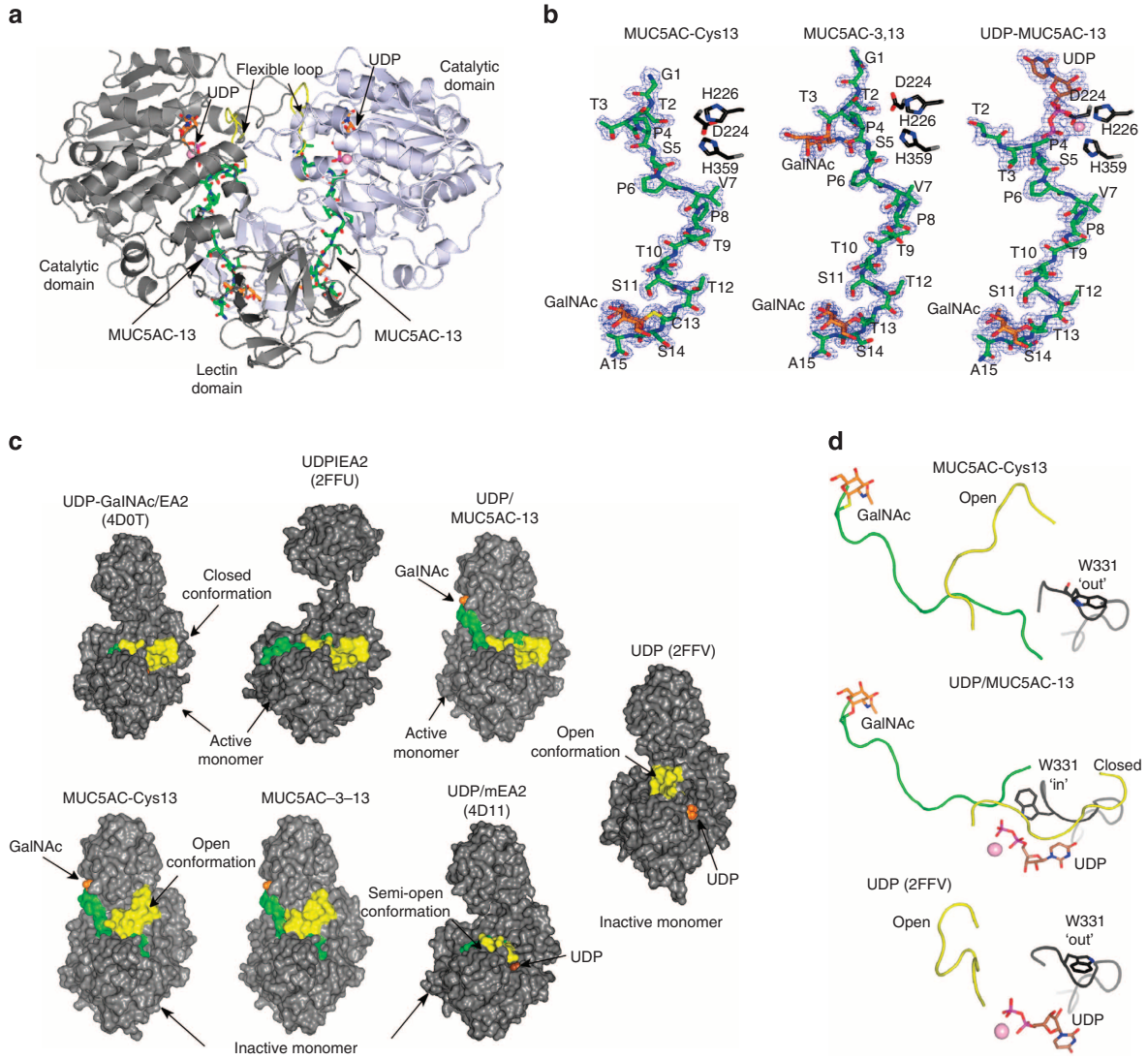


Figure 1 | Crystal structures of GalNAc-T2 in complex with glycopeptides. (**a**) Cartoon representation of the overall dimeric structure of GalNAc-T2 in complex with UDP-MUC5AC-13. The monomers are coloured in grey and blue-white, respectively. UDP, MUC5AC-13 and flexible loop in yellow are indicated by arrows. The nucleotide and the glycopeptide are depicted as orange and green carbon atoms, respectively. The GalNAc moiety covalently bound to Thr13 is depicted as orange carbon atoms. (**b**) Electron density maps are F_O–F_C syntheses (blue) contoured at 2.0σ for the ligands. The residues (Asp224, His226 and His359) coordinated to the manganese atom (shown as a pink sphere) are illustrated as black carbon atoms. Colours for the ligands are the same as above. (**c**) Surface representation of binary and ternary GalNAc-T2 complexes showing different active and inactive states. Protein, flexible loop, nucleotides, GalNAc moiety and peptides are coloured in grey, yellow, brown, orange and green, respectively. The active and inactive states are indicated as closed and open conformations, respectively. (**d**) Close-up view of GalNAc-T2-MUC5AC-Cys13, GalNAc-T2-MUC5AC-13-UDP and GalNAc-T2-UDP (PDB entry 2FFV[14]) complexes. The colours are the same as shown above. The loop in which Trp331 is located is in black. In the latter complex, UDP is shown as inverted relative to the most frequent position found for UDP and has been proposed to be a conformation ready to depart from the active site[14,16]. Trp331 is depicted as black carbon atoms and adopts 'in' and 'out' conformations.

domain (Fig. 2a). MUC5AC-13 as well as MUC5AC-Cys13 and MUC5AC-3-13 glycopeptides have a C-terminal GalNAc moiety that establishes interactions with the lectin domain-binding site (Figs 1 and 2a). The sugar moiety for the three above glycopeptides adopts a perpendicular conformation with respect to the peptide backbone. This also supports previous data suggesting that Cys residues bound to a GalNAc moiety mimic fairly well Thr residues linked to the same sugar[26].

It is noteworthy that the C-terminal of MUC5AC-13, compared with the naked EA2 peptide, follows a divergent pathway towards its C-terminus (Fig. 2a,b). Both, MUC5AC-13 and EA2, bind in a competitive manner engaging potential acceptor sites close to the UDP (Fig. 2a,b). The different binding modes found for the peptides emphasize the plastic nature of the peptide-binding groove, which makes it potentially important to sample a large number of different acceptor substrates with multiple acceptor sites.

At the peptide-binding groove level, the specificity of GalNAc-T2 for the peptides MUC5AC-13 and EA2 (same interactions are also present for MUC5AC-Cys13 and MUC5AC-3-13) is governed mainly by hydrophobic interactions with Val255, Phe361, His365, Leu270, Phe377, Phe280, Trp282 and Tyr284 that are residues located in a solvent-exposed hydrophobic pocket (Fig. 2a,b). Both peptides are also tethered by hydrogen bonds with Trp282, His365 and Ser373 (Fig. 2a,b).

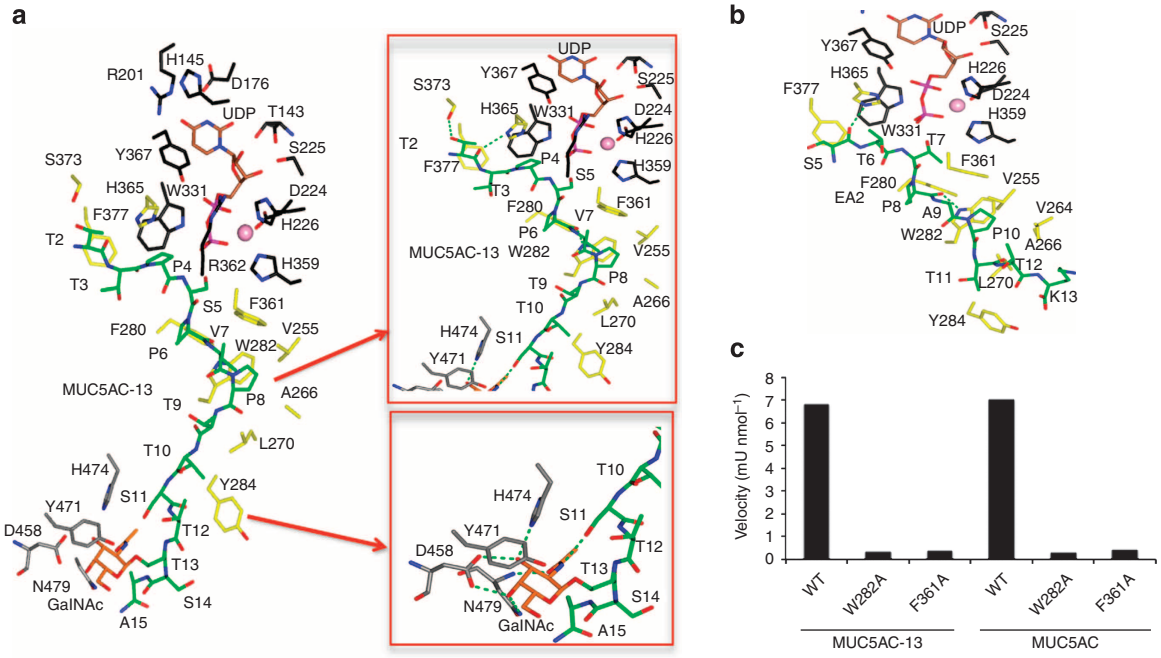


Figure 2 | Structural features of peptide and lectin domain-binding sites. (**a**) View (left panel) of complete sugar nucleotide, peptide and lectin domain-binding sites of the GalNAc-T2-UDP-MUC5AC-13 complex. Close-up view (right panel) of peptide and lectin domain-binding sites. The residues forming 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. (**b**) Close-up view of the peptide-binding site of the GalNAc-T2-UDP-EA2 complex. Colours are the same as above. (**c**) Graph that shows the velocity for the wild-type (WT) enzyme and mutants with peptides MUC5AC and MUC5AC-13. Time-course experiments were carried out and the reactions were analysed and evaluated by MALDI-TOF-MS. The specific activities under linear conditions were inferred from the mass spectrometry data (see Methods for details). One unit of enzyme is defined as the amount of enzyme that transfers 1 μmol of GalNAc in 1 min using the standard reaction mixture and conditions. The velocity values were obtained from three independent experiments and errors are <20%.

Of four amino acids widely conserved among isoforms (Phe361, Phe280, Trp282 and Tyr396; Supplementary Fig. 1a), Phe361, Phe280 and Trp282 are key residues in the recognition of common peptide motifs such as Pro-x-Pro (where x is usually a small hydrophobic residue; Fig. 2a,b) found in acceptor substrates. In fact, site-directed mutagenesis of Phe361 and Trp282 to Ala residues almost eliminates the activity of the enzyme, and thus confirms the importance of these amino acids in peptide recognition (Fig. 2c and Supplementary Fig. 1b).

In all cases, the lectin domains appear to interact only with the GalNAc moiety of the glycopeptides without any discernable interaction with the peptide backbone (Fig. 2a). The lectin domains of GalNAc-Ts consist of three pseudo repeat regions termed α, β and γ, which each potentially may bind GalNAc (Supplementary Fig. 1c). However, it is accepted that the lectin domains from human GalNAc-T1 contain two functional GalNAc-binding sites (α and β), whereas GalNAc-T2 (α), GalNAc-T4 (α) and GalNAc-T10 (β) contain only one[21]. Unlike GalNAc-T10, in which the sugar moiety is located on the β-site[15], the *C*-terminal GalNAc of the glycopeptides in GalNAc-T2 is located on the α-site of the lectin domain and is tethered by conserved residues such as Asp458, Asn479, Tyr471 and His474 (equivalent residues for β-binding site in GalNAc-T10 are Asp525, Asn544, Tyr536 and His539; Fig. 2a and Supplementary Fig. 1a). This explains why Asp458 (the only amino acid establishing two hydrogen bonds with the sugar moiety, Fig. 2a) is important for GalNAc-peptide substrate specificity and consequently in tuning the glycosylation profile of these enzymes[11,17,27].

The sugar nucleotide-binding site. Both the UDP and glycopeptides can bind to a versatile sugar nucleotide-binding site with a flexible loop adopting both open and closed conformations (Figs 1 and 3a). Although the *N*-terminal of MUC5AC-13 is covered by the flexible loop, the *N*-termini of MUC5AC-Cys13 and MUC5AC-3-13 are exposed to the solvent and inserted into the sugar nucleotide-binding site (Fig. 1c). The latter two glycopeptides adopt very similar binding modes, although some differences stand out towards the *N*-terminus (RMSD of 0.72 Å for aligned Cα atoms corresponding to residues Gly1-Ser5). The Gly1–Ser5 interactions with GalNAc-T2 are mainly governed by common hydrogen bonds with Glu334, Arg362 and Lys363, and distinct ones in particular for Asp224 and Tyr367 (Fig. 3a). There is an extra sulfate molecule in the GalNAc-T2-MUC5AC-Cys13 complex that occupies the same position found for the manganese atom in the competent structures such as the GalNAc-T2-MUC5AC-13-UDP complex (Fig. 3a; the interactions of GalNAc-T2 with UDP were discussed earlier)[14,16]. MUC5AC-Cys13 establishes further hydrogen bond interactions with this sulfate molecule (Fig. 3a). For the diglycosylated peptide, an unprecedented GalNAc-binding site, formed by aromatic residues such as Phe280, Tyr367 and Phe377, is found (Fig. 3a). This GalNAc moiety linked to Thr3 is tethered by Ser5, Phe280, Ala307, Gly333 and Phe377, and is found close to the sugar unit of UDP-GalNAc (average atomic shift of 5.29 Å; see Fig. 3a,b for further information). It is also important to emphasize that the lectin domain, through the interaction with the *C*-terminal GalNAc moiety of MUC5AC-13, imposes *N*-terminal residues of the glycopeptide

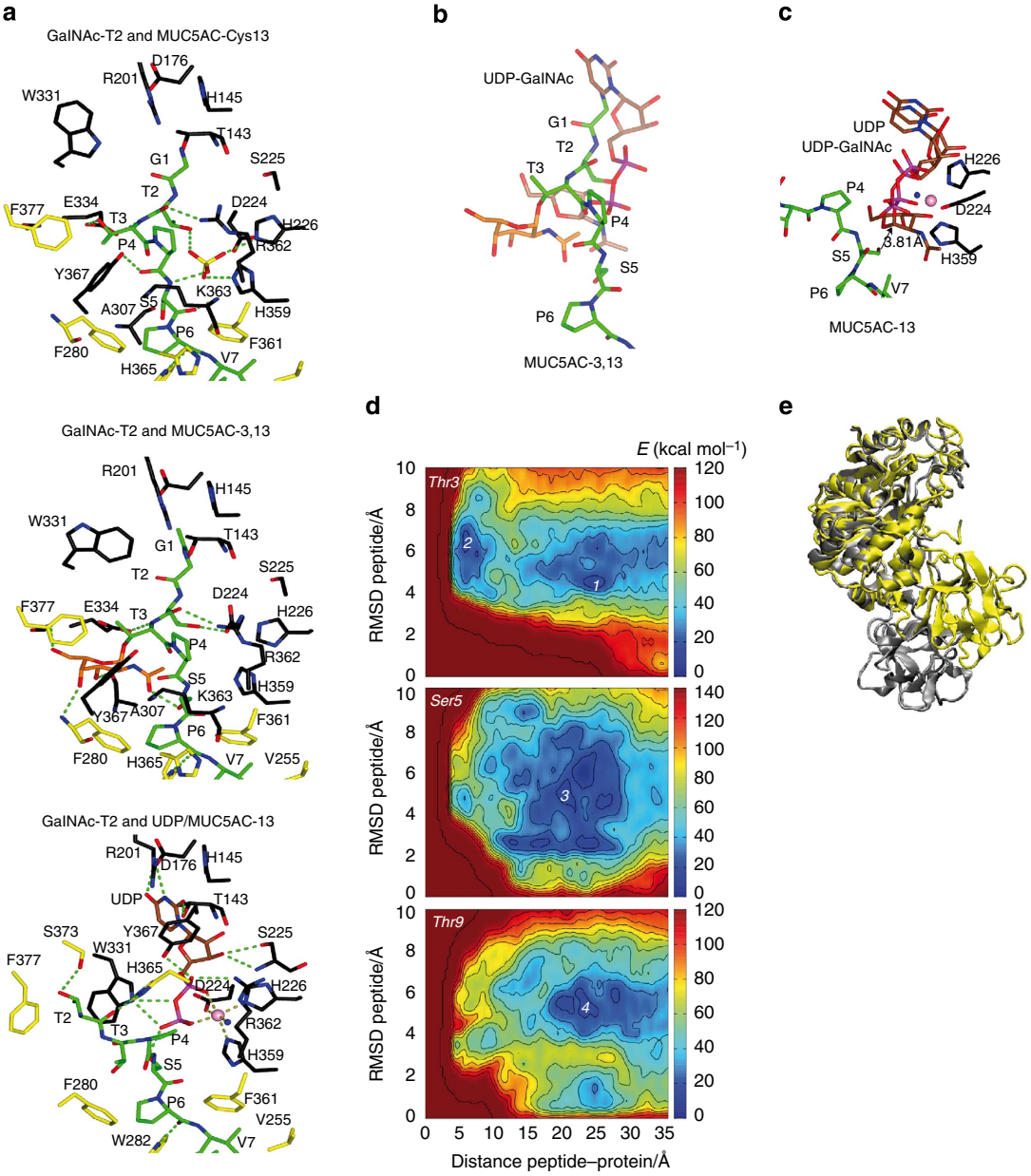


Figure 3 | Structural features of the sugar nucleotide-binding site and metadynamics simulations of substrate binding. (**a**) Close-up view of the sugar nucleotide-binding site for GalNAc-T2-MUC5AC-Cys13, GalNAc-T2-MUC5AC-3-13 and GalNAc-T2-UDP-MUC5AC complexes. Hydrogen bond interactions and Mn^{+2} coordination are shown as dotted green and brown lines, respectively. A sulfate molecule is depicted in GalNAc-T2-MUC5AC-Cys13 complex. Peptides, UDP and amino acids are shown with the same colours as in Fig. 2. (**b**) Overlay of one of the monomers of GalNAc-T2 containing UDP-GalNAc (PDB entry 4D0T[16]) in the active site with GalNAc-T2-MUC5Ac-3-13 complex. The GalNAc moiety of UDP-GalNAc and MUC5AC-3-13 is shown as brown and orange carbon atoms, respectively. (**c**) Overlay of one of the monomers of GalNAc-T2 containing UDP-GalNAc (PDB entry 4D0T[16]) in the active site with GalNAc-T2-UDP-MUC5AC-13 complex. The structure shows that Ser5 is close to the anomeric carbon of UDP-GalNAc. (**d**) Free energy landscapes of peptide binding for each one of the three-acceptor sites. (**e**) Comparison between a reactive conformation of the peptide MUC5AC-13 for the Thr9 (yellow) and Thr3 (grey) acceptor sites.

close to UDP (Fig. 3a). Particularly, Ser5 is quite close to UDP (the distance between the Ser5 oxygen and the β-phosphate is 3.81 Å) and is potentially in a position to attack the anomeric carbon of UDP-GalNAc (see Fig. 3c for further information). This structure is compatible with the fact that GalNAc-T2 prefers to glycosylate *N*-terminal acceptor sites of glycopeptides efficiently, but the structure itself does not explain why Thr3 is the most favourable acceptor site, as shown earlier by kinetic studies, or why glycosylation optimally takes place at acceptor sites located ten residues *N*-terminal of an existing site of glycosylation[17,23].

Structure-guided metadynamics. In an attempt to answer why Thr3, positioned ten residues *N*-terminal to the prior glycosylated Thr13, is the most favourable acceptor site in comparison to less efficient acceptor sites such as Ser5 and Thr9, we modelled the binding of a glycopeptide to GalNAc-T2 using an enhanced sampling molecular dynamics (MD) approach (metadynamics)[28]. A ternary complex of GalNAc-T2 with UDP-GalNAc and a glycopeptide was used as the initial structure for the calculations. Room temperature MD simulations were performed to equilibrate the enzyme complex (see details in Supplementary Methods). Afterwards, the binding of the glycopeptide to the

active site was monitored by metadynamics using two collective variables. The first variable was the distance between the centre of mass of the UDP-GalNAc and the centre of mass of the side chain of the peptide acceptor sites (Thr3, Ser5 or Thr9). This variable measures directly the penetration of the peptide into the active site. The second collective variable was taken as the RMSD of the peptide structure and measures changes in peptide conformation.

The free energy landscapes of peptide binding reconstructed from the metadynamics simulations show two main energy minima for Thr3 (minima **1** and **2** in Fig. 3d) at distances of 20–25 Å (outer minimum) and 4.5 Å (inner minimum), respectively, but only a broad outer minimum for Ser5 and Thr9 (**3** and **4**, respectively, in Fig. 3d). The outer minima (**1**, **3** and **4**) correspond to conformations in which the peptide is far from the active site and mainly interacts with the lectin domain (Supplementary Fig. 2a). Instead, the inner minimum (**2**, Thr3 glycosylation) corresponds to a reactive conformation, very close to the one found in the crystal structure of the GalNAc-T2-UDP-MUC5AC-13 complex (Supplementary Fig. 2a). Similar inner structures on the Ser5 and Thr9 maps are less stable than their corresponding outer global minimum, and thus much less populated. Consequently, the glycosylation reaction is expected to be less efficient for Ser5 and Thr9 than for Thr3.

The order of the preferred glycosylation site can be easily visualized by representing the free energy of the binding process versus the distance between the acceptor amino acid and the UDP-GalNAc donor (Supplementary Fig. 2b). It is important to note that only Thr3 glycosylation shows a deep inner minimum corresponding to reactive configurations.

A comparison of the protein structure for Thr9 and Thr3 glycosylation at short distances (Fig. 3e) shows that Thr9 glycosylation requires a highly compact and strained protein conformation in which the lectin domain approaches the catalytic domain. This conformation differs from the less compacted structure found in our crystal structure in complex with UDP-MUC5AC-13 (RMSD of 2.51 Å for 495 aligned Cα atoms). Notably, our simulation results are consistent with the previously reported solution studies, in which Thr3 was shown to be glycosylated preferentially over Ser5 and Thr9 (ref. 17), and also explains why GalNAc-T2 prefers to glycosylate *N*-terminal residues that are ten residues apart of a prior *C*-terminal glycosite. However, several questions arise on the molecular basis of how GalNAc-T2 achieves catalysis on residues close or very distant to an already attached GalNAc *O*-glycan, or whether the more compact structure shown above really exists in solution.

GalNAc-T2 adopts compact and extended conformations. To understand the conformational state of GalNAc-T2 under ligand-binding and -unbinding conditions, we applied initially time-lapse AFM experiments. A series of AFM topography images were recorded of single GalNAc-T2 molecules under different conditions and in three distinguishable conformational states: (i) a highly compact structure similar to the one above found through molecular modelling (Figs 3e and 4a, top-left panel) and in which the two domains interact closely even though a slot is still observed between them; (ii) a compact structure similar to our crystal structures that displays a clear slot (Fig. 4a, top-right panel and Supplementary Fig. 3a) and (iii) extended structures exhibiting fully separated domains (Fig. 4a, bottom panel and Supplementary Fig. 3b). Our AFM experiments (Fig. 4a, bottom panel and Supplementary Fig. 3b) support, for the first time, that the extended state visualized in the crystal structure of GalNAc-T2 in complex with UDP and the EA2 peptide (PDB entry 2FFU[14]; Fig. 1c), in which the domains are separate, is not an artefact of the crystallographic conditions.

The highly compact and compact species show height values of ≈8–10 nm (Fig. 4a, top panel), whereas the average height for the extended species is smaller, reaching ≈5 nm (Fig. 4a, bottom panel). The size of the separated domains in the images is barely larger than that estimated from the crystal structure (3.5 and 4.7 nm for the lectin and the catalytic domains, respectively, in PDB entry 2FFU[14]) mainly as a result of hydration[29]. The intermediate sizes observed for the compact species thus suggest a slight overlap between both domains of the enzyme.

The enzyme in the apo form or bound to UDP or UDP-GalNAc in the presence of Mn^{+2} adopts either the very compact or the compact structure (Fig. 4a, top panel). However, the three different conformational states are present when the enzyme is either bound to EA2 or to glycopeptides such as MUC5ACs-Cys9 (see sequence in Supplementary Table 1) or MUC5AC-13 alone or in combination with UDP/Mn^{+2} (Fig. 4a). Although our data does not provide relevant information of the distribution between different conformational states, clearly demonstrates the highly dynamic nature of the GalNAc-T2 enzyme.

GalNAc-T2 monomeric compact conformations. To quantify the GalNAc-T2 conformational states in solution, we measured SAXS data for a large number of conditions. The capacity to describe this data with available crystallographic structures for the monomeric and dimeric forms of GalNAc-T2 was tested with the programme CRYSOL[30]. The poor agreement of either structures to the experimental SAXS curves (see Supplementary Table 3), prompted us to study the enzyme flexibility using the Ensemble Optimization Method (EOM)[31]. Ensemble analysis of solution scattering curves for the apo enzyme at different concentrations revealed a unique distinguishable bimodal distribution of monomeric compact and extended structural ensembles that have not been found before in other types of proteins by using this method (Fig. 4b and Supplementary Figs 3c, 4 and 5). In addition, the ensemble analysis identified compact dimers that reached up to ≈40% at the highest concentration measured (10 mg ml^{-1}; Supplementary Table 3). Analysis of the data shows that both monomeric conformation ensembles are similarly populated but do not display a continuum of conformational states. The radius of gyration (R_g) ranges between R_g≈24–29.5 and ≈31–36 Å for the compact and extended structures, respectively, with negligible population in between (Fig. 4b). As already mentioned, a small percentage of potential structures with R_g between ≈29.5 and 31 Å are not displayed because they are likely insufficiently stable to be detected (Fig. 4b). These data reinforce the GalNAc-T2 dynamics inferred from our AFM experiments and demonstrates that GalNAc-T2 is even more flexible than shown by AFM experiments. In contrast to AFM, SAXS is also able to identify a significant population of dimers that supports the dimeric arrangement found for our crystal structures, with the exception of the PDB entry 2FFU[14], in which the enzyme is monomeric (we used the PISA server to determine the quaternary structure of GalNAc-T2 in the crystals). The SAXS fits with EOM were notably better when the three crystallographic dimeric arrangements were combined with the monomeric forms (Fig. 4b, Supplementary Figs 3c, 4 and 5, and Supplementary Methods). In particular, three different pools were built consisting in 10,000 monomeric conformations that were separately enriched with 100 copies of the theoretical curves of each of the three dimeric arrangements tested (see Supplementary Methods). The best fit was achieved when the most compact dimer that possesses a buried area of 1,689 Å^2 (PDB entry 2FFV[14]) was used. Dimers visualized in crystal structures with a tetragonal space group appear to be the most flexible compact structures of all dimers with ≈750 Å^2 of buried area. This latter

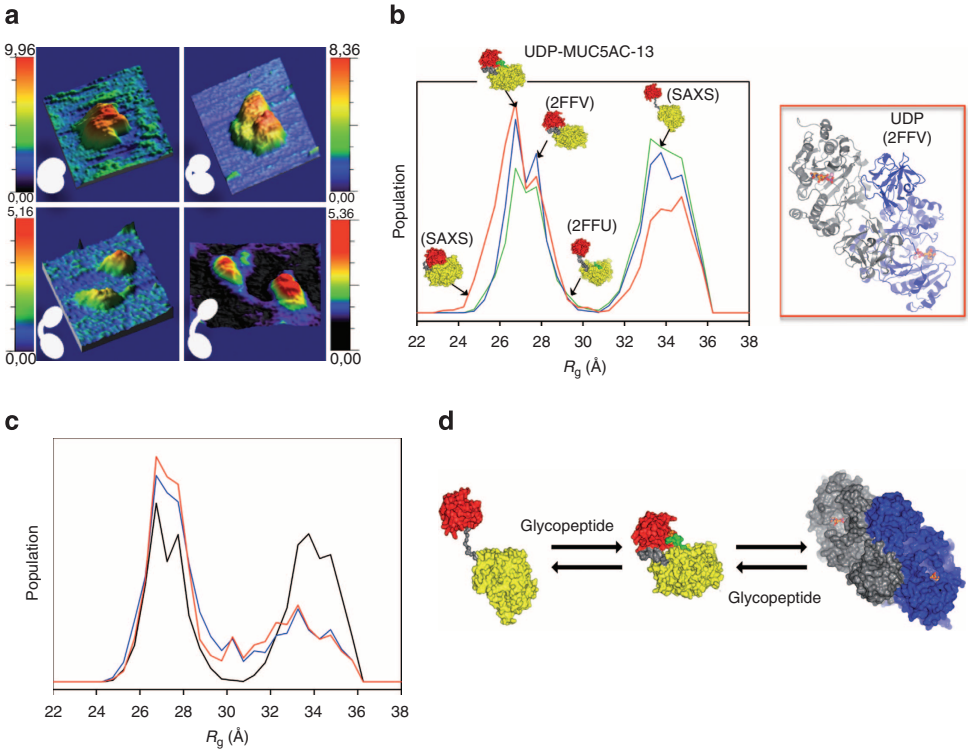


Figure 4 | Three-dimensional topography AFM images of single molecules in different conformations and SAX analysis of GalNAc-T2. (**a**) (Top-left panel) A very compact structure image of the apo form showing an overlap of a domain on the other. The average height of these features is ≈9 nm. Compact image of the enzyme bound to UDP-GalNAc/Mn^{+2} and extended structures images of the enzyme bound to UDP/Mn^{+2} and EA2 are shown in the top-right and bottom panel, respectively. Small white figures representing different conformations of GalNAc-T2 are shown for clarification purposes (insert). (**b**) (Left panel) R_g distributions derived from the EOM analysis of monomeric GalNAc-T2 apo form at the three concentrations measured (2.5 (red), 5.0 (blue) and 10.0 (green) mg ml^{-1}) identifying ensembles of monomeric compact and extended structures. Overall structures of GalNAc-T2 with different R_gs are shown with lectin and catalytic domains in red and yellow, respectively, and flexible linker is shown in grey. It is noteworthy that our structure of GalNAc-T2-UDP-MUC5AC-13 complex is located in the left peak. (Right panel) Of the three crystallographic dimers tested, the SAXS data fit better with the dimer belonging to the PDB entry 2FFV[14]. (**c**) R_g distributions for the monomeric forms derived from the EOM analysis displaying the decrease of the relative population of the extended conformation upon addition of the glycopeptide. GalNAc-T2 is at a fixed concentration of 5 mg ml^{-1} (black is for the enzyme without the peptide) with increasing amounts of the MUC5AC-13 peptide (1 (blue) and 2 mM (red) of the MUC5AC-13 peptide. (**d**) GalNAc-T2 is in equilibrium with an ensemble of compact and extended structures, and dimeric forms. This equilibrium is partly shifted towards the ensemble of compact structures in the presence of peptides and mainly glycopeptides.

arrangement may explain why this particular dimer can bind glycopeptides. In all three cases, the dimers display similar contacts between the lectin domains and the lectin of one monomer with the catalytic domain of a second monomer (Fig. 4b and Supplementary Fig. 3c).

Notably, the GalNAc-T2-UDP-MUC5AC-13 complex with an R_g of 26.41 Å is one of the most populated compact molecules in solution unlike less representative structures such as the one found for the PDB entry 2FFU[14] with R_g of 30.5 Å (Fig. 4b). This result exemplifies that GalNAc-T2 preferentially adopts conformations that might be important for catalytic aspects of this enzyme.

Further SAXS experiments were carried out in the presence of the ligands, which provided very intriguing results (Supplementary Table 3). The enzyme either without ligands or complexed with UDP or UDP-GalNAc shows a higher percentage of dimers (28–44%) relative to the enzyme in complex with glycopeptides, naked peptides or peptides-UDP (0–28%; Supplementary Table 3). In particular, dimers almost disappear from solution when GalNAc-T2 at a fixed concentration of 5 mg ml^{-1} is incubated with increasing concentrations of Muc5AC-13 (Supplementary Table 3 and Supplementary Figs 3d and 5). Moreover, similar amounts of compact and extended monomeric ensembles exist for the enzyme without ligands or complexed with UDP or UDP-GalNAc; although in some cases, a slight increase in the amount of extended conformations occurs (Supplementary Table 3). This equilibrium is partly shifted to compact monomeric forms in the presence of naked peptides and significantly shifted in the presence of glycopeptides alone or with UDP (reaching values of 59–70% in the presence of peptides; Fig. 4c and Supplementary Table 3). Overall, these results indicate a direct association between the distribution of GalNAc-T2 conformation states and acceptor substrate binding with catalysis, and consequently, the glycosylation profile of GalNAc-T2 (Fig. 4d). In particular, an increase of the monomeric compact structures ensemble appear to be required for catalysis on glycopeptides.

The role of the flexible linker. To further understand the unique dynamics of GalNAc-T2 and the consequences for catalysis, we developed a simple computational model in which the protein is treated as two interacting rigid domains, connected by a flexible linker that is described as a worm-like-chain (WLC model is used to describe the behaviour of semi-flexible polymers; Fig. 5a)[32]. The model predicts, at a qualitative level, the bimodal nature of the R_g distribution revealed by SAXS data (Fig. 5b).

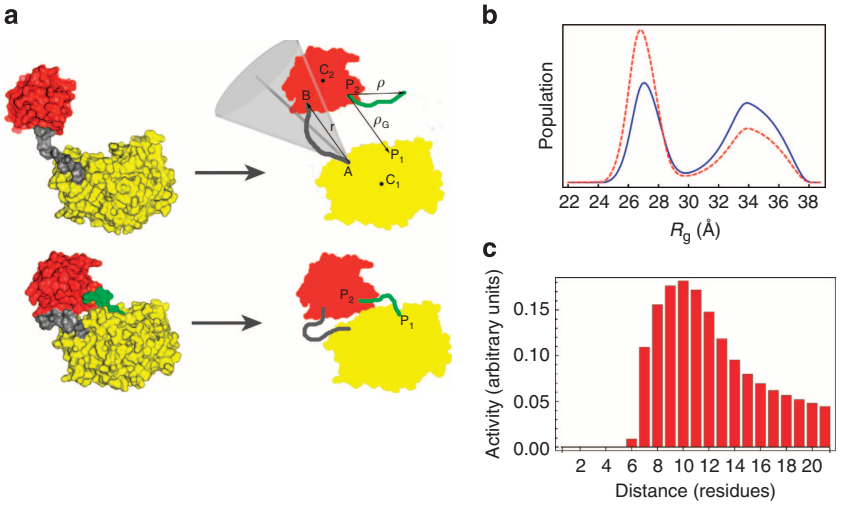


Figure 5 | Coarse-grained theoretical model of GalNAc-T2. (**a**) The crystal structure of GalNAc-T2-UDP-MUC5AC-13 (bottom left) is taken as a reference for the model (bottom, right). The apo form, without the peptide (green), is obtained by removing the latter from the coordinates file. Upon fixing the relative orientation of the two domains (red and yellow for the lectin and catalytic domains, respectively), protein conformations depend on the flexible linker (grey), modelled as a (non-extendable) semi-flexible chain with an attractive interaction between its ends. All other conformations of the protein (top left) are modelled on the basis of the previous one, by letting the end-to-end vector (r) explore the region within a cone (light grey) of angular amplitude $2\theta_0$ (see Supplementary Methods for details). In the holo-form, the peptide (green) is modelled as a WLC of length (l) residues (the distance between the glycosylating sites on the ligand), bound at P_2 on the lectin domain. The enzymatic reaction can take place in any protein conformation providing that the ligand free-end finds correct place P_1 on the catalytic domain. (**b**) Model predictions for the probability distribution [$G(R_g)$] of the radius of gyration R_g. The apo-form is shown as a blue solid line, whereas the holo-form complexed with peptide is shown as a red dashed line. (**c**) Enzymatic activity [$\sigma(l,lc)$] as a function of the residue separation (l) between potential glycosites and a prior fixed glycosite of the glycopeptides (see Supplementary Methods).

Within this model, the peak at high values of the R_g ($R_g = 31$–37 Å) is associated with the equilibrium distribution of the WLC, whereas the peak at low values of the R_g ($R_g = 24.5$–29.5 Å) corresponds to the interaction of both domains (Fig. 5b and Supplementary Fig. 6). In this context, the addition of peptides is expected to increase the effective interdomain interaction by increasing the interaction surface. Accordingly, if we strengthen the interaction energy (ε; see Supplementary Methods), we can reproduce the increase of the peak at lower R_g (Fig. 5b).

Importantly, this simple model can shed light on the different catalytic activity of GalNAc-T2 on glycopeptide substrates in which multiple acceptor sites are found at different distances from a GalNAc *O*-glycan at a fixed position (Fig. 5c). The crystal structure of the GalNAc-T2-UDP-MUC5AC-13 complex suggests that the GalNAc moiety of a glycopeptide substrate may first bind to the lectin domain, which leads to enhance the probability that GalNAc-T2 transfers another GalNAc residue to an acceptor site approximately ten residues N-terminal of the first GalNAc residue located at the lectin domain-binding site. This rationale may help to explain previous work, in which the catalytic efficiency of GalNAc-T2 is higher with glycopeptides than with naked peptides that only contain one acceptor site[23].

Despite the fact that the catalytic process is intrinsically not one at equilibrium, the equilibrium behaviour of our simple model grasps the fundamental aspects of the enzymatic activity. By keeping the first glycosylated site bound to the lectin domain at point P2 (Fig. 5a), only the probability that the acceptor site of the peptide (also described as a WLC) is found in the correct position P1 at the active site is taken into account (see Methods and Supplementary Methods). Our computational model reasonably reproduces not only the previously reported broad glycosylation profile of GalNAc-T2, but also the peak where a maximum in the enzymatic activity is achieved, which in turn corresponds to an acceptor site ten residues apart from a prior site of glycosylation[23] (Fig. 5c).

The flexible linker is a unique feature of the GalNAc-T isoenzymes whose motion is responsible for the GalNAc-T2 dynamics, and consequently, the glycopeptide glycosylation capacity of GalNAc-T2. The flexible linker is such an important structural feature that if we computationally fix it in its position in the crystal structure, the glycosylation activity profile changes and the predicted activity decreases significantly ($\approx$5,000-fold reduction; Supplementary Fig. 7).

In summary, the glycosylation activity profile of GalNAc-T2 is directly coupled to the flexible linker motion that dictates the unique dynamics of GalNAc-T2, and to the GalNAc-T2 binding to glycopeptides that increases the population of compact structures. This model, which is independent of whether the glycopeptides have prior GalNAc *O*-glycans *N*- or *C*-terminal to available to acceptor sites, might help to explain the glycosylation profile of other GalNAc-Ts isoforms.

Discussion

In the present study, we have determined the first crystal structures of GalNAc-T2 in complex with defined GalNAc-glycopeptide substrates. These structures in combination with AFM and SAXS experiments, as well as with theoretical simulations, reveal how GalNAc-T2 selectively glycosylates unused acceptor sites located in the *N*-terminal and optimally ten residues apart from a prior GalNAc glycan. Our results show that GalNAc-T2 populates an ensemble of compact and extended monomeric structures, as well as dimeric ones. In the presence of the acceptor substrates, the dimers disappear while the compact monomeric structures get more populated than the extended one, thus pointing at a prominent role of compact structures for enzymatic activity. Although flexibility in an enzyme is hardly surprising, the picture emerging from our results suggests a peculiar role of the flexible linker, which acts as a conformational dial to foster glycosylation of substrate peptides with different

distances between sites of sequential glycosylation. We demonstrate that the lectin domain, which is a unique feature of the large GalNAc-T isoenzyme family, modulates the catalytic functions and coordinates the follow-up order of glycosylation of protein substrates.

In contrast to the role of the GHs CBMs, GalNAc-Ts lectin domains are not only important to bind to the sugar moiety of glycopeptides, but also to guide catalysis to other unused acceptor sites. This study provides further support for the general model that we proposed for GalNAc-*O*-glycosylation, where a number of GalNAc-T isoenzymes orchestrate *O*-glycosylation in a coordinated manner before and in competition with subsequent elongation of *O*-glycans[2]. In addition, our analysis affords the first experimental evidence of the molecular mechanism of an exquisite additional level of control of *O*-glycosylation. Moreover, the results from the coarse-grained model suggest that the basic features, responsible for flexibility and for the activity profiles, do not depend on the sequence details, so that the proposed model might apply to other GalNAc-T isoenzymes and help explain their dynamics and activity profiles.

Together, these findings may guide the rational design of mechanism-based modulators for this relevant family of enzymes. The finding of inactive and active conformational states in GalNAc-Ts together with the motion of the flexible loop and its association with activity somehow resemble similar features found for the large family of protein kinases. In general, similar states and an activation flexible loop, marked by conserved DFG and APE motifs, are also found in protein kinases and these structural features have been exploited to discover protein kinase inhibitors[33]. We believe that modulators for this particular family of GalNAc-Ts can be developed using similar approaches that will also include the targeting of the lectin domains.

Methods

Cloning, site-directed mutagenesis and purification. The expression plasmid pPICZαA*galnact2* (K75-Q571), previously described[16], was used as a template for introducing the following single amino-acid changes by site-directed mutagenesis as follows: Trp282Ala and Phe361Ala. Site-directed mutagenesis was carried out following the QuikChange protocol (Stratagene), using the Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific). All plasmids were verified by sequencing (Sistemas Genómicos, Valencia, Spain). The mutants were purified using the purification protocol of the wild-type enzyme described previously[16].

Crystallization. In all cases, crystals were grown by hanging drop vapour diffusion at 18 °C. Crystals of the inactive GalNAc-T2 form in complex with the peptide MUC5AC-Cys13 were obtained by mixing 2 µl of protein solution (a mix formed by 7 mg ml^{-1} GalNAcT-2, 5 mM UDP, 5 mM $MnCl_2$, 5 mM MUC5AC-Cys13 peptide in 25 mM Tris (pH 8.0), 0.5 mM EDTA and 1 mM tris(2-carboxyethyl) phosphine (TCEP)) with 2 µl of precipitant solution (25% PEG 4,000, 400 mM ammonium sulfate and 100 mM sodium citrate pH 6) against 500 µl of precipitant solution. Crystals of the inactive GalNAc-T2 form in complex with the peptide MUC5AC-3-13 were obtained by mixing 2 µl of protein solution (a mix formed by 7 mg ml^{-1} GalNAcT-2, 5 mM UDP, 5 mM $MnCl_2$, 6 mM MUC5AC-3–13 peptide in the same buffer described above) with 2 µl of precipitant solution (1.6 M ammonium sulfate and 100 mM Tris, pH 9) and equilibrated against 500 µl of precipitant solution. Crystals of the GalNAc-T2 active form in complex with UDP/Mn^{+2} and the peptide MUC5AC-13 were obtained in similar conditions as the inactive forms. The protein solution consisting of 3.5 mg ml^{-1} of GalNAcT-2, 5 mM UDP, 5 mM $MnCl_2$ and 6 mM MUC5AC-13 in the same buffer described above was mixed with the precipitant solution containing 1.6 M ammonium sulfate, 100 mM sodium chloride and 100 mM HEPES (pH 7). The crystals were cryoprotected in saturated lithium sulfate and frozen in a nitrogen gas stream cooled to 100 K.

Structure determination and refinement. All data were processed and scaled using the XDS package[34] and CCP4 software[35]. Relevant statistics are given in Table 1. The crystal structures were solved by molecular replacement with Phaser[35] and using the PDB entry 4D0T as the template. Initial phases were further improved by cycles of manual model building in Coot[36] and refinement with REFMAC5 (ref. 37). The final models were validated with PROCHECK[38] and model statistics are given in Table 1. The AUs of the tetragonal crystals contain 1 molecule of GalNAc-T2 (Fig. 1). Coordinates and structure factors have been deposited in the Worldwide Protein Data Bank (wwPDB, and see Table 1 for the pdb codes).

Synthesis of peptides and glycopeptides. All peptides were synthesized by stepwise solid-phase peptide synthesis using the Fmoc strategy on Rink Amide MBHA resin (0.1 mmol). *O*-α-D-GlcNAc-L-Thr[39] or *S*-α-D-GlcNAc-L-Cys[26] building blocks (2 equiv.) were prepared as described in the literature and manually coupled. The other Fmoc amino acids were coupled in the automated mode in an Applied Biosystems 433A peptide synthesizer using 10 equiv. and HBTU as a coupling agent. The *O*-acetyl groups of the sugar moiety were deprotected in a mixture of NH_2NH_2/MeOH (7:3). The derivatives were then released from the resin, and all acid-sensitive side-chain-protecting groups simultaneously removed using 95% trifluoroacetic acid (TFA), 2.5% triisopropylsilane (TIS) and 2.5% H_2O, followed by precipitation with diethyl ether. Finally, all the compounds were purified by HPLC on a Waters Delta Prep chromatograph (Phenomenex Luna C18(2) column (10 µ, 21.20 × 250 mm^2)). Additional details are in Supplementary Methods.

Tryptophan fluorescence spectroscopy. Fluorescence spectroscopy was used to determine the dissociation constants of GalNAc-T2 against the peptides MUC5AC, MUC5AC-13 and MUC5AC-3-13 (ref. 40).

All experiments were carried out in a Cary Eclipse spectrofluorometer (Varian) at 25 °C with GalNAc-T2 at 1 µM, and concentrations of peptides varying from 1 to 500 µM in 25 mM Tris, 150 mM NaCl, pH 7.5. The same experiments were also performed in the same buffer but containing 400 µM UDP and 1 mM $MnCl_2$. Fluorescence emission spectra were recorded in the 300–400 nm range with an excitation wavelength of 280 nm, with slit width of 5 nm. The data analysis was performed in Prism (GraphPad software) considering a model with a single binding site (see equation (1), where F0 is the intrinsic fluorescence of the enzyme in the absence of quencher (*Q*), F1 is the observed fluorescence at a given quencher concentration, f_a is the fractional degree of fluorescence and K_d is the dissociation constant.

$$1-\frac{F1}{F0}=\frac{fa*[Q]}{Kd+[Q]} \quad (1)$$

Analysis of the enzymatic reaction by MALDI-TOF-MS. We carried out mass spectrometry analyses to determine the transfer activity of GalNAc-T2 apo form and mutants using UDP-GalNAc and the MUC5AC and MUC5AC-13 acceptor peptides. *In vitro* glycosylation assays were performed as product development assays in 50 µl buffer (25 mM cacodylic acid sodium, pH 7.4, 10 mM $MnCl_2$, 0.25% Triton X-100), 5 mM UDP-GalNAc (Sigma), 0.2 mM of acceptor peptides and 0.4 µM of purified GalNAc-T2 wild-type or mutant enzymes at 37 °C. For time-course evaluation, 2 µl of the reaction mixtures were taken at 10 min, 20 min, 40 min, 1 h, 2 h, 4 h, 20 h and 44 h (2 mM of UDP-GalNAc and 0.2 µM of enzyme were added into the above mixture after 20 h reaction to push the reaction for another 24 h), and analysed by MALDI-TOF-MS. Evaluation of incorporation of GalNAc residues into peptide substrates was performed by matrix-assisted laser desorption/ionization–time of flight-mass spectrometry (MALDI-TOF-MS). 2 µl of reaction mixtures were diluted with 18 µl of 0.1% TFA/H_2O, and 1 µl mixed with 1 µl of 10 mg ml^{-1} 2,5-dihydrobenzoic acid dissolved in ACN/H_2O (7:3). Acquisition of MS spectra was performed on MALDI-TOF instrument, Bruker Autoflex using Bruker FlexControl 3.4 software. Spectra were recorded in the positive ion mode and the raw spectra were processed by Bruker FlexAnalysis 3.4 software. The specific activities under linear conditions were inferred from the mass spectrometry data. One unit of enzyme is defined as the amount of enzyme that transfers 1 µmol GalNAc in 1 min using the standard reaction mixture.

Computational details. MD simulations of the enzyme were performed with the Amber11 software package. The protein was modelled with the FF99SB force field, whereas all carbohydrate molecules were modelled with the GLYCAM06 force field[41]. A snapshot of the equilibrium MD simulation was taken for the metadynamics simulations. Additional details are in Supplementary Methods.

AFM imaging. AFM measurements were performed using a MultiMode 8 AFM system (Bruker). Images were taken using the Tapping Mode with V-shaped silicon nitride cantilevers with integrated pyramidal 2 nm ultrasharp tips exhibiting a spring constant and a frequency of 0.03 N m^{-1} and 15 kHz, respectively (MSNL-D; Bruker Probes). Additional details are in Supplementary Methods.

SAXS and data analysis. SAXS measurements were performed at the European Molecular Biology Laboratory on the storage ring PETRA-III (DESY-Hamburg) on the P12 beamline equipped with a robotic sample changer and a PILATUS-2M. The analysis of the SAXS data was performed by using the EOM[31]. Additional details are in Supplementary Methods.

Coarse-grained model. We used as a template the crystal structure of the GalNAc-T2-UDP-MUC5AC-13 complex to generate our coarse-grained model. The protein was treated as two interacting rigid domains (the catalytic and the lectin

domains) joined by a 12-residue semi-flexible linker that is described by a semi-elastic WLC model[32,42,43] (Fig. 5a). We calculated the equilibrium canonical (Boltzmann) distribution of the end-to-end vector of the linker, as well as that of the glycopeptide, also considered as a WLC. The probability distribution of the R_g was derived from that of the end-to-end vector. All calculations were performed with Wolfram's Mathematica 9.0. Additional details are in Supplementary Methods.

References

1. Steentoft, C. *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32,** 1478–1488 (2013).
2. 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).
3. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42,** D490–D495 (2014).
4. 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).
5. 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).
6. Schjoldager, K. T. & Clausen, H. Site-specific protein O-glycosylation modulates proprotein processing—deciphering specific functions of the large polypeptide GalNAc-transferase gene family. *Biochim. Biophys. Acta* **1820,** 2079–2094 (2012).
7. Springer, G. F. T and Tn, general carcinoma autoantigens. *Science* **224,** 1198–1206 (1984).
8. Tarp, M. A. & Clausen, H. Mucin-type O-glycosylation and its potential use in drug and vaccine development. *Biochim. Biophys. Acta* **1780,** 546–563 (2008).
9. Springer, G. F. Immunoreactive T and Tn epitopes in cancer diagnosis, prognosis, and immunotherapy. *J. Mol. Med. (Berl.)* **75,** 594–602 (1997).
10. Ju, T., Otto, V. I. & Cummings, R. D. The Tn antigen-structural simplicity and biological complexity. *Angew. Chem. Int. Ed. Engl.* **50,** 1770–1791 (2011).
11. Gill, D. J. *et al.* Initiation of GalNAc-type O-glycosylation in the endoplasmic reticulum promotes cancer cell invasiveness. *Proc. Natl Acad. Sci. USA* **110,** E3152–E3161 (2013).
12. Radhakrishnan, P. *et al.* Immature truncated O-glycophenotype of cancer directly induces oncogenic features. *Proc. Natl Acad. Sci. USA* **111,** E4066–E4075 (2014).
13. 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 alpha-N-acetylgalactosaminyltransferase-T1. *Proc. Natl Acad. Sci. USA* **101,** 15307–15312 (2004).
14. 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).
15. Kubota, T. *et al.* Structural basis of carbohydrate transfer activity by human UDP-GalNAc: polypeptide alpha-N-acetylgalactosaminyltransferase (pp-GalNAc-T10). *J. Mol. Biol.* **359,** 708–727 (2006).
16. Lira-Navarrete, E. *et al.* Substrate-guided front-face reaction revealed by combined structural snapshots and metadynamics for the polypeptide N-acetylgalactosaminyltransferase 2. *Angew. Chem. Int. Ed. Engl.* **53,** 8206–8210 (2014).
17. 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).
18. Pedersen, J. W. *et al.* Lectin domains of polypeptide GalNAc transferases exhibit glycopeptide binding specificity. *J. Biol. Chem.* **286,** 32684–32696 (2011).
19. Herve, C. *et al.* Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proc. Natl Acad. Sci. USA* **107,** 15293–15298 (2010).
20. Ficko-Blean, E. & Boraston, A. B. Insights into the recognition of the human glycome by microbial carbohydrate-binding modules. *Curr. Opin. Struct. Biol.* **22,** 570–577 (2012).
21. Gill, D. J., Clausen, H. & Bard, F. Location, location, location: new insights into O-GalNAc protein glycosylation. *Trends Cell Biol.* **21,** 149–158 (2011).
22. Topaz, O. *et al.* Mutations in GALNT3, encoding a protein involved in O-linked glycosylation, cause familial tumoral calcinosis. *Nature Genet.* **36,** 579–581 (2004).
23. 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).
24. Sugiura, M., Kawasaki, T. & Yamashina, I. Purification and characterization of UDP-GalNAc:polypeptide N-acetylgalactosamine transferase from an ascites hepatoma, AH 66. *J. Biol. Chem.* **257,** 9501–9507 (1982).
25. Elhammer, A. & Kornfeld, S. Purification and characterization of UDP-N-acetylgalactosamine: polypeptide N-acetylgalactosaminyltransferase from bovine colostrum and murine lymphoma BW5147 cells. *J. Biol. Chem.* **261,** 5249–5255 (1986).
26. Aydillo, C. *et al.* S-Michael additions to chiral dehydroalanines as an entry to glycosylated cysteines and a sulfa-Tn antigen mimic. *J. Am. Chem. Soc.* **136,** 789–800 (2014).
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. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl Acad. Sci. USA* **99,** 12562–12566 (2002).
29. Love, R. A. *et al.* The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site. *Cell* **87,** 331–342 (1996).
30. Svergun, D. I., Barberato, C. & Koch, M. H. J. CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst* **28,** 6 (1995).
31. Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **129,** 5656–5664 (2007).
32. Kratky, O. & Porod, G. Röntgenuntersuchung gelöster Fadenmoleküle. *Rec. Trav. Chim* **68,** 18 (1949).
33. Endicott, J. A., Noble, M. E. & Johnson, L. N. The structural basis for control of eukaryotic protein kinases. *Annu. Rev. Biochem.* **81,** 587–613 (2012).
34. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 125–132 (2010).
35. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 235–242 (2011).
36. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132 (2004).
37. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 355–367 (2011).
38. Laskowski, R. A. *et al.* PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26,** 283–291 (1993).
39. Plattner, C., Hofener, M. & Sewald, N. One-pot azidochlorination of glycals. *Org. Lett.* **13,** 545–547 (2011).
40. Rubio-Ruiz, B. *et al.* Discovery of a new binding site on human choline kinase alpha1: design, synthesis, crystallographic studies, and biological evaluation of asymmetrical bispyridinium derivatives. *J. Med. Chem.* **57,** 507–515 (2014).
41. Kirschner, K. N. *et al.* GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **29,** 622–655 (2008).
42. Becker, N. B., Rosa, A. & Everaers, R. The radial distribution function of worm-like chains. *Eur. Phys. J. E Soft Matter* **32,** 53–69 (2010).
43. Zhou, H.-X. Loops in Proteins Can Be Modeled as Worm-Like Chains. *J. Phys. Chem. B* **105,** 4 (2001).

Acknowledgements

We thank synchrotron radiation sources DLS (Oxford), ALBA (Barcelona) and PETRA-III (DESY/Hamburg), and in particular beamlines I03 (experiment number MX8035-21 and MX8035-24), I04-1 (experiment number MX8035-17), XALOC and P12 (SAXS-172), respectively. We thank ARAID, the MEC (BFU2010-19504, CTQ2013-44367-C2-2-P, CTQ2011-25871, BIO2010-14983, CTQ2012-36365 and MAT2012-38318), the DNRF (DNRF107), the DGA (B89 and B18), the Generalitat de Catalunya (2014SGR-987), ANR-CHEX-2011 and ATIP-Avenir (PB) for financial support. The research leading to these results has also received funding from the FP7 (2007-2013) under BioStruct-X (grant agreement N°283570 and BIOSTRUCTX_5186). We also acknowledge the computer support, technical expertise, and assistance provided by the Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS).

Author contributions

R.H.-G. designed the crystallization construct and solved the crystal structures. E.L.-N., M.R. and R.H.-G. purified the enzymes, crystallized the different complexes and refined the crystal structures. I.C., F.C., G.J.L.B. and J.M.P. synthetized the glycopeptides. M.C.P. and A.L. performed the AFM studies. E.L.-N. and P.Be. performed the SAXS experiments. P.Br. performed the coarse-grained model. J.I.-F. and C.R. performed the metadynamics simulations. Y.K. and H.C. perfomed the analysis of the enzymatic reaction by MALDI-TOF MS. R.H.-G. wrote the article with the main contribution of F.C., G.J.L.B., H.C., C.R., P.Br, P.Be. and A.L. All authors read and approved the final manuscript.

Additional information

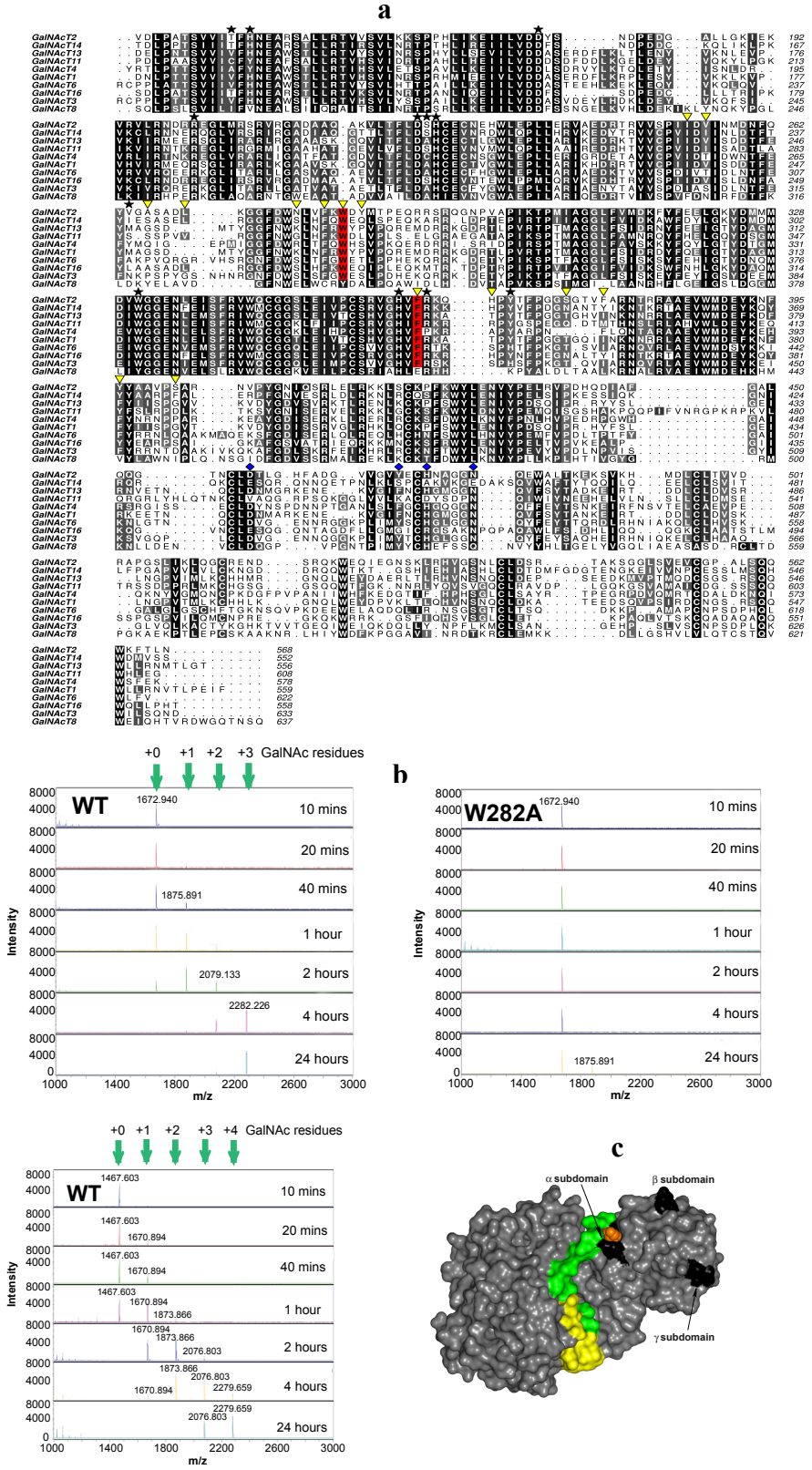
Accession codes: Coordinates and structure factors have been deposited in the World-wide Protein Data Bank (wwPDB) with accession codes **5AJN**, **5AJO** and **5AJP**.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

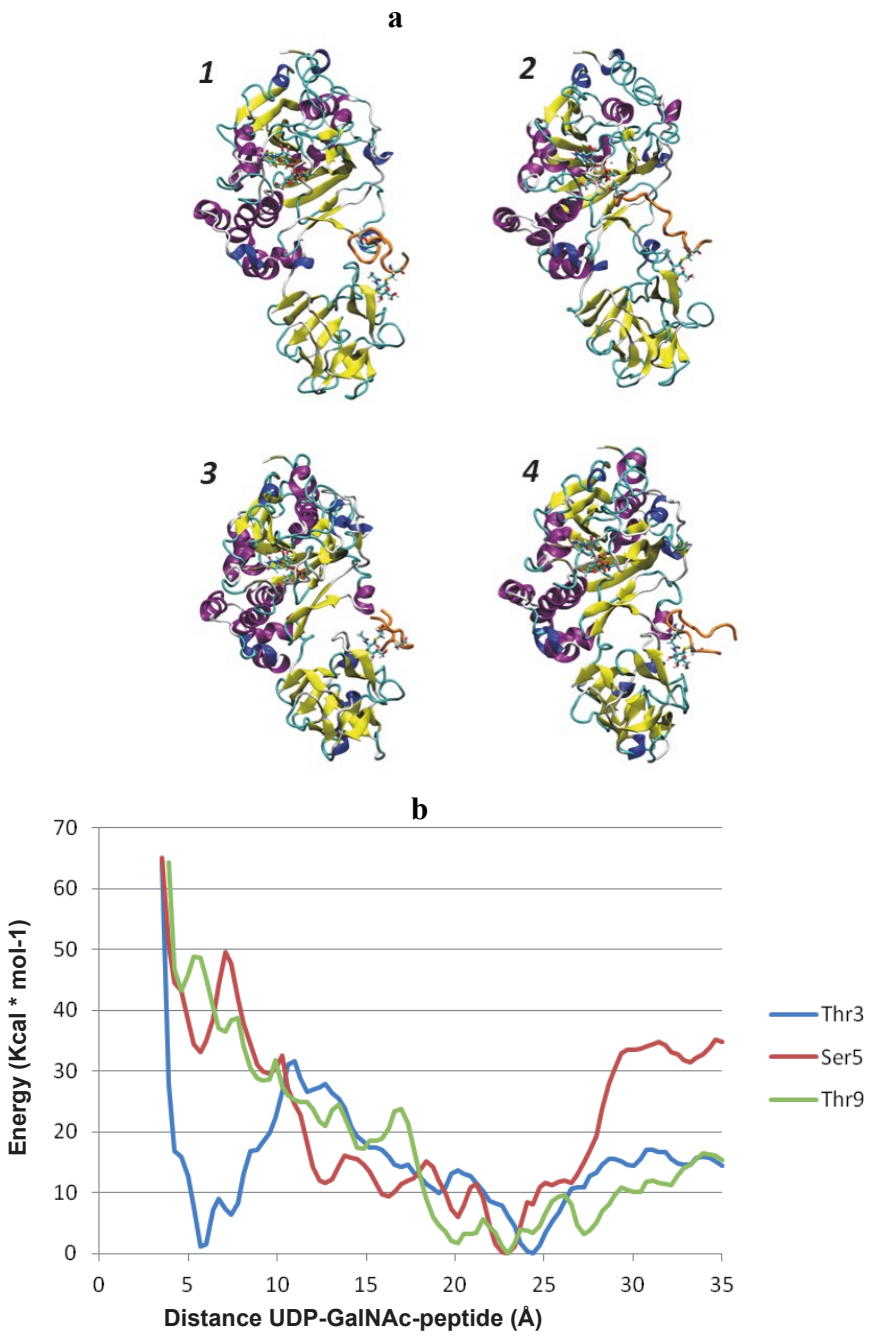
Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

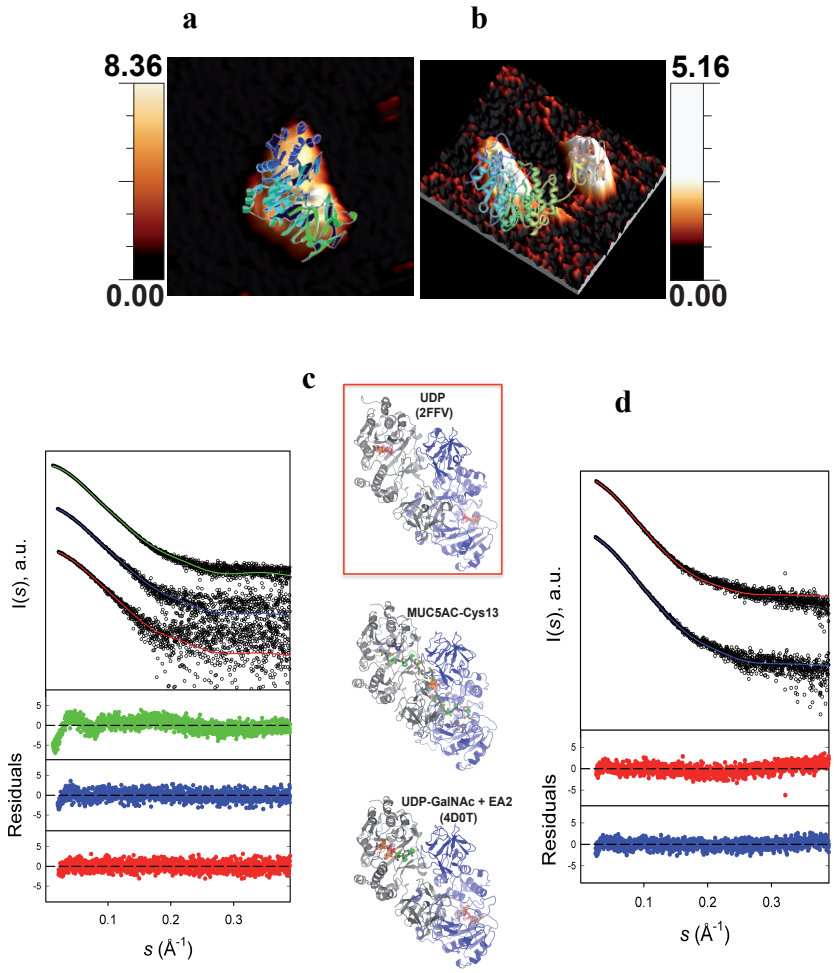
How to cite this article: Lira-Navarrete, E. *et al.* Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein *O*-glycosylation. *Nat. Commun.* 6:6937 doi: 10.1038/ncomms7937 (2015).



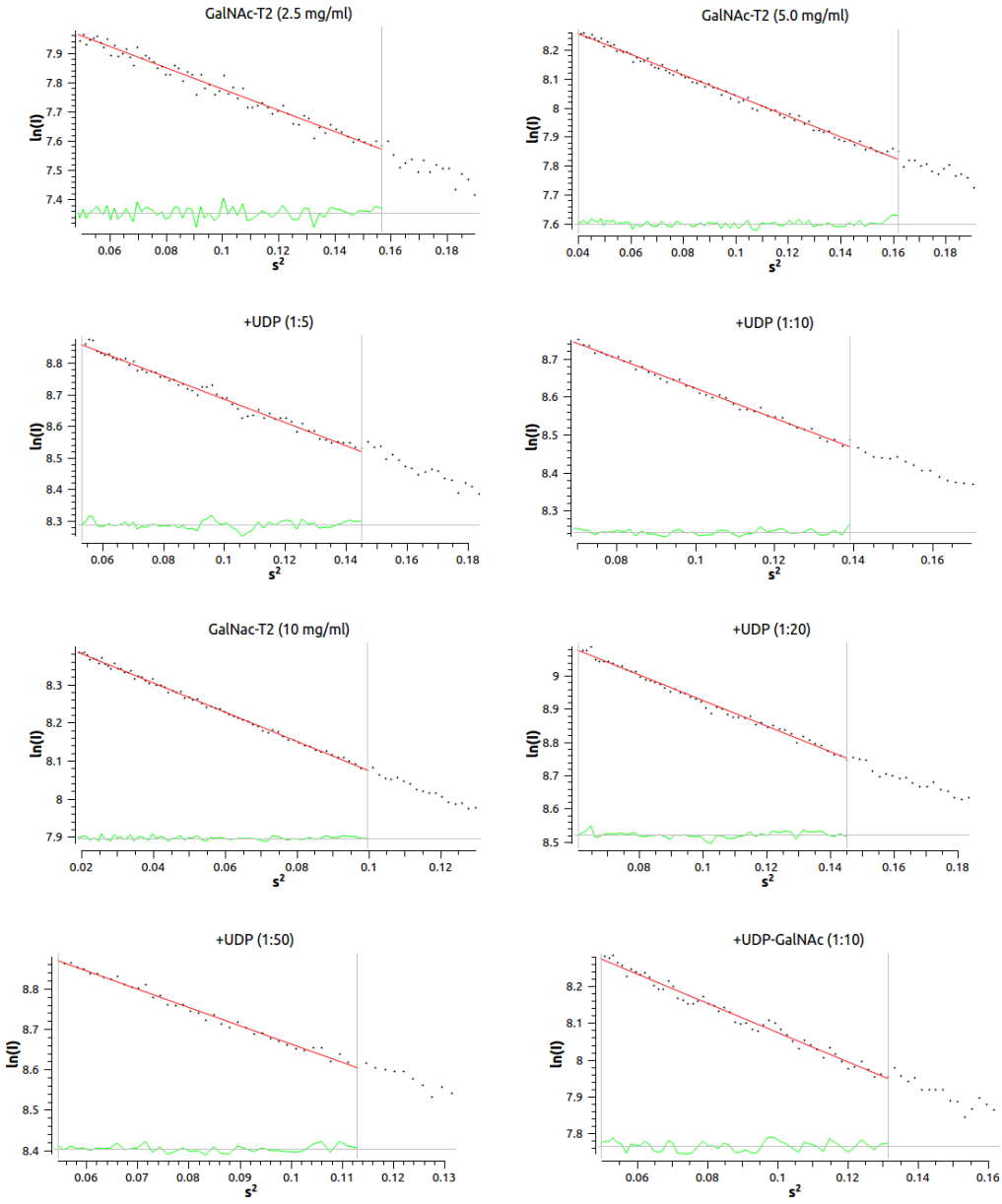
Supplementary Figure 1. (**a**) Sequence alignment of GalNAc-T2 (residues 133-568) with other GalNAc-Ts isoforms. Conserved residues are highlighted in black and similar residues are in grey. Residues forming the sugar-nucleotide, peptide and lectin domain binding sites are indicated as black stars, yellow inverted triangles and blue diamonds, respectively. Mutated residues are highlighted in red. (**b**) Time-course of the enzymatic reaction of the wild type enzyme and the mutant W282A with UDP-GalNAc and the MUC5AC-13 peptide followed by MALDI-TOF MS (Top panel). Time-course of the enzymatic reaction of the wild type enzyme with UDP-GalNAc and the MUC5AC peptide (Bottom panel). Green arrows indicate the MUC5AC-13 and MUC5AC peptides and the glycosylated forms, respectively. **(c)** Surface representation of GalNAc-T2 in complex with UDP and MUC5AC-13. Colors are indicated as **Fig. 1c** except for black colors that represent residues located in the different subdomains of the lectin domain**.**



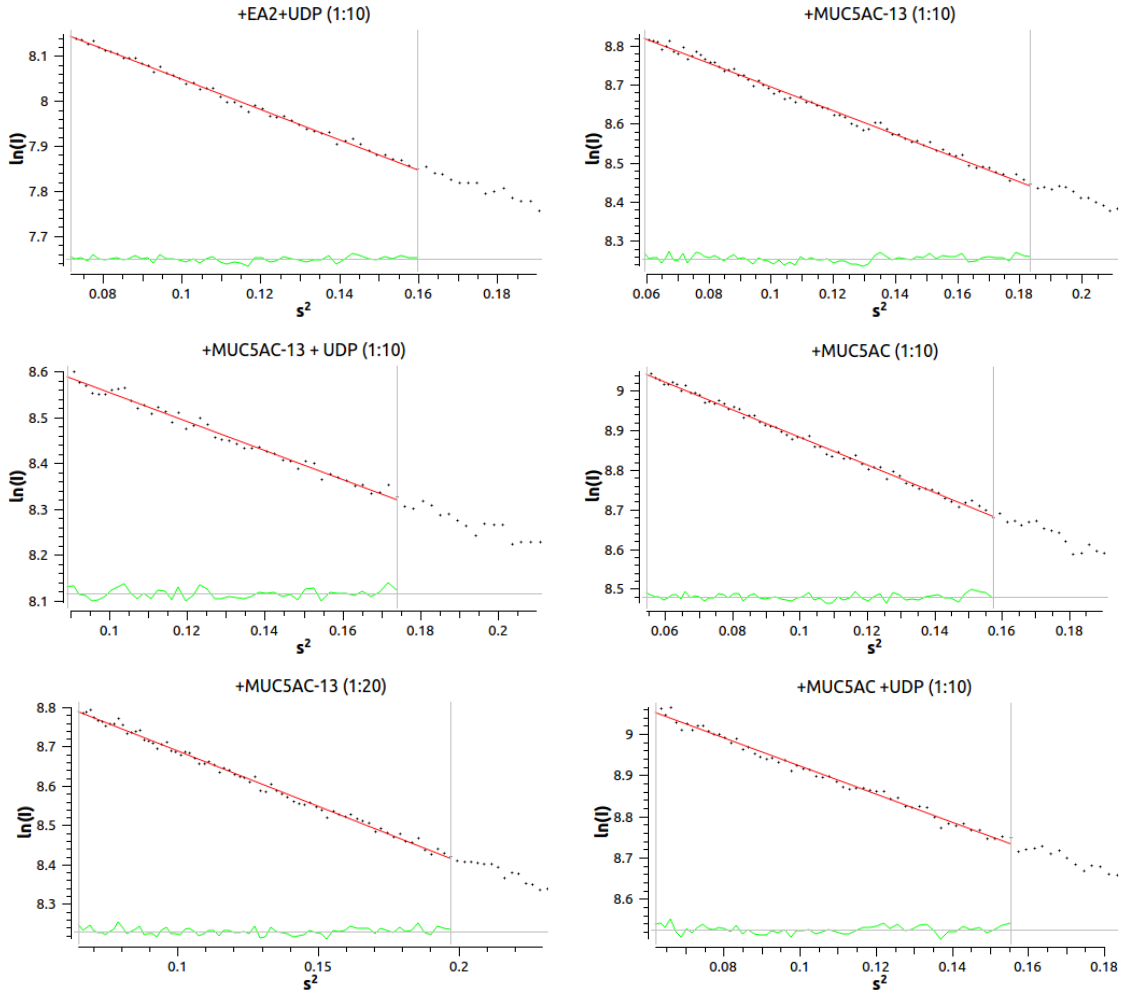
Supplementary Figure 2. (**a**) Representative conformations extracted from the metadynamics simulations. Conformations **1-4** correspond to those indicated in **Fig. 3d**. (**b**) Free energy of glycopeptide binding *versus* the distance between the acceptor amino acid and the UDP-GalNAc donor. The energy profile has been obtained by integrating the FEL with respect to the RMSD collective variable. It is clear that Thr3 is the unique glycosylation site with an energy minimum at short distances (approximately 5 Å), whereas the other acceptor sites present an energy minimum at higher distances.



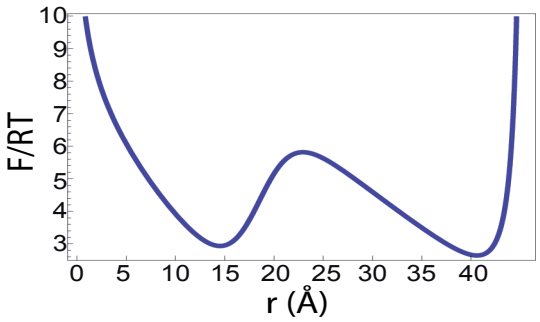
Supplementary Figure 3. (a) Overlay of the monomeric form of the GalNAc-T2-UDP complex (PDB entry 2FFV) on the 3D AFM topography image of a compact structure found in the presence of UDP-MUC5AC-13, and (**b**) overlay of GalNAc-T2-UDP-EA2 complex (PDB entry 2FFU) on the 3D AFM topography image of an extended molecule complexed with EA2 and UDP. (**c**) (Left panel) SAXS intensity profiles, [*I(s)*], as a function of the momentum transfer, (*s*), measured for GalNAc-T2 apo form at three concentrations (empty dots). EOM fits of the three profiles are displayed as solid lines for the 2.5 (red), 5.0 (blue) and 10.0 (green) mg/ml. Profiles have been displaced along the *I(s)* axis for clarity. The point-by-point residuals (same color-code) that indicate the high quality of the fit in the complete range of momentum transfer. (Right panel) Of the three crystallographic dimers tested, the SAXS data fits better with the dimer belonging to the PDB entry 2FFV. (**d**), SAXS intensity profiles [*I(s)*] as a function of the momentum transfer (*s*) for GalNAc-T2 at a fixed concentration of 5 mg/ml with increasing amounts of the MUC5AC-13 peptide. EOM fits are displayed as solid lines for the 1 (blue) and 2 mM (red) of the MUC5AC-13 peptide. The quality of the fit is shown in the point-by-point residuals displayed below with the same color-code.



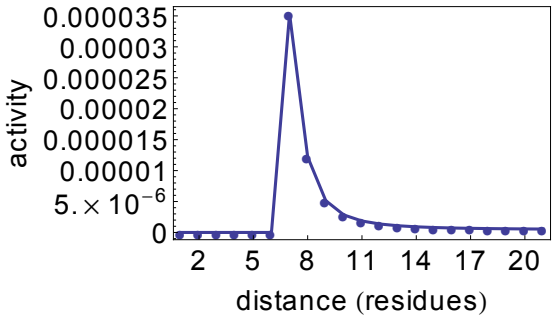
Supplementary Figure 4. Guinier's fits corresponding to samples described in **Supplementary Table 3**. In all the cases, an excellent linear fit is obtained allowing to derive precise radius of gyration, Rg, and forward scattering values, I(0).



Supplementary Figure 5. Guinier's fits corresponding to samples described in **Supplementary Table 3** . In all the cases, an excellent linear fit is obtained allowing to derive precise radius of gyration, Rg, and forward scattering values, I(0).



Supplementary Figure 6. Free-energy profile as a function of the end-to-end distance *r* of the flexible linker. The peak at high values of the R_g (R_g corresponding to values of 31-37 Å) is associated to the equilibrium distribution of the WLC model, where a balance between entropy (favoring shorter end-to-end distances *r)* and elastic energy (providing rigidity against bending of the linker) produces a free-energy minimum at high values of *r*. The peak at low values of R_g (R_g corresponding to values of 24.5-29.5 Å) is related to the interaction of the two domains, yielding the minimum at low values of *r*. The values of the parameters used are *lc* = 12 peptide bonds (1 bond = 3.8 A), lp = 0.6 *lc*, r_m = 18.5 Å, r_w =1.8 Å, θ_0 = arccos(0.42), ε/RT=6.4.



Supplementary Figure 7. Predicted enzymatic activity $\sigma(l,lc)$ (see Supplementary Experimental Procedures) as a function of the residue separation l between the potential acceptor site and a prior fixed glycosite of the glycopeptides. In this plot the flexible linker is fixed to its crystallographic structure. This modification leads to a decrease of ≈ 5,000 fold in activity with respect to that reported in Fig. 5c and changes in the glycosylation profile of the enzyme.

Supplementary Table 1. Name and sequence of the peptides used in this study

Peptide	Sequence
MUC5AC	GTTPSPVPTTSTTSA
MUC5AC-Cys13	GTTPSPVPTTSCT*SA
MUC5AC-13	GTTPSPVPTTSTT*SA
MUC5AC-3-13	GTT*PSPVPTTS TT*SA
MUC5ACs-Cys9	GTTPSPVPC*TS
EA2	DSTTPAPTTK

* indicates a GalNAc moiety linked to the underlined glycosylated amino acid.

Supplementary Table 2. Trytophan fluorescence spectroscopy data.

Peptides	K_ds (μM)
MUC5AC	7 ± 3
MUC5AC (UDP/Mn^{2+})	31 ± 16
MUC5AC-13	8 ± 3
MUC5AC-13 (UDP/Mn^{2+})	5 ± 1
MUC5AC-3-13	11 ± 4
MUC5AC-3-13 (UDP/Mn^{2+})	17 ± 9

The K_ds were obtained in the absence and presence of UDP (400 μM) and $MnCl_2$ (1 mM) (Supplementary Information).

Supplementary Table 3. Analysis of SAXS data measured for GalNAc-T2 in different experimental conditions.

Ligands (Molar Ratio of GalNAc-T2:Ligands)	GalNAc T2 (mg/ml)	R_g (Å)[e]	R_g (Å)[h]	D_{max} (Å)[f]	MW (kDa)[g]	χ_i^2 mon/dim[i]	χ_i^2 [a]	Ratio of Monomer:Dimer [b]	Ratio of Compact:Extended monomeric structures [c]
	2.5	32.9±0.5	34.8	125	65	1.11/2.03	0.92	61:39	40:60
	5	32.7±0.2	34.2	127	58	1.35/4.49	1.17	72:28	47:53
	10	33.9±0.1	34.4	127	63	3.51/10.19	1.63	62:38	40:60
UDP (1:5)	5	32.7±0.3	34.3	121	67	1.81/3.10	0.82	60:40	46:54
UDP (1:10)	5[d]	34.4±0.3	35.0	123	68	2.00/4.07	1.00	61:39	42:58
UDP (1:20)	5	34.0±0.2	34.5	122	71	2.84/4.66	1.20	56:44	40:60
UDP (1:50)	5	36.0±0.3	36.0	129	68	2.73/5.78	1.95	64:36	43:57
UDP-GalNAc (1:10)	5[d]	34.7±0.4	34.6	124	62	1.14/3.21	0.76	67:33	40:60
EA2+UDP (1:10)	5[d]	31.8±0.2	32.1	112	54	2.15/11.16	1.17	85:15	63:37
MUC5AC-13 (1:10)	5[d]	30.2±0.2	30.6	112	50	1.96/7.40	0.74	92:8	70:30
	5	29.1±0.1	29.2	102	46	2.50/8.88	1.21	100:0	68:32
MUC5AC-13 (1:20)	5[d]	30.7±0.4	31.7	112	53	1.60/5.67	1.03	89:11	70:30
MUC5AC-13+UDP (1:10)									
MUC5AC (1:10)	5[d]	32.3±0.3	33.4	120	59	1.47/5.67	0.79	72:28	59:41
	5[d]	32.0±0.2	33.3	120	55	1.39/7.04	0.81	82:18	61:39
MUC5AC+UDP (1:10)									

(a)- Quality of the EOM fit achieved by using a pool containing monomeric conformations and the PDB entry 2FFV as a dimer. Excellent fits for the SAXS bimodal curve are obtained with low values of $\chi i2$ that indicate the pool of GalNAc-T2 conformations represents a good description of the behaviour of the protein in solution.

(b)- Relative percentage of monomers and dimers of GalNAc-T2.

(c)- Relative percentage of compact (R_g < 30.5 Å) and extended (R_g > 30.5 Å) for the monomeric conformations derived from the EOM fit.

(d)- SAXS curves at 1, 2.5 and 10 mg/ml were also measured and analysed. They are omitted here for simplicity.

(e)- Radii of gyration, R_g, determined using Guinier's approach.

(f)- Maximum intramolecular distance, D_{max}, determined from the pair-wise distance distribution, $p(r)$, with the program GNOM.

(g) Estimated molecular weight of the particles in solution based on Porod's volume computed with PRIMUS divided by 1.6. Theoretical MW of GalNAc-T2 is 56.7 kDa.

(h)- Radius of gyration, R_g, derived from the $p(r)$ computed with the program GNOM.

(i)- Fitting of the crystallographic monomeric (1FFU) and dimeric (1FFV) structures of GalNAc-T2 to the experimental data using CRYSOL.

Supplementary Methods

Synthesis of peptides and glycopeptides

Our initial rationale to choose the glycopeptides in this work was inspired by an earlier work in which the glycosylation profile of GalNAc-T2 differed depending on whether naked MUC5AC or glycopeptides such as MUC5AC-13, MUC5AC-3 and MUC5AC-3-13 were used as initial substrates (**Supplementary Table 1**). Among other conclusions it was reported that the glycosylation of MUC5AC-13 took place optimally 10 residues N-terminal (Thr3) from the previous glycosylated site Thr13 and suggested to be driven by the lectin domain. Further sites such as Ser5 and Thr9 were also glycosylated but not in an efficient manner[1]. We synthesized 4 glycopeptides to understand the role of the lectin domain in catalysis. While MUC5AC-Cys13 and MUC5AC-13 share the same position of glycosylation (position 13), they differ in the identity of the glycosylated underlying amino acid, a Thr residue in the latter and a Cys residue in the former. An earlier computational study suggested that Cys-S-GalNAc mimicked fairly well the perpendicular conformation of the GalNAc moiety linked to Thr with respect to the backbone peptide[2]. Therefore we introduced a Cys in the peptides to probe experimentally the conformation of the GalNAc moiety in relation to the peptide backbone. The other two glycopeptides, MUC5AC-3-13 and MUC5ACs-C9 (a shorter peptide that also contains a Cys residue linked to a GalNAc moiety), were synthetized to understand how acceptor sites such as Thr3 of MUC5ACs-C9 and Ser5 of MUC5AC-3-13, closely located to previous C-terminal glycosites, are glycosylated.

Computational details

1.1. Initial structure

The crystal structure of GalNAc-T2 in complex with the peptide MUC5AC-Cys13 was used for the computational studies. UDP-GalNAc and the active conformation of the flexible loop, taken from the PDB entry 4D0T, were added to obtain a catalytically productive enzyme complex. To study the glycosylation capabilities of the peptide at different amino acid positions, the peptide was manually extracted from its position in the above complex and placed outside the protein, within a fully solvated environment. Protonation states and hydrogen atom positions of all ionizable amino acids residues were selected base on their hydrogen bond environment. Eight histidine residues were modeled in their neutral states and four in their protonated state. All the crystallographic water molecules were retained and extra water molecules were added to form a 20 Å water box around the protein surface. Nine chloride ions were also added to neutralize the enzyme charge.

1.2 Classical molecular dynamics simulations

MD simulations of the enzyme were performed with the Amber11 software package. The protein was modeled with the FF99SB force field, whereas all carbohydrate molecules were modeled with the GLYCAM06 force field[3]. The MD simulations were carried out in several steps. First, the system was minimized, maintaining the protein, peptide and substrate molecules fixed. In a second step, the entire system was allowed to relax. Weak spatial constraints were initially added to the protein and substrates to gradually reach the desired temperature of 300 K, while the rest of the system was allowed to move freely. The constraints were subsequently removed and the system was subjected to 100 ps of constant pressure MD simulation to adjust the

density of the water environment. Afterwards, 100 ns of constant volume MD simulation were performed. During this time, the peptide molecule remained in a solvated environment without approaching the active site of the enzyme. As no one of the three possible glycosylation positions of the peptide approached the active site during the 100 ns time-scale window, this process was subsequently activated using the metadynamics algorithm[4].

1.2. Classical metadynamics simulations of substrate binding

A snapshot of the equilibrium MD simulation was taken for the metadynamics simulations, which were performed with NAMD2.9 software. As a first approach to model the binding process, only the distance between the center of mass of the side chain of the acceptor amino acid (Thr3, Ser5 and Thr9) and the UDP-GalNAc donor was taken as a collective variable. Therefore three independent metadynamics simulations were carried out for each one of the glycosylation positions. These metadynamics simulations showed a huge variability in the free energy output and did not converge to a stable free energy profile. In an attempt to improve the results, nine additional simulations, with random initial velocities, were launched for each one of the three glycosylation sites. Unfortunately, averaging of the resulting free energy profiles (10 for each glycosylation site) showed large standard deviations. Analysis of the metadynamics trajectories showed that the variability in the results was probably due to the high flexibility of the peptide molecule and the different modes of approaching the active site. To solve this problem, a second collective variable was added to account for the different conformations of the peptide (the RMSD of the C_α carbon atoms of the peptide was chosen). Therefore, three independent metadynamics

simulations were performed with the acceptor amino acid – UDPGalNAc distance and the RMSD of the C_α peptide carbon atoms as collective variables.

The values of the height and width of the Gaussian-like biasing potential were selected as 1.0 kcal·mol^{-1} and 0.25 Å, respectively. A temperature window of ΔT = 25000 K for the well-tempered algorithm was used together with a deposition time of 1 ps. The simulation was continued until the system completely explored the free energy landscape several times which, in terms of the simulation time, corresponds to approximately 100 ns for each metadynamics run.

Atomic Force Microscopy Imaging

In Tapping Mode image operation the cantilever driven by a piezoelectric actuator vibrates near its resonance frequency. Upon approaching the sample, the tip briefly touches the surface at the bottom of each swing, resulting in a decrease in oscillation amplitude. By maintaining constant oscillation amplitude, high-resolution images of the topography of the surface may be obtained on soft samples.

AFM scanning requires the sample be immobilized on a nanoflat surface not to be swept away. Samples of GalNAc-T2 at 2.0 nM in PBS, pH 7.4, were incubated on 1 cm^2 freshly cleaved muscovite mica pieces (Electron Microscopy Sciences) for 10 min at room temperature. The concentration of the protein incubated on the mica sheets was suitable to get isolated features that could be analyzed individually. The enzyme was adsorbed electrostatically on the negatively charged mica surface. GalNAc-T2 presents a net positive charge at the working pH conditions due to its theoretical isoelectric point of 8.3. The immobilization of enzymes on mica was previously evaluated observing clearly they preserve the enzymatic activity[5].

To determine how the ligands affect the conformational dynamics of GalNAc-T2, the enzyme was also incubated with ligands for 10 min under mild stirring at room temperature. UDP, UDP-GalNAc and $MnCl_2$ were used at 20.0 nM and the different peptides were added at 10.0 nM. After sample incubation, the substrate was washed extensively with the same buffer to remove weakly joined molecules. The immobilized sample and the cantilever holder were introduced into a liquid cell (previously cleaned with 20 % isopropanol and Millipore ultrapure water). AFM measurements were conducted also in PBS, pH 7.4, at 20 °C. AFM images were further analyzed by using the WSxM software[6]. Three samples per condition were assayed. At least 10 images of 10 different areas of 500 nm^2 were analyzed for every sample and environment. Furthermore, each feature or associate was analyzed in detail with the zoom function of the WSxM program, performed without losing image information and or the discarding of artifacts.

The height discussed in the main text refers to the Z-height that has sub-nanometric resolution because of the accuracy of piezoelectric scanners. However this does not occur in the X–Y plane, where the scanned features suffer the well documented AFM tip broadening effect that arise in higher sizes[7]. This effect does not affect the comparative analysis of the width related to the size or the conformational state of the protein molecules due to proportionality.

Small-angle X-ray scattering (SAXS) and data analysis

All experiments were performed in a buffer consisting of 25 mM TRIS pH 7.5, 10 mM MnCl2. The first set of experiments contained three different protein concentrations: 10 mg/ml (180 μM), 5 mg/ml (90 μM), 2.5 mg/ml (45 μM) and 1

mg/ml (18 μM). The protein was measured alone and in the presence of ten times molar excess of UDP, UDP-GalNAc, UDP-EA2 peptide, MUC5AC-13, UDP-MUC5AC-13, MUC5AC, and UDP-MUC5AC. The second set of experiments was carried out with the protein concentration fixed at 5 mg/ml, and increasing concentrations of UDP (450 μM, 900 μM, 1.8 mM, 4.5 mM and 9 mM). The concentration of MUC5AC-13 was also changed as follows: 1 mM, 5 mM and 7.5 mM.

Synchrotron SAXS measurements were performed at the European Molecular Biology Laboratory (EMBL) on the storage ring PETRA-III (DESY-Hamburg) on the P12 beamline equipped with a robotic sample changer and a PILATUS-2M. The sample-detector distance was 3.1 m. The scattering intensity, I(s), was recorded at 10ºC in a momentum transfer range of $0.007<s<0.444$ Å^{-1}, where $s = (4\pi sin\theta)/\lambda$, 2θ is the scattering angle, and λ = 1.24 Å is the X-ray wavelength. Twenty consecutive 0.05 second measurements were performed for protein samples and buffers to monitor for radiation damage that were subsequently averaged (no systematic radiation effects were observed). Buffer scattering profiles were subtracted from the sample ones using standard procedures [8]. The values of the forward scattering, *I(0)*, and radii of gyration, R_g, were calculated from the experimental SAXS patterns using Guinier's approximation (**Supplementary Fig. 4** and **5**),

$I(s) = I(0)exp(-s^2R_g^2/3)$, which is valid in the range $sR_g < 1.3$.

Structural analysis of the SAXS data was performed using the Ensemble Optimization Method (EOM)[9]. A pool of 10,000 monomeric structures of GalNac-T2 was constructed with the program RanCh[9] by connecting three-dimensional structures of the catalytic and the lectin domains, extracted from the crystallographic structure 2FFU, with a flexible linker of 12 residues with random coil conformations. To

properly describe the data, these monomeric ensembles were enriched with 100 equivalent curves corresponding to one dimeric arrangement. As we tested three dimers, we built three different ensembles with the same 10,000 monomeric conformations and 100 copies of each one of the dimers. The three different dimers used for SAXS are shown in **Supplementary Fig. 3c** and correspond to three different crystal structures obtained under different conditions.

The scattering profiles of all members of the pool were computed with the program CRYSOL using standard parameters[10]. The pool of structures/profiles was submitted to the genetic algorithm (GA) with the aim to select a subensemble of 50 structures/curves with the capacity to collectively describe the experimental SAXS curves. The GA was repeated 200 times starting from different randomly selected subensembles, and the resulting solutions were analysed together. The same pool was used to describe all the measured data. Results were quantitatively analysed by displaying the R_g distributions of the selected ensembles and the relative population of monomeric and dimeric species.

Three different dimeric arrangements were introduced individually in the pool. In all three cases the dimers display similar contacts between the lectin domains and the lectin domain of one monomer with the catalytic domain of a second monomer (**Supplementary Fig. 3c**). We can distinguish the crystallographic dimers by the buried area parameter inferred from the PISA server. The dimers present in the crystal structure of GalNAc-T2-UDP complex (PDB entry 2FFV) are the most compact ones with a buried area of 1689 $Å^2$ followed by dimers from the PDB entry 4D0T with 1400 $Å^2$ of buried area, and finally dimers from the crystal structures with the tetragonal space group that in turn appear to be the most flexible compact structures of all dimers with ≈ 750 $Å^2$ of buried area. Taking into account these dimers for the

SAXS analysis, it was clear that the most rigid dimeric form belonging to the PDB entry 2FFV fitted better to the data (c2 = 1.3 versus values > of 1.5 for other dimeric structures).

Theoretical modeling of the SAXS results

To reproduce the SAXS distribution of the radius of gyration, we consider the crystal structure of GalNAc-T2-UDP-MUC5AC-13 complex, and define the linker corresponding to residues 437-449 (where A and B correspond to the alpha-carbon atoms of residues 437 and 449, respectively; **Fig. 5a**).

To generate the model, we made the following assumptions:

1) The catalytic (residues 75-436) and lectin domains (449-569) are rigid bodies. For the sake of simplicity and to avoid introducing three more variables corresponding to the Euler angles of rotation, their relative orientation is fixed to that found in the crystal structure of GalNAc-T2-UDP-MUC5AC-13 complex.

2) The linker residues are "phantoms", meaning that neither contribute to the radius of gyration, nor interact with each other or with the rest of the protein.

3) The linker is described as a Worm-Like-Chain (WLC)[11], which is a classical model of polymer physics for semi-flexible polymers. In WLC a polymer is represented as a continuous curve, with just an elastic energy, opposed to bending. At any temperature, its equilibrium configurations are the result of a trade-off between the elastic energy that favours stretched and loopy conformations, and entropy. Thus the

behavior of the WLC can be characterized by specifying just the contour length l_c and the persistence length l_p. The latter is a measure of the rigidity of the curve: roughly, two portions of the chains which are more than $2l_p$ away can have any relative orientation, while below that separation, their orientations will be correlated.

4) According to the coarse-grained spirit of the model, we describe the interaction between the two domains by simply adding a sigmoidal potential $E(r)$ that depends on the distance between the ends of the linker $r = |r| = |B - A|$ (**Fig. 5a**):

$$E(r) = -\epsilon \left(1 + exp\left[\frac{1}{r_w}(r - r_m)\right]\right)^{-1}.$$

Such potential is $-\epsilon$ at r=0; it halves itself at $r = r_m$ and then goes to 0 for increasing r; the smaller r_w, the sharper the transition. Here, r_m and r_w are regarded as adjustable parameters. We verified that the results, at the qualitative level, do not depend strongly on the choice of the potential.

5) The WLC is an isotropic model: setting an end of the polymer at the origin of the reference frame, there is no intrinsically angular preference for the position of the other end, that will be spherically distributed around the origin. We limit the angular freedom in our model, by restricting the second end of the linker within a cone "in front of" the catalytic domain (in a sense specified below), thus preventing configurations where the lectin domain is situated opposite to the catalytic one (**Fig. 5a**).

In this framework, the radius of gyration $R_g(X)$ of any protein conformation X will become a function of the distance vector r. Indeed, denoting $r_i as$ the position of atom

"i" of the protein in conformation X, we can write the following equation that neglects the contributions from the linker residues:

$$R_g^2(X) = \frac{1}{2N^2}\sum_{i,j=k_1}^{k_L}\left(r_j - r_i\right)^2 = \frac{N_A^2}{N^2}R_{g,A}^2 + \frac{N_B^2}{N^2}R_{g,B}^2 + \frac{1}{N^2}\sum_{i=k_1}^{k_A-1}\sum_{j=k_B+1}^{k_L}\left(r_j - r_i\right)^{2,}$$

where k_1,k_A,k_B,$k_L$are the indices of the first atom of the protein, first atom of the linker, last atom of the linker, and last atom of the protein, respectively, and

$$R_{g,A}^2 = \frac{1}{2N_A^2}\sum_{i,j=k_1}^{k_A-1}\left(r_j - r_i\right)^2,$$

$$R_{g,B}^2 = \frac{1}{2N_B^2}\sum_{i,j=k_B+1}^{k_L}\left(r_j - r_i\right)^2,$$

are the radii of gyration of just the catalytic (A) domain and lectin (B) domain, respectively; here N_A , N_B are the number of atoms in domain A and B, and $N = N_A + N_B$.

Upon introducing the coordinates of the centers of the two domains as $C_1 = \frac{1}{N_A}\sum_{i=k_1}^{k_A-1} r_i$and $C_2 = \frac{1}{N_B}\sum_{i=k_B+1}^{k_L} r_i$, and the distances of the ends of the linker from the centers: $x_A = A - C_1$,$x_B = B - C_2$, we can write:

$$r_i = \begin{cases} A + d_i = C_1 + x_A + d_i \; if \; i = k_{1,} \ldots, k_A - 1 \\ B + \delta_i = C_2 + x_B + \delta_i \; if \; i = k_B + 1, \ldots, k_L \end{cases}$$

Substituting the above equations in the expression for the radius of gyration, we can relate the latter to the end-to-end vector of the linker, $r = B - A$, as

$$R_g^2(r) = \Pi + \frac{N_A N_B}{N^2}(r^2 + 2rR),$$

where $R = x_A - x_B$and

$$\Pi = \frac{1}{N^2}\left(N_A^2 R_{g,A}^2 + N_B^2 R_{g,B}^2 + N_A N_B \sum_{i=k_1}^{k_A-1}\sum_{j=k_B+1}^{k_L}(\delta_j - d_i)^2\right),$$

is independent from r. The above expression for the radius of gyration allows to relate the probability distribution for the latter, as revealed by SAXS, to the probability density of the end-to-end vector of the linker. The latter will be given by

$$f(r|lc, lp) = \frac{1}{Z(lc, lp)} Q^{WLC}\left(\frac{r}{lc}, \frac{lp}{lc}\right) exp\left(-\frac{E(r)}{RT}\right) Y(\vartheta, \phi)$$

where $Q^{WLC}\left(\frac{r}{lc}, \frac{lp}{lc}\right)$is the distribution for the end-to-end distance in WLC as reported in Eq. 21 by Becker *et al*[12]. The exponential represents the Boltzmann factor due to the interaction and $Y(\vartheta, \phi)$ is the angular probability distribution, that we add to the isotropic WLC as discussed in item 5 above.

In the spirit of the coarse-grained model, and for the sake of simplifying the calculations, we use spherical coordinates, selecting R as the polar direction for the angle ϑ (we have verified that the angle between R and the $C_2 - C_1$axis is reasonably small, around 18 degrees), and keep isotropy for rotations of an angle ϕaround the polar axis, so that

$$Y(\vartheta, \phi) = \begin{cases} 1 \; if \vartheta < \vartheta_0 \\ 0 \; if \vartheta \geq \vartheta_0 \end{cases}$$

Finally, $Z(lc, lp) = \int r^2 \, sin(\vartheta) dr d\vartheta d\phi f(r|lc, lp)$ is the partition function, that ensure the proper normalization of the probability.

The probability distribution of the radius of gyration can be derived from $f(r|l, lc)$by performing a change of variables

$$r = (r, \vartheta, \phi) \rightarrow y = \left(y = R_g(r, \vartheta), \vartheta' = \vartheta, \phi' = \phi\right)$$

and imposing that the density-function in the new variables describe the same probability of finding the end-to-end vector within a certain volume element dV:

$$g(y|lc, lp)dV'(y, \vartheta', \phi') = f(r|lc, lp)dV(r, \vartheta, \phi),$$

whence

$$g(y|lc, lp) \equiv g(y, \vartheta'|lc, lp) = f(r(y, \vartheta), \vartheta|lc, lp)r^2(y, \vartheta)\left|\frac{\partial(r, \vartheta, \phi)}{\partial(y, \vartheta', \phi')}\right|,$$

with $\left|\frac{\partial(r,\vartheta,\phi)}{\partial(y,\vartheta',\phi')}\right|$the Jacobian determinant of the change of variables. Finally we get:

$$g(y, \vartheta'|l, lc) = f(r(y, \vartheta), \vartheta|l, lc)\frac{N^2}{N_A N_B y}\frac{r^2(y, \vartheta)}{\left(r(y, \vartheta) + Rcos(\vartheta)\right)},$$

The population of the radius of gyration observed in the experiments will be given by:

$$G\left(R_g\right) = 2\pi R_g^2 \int sen\,(\vartheta')d\vartheta' g\left(y = R_g, \vartheta'\middle|lc, lp\right).$$

Modeling the enzymatic activity

To the hypotheses above, we add another assumption:

6) To consider the enzymatic activity, we introduce a ligand peptide bound at P_2, again described as a WLC.

To this end, we consider points P_1, P_2 (**Fig. 5a**) that are located in the active site and the α subdomain of the lectin domain, respectively, and correspond to the position of the alpha carbons of the acceptor and prior glycosylated sites.

We then let the protein domains and ligand, move freely, with the only constraint that the ligand is always bound to the lectin domain at P_2. The glycosylation reaction at P_1 will take place when the linker and ligand conformation are such that the acceptor site of the ligand is found at its "active site" position P_1 in the catalytic domain.

The rate at which the reaction takes place will depend on the dynamics of the ligand and protein, as well as on the distance (in residues) between the prior glycosylated and the acceptor sites of the ligand.

We thus estimate the enzymatic activity as a function of the prior glycosylated and the acceptor sites distance that relies on the product of the equilibrium probability for the linker and the ligand.

Upon introducing, as above, $d_{P_1} = P_1 - A$,$d_{P_2} = P_2 - B$, and remembering that the linker end-to-end distance is $r = B - A$, we start by defining:

$$\rho_G(r) = P_1 - P_2 = d_{P_1} - \delta_{P_2} - r$$

which is the "correct" distance vector between the prior glycosylated and the acceptor sites on the ligand for the reaction to take place; such vector depends on the protein conformation through the linker end-to-end vector r. For a ligand peptide with the the prior glycosylated and the acceptor sites separated by l residues, the probability that the glycosylation residues are found in a volume dV_ρ around a distance vector ρ, when the protein linker end-to-end distance vector is found in volume dV_r around r, is:

$$dp = \xi(r, \rho|l, lc) dV_\rho dV_r,$$

For the sake of simplicity, we will assume that the density $\xi(r, \rho|l, lc)$ is simply the product of the separate distributions for the linker and ligand, as if they were

statistically independent, and the conformation of the former wouldn't affect the variability of the latter. Moreover, we ignore any angular bias on ρ, as if the ligand's free end were completely free to move in any direction, and assume a WLC behavior for the ligand, too.

This yields $\xi(r,\rho|l,lc) = f^{WLC}(\rho|l,\lambda)f(r|lc,lp)$, where the latter factor is the equilibrium probability for the linker, and the former is the isotropic WLC expression for the ligand, assuming a persistence length λ for the latter:

$f^{WLC}(\rho|l,\lambda) = \frac{Q^{WLC}(\rho/l,\lambda/l)}{4\pi\int d\rho\rho^2 Q^{WLC}(\rho/l,\lambda/l)}$.

Finally, the probability that, for a given length lc of the linker and a separation l between the prior glycosylated and the acceptor sites in the ligand, the latter meet the correct sites on the protein, for any protein conformation, is found by imposing $\rho = \rho_G(r)$ and integrating over the linker conformations:

$$\sigma(l,lc) = \int d^3\, r\xi(r,\rho = \rho_G(r)|l,lc)$$

To dissect the role of the linker and the ligand flexibility in determining the activity, we consider the extreme case of the linker completely frozen in its crystallographic conformation. In this case, the above equation for $\sigma(l,\lambda)$ is replaced by:

$$\sigma(l) = f^{WLC}(\rho = \rho_G(r_0)|l,\lambda)$$

and depends on the separation "l"of the prior glycosylated and the acceptor sites of the peptide (in the above equation $r_0$is the end-to-end distance of the linker in the crystal structure).

Supplementary References

1 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, doi:10.1074/jbc.M803387200 (2008).

2 Aydillo, C. *et al.* S-Michael additions to chiral dehydroalanines as an entry to glycosylated cysteines and a sulfa-Tn antigen mimic. *J Am Chem Soc* **136**, 789-800, doi:10.1021/ja411522f (2014).

3 Kirschner, K. N. *et al.* GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J Comput Chem* **29**, 622-655, doi:10.1002/jcc.20820 (2008).

4 Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc Natl Acad Sci U S A* **99**, 12562-12566, doi:10.1073/pnas.202427399 (2002).

5 Marcuello, C., de Miguel, R., Gomez-Moreno, C., Martinez-Julvez, M. & Lostao, A. An efficient method for enzyme immobilization evidenced by atomic force microscopy. *Protein Eng Des Sel* **25**, 715-723, doi:10.1093/protein/gzs086 (2012).

6 Horcas, I. *et al.* WSXM: a software for scanning probe microscopy and a tool for nanotechnology. *Rev Sci Instrum* **78**, 013705, doi:10.1063/1.2432410 (2007).

7 Meinander, K., Jensen, T. N., Simonsen, S. B., Helveg, S. & Lauritsen, J. V. Quantification of tip-broadening in non-contact atomic force microscopy with carbon nanotube tips. *Nanotechnology* **23**, 405705, doi:10.1088/0957-4484/23/40/405705 (2012).

8 M. V. Petoukhov, D. F., A. V. Shkumatov, G. Tria, A. G. Kikhney, M. Gajda, C. Gorba, H. D. T. Mertens, P. V. Konarev and D. I. Svergun. New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Cryst.* **45**, 9 (2012).

9 Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* **129**, 5656-5664, doi:10.1021/ja069124n (2007).

10 Svergun D.I., B. C. a. K. M. H. J. CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates *J. Appl. Cryst.* **28**, 6 (1995).

11 O. Kratky, G. P. Röntgenuntersuchung gelöster Fadenmoleküle. *Rec. Trav. Chim.* **68**, 18 (1949).

12 Becker, N. B., Rosa, A. & Everaers, R. The radial distribution function of worm-like chains. *Eur Phys J E Soft Matter* **32**, 53-69, doi:10.1140/epje/i2010-10596-0 (2010).

ARTÍCULO II

The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences.

El presente artículo fue publicado en la revista *Nature Communications* el año 2017, tercer año del periodo de tesis de la doctoranda. Fue realizado en coautoría con: (a) el grupo del Doctor Francisco Corzana, del Departamento de Química Orgánica de la Universidad de La Rioja; (b) el grupo de investigación del Profesor Henrik Clausen, en la Universidad de Copenhage; (c) el Departamento de Química de la Universidad de Nova de Lisboa, Portugal; (d) el Centro de Investigaciones Científicas bioGUNE, en Bilbao y (f) el Departamento de Química de la Universidad de Case Western, en Ohio.

A fecha de defensa de la presente tesis doctoral no se encuentran en disposición del título de Doctor los siguientes coautores: Ismael Compañón, del Departamento de Química Orgánica de la Universidad de la Rioja; Helena Coelho y Ana Díniz, del Departamento de Química de la Universidad de Nova de Lisboa y el Centro de Investigaciones Científicas bioGUNE, en Bilbao; y Earnest James Paul Daniel, del Departamento de Química de la Univseridad de Case Western, en Ohio. Se adjuntan sus correspondientes renuncias a presentar este artículo en la modalidad de compendio de pubilcaciones en el Apéndice 10.6 de la presente tesis.

La contribución del doctorando en esta publicación ha consistido en lo siguiente:

- Participación en el diseño experimental y en la escritura final del artículo.
- Clonaje de las diversas construcciones, quimeras y mutantes.
- Expresión y purificación de las proteínas implicadas.
- Cristalización de las proteínas y resolución de su estructura.

Se adjunta a continuación la referencia completa del artículo:

de las Rivas, M., Lira-Navarrete, E., James Paul Daniel, E., Compañón, I., Coelho, H., Diniz, A., Jiménez-Barbero, J., Peregrina, J. M., Clausen, H., Corzana, F., Marcelo, F., Jiménez-Osés, F., Gerken, . A. & Hurtado-Guerrero, R. (**2017**). The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. *Nature Communications, 8(1959),* DOI: 10.1038/s41467-017-02006-0.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-02006-0 OPEN

The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences

Matilde de las Rivas[1], Erandi Lira-Navarrete[1,2], Earnest James Paul Daniel[3], Ismael Compañón[4], Helena Coelho[5,6,7], Ana Diniz[5], Jesús Jiménez-Barbero[6,7,8], Jesús M. Peregrina[4], Henrik Clausen[2], Francisco Corzana[4], Filipa Marcelo[5], Gonzalo Jiménez-Osés[4], Thomas A. Gerken[3,9,10] & Ramon Hurtado-Guerrero[1,11]

The polypeptide GalNAc-transferases (GalNAc-Ts), that initiate mucin-type *O*-glycosylation, consist of a catalytic and a lectin domain connected by a flexible linker. In addition to recognizing polypeptide sequence, the GalNAc-Ts exhibit unique long-range N- and/or C-terminal prior glycosylation (GalNAc-*O*-Ser/Thr) preferences modulated by the lectin domain. Here we report studies on GalNAc-T4 that reveal the origins of its unique N-terminal long-range glycopeptide specificity, which is the opposite of GalNAc-T2. The GalNAc-T4 structure bound to a monoglycopeptide shows that the GalNAc-binding site of its lectin domain is rotated relative to the homologous GalNAc-T2 structure, explaining their different long-range preferences. Kinetics and molecular dynamics simulations on several GalNAc-T2 flexible linker constructs show altered remote prior glycosylation preferences, confirming that the flexible linker dictates the rotation of the lectin domain, thus modulating the GalNAc-Ts' long-range preferences. This work for the first time provides the structural basis for the different remote prior glycosylation preferences of the GalNAc-Ts.

[1] BIFI, University of Zaragoza, BIFI-IQFR (CSIC) Joint Unit, Mariano Esquillor s/n, Campus Rio Ebro, Edificio I+D, Zaragoza, 50018, Spain. [2] Copenhagen Center for Glycomics, Department of Cellular and Molecular Medicine, School of Dentistry, University of Copenhagen, Copenhagen, DK-2200, Denmark. [3] Department of Biochemistry, Case Western Reserve University, Cleveland, 44106 OH, USA. [4] Departamento de Química, Universidad de La Rioja, Centro de Investigación en Síntesis Química, E-26006 Logroño, Spain. [5] UCIBIO, REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade de Nova de Lisboa, Caparica, 2829-516, Portugal. [6] CIC bioGUNE, Bizkaia Technology Park, Building 801A, 48170 Derio, Spain. [7] Departament of Organic Chemistry II, Faculty of Science & Technology, University of the Basque Country, Leioa, Bizkaia 48940, Spain. [8] Ikerbasque, Basque Foundation for Science, Maria Diaz de Haro 13, 48009 Bilbao, Spain. [9] Department of Pediatrics, Case Western Reserve University, Cleveland, 44106 OH, USA. [10] Department of Chemistry, Case Western Reserve University, Cleveland, 44106 OH, USA. [11] Fundación ARAID, 50018 Zaragoza, Spain. Matilde de las Rivas, Erandi Lira-Navarrete and Earnest James Paul Daniel contributed equally to this work. Correspondence and requests for materials should be addressed to R.H-G. (email: rhurtado@bifi.es)

Polypeptide *N*-acetylgalactosaminyltransferases (GalNAc-Ts) are higher eukaryotic-retaining glycosyltransferases (GTs) that transfer a GalNAc moiety from uridine diphosphate *N*-acetylgalactosamine (UDP-GalNAc) onto Ser/Thr residues of proteins[1]. This family of enzymes, which initiates the most common type of *O*-glycosylation in metazoans[1–5], plays profound roles for the structure, stability, and function of proteins[6]. GalNAc-*O*-glycosylation is one of the most abundant types of protein glycosylation in eukaryotes that is chiefly found in the densely clustered heavily glycosylated domains of mucins and mucin-domain-containing glycoproteins, hence the name, mucin-type *O*-glycosylation. However, it is also clear that many other proteins contain isolated sites of GalNAc-*O*-glycosylation, where more than 5000 human proteins trafficking the secretory pathway have been identified to date containing one or more mucin-type *O*-glycans[7].

The initiation of mucin-type *O*-glycosylation by the GalNAc-Ts is one of the most complex regulated types of protein glycosylation. The GalNAc-Ts comprise a family of up to 20 isoenzymes possessing a range of kinetic properties and expression patterns. These enzymes orchestrate with high fidelity, the initial patterns of *O*-glycosylation on diverse proteins, including the high-density regions in mucins, where 30–50% of the amino acids may be glycosylated[1, 8]. The GalNAc-Ts are unique among metazoan GT enzymes because in addition to their N-terminal catalytic domain adopting a GT-A fold (family 27 in the CAZy database[9]), they possess a C-terminal GalNAc-binding lectin domain with a β-trefoil fold[4, 10–13] (classified as a carbohydrate-binding module 13 in the CAZy database), which provides additional functions to these enzymes[14] (see below for further details). Both domains are linked through a short flexible linker whose motion has been suggested to be responsible for the dynamic conformational landscape of these enzymes[4].

GalNAc-Ts can be classified according to their different glycosylation capacity on peptide and glycopeptide substrates. Three distinctive glycosylation modes have been reported: catalytic domain-dependent glycosylation on “naked” peptides (Fig. 1a) and on glycopeptides (Fig. 1b), hereafter termed short-range glycosylation[15, 16]; and remote lectin domain-dependent glycosylation on glycopeptides (Fig. 1c), termed long-range glycosylation[15, 16]. Adjacent and proximal sites relative to an existing GalNAc glycosite are glycosylated in a catalytic domain-dependent manner (Fig. 1), with GalNAc-T4 being part of a small number of GalNAc-Ts (including GalNAc-T7 and T10) that will also glycosylate contiguous sites[16]. Distant sites can be glycosylated in a lectin domain-dependent manner (Fig. 1) with individual GalNAc-Ts having distinct preferences on remote sites located N or C terminus from a prior GalNAc glycosite[15, 16]. For example, GalNAc-T1/T2/T14 tend to glycosylate preferentially N-terminal sites remote from prior C-terminal GalNAc glycosites in glycopeptides, whereas GalNAc-T3/T4/T6/T12 exhibit the opposite preference (Fig. 1c). Furthermore, some GalNAc-Ts, including GalNAc-T5/T13/T16, appear to exhibit both orientational preferences for long-range glycosylation[16]. Previous structural work with GalNAc-T2 in complex with “naked” peptides and glycopeptides has provided insight into the catalytic domain-dependent reaction, and also demonstrated how the GalNAc-T2 lectin domain guides catalysis to N-terminal acceptor sites very distant from a prior C-terminal GalNAc glycosite[4, 12, 17]. However, the molecular basis of how other GalNAc-Ts, such as GalNAc-T3/T4/T6/T12, achieve the opposite long-range glycosylation preferences remains unclear[15, 16].

The GalNAc-T4 isoform is of particular interest as it is the only isoform capable of glycosylating two out of the five acceptor sites in the partially glycosylated MUC1 mucin tandem repeat in a lectin domain-dependent manner[18]. The density of glycosylation together with the structure of *O*-glycans in the mucin repeat regions are important features for the cancer-glycoforms of MUC1 being exploited for diagnostic as well as therapeutic purposes[19]. Generally, GalNAc-T4 glycosylates very few isolated glycosylation sites in “naked” peptide acceptors, although it readily glycosylates the important Thr57 in PSGL-1 that is the essential *O*-glycan required for P-selectin-mediated leukocyte trafficking[20, 21]. We report herein a multidisciplinary approach on GalNAc-T4, combined with the characterization of different chimeras and mutants of GalNAc-T2, which begins to reveal the molecular basis of the long-range glycosylation preferences of the GalNAc-Ts on glycopeptide acceptors substrates.

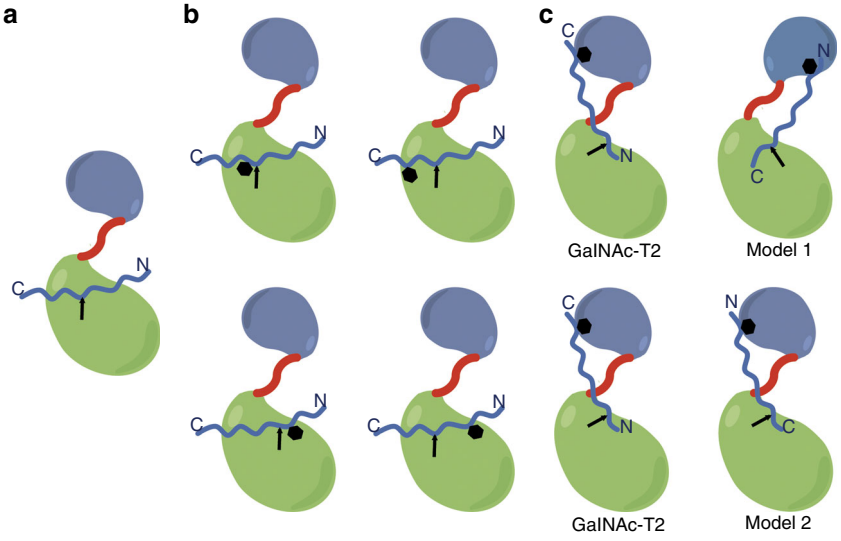


Fig. 1 Modes of *O*-glycosylation found for the GalNAc-Ts. **a**–**c** These panels denote the three distinct modes of *O*-glycosylation performed by GalNAc-Ts: the neighboring glycosylation activity by the catalytic domain in **a** and **b**, and long-range lectin domain-mediated glycosylation activity in **c**. In **c**, two plausible models are suggested that might explain the different long-distance glycosylation preferences of GTs such as GalNAc-T3/T4/T16/T12. Oval-shaped figures in blue and green depict the lectin and catalytic domains, respectively. Peptides are indicated in blue and the black hexagon-shaped figure denotes the position of prior GalNAc moieties in the glycopeptides. Arrows indicate the positions of acceptor sites

Results

Models explaining the long-range glycosylation preferences. Based on previous results, we can infer two possible models (see Fig. 1c) that might provide a plausible explanation for the GalNAc-T3/T4/T6/T12 long-range glycosylation preference in contrast to the one found for enzymes such as GalNAc-T1/T2/T14[15, 16]. In model 1, the rotation of the lectin domain GalNAc-binding site is required, thereby maintaining the same orientation of the peptide on the catalytic domain. In model 2, the lectin domain GalNAc-binding site would adopt the same orientation as found in GalNAc-T2 while the glycopeptide adopts an inverse orientation on the catalytic domain, an assumption unlikely to happen providing the high similarity of the peptide-binding groove among GalNAc-Ts[11] (Fig. 1).

GalNAc-T4-substrate interaction and kinetics studies. To test the two models proposed above, three simplified peptides enriched in glycine and alanine residues were designed, synthetized, and evaluated as substrates of GalNAc-T4 (Table 1). One of these peptides was a "naked" peptide, **1** (denoted –TT– for simplicity), whereas the other two were monoglycopeptides, **2** and **3** (denoted -TT–T*- and -T*–TT-, where T* represents a GalNAc glycosylated Thr), that contain a GalNAc moiety located either at the C or N terminus, respectively. All peptides have two potential Thr acceptor sites and also contain an adjacent PXP motif, which is recognized by most GalNAc-Ts[4, 16]. Our kinetic studies showed that GalNAc-T4 glycosylated selectively peptide **3** (with a high affinity, K_m ~ 40 μM, and catalytic efficiency, ~75 (nmole min)$^{-1}$) over peptides **1** and **2**, both of which were imperceptibly glycosylated over a range of peptide concentrations (Fig. 2a). This is consistent with previous findings[16] on GalNAc-T4 and illustrates that the lectin domain guides GalNAc-T4 to preferentially glycosylate C-terminal sites of monoglycopeptides that are remote from an N terminus GalNAc-glycosylated residue. Although, the "naked" peptide **1** was not a substrate, GalNAc-T4 has been reported to readily glycosylate other highly specific peptide sequences[20, 21].

Table 1 Peptide acceptor substrates used in this study

Peptide	Sequence
1 (-TT-)	GAGAGAGTTPGPG
2 (-TT-T*-)	AGAGTTPGPGAGAT*GA
3 (-T*-TT-)	GAT*GAGAGAGTTPGPG

T* denotes the Thr-O-GalNAc moiety

To understand why GalNAc-T4-only glycosylated monoglycopeptide **3**, we performed binding studies using saturation-transfer difference (STD) NMR experiments on peptide substrates **1–3** and α-methyl-GalNAc in the presence of UDP and $MnCl_2$ (Fig. 2b and Supplementary Fig. 1). The absence of STD-NMR signals for peptide **1** revealed that the "naked" peptide was very poorly recognized by GalNAc-T4 in agreement with the low glycosylation observed during the kinetics assays. For both monoglycopeptides **2** and **3** relatively weak STD-NMR signals were observed for the Ala and Thr methyl protons, suggesting weak recognition of the peptide by GalNAc-T4. However, the *O*-GalNAc glycosylated Thr of the monoglycopeptides **2** and **3** receive a higher STD enhancement. Remarkably, the sugar protons of the GalNAc moiety of the monoglycopeptides **2** and **3** (region 3.5 a 4.2 ppm of the spectrum together with the singlet at 2 ppm) showed unambiguous and more intense STD-NMR signals (Fig. 2b and Supplementary Fig. 1). In addition, the STD-NMR-derived epitope mapping of the GalNAc moiety is identical for both monoglycopeptide **2** and control α-methyl-GalNAc. In fact, for both compounds, the highest STD response corresponded to the H2 proton of the GalNAc moiety followed by a modest STD of protons H4, H3, and N-acetyl group, suggesting similar lectin-binding modes for these molecules. In contrast, monoglycopeptide **3** presents a somewhat different GalNAc STD-NMR-binding pattern, where H2, H4, and the N-acetyl group of the GalNAc moiety have the highest STD-NMR response closely followed by the H3 proton. This suggests that the monoglycopeptide **3** has a distinct binding mode of interaction with the enzyme, which is different to the one found for monoglycopeptide **2** and α-methyl-GalNAc control. From our kinetic and STD-NMR results, we can infer that the unique binding of the GalNAc residue of monoglycopeptide **3** to the lectin domain, facilitates the correct orientation of the peptide acceptor onto the catalytic domain for optimal C-terminal glycosylation. Thus, even though monoglycopeptide **2** shows GalNAc binding to the lectin domain, its N-terminal acceptor site must be directed away from the catalytic domain, leading to the inability of GalNAc-T4 to glycosylate this substrate. Hence, the correct location of a prior glycosite in a monoglycopeptide, with respect to potential acceptor sites, is essential for an optimal interaction of the peptide within the catalytic domain, which in turn is critical for an efficient catalysis and specificity.

Crystal structure of GalNAc-T4. To further elucidate the molecular basis of GalNAc-T4's monoglycopeptide glycosylation preference and rule out one of the two proposed models (Fig. 1),

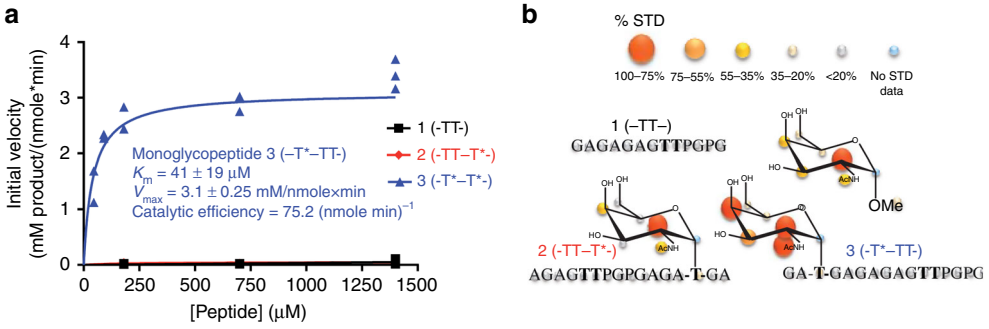


Fig. 2 Biophysical characterization of GalNAc-T4 against peptides 1–3. **a** Peptide glycosylation kinetics of GalNAc-T4 against (glyco)peptides **1–3** (black, red, and blue symbols, respectively). Michaelis–Menten kinetic values, K_m, V_{max}, and catalytic efficiency (V_{max}/K_m) for monoglycopeptide **3** were obtained from the nonlinear least square fit to the initial rate data, obtained as described in the "Methods" section and given in Supplementary Table 2. Peptide substrates **1** and **2** are largely unglycosylated by GalNAc-T4. **b** STD-NMR-derived epitope mapping. The different colored spheres indicate the normalized STD signal (in percent) observed for each proton. See "Methods" section and Supplementary Fig. 1 for further details

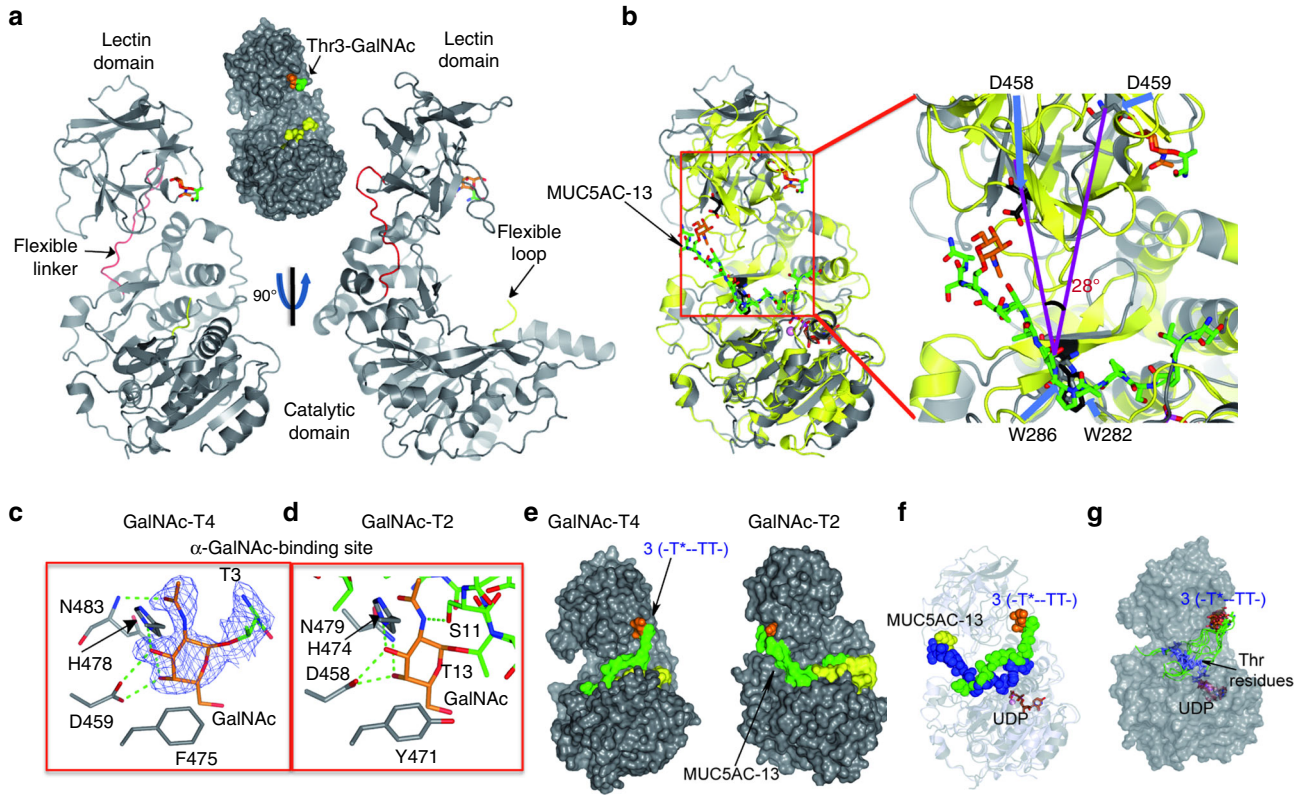


Fig. 3 Structural characterization of the interaction between GalNAc-T4 and monoglycopeptide 3. **a** Two different views of GalNAc-T4 in complex with **3**. The catalytic and lectin domains are colored in gray and the interdomain flexible linker is depicted in red. The catalytic domain flexible loop is depicted in yellow and is mostly disordered. The GlcNAc moiety of the Thr3-GalNAc is shown as orange carbon atoms while Thr3 is shown as green carbon atoms. **b** (left) Superposition of the GalNAc-T4-glycopeptide **3** (gray) and GalNAc-T2-MUC5AC-3-UDP-Mn^{+2} (yellow) structures. The glycopeptide and the GalNAc moiety are shown in green and orange carbon atoms, respectively. The lectin α-subdomain GalNAc-binding residues, Asp458, and Asp459, of GalNAc-T2 and GalNAc-T4 are shown as sticks in black and gray carbon atoms, respectively. (right panel) Close-up view of the superposition between GalNAc-T2 and GalNAc-T4. Colors for peptides, GalNAc moiety, and proteins are identical as shown above. **c**–**d** Close-up view of the lectin α-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 F_O–F_C (blue) contoured at 2.2 σ for Thr3-GalNAc. Note that both GalNAc-binding sites are depicted in the same orientation for comparison. **e** Surface representation of GalNAc-T4 (model built with monoglycopeptide **3** and UDP/Mn^{+2}), and GalNAc-T2 (with UDP/Mn^{+2}/MUC5AC-13). Both are viewed from the same orientation as in **b**. Colors for the glycopeptide and the flexible loop are the same as above. **f** Superposition of the GalNAc-T4-glycopeptide **3**-UDP-Mn^{+2} (gray) and GalNAc-T2-MUC5AC-3-UDP-Mn^{+2} (blue-white) structures. The glycopeptide and the GalNAc moiety in the structure of GalNAc-T2 and GalNAc-T4 are shown as spheres in blue/green and yellow/orange, respectively. UDP and Mn^{+2} are shown as brown carbon atoms and as a pink sphere, respectively. **g** Superimposed structures of the 200 ns MD simulation trajectory for GalNAc-T4 in complex with glycopeptide **3**/UDP/Mn^{+2}. GalNAc-T4 is depicted in a light gray surface view. The GalNAc moiety, the peptide backbone, UDP, and the acceptor Thr residues are shown as orange, green, brown, and blue carbon atoms, respectively

we obtained triclinic crystals of GalNAc-T4 that were subsequently soaked with monoglycopeptide **3** and UDP and $MnCl_2$. The resulting crystal allowed us to solve the corresponding structure at a high resolution (1.90 Å) (Supplementary Table 1). Within the asymmetric unit, two independent GalNAc-T4 molecules were present. The crystal structure shows a compact structure with the typical GT-A fold and the lectin domain located at the N- and C-terminal regions, respectively (Fig. 3a). A sequence alignment between GalNAc-T2 and T4 displays a sequence identity of ~40% and a high resemblance between these two enzymes at the secondary structure level (Supplementary Fig. 2). In fact, superposition analysis of both enzymes shows that both catalytic domains are more structurally similar (root-mean-square deviation (RMSD) of 1.23 Å for aligned Cα atoms) than the lectin domains (RMSD of 1.82 Å) (Fig. 3b and Supplementary Fig. 3). A global superposition of both enzymes renders a larger RMSD of 2.12 Å, which suggests that a major shift has occurred between the orientation of the lectin domains (see below for a further discussion; Fig. 3b). As expected and accordingly to model 1, the GalNAc-binding site in GalNAc-T4 is located on the right side of the lectin domain, which requires a rotation of 28° with respect to the homologous site in GalNAc-T2 (Fig. 3b and Supplementary Fig. 4). To calculate the angle of rotation, the Cαs of Asp458, Trp282, and Asp459 were used. Note that the angle of rotation could be also determined using the Cαs of Asp458, Trp286, and Asp459. In this case, the rotation angle was 30°, which is very close to 28°. This is due to Trp282 of GalNAc-T2 and Trp286 of GalNAc-T4 adopt almost identical positions in the catalytic domain (Fig. 3b). This different position of the GalNAc-T4 GalNAc-binding site, consistent with the STD-NMR results, may explain why GalNAc-T4 promotes a distinct type of long-range glycosylation, consisting of the glycosylation of acceptor sites located at the C-terminal from prior GalNAc glycosites in glycopeptides. This also suggests that the location of the GalNAc-binding site of the lectin domain is coupled with the long-range glycosylation preferences of these enzymes.

Lectin domain-binding site of GalNAc-T4. Even though the crystals were soaked with a high concentration of peptide **3**, UDP

and $MnCl_2$, we could only visualize well-defined density for the GalNAc-*O*-Thr moiety of monoglycopeptide **3** bound to the lectin α-subdomain GalNAc-binding site (Fig. 3a, c). Note that GalNAc-Ts lectin domains contain three potential carbohydrate-binding subdomains named α, β, and γ; however, in most GalNAc-Ts only one of them is functional[15, 16, 18, 22, 23]. The almost absence of density for the peptide backbone of monoglycopeptide **3**, is consistent with the STD-NMR results (showing no significant STD with the acceptor peptide residues) and with the weak binding affinity of this peptide (K_d estimated in the mM range), determined by surface plasmon resonance (SPR) experiments in the presence of UDP and $MnCl_2$ (Supplementary Fig. 5). The discrepancies between GalNAc-T4's high-affinity K_m obtained from our kinetics studies (Fig. 2a) versus the poor K_d values from the binding studies can be attributed to the fact that the enzyme kinetics was performed with UDP-GalNAc. It is known that this donor substrate stabilizes the flexible loop of the catalytic domain active site in a closed conformation, thus completing the formation of the peptide-binding groove and leading to an active enzyme[4, 12]. Thus, our observations in the absence of UDP-GalNAc represent the weak binding of peptide substrate to the open flexible loop conformation of the enzyme.

A closer inspection at the lectin domain GalNAc-binding site depicts that the GalNAc moiety was tethered by conserved residues, including Asp459, Asn483, Phe475, and His478, of the lectin α-subdomain GalNAc-binding site (equivalent residues in GalNAc-T2 are Asp458, Asn479, Tyr471, and His474; Fig. 3d). It is important to note that only the lectin α-subdomain GalNAc-binding site has been shown to be functional in both GalNAc-T2 and GalNAc-T4 lectin domains[4, 18]. Most of the interactions are through hydrogen bonds in both enzymes, although they also share CH–π interactions between the GalNAc moiety and Phe475/Tyr471. The importance of the Asp residue in recognition of the GalNAc moiety and its role in the lectin domain-mediated glycosylation has been previously reported[4, 18]. Furthermore, the crystal structure explained the GalNAc STD-NMR-derived epitope mapping. In particular, the structure shows that H4 and H3 of the GalNAc moiety are in closer contact with Phe475 and His478, which is further supported by the STD response (92% for H4 and 72% for H3) observed for these protons in monoglycopeptide **3**.

Molecular docking and dynamics simulations. To address how the lectin domain guides the delivery of potential peptide acceptor sites to the catalytic domain, we performed molecular docking and dynamics (MD) simulations on GalNAc-T4 in complex with UDP/Mn^{+2} and monoglycopeptide **3** (see "Methods" section). The calculations suggest that although the peptide bound at the catalytic domain was highly dynamic, the two potential peptide acceptor Thr residues were in close contact to the β-phosphate of UDP during the simulated timeframe (200 ns), thereby accounting for the glycosylation of this peptide (Fig. 3e, f, g and Supplementary Fig. 6 and Supplementary Movie 1). A comparison with the GalNAc-T2 structure in complex with MUC5AC-13 and UDP/Mn^{+2} clearly provides evidence of the two different orientations of the lectin domains and supports our conclusion on the coupling of the location of the GalNAc-binding site with the observed long-distance glycosylation preference (Fig. 3e, f).

The flexible linker dictates the lectin domain rotation. Previous studies suggested that the flexible linker allowed for the inter-domain translational-like motion in these enzymes[4]. We therefore rationalized that different flexible linkers found in these GTs might also be behind the rotation capacity of the lectin domain (Supplementary Fig. 7). To test this hypothesis, we initially performed 500 ns MD simulations in water solution of native GalNAc-T2 and GalNAc-T4. Simulations clearly showed that the lectin domains of GalNAc-T2 and GalNAc-T4 *apo* enzymes do not rotate in the considered timescale (Supplementary Movies 2 and 3). This suggests that the orientation of the GalNAc-binding site with respect to the catalytic domain may remain relatively fixed in a fully folded transferase. We then performed MD simulations on **chimeras 1** and **2** that corresponded to GalNAc-T2 with two different lengths of the flexible linker from GalNAc-T3 (Table 2). In agreement with our hypothesis, we observed the prompt rotation of the lectin domain of the GalNAc-T2 **chimera 2** toward a position similar to that in GalNAc-T4 (Fig. 4a, Supplementary Fig. 8, and Supplementary Movie 4). This stepwise motion was completed in ~30 ns and involved sequential extension of the flexible linker, rotation of the lectin domain around the *Z* and *Y* axes, and finally compression of the linker (Fig. 4a). Thus, initially the GalNAc-T3 linker was artificially forced to adopt the compact conformation of the GalNAc-T2 native linker, but early on in the simulation (after ~10 ns), the linker quickly recovered its more native extended conformation and as a consequence, the lectin domain springs and rotates sidewise to finally re-assemble in a structure similar to that of our GalNAc-T4. This motion is likely smoothed by the absence of strong interdomain

Table 2 Sequences of the chimeras and mutant GalNAc-Ts used in this study

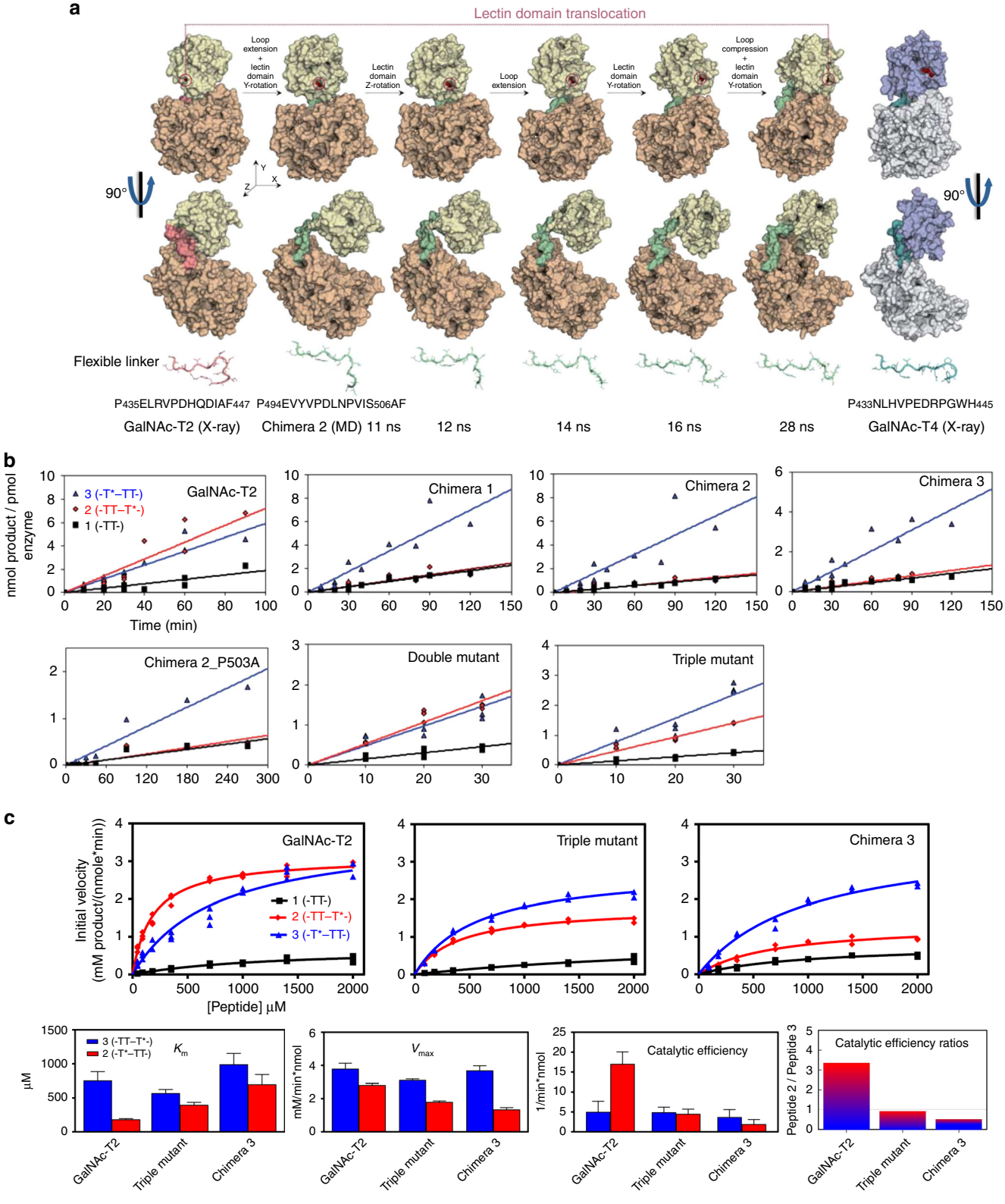
Transferase	Sequence of flexible linker[a]
GalNAc-T2	P_{435}ELRVPDHQDIAF$_{447}$
GalNAc-T2-t3$_{\text{Flexible linker}}$ (**Chimera 1**)	P_{494}EVYVPDLNPVIS$_{506}$
GalNAc-T2-t3$_{\text{Flexible linker-AF}}$ (**Chimera 2**)	P_{494}EVYVPDLNPVIS$_{506}$AF
GalNAc-T2-t4$_{\text{Flexible linker}}$ (**Chimera 3**)	P_{433}NLHVPEDRPGWH$_{445}$
GalNAc-T2-t3$_{\text{Flexible linker-AF-P503A}}$ (**Chimera 2_P503A**)	P_{494}EVYVPDLN**A**VIS$_{506}$AF
GalNAc-T2 (**double mutant**, **R438A-D444A**)	P_{435}EL**A**VPDHQ**A**IAF$_{447}$
GalNAc-T2 (**triple mutant**, **R438A-D444A-F447A**)	P_{435}EL**A**VPDHQ**A**IA**A**$_{447}$
GalNAc-T3	P_{494}EVYVPDLNPVIS$_{506}$
GalNAc-T4	P_{433}NLHVPEDRPGWH$_{445}$
GalNAc-T4-t2$_{\text{Flexible linker}}$ (**Chimera 4**)	P_{435}ELRVPDHQDI$_{445}$
GalNAc-T4-t2$_{\text{Flexible linker-AF}}$ (**Chimera 5**)	P_{435}ELRVPDHQDIAF$_{447}$
GalNAc-T4-t2$_{\text{Flexible linker-A}}$ (**Chimera 6**)	P_{435}ELRVPDHQDIA$_{446}$H

Above are the sequences of the native transferases and linker domain mutants and chimeras utilized in this work. Only the sequences of the linkers that are exchanged are shown. Shortened names in (bold) are referred to in the text and figures

[a]Note that some of the exchanged linker regions also include part of the succeeding β-strand of GalNAc-T2 and GalNAc-T4. This is due to the boundaries of flexible linkers are not well defined at the C terminus. The additional Pro in the flexible linker of GalNAc-T3/T4, that is not present in GalNAc-T2, is underlined

interactions in GalNAc-T2, which do take in place in GalNAc-T4 (see for instance the persistent salt bridge between Arg397 and Glu487 in Supplementary Movie 5). Thus, the intrinsic conformational preference(s) of a short, unfolded fragment of ~13 residues, was found to determine the relative orientation of two very large protein domains (see also Supplementary Movies 6 and 7; these movies show the intrinsic different conformational preferences of the GalNAc-T2 (folded) and GalNAc-T3 (extended) flexible linkers by performing 500 ns MD simulations of these isolated fragments).

Kinetics characterization of the chimeras. To validate our MD simulations predictions, we expressed and purified **chimeras 1–3** and performed a series of partial and full kinetic analysis against our peptide substrates (Fig. 4b, c) (also note that **chimeras 4–6**



could not be expressed in our hands, see "Methods" section and Table 2). Remarkably, the GalNAc-T2 chimeras rendered glycosylation preferences similar to GalNAc-T3 and GalNAc-T4[15, 16] (preferring monoglycopeptide **3** over monoglycopeptide **2**, by approximately fourfold, which is significantly different from native GalNAc-T2) despite comprising most of the GalNAc-T2 architecture (Fig. 4b, c and Supplementary Fig. 9). Interestingly, GalNAc-T2's preference for these substrates varied with peptide substrate concentration, ranging from a high preference for monoglycopeptide **2** at low substrate concentrations to nearly equal preferences for both monoglycopeptides **2** and **3** at higher substrate concentrations (Fig. 4c). These differences arise from monoglycopeptide **2**'s ~fourfold lower K_m compared to monoglycopeptide **3**, thus giving an over threefold increase in catalytic efficiency for monoglycopeptide **2** over monoglycopeptide **3** (Fig. 4c and Supplementary Table 2). These findings readily explain the previously reported differences in GalNAc-T2's long-range glycosylation preferences[15, 17]. This type of behavior was not observed in the kinetics of GalNAc-T4 or **chimera 3** (note that detailed kinetics were not performed on **chimera 1** or **2**). Further analysis revealed that the changes in specificity for **chimeras 1–3** were chiefly due to a decrease in the observed specific activity (**chimera 1** and **2**), and changes in V_{max} and K_m (**chimera 3**) against monoglycopeptide **2** compared to that of monoglycopeptide **3** whose specific activity and kinetic parameters remained relative constant (Fig. 4b, c, Supplementary Table 2, and Supplementary Fig. 9). It is also worth noting that the specific activities (and detailed kinetics when obtained) for **chimeras 1–3** toward the "naked" peptide **1** are quite similar to those of GalNAc-T2, indicating that the GalNAc-T3/T4 flexible linkers in these chimeras did not affect the overall architecture of the GalNAc-T2's active site (Fig. 4b, c, Supplementary Table 2, and Supplementary Fig. 9). Finally, we evaluated which Thr residues of the substrates were preferably glycosylated by Edman amino acid sequencing, finding that the Thr residue adjacent to the PXP motif was glycosylated by **chimeras 1–3**, as would be expected for the PXP motif[16] (Supplementary Fig. 10). Little to no glycosylation was observed at the proximal Thr.

Together, our results suggest that the glycosylation preferences found for the GalNAc-T2 chimeras can be reasonably explained by the rotation of the lectin domain as predicted by the MD simulations of **chimera 2**. Thus, the lectin domains of these chimeras must begin to adopt an equivalent position to that of GalNAc-T4 to account for their altered long-range glycosylation preferences. We conclude that the flexible linker modulates the rotation of the lectin domain, which subsequently determines the orientation of the functional GalNAc-binding site of the lectin domain with respect to the catalytic domain. This modulation of the lectin domain relative to the catalytic domain likely determines the distinct long-range glycosylation preferences for all of the GalNAc-T isoforms. Our studies below further examine the role of the linker in dictating these long-range preferences.

Site-directed mutagenesis of the GalNAc-T2 and chimera 2. Next, we explored whether specific particular residues within the flexible linkers might be responsible for the lectin domain orientation. A multiple alignment of the flexible linkers (Supplementary Fig. 7) shows that GalNAc-T3/T4/T6/T12, which share the same long-distance glycosylation preferences[15, 16], contain an additional Pro that is not present in GalNAc-Ts encompassing the opposed glycosylation preferences such as GalNAc-T1/T2/T14. This Pro in the flexible linkers of GalNAc-T3 ($\mathbf{P}_{494}$EVYV**P**DLN**P**VIS$_{506}$; underlined) and GalNAc-T4 ($\mathbf{P}_{433}$NLHV**P**EDR**P**GWH$_{445}$; underlined) is located at the end of the loop with respect to the flexible linker of GalNAc-T2 ($\mathbf{P}_{435}$ELRV**P**DHQDIAF$_{447}$). Considering that Pro residues are abundant in protein turns, we hypothesized that this additional Pro might be, at least in part, responsible of the different orientation found crystallographically for the GalNAc-T4 lectin domain, and computationally for **chimera 2**. To explore this hypothesis, we mutated this Pro (in the GalNAc-T3 flexible linker sequence of the GalNAc-T2 **chimera 2**) to Ala-producing **chimera 2_P503A** with the expectation that it would reverse its preference (Table 2). This chimera, however, showed an identical glycosylation preference as the starting **chimera 2** (Fig. 4b and Supplementary Fig 9), rather than reverting to the expected wild-type GalNA-T2 preference. This suggests that more complex events within the flexible linker might take place to explain the interplay between the flexible linker dynamics and its coupling with the lectin domain rotation. Also note that this mutated chimera displays a very low-specific activity compared to GalNAc-T2 (~3–10%) (Supplementary Fig. 9). This may suggest that the architecture of the catalytic site of this chimera may have been altered or perhaps blocked by an altered orientation of its lectin domain.

Next, we evaluated in detail the interactions between residues of the flexible linker in the compact and extended crystal structures of GalNAc-T2[4]. Two major interactions, namely a salt bridge between Arg438 and Asp444, and a CH–π interaction between Pro440 and Phe447, appeared to be important to maintain the compact, folded structure of its flexible linker (Supplementary Fig. 11). We reasoned that these interactions might be important for fixing the GalNAc-T2 lectin domain orientation with respect to its catalytic domain. To test this hypothesis, we generated the double mutant R438A-D444A and the triple mutant R438A-D444A-F447A (Table 2). Our assumption was that the disruption of these interactions might lead to more "flexible" linkers, which in turn could cause some degree of rotation of the lectin domain toward the orientation observed in GalNAc-T4. MD simulations showed that, whereas the lectin domain did not rotate in the appropriate direction in the double mutant R438A-D444A (Supplementary Movie 8), it did undergo complete rotation in the triple mutant R438A-D444A-F447A toward the orientation observed in **chimera 2** and GalNAc-T4, although such motion required a much longer simulation time (complete rotation observed after 160–450 ns, Supplementary Movie 9).

Fig. 4 Characterization of the GalNAc-T2 chimeras and mutants. **a** MD simulations in explicit water of **chimera 2** (500 ns total simulation time; water molecules were removed for clarity). The 28 ns time-lapse snapshots show the dynamic events occurring during the lectin domain reorientation observed for this chimera. 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 domains, 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. **b** Glycosylation time course plots of GalNAc-T2 and the GalNAc-T2 chimeras and mutants against (glyco)peptides **1**-**3** at substrate concentration of 1.4 mM. The obtained specific activities and selected substrate activity ratios are given in Supplementary Fig. 9. **c** Complete glycosylation kinetics (initial specific activity versus substrate concentration) for GalNAc-T2, the GalNAc-T2-triple mutant and the GalNAc-T2 **chimera 3** against substrate (glyco)peptides **1**-**3** (top panels) with plots of the obtained Michaelis–Menten kinetic parameters, K_m, V_{max}, catalytic efficiency (V_{max}/K_m), and catalytic efficiency ratios (monoglycopeptide **2** over **3**) (bottom panels). Kinetic parameter values are summarized in Supplementary Table 2

These results were validated by their observed substrate preferences (obtained for both mutants; Fig. 4b and Supplementary Fig. 9) and by a kinetic analysis of the triple mutant R438A-D444A-F447A (Fig. 4c and Supplementary Table 2). Whereas the double mutant gave nearly equal monoglycopeptide preferences as observed for GalNAc-T2 (at high substrate concentration), the triple mutant showed a 1.5–2-fold preference for monoglycopeptide **3** over monoglycopeptide **2**, resembling the preference of GalNAc-T4 (Fig. 4b,c, Supplementary Table 2, and Supplementary Fig. 9). As with **chimeras 1–3**, the change in preferences observed with the GalNAc-T2 triple mutant was due to a decrease in V_{max} and increase in K_m for monoglycopeptide **2** rather than due to a significant change in the kinetic properties of monogycopeptide **3** (Fig. 4b,c, Supplementary Table 2, and Supplementary Fig. 9). The specific activity (and kinetic constants, when obtained) of these linker mutants against peptide **1** are very similar to GalNAc-T2, as well as **chimeras 1–3**, further supporting the notion that their catalytic domains have not been significantly altered (Supplementary Table 2, Fig. 4b, c, and Supplementary Fig. 9). In summary, the stepwise change in preference (Fig. 4b, c) from monoglycopeptide **2** toward that of monoglycopeptide **3**, for native GalNAc-T2, the triple mutant and **chimera 3**, mostly reflects changes in the kinetic constants against monoglycopeptide **2** rather than those for monoglycopeptide **3**.

All together, these results strongly suggest that the flexible linker plays a significant role in directing the long-range glycopeptide specificity of this family of transferases by altering the relative orientations of the catalytic and lectin domains. In addition, our results suggest that specific interactions within the flexible linker are sufficient to modulate this orientation thereby dictating each isoforms long-range glycosylation preferences.

Discussion

Most types of protein *O*-glycosylation in higher eukaryotes (O-Man, O-Xyl, O-Glc, O-GlcNAc, and O-Fuc) are initiated by only one or two GT isoenzymes[1]. However, these types of protein *O*-glycosylation are only found on limited proteins and sequence motifs. On the contrary, Nature has imposed a large family of GalNAc-Ts isoforms to account for mucin-type *O*-glycosylation. Initially, these GTs were considered redundant isoenzymes but recent findings clearly demonstrate that individual GalNAc-T isoforms serve unique biological functions in health and disease. For example, deficiency in *GALNT3* and *GALNT*2 causes familial tumoral calcinosis[24] and dyslipidemia[25], respectively. How such transferase specificity may be regulated is best illustrated by GalNAc-T3, which is the only isoform capable of glycosylating Thr178 at a proprotein convertase (PC) processing site (RHTR^{179}↓) of FGF23, an important regulator of phosphate homeostasis[22]. However, glycosylation of Thr178 by GalNAc-T3 requires prior *O*-glycosylation of residue Thr171 of FGF23, demonstrating that GalNAc-T3 utilizes its long-range N-terminal preference for prior GalNAc glycosylation to target or amplify the glycosylation of Thr178[15]. Thus, the combination of local peptide sequence and remote prior glycosylation is used by these transferases to ensure the fidelity of isoform-specific glycosylation. With the recent quantitative differential *O*-glycoproteomics studies[7], it is now clear that many individual GalNAc-T isoforms serve unique non-redundant contributions to glycosylation of the proteome, presumably employing similar strategies as that of GalNAc-T3 and FGF23. The long-range lectin domain-mediated functions of GalNAc-Ts are likely instrumental for the coordinated glycosylation of additional diverse sequence motifs and in particularly high-density acceptor sequence motifs such as in mucins. Therefore, it is likely that the large number of GalNAc-Ts and their lectin domains evolved to recognize diverse peptide sequence motifs as well as to differentially recognize prior sites of GalNAc glycosylation. This would ensure the fidelity of glycosylation in both low- and high-density glycosylated regions by sequential orchestrated glycosylation reactions, coordinated by both the specificity of the catalytic domain as well as the long-range recognition of the lectin domain, that interacts with the prior attached GalNAc residue[15, 16]. Therefore, to fully understand the biological functions of these enzymes and to elucidate their roles in disease, we must understand the molecular basis, and mechanisms therein, which dictate and modulate their complex long- and short-range substrate specificity.

Here we have addressed the molecular mechanisms behind the unique long-range lectin domain-dependent glycosylation functions of GalNAc-Ts and have provided a model explaining how these transferases differentially recognize such N- or C-terminal long-range prior glycosylation[15, 16]. We previously reported that the location of the lectin α-subdomain GalNAc-binding site, relative to the catalytic domain, was key to explaining how GalNAc-T2 delivered N-terminal acceptor sites of C-terminal monoglycopeptides to the catalytic domain in a lectin domain-dependent manner[4]. We also demonstrated that the flexible linker, located ~28–36 Å from the active site of GalNAc-T2, and ~18–25 Å from GalNAc-T2's lectin α-subdomain was in charge of the conformational heterogeneity of GalNAc-T2[4]. This was clearly due to the translational capacity of the lectin domain with respect to the catalytic domain, which was further modulated in the presence of substrates[4]. Note also that in all of the GalNAc-T2 structures reported, irrespectively of whether they are compact or extended structures, the lectin domain is positioned such that its α-subdomain GalNAc-binding site is always located on the left side of the lectin domain[4, 11, 12]. However, it has remained a conundrum how other GalNAc-Ts (i.e., GalNAc-T3, T4, T12, etc.) presented the opposite long-range glycosylation preference of GalNAc-T2 and others. Our multidisciplinary approach presented herein has revealed additional conformational properties of the short flexible linker that provide the molecular basis for the distinct long-range glycosylation preferences found for these enzymes. Our crystal structure of GalNAc-T4 revealed that its lectin domain is positioned such that its lectin α-subdomain GalNAc-binding site is now located on the right side of the lectin domain, in agreement with our proposed Model 1 (Figs. 1c and 3). This is fundamental to explaining why GalNAc-T4 and similar behaving GalNAc-Ts prefer to glycosylate C-terminal acceptor sites of N-terminal glycosylated glycopeptides. Our STD-NMR studies together with the transferase kinetic studies suggested that the orientation of the lectin domain-bound GalNAc moiety was fundamentally important for the correct orientation and presentation of the acceptor residues to the catalytic site, thereby leading to more efficient catalysis. Our MD simulations on multiple chimera transferases together with their kinetic analysis, further provide compelling evidence that the flexible linkers not only allow translational motion to the lectin domain, but also rotational motion. Both types of motions likely differ between the GalNAc-T isoforms and therefore will contribute to the different glycosylation preferences found among these enzymes. All together, our studies point to the dissimilar flexible linkers as major contributors to the distinct long-range glycosylation preferences of these transferases (Supplementary Fig. 7). It is further likely that some flexible linkers will have conformational properties in solution that allow two or more conformational states of the lectin domain relative to the catalytic domain thus accounting for the dual long-range preferences observed for GalNAc-T2 (and its chimeras and mutants) as well as the several GalNAc-Ts that display nearly equal N- and C-terminal long-range glycosylation preferences (i.e., GalNAc-T5, 13,16[15]). That we did not observe

such conformational flexibility in our MD trajectories of GalNAc-T2 (and some chimeras and mutants) may suggest that the activation energy for this interconversion may be very high and/or that our simulations were not sufficiently long enough.

In conclusion, we have provided for the first time the molecular basis for the distinctive long-range glycosylation preferences of the GalNAc-Ts, which is based on a very small flexible linker that provides rotational and translational capacity to the lectin domain. This work further exemplifies how a structural feature, very distant from both the active site and the lectin domain GalNAc-binding site, is capable of tuning the activity and specificity of these biologically important multidomain enzymes.

Methods

Cloning and purification of human GalNAc-T4/chimeras. The DNA sequence encoding amino acid residues 36–578 of the human GalNAc-T4, defined as *galnact4*, was codon optimized and synthesized by GenSript (USA) for expression in *Pichia pastoris*. The DNA, containing at the 5′ end a recognition sequence for *Xho*I and a KEX2 cleavage signal, and at the 3′ end a sequence encoding a histidine tag, a stop codon and a recognition sequence for *Sac*II, was cloned into the pUC57 vector (GenScript). Following digestion with *Xho*I and *Sac*II, the construct was subcloned into the protein expression vector pPICZαA, resulting in the expression plasmid pPICZαA-*galnact4*. Subsequently, the plasmid was linearized by *Sac*I-HF and transformed into SMD1168. Transformants were selected and colonies were grown as described before[26, 27]. The DNA sequences encoding the GalNAc-T4 chimeras were also synthesized by GenScript and cloned in the same above vector. DNA linearization and transformation in SMD1168 were also performed in the same manner described above. Note that none of the GalNAc-T4 chimeras were expressed as a soluble secreted form in *P. pastoris* supernatants, impeding their biophysical characterization.

Supernatant containing the human GalNAc-T4 was dialyzed against buffer A (20 mM Na_2HPO_4 pH 7.4, 20 mM imidazol 500 mM NaCl) and loaded into a His-Trap Column (GE Healthcare). The protein was eluted with an imidazol gradient in buffer A from 20 mM to 500 mM. Buffer exchange of GalNAc-T4 into buffer B (25 mM Tris pH 8, 150 mM NaCl) was carried out using a HiPrep 26/10 desalting column (GE Healthcare). GalNAc-T4 was further purified by size exclusion chromatography using a Superdex 75 XK26/60 column (Sigma) previously equilibrated with buffer B. Fractions containing GalNAc-T4 were dialized against buffer C (25 mM Tris pH 8, 1 mM TCEP), concentrated and used for biophysical experiments.

Cloning and purification of GalNAc-T2 chimeras/mutants. The DNA sequences encoding the GalNAc-T2 chimeras were synthesized by GenScript and cloned in the vector pPICZαA vector as previously described[12]. Site-directed mutagenesis experiments were also performed by GenScript using the previous reported template pPICZαA*galnact2* (K75-Q571)[12] and the plasmid encoding the **chimera 2**. The linearization and transformation of all constructs to SMD1168 were also performed in the same manner described for GalNAc-T4. The purification of all these GalNAc-T2 chimeras and mutants were expressed and purified using the purification protocol of the wild-type enzyme, as described previously[12]. Purity of enzymes was evaluated by SDS-PAGE coomassie staining and quantification of enzymes was quantified by absorbance at 280 nm using their theoretical extinction coefficients.

Crystallization. Crystals of the GalNAc-T4 were grown by hanging drop experiments at 18 °C by mixing 0.5 µl of protein solution (4 mg/ml GalNAc-T4, 5 mM UDP, 2 mM $MnCl_2$, and 5 mM MUC5AC-3,13[4] (GTT*PSPVPTTSTT*SAP) in 25 mM Tris pH 7.5, 0.5 mM EDTA, and 1 mM tris(2-carboxyethyl)phosphine (TCEP)) with an equal volume of a reservoir solution (18% PEG3350, 0.1 M ammonium nitrate). Under these conditions, crystals appeared within 2–5 days. Note that MUC5AC-3-13 acted as an additive to improve the diffraction quality and the resolution of the crystals. The crystals were soaked for 30 min with a mix containing 20 mM monoglycopeptide **3** and 20 mM UDP in 25 mM Tris pH 7.5 and 2 mM $MnCl_2$. Then, the crystals were cryo-protected with 25% ethylene glicol, 18% PEG3350, and 0.1 M ammonium nitrate, and frozen in a nitrogen gas stream cooled to 100 K.

Structure determination and refinement. The data were collected in the beamline I03 of Diamond (DLS) at a wavelength and temperature of 0.97 Å and a temperature of 100 K, respectively. The data were processed and scaled using the XDS package[28] and CCP4[29, 30] software. Relevant statistics are given in Supplementary Table 1. The crystal structure was solved by molecular replacement with Phaser[29, 30] and using the PDB entry 5AJP as the template that corresponds to the human GalNAc-T2. Initial phases were further improved by cycles of manual model building in Coot[31] and refinement with REFMAC5[32]. Once the GalNAc-T4 catalytic domain was unambiguously built and refined, ARP/wARP[29, 30] was used to fully build the GalNAc-T4 lectin domain. Again, new rounds of Coot and refinement with REFMAC5 were performed. The crystal structure of GalNAc-T4 from crystals co-crystallized with MUC5AC-3-13 displayed no electron density for the MUC5AC-3-13 diglycopeptide. The final model of crystals soaked with monoglycopeptide **3** and UDP was validated with PROCHECK, model statistics are given in Supplementary Table 1. The asymmetric unit of the triclinic crystal contained two molecules of GalNAc-T4. Despite the long soaking and high concentration of monoglycopeptide **3** and UDP, only clear density was visualized for the presence of GalNAc-O-Thr and GalNAc moieties of monoglycopeptide **3** bound to the lectin α-subdomain GalNAc-binding site (Fig. 3a, c). The Ramachandran plot shows that 96.71, 2.89, and 0.40% of the amino acids are in most favored, allowed, and disallowed regions, respectively.

Surface plasmon resonance experiments. SPR experiments were performed at 25 °C with a Biacore X-100 apparatus (Biacore AB) in 25 mM Tris buffer, 1 mM DTT, 4 mM $MnCl_2$, 100 µM UDP, 0.01% surfactant P20, pH 7.5 (running buffer) at 25 °C. Flow cells (CM5 sensor chip; Biacore) were activated for 7 min by injecting 140 µl of 50 mM N-hydroxysuccinimide (NHS): 200 mM ethyl-3(3dimethylamino) propylcarbodiimide (EDC). GalNAc-T4 was immobilized on a flow cell 2 by injection of a 100 µg/ml protein solution diluted with 10 mM sodium acetate buffer with a flow rate of 10 µl/min for 7 min followed by an injection of 130 µl ethanolamine to block any remaining activated groups on the surface. The level of immobilization reached was about 8000 RU. Flow cell 1, used as reference, was blocked with ethanolamine at the same conditions of flow cell 2 without immobilization of protein. Affinity experiments were made using a series of different concentrations of peptide **3** in the range of 0.01–5 mM with a flow rate of 30 µl/min during 100 s. Each injection was followed by a 100 s injection of running buffer (dissociation phase). No regeneration steps were performed between injections. Response data were collected at real time and analyzed with the Biacore® X-100 Evaluation software and plotted as response shift versus analyte concentration. We could not determine a K_d value due to binding saturation was not achieved.

Synthesis of peptides. Peptides were synthesized via stepwise microwave-assisted solid-phase peptide synthesis on a Liberty Blue synthesizer using the Fmoc strategy on Rink Amide MBHA resin (0.1 mmol). The glycosylated amino acid building blocks (2.0 equiv) were synthesized as described in the literature[33] and manually coupled using HBTU, while the other Fmoc amino acids (5.0 equiv) were automatically coupled using oxyma pure/DIC. The *O*-acetyl groups of $(AcO)_3$GalNAc moiety were removed in a mixture of NH_2NH_2/MeOH (7:3). The peptides were then released from the resin, and all acid sensitive side-chain protecting groups were simultaneously removed using TFA 95%, TIS 2.5%, H_2O 2.5%, followed by precipitation with cold diethyl ether. Finally, they were purified by HPLC using a Phenomenex Luna C18(2) column (10 µ, 250 mm × 21.2 mm) and a dual absorbance detector, with a flow rate of 20 ml/min. Peptides were further subjected to Edman amino acid sequencing on a Shimadzu PPSQ-53A peptide sequencer prior to use.

Peptide **1**: HPLC: Rt = 11.53 min (Grad: water 0.1% TFA/acetonitrile (95:5) → (87:13), 13 min, λ = 212 nm). HRMS ESI + (*m/z*) calcd. for $C_{39}H_{66}N_{14}O_{15}$ $[M + 2H]^{2+}$ 485.2411, found 485.2301.

Monoglycopeptide **2**: HPLC: Rt = 11.75 min (Grad: water 0.1% TFA/acetonitrile (95:5) → (85.6:14.4), 15 min, λ = 212 nm). HRMS ESI + (*m/z*) calcd. for $C_{57}H_{95}N_{18}O_{24}$ $[M + H]^+$ 1415.6761, found 1415.6739.

Monoglycopeptide **3**: HPLC: Rt = 13.72 min (Grad: water 0.1% TFA/acetonitrile (93:7) → (86:14), 15 min, λ = 212 nm). HRMS ESI + (*m/z*) calcd. for $C_{56}H_{93}N_{18}O_{24}$ $[M + H]^+$ 1401.6605, found 1401.6546.

NMR experiments. All NMR experiments were recorded on a Bruker Avance 600 MHz spectrometer equipped with a triple channel cryoprobe head. The ^{1}H NMR resonances of the peptides **1–3** 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-d11 in 90:10 H_2O/D_2O, 7.5 mM NaCl, and 1 mM DTT, uncorrected pH 7.4. The assignments were accomplished either at 278 or 298 K. The resonance of 2,2,3,3-tetradeutero-3-trimethylsilylpropionic acid (TSP) was used as a chemical shift reference in the ^{1}H NMR experiments (δ 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.

Samples for STD experiments were prepared in perdeuterated 25 mM TRIS-d11 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 75 µM UDP, 75 µM $MnCl_2$ with ~880 µM peptide (or GalNAc-O-Me) and 13.5 µM GalNAc-T4 giving a molar ratio of 65:1 peptide:GalNAc-T4.

The STD-NMR spectra were acquired with 1920 transients in a matrix with 64 k data points in t2 in a spectral window of 12335.53 Hz centered at 2819.65 Hz. An excitation sculpting module with gradients 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° 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 in order to optimize the frequency for protein saturation (−1 ppm) and to ensure that the ligand signals were not affected. However, all the 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 taken into account, by subtraction, when quantifying the STD-NMR data in the presence of the transferase. As well, a blank STD experiment with only the protein was also recorded. The substraction of this protein STD spectrum allowed eliminating the signal background of the protein. In all cases, to accomplish the STD-NMR-derived epitope mapping of each ligand, the STD-NMR total intensities were normalized with respect to the highest STD-NMR response. For Fig. 2b and Supplementary Fig. 1, 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 of as well as, the Hα protons of the Ala amino acids of glycopeptides could not be analyzed in the STD-NMR spectra due to their close proximity to the HDO resonance. Proton resonances, from Gly and Pro, appear in the same chemical shift region of the spectrum and were not discriminated.

Molecular docking. Two different docking calculations were conducted to generate the ternary complex.

First, UDP was docked into GalNAc-T4 with the aid of of AutoDock Vina 1.1.2[34]. The predicted binding energies ranged from −6.4 to −5.2 kcal/mol. The Autogrid grid point spacing was set at 0.375 Å, center coordinates of the grid box were 5.4, −2.4, − 7.2 (*x*, *y*, *z*), and number of grid points in *xyz* was 76, 40, 42, respectively. All allowed torsional bonds were considered rotatable. The 3D structure of the docked ligand to the protein with the lowest-binding energy was used for further calculations (see below).

In a second simulation, glycopeptide T(α-*O*-GalNAc)GAGAGAGTTPGPG was docked into GalNAc-T4. The Autogrid grid point spacing was set at 0.375 Å, center coordinates of the grid box were 0.0, −9.7, 9.4 (*x*, *y*, *z*), and number of grid points in *xyz* was 86, 60, 90, respectively. All allowed torsional bonds were considered rotatable, except the bonds involved in the glycosidic linkage, which were fixed to $\phi = 64°$, $\psi = 110°$. The 3D structure of the docked ligand to the protein with lowest-binding energy and with the GalNAc unit located in the lectin domain region, was then used for further calculations.

The combination of both structures was used to build up the ternary complex, which was subjected to MD simulations (see below).

MD simulations. The starting coordinates for the complex between glycopeptide **3**, UDP/Mn^{+2}, and GalNAc-T4 were generated combining the X-ray structure of GalNAc-T4 solved in this manuscript and the crystal structure of the activated form of GalNAc-T2 in complex with UDP/Mn^{+2} and MUC5AC-13 (PDB entry 5AJP). The GalNAc unit of the glycopeptide **3** was located by superimposition to the one solved in the X-ray structure of GalNAc-T4, and UDP was located by superimposition of GalNAc-T2 and T4 proteins. The peptide backbone of compound **3** was situated so that the Thr residues were laying on the proximity of the UDP moiety.

The starting coordinates for **chimera 2** were generated by superimposing the X-ray structures of GalNAc-T2 (PDB entry: 5AJP) and GalNAc-T4. Residues 435–445 (PELRVPDHQDI) were manually deleted from GalNAc-T2 and replaced by residues 494–506 (PEVYVPDLNPVIS) of GalNAc-T3 using PyMol (http://www.pymol.org). Residues $A_{507}F_{508}$ were maintained in GalNAc-T2 to facilitate the linkage between its catalytic and lectin domains. The double (R438A-D444A) and triple mutant (R438A-D444A-F447A) were generated using PyMol. The same protocol was used for other computational chimeras.

Force field parameters for the substrates were generated with the antechamber module of Amber14 using a combination of GLYCAM06[35] parameters for the GalNAc unit and the general Amber force field (GAFF) for GDP, with partial charges set to fit the electrostatic potential generated with HF/6-31 G(d) by RESP. The charges are calculated according to the Merz–Singh–Kollman scheme using Gaussian 09[36]. Each protein was immersed in a truncated octahedral box with a 10 Å buffer of TIP3P water molecules and neutralized by adding explicit counter ions (Na^+, Cl^-). All subsequent simulations were performed using the *ff14SB* force field[37]. A two-stage geometry optimization approach was used. The first stage minimizes only the positions of solvent molecules and ions, and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems were then gently heated by incrementing the temperature from 0 to 300 K under a constant pressure of 1 atm and periodic boundary conditions. Harmonic restraints of 30 kcal/mol were applied to the solute, and the Andersen temperature coupling scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages. Water molecules are treated with the SHAKE algorithm such that the angle between the hydrogen atoms is kept fixed. Long-range electrostatic effects are modeled using the particle-mesh-Ewald method. An 8-Å cutoff was applied to Lennard–Jones and electrostatic interactions. Each system was equilibrated for 2 ns with a 2-fs time step at a constant volume and temperature of 300 K. Production trajectories were then run for additional 100–500 ns under the same simulation conditions.

Transferase assays and kinetics. Specific activity determinations: GalNAc-T glycosylation reactions against (glyco)peptides **1**–**3** were performed with 75 mM sodium cacodylate, pH 6.5, 1 mM 2-mercaptoethanol, 10 mM $MnCl_2$, 0.25 mM [^{3}H]-radiolabeled UDP-GalNAc (~6×10^8 DPM/μmole, American Radiolabeled Chemicals Inc.), and 1.4 mM (~1.25 mg/ml) of (glyco)peptide (from 4 mM stock in 1 mM TRIS, pH ~7) and varying concentrations (0.01 to 0.5 μM) of transferase (determined by OD_{280}), giving a final reaction volumes of 20 μl in 100 μl capped Eppendorf tubes. Reactions were incubated at 37 °C in a thermostated microplate shaker (Taitec Microincubator M-36) and quenched at the appropriate reaction time by the addition of 20 μl of 250 mM EDTA and placed on ice prior to final workup. Typically, for each transferase, glycosylation reactions were performed at the same time with all three (glyco)peptide substrates (including a no peptide control) using the same batch of transferase and the same UDP-[^{3}H]-GalNAc stock. Before performing time course experiments, the relative activity of each transferase was determined by trial and error against all three peptides to determine the optimal transferase concentration or incubation time that gave no more than 10% peptide glycosylation against the best substrate. Final time course experiments typically consisted of incubation times of 10, 20, and 30, or 30, 60, and 90 or 120 min (depending on activity) and were performed at least twice. Most transferases were also characterized at two different enzyme concentrations (0.01–0.03 and 0.2–0.5 μM) and the results combined. After quenching the reaction and diluting to 4 ml, free UDP and non-hydrolyzed UDP-GalNAc were removed by passage over a Dowex 1 × 8 anion exchange resin (~3 ml column). Total UDP-[^{3}H]-GalNAc utilization (transfer to peptide substrate and transfer to water, i.e, hydrolysis) was determined by difference after scintillation counting (Beckman LS5801 scintillation counter) 1/20 of the sample before and after passage over Dowex 1 × 8. The extent of [^{3}H]-GalNAc transfer to peptide and the extent of hydrolysis was determined by Sephadex G10 gel filtration analysis, typically for the longest reaction time point, as previously described[15]. Example gel filtration chromatograms are given in Supplementary Fig. 12. Significant hydrolysis was only observed for GalNAc-T4, which was corrected for in determining its specific activity. Specific activity was obtained by calculating at each individual data time point (of the combined time course plots, Fig. 4b), an individual-specific activity, which was then averaged to obtain the overall average specific activity (and standard deviation). Specific activity is reported as nmol of glycopeptide product formed per min, per pmol of enzyme. Substrate activity ratios (i.e., substrate ratios **2/1**, **3/1**, and **3/2**) were also obtained at each individual time point and averaged (with standard deviation). These data are tabulated in Supplementary Fig. 9. Note that the Microsoft Excel least squares fitting function LINEST was used to obtain the linear plots shown in Fig. 4b while fixing the *y* intercept to zero. Note that the specific activity calculated from the least square slope differed slightly (less than 10%) from that determined from the average of the individual time points.

Detailed enzyme kinetics on GalNAc-T4, GalNAc-T2, the GalNAc-T2 triple mutant and GalNAc-T2 **chimera 3**: Glycosylation reactions were performed as described above using a fixed enzyme concentration (0.06 μM for GalNAc-T4, and 0.02 μM for GalNAc-T2 and its linker constructs) with varying peptide concentrations (typically 45, 90, 180, 350, 700, 1400, and 2000 μM). Incubation times (10–90 min) were chosen such that no more than 30% of the UDP-GalNAc donor was depleted while typically giving less than 10% peptide glycosylation, after correction for UDP-GalNAc hydrolysis described above. Specific activities were obtained from one to two individual incubation time points performed in two–three separate experiments. These individual-specific activity values were used to calculate the kinetic constants of K_m and V_{max} using the nonlinear Michaelis–Menten fitting program in GraphPad Prism 7.03.

Determination of site of glycosylation. The determination of substrate glycosylation sites was performed by Edman amino acid sequencing (Applied Biosystems Procise 494 peptide sequencer) of the G10 isolated [^{3}H]-GalNAc glycosylated substrate as previously described[15, 16] where each cycle was collected on a fraction collector and scintillation counted (Beckman LS5801 scintillation counter) for [^{3}H]-GalNAc content. Although there is commonly sample-to-sample variability in the [^{3}H]-GalNAc content loaded on the sequencer (due sample losses and different initial UDP-[^{3}H]-GalNAc-specific activities), the observed sites of incorporation were found to be identical between different experiments with the same transferase and substrate. The presence of [^{3}H]-GalNAc lag after a peak of [^{3}H]-GalNAc incorporation is commonly observed in these determinations, which is due to the poor extraction from the sample filter of the glycosylated-PTH residues compared with the standard amino acid PTH derivatives[15, 16].

Data availability. The coordinate and structure factor for GalNAc-T4 in complex with monoglycopeptide 3 has been deposited in the Worldwide Protein Data Bank (wwPDB) with the accession code 5NQA. All relevant data are available from the authors on reasonable request.

Received: 8 April 2017 Accepted: 1 November 2017
Published online: 05 December 2017

References

1. 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).
2. Gill, D. J., Clausen, H. & Bard, F. Location, location, location: new insights into O-GalNAc protein glycosylation. *Trends Cell Biol.* **21**, 149–158 (2011).
3. Hurtado-Guerrero, R. Recent structural and mechanistic insights into protein O-GalNAc glycosylation. *Biochem. Soc. Trans.* **44**, 61–67 (2016).
4. Lira-Navarrete, E. et al. Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein O-glycosylation. *Nat. Commun.* **6**, 6937 (2015).
5. Tabak, L. A. The role of mucin-type O-glycans in eukaryotic development. *Semin. Cell Dev. Biol.* **21**, 616–621 (2010).
6. Schjoldager, K. T. & Clausen, H. Site-specific protein O-glycosylation modulates proprotein processing - deciphering specific functions of the large polypeptide GalNAc-transferase gene family. *Biochim. Biophys. Acta* **1820**, 2079–2094 (2012).
7. Schjoldager, K. T. et al. Deconstruction of O-glycosylation–GalNAc-T isoforms direct distinct subsets of the O-glycoproteome. *EMBO Rep.* **16**, 1713–1722 (2015).
8. Hanson, R. L. & Hollingsworth, M. A. Functional consequences of differential O-glycosylation of MUC1, MUC4, and MUC16 (downstream effects on signaling). *Biomolecules* **6**, E34 (2016).
9. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
10. 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 alpha-N-acetylgalactosaminyltransferase-T1. *Proc. Natl Acad. Sci. USA* **101**, 15307–15312 (2004).
11. 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).
12. Lira-Navarrete, E. et al. Substrate-guided front-face reaction revealed by combined structural snapshots and metadynamics for the polypeptide N-acetylgalactosaminyltransferase 2. *Angew. Chem. Int. Ed. Engl.* **53**, 8206–8210 (2014).
13. Kubota, T. et al. Structural basis of carbohydrate transfer activity by human UDP-GalNAc: polypeptide alpha-N-acetylgalactosaminyltransferase (pp-GalNAc-T10). *J. Mol. Biol.* **359**, 708–727 (2006).
14. 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).
15. 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).
16. 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).
17. 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).
18. 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).
19. Posey, A. D. Jr et al. Engineered CAR T cells targeting the cancer-associated Tn-Glycoform of the membrane mucin MUC1 control adenocarcinoma. *Immunity* **44**, 1444–1454 (2016).
20. McEver, R. P. & Cummings, R. D. Role of PSGL-1 binding to selectins in leukocyte recruitment. *J. Clin. Invest.* **100**, S97–S103 (1997).
21. 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).
22. 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).
23. Pedersen, J. W. et al. Lectin domains of polypeptide GalNAc transferases exhibit glycopeptide binding specificity. *J. Biol. Chem.* **286**, 32684–32696 (2011).
24. 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).
25. Khetarpal, S. A. et al. Loss of function of GALNT2 lowers high-density lipoproteins in humans, nonhuman primates, and rodents. *Cell Metab.* **24**, 234–245 (2016).
26. Lira-Navarrete, E. et al. Structural insights into the mechanism of protein O-fucosylation. *PLoS ONE* **6**, e25365 (2011).
27. Valero-Gonzalez, J. et al. A proactive role of water molecules in acceptor recognition by protein O-fucosyltransferase 2. *Nat. Chem. Biol.* **12**, 240–246 (2016).
28. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
29. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
30. Collaborative Computational Project, Number 4. The CCP4 suite programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 760–763 (1994).
31. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
32. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).
33. Plattner, C., Hofener, M. & Sewald, N. One-pot azidochlorination of glycals. *Org. Lett.* **13**, 545–547 (2011).
34. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
35. Kirschner, K. N. et al. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **29**, 622–655 (2008).
36. Frisch, M. J. et al. *Gaussian 09, Revision D.01* (Gaussian, Inc., Wallingford, 2009).
37. Hornak, V. et al. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725 (2006).

Acknowledgements

We thank synchrotron radiation sources DLS (Oxford) and in particular beamline I03 (experiment number MX10121-7). We thank ARAID, MEC (CTQ2013-44367-C2-2-P, BFU2016-75633-P, CTQ2015-67727-R, CTQ2015-70524-R, and RYC-2013-14706), the National Institutes of Health (GM113534, and instrument grant GM113534-01S), the Danish National Research Foundation (DNRF107), the FCT-Portugal (UID/Multi/04378/2013 and PTNMR Project No 022161), and the DGA (B89) for the financial support. I.C. thanks Universidad de La Rioja for the FPI grant. F.M. thanks FCT-Portugal for IF Investigator. E.L.-N. acknowledges her postdoctoral EMBO fellowship ALTF 1553-2015 co-funded by the European Commission (LTFCOFUND2013, GA-2013-609409) and Marie Curie Actions. H.C. and J.J.-B. thank EU for the TOLLerant project. The research leading to these results has also received funding from the FP7 (2007–2013) under BioStruct-X (grant agreement No. 283570 and BIOSTRUCTX_5186). We also thank BIFI (Memento cluster) and CESGA for computer support.

Author contributions

R.H.-G. designed the crystallization construct and solved the crystal structure. E.L.-N., M. R., and R.H.-G. purified the enzymes, crystallized the complex, and refined the crystal structure. I.C., F.C., and J.M.P. synthetized the glycopeptides. F.C. and G.J.-O. performed the MD simulations. H.Co., A.D., J.J.-B., and F.M. performed and analyzed the NMR experiments. T.A.G. and E.J.P.D. performed the kinetic studies together with the Edman amino acid sequencing. R.H.-G. wrote the article with the main contribution of T.A.G., F. C., G.J.-O., H.Cl., and F.M. All authors read and approved the final manuscript.

Additional information

Supplementary Information accompanies this paper at doi:https://doi.org/10.1038/s41467-017-02006-0.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SUPPLEMENTARY INFORMATION

The interdomain flexible linker of the polypeptide GalNAc transferases (GalNAc-Ts) dictates their long-range glycosylation preferences

Matilde de las Rivas[1,12], Erandi Lira-Navarrete[1,6,12], James Earnest[2,12], Ismael Compañón[4], Helena Coelho[3,5,11], Ana Diniz[3], Jesús Jiménez-Barbero[5,10,11], Jesús M. Peregrina[4], Henrik Clausen[6], Francisco Corzana[4], Filipa Marcelo[3], Gonzalo Jiménez-Osés[4], Thomas A. Gerken[2,8], and Ramon Hurtado-Guerrero[1,7]*

[1]. BIFI, University of Zaragoza, BIFI-IQFR (CSIC) Joint Unit, Mariano Esquillor s/n, Campus Rio Ebro, Edificio I+D, Zaragoza, Spain.

[2]. Department of Chemistry, Case Western Reserve University, Cleveland, OHIO, USA.

[3]. UCIBIO, REQUIMTE, departamento de Química, Faculdade de Ciências e Tecnologia, Universidade de Nova de Lisboa , Caparica, Portugal.

[4]. Departamento de Química, Universidad de La Rioja, Centro de Investigación en Síntesis Química, E-26006 Logroño, Spain.

[5]. CIC bioGUNE, Bizkaia Technology Park, Building 801A, 48170 Derio, Spain.

[6]. Copenhagen Center for Glycomics, Department of Cellular and Molecular Medicine, School of Dentistry, University of Copenhagen, Copenhagen, Denmark.

[7]. Fundación ARAID, 50018, Zaragoza, Spain.

[8]. Departments of Pediatrics and Biochemistry, Case Western Reserve University, Cleveland, OHIO, USA

[10]. Ikerbasque, Basque Foundation for Science, Maria Diaz de Haro 13, 48009 Bilbao, Spain

[11]. Departament of Organic Chemistry II, Faculty of Science & Technology, University of the Basque Country, 48940 Leioa, Bizkaia, Spain

[12]. These authors contributed equally to this work.

* To whom correspondence should be addressed.

E-mail: rhurtado@bifi.es

Supplementary Table 1. Data collection and refinement statistics. Values in parentheses refer to the highest resolution shell. Ramachandran plot statistics were determined with PROCHECK.

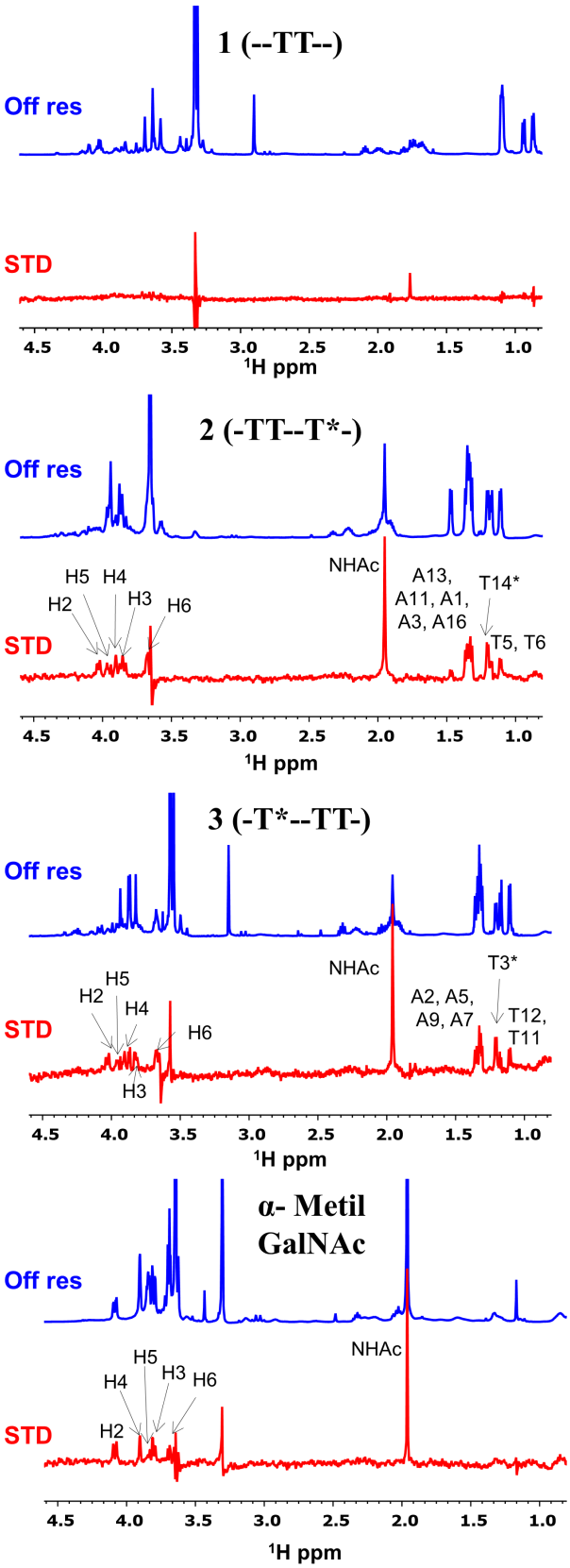
	GalNAc-T4 in complex with monoglycopeptide 3*
Space group	P1
Wavelength (Å)	0.97
Resolution (Å)	20-1.90 (2.00-1.90)
Cell dimensions a, b, c (Å) α, β, γ (°)	*a = 65.45* *b*= 79.88 *c* = 88.61 *116.47, 96.74, 104.72*
Mn(I) half-set correlation CC(1/2)	0.988 (0.560)
Unique reflections	115064
Completeness	97.5 (95.8)
R_{pim}	0.068 (0.556)
$I/\sigma(I)$	6.9 (1.3)
Redundancy	3.6 (3.2)
R_{work} / R_{free}	0.225/0.273
RMSD from ideal geometry, bonds (Å)	0.014
RMSD from ideal geometry, angles (°)	1.718
<B> GalNAc-T4 (Å^2)	32.48
<B> GalNAc (Å^2)	48.44
<B> Thr3-GalNAc (Å^2)	52.45
<*B*> solvent (Å^2)	43.22
<*B*> Ethylenglycol (Å^2)	48.40
<*B*> Glycerol (Å^2)	57.74
Ramachandran plot:	
Most favoured (%)	96.71
Additionally allowed (%)	2.89
Disallowed (%)	0.40
PDB ID	5NQA

*Note that the density is only good enough to model either GalNAc or GalNAc bound to Thr3 of monoglycopeptide **3**.

Supplementary Table 2. 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[1]	V_{max}[2]	0.7 ± 0.09	3.1 ± 0.07	3.8 ± 0.34
	K_m[3]	1.2 ± 0.31	0.2 ± 0.01	0.75 ± 0.13
	Catalytic efficiency[4]	**0.56 ± 0.34**	**17.0 ± 3.1**	**5.0 ± 2.6**
Triple Mutant	V_{max}	1.1 ± 0.35	1.8 ± 0.06	2.8 ± 0.10
	K_m	3.3 ± 1.55	0.4 ± 0.04	0.6 ± 0.05
	Catalytic efficiency	**0.33 ± 0.23**	**4.5 ± 1.5**	**4.9 ± 1.9**
Chimera 3	V_{max}	0.8 ± 0.1	1.35 ± 0.1	3.7 ± 0.3
	K_m	1 ± 0.2	0.7 ± 0.15	1 ± 0.15
	Catalytic efficiency	**0.79 ± 0.38**	**1.9 ± 0.80**	**3.7 ± 1.8**
GalNAc-T4[5]	V_{max}	-	-	3.1 ± 0.25
	K_m	-	-	0.041 ± 0.018
	Catalytic efficiency	**-**	**-**	**75.2 ± 13**

1) 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 monoglycopeptide 3 at high substrate concentrations, i.e its Vmax value. Parameters for peptides 1 and 2 were unchanged using this weighting compared to the unweighted fit.
2) V_{max} units: mM/(nmole*min)
3) K_m units: µM
4) Catalytic efficiency (V_{max}/K_m): $(nmole*min)^{-1}$. Note that the given standard deviations were estimated from the ratio of the individual V_{max} and K_m standard deviations.
5) For GalNAc-T4 the activities of peptides 1 and 2 were too low to determine their kinetic constants.



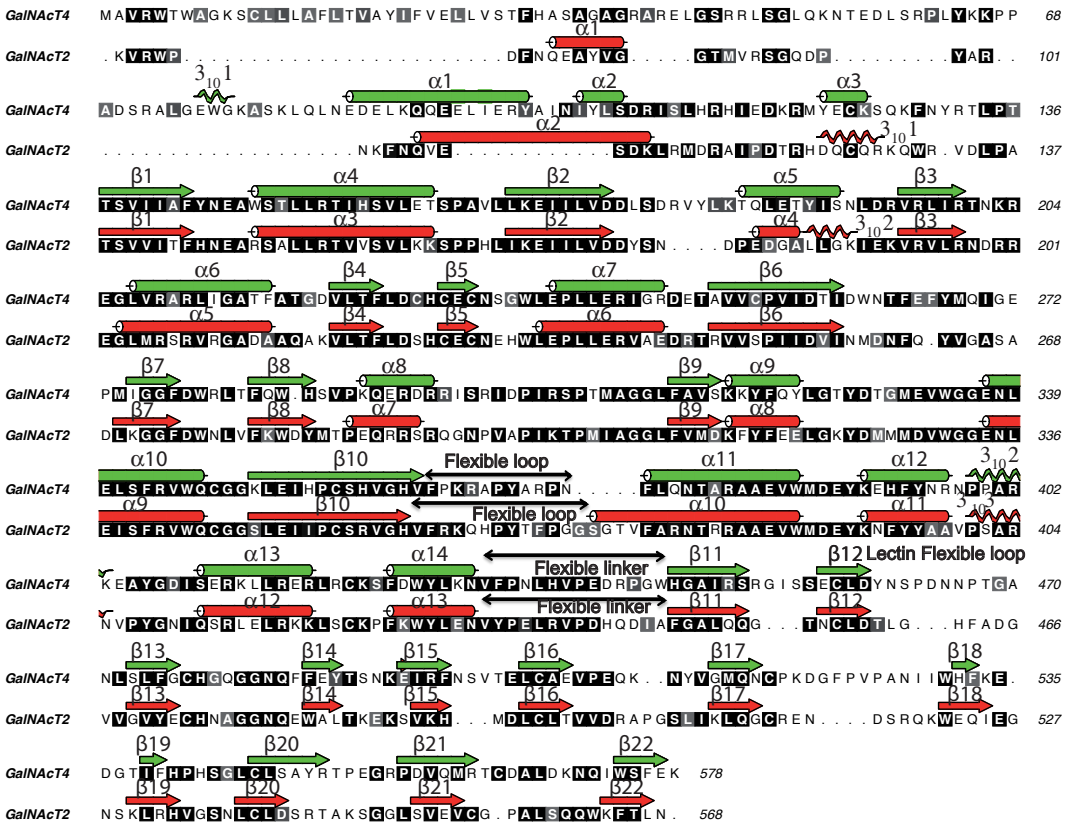
Proton	% STD relative
H2	100
H3	<20
H4	49
H5	<20
H6	<20
NHAc	45
A13, A11, A1, A3, A16	<20
T14*	21
T5, T6	<20

Proton	% STD relative
H2	100
H3	75
H4	93
H5	30
H6	29
NHAc	87
A2, A5, A9, A7	<20
T3*	31

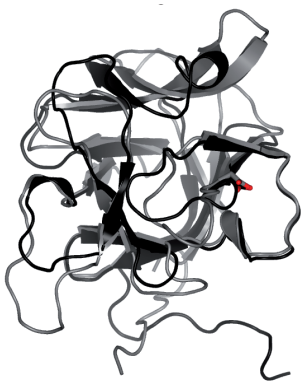
Proton	% STD relative
H2	100
H3	55
H4	42
H5	25
H6	<20
OMe	<20
NHAc	46

Supplementary Figure 1. STD-NMR Experiments. For all compounds the reference spectrum (Off res) is displayed in blue color while the STD spectrum (STD) is displayed in red. No STD response was observed in the spectrum with peptide **1**. For the other compounds (monoglycopeptides **2**, **3** and α-

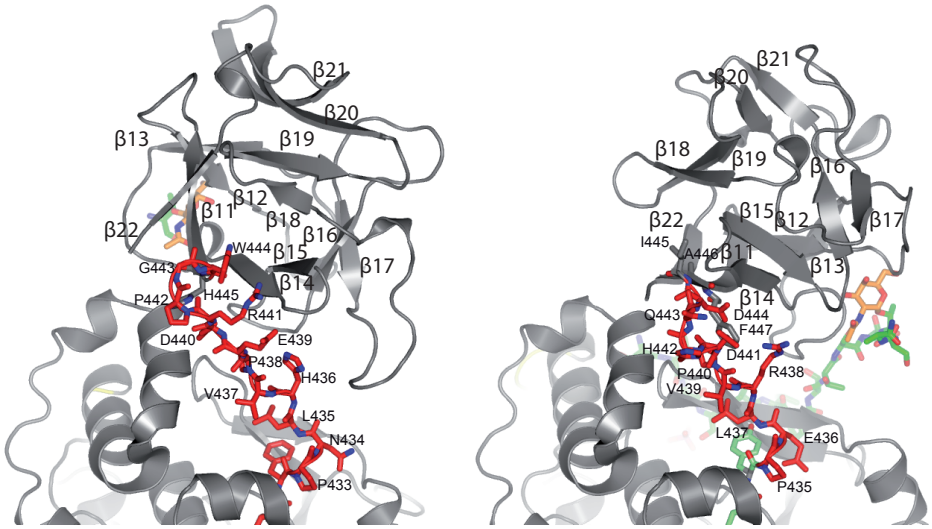
methyl-GalNAc) the key proton resonances are marked in each STD spectrum and their relative STD percentages are represented in the table at right of the spectra.



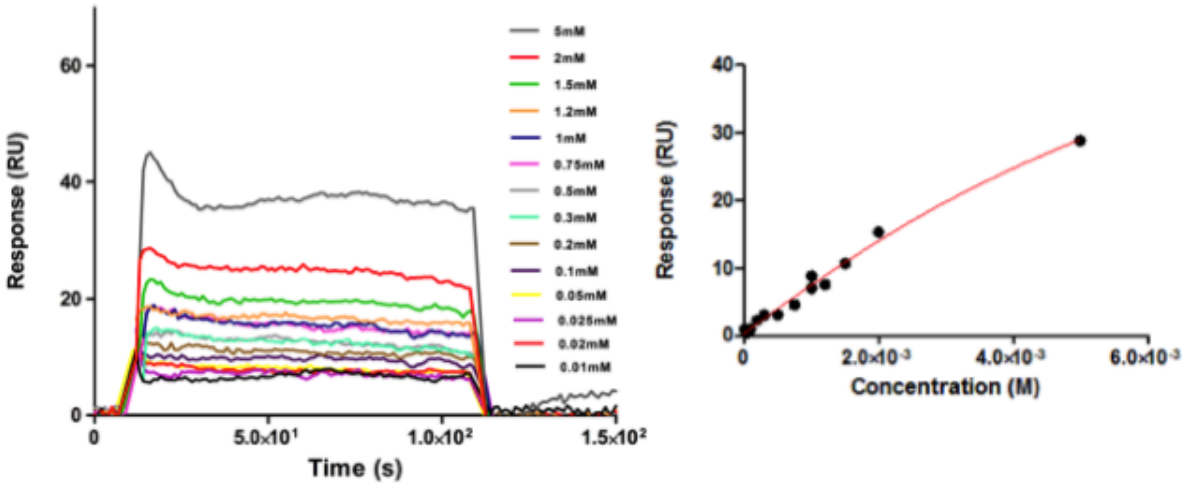
Supplementary Figure 2. Sequence alignment of human GalNAc-T2 and GalNAc-T4. Secondary structure elements from the GalNAc-T2 and T4 structures are shown, with α-helices, 3_{10}-helices and β-strands in red and green, respectively. The regions encompassing the flexible loop and linker are indicated with double sided arrows.



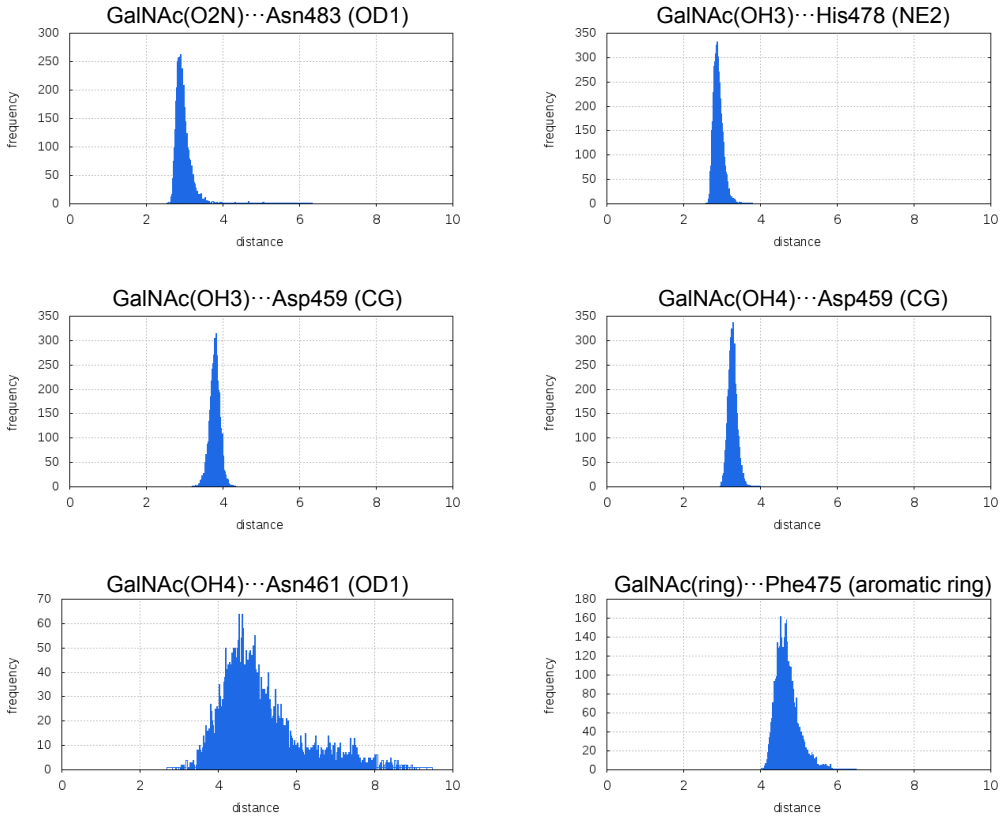
Supplementary Figure 3. Superposition analysis between GalNAc-T2 (black) and GalNAc-T4 (grey) lectin domains. Asp458 and Asp459 are shown as black and grey carbon atoms, respectively.

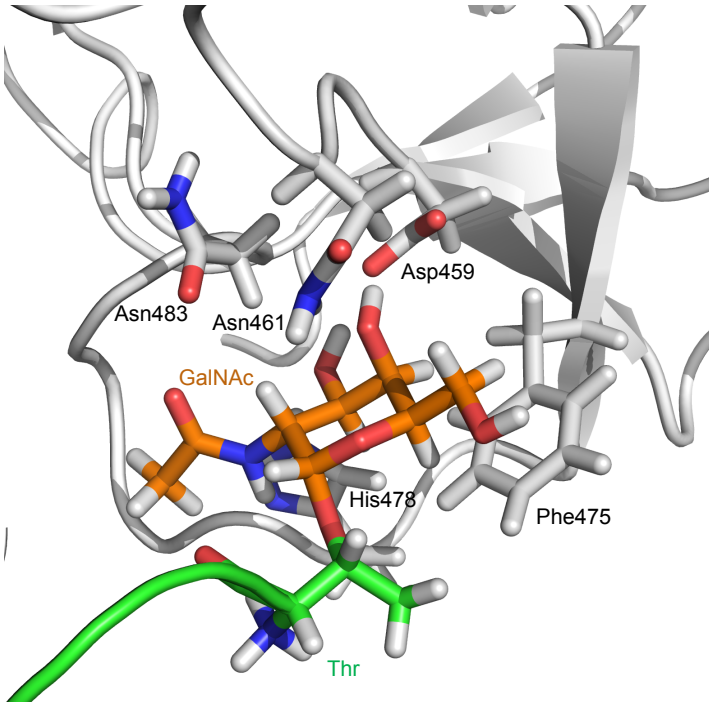


Supplementary Figure 4. Close up-view of the GalNAc-T4 (left) and GalNAc-T2 (right) lectin domains. Colours for the peptides, proteins, and the GalNAc moiety are the same as indicated in **Fig. 3**. The flexible linkers are shown as sticks in red carbon atoms. The numbering of the lectin domain secondary structural elements is also shown.

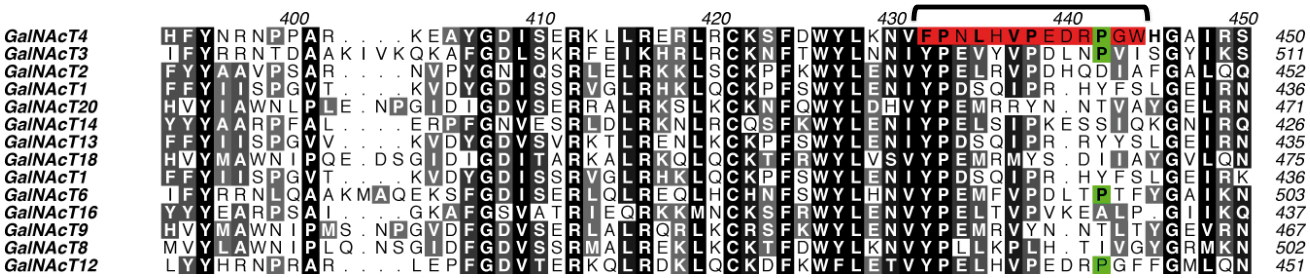


Supplementary Figure 5. SPR sensogram. (left) Sensogram for binding of the monoglycopeptide **3** to GalNAc-T4**.** Used ligand concentrations are reported in the inset legend. (right) Fitting of SPR data. The end-points of the various injections were plotted against protein concentration. Note that the Kd could not be determined because binding saturation could not be achieved.

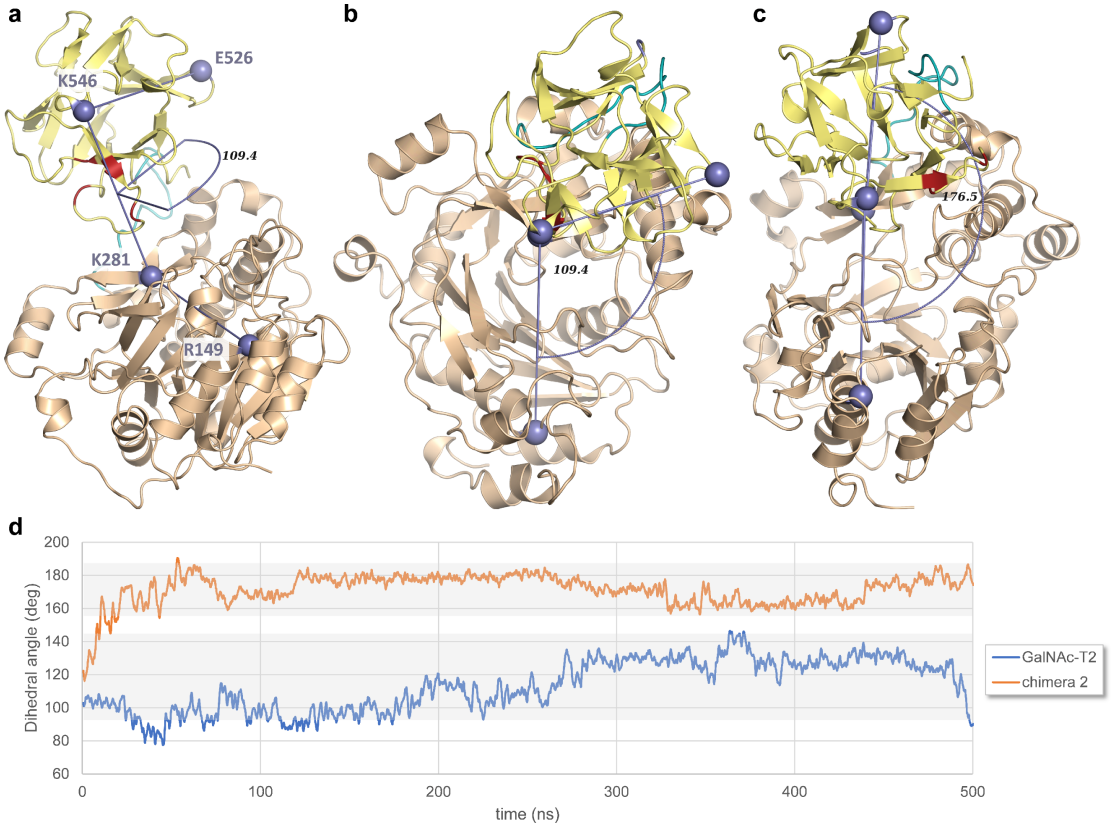




Supplementary Figure 6. Representative distance distributions obtained by 200 ns MD simulations on the complex between GalNAc-T4 and glycopeptide 3. These data indicate that the GalNAc unit is retained in the lectin domain during the simulation time.

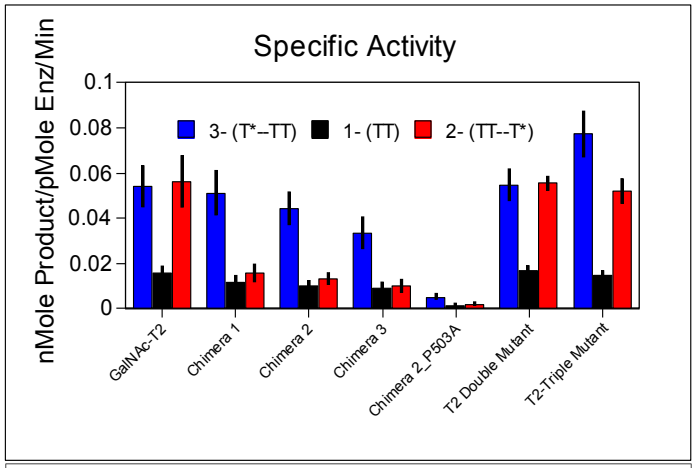


Supplementary Figure 7. Multiple sequence alignment for GalNAc-Ts spanning the regions around the flexible linker. The sequence highlighted in red denotes the flexible linker of GalNAc-T4. The additional Pro in GalNAc-T4/T3/T6/T12 is highlighted in green. The alignment shows the most conserved region between the flexible linkers is towards the N-termini whereas more dissimilarities are around the C-termini.

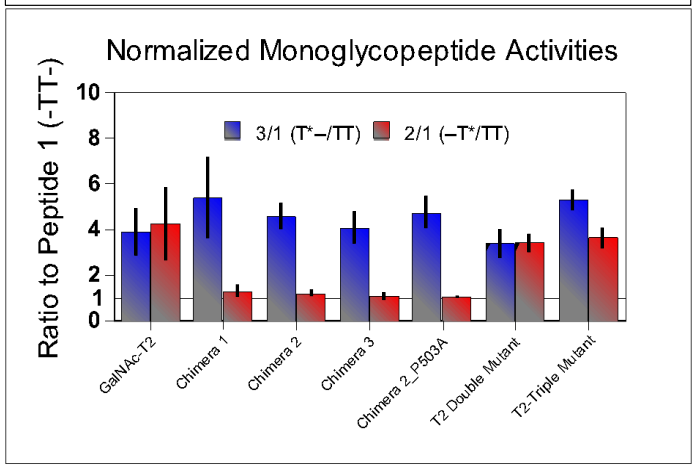


Supplementary Figure 8. Rotation of the lectin domains in GalNAc-T2 and chimera 2. **a**, Selected positions (alpha carbons, in blue) arbitrarily chosen to define a dihedral angle describing the relative positions of the lectin and catalytic domains. **b**, Top view of the first MD snapshot (time = 0 ns) showing a nearly perpendicular relative orientation (109º) of the two domains in chimera2. **c**, Top view of the last MD snapshot (time = 500 ns) showing a nearly parallel relative orientation (177º) of the two domains in chimera 2. **d**, Evolution of the rotation dihedral angle for GalNAc-T2 (in blue) and chimera 2 (in orange) as a function of the MD simulation time. For chimera 2, a quick ~70º rotation of the lectin domain is observed within the first ~30 ns, remaining stable in a nearly parallel orientation for the rest of the simulation. However, for GalNAc-T2, the initial close to perpendicular relative disposition of the domains is maintained, with fluctuations, throughout all the simulation.

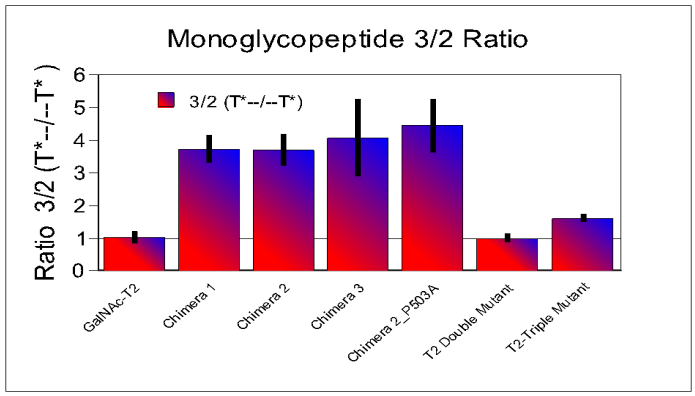
a



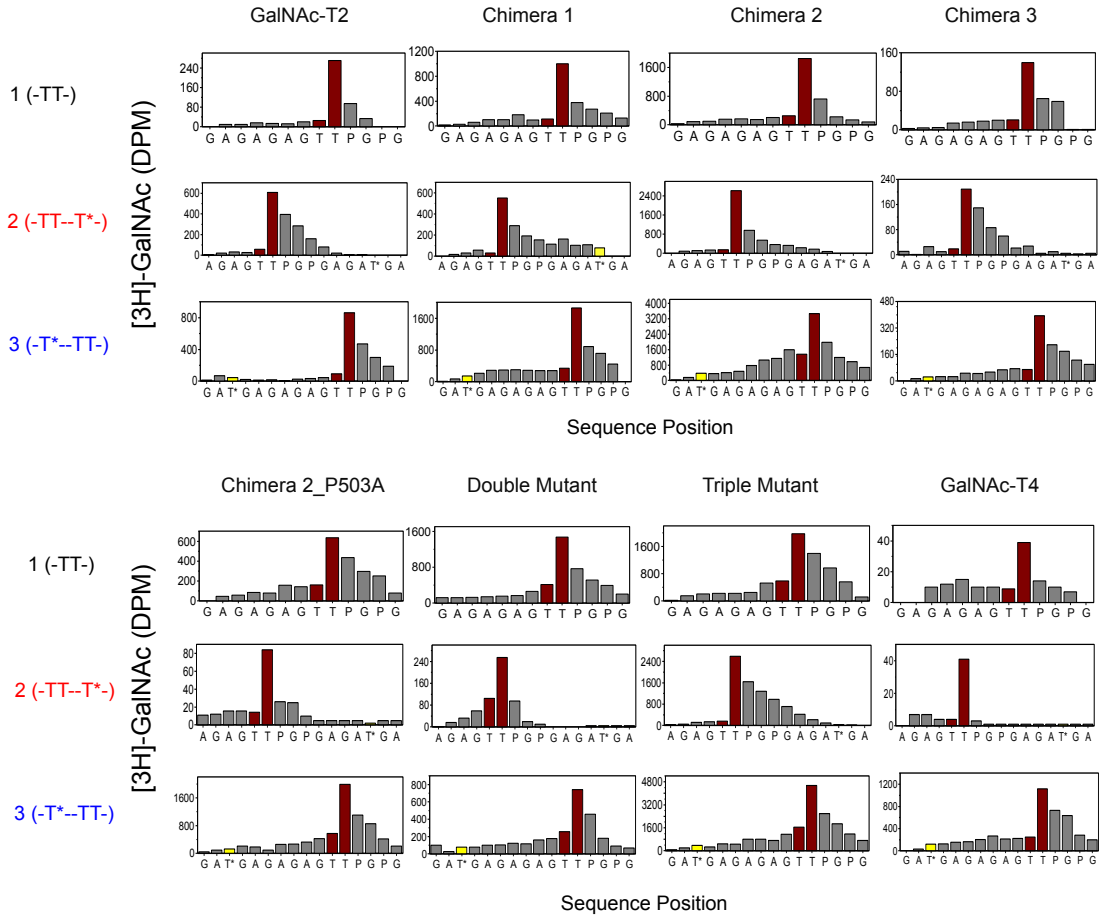
B



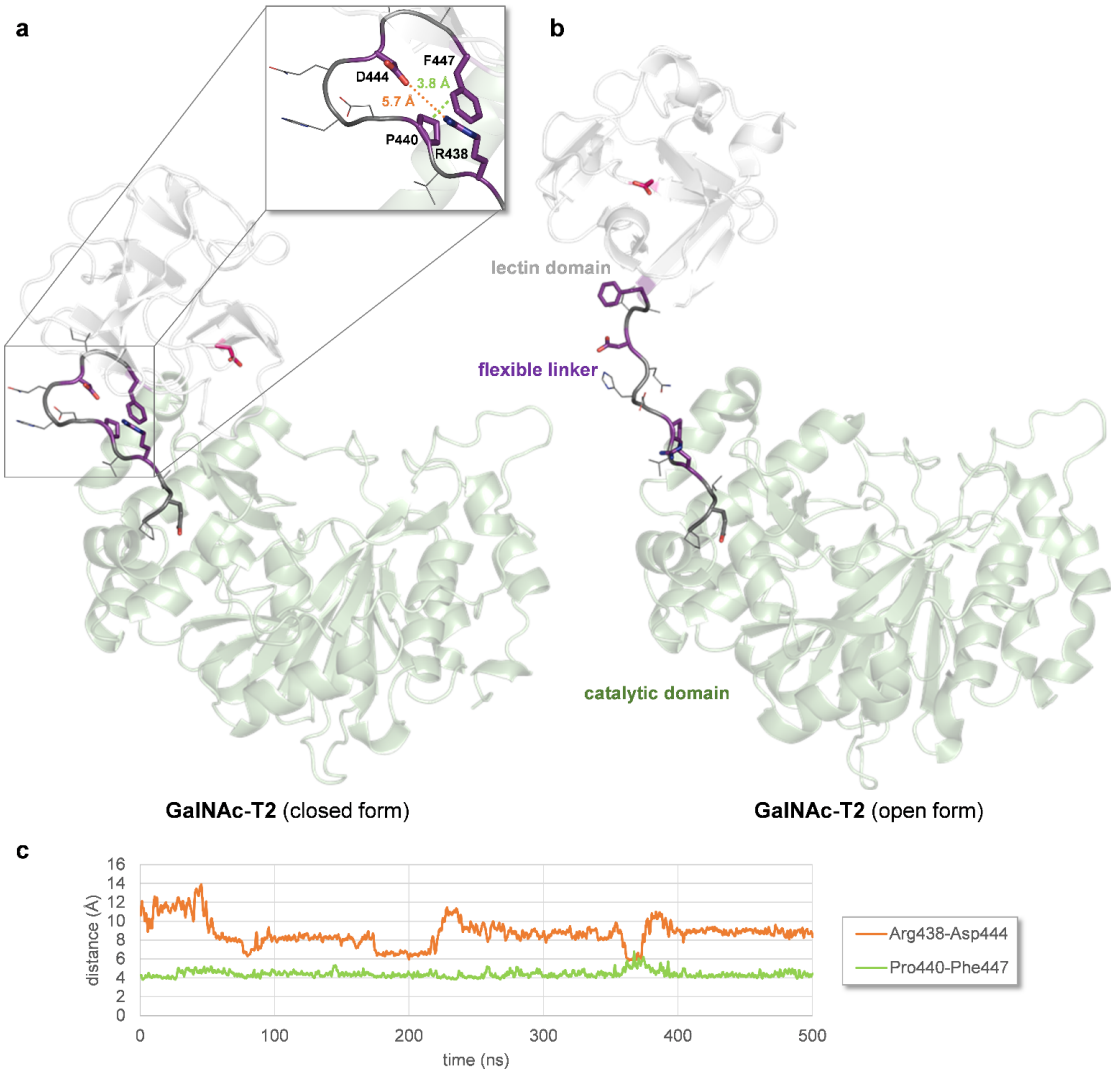
c



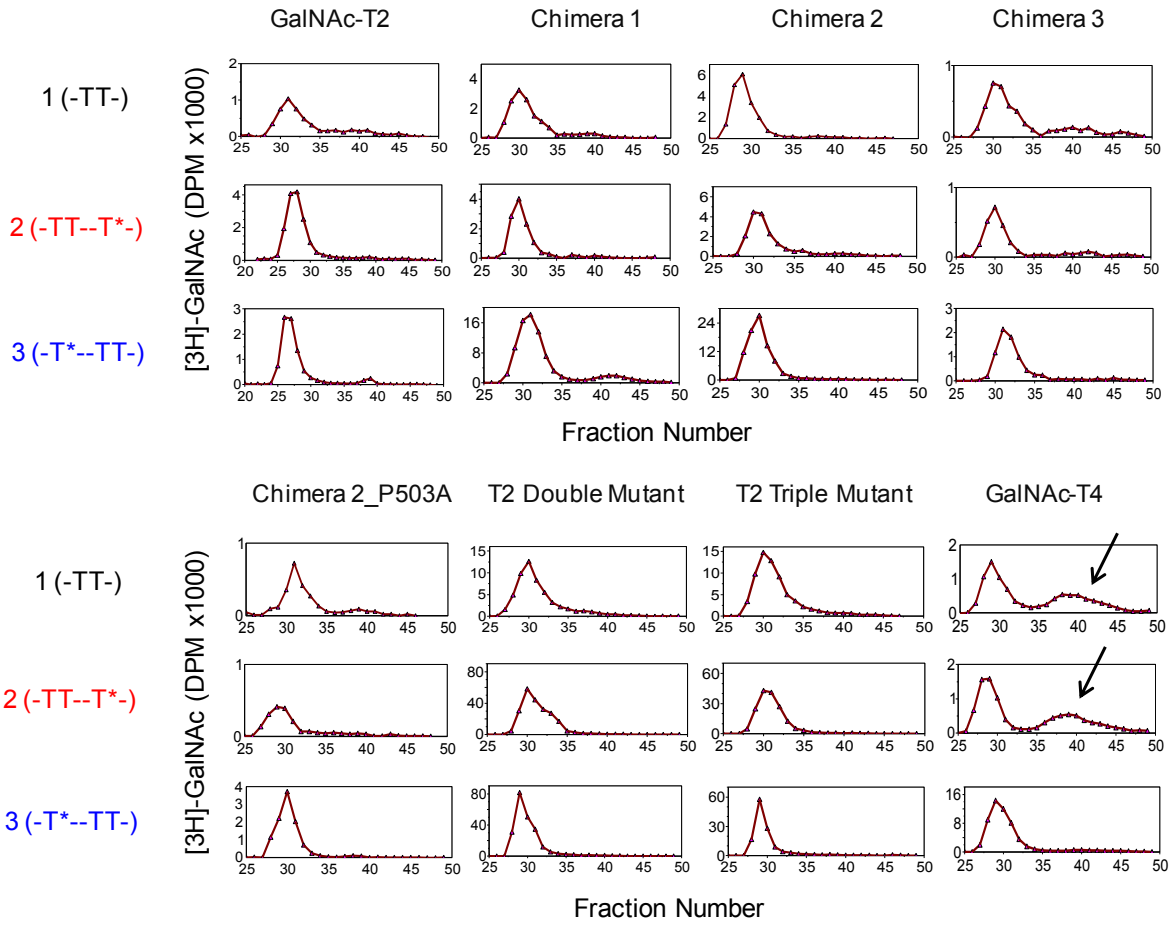
Supplementary Figure 9. Specific activity of the GalNAc-T2 chimeras and mutants and their selected substrate activity ratios. a, Plot of specific activity calculated as described in the **Methods** from 3-5 (except 2 for chimera 2P503A) independent time course reactions. **b**, Plot of monoglycopeptide specific activity normalized to the nonglycosylated substrate , i.e. peptide **3**/**1** and **2**/**1**. **c**, Plot of the ratio of monoglycopeptide **3** over monoglycopeptide **2**, i.e. peptide **3**/**2**. The latter provides an indication of the selectivity of the transferase for the 2 substrates. Black lines represent obtained standard deviations as described in the **Methods**



Supplementary Figure 10. Sites of [3H]-GalNAc incorporation into substrates 1-3 by GalNAc-T2 and its chimeras and mutants as determined by Edman amino acid sequencing. Each Edman sequencing cycle was collected and its [^{3}H]-GalNAc content determined by scintillation counting as described in the **Methods**. The bars represent the [^{3}H]-GalNAc content determined at each cycle, the dark red bars represent the two potential Thr acceptor sites, the yellow bar represents the site of prior GalNAc-Thr glycosylation (T*) and the grey bars denote the remaining residues. Sequencing was occasionally terminated a few cycles after the last potential acceptor Thr. See **Methods** for further details.



Supplementary Figure 11. 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. CH–π 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.



Supplementary Figure 12. Sephadex G10 gel filtration chromatograms of the [^{3}H]-GalNAc products of GalNAc-T2, its chimeras and mutants, and GalNAc-T4. [^{3}H]-GalNAc content at each fraction was measured by scintillation counting. The first peak of [^{3}H]-GalNAc content (fractions ~20-30) represents glycosylated substrate (glyco)peptide, while the second peak (arrow) represents the hydrolysis product free [^{3}H]-GalNAc. See **Methods** for further details.

ARTÍCULO III

Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-Polar Conjugates.

El presente artículo fue publicado en la revista *Chemistry-A European Journal* el año 2016, segundo año del periodo de tesis de la doctoranda. Fue realizado en coautoría con: (a) el Departamento de Síntesis y Estructura de Biomoléculas de la Universidad de Zaragoza (b) el Departamento de Biología Físico-Química del Centro de Investigaciones Biológicas CIB-CSIC, en Madrid y (c) el Servicio de Resonancia Magnética Nuclear de la Universidad de Zaragoza.

A fecha de defensa de la presente tesis doctoral no hay ningún otro coautor de la publicación que no se encuentre en posesión del título de Doctor. Los resultados de la misma fueron defendidos en 2017, en la modalidad clásica de defensa de tesis doctoral, por el Doctor Mattia Ghirardello.

La contribución del doctorando en esta publicación ha consistido en lo siguiente:

- Expresión y purificación de la proteína.
- Cristalización de la proteína y realización de experimentos de cristalización para obtener el complejo proteína-compuesto.
- Obtención y resolución de la estructura proteína-compuesto.
- Determinación y cálculo de las distintas constantes de disociación proteína-compuesto.
- Revisión de la escritura final del artículo.

Se adjunta a continuación la referencia completa de la publicación:

Ghirardello, M., **de las Rivas, M.**, Lacetera, A., Delso, I., Lira-Navarrete, E,. Tejero, T., Martín-Santamaría, S., Hurtado-Guerrero, R., & Merino, P. (**2016**). Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-Polar Conjugates, *Chemistry - A European Journal*, 22, 7215-7224.

ChemPubSoc Europe

DOI: 10.1002/chem.201600467

CHEMISTRY A European Journal Full Paper

Molecular Modeling

Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-Polar Conjugates

Mattia Ghirardello,[a] Matilde de las Rivas,[b] Alessandra Lacetera,[c] Ignacio Delso,[a, d] Erandi Lira-Navarrete,[a] Tomás Tejero,[a] Sonsoles Martín-Santamaría,*[c] Ramón Hurtado-Guerrero,*[b, e, f] and Pedro Merino*[a]

Abstract: The Leloir donors are nucleotide sugars essential for a variety of glycosyltransferases (GTs) involved in the transfer of a carbohydrate to an acceptor substrate, typically a protein or an oligosaccharide. A series of less-polar nucleotide sugar analogues derived from uridine have been prepared by replacing one phosphate unit with an alkyl chain. The methodology is based on the radical hydrophosphonylation of alkenes, which allows coupling of allyl glycosyl compounds with a phosphate unit suitable for conjugation to uridine. Two of these compounds, the GalNAc and galactose derivatives, were further tested on a model GT, such as GalNAc-T2 (an important GT widely distributed in human tissues), to probe that both compounds bound in the medium–high micromolar range. The crystal structure of GalNAc-T2 with the galactose derivative traps the enzyme in an inactive form; this suggests that compounds only containing the β-phosphate could be efficient ligands for the enzyme. Computational studies with GalNAc-T2 corroborate these findings and provide further insights into the mechanism of the catalytic cycle of this family of enzymes.

Introduction

Glycosyltransferases (GTs) are key enzymes responsible for the incorporation of carbohydrates into a variety of acceptor biomolecules, including proteins, lipids, oligosaccharides and different metabolites.[1] The resulting glycoconjugates mediate a wide range of functions from structure and storage to signalling and, as a consequence, they are related to important diseases. Therefore, the chemical manipulation of the activity of GTs could lead to the development of useful therapeutic drugs.[2] Thus, considerable synthetic efforts have been directed toward the preparation of efficient GT inhibitors.[3] Transfer of the sugar residue occurs from an anionic nucleotide sugar donor to the acceptor substrate; it can take place with retention or inversion of configuration at the anomeric centre of the sugar residue.[4] In this context, elucidation of the mechanisms used by GTs has been pursued with much interest.[5] Notably, only nine sugar donors are known to be involved in protein glycosylation, which is the most abundant post-translational modification in nature, in mammals.[6] Six of these sugar donors contain the uridine moiety (Figure 1), which is in agreement with the existence of GTs employing uridine diphosphate (UDP) sugars as the most predominant in nature.[7]

Most of designed inhibitors mimic nucleotide phosphate sugars by incorporating anionic groups to emulate binding of the diphosphate bridge.[8] However, such compounds are not capable of permeating into cells due to their high polarity.[9] To overcome this problem, neutral inhibitors have also been prepared,[10] and derivatives in which the phosphate group was replaced by a different apolar group showed enhanced cell internalisation,[10f,g] although some reduction in binding affinity occurred. Vocadlo and co-workers demonstrated that UDP-5SGlcNAc acted as an inhibitor of O-GT,[11] but it should be generated inside the cell from precursor 5SGlcNAc by using the salvage pathway. Nevertheless, most of the designed compounds resulted in poor biological activity because they did

[a] M. Ghirardello, Dr. I. Delso, Dr. E. Lira-Navarrete, Prof. T. Tejero, Prof. P. Merino
Departamento de Síntesis y Estructura de Biomoléculas
Instituto de Síntesis Química y Catálisis Homogénea (ISQCH)
Universidad de Zaragoza, CSIC, 50009 Zaragoza, Aragón (Spain)
E-mail: pmerino@unizar.es

[b] M. d. l. . Rivas, Dr. R. Hurtado-Guerrero
Instituto de Biocomputación y Fisica de Sistemas Complejos (BIFI)
BIFI-IQFR (CSIC) Joint Unit, Universidad de Zaragoza
50009 Zaragoza (Spain)
E-mail: rhurtado@bifi.es

[c] A. Lacetera, Dr. S. Martín-Santamaría
Departamento de Biología Físico-Química
Centro de Investigaciones Biológicas
CIB-CSIC, Ramiro de Maeztu, 9, 28040 Madrid (Spain)
E-mail: smsantamaria@cib.csic.es

[d] Dr. I. Delso
Servicio de Resonancia Magnética Nuclear
Centro de Química y Materiales de Aragón (CEQMA)
Universidad de Zaragoza, CSIC, Campus San Francisco
50009 Zaragoza (Spain)

[e] Dr. R. Hurtado-Guerrero
Fundación ARAID, 50018 Zaragoza (Spain)

[f] Dr. R. Hurtado-Guerrero
Instituto de Investigaciones Sanitarias de Aragón (IIS-A)
Zaragoza 50009 (Spain)

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under http://dx.doi.org/10.1002/chem.201600467.

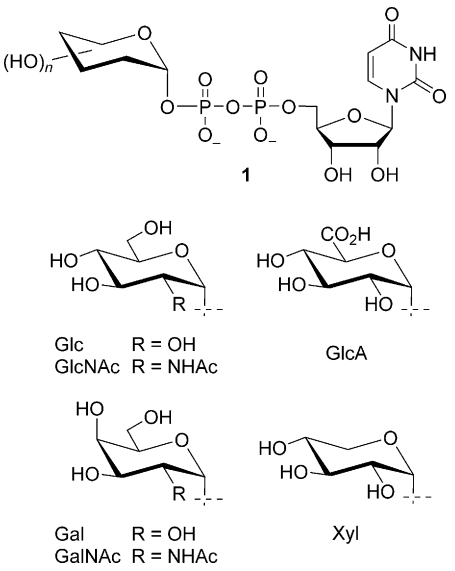


Figure 1. Uridine-based sugar donors.

not incorporate the nucleoside moiety, which diminished recognition and selectivity by the target enzymes.

Herein, we report the design, preparation and binding studies of less-polar uridyl–sugar analogues **6** as suitable GT binders in which the β-phosphate unit has been replaced by an alkyl chain (Figure 2).

Insertion of methylene (**2**)[12] and replacement of the anomeric oxygen by methylene (**3**)[13] or propylene (**4**)[14] groups has already been reported, but compounds **2–4** are still too polar due to the presence of two phosphate units (Figure 2). Compound **3a** was evaluated against three bacterial UDP–galactopyranose mutases and showed only moderate inhibition, although crystallographic studies showed that it was bound to the active site of the enzyme in a novel conformation not observed previously.[13d] As a model enzyme, we have selected GalNAc-T2; a member of the important group of *N*-acetylgalactosaminyl transferases (GalNAcTs E.C. 2.4.1.41) involved in mucin biosynthesis.[15] In particular, a previous computational study with GalNAc-T2 reported by Masgrau and co-workers anticipated that UDP-2′-deoxyGal would not be a good donor substrate, whereas UDP-Gal could be a valid sugar donor in some cases.[16]

Based on these precedents, we designed analogues **6**, in which one phosphate group was replaced by an ethylene group, as suitable GT binders with GalNAc-T2 as a model enzyme. The synthetic strategy is based on coupling between uridine and phosphoalkyl sugars, which are prepared through radical hydrophosphonylation[17] of allyl sugars.

We also demonstrate through tryptophan fluorescence spectroscopy and X-ray studies that two of these compounds bind moderately to the enzyme. In particular, trapping of a galactose moiety containing compound bound to an inactive form of GalNAc-T2 provides hints for improving their affinity through the incorporation of β-phosphate or similar negatively charged groups. Furthermore, in silico studies support the aforementioned findings and that the presence of only one phosphate

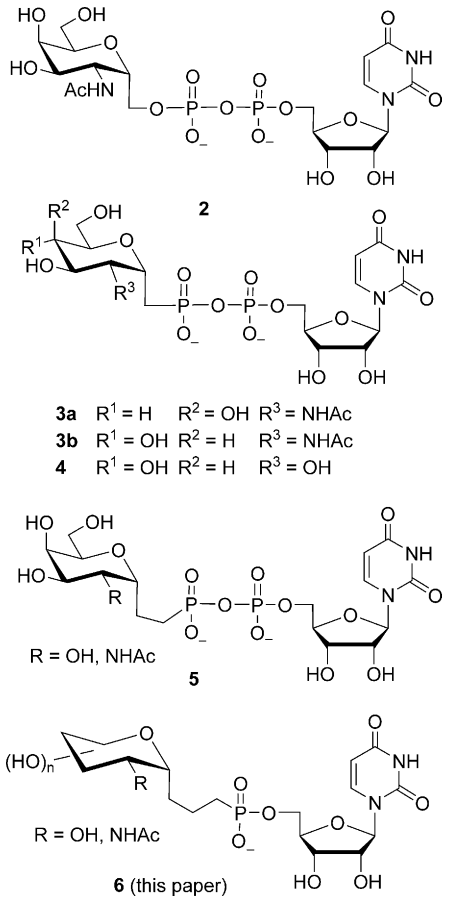


Figure 2. Structures of the C-analogues of uridine dinucleotide sugars discussed herein.

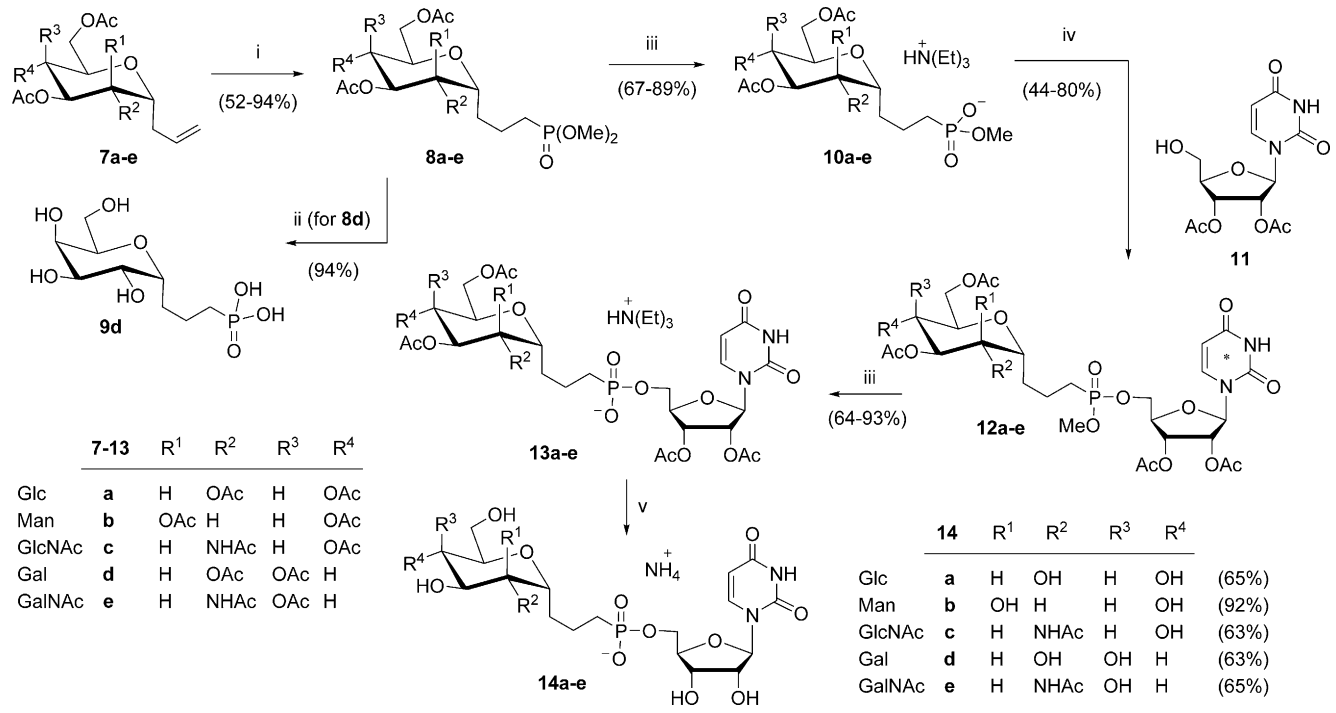
group close to the sugar moiety could be enough for the uridyl–sugar analogue to interact with the protein. These combined studies also provide new evidence of the mechanism of the catalytic cycle of GalNAc-T2.

Results and Discussion

Synthesis of glycomimetics

The starting allyl sugars **7** used herein were prepared from D-glucose for allylsugar **7a**,[18] D-mannose for allylsugar **7b**,[19] D-glucosamine for allylsugar **7c**[20] and D-galactose for allylsugars **7d**[19,21] and **7e** (Scheme 1).[22] Protected diacetyl uridine **11** was prepared from commercially available uridine via the trityl derivative.[23]

Photoinduced free-radical hydrophosphonylation of **7a–e**, following the protocol developed by Dondoni and co-workers,[24] afforded phosphonates **8a–e** with complete regioselectivity and good yields (Scheme 1). Removal of the protecting groups of **8d** under typical conditions[25] led to free phosphonate **9d** in excellent yield. Any attempt to couple compound **9d** with **10** in the presence of *N*,*N*′-dicyclohexylcarbodiimide (DCC)[26] failed, as well as deprotection of both OMe groups. Izumi and co-workers previously reported the condensation of a phosphate monoester with a nucleoside in good yield.[27]



Scheme 1. Synthesis of partially free phosphonates. i) dimethyl phosphite, dieethyl phosphate acetophenone (DPAP), *hν*, 30 min, RT; ii) Me_3SiCl, NaBr, MeCN, 1 h, 40 °C; then 2 h, RT; then 5:1 H_2O/EtOAc, 2 h, RT; then NH_4OH, 2 h, RT; iii) PhSH, Et_3N, dioxane, 4 h, RT; iv) **10** (1.0 equiv), (benzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate (BOP), DMF, *i*$PrEt_2N$, 4 h, RT; v) 25% $NH_4OH_{(aq)}$, MeOH, 7 h, RT.

Therefore, we decided to prepare monoesters **10a–e** through selective deprotection of the phosphonate group by using thiophenol and triethylamine.[28] Condensation of monoesters **10a–d** and diacetyluridine **11** by using BOP and *N*,*N*-diisopropylethylamine (DIEA) in DMF[29] yielded the fully protected nucleotide sugar analogues **12a–e** (as a mixture of diastereomers due to phosphorous chirality) with good yields. Removal of the methyl group of the phosphonate[28] gave **13a–e**.

Finally, deacetylation of **13a–e** with a 25% aqueous solution of ammonium hydroxide afforded free sugar nucleotide analogues **14a–e** in good yields after purification through a Bio-Gel P-2 column (eluted with NH_4HCO_3 in 4:1 H_2O/MeOH), followed by lyophilisation.

Crystallographic and binding studies

To validate whether our compounds might have a biological effect, we selected the GT GalNAc-T2 as an example. This enzyme is part of a large family of 20 members that transfer a GalNAc moiety from UDP–GalNAc onto Ser or Thr residues of proteins in the presence of manganese.[30] This enzyme predominantly uses UDP–GalNAc as the sugar donor, although, in some cases, it might also use UDP–galactose.[16,31] Prompted by this, we determined whether compounds **14d** and **14e**, that contained a galactose and GalNAc moiety, respectively, might bind to this enzyme. Tryptophan fluorescence spectroscopy studies for compound **14d** gave K_d values of (269±40) and (254±20) μM, and those for compound **14e** gave K_d values of (800±10) and (915±180) μM, in the absence and presence of 2 mM Mn^{2+} (Figure 3), respectively; this suggested that interactions with Mn^{2+} for these ligands might not be crucial for their binding. Contrary to the preference of GalNAc-T2 for UDP–GalNAc as the favourite donor substrate for catalysis, the results also suggest that the compound with the galactose moiety (**14d**) binds about 3.5-fold better than that with the GalNAc moiety (**14e**).

Compound **14d** traps the enzyme in an inactive state. To understand the binding mode of these compounds to GalNAc-T2, we solved the crystal structure of GalNAc-T2 in a complex

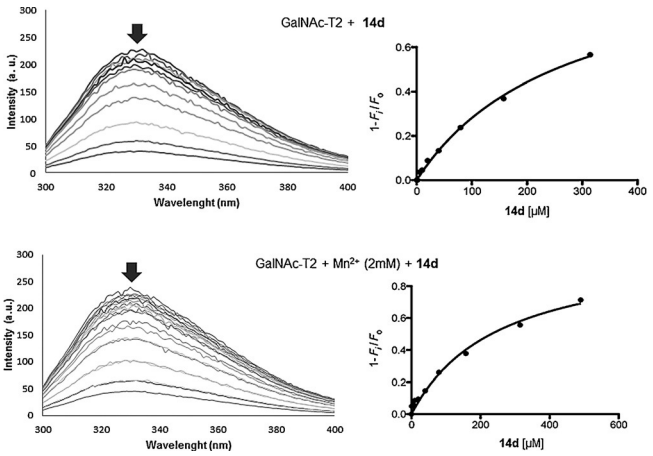


Figure 3. Quenching of intrinsic GalNAc-T2 tryptophan fluorescence measured with increasing concentrations of **14d** in the absence and presence of Mn^{2+}. All data points represent the means ± standard deviations (S.D.) for three measurements. The K_d value for **14d** was determined by fitting fluorescence intensity data against the concentration of **14d**.

with **14d**. Despite numerous attempts with both compounds (**14d** and **14e**) in cocrystallisation and soaking experiments, we only managed to obtain a complex with **14d** through soaking experiments for orthorhombic crystals previously grown with UDP/Mn^{2+}. The structure was solved at 2.07 Å, which allowed us to solve and interpret the density maps (Table S1 in the Supporting Information).

The asymmetric unit, as reported previously, displayed six molecules of GalNAc-T2.[31] Three out of six GalNAc-T2 molecules present in the asymmetric unit clearly showed density maps for the presence of both UDP/Mn^{2+} and **14d** in each active site, whereas the three remaining GalNAc-T2 molecules contained UDP (Figure S1 in the Supporting Information). The crystal structure shows that the typical GT-A fold (catalytic domain) is located in the N-terminal region and the lectin domain is located in the C-terminal region (Figure 4).

Strikingly, GalNAc-T2 in complex with **14d** adopts an inactive-state conformation, as reported previously for other binary complexes containing either glycopeptides or UDP (Figure 4).[15a,31,32] Both active and inactive states are present during the catalytic cycle of this enzyme and are associated with the motion of a flexible loop that can oscillate between closed and open conformations, respectively. In particular, UDP and **14d** are exposed to the solvent due to the open conformation of the flexible loop, whereas UDP-5S-GalNAc is covered by the flexible loop, which adopts a closed conformation, rendering the enzyme in an active state (Figure 4; PDB entry 4D0Z).

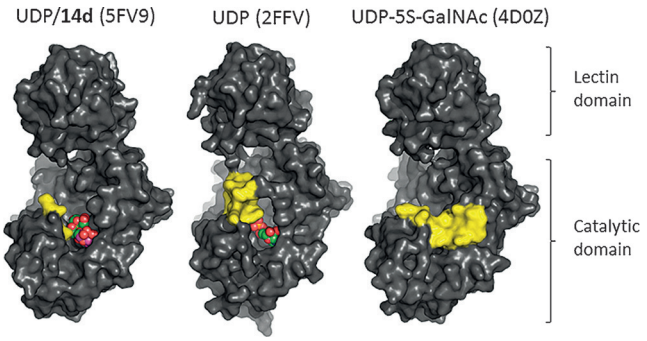


Figure 4. Surface representation of GalNAc-T2 in complex with **14d**/UDP, UDP and UDP-5S-GalNAc. Protein, flexible loop and nucleotides/**14d** are coloured black, dark grey and light grey, respectively. The active and inactive states correspond to closed (4D0Z) and open (5V9 and 2FFV) conformations, respectively.

Compound **14d** induces an inverted uridine conformation and an unusual 4-coordinate Mn^{2+} complex. Closer inspection of the active site reveals that the uridine moiety of both **14d** and UDP adopts an unusual inverted conformation; this was also found earlier in a structure of this enzyme in complex with UDP (Figure 5 and PDB entry 2FFV).[32] This atypical conformation was proposed to represent a final step during the catalytic cycle of GalNAc-T2 in which UDP was ready to exit the enzyme.[31,32] One major difference is apparent between both complexes: the metal is unusually four-coordinate in our crystal structure and five-coordinate in the GalNAc-T2-UDP

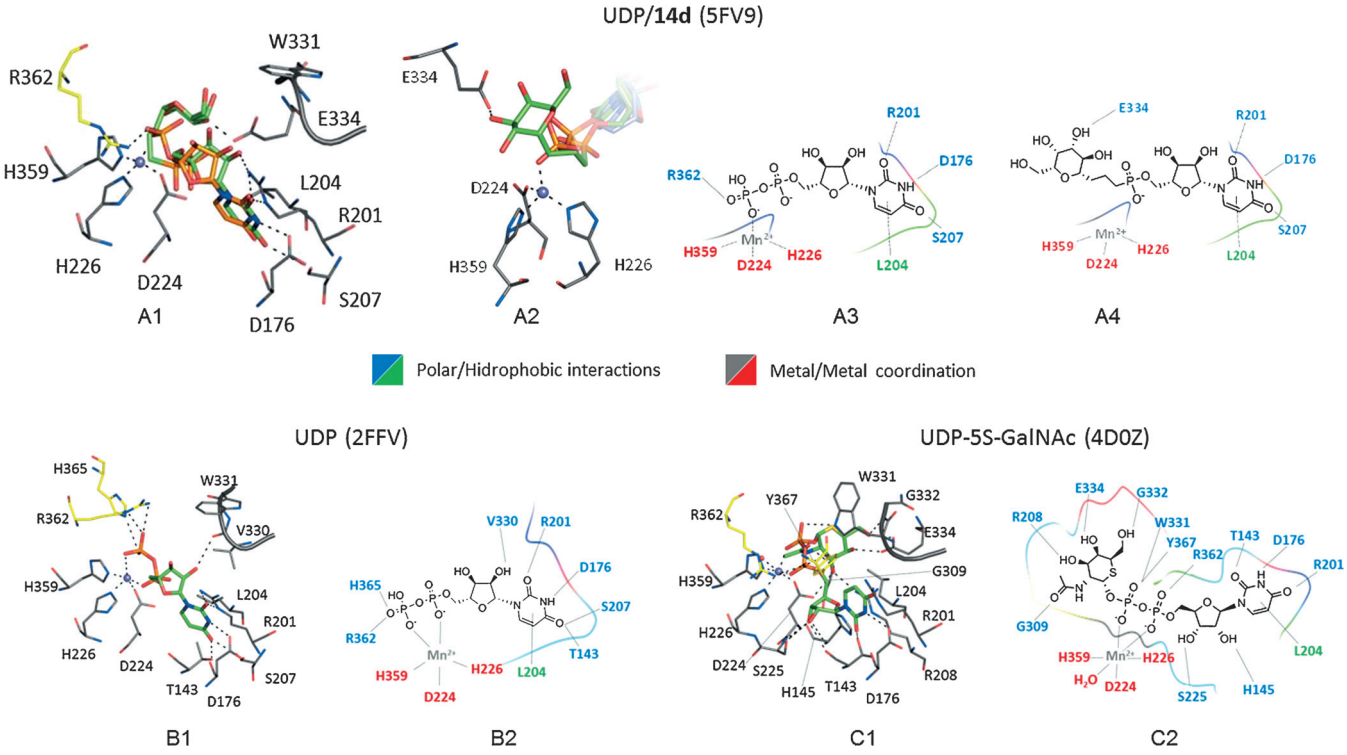


Figure 5. Structural features of the sugar nucleotide binding site for GalNAc-T2 in complex with UDP/**14d**. A1) A close-up view of the active site in complex with UDP-Mn^{2+}/**14d**. A2) Detailed view of metal interactions. Two-dimensional schemes of the interactions between **14d** (A3) and UDP (A4) and the residues of GalNAc-T2. Close-up views of the active site in complex with UDP/Mn^{2+} (B1) and UDP-5S-GalNAc-Mn^{2+} (C1). Two-dimensional schemes of the interactions between the ligands (B2) and residues of GalNAc-T2 (C2).

complex (PDB entry 2FFV; Figure 5). The metal in the former structure is coordinated by the α-phosphate, Asp224, His226 and His359 (Figure 5, A3), whereas in the latter complex the metal is additionally coordinated by the UDP β-phosphate (Figure 5, B1 and B2). However, six-coordinate Mn^{2+}, which requires the previous residues, pyrophosphate and an extra water molecule to form an octahedral complex, is not only the most common coordination in crystal structures of this enzyme, but also the most abundant in proteins in general.[15a,31–34] These results suggest that these UDP–phosphonate scaffolds, with only the α-phosphate, are weaker ligands in terms of binding to GalNAc-T2, but they are still suitable binders for this enzyme.

Previously, different crystal structures obtained from the same type of orthorhombic crystals suggested that UDP might be present in both active and inactive states of the enzyme. On the contrary to our structure, UDP in these crystals always adopted the typical conformation found for UDP-GalNAc, which was required for efficient turnover.[31] Thus, it is tempting to speculate that **14d** has allowed us to trap a conformation in the final step of the catalytic cycle of this enzyme, in which UDP would be ready for immediate departure from the enzyme.

The crystal structure also supports the aforementioned K_d values; this confirms that **14d** does not rely on Mn^{2+} for binding to GalNAc-T2. With regard to direct interactions between the enzyme and nucleotides/**14d** in the different complexes, it is important to note that **14d** shows fewer interactions (Figure 5) than UDP (PDB entry 2FFV) and UDP-5S-GalNAc (PDB entry 4D0Z). It is important to note that Trp331, a critical residue in the catalytic cycle of these enzymes,[32] adopts an "out" (out of the active site) conformation in the complex with **14d**/UDP (Figure 5, A1), which correlates well with the inactive state of GalNAc-T2. On the contrary, Trp331 adopts an "in" (inside the active site) conformation in the complex with UDP-GalNAc that sets up the enzyme in an active state (Figure 5, C1).

Most of the direct interactions between the uridine moiety of **14d** and the enzyme are kept as the UDP from the reported PDB entry 2FFV, whereas a larger number of direct interactions are established between UDP-GalNAc and the enzyme, presumably because the active site is covered by the flexible loop (Figures 4–6). Finally, the phosphonate of **14d** does not interact directly with any residue and the galactose moiety is recognised exclusively by Glu334 (Figure 5, A2); this is likely to be an explanation for the poor binding of this compound to the enzyme. In the docking studies, we found that the Gal moiety adopted several alternative binding poses within the sugar site.

Docking studies

To obtain a 3D picture of the putative binding ability of this family of compounds, docking calculations were performed for analogues **14** in both conformations of GalNAc-T2: closed (active, PDB ID 4D0Z) and open (inactive, PDB ID 2FFV). The calculations were able to predict reasonable docked binding poses for all ligands, although with some slight differences. UDP, UDP-GalNAc and UDP-5S-GalNAc were also docked for comparison purposes to validate the computational protocol. Interestingly, all compounds were predicted to bind in both conformations (open and closed) inside the active site of the enzyme with binding poses similar to those found in the crystallographic structures.

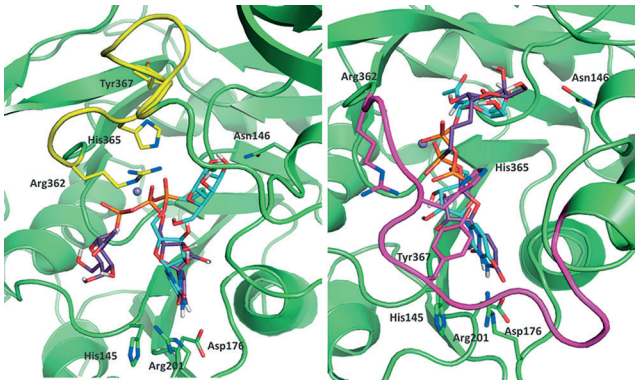


Figure 6. Left: Superposition of docked poses in GalNAc-T2 (PDB-ID 2FFV) of **14d** (cyan) and UDP-GalNAc (purple). The open loop is shown in yellow. Right: Superposition of docked poses in GalNAc-T2 (PDB-ID 4D0Z) of **14d** (cyan) and UDP-GalNAc (purple). The closed loop is shown in magenta.

Predicted binding energies were higher than those of reference diphosphate compounds due to the consequently decreased ionic phosphate–manganese interactions caused by the lack of the β-phosphate group. Nevertheless, docked binding energies were favourable; thus validating the phosphonate scaffold of analogues **14** as a possible platform for the optimisation of GT binders. The best binding energies were always predicted for the active (closed) conformation, which was in agreement with a binding pose with more GalNAc-T2/ligand interactions, although the docking results indicated the ability to bind also to the open conformation, which was preferred in the crystal structure. Both possibilities might coexist in the equilibrium of conformational ensembles. Docking calculations are able to estimate both binding poses, whereas the limitations of X-ray, in this case, have not allowed us to obtain the closed 3D structure. No major differences in the predicted binding energy were found among the phosphonate analogues **14**, which indicated a lack of preferred binding. For compounds **14**, the main contributions to binding are provided by the uridine–enzyme hydrogen bonds, as in the 4D0Z complex, and, to a lesser extent, from ionic phosphonate–Mn^{2+} interactions through one or two of the phosphonate oxygen atoms.

In the case of the active conformation of the GT, an interesting difference was found with regard to binding of the sugar moiety of analogues **14**, which also showed alternative binding poses in which the sugar moiety reached the Ala307 CO group and Asn146 side chain, according to the highly flexible aliphatic chain. For the inactive conformation of GalNAc-T2, the docking calculations placed the synthetic analogues in the equivalent region to that of the active conformation (Figure 6).

The uridine moiety is observed to rotate in a similar way to the crystallographic pose, and the overall conformation of the ligand was found to adopt a more bended conformation, with the sugar moiety placed in front of the uridine plane. For comparison purposes we performed docking calculations for **14d** to our crystallographic structure (PDB 5V9, chain "B ") finding similar results between the crystallographic pose for **14d** and the docked pose for UDP-GalNAc.

We conclude, from the overall docking study, that compounds **14** are putative GalNAc-T2 binders in which the uridine moiety plays a major role in binding, in agreement with the crystallographic binding poses, whereas the phosphonate group and sugar moiety provide additional interactions; there are no differences in the types of sugar.

Targeted molecular dynamics (tMD) simulations

To shed light on the conformational changes that the protein needs to undergo between the inactive and active conformations (opening/closing process), and to analyse the required interactions to host the substrate, targeted molecular dynamics (tMD) simulations were undertaken for the GalNAc-T2/UDP-GalNAc complex. The simulation was started from the complex in the inactive open conformation (from the docked binding pose) towards the complex in the active closed conformation (from the docked binding pose of UDP-GalNAc, very similar to the crystallographic complex with UDP-5S-GalNAc; Figure 7). The analysis of the root-mean-square fluctuation (Figure S2 in the Supporting Information) during the tMD simulation time identified a main fluctuation that corresponded to the flexible loop Arg362–Ser373 with a maximum value for Pro370. This main change corresponds to the folding of the loop over the substrate, sealing it in the active site, and leading to the closed conformation. Additionally to the Asp224, His226 and His359 side chains, the Mn^{2+} ion was found to coordinate to two phosphate oxygen atoms (Figure S3 in the Supporting Information). During the progression of the tMD simulation, one water molecule entered into the coordination sphere, according to a six-coordinate Mn^{2+} model (Figure S4 in the Supporting Information).

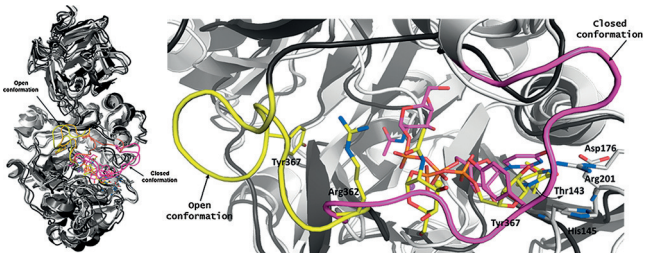


Figure 7. Superposition of five structures of the GalNAc-T2/UDP-GalNAc complex along the tMD simulation of the transition from the open (black, loop in yellow) to the closed (white, loop in magenta) conformation. Intermediate protein structures are shown as a gradient from yellow (starting geometry) to orange (intermediate geometries) to magenta (final target geometry). The ligand UDP-GalNAc is shown in yellow for the open complex and in magenta for the closed complex. Left: Full perspective. Right: Detailed image showing some representative residues.

Several functional groups from GalNAc-T2 establish interactions, especially those involving the uridine moiety. In particular, the NH in position 3 of the uracyl ring establishes a stable hydrogen bond along the simulation with the Asp176 side chain (Figure 7, right, and Figure S5 in the Supporting Information). The CO groups at positions 2 and 4 of the uracyl ring establish alternating hydrogen bonds with Arg201 and Thr143 side chains (Figure 7, right, and Figure S6 in the Supporting Information) during the rotation of the uridine. This rotation allows the torsional change undergone by the diphosphate linker. Also, stacking interactions between the His145 side chain and the uracyl moiety are identified. Most of these interactions are found in the crystallographic structures, in agreement with a specific anchorage for the uracil moiety. On the contrary, the flexibility of the diphosphate linker allows GalNAc to rotate and move during the closing process, which changes the interactions along the simulation to adapt to the new surrounding environment.

For example, in the inactive conformation, GalNAc is involved in two hydrogen bonds: one between the endocyclic GalNAc oxygen and the Arg362 backbone NH, and another one between the N-acetylamido CO group and the Arg362 guanidinium group. After closing, these interactions with Arg362 are lost and new hydrogen bonds are established: one between GalNAc OH-3 and the backbone NH of Gly309, and another one between OH-4 from the GalNAc moiety and the Ala307 CO group. Moreover, new CH–π interactions form between the ribose and the Tyr367 indole ring.

In summary, tMD results may give us an approximation of the binding pose changes that the UDP-GalNAc undergoes during the closing/opening process when bound to the GT. Specific interactions for GalNAc during the closing/opening process may be useful to understand the mechanism of the catalytic cycle of this enzyme, and to help in the design of selective GT binders. In fact, we provide a plausible explanation for the role of the N-acetylamido group in binding, which supports its higher affinity to the enzyme in comparison to UDP-Gal.

Conclusion

GalNAc-T2 has been used as a prototype GT for studying the binding mode of less-polar glycomimetics. Easily accessible sugar nucleotides analogues, in which one phosphate group has been replaced by an apolar alkyl chain, are weaker binders of GalNAc-T2 than the natural substrate UDP-GalNAc, as demonstrated by combined structural and computational studies. However, compound **14d** is capable of trapping the enzyme in an inactive form. Interestingly, docking studies reveal that both compounds can bind to the inactive and active forms of GalNAc-T2, which suggests a more complex scenario. The crystal structure provides additional useful information to design better compounds that might inhibit GalNAc-T2. In particular, it is inferred from the crystal structure that the β-phosphate is required for binding to Mn^{2+}, and thus, replacing the α-phosphate by a propylene group, while keeping the phosphate or

similar groups in the β position, should maintain or improve the binding of these compounds.

Finally, tMD simulations of the complex of GalNAc-T2 with UDP-GalNAc has provided new insights into substrate recognition events involved in the preliminary steps of the catalytic cycle of GalNAc-T2. We have simulated the change from the open, inactive conformation to the closed, active conformation to identify the overall change, with regard to the conformation and interactions with the protein, as experienced by the substrate to finally lead to the active, closed complex. This transition involves the rotation of the uridine and GalNAc moieties and a conformational change of a very dynamic flexible loop. Our studies, combining the synthesis of UDP-GalNAc-based analogues with structural and computational techniques, allowed us a better understanding of the substrate recognition and catalytic cycle of GalNAc-T2.

Experimental Section

General methods

The reaction flasks and other glass equipment were heated in an oven at 130 °C overnight and assembled in a stream of argon. All reactions were monitored by TLC on silica gel 60 F254; the positions of the spots were detected by $\lambda = 254$ nm UV light or by spraying with either a 5% solution of phosphomolybdic acid in ethanol or Mostain solution. Column chromatography was carried out in a Buchi 800 MPLC system or a Combiflash apparatus by using silica gel 60 microns and with solvents that were distilled prior to use. Melting points were uncorrected. Purification by semi-preparative HPLC (column Atlantis® DC18 5 μm, 19×100 mm, flow: 12.5 mL min^{-1}) was carried out in a Waters 515 pump system with photodiode array (PDA) detection. ^{1}H and ^{13}C NMR spectra were recorded on Bruker Avance 400 MHz or AVANCE II 300 MHz instruments in the stated solvent. Chemical shifts are reported in ppm (δ) relative to $CHCl_3$ ($\delta = 7.26$) in $CDCl_3$. NMR assignments were made by using standard 2D experiments. For numbering of NMR data, see Figure 8. Optical rotations were recorded on a JASCO DIP-370 polarimeter. Elemental analyses were performed on a PerkinElmer 240B microanalyzer or with a PerkinElmer 2400 instrument. High-resolution mass spectra were recorded on a QToF spectrometer equipped with an ESI source (microTOF-Q, Bruker Daltonik) by using sodium formate as an external reference. Hydrophosphonylation was carried out in a glass vial (diameter: 1 cm; wall thickness: 0.65 mm), closed with a natural rubber septum, located 2.5 cm away from the household UVA lamp apparatus equipped with four 15 W tubes (1.5×27 cm each).

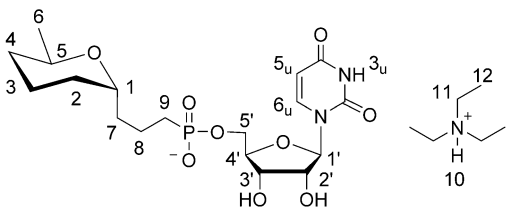


Figure 8. Numbering used for NMR spectroscopy assignments of compounds.

Hydrophosphonylation of allylglycosides 7 a–e: General procedure

DPAP (37.8 mg, 0.148 mmol) was added to a well-stirred solution of allyl glycoside **7** (0.295 mmol) in dimethyl phosphite (2.7 mL, 29.5 mmol). The solution was irradiated at room temperature under stirring for 30 min and then concentrated. The residue was purified by column chromatography on silica gel (EtOAc/MeOH, 1:0 to 95:5 containing 0.2% of Et_3N) to give **8** as an oil.

Dimethyl [(2,3,4,6-tetra-*O*-acetyl-α-D-glucopyranosyl)propyl]-phosphonate (8 a): Yield: 367 mg, 94%; $[\alpha]_D^{27} = +53$ ($c = 1.0$, $CHCl_3$); ^{1}H NMR (400 MHz, $CDCl_3$): $\delta = 5.32$ (dd, $J = 9.5$, 8.7 Hz, 1H; H-3), 5.09 (dd, $J = 9.5$, 5.8 Hz, 1H; H-2), 4.99 (dd, $J = 9.6$, 8.7 Hz, 1H; H-4), 4.27 (dd, $J = 12.0$, 5.7 Hz, 1H; H-6), 4.18–4.11 (m, 1H; H-1), 4.09 (dd, $J = 12.0$, 2.6 Hz, 1H; H-6), 3.83 (ddd, $J = 9.6$, 5.7, 2.6 Hz, 1H; H-5), 3.77 (d, $J = 11.0$ Hz, 6H; OCH_3), 2.08 (s, 3H; Ac), 2.04 (s, 3H; Ac), 2.01 (s, 3H; Ac), 2.00 (s, 3H; Ac), 1.94–1.51 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, $CDCl_3$): $\delta = 170.1$, 170.0, 169.9, 169.8, 72.5, 70.5 (2 C), 69.0 (2 C), 62.7, 52.6 (d, $J = 6.3$ Hz), 26.6 (d, $J = 15.5$ Hz), 25.0 (d, $J = 140.5$ Hz), 20.9 (2 C), 20.8 (2 C), 18.7 ppm (d, $J = 5.0$ Hz); ^{31}P NMR (162 MHz, $CDCl_3$): $\delta = 32.0$ ppm; elemental analysis calcd (%) for $C_{19}H_{31}O_{12}P$: C 47.31, H 6.48, P 6.42; found: C 47.56, H 6.30, P 6.24.

Dimethyl [(2,3,4,6-tetra-*O*-acetyl-α-D-mannopyranosyl)propyl]-phosphonate (8 b): Yield: 254 mg, 42%; $[\alpha]_D^{28} = +7$ ($c = 0.9$, $CHCl_3$); ^{1}H NMR (400 MHz, $CDCl_3$): $\delta = 5.22$ (dd, $J = 8.9$, 3.3 Hz, 1H; H-3), 5.17 (dd, $J = 8.9$, 8.5 Hz, 1H; H-4), 5.15 (dd, $J = 3.3$, 3.2 Hz, 1H; H-2), 4.36 (dd, $J = 12.0$, 6.5 Hz, 1H; H-6), 4.10 (dd, $J = 12.0$, 4.0 Hz, 1H; H-6), 3.95 (ddd, $J = 10.6$, 3.8, 3.2 Hz, 1H; H-1), 3.87 (ddd, $J = 8.5$, 6.5, 4.0 Hz, 1H; H-5), 3.74 (d, $J = 9.5$ Hz, 6H; OCH_3), 2.12 (s, 3H; Ac), 2.11 (s, 3H; Ac), 2.06 (s, 3H; Ac), 2.03 (s, 3H; Ac), 1.92–1.60 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, $CDCl_3$): $\delta = 170.7$, 170.5, 170.0, 169.8, 73.8, 70.6 (2 C), 68.8, 66.9, 62.3, 52.2 (d, $J = 9.8$ Hz), 29.2 (d, $J = 15.2$ Hz), 24.0 (d, $J = 140.5$ Hz), 20.8, 20.7, 20.6, 20.5, 18.5 ppm (d, $J = 4.5$ Hz); ^{31}P NMR (162 MHz, $CDCl_3$): $\delta = 31.8$ ppm; elemental analysis calcd (%) for $C_{19}H_{31}O_{12}P$: C 47.31, H 6.48, P 6.42; found: C 47.22, H 6.38, P 6.51.

Dimethyl [(2-acetamido-2-deoxy-3,4,6-tri-*O*-acetyl-α-D-glucopyranosyl)propyl]phosphonate (8 c): Yield: 213 mg, 73%; $[\alpha]_D^{28} = +29$ ($c = 0.9$, $CHCl_3$); ^{1}H NMR (400 MHz, $CDCl_3$): $\delta = 6.00$ (d, $J = 8.5$ Hz, 1H; N*H*Ac), 4.97 (dd, $J = 8.5$, 7.5 Hz, 1H; H-3), 4.87 (dd, $J = 7.5$, 7.0 Hz, 1H; H-4), 4.26 (dd, $J = 12.0$, 5.6 Hz, 1H; H-6), 4.19 (dt, $J = 8.5$, 4.7 Hz, 1H; H-2), 4.09–4.01 (m, 2H; H-1, H-6), 3.79 (ddd, $J = 7.0$, 5.6, 3.5 Hz, 1H; H-5), 3.67 (d, $J = 10.6$ Hz, 6H; OCH_3), 2.04 (s, 3H; Ac), 2.01 (s, 3H; Ac), 2.00 (s, 3H; Ac), 1.91 (s, 3H; Ac), 1.79–1.44 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, $CDCl_3$): $\delta = 170.9$, 170.6, 169.9, 169.0, 71.3, 70.2, 70.1, 68.5, 61.7, 52.3 (d, $J = 6.7$ Hz), 52.3 (d, $J = 6.7$ Hz), 50.7, 27.3 (d, $J = 14.3$ Hz), 24.0 (d, $J = 141.5$ Hz), 23.1, 20.8, 20.7, 20.6, 18.5 ppm (d, $J = 4.9$ Hz); ^{31}P NMR (162 MHz, $CDCl_3$): $\delta = 34.2$ ppm; elemental analysis calcd (%) for $C_{19}H_{32}NO_{11}P$: C 47.40, H 6.70, N 2.91, P 6.43; found: C 47.31, H 6.55, N 3.31, P 6.35.

Dimethyl [(2,3,4,6-tetra-*O*-acetyl-α-D-galactopyranosyl)propyl]-phosphonate (8 d): Yield: 561 mg, 78%; $[\alpha]_D^{28} = +60$ ($c = 1.0$, $CHCl_3$); ^{1}H NMR (400 MHz, $CDCl_3$): $\delta = 5.36$ (dd, $J = 3.4$, 2.1 Hz, 1H; H-4), 5.21 (dd, $J = 9.5$, 5.1 Hz, 1H; H-2), 5.14 (dd, $J = 9.5$, 3.4 Hz, 1H; H-3), 4.22–4.13 (m, 2H; H-1, H-6), 4.10–3.99 (m, 2H; H-5, H-6), 3.70 (d, $J = 10.7$ Hz, 6H; OCH_3), 2.08 (s, 3H; Ac), 2.04 (s, 3H; Ac), 2.03 (s, 3H; Ac), 1.98 (s, 3H; Ac), 1.81–1.45 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, $CDCl_3$): $\delta = 170.3$, 170.1, 170.0, 169.9, 71.5, 68.4, 68.3, 68.1, 68.7, 61.5, 52.3 (d, $J = 11.1$ Hz), 52.2 (d, $J = 11.1$ Hz), 26.3 (d, $J = 15.0$ Hz), 24.2 (d, $J = 141.0$ Hz), 20.7, 20.7, 20.6, 20.6, 19.1 ppm (d, $J = 4.4$ Hz); ^{31}P NMR (162 MHz, $CDCl_3$): $\delta = 34.0$ ppm;

elemental analysis calcd (%) for $C_{19}H_{31}O_{12}P$: C 47.31, H 6.48, P 6.42; found: C 47.45, H 6.37, P 6.29.

Dimethyl [(2-acetamido-2-deoxy-3,4,6-tri-*O*-acetyl-α-D galactopyranosyl)propyl]phosphonate (8e): Yield: 82.8 mg, 58%; $[\alpha]_D^{28} = 31$ ($c = 1.0$, $CHCl_3$); ^{1}H NMR (400 MHz, $CDCl_3$): $\delta = 6.14$ (d, $J = 8.3$ Hz, 1H; N*H*Ac), 5.30 (t, $J = 3.0$ Hz, 1H; H-4), 5.10 (dd, $J = 9.7$, 3.0 Hz, 1H; H-3), 4.52–4.42 (m, 1H; H-2), 4.29–4.14 (m, 2H; H-1, H-6), 4.07 (dd, $J = 11.6$, 4.9 Hz, 1H; H-6), 3.99 (m, 1H; H-5), 3.72 (d, $J = 10.8$ Hz, 6H; OCH_3), 2.10 (s, 3H; Ac), 1.05 (s, 3H; Ac), 2.03 (s, 3H; Ac), 1.96 (s, 3H; Ac), 1.79–1.44 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, $CDCl_3$): $\delta = 170.7$ (2 C), 170.5, 170.2, 71.8, 68.5, 68.3, 67.0, 61.6, 52.3 (d, $J = 6.7$ Hz), 48.6, 26.7, 24.0 (d, $J = 140.5$ Hz), 23.0, 20.9, 20.7, 20.7, 18.8 ppm (d, $J = 5.2$ Hz); ^{31}P NMR (121 MHz, $CDCl_3$): $\delta = 27.0$ ppm; elemental analysis calcd (%) for $C_{19}H_{32}NO_{11}P$: C 47.40, H 6.70, N 2.91, P 6.43; found: C 47.58, H 6.82, N 3.05, P 6.34.

Deprotection of compounds 13a–e: General procedure

A stirred solution of the corresponding derivative **13** (0.034 mmol) in MeOH (1.7 mL) was treated with a 25% solution of NH_4OH (1.7 mL) and stirred for 7 h at room temperature. The solution was concentrated in vacuo, and the residue was purified on a Bio-Gel P-2 column (1×70 cm), eluted with 100 mmol NH_4HCO_3 in 4:1 H_2O/MeOH, and lyophilised to give pure **14** as a foam.

Ammonium uridin-5′-yl [(α-D-glucopyranosyl)-propyl]phosphonate (14a): Yield: 13.1 mg, 65%; m.p. 153–155 °C; $[\alpha]_D^{27} = +31$ ($c = 0.7$, H_2O); ^{1}H NMR (400 MHz, D_2O, 25 °C): $\delta = 8.00$ (d, $J = 8.1$ Hz, 1H; H-6_u), 6.00 (d, $J = 4.7$ Hz, 1H; H-1′), 5.99 (d, $J = 8.1$ Hz, 1H; H-5_u), 4.39 (dd, $J = 5.0$, 4.7 Hz, 1H; H-2′), 4.36 (t, $J = 5.0$ Hz, 1H; H-3′), 4.32–4.29 (m, 1H; H-4′), 4.18 (ddd, $J = 11.8$, 4.2, 2.5 Hz, 1H; H-5′), 4.10 (ddd, $J = 11.8$, 5.3, 2.9 Hz, 1H; H-5), 4.03 (ddd, $J = 11.4$, 5.9, 2.8 Hz, 1H; H-1), 3.84 (dd, $J = 12.2$, 2.3 Hz, 1H; H-6), 3.75–3.63 (m, 3H; H-6, H-2, H-3), 3.53 (ddd, $J = 9.8$, 5.6, 2.3 Hz, 1H; H-5), 3.37 (dd, $J = 9.8$, 8.9 Hz, 1H; H-4), 1.88–1.50 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, D_2O, 25 °C): $\delta = 166.6$, 152.0, 141.6, 102.4, 88.8, 83.2 (d, $J = 16.2$ Hz), 75.4, 73.9, 73.3, 72.4, 71.3, 70.4, 69.6, 62.6 (d, $J = 5.1$ Hz), 61.1, 25.7 (d, $J = 134.5$ Hz), 25.0 (d, $J = 16.8$ Hz), 19.1 ppm (d, $J = 4.3$ Hz); ^{31}P NMR (121 MHz, D_2O, 25 °C): $\delta = 28.4$ ppm; elemental analysis calcd (%) for $C_{18}H_{32}N_3O_{13}P$: C 40.84, H 6.09, N 7.94, P 5.85; found: C 40.77, H 6.18, N 8.14, P 5.97

Ammonium uridin-5′-yl [(α-D-mannopyranosyl)propyl]phosphonate (14b): Yield: 16.6 mg, 92%; m.p. 134–136 °C; $[\alpha]_D^{28} = +9$ ($c = 0.9$, H_2O); ^{1}H NMR (400 MHz, D_2O, 25 °C): $\delta = 7.96$ (d, $J = 8.1$ Hz, 1H; H-6_u), 5.96 (d, $J = 4.4$ Hz, 1H; H-1′), 5.94 (d, $J = 8.1$ Hz, 1H; H-5_u), 4.35 (dd, $J = 5.0$, 4.4 Hz, 1H; H-2′), 4.32 (t, $J = 5.0$ Hz, 1H; H-3′), 4.28–4.24 (m, 1H; H-4′), 4.14 (ddd, $J = 11.7$, 4.1, 2.4 Hz, 1H; H-5′), 4.06 (ddd, $J = 11.7$, 5.3, 2.9 Hz, 1H; H-5′), 3.95–3.89 (m, 1H; H-1), 3.87 (dd, $J = 3.3$, 1.8 Hz, 1H; H-2), 3.82 (dd, $J = 12.2$, 2.3 Hz, 1H; H-6), 3.81 (dd, $J = 9.4$, 4.3 Hz, 1H; H-3), 3.70 (dd, $J = 12.2$, 6.0 Hz, 1H; H-6), 3.62 (t, $J = 9.4$ Hz, 1H; H-4), 3.50 (ddd, $J = 9.4$, 6.0, 2.3 Hz, 1H; H-5), 1.94–1.46 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, D_2O, 25 °C): $\delta = 166.3$, 151.8, 141.6, 102.3, 88.8, 83.2 (d, $J = 7.9$ Hz), 78.0, 73.9, 73.4, 71.5, 70.8, 69.5, 67.3, 62.6 (d, $J = 5.2$ Hz), 61.3, 28.5 (d, $J = 16.6$ Hz), 25.5 (d, $J = 134.8$ Hz), 19.5 ppm (d, $J = 4.4$ Hz); ^{31}P NMR (121 MHz, D_2O, 25 °C): $\delta = 28.4$ ppm; elemental analysis calcd (%) for $C_{18}H_{32}N_3O_{13}P$: C 40.84, H 6.09, N 7.94, P 5.85; found: C 40.86, H 6.23, N 7.89, P 5.74.

Ammonium uridin-5′-yl [(2-acetamido-2-deoxy-α-D-glucopyranosyl)propyl]phosphonate (14c): Yield: 12.2 mg, 63%; m.p. 161–163 °C; $[\alpha]_D^{25} = +38$ ($c = 1.1$, H_2O); ^{1}H NMR (400 MHz, D_2O, 25 °C): $\delta = 7.85$ (d, $J = 8.1$ Hz, 1H; H-6_u), 5.84 (d, $J = 4.4$ Hz, 1H; H-1′), 5.83 (d, $J = 8.1$ Hz, 1H; H-5_u), 4.24 (dd, $J = 4.9$, 4.4 Hz, 1H; H-2′), 4.20 (t, $J = 4.9$ Hz, 1H; H-3′), 4.17–4.13 (m, 1H; H-4′), 4.06–3.99 (ddd, $J = 11.8$, 4.3, 2.5 Hz, 1H; H-5′), 3.98–3.89 (m, 2H; H-5′, H-1), 3.84 (dd, $J = 10.7$, 5.8 Hz, 1H; H-2), 3.69 (dd, $J = 12.2$, 2.2 Hz, 1H; H-6), 3.64–3.57 (m, 2H; H-3, H-6), 3.43–3.36 (ddd, $J = 9.8$, 5.5, 2.2 Hz, 1H; H-5), 3.29 (dd, $J = 9.8$, 8.8 Hz, 1H; H-4), 1.92 (s, 3H; NAc), 1.79–1.31 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, D_2O, 25 °C): $\delta = 174.3$, 166.3, 151.8, 141.5, 102.3, 88.8, 83.2 (d, $J = 7.9$ Hz), 73.9, 73.3, 72.4, 70.8, 70.7, 69.5, 62.5 (d, $J = 5.2$ Hz), 61.0, 53.4, 25.7 (d, $J = 16.8$ Hz), 25.6 (d, $J = 134.7$ Hz), 21.8, 19.1 ppm (d, $J = 4.3$ Hz); ^{31}P NMR (162 MHz, D_2O, 25 °C): $\delta = 28.3$ ppm; elemental analysis calcd (%) for $C_{20}H_{35}N_4O_{13}P$: C 42.11, H 6.18, N 9.82, P 5.43; found: C 42.28, H 6.23, N 9.71, P 5.65.

Ammonium uridin-5′-yl [(α-D-galactopyranosyl)propyl]phosphonate (14d): Yield: 11.3 mg, 63%; m.p. 152–155 °C; $[\alpha]_D^{29} = +34$ ($c = 0.9$, H_2O); ^{1}H NMR (400 MHz, D_2O, 25 °C): $\delta = 7.93$ (d, $J = 8.1$ Hz, 1H; H-6_u), 5.94 (d, $J = 8.1$ Hz, 1H; H-5_u), 5.94 (d, $J = 4.6$ Hz, 1H; H-1′), 4.27 (t, $J = 4.6$ Hz, 1H; H-2′), 4.23 (t, $J = 4.6$ Hz, 1H; H-3′), 4.20–4.16 (m, 1H; H-4′), 4.05 (ddd, $J = 11.8$, 4.0, 2.5 Hz, 1H; H-5′), 3.97 (ddd, $J = 11.8$, 5.2, 2.8 Hz, 1H; H-5′), 3.93 (dd, $J = 6.1$, 2.8 Hz, 1H; H-1), 3.87 (dd, $J = 9.7$, 5.8 Hz, 1H; H-3), 3.84 (dd, $J = 3.5$, 0.7 Hz, 1H; H-2), 3.68 (dd, $J = 9.9$, 3.5 Hz, 1H; H-4), 3.54–3.63 (m, 3H; H-5, H-6), 1.76–1.35 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, D_2O, 25 °C): $\delta = 166.2$, 151.7, 141.6, 102.2, 88.9, 88.2 (d, $J = 8.0$ Hz), 75.1, 73.9, 71.5, 69.7, 69.4, 69.1, 68.3, 62.3 (d, $J = 5.2$ Hz), 61.1, 25.7 (d, $J = 134.8$ Hz), 24.9 (d, $J = 17.1$ Hz), 19.3 ppm (d, $J = 4.3$ Hz); ^{31}P NMR (162 MHz, D_2O, 25 °C): $\delta = 28.6$ ppm; elemental analysis calcd (%) for $C_{18}H_{32}N_3O_{13}P$: C 40.84, H 6.09, N 7.94, P 5.85; found: C 40.82, H 5.94, N 7.90, P 5.93.

Ammonium uridin-5′-yl [(2-acetamido-2-deoxy-α-D-galactopyranosyl)propyl]phosphonate (14e): Yield: 12.6 mg, 65%; m.p. 159–161 °C; $[\alpha]_D^{26} = +57$ ($c = 1.0$, H_2O); ^{1}H NMR (400 MHz, D_2O, 25 °C): $\delta = 7.88$ (d, $J = 8.1$ Hz, 1H; H-6_u), 5.83 (d, $J = 8.1$ Hz, 1H; H-5_u), 5.82 (d, $J = 5.1$ Hz, 1H; H-1′), 4.25 (t, $J = 5.1$ Hz, 1H; H-2′), 4.20 (t, $J = 5.1$ Hz, 1H; H-3′), 4.17–4.14 (m, 1H; H-4′), 4.09 (dd, $J = 11.0$, 5.9 Hz, 1H; H-2), 4.04 (ddd, $J = 11.7$, 4.0, 2.4 Hz, 1H; H-5′), 4.01–3.92 (m, 2H; H-5′, H-1), 3.83–3.81 (m, 1H; H-4), 3.77 (dd, $J = 11.0$, 3.3 Hz, 1H; H-3), 3.62–3.53 (m, 3H; H-5, H-6), 1.91 (s, 3H; NAc), 1.78–1.32 ppm (m, 6H; H-7, H-8, H-9); ^{13}C NMR (100 MHz, D_2O, 25 °C): $\delta = 174.5$, 166.3, 151.7, 141.6, 102.1, 88.9, 83.1 (d, $J = 7.9$ Hz), 73.9, 73.2, 71.4, 69.3, 68.4, 67.3, 62.3 (d, $J = 5.2$ Hz), 61.2, 49.7, 25.7 (d, $J = 134.9$ Hz), 25.5 (d, $J = 17.3$ Hz), 21.9, 19.2 ppm (d, $J = 4.2$ Hz); ^{31}P NMR (162 MHz, D_2O, 25 °C): $\delta = 28.4$ ppm; elemental analysis calcd (%) for $C_{20}H_{35}N_4O_{13}P$: C 42.11, H 6.18, N 9.82, P 5.43; found: C 42.02, H 6.21, N 9.90, P 5.52.

Expression and purification of *Hs*GalNAc-T2

Homo sapiens GalNAc-T2 (*Hs*GalNAc-T2) was expressed in *Pichia pastoris* and purified as described previously.[15a,32]

Crystallisation

Protein drops were prepared by mixing protein solution (2 μL; the mix contained 7 mg mL^{-1} GalNAc-T2, 5 mM UDP, 5 mM $MnCl_2$ in 25 mM Tris-HCl pH 7.5) and reservoir solution (2 μL; 8–10% polyethylene glycol (PEG) 8000, 8% ethylene glycol, 100 mM Hepes pH 7.0–7.5), as reported previously.[32] Crystals were grown by hanging drop diffusion at 18 °C for 2 days. Then, compound **13d**, as a powder, was soaked onto these crystals for 25 min. The crystals were cryoprotected in the precipitant solution containing an additional 18% ethylene glycol. All crystals were frozen in a nitrogen gas stream cooled to 100 K.

Structure determination and refinement

All data were processed and scaled by using the XDS package[35]and CCP4[36] software; relevant statistics are given in Table S1 in the Supporting Information. The crystal structure was solved by molecular replacement with Phaser[37] and by using PDB entry 2FFV as a template. Initial phases were further improved by cycles of manual model building in Coot[38] and refinement with REFMAC5:[39] The final model was validated with PROCHECK (model statistics are given in Table S1 in the Supporting Information). The asymmetric units of these crystals showed six molecules of GalNAc-T2, forming three independent dimers, as confirmed by the PISA server.[31] Only chain D showed disordered density towards the end of the lectin domain, and thus, we did not include residues with numbering 555 to 569 in the final model. Coordinates and structure factors have been deposited in the Worldwide Protein Data Bank (wwPDB; see Table S1 in the Supporting Information for the pdb codes).

Tryptophan fluorescence spectroscopy

Fluorescence spectroscopy was used to determine the dissociation constant of GalNAc-T2 against the compounds. All experiments were carried out in a Cary Eclipse spectrofluorometer (Varian) at 25 °C with GalNAc-T2 at 1 μM, and concentrations of the compound varying from 1 to 1000 μM in 25 mM Tris, 150 mM NaCl, pH 7.5. Fluorescence emission spectra were recorded in the $\lambda =$ 300–400 nm range with an excitation wavelength of $\lambda = 280$ nm and a slit width of 5 nm. Data analysis was performed in Prism (GraphPad software) by considering a model with a single binding site [see Eq. (1), in which F_0 is the intrinsic fluorescence of the enzyme in the absence of quencher [Q], F_1 is the observed fluorescence at a given quencher concentration, f_a is the fractional degree of fluorescence, and K_d is the dissociation constant].

$$1 - \frac{F_1}{F_0} = \frac{f_a[Q]}{K_d + [Q]} \quad (1)$$

Computational studies

Macromolecule preparation: Two X-ray crystal structures of GalNAc-T2 were selected: one in the closed active conformation (PDB entry 4D0Z, in complex with UDP-5S-GalNAc, EA2 peptide and Mn^{2+}), and the second one in the open inactive conformation (PDB entry 2FFV, in complex with UDP and Mn^{2+}). Chain A was selected in both PDB structures. All water molecules, ethylene glycol molecules and ligands were removed, but Mn^{2+} ions were retained. By using the Maestro package,[40] missing residues were added and modelled, and neutral terminal N and C groups and hydrogen atoms were added. Charges were assigned to all atoms with the OPLS_2005 force field, and the final protein structure was finally minimised with the same force field.

Ligand preparation: The 3D structure for UDP was extracted from PDB 2FFV. The 3D structure for UDP-GalNAc was built from the 3D structure of UDP-5S-GalNAc from PDB 4D0Z. UDP-5S-GalNAc was also prepared for docking as a reference compound. The 3D structures of 16 compounds were built from their SMILES codes. All ligands were subjected to geometry optimisation by using the MMFFs force field implemented in Maestro. This method was chosen by comparing the docking results for reference compounds by applying different force fields.

Docking calculations: Docking was performed by means of the Glide[41] and AutoDock4 programs.[42] In Glide, a cubic grid box of 25^3 Å^3 was generated for each protein, with the centre defined as the centre of mass among Asp176, His145, Arg362 and His226 residues. For all ligands, the torsional angles were allowed to rotate with no restrictions. A standard precision docking was performed for both structures. In the case of AutoDock, different conformers of the starting geometries of the ligands were docked by using the Lamarckian genetic algorithm, by randomly changing the torsion angles and overall orientation of the molecule. A volume for exploration was defined with a grid spacing of 0.375 Å, centred on the centre of mass among Asp176, His145, Arg362 and His226 residues, and a 3D grid of $21\times18\times19$ Å^3 for 4D0Z and $21\times21\times18$ Å^3 for 2FFV. After docking, the 200 solutions were clustered in groups with root-mean-square deviations of less than 2.0 Å. The clusters were ranked by the lowest energy representative of each cluster. Docking results from both Glide and AutoDock methods were analysed and taken together for the discussion.

tMD simulations: The tMD simulation was performed with the AMBER12 software[43] by using the ff10, gaff and Glycam force fields. The Cα atoms of the starting structure (GalNAc-T2 in complex with UDP-GalNAc in the open conformation) were restrained to the corresponding positions of the target structure (GalNAc-T2 in complex with UDP-GalNAc in the closed conformation). Explicit water solvent (TIP3BOX model) was considered by adding the same number of water molecules to both complexes. Also, the same number of counterions (4 chlorine ions) was added. Energy minimisation of all initial structures (closed and open states) was carried out by using the steepest descent method in AMBER12 for 500 steps and the non-bonded cutoff was 9 Å. For the tMD simulation, the SHAKE algorithm was used, the time step was 0.002 ps, and the non-bonded cutoff was 15 Å. The temperature coupling method was used to keep the temperature constant at 300 K and Berendsen temperature coupling scheme was used. No positional restraints were applied. The total simulation time was 5 ns.

Acknowledgements

This work was supported by the Spanish MINECO contracts CTQ2013-44367-C2-1-P (to P.M.), CTQ2013-44367-C2-2-P (to R.H.G.), and CTQ2011-22724 and CTQ2014-57141-R (to S.M.S.). We also acknowledge the Government of Aragón (Spain) (Bioorganic Chemistry group E-10 and Protein Targets group B-89) for financial support. We acknowledge the Institute of Biocomputation and Physics of Complex Systems (BIFI) at the University of Zaragoza (Spain) for computer time at clusters Terminus and Memento. We thank synchrotron radiation sources DLS (Oxford) and, in particular, beamline I04 (experiment number MX8035-26). M.G. thanks the Spanish Ministry of Education (MEC) for a predoctoral grant (FPU program). The European Commission is gratefully acknowledged (Grant GLYCOPHARM FP7-PITN-GA-2012-317297 and BioStruct-X grant agreement no. 283570 and BIOSTRUCTX_5186).

Keywords: carbohydrates · glycomimetics · glycosyltransferases · molecular modeling · proteins

[1] a) J. Voglmeir, S. L. Flitsch, in *Science of Synthesis*, Biocatalysis in Organic Synthesis (Eds.: K. Faber,W.-D. Fessner, N. J. Turner), Thieme, Stuttgart, **2015**, *Vol. 1*, pp. 507 – 542; b) T. M. Gloster, *Curr. Opin. Struct. Biol.* **2014**, *28*, 131 – 141; c) R. Roychoudhury, N. L. B. Pohl, *Curr. Opin. Chem. Biol.*

2010, *14*, 168 – 173; d) L. L. Lairson, B. Henrissat, G. J. Davies, S. G. Withers *Ann. Rev. Biochem.* **2008**, *77*, 521 – 555.
[2] X. Chen, *ACS Chem. Biol.* **2011**, *6*, 14 – 17.
[3] a) S. Wang, S. Vidal, *Carbohydr. Chem.* **2013**, *368*, 78 – 101; b) P. Merino, T. Tejero, I. Delso, R. Hurtado-Guerrero, A. Gomez-SanJuan, D. Sadaba, *Mini-Rev. Med. Chem.* **2012**, *12*, 1455 – 1464; c) T. Kajimoto, M. Node, *Synthesis* **2009**, 3179 – 3210; d) L. J. Whalen, W. A. Greenberg, M. L. Mitchell, C.-H. Wong in *Iminosugars* (Eds.: P. Compain, O. R. Martin), Wiley, Hoboken, **2007**, pp. 153 – 176; e) K.-H. Jung, R. R. Schmidt in *Carbohydrate-Based Drug Discovery, Vol. 2* (Ed.: C.-H. Wong), Wiley-VCH, Weinheim, **2003**, pp. 609 – 659; f) X. Qian, M. M. Palcic, *Carbohydr. Chem. Biol.* **2000**, *3*, 293 – 312; g) L. Tedaldi, G. K. Wagner, *MedChemComm* **2014**, *5*, 1106 – 1125.
[4] D.-M. Liang, J.-H. Liu, H. Wu, B.-B. Wang, H.-J. Zhu, J.-J. Qiao, *Chem. Soc. Rev.* **2015**, *44*, 8350 – 8374.
[5] a) D. Albesa-Jové, M. F. Mendoza, A. Rodrigo-Unzueta, F. Gomollón-Bel, J. Cifuente, S. Urresti, N. Comino, H. Gómez, J. Romero-García, J. M. Lluch, E. Sancho-Vaello, X. Biarnés, A. Planas, P. Merino, L. Masgrau, M. E. Guerin, *Angew. Chem. Int. Ed.* **2015**, *54*, 9898 – 9902; *Angew. Chem.* **2015**, *127*, 10036 – 10040; b) A. Ardèvol, C. Rovira, *J. Am. Chem. Soc.* **2015**, *137*, 7528 – 7547; c) H. Gómez, I. Polyak, W. Thiel, J. M. Lluch, L. Masgrau, *J. Am. Chem. Soc.* **2012**, *134*, 4743 – 4752; d) R. Hurtado-Guerrero, G. J. Davies, *Curr. Opin. Chem. Biol.* **2012**, *16*, 479 – 487; e) I. Tvaroska, *Mini-Rev. Org. Chem.* **2011**, *8*, 263 – 269; f) D. L. Jakeman, *ChemBioChem* **2011**, *12*, 2540 – 2542.
[6] J. Rini, J. Esko, A. Varki in *Essentials of Glicobiology. 2nd ed.* (Eds.: A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, M. E. Etzler), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, **2009**, Available from: http://www.ncbi.nlm.nih.gov/books/NBK1921/.
[7] V. Malik, W. Black Gary, *Adv. Protein Chem. Struct. Biol.* **2012**, *87*, 87 – 115.
[8] a) K. Descroix, G. K. Wagner, *Org. Biomol. Chem.* **2011**, *9*, 1855 – 1863; b) M. S. M. Timmer, M. V. Chumillas, W. E. Donker-Koopman, J. Alerts, G. A. van der Marel, H. S. Overkleeft, J. H. van Boom, *J. Carbohydr. Chem.* **2005**, *24*, 335 – 351; c) H. C. Dorfmueller, V. S. Borodkin, D. E. Blair, S. Pathak, I. Navratilova, D. M. F. van Aalten, *Amino Acids* **2011**, *40*, 781 – 792; d) F. Stolz, M. Reiner, A. Blume, W. Reutter, R. R. Schmidt, *J. Org. Chem.* **2004**, *69*, 665 – 679.
[9] J. Jiang, M. B. Lazarus, L. Pasquina, P. Sliz, S. Walker, *Nat. Chem. Biol.* **2012**, *8*, 72 – 77.
[10] a) R. Jørgensen, L. L. Grimm, N. Sindhuwinata, T. Peters, M. M. Palcic, *Angew. Chem. Int. Ed.* **2012**, *51*, 4171 – 4175; *Angew. Chem.* **2012**, *124*, 4247 – 4251; b) K. Schaefer, J. Albers, N. Sindhuwinata, T. Peters, B. Meyer, *ChemBioChem* **2012**, *13*, 443 – 450; c) J. C. Errey, S. S. Lee, R. P. Gibson, C. M. Fleites, C. S. Barry, P. M. J. Jung, A. C. O'Sullivan, B. G. Davis, G. J. Davies, *Angew. Chem. Int. Ed.* **2010**, *49*, 1234 – 1237; *Angew. Chem.* **2010**, *122*, 1256 – 1259; d) S. Wang, J. A. Cuesta-Seijo, D. Lafont, M. M. Palcic, S. Vidal, *Chem. Eur. J.* **2013**, *19*, 15346 – 15357; e) S. Wang, J. A. Cuesta-Seijo, A. Striebeck, D. Lafont, M. M. Palcic, S. Vidal, *ChemPlusChem* **2015**, *80*, 1525 – 1532; f) S. Wang, D. L. Shen, D. Lafont, A.-S. Vercoutter-Edouart, M. Mortuaire, Y. Shi, O. Maniti, A. Girard-Egrot, T. Lefebvre, B. M. Pinto, D. Vocadlo, S. Vidal, *MedChemComm* **2014**, *5*, 1172 – 1178; g) J. Jiang, V. Kanabar, B. Padilla, F. Man, S. C. Pitchford, C. P. Page, G. K. Wagner *Chem. Commun* **2016**, *52*, 3955 – 3958.
[11] T. M. Gloster, W. F. Zandberg, J. E. Heinonen, D. L. Shen, L.-H. Deng, D. J. Vocadlo, *Nat. Chem. Biol.* **2011**, *7*, 174 – 181.
[12] A. Schäfer, J. Thiem, *J. Org. Chem.* **2000**, *65*, 24 – 29.
[13] a) R. Chang, T.-T. Vo, N. S. Finney, *Carbohydr. Res.* **2006**, *341*, 1998 – 2004; b) J. Hajduch, G. Nam, E. J. Kim, R. Froehlich, J. A. Hanover, K. L. Kirk, *Carbohydr. Res.* **2008**, *343*, 189 – 195; c) A. J. Clarke, R. Hurtado-Guerrero, S. Pathak, A. W. Schuttelkopf, V. Borodkin, S. M. Shepherd, A. F. M. Ibrahim, D. M. F. van Aalten, *Embo J.* **2008**, *27*, 2780 – 2788; d) S. K. Partha, A. Sadeghi-Khomami, K. Slowski, T. Kotake, N. R. Thomas, D. L. Jakeman, D. A. R. Sanders, *J. Mol. Biol.* **2010**, *403*, 578 – 590; e) S. M. Forget, A. Jee, D. A. Smithen, R. Jagdhane, S. Anjum, S. A. Beaton, D. R. J. Palmer, R. T. Syvitskia, D. L. Jakeman, *Org. Biomol. Chem.* **2015**, *13*, 866 – 875.
[14] S. Vidal, I. Bruyere, A. Malleron, C. Auge, J.-P. Praly, *Bioorg. Med. Chem.* **2006**, *14*, 7293 – 7301.
[15] a) E. Lira-Navarrete, M. d. L. Rivas, I. Companon, M. C. Pallares, Y. Kong, J. Iglesias-Fernandez, G. a. J. L. Bernardes, J. M. Peregrina, C. Rovira, P. Bernado, P. Bruscolini, H. Clausen, A. Lostao, F. Corzana, R. Hurtado-Guerrero, *Nat. Commun.* **2015**, *6*, 6937; b) B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, *Nucleic Acids Res.* **2009**, *37*, D233 – D238; c) K. G. T. Hagen, T. A. Fritz, L. A. Tabak, *Glycobiology* **2003**, *13*, 1R – 16R.
[16] H. Gómez, R. Rojas, D. Patel, L. A. Tabak, J. M. Lluch, L. Masgrau, *Org. Biomol. Chem.* **2014**, *12*, 2645 – 2655.
[17] A. Dondoni, A. Marra, *Org. Biomol. Chem.* **2015**, *13*, 2212 – 2215.
[18] a) D. Horton, T. Miyake, *Carbohydr. Res.* **1988**, *184*, 221 – 229; b) J. R. Kramer, T. J. Deming, *J. Am. Chem. Soc.* **2010**, *132*, 15068 – 15071.
[19] R. Y. Tam, S. S. Ferreira, P. Czechura, J. L. Chaytor, R. N. Ben, *J. Am. Chem. Soc.* **2008**, *130*, 17494 – 17501.
[20] a) A. Babič, S. Gobec, C. Gravier-Pelletier, Y. Le Merrer, S. Pečar, *Tetrahedron* **2008**, *64*, 9093 – 9100; b) J. Cui, D. Horton, *Carbohydr. Res.* **1998**, *309*, 319 – 330.
[21] S. Liu, R. N. Ben, *Org. Lett.* **2005**, *7*, 2385 – 2388.
[22] a) V. R. Bouvet, R. N. Ben, *J. Org. Chem.* **2006**, *71*, 3619 – 3622; b) C. H. Röhrig, M. Takhi, R. R. Schmidt, *Synlett* **2001**, 1170 – 1172.
[23] a) M. J. Damha, N. Usman, K. K. Ogilvie, *Can. J. Chem.* **1989**, *67*, 831 – 839; b) P. Barthelemy, C. A. H. Prata, S. F. Filocamo, C. E. Immoos, B. W. Maynor, S. A. N. Hashmi, S. J. Lee, M. W. Grinstaff, *Chem. Commun.* **2005**, 1261 – 1263.
[24] A. Dondoni, S. Staderini, A. Marra, *Eur. J. Org. Chem.* **2013**, 5370 – 5375.
[25] H. Pizova, P. Bobal, *Tetrahedron Lett.* **2015**, *56*, 2014 – 2017.
[26] P. J. Casara, K. C. Jund, A. Clauss, J.-F. Nave, R. D. Snyder, *Bioorg. Med. Chem. Lett.* **1992**, *2*, 145 – 148.
[27] M. Izumi, K. Wada, H. Yuasa, H. Hashimoto, *J. Org. Chem.* **2005**, *70*, 8817 – 8824.
[28] G. W. Daub, E. E. J. Van Tameken, *J. Am. Chem. Soc.* **1977**, *99*, 3526 – 3528.
[29] J.-M. Campagne, J. Coste, P. Jouin, *J. Org. Chem.* **1995**, *60*, 5214 – 5223.
[30] E. P. Bennett, U. Mandel, H. Clausen, T. A. Gerken, T. A. Fritz, L. A. Tabak, *Glycobiology* **2012**, *22*, 736 – 756.
[31] E. Lira-Navarrete, J. Iglesias-Fernandez, W. F. Zandberg, I. Companon, Y. Kong, F. Corzana, B. M. Pinto, H. Clausen, J. M. Peregrina, D. J. Vocadlo, C. Rovira, R. Hurtado-Guerrero, *Angew. Chem. Int. Ed.* **2014**, *53*, 8206 – 8210; *Angew. Chem.* **2014**, *126*, 8345 – 8349.
[32] T. A. Fritz, J. Raman, L. A. Tabak, *J. Biol. Chem.* **2006**, *281*, 8613 – 8619.
[33] C. Andreini, I. Bertini, G. Cavallaro, G. L. Holliday, J. M. Thornton, *J. Biol. Inorg. Chem.* **2008**, *13*, 1205 – 1218.
[34] C. Andreini, I. Bertini, G. Cavallaro, G. L. Holliday, J. M. Thornton, *J. Biol. Inorg. Chem.* **2008**, *13*, 1205 – 1218.
[35] W. Kabsch, *Acta Crystallogr. Sect. D* **2010**, *66*, 125 – 132.
[36] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, *Acta Crystallogr. Sect. D* **2011**, *67*, 235 – 242.
[37] Collaborative Computational Project, Number 4, *Acta Crystallogr. Sect. D* **1994**, *50*, 760 – 763.
[38] P. Emsley, K. Cowtan, *Acta Crystallogr. Sect. D* **2004**, *60*, 2126 – 2132.
[39] G. N. Murshudov, A. A. P. Skubák, N. S. Lebedev, R. A. Pannu, R. A. Steiner, M. D. Nicholls, F. L. Winn, A. A. Vagin, *Acta Crystallogr. Sect. D* **2011**, *67*, 355 – 367.
[40] Schrödinger Release 2014-4, Maestro, version 10.0, Schrödinger, LLC, New York, NY, **2014**.
[41] Glide, version 6.5, Schrödinger, LLC, New York, NY, **2014**.
[42] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem.* **1998**, *19*, 1639 – 1662.
[43] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Götz, I. Kolossváry, K. F.Wong, F. Paesani, J. Vanicek, R. M.Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P. A. Kollman, *AMBER 12*, **2012**, University of California, San Francisco.

Received: January 31, 2016
Published online on April 13, 2016

Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-polar Conjugates

Mattia Ghirardello,[a] Matilde de las Rivas,[b] Ignacio Delso,[c] Erandi Lira-Navarrete[b], Tomás Tejero,[a] Sonsoles Martín-Santamaría,*[d] Ramón Hurtado-Guerrero*[b,e,f] and Pedro Merino*[a]

[a] Departamento de Síntesis y Estructura de Biomoléculas, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH), Universidad de Zaragoza, CSIC, 50009 Zaragoza, Spain
[b] Instituto de Biocomputación y Fisica de Sistemas Complejos (BIFI), BIFI-IQFR (CSIC) Joint Unit, Universidad de Zaragoza, 50009 Zaragoza, Spain
[c] Servicio de Resonancia Magnética Nuclear, Centro de Química y Materiales de Aragón (CEQMA), Universidad de Zaragoza, CSIC, Campus San Francisco, 50009 Zaragoza , Spain
[d] Departmento de Biología Físico-Química, Centro de Investigaciones Biológicas, CIB-CSIC, Ramiro de Maeztu, 9, 28040 Madrid, Spain
[e] Fundación ARAID, 50018, Zaragoza, Spain.
[f] Instituto de Investigaciones Sanitarias de Aragón (IIS-A), Zaragoza 50009, Spain.

Supporting Information

Crystallographic Data

Supplementary Table XXX. Data collection and refinement statistics. Values in parentheses refer to the highest resolution shell. Ramachandran plot statistics were determined with PROCHECK.

	*GalNAc-T2-**16d***
Space group	$P2_12_12_1$
Wavelength (Å)	0.97
Resolution (Å)	20-2.07 (2.124-2.07)
Cell dimensions (Å)	a =116.49 b = 121.75 c = 250.15
Unique reflections	215817
Completeness	99.8 (99.84)
R_{sym}	0.061 (0.632)
$I/\sigma(I)$	15.6 (2.6)
Redundancy	5.7 (5.7)
R_{work} / R_{free}	0.189/0.246
RMSD from ideal geometry, bonds (Å)	0.017
RMSD from ideal geometry, angles (º)	1.996
<*B*> protein (Å^2)	48.81
X1<*B*> Ethlenglycol (Å^2)	64.24
<*B*> Glycerol (Å^2)	71.18
<*B*> DRG(Å^2)	64.33
<*B*> UDP (Å^2)	47.52
<*B*> Mn^{+2} (Å^2)	32.78
<*B*> solvent (Å^2)	49.87
Ramachandran plot: Most favoured (%) Additionally allowed (%) Disallowed (%)	 93.37 5.09 1.53
PDB ID	

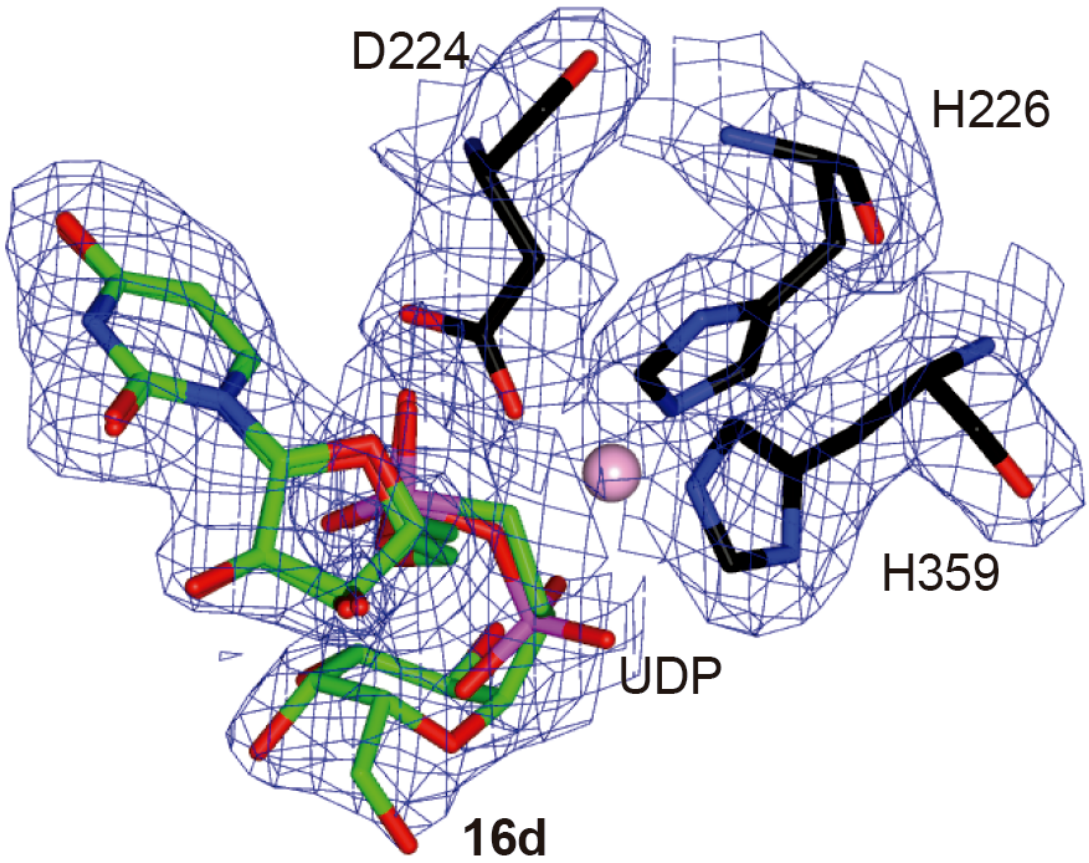


Figure S1. Compound **16d** electron density map. Electron density maps are F_O–F_C syntheses (blue) contoured at 2.2 σ for UDP/**16d**, and $2F_O$–F_C syntheses (blue) contoured at 1.0 σ for the residues Asp224, His226 and His359. The manganese atom is shown as a pink sphere.

Computational Data

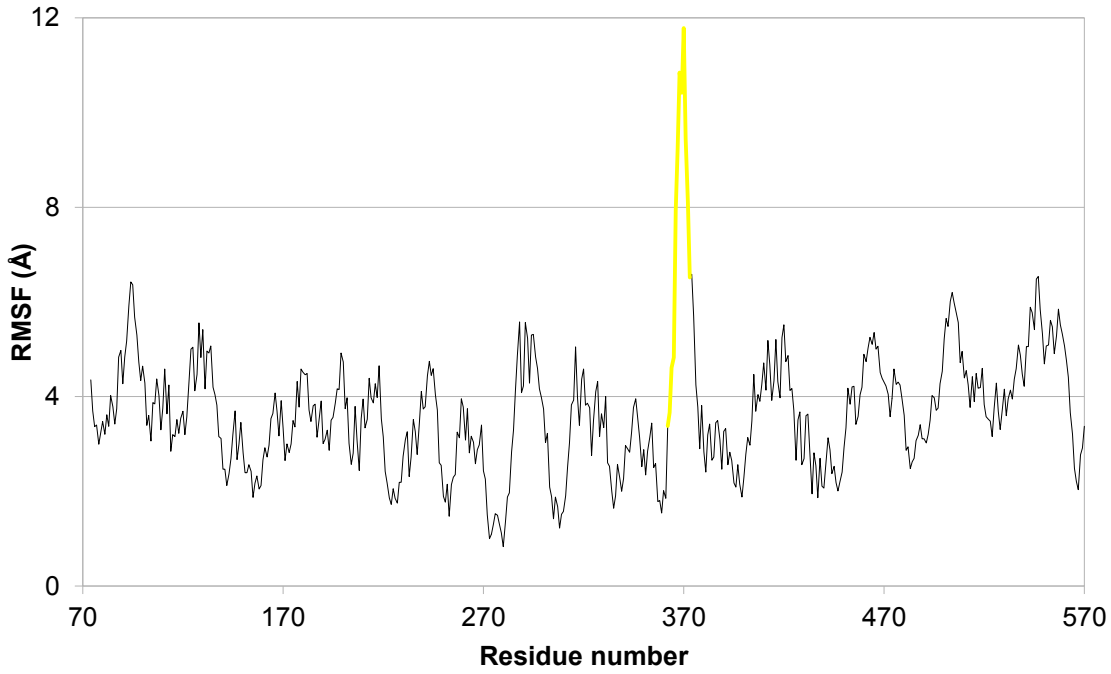


Figure S2. RMSF for the Cα atoms for GalNAc-T2 in complex with UDP-GalNAc during the tMD simulation. Values corresponding to the flexible loop Arg362-Ser373 are marked in yellow.

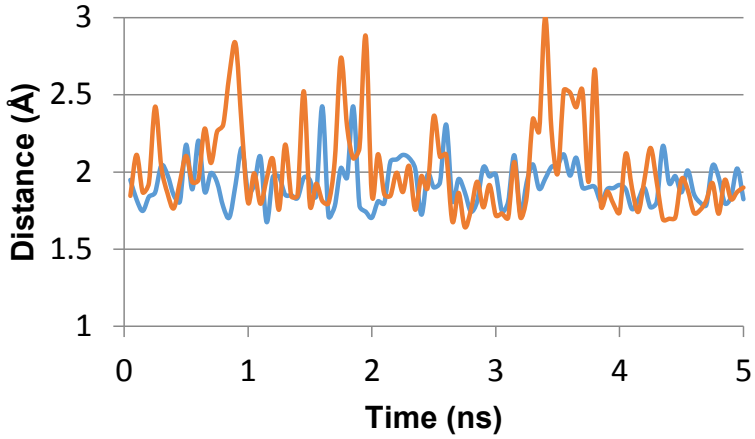


Figure S3. Distances between the carboxylic oxygens from Asp176 side chain and the NH in position 3 from the uracyl (blue), and between the Arg201 guanidinium hydrogen atoms and the CO in position 2 from the uracyl (orange) along the tMD simulation time.

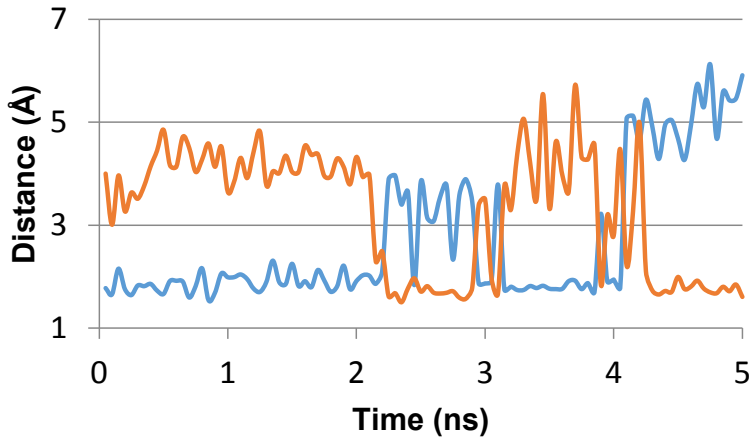


Figure S4. Distances between the CO in position 4 from the uracyl, and the OH group from the side chains of Thr143 (blue) and Ser207 (orange) along the tMD simulation time.

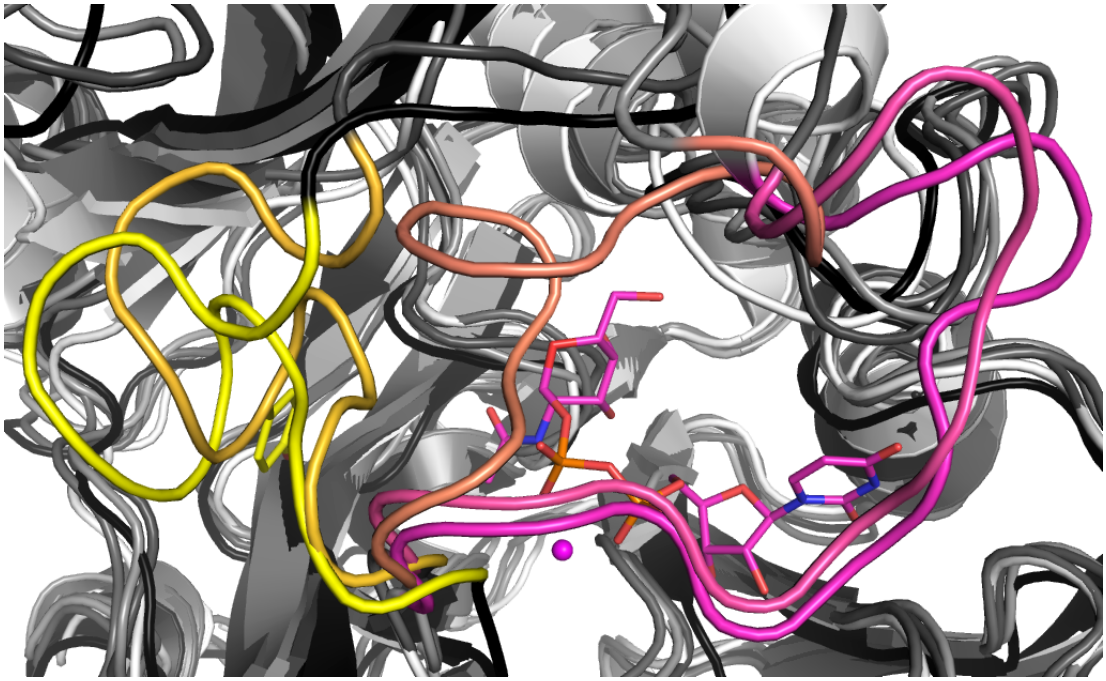


Figure S5. Superimposition of five structures of the GalNAc-T2/UDP-GalNAc complex along the tMD simulation of the transition from the open (black, loop in yellow) to the closed (white, loop in magenta) conformation. Intermediate structures are shown grading from yellow (starting geometry), orange (intermediate geometries) to magenta (final target geometry). Ligand UDP-GalNAc is shown in yellow for the open complex, and in magenta for the closed complex. A) Full perspective. B) Detail showing UDP-GalNAc in the closed complex. C) Detail showing the open (yellow) and the closed (magenta) structures and some representative residues.

ARTÍCULO IV

The small molecule luteolin inhibits *N*-acetyl-alpha-galactosaminyltransferases and reduces mucin-type O-glycosylation of amyloid precursor protein.

El presente artículo fue publicado en la revista *Journal of Biological Chemistry* el año 2017, tercer año del periodo de tesis de la doctoranda. Fue realizado en coautoría con: (a) el Laboratorio de Sistemas de Biomedicina de la Universidad Jiao Tong de Shanghai; (b) el Instituto de Síntesis Química y Catálisis Homogénea de la Univseridad de Zaragoza y (c) el Laboratorio para la Exploración y Utilización de Recursos Genéticos Acuáticos de la Universidad Oceánica de Shanghai.

A fecha de la presente tesis doctoral, el único de los coautores del artículo que no está en posesión del Grado de Doctor es Feng Liu, del Laboratorio de Sistemas de Biomedicina de la Universidad Jiao Tong de Shanghai. Se adjunta su renuncia a presentar este artículo en la modalidad de compendio de pubilcaciones en el Apéndice 10.6 de la presente tesis.

La contribución del doctorando en esta publicación ha consistido en lo siguiente:

- Expresión y purificación de la proteína silvestre.
- Clonación, expresión y purificación de los mutantes de la proteína silvestre empleados para verificar la unión de ésta al compuesto inhibidor.
- Cristalización de la proteína en complejo con el compuesto inhibidor y resolución de su estructura.

Se adjunta a continuación la referencia completa del artículo:

Liu, F., Xu, K., Xu, Z., **de las Rivas, M**., Wang, C., Li, X., Lu, J., Zhou, Y., Delso, I., Merino, P., Hurtado-Guerrero, R., Zhang, Y. & Wu, F. (**2017**). The small molecule luteolin inhibits *N*-acetyl-alpha-galactosaminyltransferases and reduces mucin-type O-glycosylation of amyloid precursor protein. *Journal of Biological Chemistry*, 292(52), pp. 21304-21319.

JBC Papers in Press. Published on October 23, 2017 as Manuscript M117.814202
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.M117.814202

The small molecule luteolin inhibits *N*-acetyl-α-galactosaminyltransferases and reduces mucin-type *O*-glycosylation of amyloid precursor protein

Feng Liu[1], Kai Xu[5], Zhijue Xu[1], Matilde de las Rivas[2], Xing Li[1], Jishun Lu[1], Ignacio Delso[3], Pedro Merino[2], Ramon Hurtado-Guerrero[2, 4, *] and Yan Zhang[1, *]

[1]Key Laboratory of Systems Biomedicine (Ministry of Education) and Collaborative Innovation Center of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China.
[2]Instituto de Biocomputación y Fisica de Sistemas Complejos (BIFI), BIFI-IQFR (CSIC) Joint Unit, Universidad de Zaragoza, 50009, Zaragoza, Spain.
[3]Instituto de Síntesis Química y Catálisis Homogénea (ISQCH). Universidad de Zaragoza. CSIC. E-50009 Zaragoza. Aragón, Spain.
[4]Fundación ARAID, 50018 Zaragoza, Spain.
[5]Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Ministry of Education, College of Fisheries and Life Science, Shanghai Ocean University, Shanghai 201306, China.

Running title: Luteolin inhibits ppGalNAc-Ts

[*]**To whom correspondence should be addressed:**
Yan Zhang, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dongchuan Rd., 200240, Shanghai, China; Tel./Fax: 86-21-34206778; E-mail: yanzhang2006@sjtu.edu.cn.
Ramon Hurtado-Guerrero, Instituto de Biocomputación y Fisica de Sistemas Complejos (BIFI), BIFI-IQFR (CSIC) Joint Unit, Universidad de Zaragoza, 50009, Zaragoza, Spain; Tel.: +34 976 762997, fax: + 34 976 762990; E-mail: rhurtado@bifi.es.

Key words: *O*-glycosylation, ppGalNAc-T, glycosylation inhibitor, glycoprotein, glycosyltransferase, luteolin, amyloid precursor protein (APP), crystal structure

Mucin-type *O*-glycosylation is the most abundant type of *O*-glycosylation. It is initiated by the members of polypeptide *N*-acetyl-α-galactosaminyltransferase (ppGalNAc-T) family and closely associated with both physiological and pathological conditions such as coronary artery disease or Alzheimer's disease. The lack of direct and selective inhibitors of ppGalNAc-Ts has largely impeded research progress in understanding the molecular events in mucin-type *O*-glycosylation. Here, we report that a small molecule, the plant flavonoid luteolin, selectively inhibits ppGalNAc-Ts *in vitro* and in cells. We found that luteolin inhibits ppGalNAc-T2 through a peptide/protein-competitive manner but not promiscuously, *e.g.* via aggregation-based activity. X-ray structural analysis revealed that luteolin binds to the PxP motif–binding site found in most protein substrates, which was further validated by comparing the interactions between luteolin with wildtype enzyme and mutants using ^{1}H NMR-based binding experiments. Functional studies disclosed that luteolin at least partially

reduced production of β-amyloid (Aβ) protein by selectively inhibiting the activity of ppGalNAc-T isoforms. In conclusion, our study provides key structural and functional details on luteolin inhibiting ppGalNAc-T activity, opening up the way for further optimization of more potent and specific ppGalNAc-T inhibitors. Moreover, our findings may inform future investigations into site-specific *O*-GalNAc glycosylation and into the molecular mechanism of luteolin-mediated ppGalNAc-T inhibition.

INTRODUCTION

Mucin-type *O*-glycosylation (hereafter, also referred to as *O*-GalNAc glycosylation) is the most abundant type of *O*-glycosylation and is closely relevant in physiological and pathological conditions such as coronary artery disease (1), familial tumoral calcinosis (2) and Alzheimer's disease (3,4). It has been clearly demonstrated that *O*-GalNAc glycosylation of Amyloid precursor protein (APP) is important for its cell surface localization, endocytosis and the production of Amyloid β (Aβ) (5-7), which is the primary component of the parenchymal amyloid deposits in Alzheimer's disease (8). In mammals, *O*-GalNAc glycosylation is initiated by a family of 20 polypeptide *N*-acetyl-α-galactosaminyltransferases (ppGalNAc-Ts) (9,10). These enzymes transfer GalNAc from the donor substrate, uridine diphosphate *N*-acetyl-α-galactosamine (UDP-GalNAc), to an acceptor substrate-protein, forming GalNAc-Ser/Thr (also known as the Tn antigen) and release the free UDP (9) (Fig. 1A). The acceptor substrate specificities of these isoforms are distinct but partly overlapping (11-13). This is reflected by increasing evidence that different ppGalNAc-Ts could also elicit distinct biological outcomes (14). The site-specific *O*-GalNAc glycans on proteins catalyzed by different ppGalNAc-T isoforms are closely related with protein processing, stability and secretion (14). Change of *O*-GalNAc glycans on specific sites of protein substrates such as angiopoietin-like protein 3 (ANGPTL3) or APP is deeply involved in coronary artery disease or Alzheimer's disease (1,4,7,15). Therefore, development of ppGalNAc-T inhibitor could provide a useful tool for research on the molecular mechanism of *O*-GalNAc glycosylation in human development and disease.

While several small molecules are reported to perturb *O*-GalNAc glycosylation in cells (16-18), such as the uridine derivative 1-68A and GalNAc analogue Ac_5GalNTGc, most of these compounds are donor substrate derivatives (16,18). We have described UDP-GalNAc derivatives that acted as poor binders of ppGalNAc-T2 (19). Recently, a mixed-mode selective inhibitor of ppGalNAc-T3 has been reported (17). Natural products and approved drugs are structurally diverse and have well-characterized bioactivity, safety and pharmacological function. Luteolin is an effective ingredient and natural product found in African and Chinese traditional herbal medicines with safe profile (20), (21). As a flavonoid, several co-crystal studies revealed that luteolin interacts with protein kinase, tankyrases or receptor, etc (22-27). However, it is unknown whether luteolin could affect protein *O*-glycosylation.

In this context, we identified luteolin as a direct protein substrate-competitive inhibitor of ppGalNAc-T2. The competitive nature of this flavonoid versus peptide substrates was also supported by the crystal structure of ppGalNAc-T2 in complex with UDP and luteolin, which allowed us to pinpoint the flavonoid in the peptide-binding groove and in particular where the PxP motif (where x is usually a small hydrophobic residue) of protein

substrates is recognised. The location of luteolin was also supported by ^{1}H NMR binding experiments. Moreover, we found that luteolin reduced the production of Aβ through selectively inhibiting ppGalNAc-Ts, which provided a putative molecular mechanism of luteolin in reducing the Alzheimer's disease pathologies.

RESULTS

Identification of luteolin as an inhibitor of ppGalNAc-T2 and reducing *O*-GalNAc glycosylation in cells

Based on the high-performance liquid chromatography (HPLC) ppGalNAc-T assay (28-30), we found luteolin could inhibit ppGalNAc-T2 catalytic activity with an IC_{50} of ~15 μM using EA2 peptide (PTTDSTTPAPTTK) as acceptor substrate (Fig. 1B). ppGalNAc-T2 is ubiquitously expressed in a variety of cells and several structural studies revealed that EA2 is an effective substrate of ppGalNAc-T2 (31-33).

To evaluate whether luteolin had specific inhibition against *O*-GalNAc glycosylation in cell, firstly, we detected the metabolic labelling of GalNAz in CHO-ldlD cells with or without luteolin treatment by using click chemistry. CHO-ldlD is an *O*-GalNAc glycosylation deficient cell line due to the lack of the epimerase that converts UDP-Glc/GlcNAc to UDP-Gal/GalNAc (34), then extra addition of peracetylated *N*-azidoacetylgalactosamine (Ac_4GalNAz) in cell culture will lead to specific formation of GalNAz modification (35). We incubated CHO-ldlD cells with Ac_4GalNAz and treated with or without indicated concentrations of luteolin. The cells were collected for click-chemistry reaction and analysed by streptavidin blotting. We found that luteolin showed a dose-dependent inhibitory effect on the GalNAz signal, indicating a potent inhibitory effect of luteolin on *O*-GalNAc glycosylation in cells (Fig. 1C).

Next, we investigated the inhibitory selectivity of luteolin in several cell lines. Three lectins, namely, Jacalin (binds to Gal β1-3GalNAc, GalNAc; *O*-GalNAc glycans), *Helix pomatia agglutinin* (HPA, binds to α-linked terminal GalNAc) and *concanavalin A* (ConA, binds to high Man) were used. Lectin blots of total lysates of HEK 293T cells showed a substantial reduction in the Jacalin signal but not ConA signal (Fig. 1D). In Jurkat cells, immunofluorescence detection indicated that luteolin could also suppress the HPA but not ConA signal on the cell surface (Fig. 1E). These results demonstrated that luteolin selectively reduced the global levels of *O*-GalNAc glycosylation in cells. In agreement with the results above, we detected the inhibitory effects of luteolin on the endogenous APP, a typical *O*-GalNAc glycosylated protein (5,7) in HEK 293T cells and A549 cells. We found that luteolin could significantly reduce the ratio of high molecular weight band of APP (APP-H, *O*-GalNAc glycosylated APP (7)) with 75% and 80% reduction in HEK 293T cells and A549 cells, respectively (Fig. 1F and G). Meanwhile, we transfected APP and another typical *O*-GalNAc glycosylated protein podoplanin (PDPN) (36) in HEK 293T cells to test the inhibitory effects of luteolin. After luteolin treatment, the APP-H as well as its Jacalin signal was significantly decreased (Fig. 1H). In contrast, there was little change in APP-L (non-*O*-GalNAc glycosylated APP) or ConA signal (Fig. 1H). Besides APP, luteolin treatment also led to a shift of podoplanin (PDPN) to a lower molecule weight, which is consistent with the effect of *O*-GalNAc glycans depletion (37) (Fig. 1I). Meanwhile, similar expression level of ppGalNAc-T2 was observed in HEK 293T treated with or without luteolin, suggesting that the inhibitory effects of luteolin should not be attributed to the alteration in the ppGalNAc-T2

expression (Fig. 1H and I). Those experiments suggested that luteolin exhibited selective inhibition towards *O*-GalNAc glycosylation on proteins in cells.

Luteolin is a selective and protein substrate-competitive inhibitor of ppGalNAc-T2

Flavonoids have been reported as a family of molecules with multiple targets through promiscuous mechanism, in particular owing to colloidal aggregation (38). To rule out this possibility on the inhibitory effects of luteolin against ppGalNAc-T2 in our enzyme system, dynamic light scattering (DLS) analysis and spin-down experiments were performed. The DLS analysis showed that the particles in the 50–1000 nm size of luteolin were gradually reduced with the increased addition of TritonX-100 and almost undetectable when TritonX-100 reached 0.2% (Fig. 2A and Supplementary Fig. S1). And then the spin-down experiments using our ppGalNAc-T assay containing 0.2% TritonX-100 also showed that the inhibition of luteolin against ppGalNAc-T2 had no significant difference after centrifugation (Fig. 2B). These results strongly suggested that the inhibition of luteolin on ppGalNAc-T2 with 0.2% TritonX-100 was not due to the aggregation-based mechanism.

To address the inhibitory mechanism of luteolin on ppGalNAc-T2, we performed enzyme kinetics assays (Fig. 2C and D). The enzyme kinetics experiments showed that luteolin exhibited competitive inhibition with respect to EA2 substrate (K_i~1.4 μM; Fig. 2C) and non-competitive inhibition with respect to the UDP-GalNAc (αK_i~4.1 μM; Fig. 2D). Luteolin bound ~100 fold greater to ppGalNAc-T2 than EA2 peptide (K_m~160 μM; Fig. 2E).

Prompted by the potent inhibition of luteolin on ppGalNAc-T2, we also measured the IC_{50} values of luteolin in the presence of various peptide substrates and ppGalNAc-T isoforms. We found out that luteolin displayed poor substrate selectivity (Fig. 2E and Supplementary Fig. S2) . In addition to ppGalNAc-T2, we also evaluated the effect of luteolin on other members of ppGalNAc-T family and human *O*-GlcNAc transferase (OGT) (39-41). Luteolin inhibited with similar IC_{50}s, around low μM values, both members of subfamilies Ib and Ic (42) such as ppGalNAc-T14 (note that ppGalNAc-T2 also belongs to Ib subfamily) and ppGalNAc-T3/T6, respectively, whereas it showed a more potent inhibition on ppGalNAc-T10 (subfamily IIb) with an IC_{50} of 1.7 μM (Fig. 2E and Supplementary Fig. S2). In contrast to the effects on these ppGalNAc-Ts, luteolin displayed a poor inhibition on ppGalNAc-T1 and T13, two ppGalNAc-Ts from subfamily Ia with high homology between their amino acid sequences (Fig. 2E and Supplementary Fig. S2). Likewise, luteolin had little effect on OGT either. These results suggested that luteolin displayed selective inhibition on ppGalNAc-Ts from subfamilies Ib, Ic and IIb, and was a poor inhibitor of subfamily Ia and other distant glycosyltransferases such as OGT (Fig. 2E).

X-ray structure and ^{1}H NMR binding experiments of ppGalNAc-T2/mutants with luteolin

To further elucidate the luteolin-binding mode, we obtained crystals of ppGalNAc-T2 in complex with UDP/Mn^{2+} that were subsequently soaked with the flavonoid. Crystals of the complex diffracted to a resolution of 2.3 Å, which allowed us to build a final model with good refinement statistics (Supplementary Table. S1). Consistent with the above kinetics data, the crystal structure revealed that this flavonoid bound to the peptide-binding groove (Fig. 3A). The presence of UDP in the active site at a distance of ~8 Å from luteolin also supported the

non-competitive character of this flavonoid versus UDP-GalNAc inferred from kinetics data (Fig. 2D).

A closer inspection of the luteolin-binding site revealed that the flavonoid is sandwiched between β6, the loop between β6-β7, β8 and the flexible loop between β10-α11 (Supplementary Fig. S3). The compound is also close to the lectin domain from the neighbouring monomer present in the asymmetric unit (AU). Note that in these orthorhombic crystals, six molecules of ppGalNAc-T2 are arranged in three independent dimers in the AU (33) . However, luteolin only accomplishes direct interactions with residues located in the peptide-binding groove of one of the monomers, ruling out the potential contribution of the lectin domain of the neighbouring ppGalNAc-T2 to the binding of the flavonoid (Supplementary Fig. S4). Though dimers have been described by small-angle X-ray scattering experiments (32), it is not clear if this particular crystallographic dimer is present in solution because part of the peptide-binding groove is partly occluded by residues of the neighbouring monomer, impeding the binding of long peptides to the enzyme (33). A more detailed evaluation of the binding site reveals the flavonoid competes specifically with the PxP motif found on most protein substrates (Fig. 3A). The PxP motif-binding site is formed by Phe361, Phe280 and Trp282, which are conserved residues along the ppGalNAc-T family. These are key residues in the recognition of common peptide motifs such as Pro-x-Pro (where x is usually a small hydrophobic residue) found in acceptor substrates (31,32). The 5,7-dihydroxy-4-chromenone ring is engaged in parallel and edge-to-face π-π interactions with Trp282 and Phe361 respectively, and π-CH interactions with Val255 (Fig. 3A). Phe361 also establishes CH-π interactions with the dihydroxyphenyl moiety. In addition, there are direct hydrogen bonds between the 7-hydroxyl group and Ile253 backbone, and water molecules-mediated hydrogen bonds between the 5-hydroxyl group and Ser267 backbone, the 3'-hydroxyl group and Phe361/Lys363, and the oxygen ring and Trp282 (Fig. 3A). The 7-hydroxyl group also interacts with an ethylenglycol molecule, which is located in a hydrophobic pocket formed by Ala266, Ile252, Ile253, Leu270, Ile301 and Ile351 (Fig. 3A).

The AU also depicted the presence of two additional luteolin molecules (Fig. 3B) with less defined density compared to the one displayed in Fig. 3A. The 5,7-dihydroxy-4-chromenone adopted almost identical positions (atomic shift average of 0.62 Å; Fig. 3C) whereas the dihydroxyphenyl moiety was more flexible (atomic shift average of 1.50 Å; Fig. 3C). Overall, the major interactions with the enzyme are conserved whereas the water molecules-mediated hydrogen bonds between Trp282 and residues of the flexible loop are lost (Fig. 3A), supporting again the flexibility and the potential weaker binding affinity of the dihydroxyphenyl moiety in opposition to the chromenone ring. Strikingly, luteolin can only access the active form of ppGalNAc-T2 (Fig. 3D), which suggests the flexible loop must adopt its closed conformation, leading to the unblocking of the PxP motif binding site and consequently allowing the entrance of luteolin (Fig. 3D).

To further validate the location of luteolin inferred from the crystal structure, ^{1}H NMR experiments were conducted. ^{1}H NMR spectra were recorded to compare the effect of binding on the luteolin signal broadening (43-45). Due to its low solubility, luteolin concentration was set to 250 μM (final concentration of DMSO was 8%). Samples with different protein concentrations were assayed. However, best

results were obtained at 2.5 μM in the presence of 50 μM of UDP and $MnCl_2$.

Fig. 3E and Supplementary Fig. S5 shows the ^1H spectra of free luteolin, luteolin with ppGalNAc-T2 wild-type and the mutants W282A and F463A. When ppGalNAc-T2 wild type is incubated with a solution containing luteolin, all signals get broad and less intense due to tight binding to the protein. However, no signal broadening was observed when luteolin is in the presence of W282A mutant, implying that a significant decrease of the affinity is achieved. This supports the crystal structure where the benzopyronic moiety interacts with W282. In the sample containing luteolin and the mutant F463A (located in the lectin domain from the neighbouring ppGalNAc-T2), the broad signals indicate strong binding with the protein, suggesting the lectin domain from the neighbouring monomer forming the dimer in the asymmetric unit, do not play any role in the interaction with this flavonoid. Note that F463 is the closest residue to luteolin coming from the lectin domain of the neighbouring ppGalNAc-T2.

Structure-function analysis of luteolin-analogues inhibition on ppGalNAc-T2

Considering flavonoid has drawn great attention for their beneficial effect (46), we further examined the inhibitory effects of another 16 flavonoid analogues (Fig. 4). Most of them showed less than 10% inhibition on the activity of ppGalNAc-T2 at 50 μM in our HPLC-based ppGalNAc-T assay, suggesting that the inhibition on ppGalNAc-T2 is only achieved by specific flavonoids. To explore the structure and activity relationship of luteolin, we also compared the inhibitory effects of flavonoid analogues with or without these hydroxyl groups. For example, isorhamnetin and chrysoeriol, which contained bulkier substituents at position 3' in the B ring exhibited poor inhibitory effect; fisetin which had no OH-5 in A ring exhibit no inhibitory effect at 50 μM and only ~20 % effects when we raised its amount up to 200 μM. Both OH-3' and OH-5 do not make direct interactions with the enzyme but are engaged in water-mediated hydrogen bonds with the enzyme, explaining their importance in the inhibitory effects on ppGalNAc-T2. In addition, bulkier substituents at position 3' might cause the destabilization of the flexible loop (Fig. 3A and Fig. 4), explaining the lack of potency of these compounds. Other flavonoids such as rhoifolin, which contains a polar dissacharide coupled to the 7-hydroxyl group in A ring, is a poor inhibitor of ppGalNAc-T2. In this case, the sugar moiety might cause steric hindrance with residues within the pocket or lead to the destabilization of the hydrophobic pocket (Fig. 3A and Fig. 4).

Although rutin, quercitrin and myricitrin have both hydroxyl groups, they also bear large substituents in position 3 of C ring as sugar moieties that greatly reduced their inhibitory potency (Fig. 4). In these compounds, the breaking of the planarity between the chromenone and phenyl rings (47), imposed by the sugar moieties, might lead to constraints in the structural disposition of the rings. This might disfavour an optimal interaction with Trp282 and Phe361 or cause steric hindrance with the mobile flexible loop, which might explain why these compounds are poor inhibitors of ppGalNAc-T2. On the contrary, quercetin or myricetin, luteolin-like compounds containing small substituents in position 3 of C ring, rendered similar inhibitory potencies to luteolin, suggesting that these flavonoids not only adopted a similar structure as luteolin but also the 3-hydroxyl group or small substituents in this position did not contribute to the binding of the enzyme (Fig. 3A and Fig. 4).

Luteolin selectively inhibits the catalytic activity of ppGalNAc-Ts on APP in cells

Luteolin exhibited enzyme selectivity on EA2 peptide in the *in vitro* enzyme assay and then we wondered whether it also has same selectivity in cells. Luteolin has been reported to reduce Aβ production in Alzheimer's disease (AD) model mouse (48,49). APP is the precursor of Aβ and also is a well-known *O*-GalNAc glycosylated protein in AD (5-7,50). Therefore, we chose APP as a protein substrate to test the specificity and selectivity of luteolin in cells.

We transiently co-expressed APP and different ppGalNAc-Ts into HEK 293T cells and treated the cells with luteolin (Fig. 5A). We found that all the tested ppGalNAc-Ts could glycosylate APP but the bands of *O*-GalNAc glycosylated APP (APP-H bands) for each ppGalNAc-Ts were different. Luteolin decreased the APP-H bands catalyzed by ppGalNAc-T2 or T3 but not by T1 or T13 (Fig. 5A), suggesting that luteolin selectively inhibited ppGalNAc-T isoforms in cells. In order to further confirm this observation, we synthesized three peptides from APP as the acceptor substrates for ppGalNAc-T activity assay (Fig. 5B and C). These peptides contained the dominant glycosylation sites (Thr^{291}, Thr^{292}, Thr^{576} and Thr^{577} for APP695; Thr^{353} for APP770), which have been reported crucial for APP processing (7,50). HPLC-based ppGalNAc-T assay showed that both peptide 1 and peptide 3 could be efficiently glycosylated by ppGalNAc-T2 and ppGalNAc-T3, while ppGalNAc-T1 showed very weak catalytic activity towards them (Fig. 5C). The peptide 2 could be only catalyzed by ppGalNAc-T3. ppGalNAc-T13 could not glycosylate all the three peptides (Fig. 5C). This result is consistent with Yun Kong 's report (12) and clearly indicated that ppGalNAc-T isoforms have different preference for glycosylation sites on APP. Here, ppGalNAc-T13 may glycosylate other sites in cells that are not covered in our peptides, such as sequential triple T (28,29) on APP. Furthermore, upon the treatment with luteolin, the product peaks on these peptides resulting from the ppGalNAc-T2 and ppGalNAc-T3 were largely blocked (Fig. 5C). However, there was no big difference in the peak produced by ppGalNAc-T1 after luteolin treatment (Fig. 5C). Moreover, we also co-expressed ppGalNAc-T1, T2 and APP in HEK 293T cells and treated with luteolin. The inhibitory effects of luteolin on the ppGalNAc-T2-catalyzed *O*-GalNAc glycans on APP could be masked by co-expressing ppGalNAc-T1 (Fig. 5D). These results are consistent with our observation in Fig. 5A and Fig. 2E and indicated that luteolin blocked the glycosylation on APP by inhibiting specific ppGalNAc-T isoforms not only in the *in vitro* enzyme assay but also in cells.

In addition to luteolin, we also compared the inhibitory effects of its analogues on the ppGalNAc-T2 using APP as the protein substrate in cell. Luteolin and quercetin showed pronounced inhibition (Fig. 5E). However, fisetin, acacetin, rutin and isorhamnetin had no effect on the *O*-GalNAc glycosylation of APP (Fig. 5E), which is consistent with our crystal structure and structure-function analysis (Fig. 3A and Fig. 4). The results above suggested that luteolin could selectively inhibit the activity of certain ppGalNAc-Ts on APP in cells.

Decrease in *O*-GalNAc glycosylation by luteolin substantially reduces the generation of Aβ

O-GalNAc glycosylation has been reported to influence APP processing (4,6,7), so we wonder whether the inhibition of ppGalNAc-T by luteolin can affect Aβ production. We treated the Swedish-mutation APP stable cell line (HEK 293T-APP Swe cell) with luteolin and found the production of APP-H and the amounts of total

Aβ, Aβ40 and Aβ42 were decreased in a dose-dependent manner (Fig. 6A and B). Meanwhile, the sAPPα and sAPPβ were also decreased as Aβ (Fig. 6B). Then we tested whether the reduction of Aβ was due to the decreased level of its *O*-GalNAc glycosylation. We analysed the levels of Aβ in the CHO-K1 cells and its *O*-GalNAc glycan-deficient derivative CHO-ldlD cells. The *O*-GalNAc glycosylation of proteins was different in various cell lines (51). Notably, only one band of APP was observed in CHO-K1 and CHO-ldlD cells, suggesting that the amount of *O*-GalNAc glycosylation of APP in CHO-K1 and CHO-ldlD cells may be lower than that in HEK 293T cells. The immunoprecipitation analysis with Jacalin clearly demonstrated that the APP was glycosylated in the CHO-K1 cells but not in CHO-ldlD cells (Fig. 6C). Consistent with the effects in HEK 293T cells, we found luteolin led to a significant decrease in the *O*-GalNAc glycans on APP and Aβ production in CHO-K1 cells (Fig. 6C and D). However, the Aβ generation in transfected CHO-ldlD cells was at least 60% lower than that in CHO-K1 cells (Fig. 6D). After the restoration of *O*-GalNAc glycosylation by adding exogenous Gal and GalNAc into CHO-ldlD cell culture medium, the production of Aβ was partly enhanced (Fig. 6E and F). However, luteolin largely decelerated the restoration process in both *O*-GalNAc glycosylation and Aβ production (Fig. 6E and F). The total Aβ, Aβ40 and 42 were reduced ~30% by luteolin treatment with 800 μM Gal&GalNAc (Fig. 6E and F). Additionally, we also treated APP/PS1 transgenic mice with luteolin at a dose of 50 mg/kg/day for two months and checked the *O*-GalNAc glycosylation level on APP and Aβ production in their brain homogenates. We found that *O*-GalNAc glycosylation signal on APP was significantly reduced as indicated by the immunoprecipitation analysis (Fig. 7A). The amounts of soluble Aβ40 and Aβ42 in the brains of APP/PS1 mice were reduced by nearly 30% (Fig. 7B). It preliminarily suggested that luteolin could affect *O*-GalNAc glycosylation of APP in mice. Taken together, these results suggested that decreased *O*-GalNAc glycosylation is a pivotal mechanism for the reduction of Aβ production after luteolin treatment.

DISCUSSION

Considering the increasing evidence for the involvement of ppGalNAc-Ts in the regulation of normal physiological functions as well as pathogenesis of various diseases (9,14,52), direct and bioactive ppGalNAc-T inhibitors could serve as useful tools for research and may have beneficial therapeutic effects for the treatment of some diseases. In this study, we reported for the first time that luteolin could directly inhibit ppGalNAc-Ts in a protein/peptide substrate-competitive manner and show selectivity in the ppGalNAc-T family. Additionally, we put forward that luteolin reduced the Aβ production at least partially by decreasing the *O*-GalNAc glycosylation via selectively inhibiting the activity of ppGalNAc-T isoforms, providing a putative mechanism for luteolin in modulating Alzheimer's disease pathologies.

ppGalNAc-T2 is a membrane-bound protein that needs detergent to make itself into active state (53). Luteolin could aggregate in solution, leading to promiscuous inhibition in the absence of detergent. Our experiments proved that in our reaction system with 0.2% TritonX-100, luteolin did not function in aggregate-based mechanism. In addition, based on our kinetic and crystal structure analysis, we clearly demonstrated that luteolin exhibited a classical competitive inhibitory mechanism against ppGalNAc-T2.

Our inhibition kinetics and X-ray structural demonstrated that luteolin acted as a protein

substrate-competitive inhibitor of ppGalNAc-T2 active form and competed specifically with the PxP motif present in most protein substrates (Fig. 3). The structure together with ^{1}H NMR experiments also revealed that two major residues, Trp282 and Phe361, contribute mainly to the binding and are responsible for the likely better affinity of the chromenone ring in opposition to the more flexible dihydroxyphenyl ring. The structure also provided the discovery of a druggable pocket that might lead to the design and synthesis of more potent luteolin-derivatives containing hydrophobic moieties in the 7-hydroxyl group or in position 6 of the chromenone ring. A combination of the crystal structure analysis together with the structure-inhibition relationship studies of the flavonoids on ppGalNAc-T2 provided more insights into the structural features responsible for an optimal inhibition, and offered other exploitable positions in luteolin for the design of more potent compounds (e.g. the OH-3' might be modified with polar groups to replace the water molecule visualized in the crystal structure, maximizing the interaction with the flexible loop).

Although several co-crystal studies showed that luteolin could also interact with other enzymes/proteins (22-27), there exits some difference between the interaction of luteolin with ppGalNAc-T2 and other targets. The characteristics of the surrounding amino acids are also quite different. Utilizing the differential interaction of luteolin with ppGalNAc-T2 and other targets could largely contribute to developing specific inhibitor.

The enzymatic activity assay *in vitro* and in cells showed that luteolin behaved as a pan inhibitor of different ppGalNAc-Ts isoforms including subfamilies Ib, Ic and IIb (Fig. 2E). The key Trp (Trp282 in ppGalNAc-T2) establishing π-π interactions with the 5,7-dihydroxy-4-chromenone ring is conserved in all above ppGalNAc-Ts except for ppGalNAc-T10, which contains an Arg instead (Arg296). This residue might be engaged in a stronger cation-π interaction with the larger ring of luteolin, which can explain why ppGalNAc-T10 is the most sensitive to luteolin. However, it was not expected that luteolin showed a poor inhibition of subfamily Ia (Fig. 2E), because a closer evaluation of multiple sequence alignments and structural conservation analysis between ppGalNAc-T2 and the above isoforms revealed that the luteolin-binding site was more conserved between subfamilies Ia and Ib as compared with the members of Ic and IIb subfamilies (42). This suggested that other more complex features such as the differences in the mobility of the flexible loop between different ppGalNAc-Ts, which in turn will likely depend on the surrounding residues, might account for the different selectivity of luteolin versus different isoforms.

The inhibitory effect of luteolin on the total level of *O*-GalNAc glycosylation in cells was modest. This is highly likely due to the selective inhibition of luteolin in ppGalNAc-Ts. Each ppGalNAc-T isoform has overlapping and unique substrate specificity. In spite of the modest inhibition on *O*-GalNAc glycosylation in cells, our work paved the way for further optimization based on the scaffold of luteolin to develop selective ppGalNAc-T inhibitor. Recently, increasing evidence indicates the importance of some specific glycosylation sites on certain glycoprotein during disease progression (13,51). *O*-GalNAc glycosylation is universal and important even for the normal cells. Complete abolishment of *O*-GalNAc glycosylation may lead to side-effect to cells. Therefore, the clear selectivity of luteolin in ppGalNAc-Ts may serve as a potential candidate for further studies of the site-specific inhibition

of *O*-GalNAc glycosylation on disease-associated glycoproteins. In that context, modest inhibitory effect at the total level in cells will not be an obstacle for its further application.

The excess generation of Aβ is regarded as a contributor to the dysfunction and degeneration of neurons that occurs in AD (8). APP is the precursor of Aβ and *O*-GalNAc glycosylation of APP is needed for its cell localization, endocytosis and plays critical roles in the Aβ production (4,6,7). In this study, we proved that ppGalNAc-T isoforms have different preference for glycosylation sites on APP and luteolin selectively inhibited a set of ppGalNAc-T isoforms including ppGalNAc-T2 and T3 but not T1 and T13 both *in vitro* and in cells (Fig. 5). Moreover, our restoration experiments with exogenous Gal and GalNAc in CHO-ldlD cells indicated that luteolin treatment significantly inhibited *O*-GalNAc glycosylation on APP and then led to the reduction of Aβ in cells (Fig. 6). These findings go in line with previous reports (48,49). Therefore, we inferred that luteolin may reduce the Aβ production at least partially by selectively inhibiting the activity of ppGalNAc-T isoforms. Since luteolin might also reduce the *O*-GalNAc glycosylation levels on other glycoproteins apart of APP, we cannot rule out other potential mechanisms of action for the reduction of Aβ. Actually, the sAPPα and sAPPβ were also reduced as Aβ (Fig. 6B) of HEK 293T-APP Swe cells with luteolin treatment, so we inferred that luteolin acted on APP and affected APP secretion before the activity of α, β and γ-secretase. It is possible that luteolin may regulate APP transporting out of the Golgi. Further experiments were needed to settle this issue. Nevertheless, we have proved that the *O*-GalNAc glycosylation inhibition is an important mechanism for luteolin in modulating AD pathologies, which may also applicable for other diseases such as cancer or inflammation.

In summary, although luteolin has been reported as interacting with several targets (22-27), we first showed that luteolin could directly inhibit ppGalNAc-Ts and reduce Aβ production by decreasing the *O*-GalNAc glycosylation. Our study represents a candidate scaffold structure for further optimization to develop specific ppGalNAc-T inhibitor. Considering that luteolin exhibited selectivity in ppGalNAc-T isoforms, further optimization based on this specific flavonoid may provide potent inhibitors to explore site-specific *O*-GalNAc glycosylation and a new avenue for developing drugs for AD and other diseases.

EXPERIMENTAL PROCEDURES

Chemicals and reagents — Uridine 5'-diphospho-N-acetylglucosamine sodium salt (UDP-GalNAc), *N*-acetylgalactosamine (GalNAc) and galactose (Gal) were purchased from Sigma-Aldrich (Steinheim, Germany). EDTA and manganese chloride from Sinopharm (Shanghai, China). EA2 and other peptides were synthesized by Scilight-Peptide (Beijing, China). Luteolin was from PI & PI Technology (Guangdong, China) in the highest purity available as powder.

Cloning, expression and purification of recombinant enzymes — The putative catalytic region and lectin domain of ppGalNAc-T2 (amino acids 52-571) were amplified by high-fidelity polymerase chain reaction (PCR) and cloned into the pFLAG-CMV-3 expression vector (Sigma) as described previously (10,28-30). FLAG-tagged ppGalNAc-T2 was expressed in HEK 293T cells, and the recombinant enzyme was purified from the cell medium using anti-FLAG M2 affinity gel (Sigma). Similarly, recombinant human ppGalNAc-T1 (amino acids 34-559), T3 (amino acids 50-633), T6 (amino acids 53-622), T10 (amino acids 34-604), T13 (amino acids 40-556)

and T14 (amino acids 39-552) were cloned and purified in the same way. The purity of recombinant ppGalNAc-Ts was tested by silver staining.

For the ppGalNAc-Ts used in the cellular experiments, full length ppGalNAc-Ts were cloned into vector-pcDNA3.1 and detected with Myc antibody in the lysates (Abmart, Shanghai, China).

HPLC-based ppGalNAc-T activity assay — This assay used a FAM-labelled EA2 peptide as the substrate and was performed as described previously (30). The reaction mixture contained the following components in a final volume of 20 μL: 10 ng of ppGalNAc-T2, 100 μM EA2, 100 μM UDP-GalNAc, 0.2% TritonX-100 (v/v), 5 mM $MnCl_2$, and 0.5 μL of DMSO or the tested compound in 25 mM HEPES buffer (the concentrations of both substrates approximate to the K_m). Under these conditions, mixture was incubated at 37 °C for 30 min in which the reaction rates were linear. The reaction mixture was boiled at 95 °C for 5 min to terminate the reaction and then seperated by reverse-phase-HPLC (Shimadzu, Kyoto, Japan) using the C18 analytical column (COSMOSIL 5C18-AR-II, 4.6 × 250 mm). The experiments were independently performed in triplicate, and the activity of the positive control (DMSO) was defined as 100%. In the dose-dependence experiments, selected compounds were tested at the indicated concentrations, and IC_{50} values were obtained using GraphPad Prism 5 software.

Similarly, the inhibitory activity of luteolin on ppGalNAc-T1, T3, T6, T10, T13 or T14 in the presence of the substrate EA2 peptide (Mono-GalNAc EA2 peptide for T10 (30)) and on ppGalNAc-T2 in the presence of the peptide substrate Muc1a (FAM-AHGVTSAPDTR), Muc5AC (SAPTTSTTSAPTK-FAM) or Muc7 (PTPSATTPAPPSSSAPPETTAAK-FAM) was accordingly measured under the same conditions.

Kinetic analysis — To explore the mode of action of luteolin on ppGalNAc-T2, kinetic analysis was performed using the HPLC-based ppGalNAc-T assay described above. The reaction rate was determined with ppGalNAc-T2 at the indicated concentrations of luteolin in the presence of the increasing concentrations of EA2 (0-1200 μM) with 1200 μM UDP-GalNAc or in the presence of the increasing concentrations of UDP-GalNAc (0-1200 μM) with 1200 μM EA2. The data were fitted using the Michaelis-Menten equation to determine the values of K_i or αK_i using GraphPad Prism 5. To illustrate the type of inhibition (competitive, non-competitive or mixed-type inhibition), the data were analysed via re-plots of the slopes of Lineweaver-Burk plots against varying inhibitor concentrations (Fig. 2C and D).

Purification of ppGalNAc-T2, crystallization, structure determination and refinement — The ppGalNAc-T2 was expressed in SMD1168 *Pichia pastoris*' strain, and purified using the purification protocol described previously (33). Crystals of the ppGalNAc-T2 were grown by hanging drops experiments at 18 ºC through mixing 1 μl of protein solution (a mix formed by 7 mg/ml ppGalNAc-T2, 5 mM UDP and 5 mM $MnCl_2$ in 25 mM Tris pH 8, 0.5 mM EDTA and 1 mM tris (2-carboxyethyl) phosphine (TCEP) with an equal volume of a reservoir solution (10% PEG 8000, 6% ethylene glycol, 100 mM Hepes pH 7). Under these conditions, crystals appeared within 2 days, and they were soaked overnight in a solution containing 7.5 mM of luteolin (the final DMSO concentration was 5%), 10% PEG 8000, 6% ethylene glycol, 100 mM HEPES pH 7. Then, the crystals were cryo-protected with 25% ethylene glycol, 10% PEG 8000, 100 mM HEPES pH 7, and frozen in a nitrogen gas stream cooled to 100 K. Processing and scaling were performed as before (33). The crystal

structure was solved and refined as explained previously for alike crystals (33).

NMR experiments — NMR samples were prepared in perdeuterated 25 mM TRIS-d11 in deuterated water, 7.5 mM NaCl, 1 mM β-mercaptoethanol, uncorrected pH 7.4, 50 μM UDP, 50 μM $MnCl_2$ and 8% deuterated dimethylsulfoxide. Luteolin was solved at 250 μM. The final concentration of the wild type protein and mutants was 2.5 μM. Experiments were performed on a Bruker AVANCE 500 MHz, provided with a TBO probehead and temperature was set at 303 K. Proton spectra were recorded with solven supression (Bruker sequence zgesgp), residual signal of HDO (4.70 ppm) was used for referencing chemical shifts (see Supplementary Fig. S5) and only the aromatic region is shown. Luteolin ^{1}H NMR (D_2O, 500 MHz): δ (ppm) 6.16 (s, 1H, H_A), 6.43 (s, 1H, H_B), 6.56 (s, 1H, H_C), 6.94-7.00 (m, 1H, H_E), 7.41-7.49 (m, 2H, H_D, H_F). The mutants W282A and F463A were introduced by site-directed mutagenesis by GenScript and using the template pPICZαA*galnact2* (K75-Q571). Both proteins were expressed and purified as the wild type enzyme (33).

Cell culture and transfection — HEK 293T-APP Swedish mutant cells (HEK 293T-APP Swe; a generous gift from Dr. Fang Wu) were maintained in Dulbecco's modified Eagle's medium (DMEM) (Life Technologies, Carlsbad, CA, USA), 10% foetal bovine serum (FBS; Life Technologies), 1% (w/v) penicillin and streptomycin (P/S; Life Technologies) and 200 μg/ml G418 (Life Technologies) in a humidified 5% CO_2 atmosphere at 37 °C as previously described (54). Wild-type HEK 293T and Jurkat cells were cultured in DMEM and RPMI-1640 (Life Technologies), respectively, supplemented with 10% FBS or 1% P/S. Chinese hamster ovary (CHO-K1) cells and CHO-ldlD cells were cultured in DME/F-12 (Life Technologies).

Plasmids containing human ppGalNAc-T1, T2, T3, T13 or FLAG-tagged wild-type APP695 (FLAG-tagged APP) were transfected into HEK293T cells using X-treme GENE 9 DNA Transfection Reagent (Roche, Mannheim, Germany) according to the manufacturer's instructions.

Evaluation of the activity of ppGalNAc-T in cells by using click chemistry to capture GalNAz-labelled proteins — CHO-ldlD cells were incubated with 100 μM Ac_4GalNAz (Thermo) for 12 h before treatment with indicated concentrations of luteolin for another 36 h. After the incubation, cells were lysed in 50 mM Tris-HCl buffer (pH 7.4) containing 1% Triton-X100 and the extracted proteins were quantified by BCA protein assay (Pierce). 1 mg cell lysate was mixed with 20 μM Biotin Alkyne (Invitrogen) and a cocktail of catalysts in PBS (pH 7.5), and rotated for 3 h at 25 °C. The ingredients of catalyst were 0.5 mM TBTA (tert-butyl 2,2,2-trichloroacetimidate, Sigma), 2 mM L-ascorbic acid sodium salt (Sinopharm Chemical Reagent Inc.) and 1 mM $CuSO_4$•$5H_2O$ (Sigma). Biotin-labelled proteins were then subjected to 10% SDS-PAGE and blotted with strepavidin (LI-COR Biosciences, Lincoln, Nebraska).

Western blot and immunoprecipitation (IP) — Cells were collected and lysed in 50 mM Tris-HCl buffer (pH 7.4) containing 1% Triton-X100 or 1% Nonidet P-40, 150 mM NaCl, and 1 × protease inhibitor cocktail (Roche). The extracted proteins were quantified by bicinchoninic acid (BCA) protein assay (Pierce, Rockford, IL, USA). Equal amounts of protein were subjected to 10% SDS-PAGE, transferred to membranes (GE Healthcare) and blotted with the indicated antibodies: anti-APP C-terminus antibody (C-T15, Sigma); anti-FLAG antibody

against FLAG-tagged APP or PDPN (Sigma), anti-Myc antibody against Myc-tagged ppGalNAc-Ts (Abmart), anti-ppGalNAc-T2 antibody (Sigma), anti-β-actin antibody or α-tubulin antibody (Sigma).

The IP of FLAG-tagged and endogenous APP was performed by incubating 1 mg of total protein with anti-FLAG affinity resin (Sigma) or C-T15 (Sigma), respectively, at 4 °C overnight. The C-T15 bound proteins were immunoprecipitated for 3 h at 4 °C using Protein A-agarose beads (Roche). The agarose beads were then washed three times with Tris-buffered saline (TBS), and the bound proteins were eluted with 2×Laemmli buffer. The resuspended proteins were subjected to Western blot analysis.

Lectin blot — The cell lysates were separated by 10% SDS-PAGE and transferred to cellulose acetate membranes (GE Healthcare). The membranes were washed with PBS containing 1% Tween-20 at room temperature (RT) prior to incubation with the Jacalin (primarily recognizing *O*-GalNAc glycans) and concanavalin A (ConA, primarily recognizing *N*-glycans) lectin conjugated with HRP (Vector Laboratories, Burlingame, CA, USA) in 1‰ Tween-20-PBS buffer at 4 °C overnight. The blots were visualized by enhanced chemiluminescence (ECL, Pierce).

Fluorescence imaging — After incubation with DMSO or luteolin at the indicated concentrations, Jurkat cells were collected and fixed on coverslips using Cytospin (Thermo, San Jose, CA, USA). The fixed cells were treated with 4% paraformaldehyde (Sangon Biotech, Shanghai, China) for 30 min at RT, washed with PBS, blocked with 5% bovineserum albumin (BSA, Sangon Biotech) in PBS for 1 h at RT and incubated with the *Helix pomatia* agglutinin (HPA, primarily recognizing *O*-GalNAc glycans) and ConA lectin conjugated with fluorescein isothiocyanate (FITC, Vector Laboratories) at 4 °C overnight. The cells were then incubated with 0.5 mL of 4', 6-diamidino-2-phenylindole (DAPI) (1.25 μg/mL) for 3 min and washed 3 times in PBS prior to acquiring images using an inverted fluorescence microscope (Nikon, Japan).

OGT activity assay with HPLC — OGT was expressed in *E. coli* and purified by nickel affinity chromatography and gel filtration as previous methods (40,41). The reaction mixtures containing 200 μM CKII peptide (KKKYPGGSTPVSSANMM), 100 nM ncOGT, 12.5 mM $MgCl_2$, 1 mM UDP-GlcNAc, buffer (150 mM NaCl, 20 mM Tris-HCl, 1 mM EDTA, 2.5 mM tris(hydroxypropyl)phosphine, pH 7.4) and indicated concentrations of luteolin were incubated at 37 °C for 30 min. After quenched by adding equal volume of methanol, the reaction mixtures were centrifuged at 12000 × g for 30 min. The supernatants (40 μL) were loaded onto an Agilent 1260 Infinity HPLC system to quantify the yield based on the integrated areas of glycopeptide product and peptide substrate. The reverse-phase chromatographic column was a Zorbax SB-C18 stablebond analytical column (4.6 mm × 250 mm, 5 μm; Agilent), preceded by a Zorbax SB-C18 analytical guard column (4.6 mm × 12.5 mm, 5 μm; Agilent). Mobile phase A consisted of 0.1% TFA in H_2O and mobile phase B consisted of 0.1% TFA in MeCN. The components were eluted using a gradient (flow rate at 1 mL/min; at 0 min elution solvent mixture A/B = 90/10; at 20 min elution solvent mixture A/B = 70/30; wavelength = 214 nm). Each reaction was repeated three times.

Dynamic Light Scattering — Particle formation was measured using Particle Size Analyzer, Z90s, Malvern Instruments Ltd., UK. Aggregator (50-1000 nm) sizes were measured in 25 mM HEPES buffer with different concentrations of TritonX-100 (v/v).

Experiments were repeated thrice and Each histogram shows a single representative sample.

Aβ, sAPPα and sAPPβ ELISA — HEK 293T, HEK 293T-APP Swe, CHO-K1, CHO-ldlD and other cells expressing exogenous APP or not; the corresponding cell medium or lysates were then harvested, and Aβ was quantified using Human Amyloid Assay Kits for Aβ40, Aβ42 and Aβ 1-x (IBL, Gumma, Japan) according to the manufacturer's protocols. sAPPα and sAPPβ were tested using Human sAPPα and sAPPβ kit (ExCell Bio, Shanghai, China) according to the protocols.

Mice — APP/PS1 transgenic mice (a total of 10, ♂) were obtained from the Model Animal Research Center of Nanjing University (B6C3-Tg (APPswe, PSEN1dE9)85Dbo/MmJNju). All the experiments were performed according to the guidelines established by the ethics committee of Shanghai Jiao Tong University. The mice were housed two or three per cage with free access to water and food in a temperature-controlled room and under natural lighting conditions with a 12/12 h light-dark cycle (lights on at 7:00). Mice were intraperitoneally injected with luteolin (50 mg/kg/day) or PBS from 4 months to 6 months of age, a period in which Aβ has been substantially emerged. The mice were then sacrificed, and the obtained brain homogenates of each mouse was extracted with 1% TritonX-100 and analysed for APP by Immunoprecipitation (anti-C-T15 antibody) and total Aβ, Aβ40 and Aβ42 by enzyme-linked immunosorbent assay (ELISA).

Statistical analysis and Western blot quantification — All experiments were performed at least twice in duplicate or triplicate with comparable results; the data are presented as the mean ± SD. Statistical analysis was conducted using one-way or two-way ANOVA with Bonferroni's multiple comparisons tests or two-tailed Student's t test; statistical significance is indicated as * ($P<0.05$), ** ($P<0.01$), ***($P<0.001$). The density of the bands was obtained in the linear range by choosing the suitable exposure time and followed by quantification by using Image J (Rawak Software, Inc.).

ACKNOWLEDGEMENTS

We thank Dr. Fang Wu (Shanghai Jiao Tong university) for cooperating with us in primary screening, gift of APP plasmid and HEK 293-APP Swe cells and helpful discussion. We thank Dr. Xuedong Liu (University of Colorado-Boulder) for his helpful discussion. This work was supported by the National Basic Research Program of China (No. 2012CB822103 and 2011CB910603 to Y. Z.), the National High Technology Research and Development Program of China (No. 2012AA020203 to Y. Z.), the National Natural Science Foundation (Grant Numbers 31170771 to Y.Z., 31370806 to Y.Z., 31570796 to Y.Z.), Agencia Aragonesa para la Investigación y Desarrollo (ARAID), Ministerio de Economía y Competitividad (CTQ2013-44367-C2-2-P and BFU2016-75633-P to R.H.-G.), Diputación General de Aragón (DGA; B89 to R.H.-G.) and the EU Seventh Framework Programme (2007–2013) under BioStruct-X (grantagreement 283570 and BIOSTRUCTX 5186, to R.H.-G.).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest with the contents of this article.

AUTHOR CONTRIBUTIONS

YZ and R H-G conceived the idea and supervised the study; FL, KX, ZX, MR, XL and ID performed the experiments; FL, PM, R H-G and YZ analyzed the experimental data; FL, JL, R H-G and YZ wrote and edited the manuscript. All authors approved the final version of the manuscript.

References

1. Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., Cooper, G. M., Roos, C., Voight, B. F., Havulinna, A. S., Wahlstrand, B., Hedner, T., Corella, D., Tai, E. S., Ordovas, J. M., Berglund, G., Vartiainen, E., Jousilahti, P., Hedblad, B., Taskinen, M. R., Newton-Cheh, C., Salomaa, V., Peltonen, L., Groop, L., Altshuler, D. M., and Orho-Melander, M. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189-197
2. Kato, K., Jeanneau, C., Tarp, M. A., Benet-Pages, A., Lorenz-Depiereux, B., Bennett, E. P., Mandel, U., Strom, T. M., and Clausen, H. (2006) Polypeptide GalNAc-transferase T3 and familial tumoral calcinosis. Secretion of fibroblast growth factor 23 requires O-glycosylation. *J Biol Chem* **281**, 18370-18377
3. Halim, A., Brinkmalm, G., Ruetschi, U., Westman-Brinkmalm, A., Portelius, E., Zetterberg, H., Blennow, K., Larson, G., and Nilsson, J. (2011) Site-specific characterization of threonine, serine, and tyrosine glycosylations of amyloid precursor protein/amyloid beta-peptides in human cerebrospinal fluid. *Proc Natl Acad Sci U S A* **108**, 11848-11853
4. Schedin-Weiss, S., Winblad, B., and Tjernberg, L. O. (2014) The role of protein glycosylation in Alzheimer disease. *FEBS J* **281**, 46-62
5. Weidemann, A., Konig, G., Bunke, D., Fischer, P., Salbaum, J. M., Masters, C. L., and Beyreuther, K. (1989) Identification, biogenesis, and localization of precursors of Alzheimer's disease A4 amyloid protein. *Cell* **57**, 115-126
6. Tomita, S., Kirino, Y., and Suzuki, T. (1998) Cleavage of Alzheimer's amyloid precursor protein (APP) by secretases occurs after O-glycosylation of APP in the protein secretory pathway. Identification of intracellular compartments in which APP cleavage occurs without using toxic agents that interfere with protein metabolism. *J Biol Chem* **273**, 6277-6284
7. Kitazume, S., Tachida, Y., Kato, M., Yamaguchi, Y., Honda, T., Hashimoto, Y., Wada, Y., Saito, T., Iwata, N., Saido, T., and Taniguchi, N. (2010) Brain endothelial cells produce amyloid {beta} from amyloid precursor protein 770 and preferentially secrete the O-glycosylated form. *J Biol Chem* **285**, 40097-40103
8. Glenner, G. G., Wong, C. W., Quaranta, V., and Eanes, E. D. (1984) The amyloid deposits in Alzheimer's disease: their nature and pathogenesis. *Appl Pathol* **2**, 357-369
9. Bennett, E. P., Mandel, U., Clausen, H., Gerken, T. A., Fritz, T. A., and Tabak, L. A. (2012) Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* **22**, 736-756
10. Peng, C., Togayachi, A., Kwon, Y. D., Xie, C., Wu, G., Zou, X., Sato, T., Ito, H., Tachibana, K., Kubota, T., Noce, T., Narimatsu, H., and Zhang, Y. (2010) Identification of a novel human UDP-GalNAc transferase with unique catalytic activity and expression profile. *Biochem Biophys Res Commun* **402**, 680-686
11. Gerken, T. A., Jamison, O., Perrine, C. L., Collette, J. C., Moinova, H., Ravi, L., Markowitz, S. D.,

Shen, W., Patel, H., and Tabak, L. A. (2011) 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

12. Kong, Y., Joshi, H. J., Schjoldager, K. T., Madsen, T. D., Gerken, T. A., Vester-Christensen, M. B., Wandall, H. H., Bennett, E. P., Levery, S. B., Vakhrushev, S. Y., and Clausen, H. (2015) Probing polypeptide GalNAc-transferase isoform substrate specificities by in vitro analysis. *Glycobiology* **25**, 55-65
13. Schjoldager, K. T., Joshi, H. J., Kong, Y., Goth, C. K., King, S. L., Wandall, H. H., Bennett, E. P., Vakhrushev, S. Y., and Clausen, H. (2015) Deconstruction of O-glycosylation--GalNAc-T isoforms direct distinct subsets of the O-glycoproteome. *EMBO Rep* **16**, 1713-1722
14. Schjoldager, K. T., and Clausen, H. (2012) Site-specific protein O-glycosylation modulates proprotein processing - deciphering specific functions of the large polypeptide GalNAc-transferase gene family. *Biochim Biophys Acta* **1820**, 2079-2094
15. Schjoldager, K. T., Vester-Christensen, M. B., Bennett, E. P., Levery, S. B., Schwientek, T., Yin, W., Blixt, O., and Clausen, H. (2010) O-glycosylation modulates proprotein convertase activation of angiopoietin-like protein 3: possible role of polypeptide GalNAc-transferase-2 in regulation of concentrations of plasma lipids. *J Biol Chem* **285**, 36293-36303
16. Agarwal, K., Kaul, R., Garg, M., Shajahan, A., Jha, S. K., and Sampathkumar, S. G. (2013) Inhibition of mucin-type O-glycosylation through metabolic processing and incorporation of N-thioglycolyl-D-galactosamine peracetate (Ac5GalNTGc). *J Am Chem Soc* **135**, 14189-14197
17. Song, L., and Linstedt, A. (2017) Inhibitor of ppGalNAc-T3-mediated O-glycosylation blocks cancer cell invasiveness and lowers FGF23 levels. *Elife* **6**, e24051
18. Hang, H. C., Yu, C., Ten Hagen, K. G., Tian, E., Winans, K. A., Tabak, L. A., and Bertozzi, C. R. (2004) Small molecule inhibitors of mucin-type O-linked glycosylation from a uridine-based library. *Chem Biol* **11**, 337-345
19. Ghirardello, M., de Las Rivas, M., Lacetera, A., Delso, I., Lira-Navarrete, E., Tejero, T., Martin-Santamaria, S., Hurtado-Guerrero, R., and Merino, P. (2016) Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-Polar Conjugates. *Chemistry* **22**, 7215-7224
20. Joel Mjiqiza, S., Abraham Syce, J., and Chibuzo Obikeze, K. (2013) Pulmonary effects and disposition of luteolin and Artemisia afra extracts in isolated perfused lungs. *J Ethnopharmacol* **149**, 648-655
21. Wang, S., Ruan, J., Zhuang, J., and Jiang, Q. (2000) The effect of Xiaofu Oral liquid on anti-inflammation, sedation and acute toxicity test. *Strait Pharmaceutical Journal* **12**, 23-25
22. Puhl, A. C., Bernardes, A., Silveira, R. L., Yuan, J., Campos, J. L., Saidemberg, D. M., Palma, M. S., Cvoro, A., Ayers, S. D., Webb, P., Reinach, P. S., Skaf, M. S., and Polikarpov, I. (2012) Mode of peroxisome proliferator-activated receptor gamma activation by luteolin. *Mol Pharmacol* **81**, 788-799
23. Trivella, D. B., dos Reis, C. V., Lima, L. M., Foguel, D., and Polikarpov, I. (2012) Flavonoid interactions with human transthyretin: combined structural and thermodynamic analysis. *J Struct Biol* **180**, 143-153
24. Yokoyama, T., Kosaka, Y., and Mizuguchi, M. (2015) Structural Insight into the Interactions between Death-Associated Protein Kinase 1 and Natural Flavonoids. *J Med Chem* **58**, 7400-7408
25. Narwal, M., Haikarainen, T., Fallarero, A., Vuorela, P. M., and Lehtio, L. (2013) Screening and

structural analysis of flavones inhibiting tankyrases. *J Med Chem* **56**, 3507-3517

26. Lolli, G., Cozza, G., Mazzorana, M., Tibaldi, E., Cesaro, L., Donella-Deana, A., Meggio, F., Venerando, A., Franchin, C., Sarno, S., Battistutta, R., and Pinna, L. A. (2012) Inhibition of protein kinase CK2 by flavonoids and tyrphostins. A structural insight. *Biochemistry* **51**, 6097-6107
27. Iakovleva, I., Begum, A., Pokrzywa, M., Walfridsson, M., Sauer-Eriksson, A. E., and Olofsson, A. (2015) The flavonoid luteolin, but not luteolin-7-O-glucoside, prevents a transthyretin mediated toxic response. *PLoS One* **10**, e0128222
28. Zhang, Y., Iwasaki, H., Wang, H., Kudo, T., Kalka, T. B., Hennet, T., Kubota, T., Cheng, L., Inaba, N., Gotoh, M., Togayachi, A., Guo, J., Hisatomi, H., Nakajima, K., Nishihara, S., Nakamura, M., Marth, J. D., and Narimatsu, H. (2003) Cloning and characterization of a new human UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase, designated pp-GalNAc-T13, that is specifically expressed in neurons and synthesizes GalNAc alpha-serine/threonine antigen. *J Biol Chem* **278**, 573-584
29. Xu, Y., Pang, W., Lu, J., Shan, A., and Zhang, Y. (2016) Polypeptide N-acetylgalactosaminyltransferase 13 contributes to neurogenesis via stabilizing the mucin-type O-glycoprotein podoplanin. *J Biol Chem* **291**, 23477-23488
30. Li, X., Wang, J., Li, W., Xu, Y., Shao, D., Xie, Y., Xie, W., Kubota, T., Narimatsu, H., and Zhang, Y. (2012) Characterization of ppGalNAc-T18, a member of the vertebrate-specific Y subfamily of UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferases. *Glycobiology* **22**, 602-615
31. Fritz, T. A., Raman, J., and Tabak, L. A. (2006) Dynamic association between the catalytic and lectin domains of human UDP-GalNAc:polypeptide alpha-N-acetylgalactosaminyltransferase-2. *J Biol Chem* **281**, 8613-8619
32. Lira-Navarrete, E., de Las Rivas, M., Companon, I., Pallares, M. C., Kong, Y., Iglesias-Fernandez, J., Bernardes, G. J., Peregrina, J. M., Rovira, C., Bernado, P., Bruscolini, P., Clausen, H., Lostao, A., Corzana, F., and Hurtado-Guerrero, R. (2015) Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein O-glycosylation. *Nat Commun* **6**, 6937
33. Lira-Navarrete, E., Iglesias-Fernandez, J., Zandberg, W. F., Companon, I., Kong, Y., Corzana, F., Pinto, B. M., Clausen, H., Peregrina, J. M., Vocadlo, D. J., Rovira, C., and Hurtado-Guerrero, R. (2014) Substrate-guided front-face reaction revealed by combined structural snapshots and metadynamics for the polypeptide N-acetylgalactosaminyltransferase 2. *Angew Chem Int Ed Engl* **53**, 8206-8210
34. Kingsley, D. M., Kozarsky, K. F., Hobbie, L., and Krieger, M. (1986) Reversible defects in O-linked glycosylation and LDL receptor expression in a UDP-Gal/UDP-GalNAc 4-epimerase deficient mutant. *Cell* **44**, 749-759
35. Slade, P. G., Hajivandi, M., Bartel, C. M., and Gorfien, S. F. (2012) Identifying the CHO secretome using mucin-type O-linked glycosylation and click-chemistry. *J Proteome Res* **11**, 6175-6186
36. Kaneko, M. K., Kato, Y., Kameyama, A., Ito, H., Kuno, A., Hirabayashi, J., Kubota, T., Amano, K., Chiba, Y., Hasegawa, Y., Sasagawa, I., Mishima, K., and Narimatsu, H. (2007) Functional glycosylation of human podoplanin: glycan structure of platelet aggregation-inducing factor. *FEBS Lett* **581**, 331-336
37. Pan, Y., Yago, T., Fu, J., Herzog, B., McDaniel, J. M., Mehta-D'Souza, P., Cai, X., Ruan, C., McEver, R. P., West, C., Dai, K., Chen, H., and Xia, L. (2014) Podoplanin requires sialylated O-glycans for stable

expression on lymphatic endothelial cells and for interaction with platelets. *Blood* **124**, 3656-3665

38. Pohjala, L., and Tammela, P. (2012) Aggregating behavior of phenolic compounds--a source of false bioassay results? *Molecules* **17**, 10774-10790
39. Wang, Y., Zhu, J., and Zhang, L. (2016) Discovery of Cell-Permeable O-GlcNAc Transferase Inhibitors via Tethering in Situ Click Chemistry. *J Med Chem* **60**, 263-272
40. Lazarus, M. B., Nam, Y., Jiang, J., Sliz, P., and Walker, S. (2011) Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* **469**, 564-567
41. Kreppel, L. K., Blomberg, M. A., and Hart, G. W. (1997) Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. *J Biol Chem* **272**, 9308-9315
42. Revoredo, L., Wang, S., Bennett, E. P., Clausen, H., Moremen, K. W., Jarvis, D. L., Ten Hagen, K. G., Tabak, L. A., and Gerken, T. A. (2016) 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
43. Sarrazin, M., Sari, J. C., Bourdeaux-Pontier, M., and Briand, C. (1979) NMR study of the interactions between flurazepam and human serum albumin. The nature of the complexation site on the benzodiazepin molecule. *Mol Pharmacol* **15**, 71-77
44. Fischer, J. J., and Jardetzky, O. (1965) Nuclear Magnetic Relaxation Study of Intermolecular Complexes. The Mechanism of Penicillin Binding to Serum Albumin. *J Am Chem Soc* **87**, 3237-3244
45. Fielding, L. (2003) NMR methods for the determination of protein-ligand dissociation constants. *Curr Top Med Chem* **3**, 39-53
46. Singh, M., Kaur, M., and Silakari, O. (2014) Flavones: an important scaffold for medicinal chemistry. *Eur J Med Chem* **84**, 206-239
47. Cai, W., Chen, Y., Xie, L., Zhang, H., and Hou, C. (2014) Characterization and density functional theory study of the antioxidant activity of quercetin and its sugar-containing analogues. *European Food Research and Technology* **238**, 121-128
48. Rezai-Zadeh, K., Douglas Shytle, R., Bai, Y., Tian, J., Hou, H., Mori, T., Zeng, J., Obregon, D., Town, T., and Tan, J. (2009) Flavonoid-mediated presenilin-1 phosphorylation reduces Alzheimer's disease beta-amyloid production. *J Cell Mol Med* **13**, 574-588
49. Theoharides, T. C., Stewart, J. M., Hatziagelaki, E., and Kolaitis, G. (2015) Brain "fog," inflammation and obesity: key aspects of neuropsychiatric disorders improved by luteolin. *Front Neurosci* **9**, 225
50. Perdivara, I., Petrovich, R., Allinquant, B., Deterding, L. J., Tomer, K. B., and Przybylski, M. (2009) Elucidation of O-glycosylation structures of the beta-amyloid precursor protein by liquid chromatography-mass spectrometry using electron transfer dissociation and collision induced dissociation. *J Proteome Res* **8**, 631-642
51. Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T., Lavrsen, K., Dabelsteen, S., Pedersen, N. B., Marcos-Silva, L., Gupta, R., Bennett, E. P., Mandel, U., Brunak, S., Wandall, H. H., Levery, S. B., and Clausen, H. (2013) Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J* **32**, 1478-1488
52. Schjoldager, K. T., Vakhrushev, S. Y., Kong, Y., Steentoft, C., Nudelman, A. S., Pedersen, N. B., Wandall, H. H., Mandel, U., Bennett, E. P., Levery, S. B., and Clausen, H. (2012) Probing isoform-specific functions of polypeptide GalNAc-transferases using zinc finger nuclease

glycoengineered SimpleCells. *Proc Natl Acad Sci U S A* **109**, 9893-9898

53. Sugiura, M., Kawasaki, T., and Yamashina, I. (1982) Purification and characterization of UDP-GalNAc:polypeptide N-acetylgalactosamine transferase from an ascites hepatoma, AH 66. *J Biol Chem* **257**, 9501-9507

54. Haass, C., Capell, A., Citron, M., Teplow, D. B., and Selkoe, D. J. (1995) The vacuolar H(+)-ATPase inhibitor bafilomycin A1 differentially affects proteolytic processing of mutant and wild-type beta-amyloid precursor protein. *J Biol Chem* **270**, 6186-6192

FOOTNOTES

The abbreviations used in this article are: APP, amyloid precursor protein; AD, Alzheimer's disease; Aβ, amyloid β; Gal, galactose; GalNAc, α-N-acetylgalactosamine; Man, mannose; ppGalNAc-T, polypeptide *N*-acetyl-α-galactosaminyltransferase; PDPN, podoplanin.

FIGURE LEGENDS

Fig. 1 Luteolin functions as a potent inhibitor *in vitro* and in cells. (**A**) Schematic representation of the initiation of *O*-GalNAc glycosylation by ppGalNAc-Ts. (**B**) Chemical structure and IC_{50} of luteolin against ppGalNAc-T2 detected by HPLC-based ppGalNAc-T assay. The data are expressed as percentages of the control (DMSO, 100%) and are shown as the mean ± SD (n=3). (**C**) Effects of luteolin on the *O*-GalNAc glycosylation in CHO-ldlD cells detected by the metabolic labelling and click chemistry. CHO-ldlD cells were pre-incubated with 100 μM Ac4GalNAz for 12 h and then treated with indicated concentrations of luteolin for additional 36 h. Total cell lysates were collected to react with biotin alkyne for 3 h before analysed by streptavidin blotting. (**D**) Effects of luteolin on glycans of proteins in HEK 293T cells. Cells were incubated with luteolin at 25 μM for 36 h and total lysates were analysed by lectin blot. Control, DMSO-treated cells. (**E**) Effects of luteolin on glycans in Jurkat cells detected by immunofluorescence. Cells were treated with 25 μM luteolin for 24 h and stained with different lectins. HPA and ConA, green; nuclei were stained with DAPI, blue. Scale bars, 50 μm. Control, DMSO-treated cells. (**F, G**) Effects of luteolin on the glycosylation of endogenous APP in HEK 293T cells (**F**) and A549 cells (**G**). Cell were incubated with 25 μM luteolin for 36 h and collected and analysed by western blot. L, long time exposure; S, short time exposure. (**H, I**) Effects of luteolin on the glycosylation of transfected APP (**H**) and PDPN (**I**) in HEK 293T cells. FLAG-tagged APP and PDPN were transfected for 24 h and then incubated with 25 μM luteolin for another 24 h**.** APP was then immunoprecipitated from the cell lysates (Input) using an anti-FLAG M2 affinity resin. The mean density of APP-H or PDPN-H was quantified and normalized with APP-L or PDPN-L. Cells expressing APP or PDPN without luteolin treatment, 100%. IP, immunoprecipitation. 50 μg total protein was used to analyse endogenous APP and 10 μg to transfected APP and PDPN. The data shown represent as mean ± SD (n=3). *, $P<0.05$; **, $P<0.01$; ***, $P< 0.001$. Two-tailed Student's t-test for (**F-I**). β-actin and α-tubulin were used as loading control. EV, empty vector.

Fig. 2 Kinetics and selectivity analyses of luteolin inhibition mechanism on ppGalNAc-T2. (**A**) The change of percentage of particles in 50-1000 nm size range with luteolin and different concentrations of TritonX-100 (v/v). The size of particles was measured by dynamic light scattering (DLS). (**B**) Spin down experiments for luteolin in our HPLC-based ppGalNAc-T assay with 0.2%

TritonX-100 (v/v). Enzyme reactions were performed before or after centrifugation (16000×g, 1 h). The data are expressed as percentages of the control (DMSO, 100%). Data above are shown as the mean ± SD (n=3). Two-tailed Student's t-test, n.s., no significant difference. (**C**) Luteolin shows competitive inhibition (K_i=1.4 ± 0.05 μM) with respect to EA2 substrate in the presence of 1200 μM UDP-GalNAc. (**D**) UDP-GalNAc shown non-competitive inhibition (αK_i=4.1 ± 1.1 μM) in the presence of 1200 μM EA2. The concentration (μM) of luteolin was 0 (●), 5 (▲), 10 (■) and 20 (◆). (**E**) The kinetics parameter of luteolin inhibition on ppGalNAc-T2 with various peptide substrates and the selectivity of luteolin within ppGalNAc-T family with EA2 peptide. [a]Competitive inhibition constant; [b]Non-competitive inhibition constant; OGT, *O*-GlcNAc transferase. The data shown in (**E**) are represented as mean ± SD (n=3).

Fig. 3 X-ray structure and ^{1}H NMR binding experiments of ppGalNAc-T2/mutants with luteolin. (**A**) View (left panel) of luteolin-binding site of the ppGalNAc-T2-UDP-luteolin complex. The crystal structure of ppGalNAc-T2 in complex with UDP and MUC5AC-13 (32) is shown for comparison purposes (insert). A surface representation (right panel) of the ppGalNAc-T2-UDP-luteolin complex depicts the pocket present in these enzymes. Part of the residues forming the luteolin-binding site and the pocket are shown in grey whereas residues of the flexible loop are depicted in yellow. Luteolin/MUC5AC-13 and UDP are shown as green and brown carbon atoms, respectively. Ethylenglycol is shown as grey carbon atoms whereas water molecules are depicted as red spheres. Hydrogen bond interactions are shown as dotted deep olive lines. Electron density maps are F_O–F_C syntheses (blue) contoured at 2.2 σ for luteolin. (**B**) View of the two additional luteolin molecules present in the AU. (**C**) Overlay of the ppGalNAc-T2 molecules in complex with luteolin showing the flexibility mainly in the dihydroxyphenyl moiety. Luteolin and Trp282/Phe361 are depicted as green, yellow and brown carbon atoms. (**D**) Overlay of one of the monomers containing luteolin with ppGalNAc-T2 inactive form in complex with UDP (PDB entry 2FFV (31)) and ppGalNAc-T2 inactive form in complex with MUC5AC-Cys13 (PDB entry 5AJN (32)). This figure depicts that luteolin has steric hindrance with both inactive forms of ppGalNAc-T2. The flexible loop for the active and inactive forms is shown in yellow and blue/brown, respectively. (**E**) ^{1}H NMR spectra were recorded to compare the effect of binding on the luteolin signal broadening. ^{1}H spectra of free luteolin, luteolin with ppGalNAc-T2 wild-type and the mutants W282A and F463A.

Fig. 4 Structure-function analysis of luteolin-analogues using HPLC-based ppGalNAc-T assay with EA2 peptide. [a]The IC_{50} are calculated as percentages of the control (DMSO, 100%) and shown as the mean ± SD (n=3). [b]The inhibitory ratio (%) was calculated with the positive control (DMSO, 0%) (50 μM compounds). [c]Isorhamnetin did not show any inhibitory activity at 200 μM. [d]The inhibitory ratio of fisetin at 200 μM is 23.87 ± 0.4 %. N.I., no inhibition.

Fig. 5 Luteolin could inhibit the catalytic activity of ppGalNAc-Ts on APP in cells. (**A**) Effects of luteolin on the glycosylation on APP in HEK 293T cells overexpressing ppGalNAc-T1, T2, T3 or T13. FLAG-tagged APP alone or together with T1, T2, T3 or T13 were transfected into cells for 24 h before treating with 20 μM luteolin for additional 24 h. The cell lysates were collected and analysed for APP. (**B**) Schematic diagrams of synthesized APP-derived peptides. *O*-GalNAc glycosylation sites

are marked with *black dots* (7). KPI, Kunitz-type protease inhibitor domain; OX2, OX2 domain (7); α, β and γ, cleavage sites of α, β and γ-secretases. (**C**) Luteolin (50 μM) can inhibit the glycosylation on APP-derived peptides catalyzed by ppGalNAc-T2 or T3. S, substrate peak; P, product peak. (**D**) ppGalNAc-T1 attenuates the inhibitory effect of luteolin on *O*-GalNAc glycosylation of APP catalyzed by T2. (**E**) Effects of luteolin analogues on the *O*-GalNAc glycosylation of APP catalyzed by ppGalNAc-T2. For (**D**, **E**), different combinations of APP, ppGalNAc-T2 or T1 plasmids were transfected into HEK 293T cells and then treated with indicated compounds (20 μM) and analysed as (**A**). For (**A**), (**D**) and (**E**), the mean density of APP-H was quantified and normalized with APP-L, and compared with the control (DMSO-treated cells co-transfected with APP and ppGalNAc-Ts without luteolin treatment, 100%). The data shown in (**A**), (**D**), (**E**) represent as mean ± SD (n=3). *, P<0.05; **, P<0.01; ***, P< 0.001; two-tailed Student's t-test for (**A**) and one-way ANOVA for (**D**) and (**E**). EV, empty vector.

Fig. 6 Luteolin reduces the *O*-GalNAc glycosylation on APP and leads to decrease of the Aβ production in cells. (**A, B**) Dose-dependent effect of luteolin on the *O*-GalNAc glycosylation of APP and Aβ production in HEK 293T-APP Swe cells. Cells were incubated with luteolin for 24 h and then lysed and blotted for APP. The mean density of APP-H was quantified and normalized with APP-L and presented as percentages of control (DMSO, 100%). Aβ, sAPPα and sAPPβ in the cell medium above were quantified by ELISA and normalized to the amount of total protein and presented as percentages of control (cells without luteolin treatment, 100%). (**C**) Effects of luteolin on *O*-GalNAc glycans of APP in CHO-K1 and CHO-ldlD cells. Cells were transfected with FLAG-tagged APP for 24 h followed by incubation with 25 μM luteolin for additional 24 h. APP was then immunoprecipitated and detected by Jacalin staining**.** (**D**) Aβs in the cell medium above were analysed by ELISA. Control, CHO-K1 cells expressing APP without luteolin treatment, 100%. (**E**) Adding exogenous Gal and GalNAc restores the *O*-GalNAc glycosylation on APP in CHO-ldlD cells but these effects could be attenuated by luteolin treatment (25 μM). (**F**) Aβs in the cell medium above were quantified by ELISA and represented as Aβ (pg) per total protein (mg). ○, DMSO treatment; ●, 25 μM luteolin treatment. The mean density of Jacalin in (**C**) and (**E**) was quantified and normalized with APP and presented as percentages of control (cells without luteolin treatment, 100%). Data represent as mean ± SD, n=3. *, P<0.05; **, P<0.01; ***, P< 0.001; for (**A**), one-way ANOVA; for Aβ in (**B**), (**D**) and (**F**), two-way ANOVA; for sAPPα and sAPPβ in (**B**), (**C**) and (**E**), two-tailed Student's t-test. EV, empty vector.

Fig. 7 The change of *O*-GalNAc glycosylation of APP and Aβ production in APP/PS1 transgenic mice with luteolin treatment. (**A**) The *O*-GalNAc glycosylation of APP in extracts from brain homogenates were analysed by immunoprecipitation (IP) of anti-APP antibody (C-T15). The mean density of Jacalin was quantified. (**B**) Soluble Aβs from brain homogenates of mice were detected by ELISA. Data are presented as percentage of control (PBS-treated mice, 100%) and shown as mean ± SD (n=3 for Jacalin staining and n=5 for Aβ detection). *, P<0.05; **, P<0.01; two-tailed Student's t-test.

Figure 1

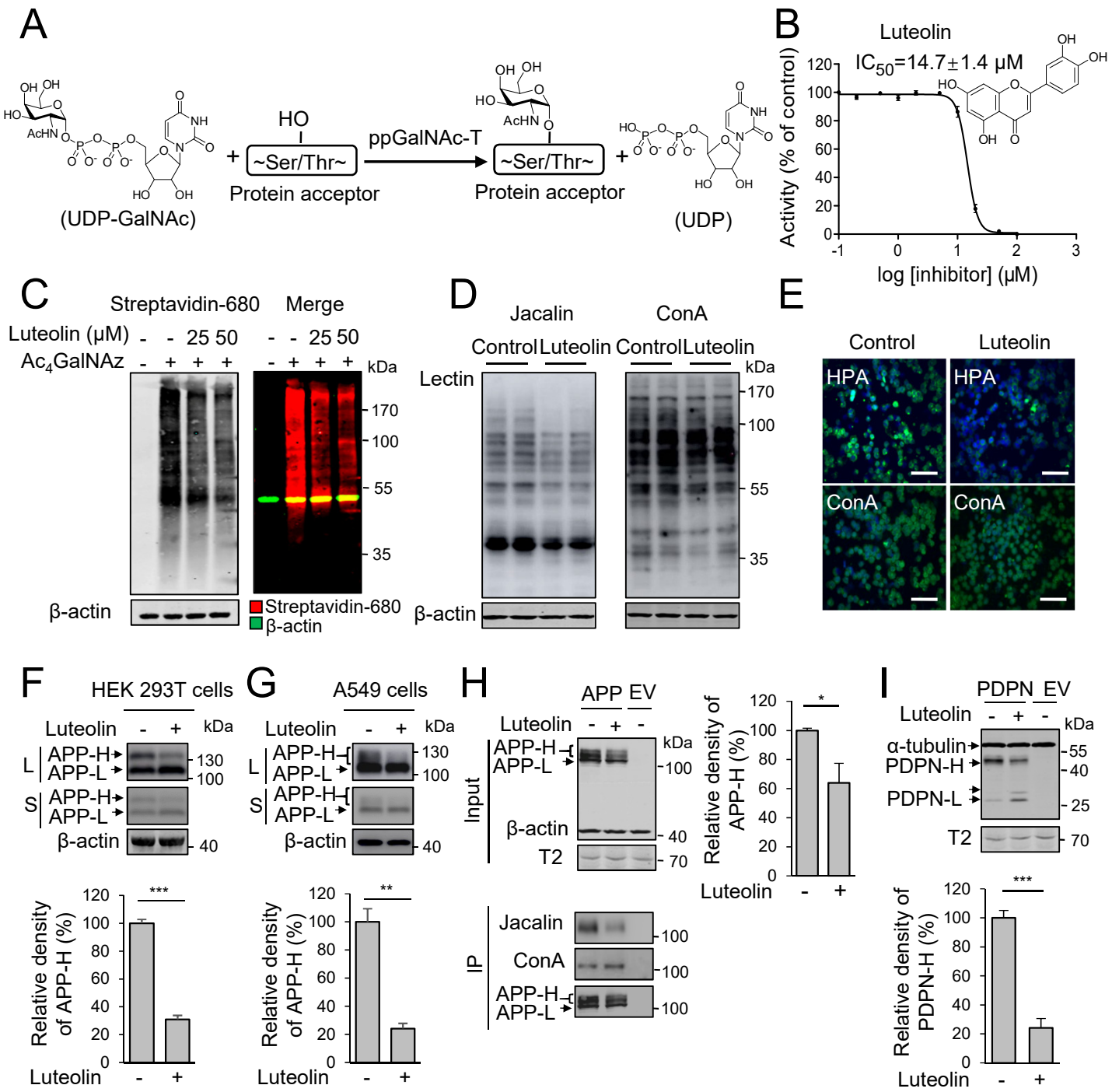
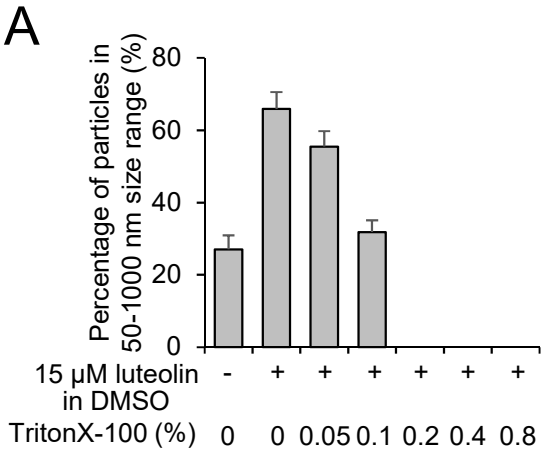
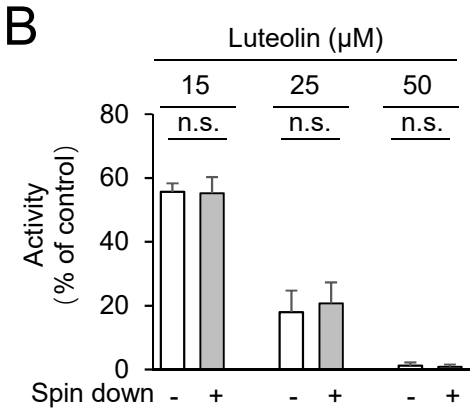
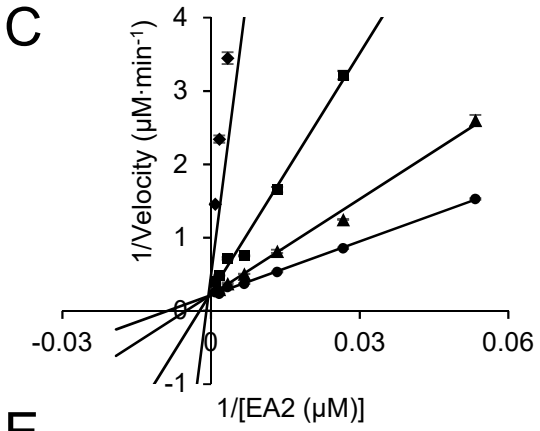
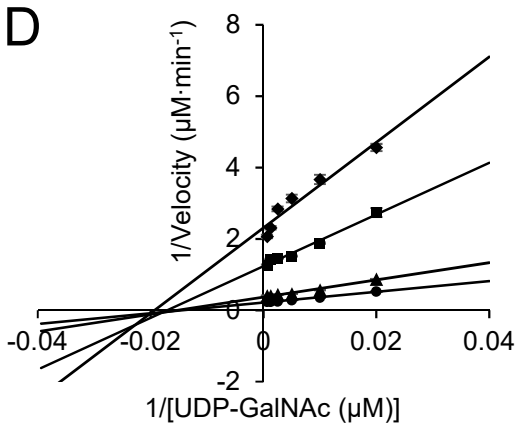


Figure 2









E

Inhibition of ppGalNAc-T2 with different peptides			Enzyme specificity		
Parameter	Peptide	Value (μM)	Subfamily	Enzyme	IC_{50} (μM)
K_m (EA2)	EA2	160.2 ± 11.5	Ib	ppGalNAc-T2	14.7 ± 1.4
K_m (UDP-GalNAc)	EA2	59.7 ± 4.6		ppGalNAc-T14	6.0 ± 0.2
K_i [a]	EA2	1.4 ± 0.05	Ic	ppGalNAc-T3	13.9 ± 1.1
αK_i [b]	EA2	4.1 ± 1.1		ppGalNAc-T6	15.4 ± 0.7
IC_{50}	EA2	14.7 ± 1.4	IIb	ppGalNAc-T10	1.7 ± 0.3
IC_{50}	Muc1a	4.9 ± 0.1	Ia	ppGalNAc-T1	>200
IC_{50}	Muc5AC	13.7 ± 2.0		ppGalNAc-T13	>200
IC_{50}	Muc7	6.6 ± 0.2		OGT	>200

Figure 3

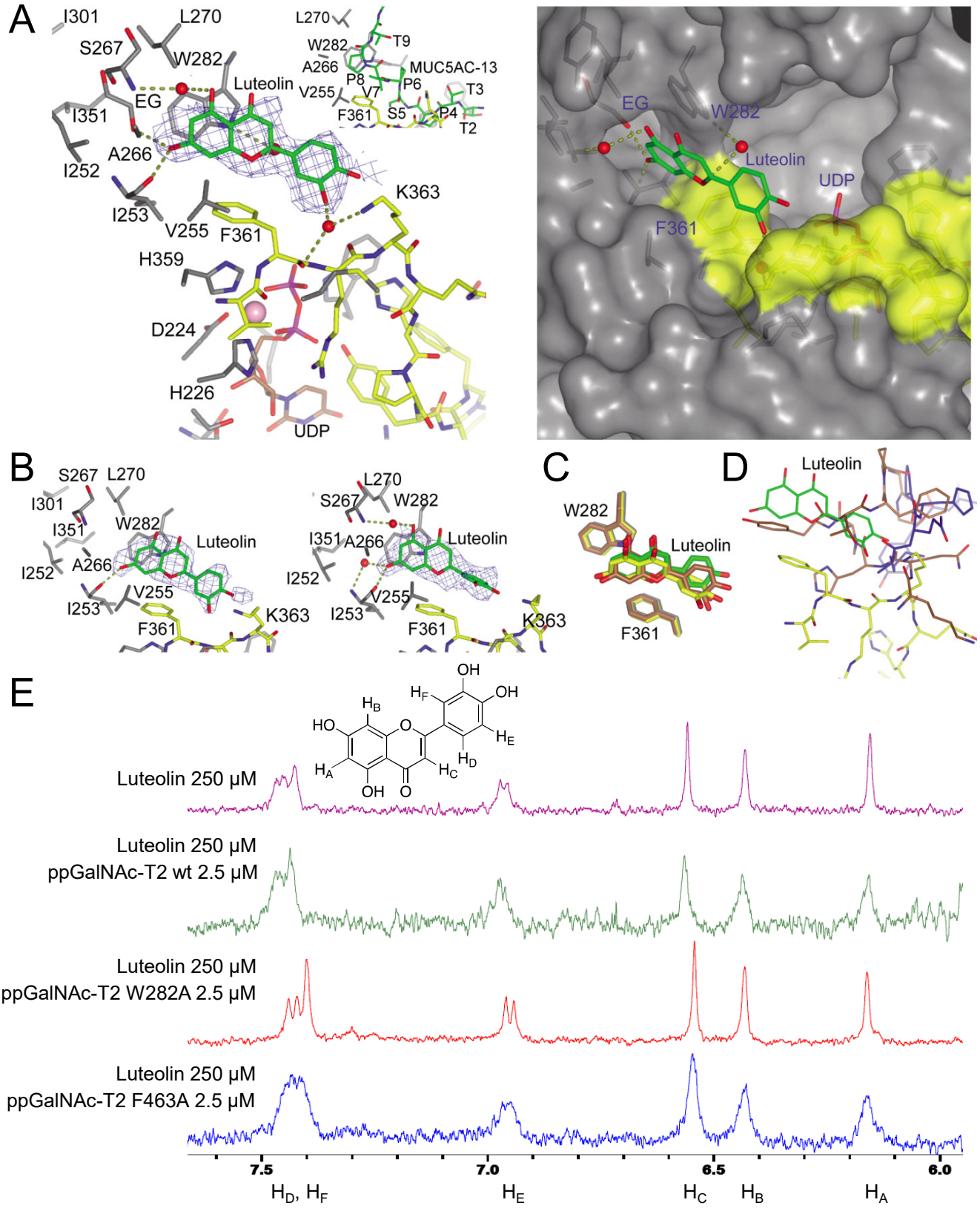


Figure 4

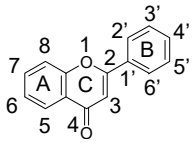
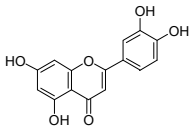
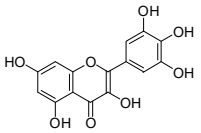
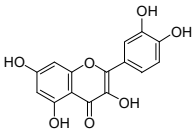
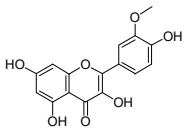
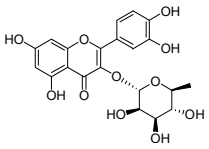
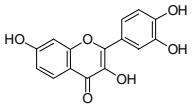
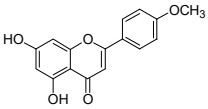
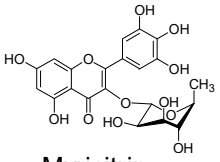
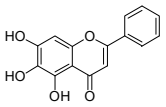
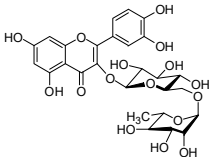
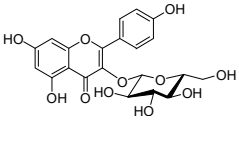
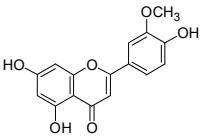
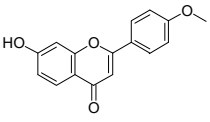
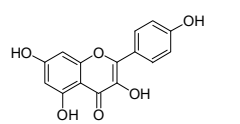
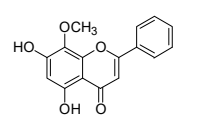
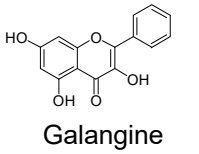
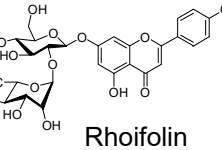
	IC_{50}[a]	Inhibitory ratio (%)[b]	IC_{50}[a]	Inhibitory ratio (%)[b]	IC_{50}[a]	Inhibitory ratio (%)[b]
Active compounds						
	14.7 ± 1.4	98.7 ± 1.8		1.9 ± 0.1	98.7 ± 0.67		11.7 ± 1.6	97.5 ± 1.8
Luteolin			Myricetin			Quercetin		
Inactive compounds								
	>50	N.I.		>50	15.2 ± 0.3		>50	N.I.
Isorhamnetin[c]			Quercitrin			Fisetin[d]		
	>50	N.I.		>50	6.2 ± 1.1		>50	N.I.
Acacetin			Myricitrin			Baicalein		
	>50	N.I.		>50	N.I.		>50	N.I.
Rutin			Kaempferol-3 -O-glucoside			Chrysoeriol		
	>50	N.I.		>50	N.I.		>50	N.I.
Pratol			Kaempferol			Wogonin		
	>50	N.I.		>50	N.I.		>50	N.I.
Galangine			Rhoifolin					

Figure 5

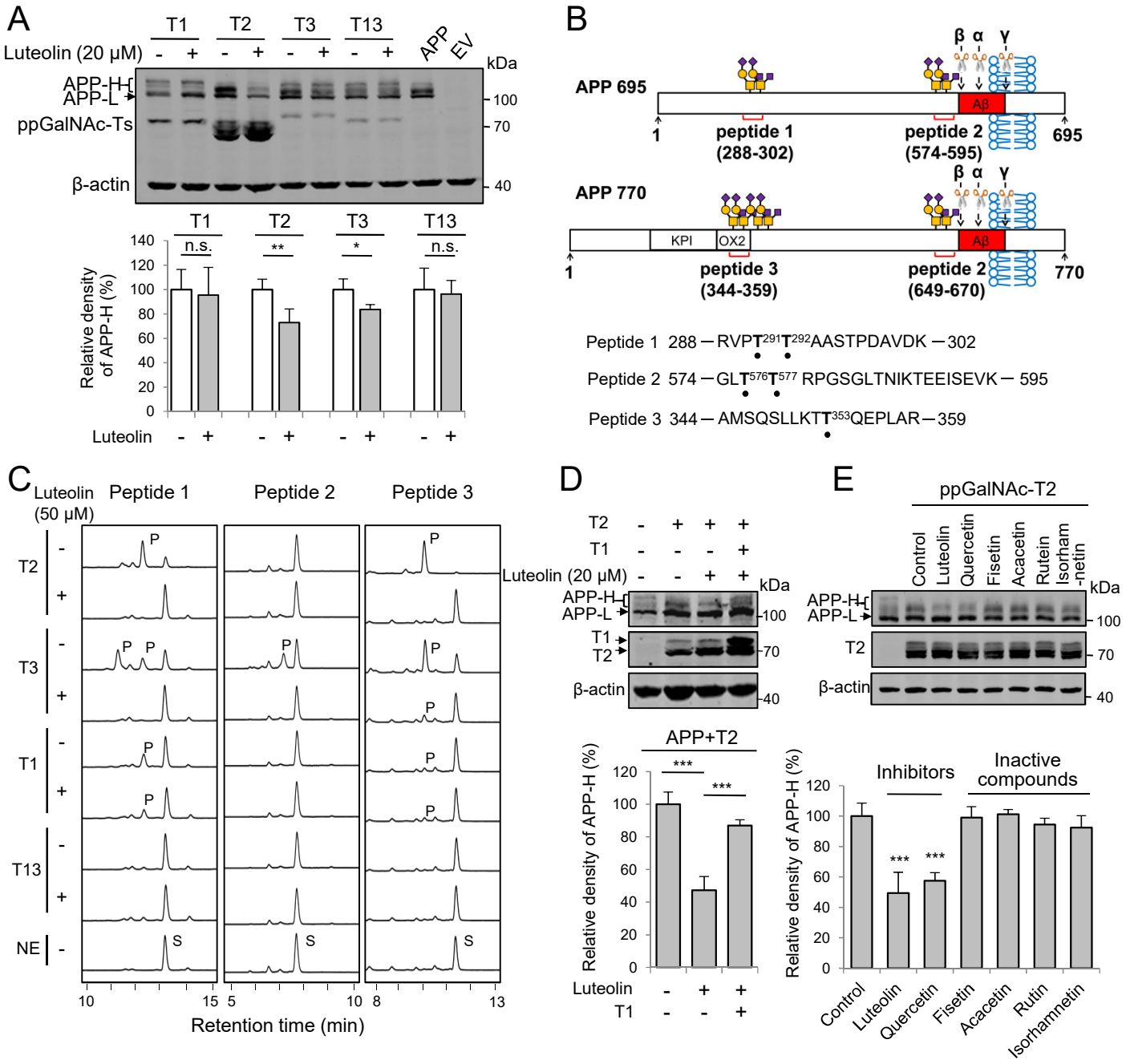


Figure 6

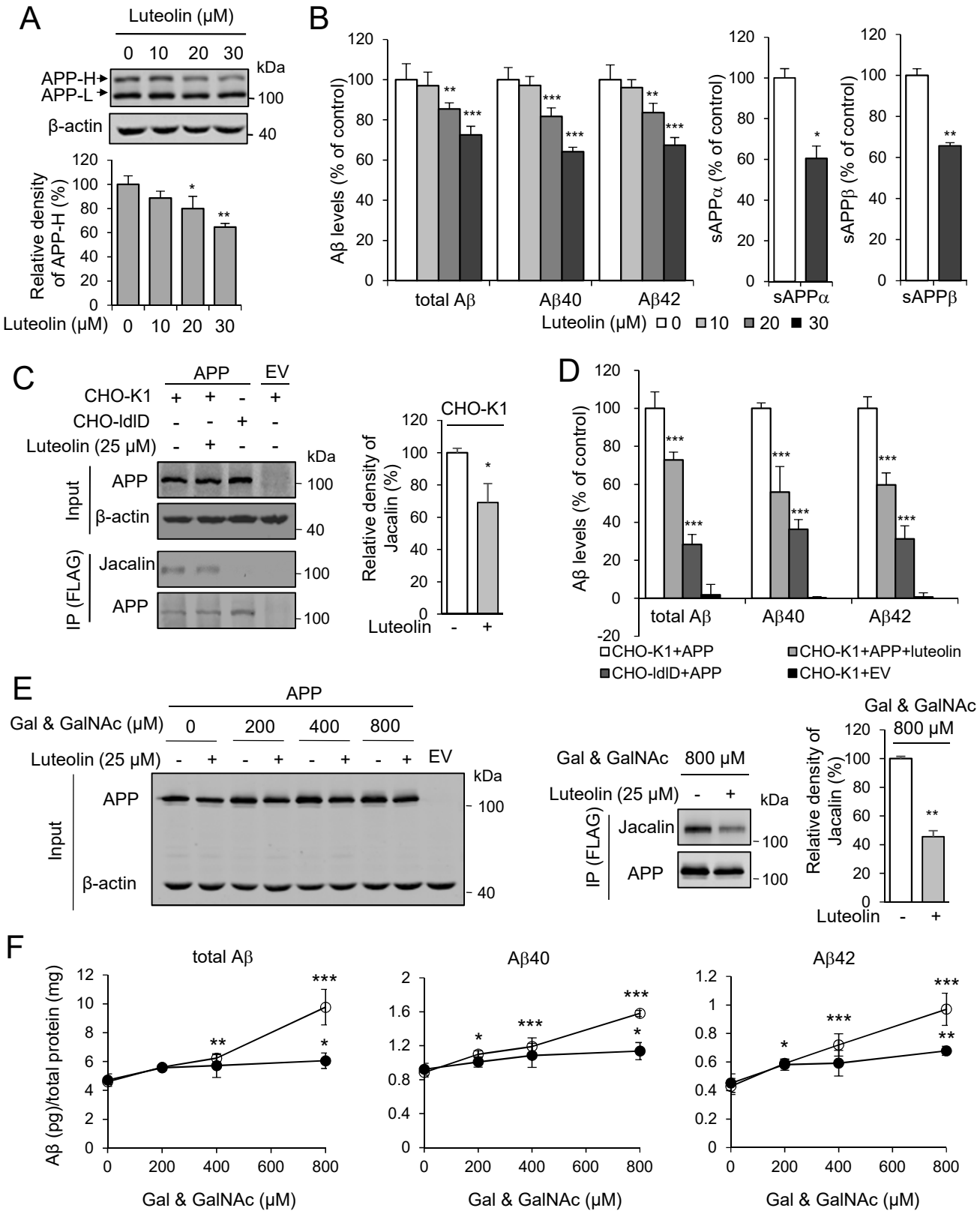
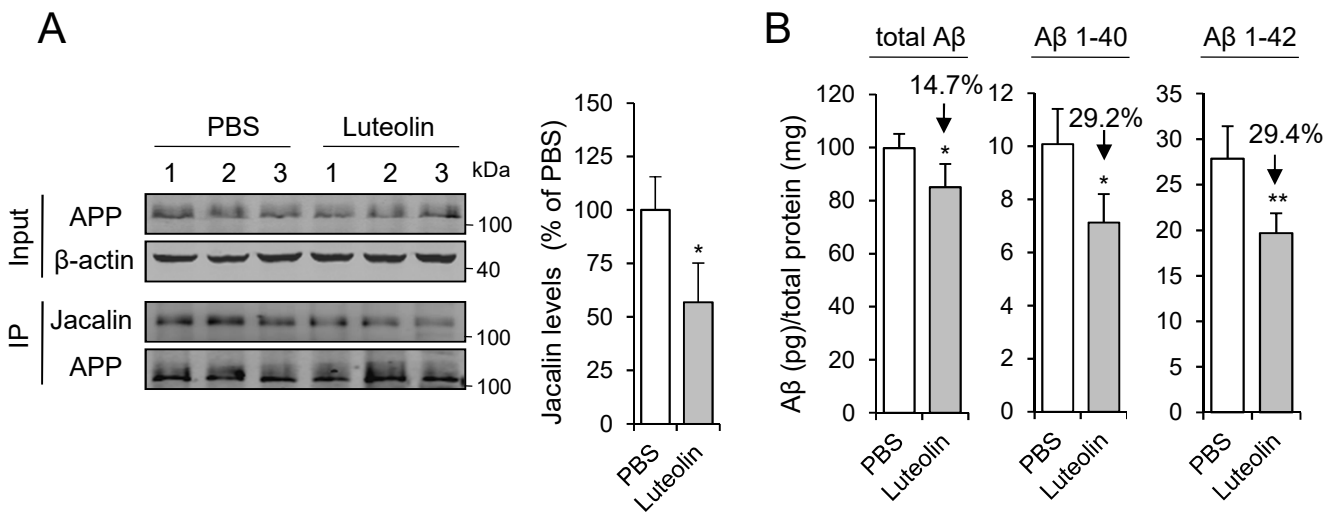


Figure 7



The small molecule luteolin inhibits *N*-acetyl-α-galactosaminyltransferases and reduces mucin-type *O*-glycosylation of amyloid precursor protein

Feng Liu[1], Kai Xu[5], Zhijue Xu[1], Matilde de las Rivas[2], Xing Li[1], Jishun Lu[1], Ignacio Delso[3], Pedro Merino[2], Ramon Hurtado-Guerrero[2, 4, *] and Yan Zhang[1, *]

[1]Key Laboratory of Systems Biomedicine (Ministry of Education) and Collaborative Innovation Center of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China.
[2]Instituto de Biocomputación y Fisica de Sistemas Complejos (BIFI), BIFI-IQFR (CSIC) Joint Unit, Universidad de Zaragoza, 50009, Zaragoza, Spain.
[3]Instituto de Síntesis Química y Catálisis Homogénea (ISQCH). Universidad de Zaragoza. CSIC. E-50009 Zaragoza. Aragón, Spain.
[4]Fundación ARAID, 50018 Zaragoza, Spain.
[5]Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Ministry of Education, College of Fisheries and Life Science, Shanghai Ocean University, Shanghai 201306, China.

Running title: Luteolin inhibits ppGalNAc-Ts

[*]**To whom correspondence should be addressed:**
Yan Zhang, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dongchuan Rd., 200240, Shanghai, China; Tel./Fax: 86-21-34206778; E-mail: yanzhang2006@sjtu.edu.cn.
Ramon Hurtado-Guerrero, Instituto de Biocomputación y Fisica de Sistemas Complejos (BIFI), BIFI-IQFR (CSIC) Joint Unit, Universidad de Zaragoza, 50009, Zaragoza, Spain; Tel.: +34 976 762997, fax: + 34 976 762990; E-mail: rhurtado@bifi.es.

Key words: *O*-glycosylation, ppGalNAc-T, glycosylation inhibitor, glycoprotein, glycosyltransferase, luteolin, amyloid precursor protein (APP), crystal structure

TABLE OF CONTENTS

Table S1. Data collection and refinement statistics. Values in parentheses refer to the highest resolution shell. Ramachandran plot statistics were determined with PROCHECK.

	ppGalNAc-T2 in complex with the luteolin and UDP/Mn^{+}
Space group	$P2_12_12_1$
Wavelength (Å)	0.97
Resolution (Å)	20-2.30 (2.42-2.30)
Cell dimensions a, b, c (Å) α, β, γ (°)	a = 116.98 b = 123.11 c = 248.15 90, 90, 90
Mn(I) half-set correlation CC (1/2)	0.997 (0.590)
Unique reflections	151903
Completeness	95.5 (99.5)
R_{pim}	0.050 (0.535)
I/σ(I)	10.3 (1.6)
Redundancy	6.7 (6.9)
R_{work} / R_{free}	0.213/0.251
RMSD from ideal geometry, bonds (Å)	0.013
RMSD from ideal geometry, angles (º)	1.710
<B> GalNAc-T2 ($Å^2$)	54.66
<B> UDP ($Å^2$)	49.23
<B> luteolin ($Å^2$)	103.02
<B> solvent ($Å^2$)	44.38
<B> ethylenglycol($Å^2$)	64.17
Ramachandran plot: Most favoured (%) Additionally allowed (%) Disallowed (%)	 96.50 3.03 0.48
PDB ID	5NDF

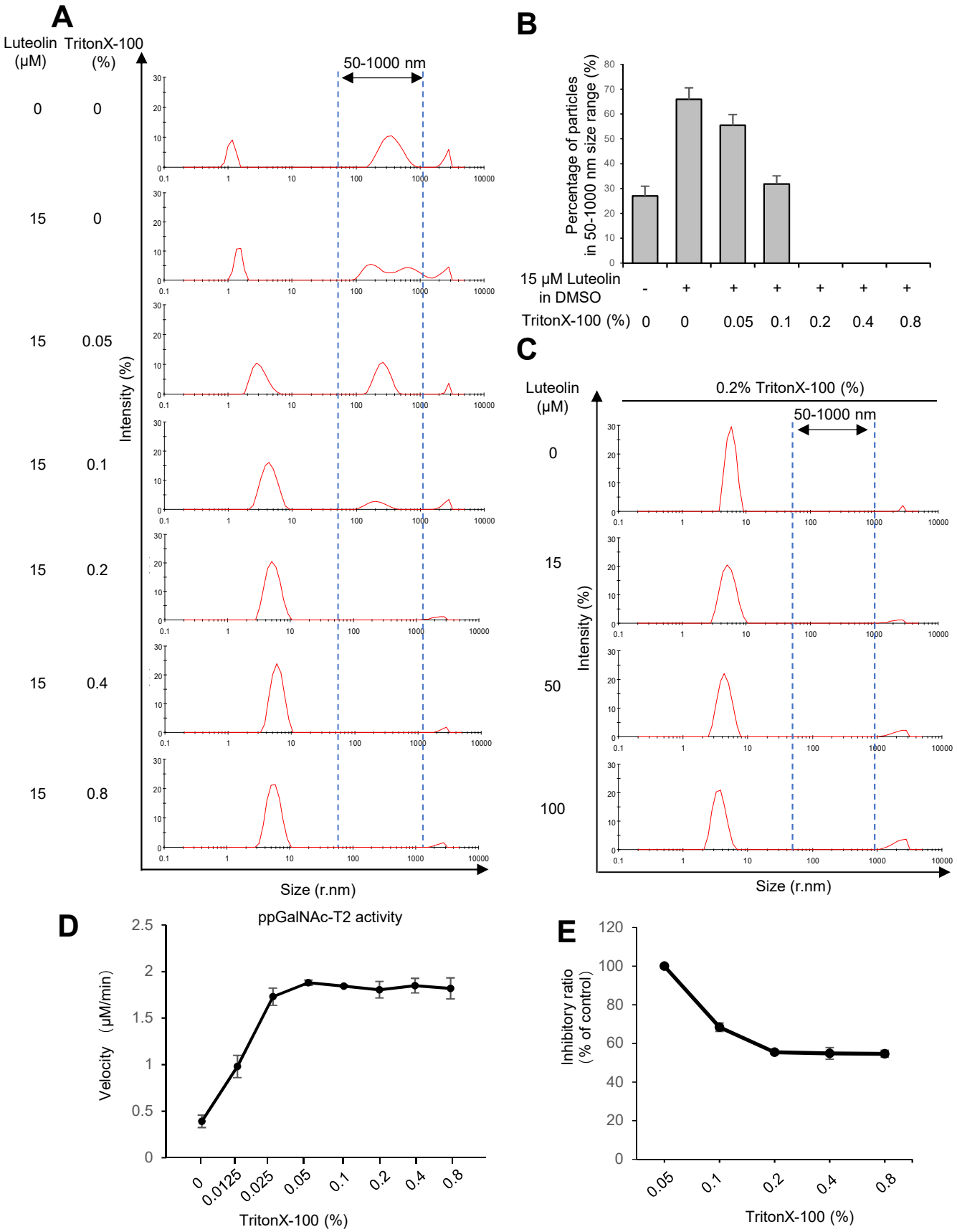


Fig. S1 Aggregation-based activity of luteolin with ppGalNAc-T2. (**A-C**) The particle spectrum change of luteolin in different concentration of TritonX-100 (v/v). The particles spectrum was measured by dynamic light scattering (DLS, Particle Size Analyzer, Z90s, Malvern Instruments Ltd., UK). (**D**) The change of ppGalNAc-T2 catalytic velocity with different concentrations of TritonX-100 (v/v). (**E**) The inhibitory potency of luteolin (15 mM) under the different concentrations of TritonX-100 (v/v). 10 mM EDTA was used as 100% inhibitor control. Data are shown as the mean ± SD (n=3). Two-tailed Student's t-test, **, P<0.01; ***, P<0.001; n.s., no significant difference.

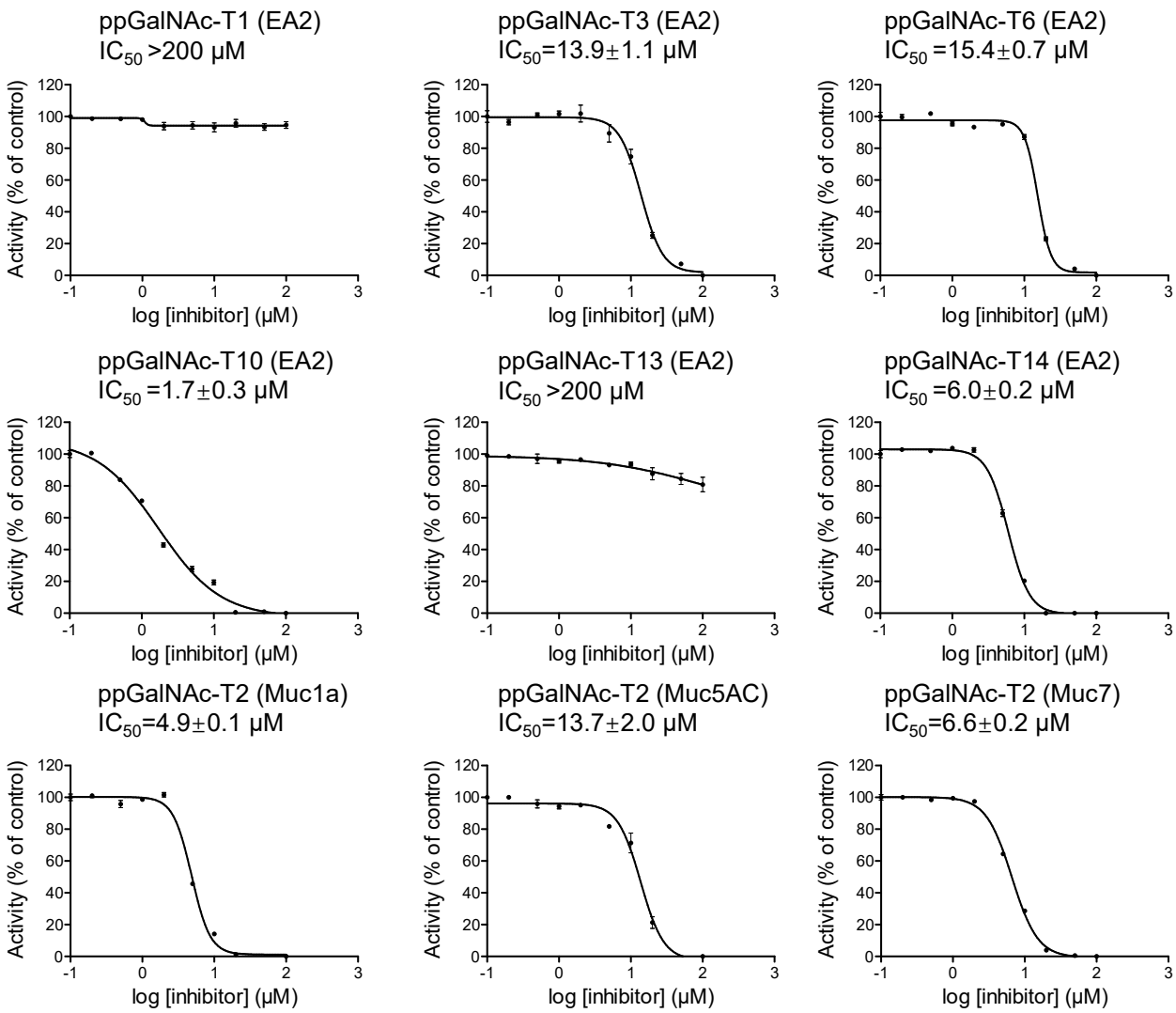


Fig. S2 Dose-dependent inhibition of ppGalNAc-Ts by luteolin in the presence of EA2 substrate and inhibition of ppGalNAc-T2 by luteolin in the presence of various substrates. The IC_{50} values of luteolin for ppGalNAc-Ts were determined under the standard conditions of the HPLC-based assay with EA2, Muc1a, Muc5AC or Muc7 as the substrate (Experimental Procedures). The data are expressed as percentages of the control (DMSO, 100%) and shown as the mean ± SD (n=3).

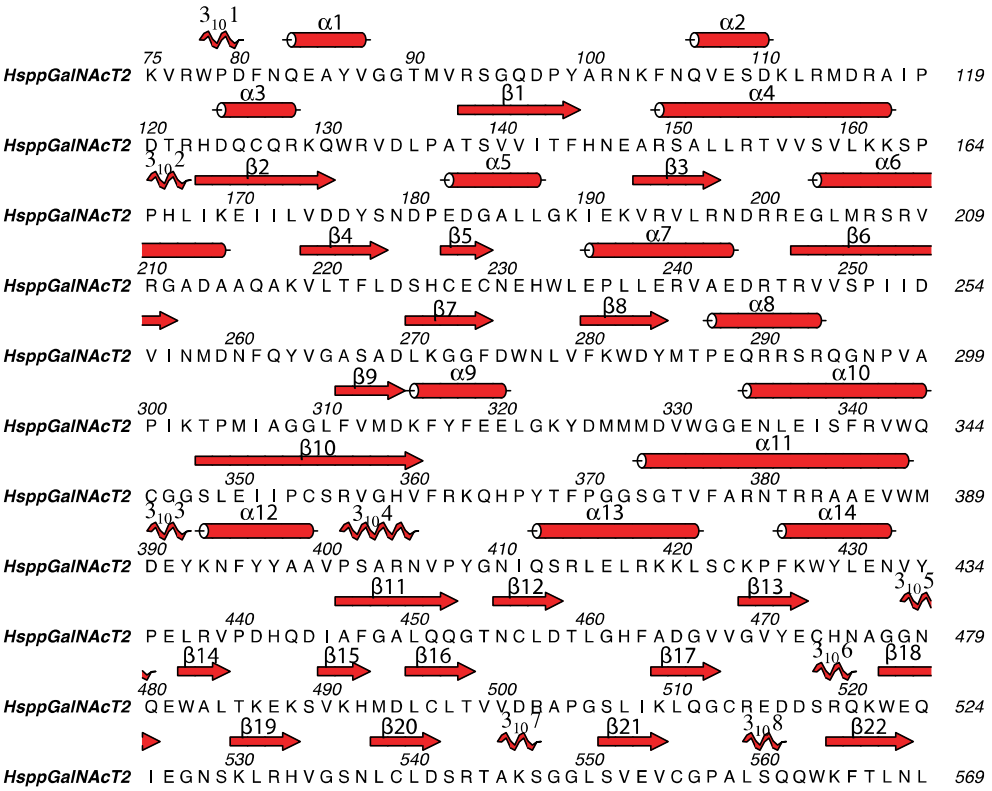


Fig. S3 Sequence of ppGalNAc-T2. Secondary structure elements from the ppGalNAc-T2 structure are shown, with a-helices, 3_{10}-helices and b-strands in red.

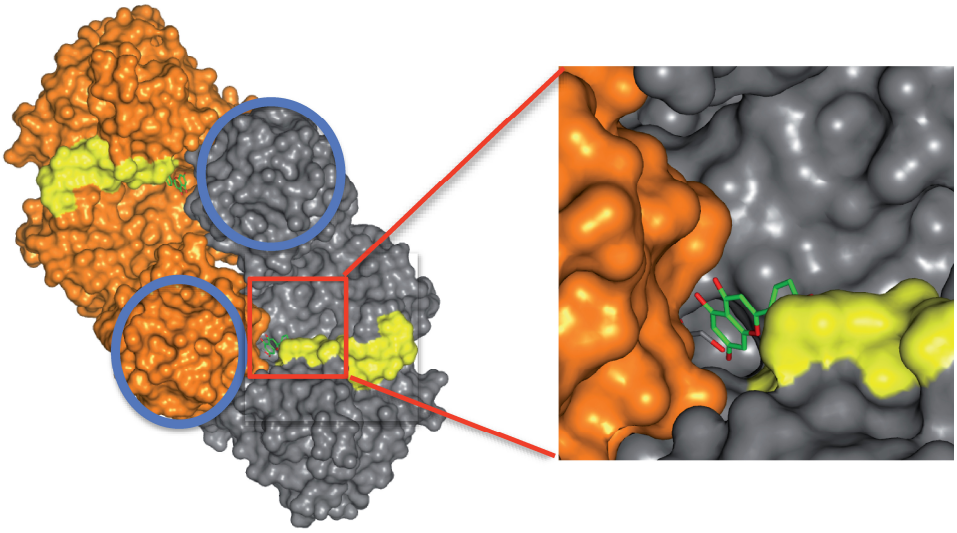


Fig.S4 Surface representation of one of the dimers present in the asymmetric unit. The surface of one of the monomers is depicted in grey whereas the other one is in orange. The flexible loop is shown in yellow in both ppGalNAc-T2 molecules. The two circles in blue and the red square point out the lectin domain position of both molecules, and the position of luteolin in one of the monomers, respectively.

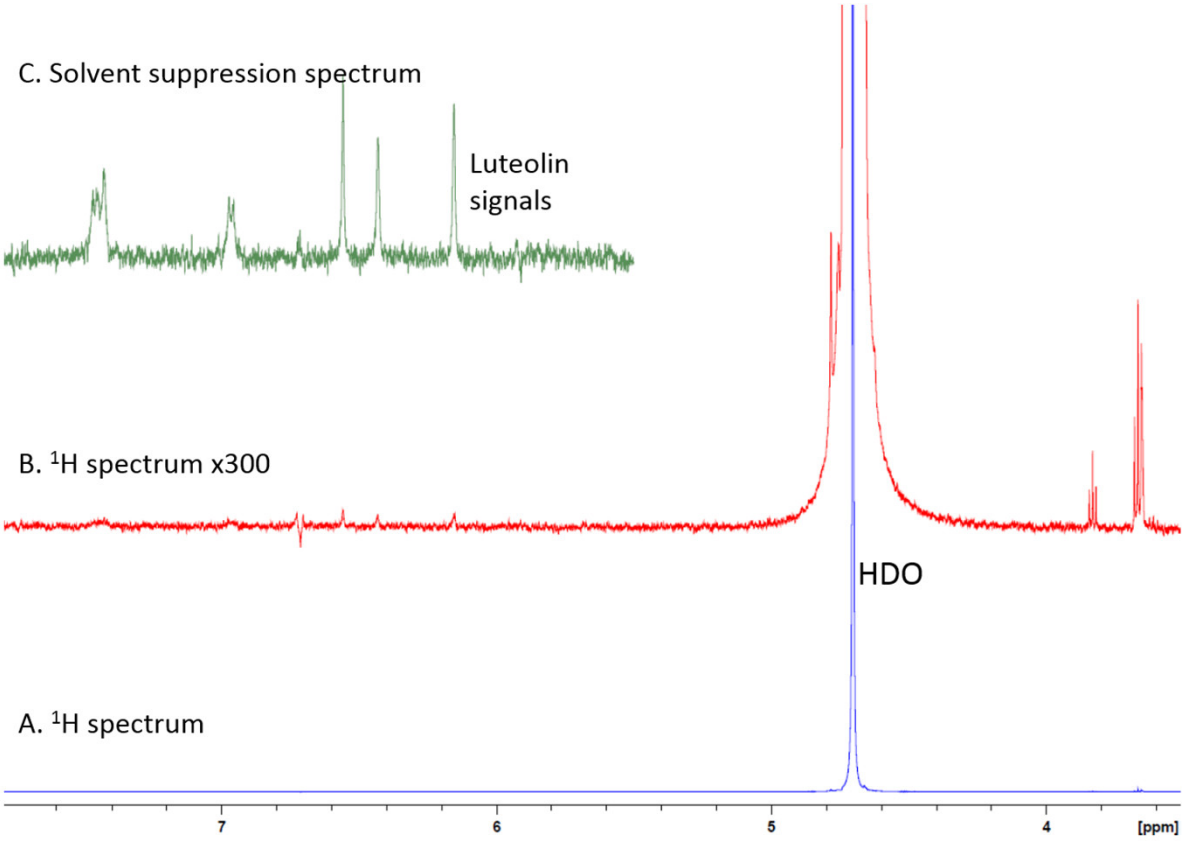


Fig.S5 Stacked ^{1}H NMR spectra of luteolin in deuterated buffer: A. Normal-scale spectrum. HDO is used as standard for chemical shift (4.70 ppm), B. Spectrum with increased intensity (300-fold), where luteolin signals arise, C. Aromatic region of the solvent suppression spectrum.

ARTÍCULO V

Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts

El presente artículo fue publicado en la revista *Chemistry-A European Journal* el año 2018, cuarto y último año del periodo de tesis de la doctoranda. Fue realizado en coautoría con: (a) el Departamento de Química de la Universidad de Nova de Lisboa, Portugal; (b) el Centro de Investigaciones Científicas bioGUNE, en Bilbao; (c) el grupo del Doctor Francisco Corzana, del Departamento de Química Orgánica de la Universidad de La Rioja y (d) el grupo de investigación del Profesor Henrik Clausen, en la Universidad de Copenhage.

A fecha de defensa de la presente tesis doctoral no se encuentran en disposición del título de Doctor los siguientes coautores: Helena Coelho y Ana Díniz, del Departamento de Química de la Universidad de Nova de Lisboa y el Centro de Investigaciones Científicas bioGUNE, en Bilbao, e Ismael Compañón, del Departamento de Química Orgánica de la Universidad de la Rioja. Se adjuntan sus correspondientes renuncias a presentar este artículo en la modalidad de compendio de pubilcaciones en el Apéndice 10.6 de la presente tesis.

La contribución del doctorando en esta publicación ha consistido en lo siguiente:

- Participación en el diseño experimental y en la escritura final del artículo.
- Clonaje, expresión y purificación de las diversas proteínas y mutantes.
- Cristalización del mutante patogénico y resolución de su estructura.
- Determinación y cálculo de las diversas constantes de disociación entre las proteínas y sus sustratos.
- Determinación de las temperaturas de fusión de las proteínas.
- Marcaje de los correspondientes mutantes con el fluoróforo 3-bromo-1,1,1,-trifluoroacetona.

Se adjunta a continuación la referencia completa del artículo:

de las Rivas, M., Coelho, H., Diniz, A., Lira-Navarrete, E., Compañón, I., Jiménez-Barberó, J., Schjoldager, K., T., Bennett, E. P., Vakhrushev, S. Y., Clausen, H., Corzana, F., Marcelo, F. & Hurtado-Guerrero, R. (**2018**). Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts. *Chemistry - A European Journal.* DOI: 10.1002/chem.201800701.

Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts

Matilde de las Rivas,[a]^ Helena Coelho, [b],[c]^ Ana Diniz,[b] Erandi Lira-Navarrete,[d] Ismael Compañón,[e] Jesús Jiménez-Barbero,[c],[f] Katrine. T. Schjoldager,[d] Eric. P. Bennett,[d] Sergey Y. Vakhrushev,[d] Henrik Clausen,[d] Francisco Corzana,[e] Filipa Marcelo,*[b] and Ramon Hurtado-Guerrero*[a],[g]

Abstract: The family of polypeptide GalNAc-transferases (GalNAc-Ts) orchestrates the initiating step of mucin-type protein *O*-glycosylation by transfer of GalNAc moieties to serine and threonine residues in proteins. Deficiencies and dysregulation of GalNAc-T isoenzymes have been found to be related to different diseases. Recently, we have demonstrated that an inactive GalNAc-T2 mutant (F104S), which is not located at the active site, induces low levels of high-density lipoprotein cholesterol (HDL-C) in humans. Here, we have deciphered the molecular basis for F104S mutant inactivation. Saturation transfer difference NMR experiments demonstrate that the mutation induces loss of binding to peptide substrates. The analysis of the crystal structure of the F104S mutant bound to UDP-GalNAc, combined with molecular dynamics (MD) simulations, has revealed that the flexible loop is disordered and displays larger conformational changes in the mutant enzyme than in the wild-type (WT) enzyme. ^{19}F-NMR experiments reveal that the WT enzyme reaches the active state only in the presence of UDP-GalNAc, providing compelling evidences that GalNAc-T2 adopts an UDP-GalNAc-dependent induced-fit mechanism. The F104S mutation precludes the enzyme to achieve the active conformation and concomitantly to bind peptide substrates. In summary, the present study provides new insights into the catalytic mechanism of the large family of GalNAc-Ts and how these enzymes orchestrate protein *O*-glycosylation.

[a] Mrs. M de las Rivas, Dr. R. Hurtado-Guerrero
Instituto de Biocomputación y Fisica de Sistemas Complejos (BIFI), BIFI-IQFR (CSIC) Joint Unit, Universidad de Zaragoza, 50018 Zaragoza, Spain. E-mail: rhurtado@bifi.es
[b] Mrs. H. Coelho, Mrs. A. Diniz, Dr. F. Marcelo
UCIBIO, REQUIMTE, departamento de Química, Faculdade de Ciências e Tecnologia, Universidade de Nova de Lisboa, 2829-516 Caparica, Portugal. E-mail: filipa.marcelo@fct.unl.pt
[c] Mrs. H. Coelho, Prof. J. Jiménez-Barbero
CIC bioGUNE, Bizkaia Technology Park, Building 801A, 48170 Derio, Spain; Departament of Organic Chemistry II, Faculty of Science & Technology, University of the Basque Country, 48940 Leioa, Bizkaia, Spain
[d] Mrs. E. Lira-Navarrete, Dr. K. T. Schjoldager, Dr. E. P. Bennett, Dr. S. Y. Vakhrushev, Prof. H. Clausen
Copenhagen Center for Glycomics, Department of Cellular and Molecular Medicine, School of Dentistry, University of Copenhagen, Copenhagen, Denmark.
[e] Mr. I. Compañón, Dr. F. Corzana
Departamento de Química, Universidad de La Rioja, Centro de Investigación en Síntesis Química, E-26006 Logroño, Spain.
[f] Prof. J. Jiménez-Barbero
Ikerbasque, Basque Foundation for Science, Maria Diaz de Haro 13, 48009 Bilbao, Spain.
[g] Dr. R. Hurtado-Guerrero
Fundación ARAID, 50018, Zaragoza, Spain.
[^] Mrs. M de las Rivas and H. Coelho contributed equally to this work
Supporting information for this article is given via a link at the end of the document.

Introduction

Mucin-type (GalNAc-type) *O*-glycosylation is the most differentially and complex regulated type of protein glycosylation, and likely the most abundant, with over 80% of all proteins predicted to be *O*-glycosylated.[1] This post-translational modification is initiated by a large family (20 in humans) of enzymes named polypeptide *N*-acetylgalactosaminyltransferases (GalNAc-Ts), which catalyze the attachment of a a-GalNAc residue of the substrate UDP-GalNAc to Ser/Thr residues of the peptide backbone.

Although these enzymes have a partial overlap in recognition and glycosylation of substrates, a subset of glycosylation sites is subjected to site-specific *O*-glycosylation by individual GalNAc-Ts. For instance, GalNAc-T3 is responsible for the site-specific glycosylation of Thr178 of FGF23, a major regulator of phosphate homeostasis[2]. This site-specific glycosylation of Thr178 prevents the inactivation of FGF23 by proprotein convertase proteases[3], thus mutations in GalNAc-T3 leading to a decrease or a complete abolishment of its activity are known to cause familial tumoral calcinosis.[3] By a similar mechanism, we have recently demonstrated that the loss of GalNAc-T2 function lowers high-density lipoprotein cholesterol (HDL-C) levels in humans.[4] We also identified two independent deficiencies in the *GALNT2* gene, and one of these involved the homozygous missense mutation F104S[4], which was not predicted to be deleterious based on conservation and prior structural models of the enzyme.[5] Those examples illustrate how GalNAc-T isoforms, with seemingly many roles in protein glycosylation, have in fact limited non-redundant biological functions.[3-4]

From a structural perspective, GalNAc-Ts are unique among glycosyltransferases because in addition to their N-terminal catalytic domain, they also possess a unique C-terminal lectin domain with both domains being connected by a short flexible linker (Figure 1a and Figure S1). There is also an interesting structural feature in the catalytic domains of these enzymes, a so-called flexible loop, which comprises residues Val360 to Gly372 in GalNAc-T2. This loop may adopt different conformations during the catalytic cycle, thus rendering the enzyme catalytically inactive or active.[5-6].

Crystal structures of GalNAc-Ts can be used as a platform to infer how mutations on these enzymes might lead to loss-of-function.[5-7] However, this is not always straightforward and further experimental work in solution to infer the dynamic

behavior of the global system is required. The F104S 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 and the critical Mn^{+2} binding site, which is formed by the $D_{224}xH_{226}$ residues and His359 (Figure 1b).[5] Thus, it is not obvious how mutation of Phe to Ser impairs GalNAc-T2 catalytic activity.

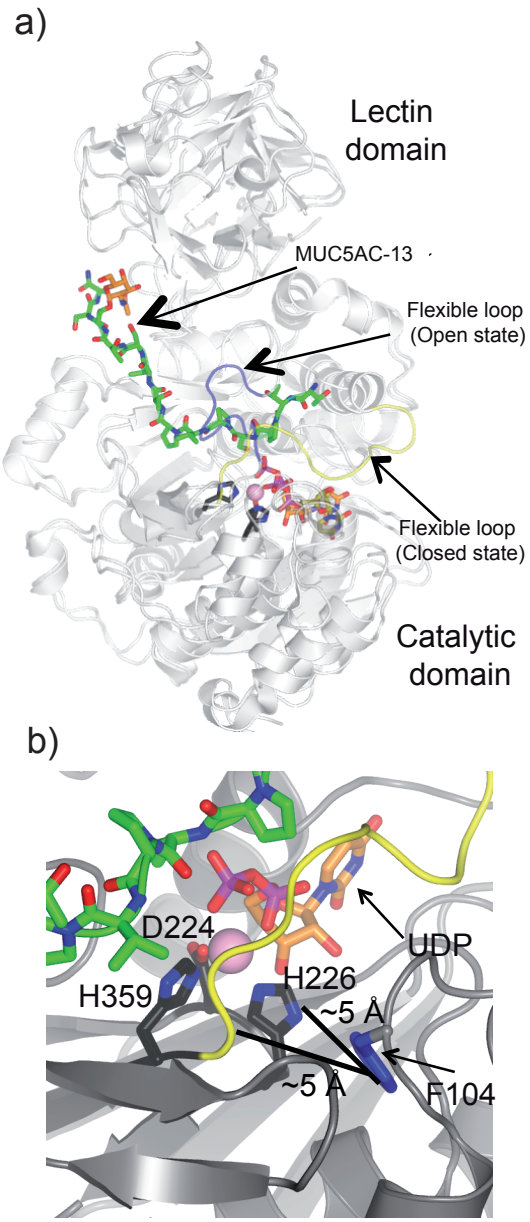


Figure 1. States of the flexible loop and location of Phe104: a) Superposition of the overall crystal structure of GalNAc-T2 in complex with UDP and the glycopeptide MUC5AC-13[5] (closed conformation shown in white color; see PDB entry 5AJP), with the crystal structure of GalNAc-T2 in complex with UDP[7b] (open conformation also shown in white color; see PDB entry 2FFV). The GalNAc moiety of MUC5AC-13 is colored in orange. UDP and MUC5AC-13 are depicted as orange and green carbon atoms, respectively. The Mn^{+2} binding $D_{224}xH_{226}$ motif and His359 are 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^{+2} is shown as a pink sphere. b) Close-up view of the active site (closed conformation; PDB entry 5AJP) showing the distance of Phe104 (in blue) to the flexible loop (in yellow) and the His226 (in black). The backbone structure is shown in grey while the other structural features and amino acids are shown in the same colors as in panel a.

Herein, we show that the F104S mutation in GalNAc-T2 constrains the flexible loop and renders the enzyme locked in an inactive state unable to bind the acceptor peptide substrate. We have combined saturation transfer difference NMR (STD-NMR), ^{19}F-NMR spectroscopy, X-ray crystallography and molecular dynamics simulations to demonstrate that Phe104 fine-tunes the dynamic behavior of the flexible loop by establishing a CH-π interaction with the side-chain of Arg362, which is located at the flexible loop. This interaction appears to be the key to maintain the flexible loop in the required closed conformation that leads to the active state of the enzyme. Therefore, the inability of the mutant to adopt the active state disrupts the peptide substrate recognition and prevents *O*-glycosylation. Finally, we have experimentally demonstrated that GalNAc-T2 adopts an UDP-GalNAc-dependent induced-fit mechanism for catalysis and suggest that the behavior of the flexible loop can be extrapolated to the rest of the members of the GalNAc-T family.

Results and Discussion

The GalNAc-T2 F104S mutant does not bind to peptides. The available data based on kinetic assays of the F104S mutant cannot explain the molecular basis of the inactivation of the GalNAc-T2 mutant F104S against protein substrates.[4] Since Phe104 is not located at the active site (Figure 1b), its mutation to Ser must induce conformational changes of the residues located in the active site, affecting either the binding of the mutant to the substrates or the catalytic turnover. To rule out one of these two hypotheses, we have determined the binding of the GalNAc-T2 WT enzyme and the F104S mutant to UDP-GalNAc and a peptide substrate applying tryptophan fluorescence spectroscopy and STD-NMR experiments, respectively. In addition, we have also conducted a thermal shift-assay experiment on both enzymes to evaluate the differences in the protein stability.[8] The data in Figure 2 shows that the WT enzyme is more stable than the F104S mutant (denaturation temperatures of T_0 of 47 ± 0.09 °C and 42 ± 0.05 °C, respectively), implying that the mutation leads to a significant decrease on the stability of the mutant (Figure 2a). The tryptophan fluorescence spectroscopy analysis of UDP-GalNAc binding revealed that both the WT enzyme and the F104S mutant bind to the donor sugar nucleotide, with the K_d of UDP-GalNAc to the mutant being 1.7-fold higher than the K_d determined for the WT enzyme (Figure 2b and Figure S2). In contrast, STD-NMR experiments recorded for the peptide substrate MUC5AC only showed WT GalNAc-T2 binding to the MUC5AC peptide (Figure 2c, Figure S3). No STD signal response was observed for MUC5AC in presence of the mutant F104S (Figure 2c, Figure S3). This result points out that the

mutant is not able to properly recognize and bind the peptide substrate, thus explaining the previously observed inactivity of the mutant towards protein substrates such as phospholipid transfer protein (PLTP) and ApoC-III[4]. In addition, the WT enzyme showed stronger STD-NMR signals for the sequence "PSPVPT" (where the underlined Thr is the main acceptor site on this peptide for GalNAc-T2) than for the rest of peptide sequence. The identified STD-NMR epitope is consistent with the sequence preferences of GalNAc-T2, particularly having a Pro residue preceding the site of glycosylation[9]. For the WT GalNAc-T2, the underlined Thr9 of MUC5AC is the best-recognized acceptor site of all possible potential sites of MUC5AC, which also fits with the observations that this residue is the most glycosylated in this peptide.[10] Hence, while there are no major differences in UDP-GalNAc recognition between the WT enzyme and the F104S mutant, the dramatic differences in activity are accounted for either the inability or poor affinity of the F104S mutant to bind the peptide substrates.

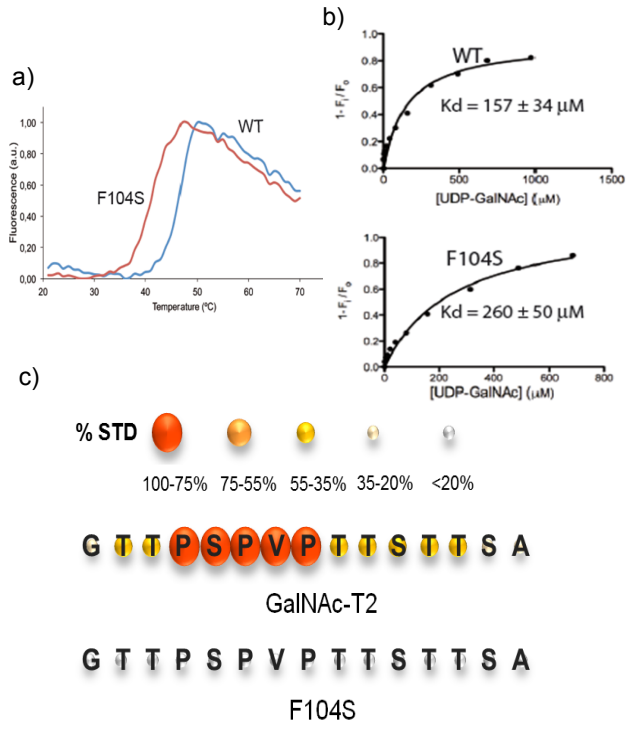


Figure 2. Stability and binding properties of the WT enzyme and the F104S mutant: a) Thermal denaturation curves of the WT enzyme and the F104S mutant as monitored by 1-anilino-8-naphthalene sulfonate (ANS) fluorescence. b) K_d values for sugar nucleotide binding determined from the tryptophan fluorescence intensity as a function of UDP-GalNAc concentrations. c) STD-NMR-derived epitope mapping of MUC5AC (see Table S1 for the NMR assignment of the peptide). The different colored spheres indicate the normalized STD signal (in percent) observed for each proton. See Supporting Information (SI) for the NMR experiments and Figure S3 for further details.

Phe104 controls the inactive-to-active state transition. To elucidate the molecular basis for the loss of acceptor peptide substrate binding induced by the F104S mutation, we obtained tetragonal crystals of this mutant in complex with UDP that were subsequently soaked with UDP-GalNAc. The resulting crystals allowed us to solve the structure at 2.70 Å resolution (Table S2). Within the asymmetric unit, two molecules of the F104S mutant were present. The crystal structure shows a compact structure with the typical N-terminal GT-A fold and a C-terminal lectin domain (Figure 3).

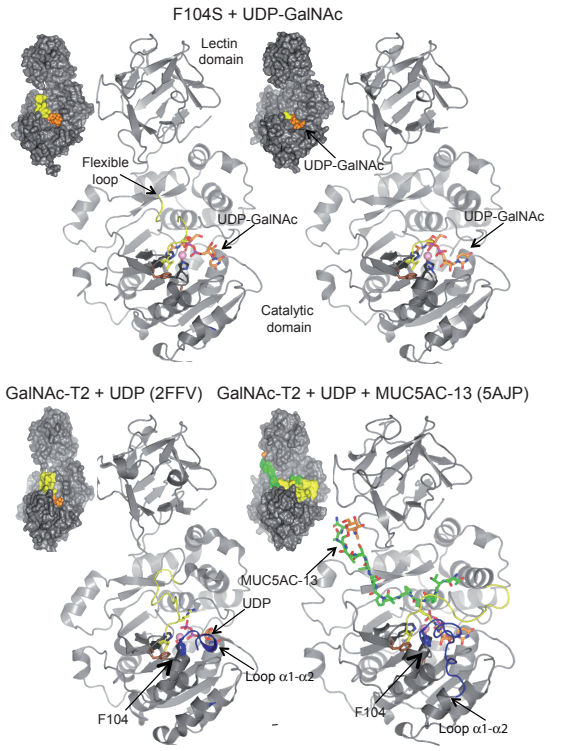


Figure 3. Overall crystal structure of the F104S mutant bound to UDP-GalNAc and molecular dynamics simulations of the flexible loop: Cartoon representation of the overall F104S mutant in complex with UDP-GalNAc and the WT enzyme either in complex with UDP[7b] (PDB entry 2FFV) or UDP/MUC5AC-13[5] (PDB entry 5AJP). The proteins and the GalNAc moiety of MUC5AC-13 are colored in grey and orange, respectively. UDP/UDP-GalNAc, MUC5AC-13, the flexible loop, and the α1-α2 loop are depicted as orange, green, yellow and blue carbon atoms, respectively. The Phe104, Val360/Arg362 and the $D_{224}xH_{226}$ motif and His359 are shown as sticks in blue, yellow and black carbon atoms, respectively. Other residues interacting with Phe104 are shown as brown carbon atoms. Mn^{+2} is depicted as a pink sphere. (insert) Surface representation of the above complexes showing inactive and active states. Colours are the same as above.

The flexible loop in both F104S molecules of the asymmetric unit is highly disordered (Figure 3, 4, 5 and Figure S4), exposing the substrate UDP-GalNAc to the solvent. In these structures the flexible loop displays the open conformation that restraints the enzyme in an inactive state, resembling the structure of the WT enzyme in its inactive state (see PDB entry 2FFV; Figure 3, 4 and 5). This conformation is opposed to that presenting the active state. In the active state the flexible loop displays a closed conformation that functions to cover UDP-GalNAc or UDP from

the solvent (e.g. see PDB entries 5AJP, 4D0T and 4D0Z; Figure 3). This geometrical arrangement is required for the binding of the peptide/protein substrates[5, 6c]. This closed conformation of the flexible loop is further stabilized by interactions of His365 and Phe369 with Trp331 (in-conformation) and by additional interactions with the donor substrate UDP-GalNAc (Figure 5)[6c]. In fact, the fully active state can only be reached when both the flexible loop and Trp331 adopt the closed geometry and the "in-conformation", respectively.

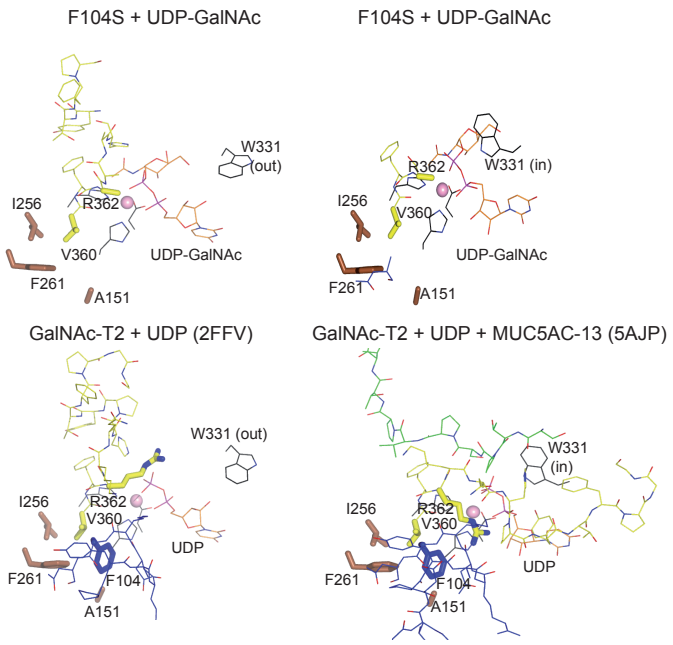


Figure 4. Close-up view of the complete sugar nucleotide and partial peptide-binding site of the complexes in Figure 3. Amino acids are shown as sticks with the same colors indicated above. Both Phe104 and residues interacting with it are magnified. Trp331 is shown in an "out and in conformations". Colours are the same as shown in Figure 3.

UDP-GalNAc in the inactive F104S mutant is shown adopting two potential conformations as previously described for UDP (Figure 3, 4 and 5)[5, 7b]. The first conformation is where the uridine moiety is bound in an "inverted" conformation, compatible with UDP-GalNAc entrance in the active site. The second conformation is for the enzyme in an active state, having a conformation suited for catalysis (see also PDB entry 5AJP; Figure 4). It should also noted that, the crystal structure of GalNAc-T2 in complex with UDP shown in the PDB entry 2FFV, shows the uridine moiety in the "inverted" conformation, although this geometry is likely related to product release (Figure 4).

A detailed analysis of the F104S mutant structure reveals a lack of density for Ser104 and for most of the residues of the loop between helical α1 and α2 (named hereafter α1-α2 loop). In the WT enzyme, Phe104 establishes hydrophobic and stacking interactions with Ala151/Ile256/Val360 and Phe261, likely stabilizing the location of both the α1-α2 and the flexible loops, in the inactive state (Figure 4). The disordered states of both loops in the inactive mutant might also explain why this structure is less stable than the WT enzyme (see above). In the active state of the enzyme, Phe104 also establishes a CH-π interaction with the side-chain of Arg362 located in the flexible loop (Figure 4). This stabilizing interaction appears to be a key feature rendering the closed conformation of the flexible loop. In contrast, the polar Ser104 residue likely disrupts the interactions with the hydrophobic/aromatic residues, leading to the destabilization of the α1-α2 loop and the inability for the flexible loop to reach the closed conformation. Therefore, the obtained data suggest that when Phe104 is replaced by a serine residue

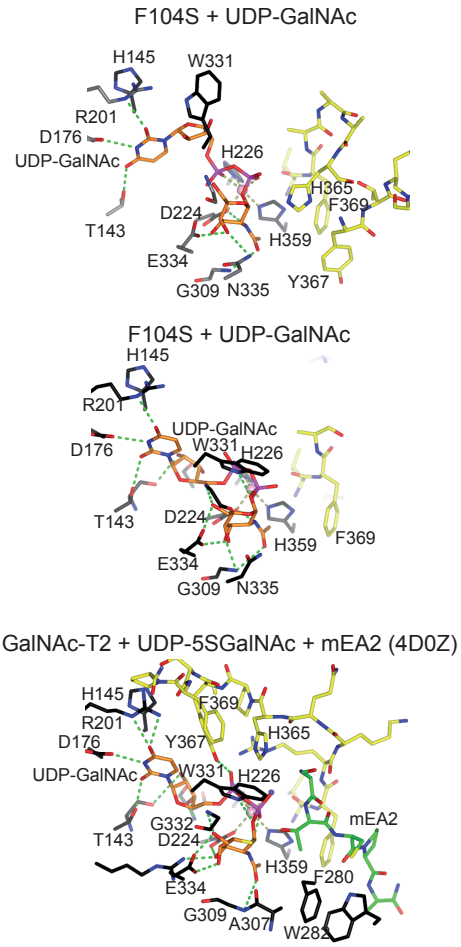


Figure 5. Close-up view of the sugar nucleotide and peptide binding site of the F104S-UDP-GalNAc and WT-UDP-5S-GalNAc-mEA2 complexes. Residues and the sugar nucleotides are depicted with the same colors as shown in Figure 3 and 4. Hydrogen bond interactions are shown as dotted green lines.

(Ser104), the enzyme is not able to reach the active state, even in the presence of UDP-GalNAc due to the lack of the essential Phe104/Arg362 CH-π interaction, which provides the additional stabilization required to induce the closed conformation of the flexible loop. Interestingly, most of the residues interacting with

Phe104, Ile256/Val360/Phe261/Arg362, are highly conserved among different GalNAc-Ts (Figure S5). This fact suggests that the flexible loop behavior and the corresponding activation mechanism via the Phe104/Arg362 CH-π interaction is conserved across most GalNAc-T isoenzymes.

Regarding the UDP-GalNAc recognition, the binding sites of both the F104S mutant and the WT enzyme in the active form show almost equivalent interactions, explaining the similar K_d values obtained for this sugar nucleotide (Figure 2b). The interactions with UDP-GalNAc are mostly based on hydrogen bonds with the surrounding protein residues, whereas the pyrophosphate is involved in the coordination of the Mn^{+2} ion, together with the residues of the $D_{224}xH_{226}$ motif and His359 (Figure 5). Minor additional interactions are found for the WT enzyme in the active form (e.g. hydrogen bond interactions with Gly332 and Tyr367; Figure 5), which is in agreement with the slightly better affinity for UDP-GalNAc on the WT enzyme versus the mutant (1.7-fold).

The flexible loop in the F104S mutant displays large conformational changes. To understand how the mutation F104S disturbs the dynamics of the flexible loop, we ran 200 ns molecular dynamics (MD) simulations in explicit water on both the WT enzyme and the F104S mutant in the presence of UDP-GalNAc. The data revealed that the flexible loop of the mutant displayed 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 enzyme (Figure 6a). This trend was also shown in the variation of the Cα distances between the Phe104/Ser104 residues and Val360. While a shorter distance of 5.7 ± 1.4 Å was found for Phe104-Val360, highly variable and larger distances (ranging from 4.5 to 16 Å) were found for Ser104-Val360 (Figure 6b and 6c**)**. We also determined the distance distributions for Arg362/Phe104, and Arg362/Ser104-Cα pairs (Figure 6b). The average distance of 4.7 ± 1.4 Å found for the WT enzyme suggests the presence of stabilizing hydrophobic interactions between the aromatic ring of Phe104 and the side-chain of Val360. In contrast, two populations were found for the mutant: a minor population, with distances ranging between 3.5-7 Å, and a major population, with highly oscillating distances, between 7.5-17 Å (Figure 6b). These results also point out that the flexible loop and the α1-α2 loop are much more structured in the WT enzyme than in the mutant and that these structural entities remain at a closer distance in the WT enzyme than in the mutant. Therefore, according to the MD analysis, the closed conformation of the flexible loop in the inactive mutant would rarely form.

^{19}F labeling of the T375C and F104S-T375C variants for NMR experiments. To experimentally monitor the dynamics of flexible loop, we also performed ^{19}F-NMR spectroscopy experiments on the WT enzyme and on the F104S mutant. ^{19}F chemical shifts are highly sensitive to changes in the local conformational environment and provide a simple and well-established approach for studying biomolecular structure, conformation, and dynamics.[11] To introduce the required ^{19}F label, Thr375 was mutated to the more nucleophilic Cys residue, in both the WT enzyme and the F104S mutant. This residue was selected for several reasons: i) it is exposed and accessible to be modified

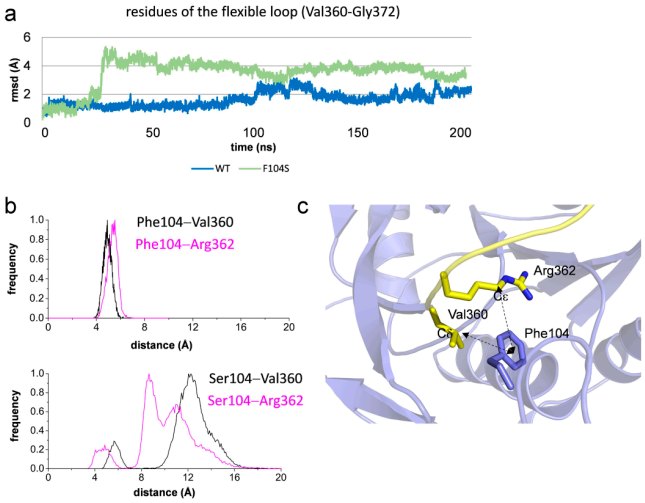


Figure 6. Molecular dynamics simulations on the WT enzyme and the F104S mutant: a) The RMSD (root-mean-square deviation) values calculated from 200 ns molecular dynamics (MD) simulations for the flexible loop indicates that this structural motif is more flexible in the mutant (green plot) than in the WT enzyme (blue plot). Note that the RMSD was obtained using the fixed coordinates of the flexible loop from the PDB entry 4D0T. b) This is also evidenced by larger distance distribution between Ser104 (Cα) and Val360 (Cd)/Arg362 (Cε) in the mutant (bottom) compared to the corresponding distances (Phe104-Val360 and Phe104-Arg362) found in the WT enzyme (top). c) The center of mass of the π-electron system of Phe104 was used to calculate the distances.

and WT GalNAc-T2 and the F104S mutant contain only two free buried cysteines at the active site; ii) it is fairly close to the flexible loop but sufficiently far away from the Mn^{2+} ion to minimize strong paramagnetic relaxation enhancements; and iii) the crystallographic analyses indicates that Thr375 would be sensitive to the flexible loop dynamics and may adopt different conformations depending on the geometry of the flexible loop (open or closed). It should be noted that Thr375 in the PDB entry 2FFV is 10 Å away from the same residue in PDB entry 5AJP (Figure 7a). Thus, both the WT enzyme and the F104S mutant variants containing the T375C mutation were produced (see SI for expression and purification methodology). The ^{19}F label was subsequently introduced into the T375C and the F104S-T375C variants using 3-bromo-1,1,1-trifluoroacetone (BFA) **(**see SI).[11] A single $SCH_2(CO)CF3$ adduct was obtained for each enzymes (GalNAc-T2* and F104S*) as shown by trypsin digestion and MS fragmentation studies (Figure 7b). While the GalNAc-T2* was 100% labeled, the F104S* was only 93% modified (Figure 7b). The activity of GalNAc-T2* was evaluated and compared with that obtained for the WT enzyme (Figure 7c and SI). GalNAc-T2* still remained active giving ~50% activity against MUC1 peptide compared to the WT enzyme.

^{19}F NMR experiments reveal that GalNAc-T2 adopts an UDP-GalNAc dependent induced-fit mechanism. With both WT and mutant proteins labeled with BFA, we first screened enzymes activity against MUC5AC by ^{1}H-NMR experiments confirming that only GalNAc-T2* and not the F104S* mutant was able to glycosylate MUC5AC (Figure S6).

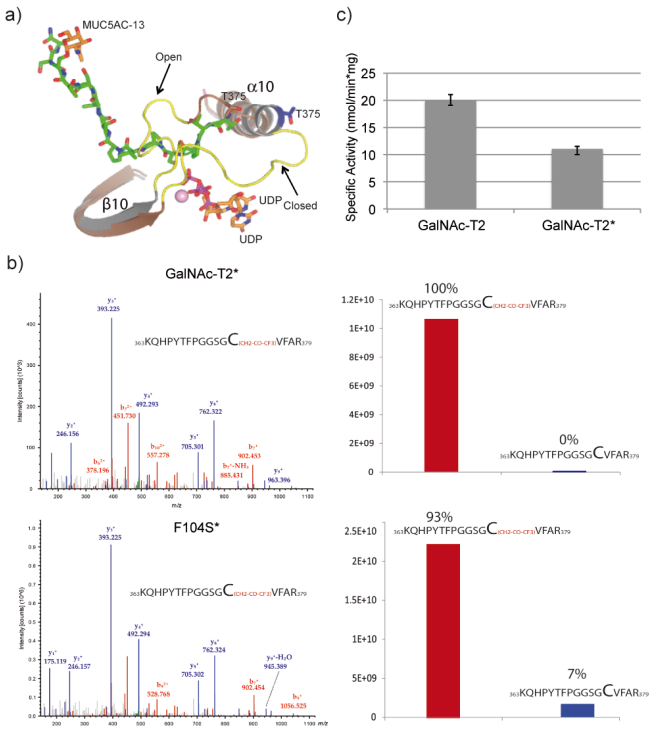


Figure 7. Labeling of the T375C and F104S-T375C variants with BFA: a) Close-up view of the active sites of the PDB entries 2FFV and 5AJP, which are in an inactive and active states, respectively. The flexible loop adopting an open (2FFV) and closed conformation (5AJP) is colored in yellow. The MUC5AC-13 is depicted as sticks in green carbon atoms whereas the GalNAc-moiety is in orange carbon atoms. UDP is colored in orange. The a and b secondary structures are brown and grey for the PDB entries 2FFV and 5AJP, respectively. Thr375 is depicted as sticks as brown (2FFV) and blue (5AJP) carbon atoms. Note that DSSP[12] predicts that Thr375 in the PDB entry 2FFV is part of the flexible loop whereas in the PDB entry 5AJP is located in α10. b) MS/MS spectrum of the precursor ions at *m/z* 654.6346, z=3+ assigned to $_{363}$KQHPYTFPGGSGCVFAR$_{379}$ peptide containing the ^{19}F label at the position C375 (left) and relative quantification based on the extracted ion chromatogram between peptides with and without the ^{19}F label at the position C375 (right). The experiments were performed for the GalNAc-T2* (top) and F104S* (bottom). c) Specific activity of the WT enzyme and GalNAc-T2* using the MUC1 peptide as acceptor substrate. The data represent means ± S.D. for three independent experiments.

To evaluate whether the Cys375 ^{19}F label was sensitive to the dynamics of the flexible loop upon substrate binding, we firstly performed ^{19}F-NMR experiments on GalNAcT2*. A set of experiments using distinct combinations of UDP-GalNAc, $MnCl_2$ and MUC5AC were used (Figure 8 left and middle panel and Figures S7 and S8 in SI). The ^{19}F-NMR spectra of unbound GalNAcT2* (Figure 8, left and middle panels blue tracings) showed two ^{19}F resonances at δ = -84.0 ppm and δ = -84.3 ppm. The upfield ^{19}F signal at δ = -84.3 ppm remained unperturbed for all combinations of added substrates, which we attribute to the oligomeric GalNAc-T2* state. The linewidth of the upfield oligomer ^{19}F peak (δ = -84.3 ppm) is 161 Hz, significantly larger than that observed for the monomer at δ = -84.0 ppm (111 Hz), which is consistent with a reduced relaxation time of the oligomeric enzyme[13]. We have previously demonstrated by small-angle X-ray scattering experiments that GalNAc-T2 forms oligomers that are not fully catalytically active.[5] This may explain the lack of perturbation of this particular NMR signal under different substrate combinations. After addition of UDP-GalNAc and Mn^{+2} to GalNAc-T2*, the ^{19}F-NMR resonance at δ -84.0 ppm almost vanished and a new signal appeared at δ = -83.7 ppm (Figure 8, left panel and middle panel red tracing). It is remarkable that the exchange between these two ^{19}F-NMR resonance (-84.0 ppm and -83.7 ppm) only occurs in the presence of UDP-GalNAc and Mn^{+2}. Indeed, in the absence of UDP-GalNAc, the addition of Mn^{+2} (Figure 8, left panel and middle panel green tracing) or the peptide MUC5AC (Figure 8, left panel and middle panel grey tracing) did not induce any appreciable perturbation in the ^{19}F-NMR spectrum of GalNAc-T2*. This result strongly suggests that the ^{19}F-NMR resonances at δ = -84.0 ppm and δ = -83.7 ppm correspond to the open and closed conformations of the flexible loop, respectively. These two populations are in slow exchange on the NMR chemical shift timescale and thus detected as two resonances. An additional set of experiments was also recorded in the presence of UDP and Mn^{+2} (Figure S9 in SI), giving similar results as obtained for addition of UDP-GalNAc. However, the exchange between the open and closed conformations of the flexible loop was not so pronounced compared to that in the presence of UDP-GalNAc. All together, our data show that UDP-GalNAc is absolutely required for the interconversion between the open and closed conformations of the flexible loop and that UDP-GalNAc stabilizes the closed conformation better than UDP (see Figure S8-S9). In fact, as deduced from the analysis of the crystal structure (see Figure 5 and panel on the right corresponding to the active site of the PDB entry 4D0Z), UDP-GalNAc displays additional interactions with the enzyme through the GalNAc moiety. The fact that UDP does not completely exchange the population of the enzyme conformations towards the fully active state is consistent with its role as a reaction product after catalysis.

The described ^{19}F-NMR experiments provide a compelling evidence that GalNAc-T2 follows an induced-fit mechanism and support our previously proposed ordered bi–bi kinetic mechanism for this enzyme.[6c] The induced-fit mechanism has been also proposed for distant glycosyltransferases such as B4GALT1 and the lactose synthase by applying X-ray crystal structure studies[14]. In this mechanism, GalNAc-T2 is always in the inactive state (open flexible loop) and only exchanges to the active state (closed flexible loop) in the presence of UDP-GalNAc/$MnCl_2$. In contrast to the earlier hypothesis,[6c] our new data does not support the existence of an equilibrium between the inactive and active states in the absence of UDP-GalNAc/$MnCl_2$. The uridine moiety of UDP-GalNAc might also shift its orientation from the “inverted” conformation to the typical geometry required for catalysis during the transition from the inactive to the active state of the enzyme. Only after the enzyme is in the active state, the peptide can bind, and the catalysis can take place.

The inactive-to-active state transition of the flexible loop is hindered in the F104S mutant.

The same set of experiments was performed for the F104S* construct (Figure 8 right panel and Figure S10-S11 in SI). In this case, no alterations in the ^{19}F-NMR spectrum were observed for any substrate addition. In particular, the addition of UDP-GalNAc or UDP did not induce any perturbation of the ^{19}F resonance at δ = -84.0 ppm (Figure 8 and S10-S11). This result also confirms the assignment of the peak at δ = -84.0 ppm to the open conformation of the flexible loop and highlights that this conformation is associated to the inactive form of GalNAc-T2. Furthermore, the ^{19}F-NMR experiments demonstrate that the flexible loop in F104S* is fixed in the open conformation, precluding the binding to peptide substrates and their glycosylation. Overall, the ^{19}F-NMR experiments provide the molecular basis to explain why the F104S mutant does not recognize peptides and in turn is inactive.

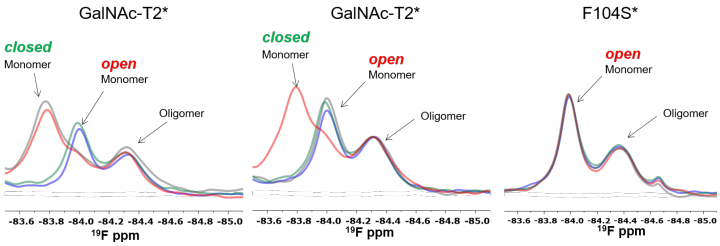


Figure 8. ^{19}F-NMR Spectra of GalNAc-T2* and F104S* showing the open and closed loop conformations as a function of donor and acceptor substrate additions: (Left panel) Blue color corresponds to unliganded GalNAc-T2*. The green color corresponds to GalNAc-T2* in the presence of $MnCl_2$ (225 μM). Red color corresponds to GalNAc-T2*, $MnCl_2$ (225 μM) and UDP-GalNAc (225 μM). Grey color corresponds to GalNAc-T2*, $MnCl_2$ (225 μM), UDP-GalNAc (225 μM) and MUC5Ac (150 μM). (Middle panel) Blue color corresponds to unliganded GalNAc-T2*. Grey color corresponds to GalNAc-T2* and MUC5AC (150 μM). Green color corresponds to GalNAc-T2*, MUC5AC (150 μM) and $MnCl_2$ (225 μM). Red color corresponds to GalNAc-T2* (135 μM), MUC5Ac (150 μM), $MnCl_2$ (225 μM) and UDP-GalNAc (225 μM). (Right panel) Blue color corresponds to unliganded F104S*. Green color corresponds to F104S*, $MnCl_2$ (130 μM). Red color corresponds to F104S*, $MnCl_2$ (130 μM) and UDP-GalNAc (130 μM). Grey color corresponds to F104S*, $MnCl_2$ (130 μM), UDP-GalNAc (130 μM) and MUC5Ac (100 μM). The concentration of GalNAc-T2* and F104S was 135 μM and 86 μM, respectively. The experiments were recorded at 298K and at 600MHz (see SI for experimental details).

Conclusions

Studies of naturally occurring mutations in glycosyltransferases that underlie human diseases or predispositions to diseases or other phenotypes have brought important insights into our understanding of this diverse group of enzymes.[15] In particular, studies of mutations that are not clear about how they affect protein structure and function have resulted in surprisingly new discoveries. From this perspective, our studies with the GalNAc-T2 F104S mutant provides another excellent example. Here, we provide compelling evidence that the inactive-to-active state transition of the flexible loop in the GalNAc-T2 F104S mutant is hindered. The presence of the polar residue (Ser104) causes instability of both the flexible and the α1-α2 loops and prevents the flexible loop from adopting the closed conformation required for the active form of the enzyme. This result is further supported by the conservation of the Phe104, Val360 and Arg362 residues of the flexible loop in most GalNAc-Ts. We also predict that this behavior of the flexible loop will be conserved among this family of enzymes despite their different lengths and their partially different amino acid sequences. Our study has also experimentally demonstrated from our ^{19}F NMR studies that GalNAc-T2 follows an induced-fit-mechanism in which UDP-GalNAc is absolutely required to activate the enzyme.

The finding that the GalNAc-T isoenzymes undergo an activation process induced by the donor substrate UDP-GalNAc may be an important factor in the regulation of the complex *O*-glycosylation process. The GalNAc-T isoenzymes not only have different, albeit partly overlapping, acceptor substrate specificities[9], but they also exhibit rather different apparent K_m values for UDP-GalNAc ranging from 10-160 μM for GalNAc-T1, T2, T3, and T4.[16] Provided that all the GalNAc-T isoenzymes undergo similar donor-induced activation, the existing differences in affinities for the donor may serve as a regulatory mechanism to ensure that the many isoenzymes act in a required order and without a premature competitive binding to the acceptor proteins. This hypothesis is also supported by the lower K_m values for UDP-GalNAc of the first enzymes that are thought to initiate the *O*-glycosylation process, GalNAc-T1 and T2. These enzymes display low apparent K_m values (27 and 10 μM, respectively), while the isoenzymes involved in the follow-up glycosylation of the partially (GalNAc) glycosylated substrates, e.g. GalNAc-T4, show a much higher apparent K_m for UDP-GalNAc (160 μM).[16]

The hexosamine biosynthetic pathway (HBP) regulates the cellular levels of UDP-GlcNAc and the cellular levels of UDP-GalNAc are co-regulated with UDP-GlcNAc through epimerization by the UDP-Gal/UDP-GalNAc 4-epimerase.[17] As the levels of UDP-GlcNAc are highly dynamic, HBP is regarded as a nutrient-sensing pathway.[18] Therefore, it is conceivable that the nutrient-sensing pathway may also influence *O*-GalNAc glycosylation via changes in UDP-GalNAc concentration.

Finally, our results have exploited the knowledge obtained from the structural analysis of a mutant associated with a known disease to obtain fundamental insights into the molecular mechanism of the GalNAc-T substrate recognition. Our multidisciplinary approach may furthermore provide the basis to treat low HDL-C levels through the design and implementation of new therapeutic strategies.

Experimental Section

Cloning, site-directed mutagenesis and purification. The expression plasmid pPICZαA*galnact2* (K75-Q571), previously described[6c] was used as a template for introducing the following single amino-acid changes by site-directed mutagenesis as follows: T375C, F104S and F104S-T375C. Site-directed mutagenesis was carried out using the QuikChange protocol (Stratagene) as recommended by the supplier. In brief, F104S mutation was introduced into *GALNT2* codon optimized K75-Q571 ORF using the template pPICZαA*galnact2* (K75-Q571) and the oligo's T2F104SF (3'-

TACGCAAGAAATAAGTCTAATCAGGTTGAAAGTGATAA-3') and T2F104SR (5'-TTATCACTTTCAACCTGATTAGACTTATTTCTTGCGTA-3'), generating the plasmid pPICZαA*galnact2*-F104S. Note that the position of the mutation is underlined in the above oligos. Hereafter the mutation T375C was introduced by custom based mutagenesis service from Genewiz generating pPICZαA*galnact2*-T375C and pPICZαA*galnact2*-F104S/T375C. The expression plasmid pPICZαA*galnact2* (K75-Q571), previously described[6c] was used as a template for introducing the following single amino-acid changes by site-directed mutagenesis as follows: F104S, T375C and F104S-T375C. Site-directed mutagenesis was carried out following the QuikChange protocol (Stratagene), using the Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific). All plasmids were verified by sequencing The mutants were purified using the purification protocol of the wild-type enzyme described previously.[6c]

Crystallization. Crystals of the F104S mutant in complex with UDP were grown by sitting drop vapor diffusion at 18ºC. Crystals were obtained by mixing 0.5 µl of protein solution (a mix formed by 7 mg/ml F104S mutant, 5 mM UDP, 5 mM $MnCl_2$, 25 mM Tris pH 7.5, 0.5 mM EDTA and 1 mM TCEP-HCl) with 0.5 µl of precipitant solution (10% PEG6000 and 100 mM Hepes pH 7) against 60 µl of precipitant solution. The crystals were soaked within the mother liquor containing UDP-GalNAc at 50 mM for 60 minutes. Subsequently, they were cryoprotected in mother liquor containing 25% ethylene glycol, 10% PEG6000 and 100 mM Hepes pH 7, and frozen in a nitrogen gas stream cooled to 100 K.

Structure determination and refinement. All data were processed and scaled using the XDS package[19] and CCP4 software,[20] relevant statistics are given in **Table S2**. The crystal structures were solved by molecular replacement with Phaser[20b] and using the PDB entry 4D0T as the template. Initial phases were further improved by cycles of manual model building in Coot and refinement with REFMAC5.[21] The final models were validated with PROCHECK,[22] model statistics are given in **Table S2**. The asymmetric unit of the tetragonal crystals contain 2 molecules of the inactive mutant in the AU (**Figure 1**). Coordinates and structure factors have been deposited in the Worldwide Protein Data Bank (wwPDB, and see **Table S2** for the pdb code).

Labeling of the mutants T375C and F104S-T375C. Both the T375C and the F104S-T375C variants were site-selectively modified by treatment with 3-bromo-1,1,1-trifluoroacetone (BFA) under mild conditions (5 min, phosphate buffer pH 7.0, room temperature), as previously described by Rydzik *et al*.[11]

Solid-phase peptide synthesis (SPPS). MUC5AC was synthetized as previously described.[5]

Tryptophan fluorescence spectroscopy. Fluorescence spectroscopy was used to determine the dissociation constants of the WT enzyme and the F104S mutant against UDP-GalNAc. All experiments were carried out in a Cary Eclipse spectrofluorometer (Varian) at 25 °C with the WT enzyme and the F104S mutant at 1 µM, and concentrations of UDP-GalNAc varying from 1 to 1000 µM in 25 mM Tris, 150 mM NaCl, 1 mM $MnCl_2$, pH 7.5. Fluorescence emission spectra were recorded in the 300–400 nm range with an excitation wavelength of 280 nm, with slit width of 5 nm. The data analysis was performed in Prism (GraphPad software) considering a model with a single binding site (see equation (1), where F0 is the intrinsic fluorescence of the enzyme in the absence of quencher (Q), F1 is the observed fluorescence at a given quencher concentration, f_a is the fractional degree of fluorescence and K_d is the dissociation constant.

Eq 1.

$$1 - \frac{F1}{F0} = \frac{fa * [Q]}{Kd + [Q]}$$

Activity assay The WT GalNAc-T2 and GalNAc-T2* activities were tested by incubating 100 nM of each enzyme with 10 µM MUC1 peptide[23] and 20 µM UDP-[^{14}C]GalNAc (2000 cpms/nm), in 25 mM cacodylate pH 7.4, 10 mM $MnCl_2$, 0.25% Triton X-100. Excess of UDP-GalNAc was separated from the glycosylated peptide trough DOWEX 1 chromatography. The product was analyzed by scintillation counting.

Thermal shift assays. The protein unfolding process is monitored using the environmentally sensitive dye 1-anilino-8-naphthalene sulfonate (ANS). Its quantum yield increases upon binding to hydrophobic surfaces exposed during protein unfolding and so does the fluorescence signal. The thermal-shift assay was conducted in the FluoDia T70 (Photal Otsuka Electronics). Solutions of 2 µM WT GalNAc-T2 or the F104S mutant, 100 µM ANS, 20 mM HEPES pH 7.0, and 200 mM NaCl final concentration and a volume of 100 µl were added to the wells of a 96-well PCR plate. Mineral oil was last added to avoid evaporation during the experiment. Data were analysed with the Origin software. To obtain T_0, a Boltzmann model from Origin was used to fit the fluorescence imaging data.[8]

Molecular Dynamics simulations. MD simulations were performed using Amber16 and the force-field ff14SB[24] Each protein was immersed in a box with a 10 Å buffer of TIP3P[25] water molecules and neutralized by adding explicit counter ions (Cl^-). A two-stage geometry optimization approach was used. The first stage minimizes only the positions of solvent molecules and ions, and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems were then gently heated by incrementing the temperature from 0 to 300 K under a constant pressure of 1 atm and periodic boundary conditions. Harmonic restraints of 30 kcal/mol were applied to the solute, and the Andersen temperature coupling scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages. Water molecules are treated with the SHAKE algorithm such that the angle between the hydrogen atoms is kept fixed. Long-range electrostatic effects are modeled using the particle-Mesh-Ewald method.[26] An 8-Å cut-off was applied to Lennard-Jones interactions. Each system was equilibrated for 2 ns with a 2-fs time step at a constant volume and temperature of 300 K. Production trajectories were then run for additional 200 ns under the same simulation conditions.

nLC MS/MS and data analysis. To preserve the trifluorinated compound at the peptide backbone, samples were not reduced and alkylated. GalNAcT* and F104S* were directly digested with trypsin and after StageTip (Empore 3M) desalting, submitted to nLC-MS/MS analysis. EASY-nLC 1000 UHPLC (Thermo Scientific) interfaced via nanoSpray Flex ion source to an Orbitrap Fusion MS (Thermo Scientific) were used for analysis. Briefly, the nLC was operated in a single analytical column set up using PicoFrit Emitters (New Objectives, 75 µm inner diameter) packed in-house with Reprosil-Pure-AQ C18 phase (Dr. Maisch, 1.9-µm particle size, 19-21 cm column length). Each sample was injected onto the column and eluted in a gradient from 2 to 25 % B in 45 min at 200 nL/min (Solvent A, 100 % H2O; Solvent B, 100 % acetonitrile; both containing 0.1 % (v/v) formic acid). A precursor MS1 scan (m/z 350–2,000) of intact peptides was acquired in the Orbitrap at the nominal resolution setting of 120.000, followed by Orbitrap HCD-MS2 and at the nominal resolution setting of 60.000 of the five most abundant multiply charged precursors in the MS1 spectrum; a minimum MS1 signal

threshold of 50,000 was used for triggering data-dependent fragmentation events. Activation times were 75 ms for HCD fragmentation; isolation width was 1.2 mass units, and usually 1 microscan was collected for each spectrum. Polysiloxane ions at *m/z* 445.12003 were used as a lock mass in all runs. Data processing was performed using Proteome Discoverer 1.4 software (Thermo Scientific) using Sequest HT node as a search engine. In all cases the precursor mass tolerance was set to 10 ppm and fragment ion mass tolerance to 20 mmu. Methionine oxidation and the trifluorinated compound at Ser/Thr were used as variable modifications. All spectra were searched against GalNAcT2* and F104S* sequences with and without mutations at the position 375.

NMR experiments. NMR experiments for peptide characterization and saturation transfer difference NMR experiments were recorded on a Bruker Avance 600 MHz spectrometer equipped with a triple channel cryoprobe head. The ^{1}H-NMR resonances of the peptide were 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 0.5 mM peptide in 90:10 H_2O/D_2O. The assignments were accomplished either at 278K. As a chemical shift reference in the ^{1}H NMR experiments we used the resonance of 2,2,3,3-tetradeutero-3-trimethylsilylpropionic acid (TSP) (δ TSP = 0 ppm). Peak lists for the 2D-TOCSY and 2D-NOESY spectra were generated by interactive peak picking using the **C**omputer **A**ided **R**esonance **A**ssignment (CARA) software. Table S1 contains the 1H-NMR assignment for the MUC5AC peptide. Samples for STD-experiments were prepared in perdeuterated 25 mM TRIS-d11 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 75 μM UDP, 75 μM $MnCl_2$ with a molar ratio of 40:1 peptide:protein. The STD-NMR spectra were acquired with 1920 transients in a matrix with 64k data points in t2 in a spectral window of 12335.5 Hz centered at 2820.6 Hz. An excitation sculpting module with gradients 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° 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 STD-NMR control experiments were performed: i) with the ligand in absence of the protein to ensure that the ligand signals were not affected; ii) with the protein in absence of the ligand (data not shown). The peptide MUC5AC when irradiated at -1 ppm in the absence of protein showed residual saturation on the aliphatic methyl groups in the STD-NMR spectra. The STD-NMR experiments shown in this work take into account the STD control experiments. To obtain the STD-NMR-derived epitope mapping the STD-NMR total intensities were normalized with respect the highest STD-NMR response. ^{19}F-NMR experiments were carried out to monitor the ^{19}F reporter $(CF_3)_2COCH2$- selectively introduced at the SH- group of the mutants T375C from both constructs of the WT GalNAc-T2 and the F104S mutant. The ^{19}F experiments of GalNAcT2* and F104S* constructs were recorded at distinct conditions in absence and presence of $MnCl_2$, UDP or UDP-GalNAc and peptide. The UDP or UDP-GalNAc and $MnCl_2$ 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 μM), $MnCl_2$ (500μM) and peptide (protein:peptide molar ratio 1:2.2) do not induce higher ^{19}F chemical shift perturbations. The ^{19}F-NMR experiments were acquired at 298 K using a tdF.Hdec pulse sequence by acquisition of 640 scans in a matrix with 8k data points in a spectral window of 11295.2 Hz centered at $(CF_3)_2CHOH$ -45174.9 Hz. The resonance of 1,1,1,3,3,3-hexafluoro-2-propanol was used as a chemical shift reference in the ^{19}F NMR experiments (δ = -75.7 ppm). Samples for ^{19}F-NMR experiments were prepared in 25 mM TRIS-d11 in 90:10 H_2O/D_2O, uncorrected pH 7.5. In the ^{19}F-NMR spectra, trace amounts of the 3-Bromo-1,1,1-trifluoroacetone reagent used in protein labelling (δ CF_3COCH_2Br= -83.3 ppm) were observed. In addition, trace amounts of trifluoroacetic acid (TFA), used in the synthesis of peptides, were observed in the ^{19}F-NMR spectra i (δ TFA= -75.4 ppm).To monitor the glycosylation of MUC5AC by GalNAcT2* and F104S* two ^{1}H-NMR spectra were recorded before and after the addition of UDP-GalNAc.

Acknowledgements

We thank the Ministerio de Economía y Competitividad (CTQ2013-44367-C2-2-P and BFU2016-75633-P to R.H-G., CTQ2015-67727-R to F.C., CTQ2015-64597-C2-1P to J.J-B). F.M. thanks FCT-Portugal to IF grant and UCIBIO funding UID/Multi/04378/2013 co-financed by the FEDER (POCI-01-0145-FEDER-007728). The NMR spectrometers are part of PTNMR supported by Project No 022161. H.C. thanks the Lundbeck Foundation and the Danish National Research Foundation (DNRF107). E.L-N. acknowledges her postdoctoral EMBO fellowship ALTF 1553-2015 co-funded by the European Commission (LTFCOFUND2013, GA-2013-609409) and Marie Curie Actions. H.C. and J.J-B. thank EU for the TOLLerant project. R.H-G. thanks Agencia Aragonesa para la Investigación y Desarrollo (ARAID) and the Diputación General de Aragón (DGA, B89) for financial support. The research leading to these results has also received funding from the FP7 (2007-2013) under BioStruct-X (grant agreement N°283570 and BIOSTRUCTX_5186). We thank synchrotron radiation source DIAMOND (Oxford) and beamline I04 (number of experiment MX10121-19).

Keywords: GalNAc-T2• F104S mutant • X-ray crystallography • missense mutations • NMR

[1] C. Steentoft, S. Y. Vakhrushev, H. J. Joshi, Y. Kong, M. B. Vester-Christensen, K. T. Schjoldager, K. Lavrsen, S. Dabelsteen, N. B. Pedersen, L. Marcos-Silva, R. Gupta, E. P. Bennett, U. Mandel, S. Brunak, H. H. Wandall, S. B. Levery, H. Clausen, *Embo J* **2013**, *32*, 1478-1488.

[2] S. Fukumoto, *Curr Opin Nephrol Hypertens* **2014**, *23*, 346-351.

[3] K. Kato, C. Jeanneau, M. A. Tarp, A. Benet-Pages, B. Lorenz-Depiereux, E. P. Bennett, U. Mandel, T. M. Strom, H. Clausen, *J Biol Chem* **2006**, *281*, 18370-18377.

[4] S. A. Khetarpal, K. T. Schjoldager, C. Christoffersen, A. Raghavan, A. C. Edmondson, H. M. Reutter, B. Ahmed, R. Ouazzani, G. M. Peloso, C. Vitali, W. Zhao, A. V. Somasundara, J. S. Millar, Y. Park, G. Fernando, V. Livanov, S. Choi, E. Noe, P. Patel, S. P. Ho, S. Myocardial Infarction Exome Sequencing, T. G. Kirchgessner, H. H. Wandall, L. Hansen, E. P. Bennett, S. Y. Vakhrushev, D. Saleheen, S. Kathiresan, C. D. Brown, R. Abou Jamra, E. LeGuern, H. Clausen, D. J. Rader, *Cell Metab* **2016**, *24*, 234-245.

[5] E. Lira-Navarrete, M. de Las Rivas, I. Companon, M. C. Pallares, Y. Kong, J. Iglesias-Fernandez, G. J. Bernardes, J. M. Peregrina, C. Rovira, P. Bernado, P. Bruscolini, H. Clausen, A. Lostao, F. Corzana, R. Hurtado-Guerrero, *Nat Commun* **2015**, *6*, 6937.

[6] aM. Ghirardello, M. de Las Rivas, A. Lacetera, I. Delso, E. Lira-Navarrete, T. Tejero, S. Martin-Santamaria, R. Hurtado-Guerrero, P. Merino, *Chemistry* **2016**, *22*, 7215-7224; bR. Hurtado-Guerrero, *Biochem Soc Trans* **2016**, *44*, 61-67; cE. Lira-Navarrete, J. Iglesias-

Fernandez, W. F. Zandberg, I. Companon, Y. Kong, F. Corzana, B. M. Pinto, H. Clausen, J. M. Peregrina, D. J. Vocadlo, C. Rovira, R. Hurtado-Guerrero, *Angew Chem Int Ed Engl* **2014**, *53*, 8206-8210.

[7] aT. A. Fritz, J. H. Hurley, L. B. Trinh, J. Shiloach, L. A. Tabak, *Proc Natl Acad Sci U S A* **2004**, *101*, 15307-15312; bT. A. Fritz, J. Raman, L. A. Tabak, *J Biol Chem* **2006**, *281*, 8613-8619; cT. Kubota, T. Shiba, S. Sugioka, S. Furukawa, H. Sawaki, R. Kato, S. Wakatsuki, H. Narimatsu, *J Mol Biol* **2006**, *359*, 708-727; dM. L. Rivas, E. Lira-Navarrete, E. J. P. Daniel, I. Companon, H. Coelho, A. Diniz, J. Jimenez-Barbero, J. M. Peregrina, H. Clausen, F. Corzana, F. Marcelo, G. Jimenez-Oses, T. A. Gerken, R. Hurtado-Guerrero, *Nat Commun* **2017**, *8*, 1959.

[8] E. Lira-Navarrete, J. Valero-Gonzalez, R. Villanueva, M. Martinez-Julvez, T. Tejero, P. Merino, S. Panjikar, R. Hurtado-Guerrero, *PLoS One* **2011**, *6*, e25365.

[9] T. A. Gerken, O. Jamison, C. L. Perrine, J. C. Collette, H. Moinova, L. Ravi, S. D. Markowitz, W. Shen, H. Patel, L. A. Tabak, *J Biol Chem* **2011**, *286*, 14493-14507.

[10] J. Raman, T. A. Fritz, T. A. Gerken, O. Jamison, D. Live, M. Liu, L. A. Tabak, *J Biol Chem* **2008**, *283*, 22942-22951.

[11] A. M. Rydzik, J. Brem, S. S. van Berkel, I. Pfeffer, A. Makena, T. D. Claridge, C. J. Schofield, *Angew Chem Int Ed Engl* **2014**, *53*, 3129-3133.

[12] W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten, G. Vriend, *Nucleic Acids Res* **2015**, *43*, D364-368.

[13] J. M. Aramini, K. Hamilton, L. C. Ma, G. V. T. Swapna, P. G. Leonard, J. E. Ladbury, R. M. Krug, G. T. Montelione, *Structure* **2014**, *22*, 515-525.

[14] B. Ramakrishnan, E. Boeggeman, P. K. Qasba, *Biochem Biophys Res Commun* **2002**, *291*, 1113-1118.

[15] H. H. Freeze, H. Schachter, T. Kinoshita, in *Essentials of Glycobiology*, 3rd ed. (Eds.: A. Varki, R. D. Cummings, J. D. Esko, P. Stanley, G. W. Hart, M. Aebi, A. G. Darvill, T. Kinoshita, N. H. Packer, J. H. Prestegard, R. L. Schnaar, P. H. Seeberger), Cold Spring Harbor (NY), **2015**.

[16] E. P. Bennett, H. Hassan, U. Mandel, E. Mirgorodskaya, P. Roepstorff, J. Burchell, J. Taylor-Papadimitriou, M. A. Hollingsworth, G. Merkx, A. G. van Kessel, H. Eiberg, R. Steffensen, H. Clausen, *J Biol Chem* **1998**, *273*, 30472-30481.

[17] aB. A. Lewis, J. A. Hanover, *J Biol Chem* **2014**, *289*, 34440-34448; bJ. B. Thoden, T. M. Wohlers, J. L. Fridovich-Keil, H. M. Holden, *J Biol Chem* **2001**, *276*, 15131-15136.

[18] S. Marshall, V. Bacote, R. R. Traxinger, *J Biol Chem* **1991**, *266*, 4706-4712.

[19] W. Kabsch, *Acta Crystallogr D Biol Crystallogr* **2010**, *66*, 125-132.

[20] a*Acta Crystallogr D Biol Crystallogr* **1994**, *50*, 760-763; bM. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, *Acta Crystallogr D Biol Crystallogr* **2011**, *67*, 235-242.

[21] G. N. Murshudov, P. Skubak, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, *Acta Crystallogr D Biol Crystallogr* **2011**, *67*, 355-367.

[22] R. A. e. a. Laskowski, *J. Appl. Cryst* **1993**, *26*

[23] H. H. Wandall, F. Irazoqui, M. A. Tarp, E. P. Bennett, U. Mandel, H. Takeuchi, K. Kato, T. Irimura, G. Suryanarayanan, M. A. Hollingsworth, H. Clausen, *Glycobiology* **2007**, *17*, 374-387.

[24] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins* **2006**, *65*, 712-725.

[25] T. A. Andrea, W. C. Swope, H. C. Andersen, *J. Chem. Phys.* **1983**, *79*, 4576–4458.

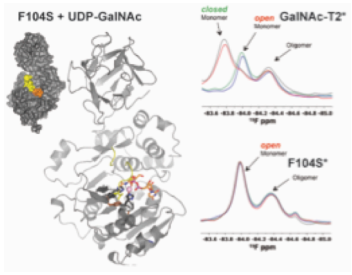
[26] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1993**, *98*, 10089-10092.

Entry for the Table of Contents (Please choose one layout)

Layout 1:

FULL PAPER

Decoding disease-causing mechanisms of missense mutations: Here, we have studied the GalNAc-T2 F104S mutant, which causes low levels of HDL-C and predisposition to cardiovascular diseases. The mutation impedes the shifting of the flexible loop from the inactive to the active conformation, affecting the enzyme ability to bind its peptide substrates. We also demonstrate that UDP-GalNAc induces the active conformation to promote *O*-glycosylation, suggesting that UDP-GalNAc acts as a regulatory molecule of GalNAc-T isoenzymes.



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, and Ramon Hurtado-Guerrero**

Page No. – Page No.

Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts

Supporting Information for

Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts

Matilde de las Rivas,[a]^ Helena Coelho, [b],[c]^ Ana Diniz,[b] Erandi Lira-Navarrete,[d] Ismael Compañón,[e] Jesús Jiménez-Barbero,[c],[f] Katrine. T. Schjoldager,[d] Eric. P. Bennett,[d] Sergey Y. Vakhrushev,[d] Henrik Clausen,[d] Francisco Corzana,[e] Filipa Marcelo,*[b] and Ramon Hurtado-Guerrero*[a],[g]

Supplementary Figures and Tables

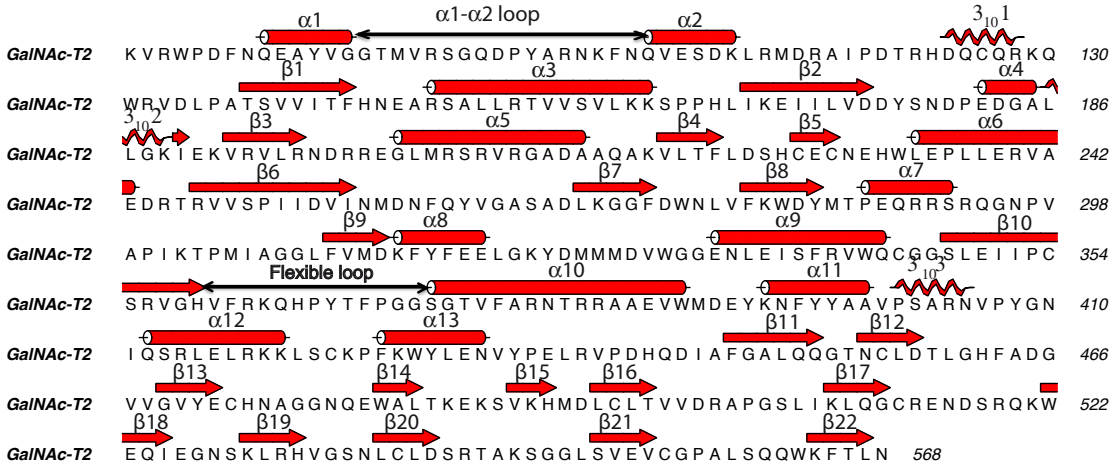
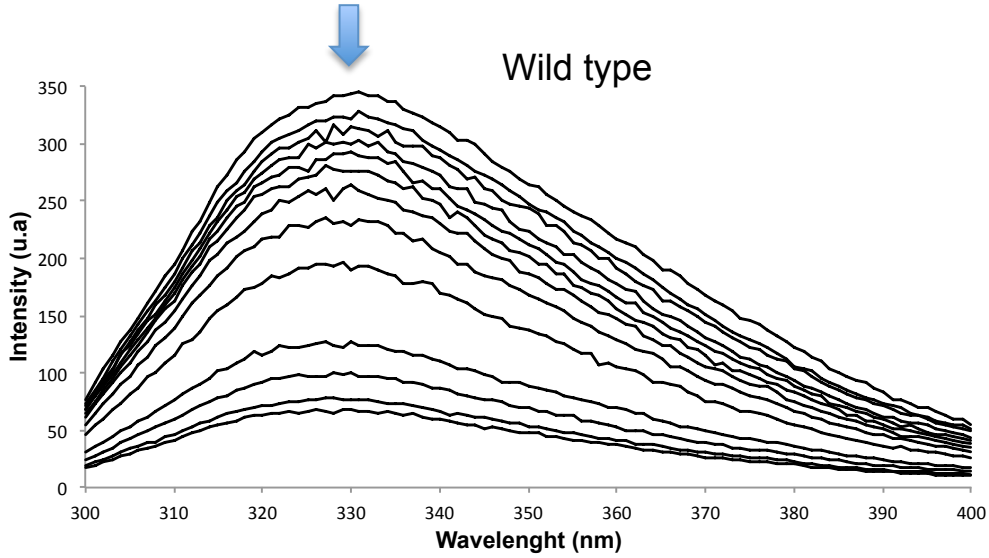


Figure S1. Sequence of GalNAc-T2 showing the secondary structure elements together with the α1- α2 loop and the flexible loop**.**



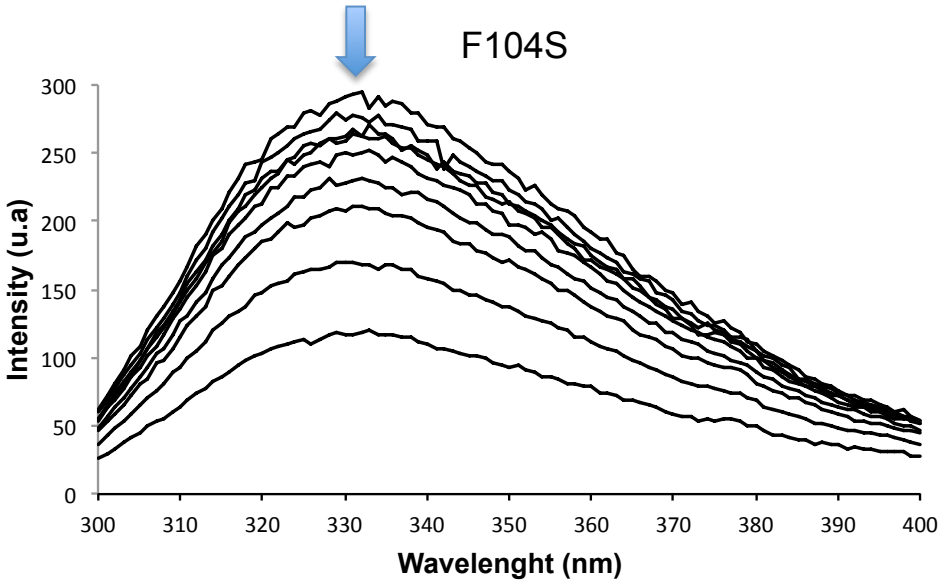


Figure S2. Quenching of intrinsic WT GalNAc-T2 and F104S mutant tryptophan fluorescence measured at increasing concentrations of UDP-GalNAc. All data points represent the means ± S.D. for three measurements.

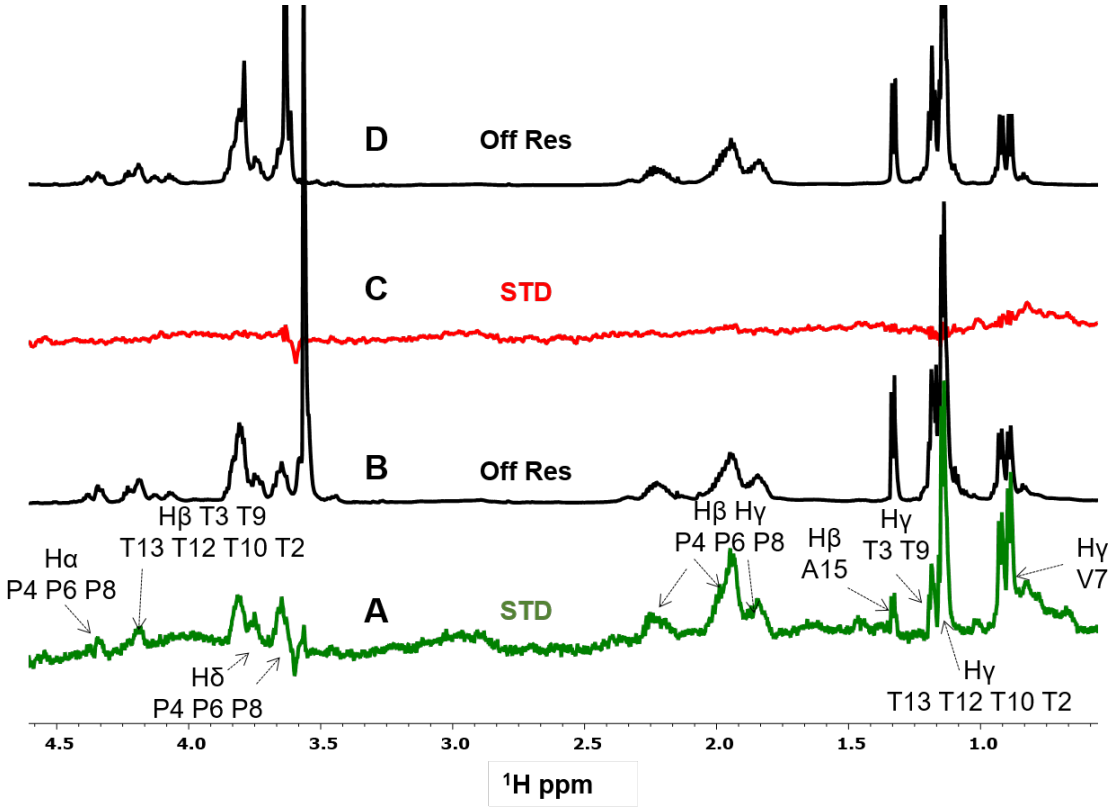


Figure S3. STD-NMR experiment recorded at 298K and 600MHz**:** a**)** STD spectrum of MUC5AC (560 µM) in the presence of WT GalNAc-T2 (14 µM), UDP (75 µM) and $MnCl_2$ (75 µM). b) The reference spectrum (Off Res) of MUC5AC (560 µM) in the presence of WT GalNAc-T2 (14 µM), UDP (75 µM) and $MnCl_2$ (75 µM). c**)** STD spectrum of MUC5AC (740 µM) in the presence of the F104S mutant (18.5 µM), UDP (75 µM) and $MnCl_2$ (75 µM). d) The reference spectrum (Off Res) of MUC5AC (740 µM) in the presence of the F104S mutant (18.5 µM), UDP (75 µM) and MnCl2 (75 µM). The key proton resonances are marked in the figure.

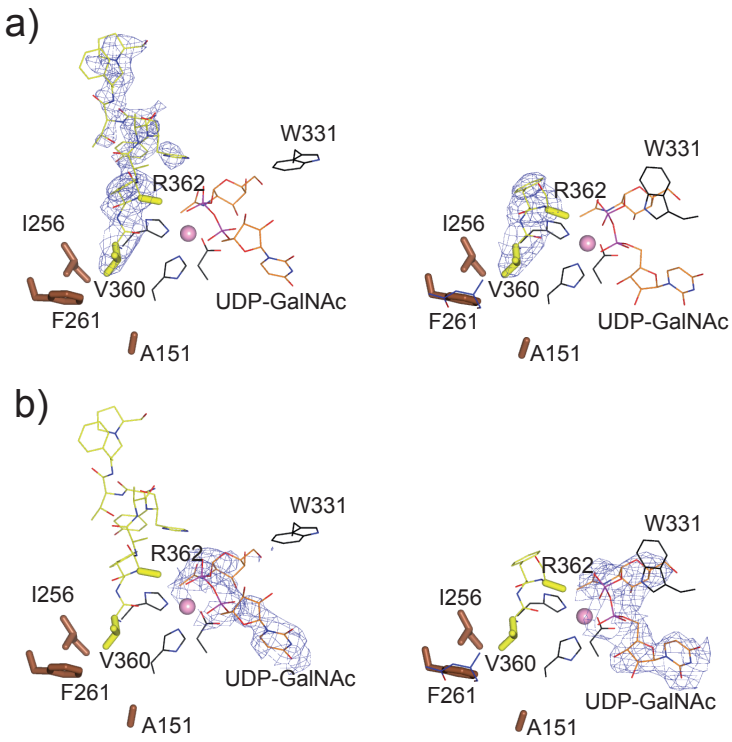


Figure S4. a) Electron density map is $2F_O–F_C$ syntheses (blue) is shown at 1.0 σ for amino acids from the flexible loop. b) Unbiased difference electron density map ($F_O–F_C$ syntheses) in blue is shown at 2.2 σ for UDP-GalNAc. Colors for all the amino acids and UDP-GalNAc are the same as shown in Figure 4**.** Note that the active site of the two molecules of the F104S in complex with UDP-GalNAc is shown. This is because there are two molecules of F104S within the asymmetric unit.

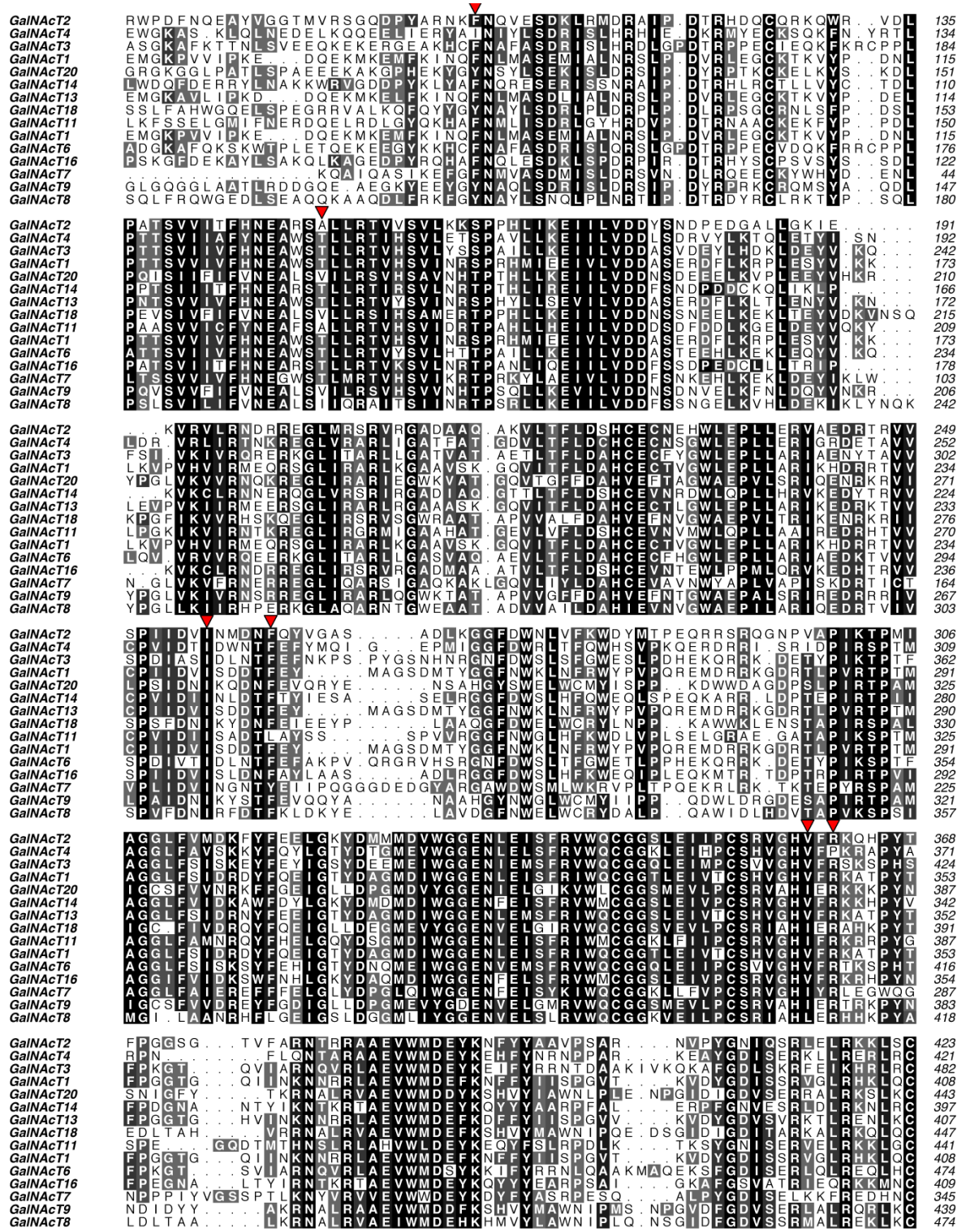


Figure S5. Multiple alignment between GalNAc-Ts. The inverted triangles in red indicate the location of Phe104 and the residues interacting with it. Note that Phe104, Ile256, Val360, Phe261 and Arg362 are highly conserved along the different GalNAc-Ts. In the case of Ala151, other GalNAc-Ts in the equivalent position contain hydrophobic or Thr residues. This latter residue might interact with the Phe residue through the methyl group, thus keeping the same hydrophobic interaction between Phe104 and Ala151. The full-length sequence of each GalNAc-T isoenzyme is not shown for clarification purposes. Note that 15 out of 20 GalNAc-T isoenzymes are aligned.

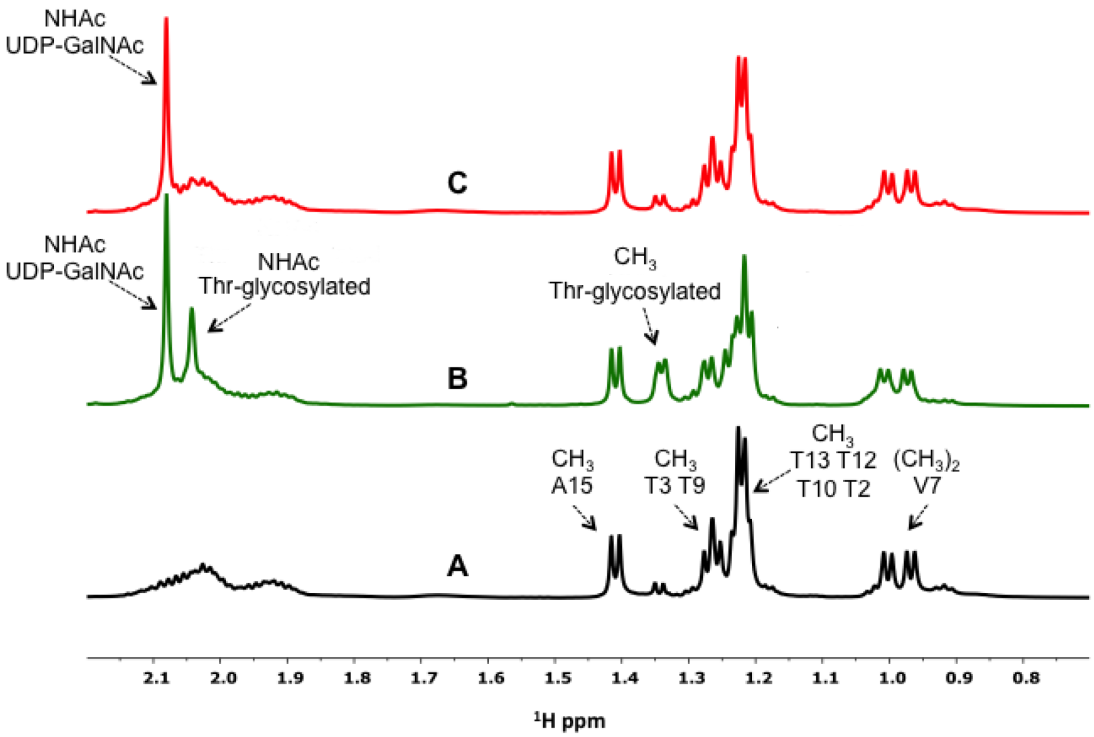


Figure S6. Glycosylation experiment of MUC5AC at 298 K and 600 MHz: a) ^{1}H-NMR of MUC5AC (1760 µM). b) ^{1}H-NMR of MUC5AC (1760 µM) in presence of GalNAc-T2* (22 µM), $MnCl_2$ (150 µM) and 5280 µM UDP-GalNAc after 12 h. The appearance of new NMR signals corresponding to the NHAc methyl group of the GalNAc-Thr motif (δ 2.05 ppm) and dramatic alteration of the methyl protons of the Thr amino acids of the peptide (δ 1.20-1.35 ppm) clearly indicate glycosylation of the peptide. c) ^{1}H-NMR of MUC5AC (1760 µM) in presence of F104S* (22 µM), $MnCl_2$ (150 µM) and 5280 µM UDP-GalNAc after 12 h. ^{1}H-NMR were recorded at distinct times, 5 min, 1 hour and after 12h after addition of UDP-GalNAc (data not shown). Alterations in the ^{1}H-NMR spectrum of MUC5AC were exclusively observed in the presence of GalNAc-T2*.

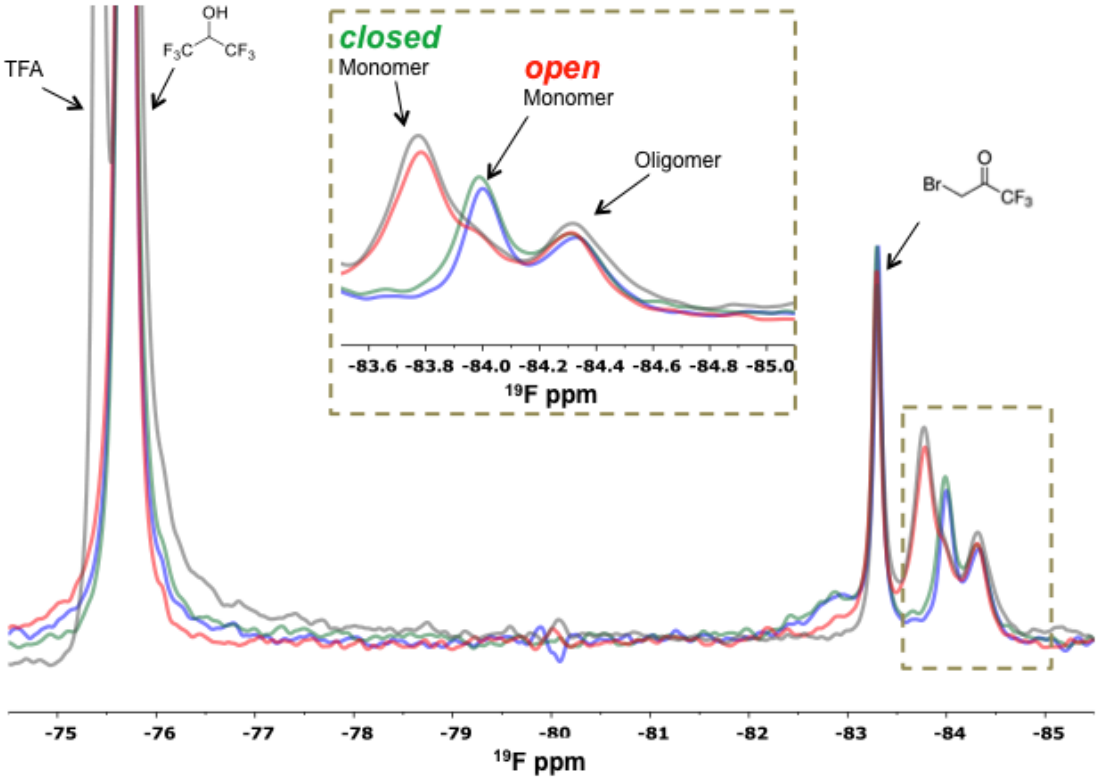


Figure S7. ^{19}F-NMR Spectra of GalNAc-T2* at 298 K and 600 MHz. Blue color corresponds to GalNAc-T2* (135 µM). Green color corresponds to GalNAc-T2* (135 µM) and $MnCl_2$ (225 µM). Red color corresponds to GalNAc-T2* (135 µM), $MnCl_2$ (225 µM) and UDP-GalNAc (225 µM). Grey color corresponds to the GalNAc-T2* (135 µM), $MnCl_2$ (225 µM), UDP-GalNAc (225 µM) and MUC5AC (150 µM). Extra signals in the ^{19}F-NMR spectra can be observed that corresponds to the external reference 1,1,1,3,3,3-hexafluoro-2-propanol δ = -75.7 ppm (added to the samples to calibrate the ^{19}F-NMR experiments), the 3-bromo-1,1,1-trifluoroacetone reactive used in protein labelling (δ = -83.3 ppm) and trifluoracetic acid used in peptide synthesis (δ = -75.4 ppm). These extra signals can be also found in Figures S8-S11.

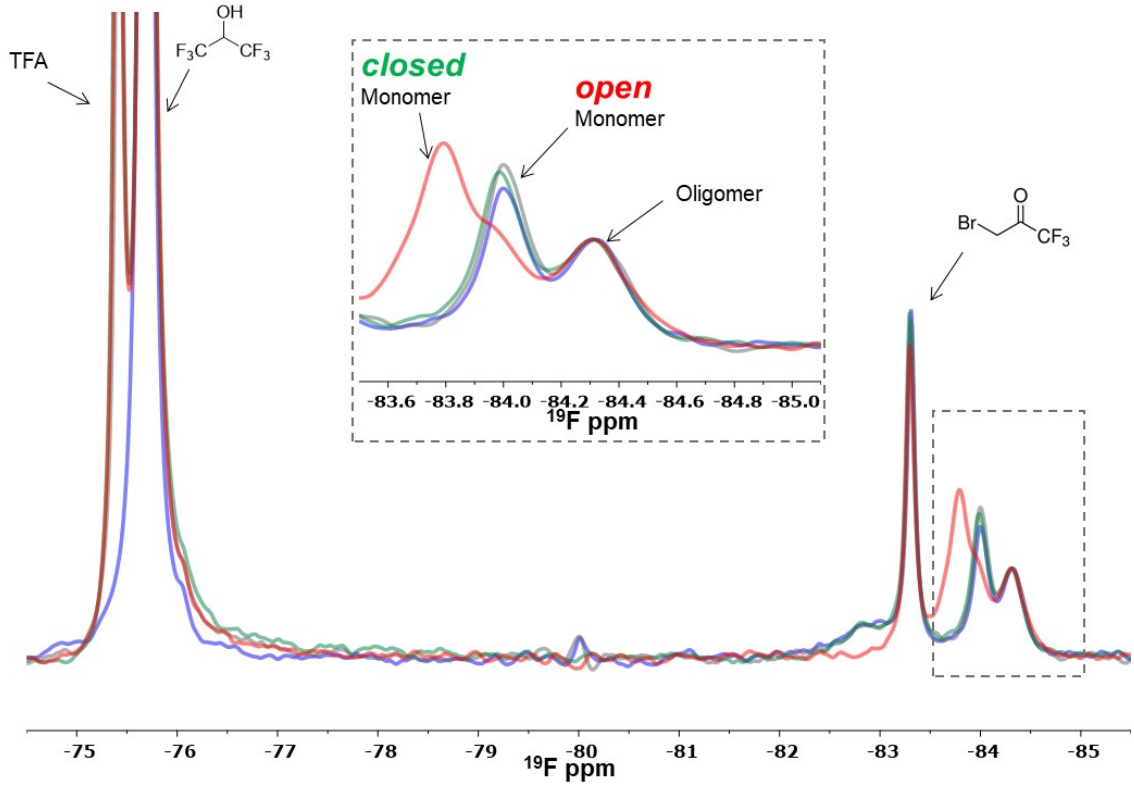


Figure S8. ^{19}F-NMR Spectra of GalNAc-T2* at 298 K and 600 MHz. Blue color corresponds to GalNAc-T2* (135 µM). Grey color corresponds to GalNAc-T2* (135 µM), MUC5AC (150 µM). Green color corresponds to GalNAc-T2* (135 µM), MUC5AC (150 µM) and $MnCl_2$ (225 µM). Red color corresponds to GalNAc-T2* (135 µM), MUC5AC (150 µM), $MnCl_2$ (225 µM) and UDP-GalNAc (225 µM).

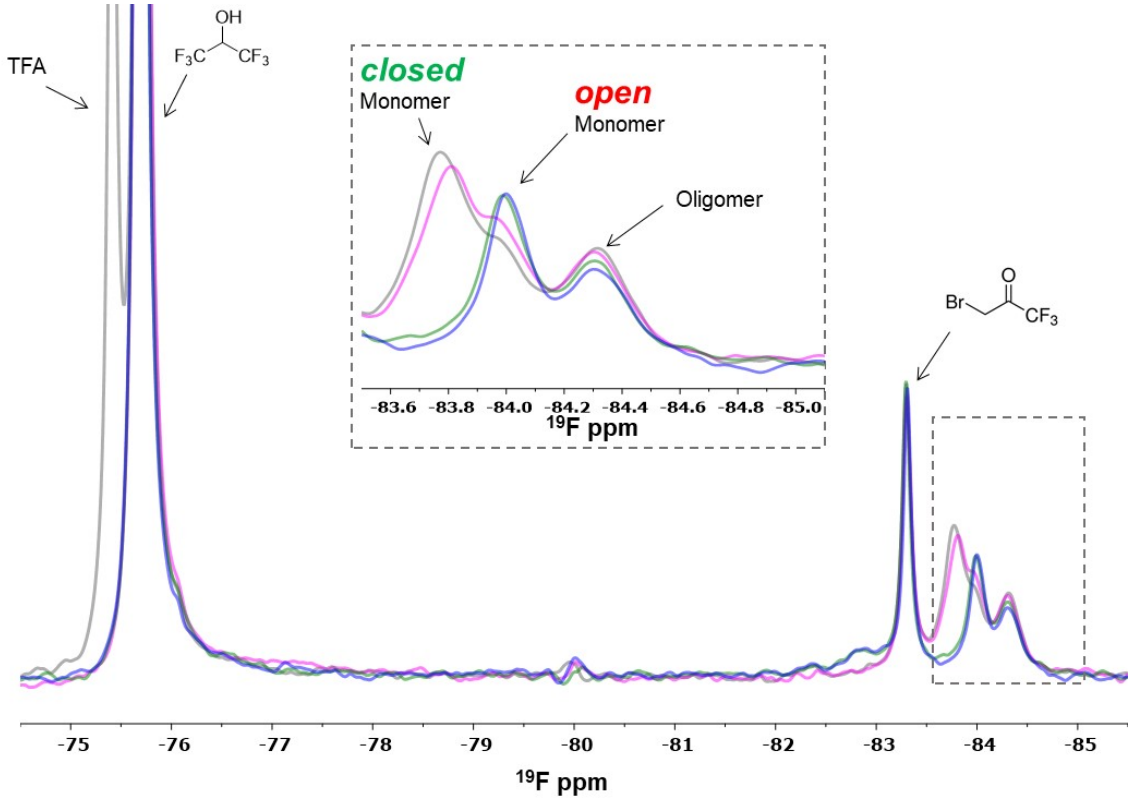


Figure S9. ^{19}F-NMR Spectra of GalNAc-T2* at 298K and 600 MHz. Blue corresponds to GalNAc-T2* (135 µM). Green corresponds to GalNAc-T2* (135 µM), $MnCl_2$ (225 µM). Magenta corresponds to GalNAc-T2* (135 µM), $MnCl_2$ (225 µM) and UDP (225 µM). Gray corresponds to GalNAc-T2* (135 µM), $MnCl_2$ (225 µM), UDP (225 µM) and MUC5AC (150 µM).

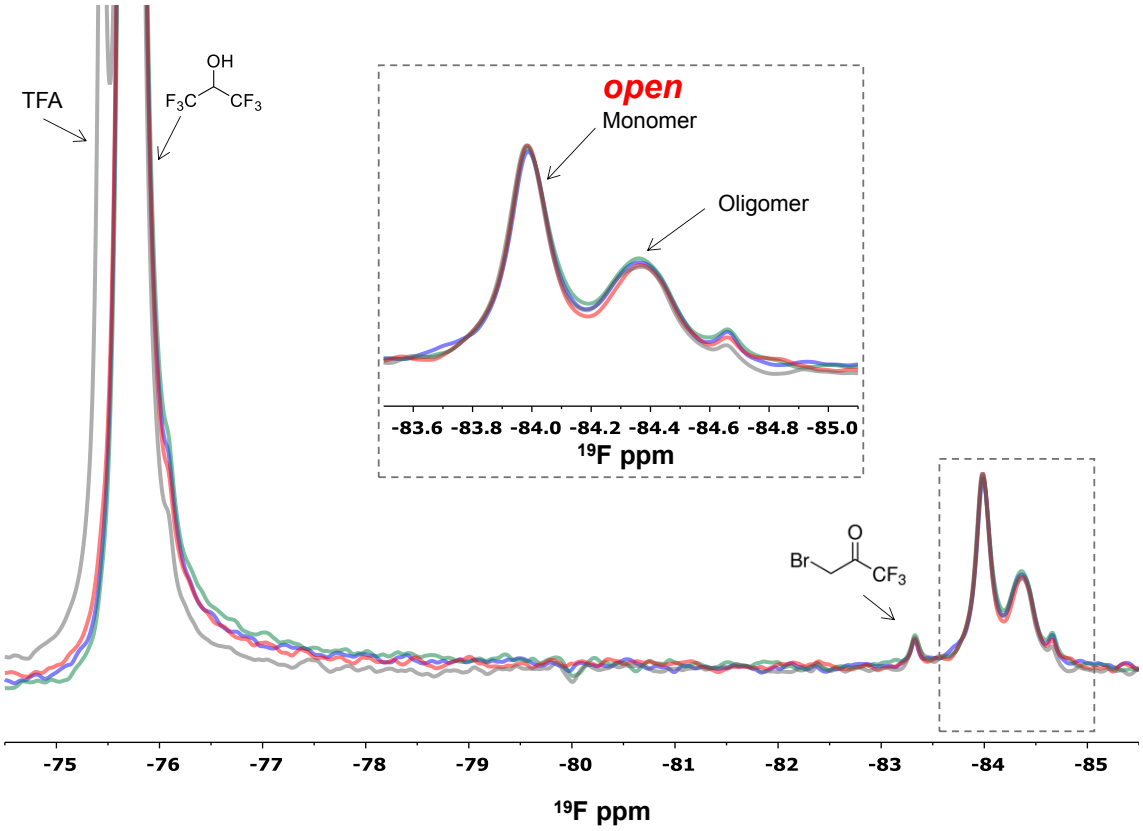


Figure S10. ^{19}F-NMR Spectra of F104S* at 298 K and 600 MHz. Blue corresponds to F104S* (86 µM). Green corresponds to F104S* (86 µM) and $MnCl_2$ (130 µM). Red corresponds to F104S* (86 µM), $MnCl_2$ (130 µM) and UDP-GalNAc (130 µM). Grey corresponds to F104S* (86 µM), $MnCl_2$ (130 µM), UDP-GalNAc (130 µM) and MUC5AC (100 µM).

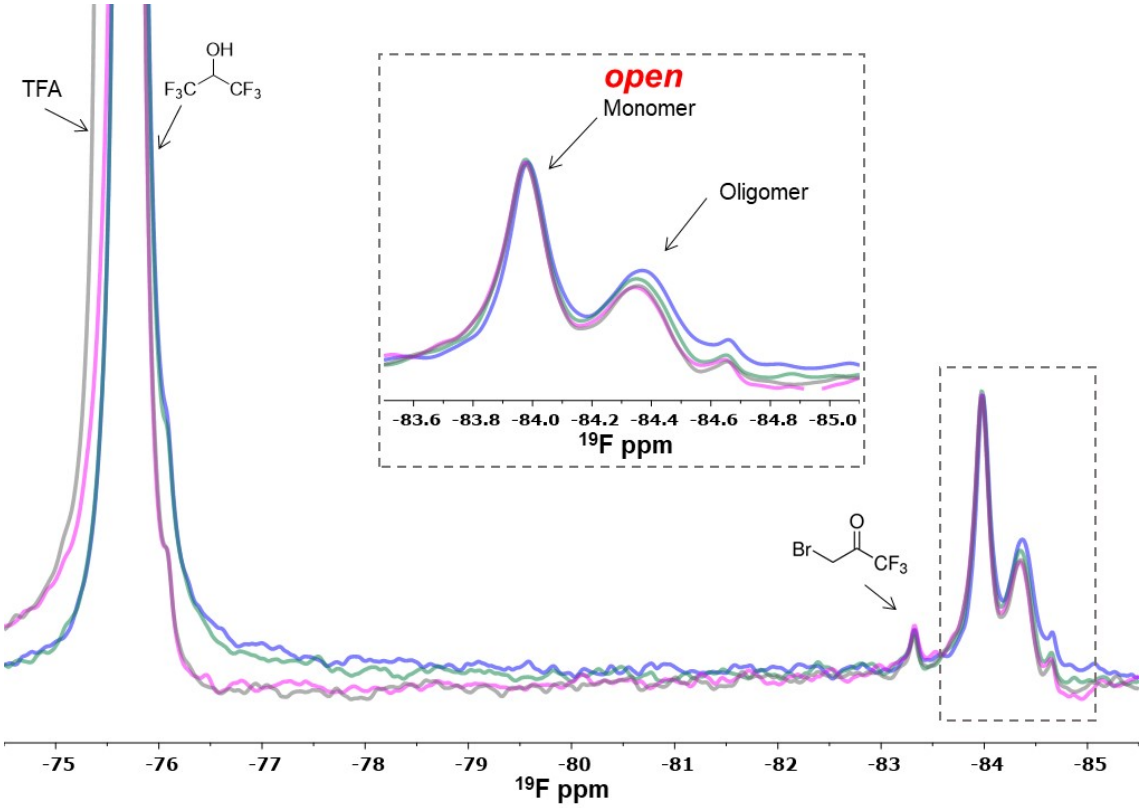


Figure S11. ^{19}F-NMR Spectra of F104S* at 298 K and 600 MHz. Blue corresponds to F104S* (86 µM). Green corresponds to F104S* (86 µM) and $MnCl_2$ (130 µM). Magenta corresponds to F104S* (86 µM), $MnCl_2$ (130 µM) and UDP (130 µM). Grey corresponds to F104S* (86 µM), $MnCl_2$ (130 µM), UDP (130 µM) and MUC5AC (100 µM).

Table S1. ^{1}H-NMR assignments of the MUC5AC peptide (ppm), recorded at Bruker Avance 600 MHz spectrometer equipped with a triple channel cryoprobe head with H_20/D_2O (90%/10%) as solvent.

Residue	*Proton*	*Chemical shift (ppm)*
G1	NH	8.143
	Hα	3.674
T2	NH	8.275
	Hα	3.984
	Hγ	0.691
T3	NH	8.155
	Hα	3.952
	Hβ	3.952
	Hγ	0.957
P4	Hα	4.135
	Hβ	1.960
	Hγ	1.686
	Hδ	3.575
S5	NH	8.333
	Hα	4.403
	Hβ	3.533
P6	Hα	4.149
	Hβ	1.942
	Hγ	1.674
	Hδ	3.539
V7	NH	8.084
	Hα	4.069
	Hβ	1.763
	Hγ	0.668
P8	Hα	4.181
	Hβ	2.002
	Hγ	1.644
	Hδ	3.598
T9	NH	8.214
	Hα	4.097
	Hβ	3.959
	Hγ	0.945
T10	NH	7.998
	Hα	4.107
	Hβ	3.910
	Hγ	0.851
S11	NH	7.907
	Hα	4.255
	Hβ	3.517
T12	NH	7.988
	Hα	4.107
	Hβ	3.913
	Hγ	0.867
T13	NH	7.981
	Hα	4.117
	Hβ	3.906
	Hγ	0.886
S15	NH	8.159
	Hα	4.321
	Hβ	3.704
A15	NH	8.100
	Hα	3.967
	Hβ	1.089

Table S2. Data collection and refinement statistics. Values in parentheses refer to the highest resolution shell. Ramachandran plot statistics were determined with PROCHECK.

	F104S mutant in complex with UDP-GalNAc
Space group	$P4_12_12$
Wavelength (Å)	0.97
Resolution (Å)	20-2.70 (2.85-2.70)
Cell dimensions a, b, c (Å) α, β, γ (°)	*a = 154.04* *b* = 154.04 *c* = 109.82 *90, 90, 90*
Mn(I) half-set correlation CC(1/2)	0.995 (0.371)
Unique reflections	70380
Completeness	99.7 (100)
R_{pim}	0.064 (0.720)
$I/\sigma(I)$	7.8 (1.1)
Redundancy	5.6 (5.5)
R_{work} / R_{free}	0.214/0.230
RMSD from ideal geometry, bonds (Å)	0.011
RMSD from ideal geometry, angles (º)	1.56
<*B*> protein (Å^2)	74.50
<*B*> UDP-GalNAc (Å^2)	91.05
<*B*> solvent (Å^2)	55.30
<*B*> Mn^{+2}(Å^2)	59.90
Ramachandran plot: Most favoured (%) Additionally allowed (%) Disallowed (%)	 94.86 3.75 1.39
PDB ID	6EGS

CAPÍTULO 6- DISCUSIÓN Y CONCLUSIONES

El hecho de que la gran mayoría de procesos de *O*-glicosilación en eucariotas superiores estén iniciados por una o dos glicosiltransferasas (Bennett *et al*. 2012; Joshi, Narimatsu, Schjoldager *et al*., 2018) contrasta enormente con el abrumador número de isoformas de ppGalNAc-Ts y sus, aparentemente, actividades redundantes. No obstante, en las dos últimas décadas se ha ido demostrando cómo el papel individual de las diferentes isoformas de estas *O*-GalNAc transferasas juega muchas veces un papel biológico fundamental, con enormes implicaciones tanto a nivel fisiológico como patológico (Kato *et al*., 2006; Topaz *et al*., 2004; Khetarpal *et al*., 2016; Varki *et al*., 2017; Chia *et al*., 2016, entre otros). El estudio de los mecanismos moleculares que justifican tanto las especificidades individuales de cada isoforma como su capacidad para trabajar de forma orquestada a la hora de glicosilar grandes dominios como los de tipo mucina (Chugh *et al*., 2015; Bennett *et al*., 2012; Varki *et al*., 2017; Hurtado-Guerrero, 2016) son imprescindibles si se pretende descifrar la función, relevancia y características de esta familia de glicosiltransferasas, así como desarrollar herramientas para la posible modulación de su actividad.

No es sino en este contexto en el que, en el capítulo Resultados, se han presentado los cinco trabajos que conforman esta tesis doctoral, con los que se ha logrado ahondar en algunos de los principales mecanismos estructurales y dinámicos que subyacen detrás de las preferencias de glicosilación de diferentes tipos de ppGalNAc-Ts (Bennett *et al*., 2012; Revoredo *et al*., 2016), así como en la justificación de la inactividad de un mutante patogénico de una de ellas (Khetarpal *et al*., 2016) y en el desarrollo de posibles moduladores de su actividad.

Si se compara con los conocimientos que sobre esta familia de glicosil transferasas existían antes del inicio de esta tesis doctoral, el conjunto de aportaciones que se pueden extraer de estos trabajos es destacadamente significativa. Así, en un primer acercamiento (Lira-Navarrete, de las Rivas *et al*., 2015), cuyos resultados ya formaron parte de la defensa de la tesis doctoral de su primera autora (Lira-Navarrete, 2015), se logró entender por primera vez cómo el dominio de lectina es capaz de dirigir la catálisis a sitios aceptores remotos y se reveló la destacada flexibilidad y dinamismo estructural de la isoforma *Hs*GalNAc-T2, así como la directa implicación del *flexible linker* en la movilidad de la misma. Las observaciones realizadas en este trabajo se complementaron ampliamente con las realizadas en de las Rivas *et al*. (2017), donde no sólo se determinó cuál era el modelo estructural que justificaba las distintas preferencias de glicosilación entre las isoformas que glicosilaban péptidos previamente glicosilados, sino que además se demostró cómo estas preferencias venían determinadas por los escasos residuos que conformaban el *flexible linker*, responsable del movimiento rotacional del dominio de lectina respecto al catalítico.

En lo tocante al desarrollo e identificación de inhibidores capaces de modular la actividad de diversas ppGalNAc-Ts, los dos estudios realizados han permitido conocer de forma más cercana las características del mecanismo de reconocimiento que la *Hs*GalNAc-T2 realiza sobre los diversos sustratos de la reacción de glicosilación que cataliza. Así, en Ghirardello, de las Rivas *et al*. (2016) se profundizó en el desarrollo de inhibidores lo suficientemente apolares como para poder entrar en la célula y que fuesen además capaces de interaccionar competitivamente con el UDP-GalNAc, para lo que a su vez fue fundamental el conocimiento pormenorizado previo acerca de las interacciones mayoritarias que tenían lugar en dicha región (Fritz *et al*., 2006; Lira-Navarrete, de las Rivas *et al*., 2015). Gracias a este trabajo se evidenció la importancia de que los compuestos análogos a nucleótidos que se diseñaron como posibles inhibidores mantuviesen un grupo fosfato en posición β, lo que confiere particular estabilidad a la unión con el catión Mn^{2+} y potencia una unión algo más fuerte enzima-compuesto. Tanto este descubrimiento como el estudio de la dinámica del *flexible loop* del dominio catalítico y su transición entre los estados activo e inactivo de la *Hs*GalNAc-T2 (Lira-Navarrete *et al*., 2014) constituyen una piedra angular para el entendimiento completo del mecanismo del ciclo catalítico de la enzima, así como para el desarrollo de futuras generaciones de inhibidores menos polares y que compitan de forma eficaz con UDP-GalNAc.

El trabajo sobre la inhibición de la *Hs*GalNAc-T2 por la luteolina (Liu, Xu, Xu, de las Rivas *et al*., 2017) nos ofrece un modelo distinto de inhibición, en el que el pequeño flavonoide inhibe de forma competitiva esta enzima al competir con los posibles péptidos aceptores. Si bien la extrapolación de los estudios de inhibición de este compuesto a sistemas *in vivo* ha permitido verificar cómo la inhibición de ciertos tipos de ppGalNAc-Ts parece tener lugar de forma selectiva, uno de los mayores inconvenientes de la luteolina es la presencia de numerosos estudios previos que habían demostrado su eficacia a la hora de inhibir otras moléculas tales como tirosín-kinasas, glucosidasas, etc (Harris *et al*., 2006; Kim *et al*., 2000; Huang *et al*., 1999). Esta falta de selectividad podría suponer indeseados efectos secundarios a nivel clínico, si bien no resta la importancia derivada del estudio estructural de su interacción con el dominio de unión a péptidos de la *Hs*GalNAc-T2. El detallado análisis estructural y dinámico de las interacciones entre la enzima y este compuesto permite identificar los elementos estructurales clave que hacen las veces de punto de partida para el desarrollo de inhibidores más específicos y que al competir, además, por el péptido de unión y no por el UDP-GalNAc (Ghirardello, de las Rivas *et al*., 2016), el más conservado entre las diferentes isoformas, posibilitarían una inhibición más selectiva entre estas últimas.

El estudio en el que, por último, se han desentrañado las bases moleculares subyacentes tras la inactividad del mutante Phe104Ser de la *Hs*GalNAc-T2, directamente asociado a un severo trastorno del metabolismo lipídico (Khetarpal

et al., 2016), nos ha permitido evidenciar el delicado equilibrio estructural de esta familia de enzimas. De igual modo, a través del estudio del mutante inactivo hemos podido corroborar el mecanismo catalítico bi-secuencial propuesto previamente por nuestro grupo (Lira-Navarrete *et al*., 2014), si bien hemos puntualizado cómo el equilibrio entre la forma activa e inactiva, determinado por las conformaciones respectivamente cerrada y abierta del *flexible loop* del dominio catalítico, no existe en ausencia de UDP-GalNAc. Cuando éste no está presente, la enzima adopta una conformación abierta para el *flexible loop* y está permanentemente inactiva. Solamente en presencia de UDP-GalNAc se induce el cambio a la conformación cerrada y, por tanto, activa y catalítica, del *flexible loop* y de la enzima.

Estudios de este tipo permiten evidenciar la importancia del conocimiento estructural y mecanístico a nivel molecular de las diversas enzimas como punto de partida imprescindible para la predicción y modulación racional de su comportamiento y actividad. En el caso concreto de las ppGalNAc-Ts, su sutil y heterogéneo perfil de actividad, los descubrimientos acerca de su implicación individual en el desarrollo de diversas patologías y su asociación directa en procesos tumorales y metastásicos hacen que siga siendo de relevante importancia el profundizar en el conocimiento integrado y multidisciplinar de sus mecanismos de acción, así como de su capacidad para orquestar de forma tan eficiente el entramado biológico que supone la *O*-glicosilación de tipo GalNAc.

Y no ha sido sino en este contexto en el que se ha desarrollado la presente tesis doctoral, de la que se pueden extraer las siguientes conclusiones, divididas según los correspondientes artículos:

Artículo I: Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein *O*-glycosylation.

- Se ha resuelto la estructura cristalográfica de complejos ternarios formados por la enzima *Hs*GalNAc-T2 y diversos péptidos glicosilados. Gracias a ello, se ha elucidado el papel del dominio de lectina de las ppGalNAc-Ts en la dirección de la glicosilación hacia sitios aceptores alejados de las glicosilaciones previas.
- Experimentos de SAXS y AFM han permitido evidenciar el gran dinamismo conformacional de la enzima *Hs*GalNAc-T2, que simulaciones computacionales han asociado con la flexibilidad y movilidad del *flexible linker* entre los dominios catalítico y de lectina.

Artículo II: The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences.

- Se ha resuelto la estructura cristalográfica del complejo ternario formado por la *Hs*GalNAc-T4, UDP/Mn^{2+} y un monoglicopéptido.
- La resolución de esta estructura cristalográfica y la realización de experimentos de RMN por diferencia de transferencia de saturación (STD) han permitido establecer un modelo, basado en la orientación de la subunidad α del dominio de lectina, que justifica la preferencia de la *Hs*GalNAc-T4 por glicosilar péptidos previamente glicosilados en el extremo N-terminal y no C-terminal, como era el caso de la *Hs*GalNAc-T2.
- Experimentos de docking y dinámica molecular han postulado que el *linker flexible* es el responsable de la diferente orientación del dominio de lectina y, por tanto, de las distintas preferencias de glicosilación entre este tipo de isoformas.
- Se han diseñado, expresado y purificado una serie de quimeras en las que se intercambia el *flexible linker* de la *Hs*GalNAc-T2 por el de la *Hs*GalNAc-T3 y *Hs*GalNAc-T4.
- Se han diseñado, expresado y purificado mutantes del *flexible linker* que ponen de manifiesto cómo pequeños cambios en el mismo conllevan variaciones en la preferencia de glicosilación remota de la *Hs*GalNAc-T2.
- La caracterización cinética de las quimeras y de los mutantes del *flexible linker* ha demostrado cómo, en efecto, las preferencias de glicosilación remota de las ppGalNAc-Ts vienen determinadas por la orientación de su dominio de lectina, establecida a su vez por el *flexible linker*.

Artículo III: Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-Polar Conjugates.

- Se han diseñado y sintetizado una serie de compuestos glicomiméticos con características menos polares que los de generaciones previas y destinados a unirse al dominio de unión a nucleótidos de las ppGalNAc-Ts. Para ello, se ha sustituido uno de los dos grupos fosfato por una cadena alquílica.
- Se han determinado las constantes de disociación entre los compuestos diseñados y la *Hs*GalNAc-T2 en presencia y ausencia de Mn^{2+} y se ha visto que se unen en el rango micromolar. Se ha comprobado que el compuesto derivado de galactosa es el que mejor se une a la enzima, si bien la interacción con el cofactor no parece determinante.
- Se ha logrado resolver la estructura cristalográfica de la *Hs*GalNAc-T2, previamente cocristalizada con UDP y Mn^{2+}, en complejo con el compuesto derivado de galactosa.

- Se ha observado que la sustitución del grupo fosfato en β respecto al carbono anomérico del residuo azucarado del compuesto glicomimético conlleva una interacción menos estable de lo deseado con el Mn^{2+}, por lo que los futuros compuestos habrán de diseñarse con sustituciones en la posición α.
- La estructura cristalográfica resuelta ha permitido ver que el compuesto diseñado atrapa la forma inactiva de la enzima *Hs*GalNAc-T2, con el *flexible loop* del dominio catalítico en conformación abierta. Experimentos de docking y de dinámica molecular han permitido caracterizar la transición entre la forma inactiva y activa de la enzima.

Artículo IV: The small molecule luteolin inhibits N-acetyl-alpha-galactosaminyltransferases and reduces mucin-type *O*-glycosylation of amyloid precursor protein.

- Se ha demostrado *in vitro* e *in vivo* que la luteolina inhibe de forma competitiva la unión entre el péptido sustrato y diversas isoformas de las ppGalNAc-Ts. Se ha comprobado, además, que esta inhibición es selectiva para la *Hs*GalNAc-T2 cuando se compara con otras isoformas ubicuas, como la *Hs*GalNAc-T1.
- Mediante la adición de luteolina se ha conseguido reducir la *O*-GalNAc glicosilación de los péptidos precursores del péptido amiloide y, con ella, el depósito de péptido β-amiloide en las líneas celulares estudiadas.
- Se ha logrado cristalizar el complejo formado por la *Hs*GalNAc-T2 y la luteolina y caracterizar las interacciones que demuestran que la inhibición tiene lugar en el sitio de unión del péptido.
- Pese a su inhibición de otras dianas terapéuticas, se ha comprobado cómo la luteolina es capaz de inhibir de forma más selectiva la *O*-GalNAc glicosilación llevada a cabo por las ppGalNAc-Ts y se ha propuesto como compuesto de partida para el desarrollo de futuros inhibidores que combatan enfermedades como el Alzheimer.

Artículo V: Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts.

- Se ha expresado y purificado el mutante inactivo y patogénico de la *Hs*GalNAc-T2, Phe104Ser.
- Se ha comprobado por desnaturalización térmica cómo el mutante es mucho más inestable que la enzima silvestre.
- Se ha caracterizado cinéticamente y por experimentos de RMN de STD la unión del mutante a sus sustratos y se ha observado que la

inactividad del mismo viene dada por su incapacidad total para unirse a péptidos, si bien sí se produce su unión a UDP-GalNAc.

- Se han obtenido cristales del mutante en complejo con UDP-GalNAc y se ha comprobado cómo el *flexible loop* del dominio catalítico se encuentra altamente desordenado y en su conformación inactiva (abierta). Se ha observado también que la presencia de la Phe104 en la enzima silvestre es un elemento clave para el establecimiento de interacciones de tipo CH-π con la Arg362, las cuales permiten fijar la conformación activa (cerrada) del *flexible loop*.
- Experimentos de dinámica molecular han permitido determinar que en el mutante Phe104Ser el *flexible loop* es mucho más móvil y desordenado que en la forma silvestre de la enzima, sin llegar a alcanzar nunca la conformación activa.
- Se han expresado mutantes activos de la enzima silvestre y del mutante Phe104Ser con una cisteína expuesta al solvente y situada en un lugar susceptible a la dinámica del *flexible loop*.
- Se ha marcado esta cisteína de los dos mutantes de la proteína silvestre y su variación patogénica con el fluoróforo 3-bromo-1, 1, 1-trifluoroacetona, gracias al cual se han realizado experimentos de RMN que han corroborado que el mutante patogénico no es capaz de adquirir la conformación cerrada del *flexible loop*.
- Estos mismos experimentos de RMN con los mutantes fluorados han corroborado el mecanismo bi-bi secuencial propuesto para el ciclo catalítico de la *Hs*GalNAc-T2 (Lira-Navarrete *et al*., 2014), si bien han demostrado también que la conformación cerrada de la enzima se induce al unirse ésta a UDP-GalNAc. A partir de la unión de éste, el *flexible loop* del dominio catalítico es capaz de pasar a su conformación cerrada (activa), y la enzima puede reconocer y unir a los posibles péptidos sustrato.

CAPÍTULO 7- PERSPECTIVAS

Como se ha ido viendo a lo largo de esta memoria, la realización de esta tesis doctoral ha permitido profundizar en algunos de los aspectos más fundamentales de la caracterización estructural y mecanísitica de la gran familia de las ppGalNAc-Ts. A través de los trabajos que la componen nos hemos adentrado en el conocimiento y justificación de las hasta ahora desconocidas preferencias de glicosilación de distintas isoformas, así como en las características estructurales que puedan servir de base para el desarrollo de futuros inhibidores específicos. Así mismo, el estudio del sutil mecanismo por el que un mutante aparentemente inocuo está directamente asociado a patogenicidad, ha evidenciado una vez más el intrincado equilibrio estructural y dinámico de la familia más amplia de glicosiltransferasas.

No obstante, son todavía numerosas las preguntas que quedan por responder, y es en ellas donde se centra el trabajo que ha continuado y continuará el mostrado en esta tesis. En primer lugar, y una vez desentrañados los mecanismos moleculares de la glicosilación de glicopéptidos dependiente del dominio de lectina o de largo alcance, quedan por desvelar los relacionados con la glicosilación de glicopéptidos de corto alcance, o dependiente del dominio catalítico. Este modelo de glicosilación de las ppGalNAc-Ts es, hasta ahora, el más desconocido, y para su comprensión se ha concluido en nuestro grupo un pormenorizado estudio cinético de la GalNAc-T4, que es una de las isoformas que muestran este tipo de reconocimiento. Aún en estado de escritura, este trabajo completa el conocimiento acerca de los diversos mecanismos de reconocimiento de sustratos de las distintas isoformas de ppGalNac-Ts.

Por otro lado, otra de las grandes incógnitas de esta familia de glicosiltransferasas es la *Hs*GalNAc-T3, la primera isoforma cuyo mal funcionamiento al no glicosilar la Thr178 del FGF23 se asoció directamente a una desregulación del metabolismo del fosfato y al desarrollo de una severa enfermedad, la calcinosis tumoral (Topaz et al., 2004; Duncan et al., 2011; Kato et al., 2006; Rafaelsen et al., 2014; Folsom e Imel, 2015). Es por ello por lo que los esfuerzos de nuestros grupo se han centrado también en el estudio estructural y molecular de esta isoforma en complejo con sustratos glicopeptídicos, esfuerzos que esperamos tomen forma de publicación a lo largo del año 2019.

De forma similar, está en marcha un proyecto acerca de las bases moleculares de la interacción de la ppGalNAc-T11 con los receptores de la lipoproteína de baja densidad (LDL; Wang, Mao *et al*., 2018), así como estudios que determinen las bases moleculares de una de las isoformas perteneciente al subgrupo de ppGalNAc-Ts IIb (Revoredo *et al*., 2016), aquéllas que únicamente reconocen, sin excepciones, sustratos glicopeptídicos, como es el caso de la GalNAc-T7. Si todos estos trabajos concluyen, serán pocas las incógnitas que queden sin responder acerca de las bases moleculares y estructurales del

reconocimiento de los sustratos por parte de las distintas isoformas de las ppGalNAc-Ts. Ahora bien, quedará por descifrar de qué forma se orquestan en la glicosilación secuencial y conjunta de dominios altamente glicosilados, proyecto en el que también formamos parte junto a nuestros colaboradores de Cohimbra, encargados de los experimentos de STD. Así mismo, otro de los proyectos principales con los que se espera continuar el trabajo recogido en esta tesis es el de la caracterización estructural y cinética de la enzima C1GalT1, encargada del siguiente paso de la *O*-glicosilación en eucariotas superiores y de cuyo mecanismo de acción y bases moleculares queda todavía mucho por descubrir (Varki *et al*., 2017; Chia *et al*., 2016).

En cualquier caso, llegar a entender el conjunto de glicanos de tipo *O*-GalNAc que se generan en el glicoproteoma humano, identificar los miembros individuales de los mismos que se asocian con enfermedades y caracterizar poco a poco los diversos estadíos del complicado entramado enzimático que completa la ruta de *O*-glicosilación de tipo mucina supone un largo y trabajoso camino que, no obstante, continuará sin duda revelando nuevas e importantes dianas terapeúticas, así como herramientas de diagnóstico, tratamiento y prevención de indudable utilidad: todo ello justifica sobradamente el esfuerzo e inversión humana y material depositado en trabajos como el recogido en esta tesis doctoral.

CAPÍTULO 8- REFERENCIAS BIBLIOGRÁFICAS

Agata, N., Ahmad, R., Kawano, T., Raina, D., Kharbanda, S. & Kufe, D. (2008). MUC1 oncoprotein blocks death receptor–mediated apoptosis by inhibiting recruitment of caspase-8. *Cancer research*, *68*(15), 6136-6144.

Ahmad, M., Hirz, M., Pichler, H. & Schwab, H. (2014). Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production. *Applied microbiology and biotechnology*, *98*(12), 5301-5317.

Angulo, J., & Nieto, P. M. (2011, October). Espectroscopía RMN de Diferencia de Transferencia de Saturación (STD NMR). Aplicaciones en Química del Reconocimiento Molecular. *Anales de Química*, 107(4), 358-366.

Apweiler, R., Hermjakob, H. & Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database1. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1473*(1), 4-8.

Bafna, S., Kaur, S. & Batra, S. K. (2010). Membrane-bound mucins: the mechanistic basis for alterations in the growth and survival of cancer cells. *Oncogene*, *29*(20), 2893-2904.

Bennett, E. P., Hassan, H., Mandel, U., Mirgorodskaya, E., Roepstorff, P., Burchell, J., Taylor-Papadimitriou, J., hollingsworth, M. A., Merkx, G., van Kessel, A. G., Eiberg, H., Steffensen, R. & Clausen, H. (1998). Cloning of a human UDP-*N*-acetyl-α-D-galactosamine: polypeptide*N*-acetylgalactosaminyltransferase that complements other GalNAc-transferases in complete *O*-glycosylation of the MUC1 tandem repeat. *Journal of Biological Chemistry*, *273*(46), 30472-30481.

Bennet, E. P., Mandel, U., Clausen, H., Gerken, T. A., Fritz, T. A. & Tabak, L. A. (2012). Control of mucin-type *O*-glysocylation: a classification of the polypeptide GalANc-transferase gene family. *Glycobiology 22(*6), pp. 736-76.

Bergstrom, K., Liu, X., Zhao, Y., Gao, N., Wu, Q., Song, K., Cui, Y., Li, Y., McDaniel, M-, McGee, S., Chen, W., Huycke, M. M., Houchen, C. W., Zenewick, L A., West, C. M., Chen, D., Braun, J., Fu, J., & Xia L. (2016). Defective intestinal mucin-type *O*-glycosylation causes spontaneous colitis-associated cancer in mice. *Gastroenterology*, *151*(1), 152-164.

Boskovski, M. T., Yuan, S., Pedersen, N. B., Goth, C. K., Makova, S., Clausen, H., Brueckner & Khokha, M. K. (2013). The heterotaxy gene GALNT11 glycosylates Notch to orchestrate cilia type and laterality. *Nature*, *504*(7480), 456.

Breton, C., Fournel-Gigleux, S., & Palcic, M. M. (2012). Recent structures, evolution and mechanisms of glycosyltransferases. *Current opinion in structural biology*, *22*(5), 540-549.

Brewer, C., Haugstad, K., Gerken, T., Stokke, B., & Sletmoen, M. (2015). Enhanced Self-Association of Mucins Possessing the Tn and Sialyl Tn Carbohydrate Cancer Antigens. *The FASEB Journal*, *29*(1 Supplement), 890-17.

Brooks, S. A. (2009). Strategies for analysis of the glycosylation of proteins: current status and future perspectives. *Molecular biotechnology*, *43*(1), 76.

Brooks, S. A., Carter, T. M., Bennett, E. P., Clausen, H., & Mandel, U. (2007). Immunolocalisation of members of the polypeptide *N*-acetylgalactosaminyl transferase (ppGalNAc-T) family is consistent with biologically relevant altered cell surface glycosylation in breast cancer. *Acta histochemica*, *109*(4), 273-284

Chaturvedi, P., Singh, A. P., Chakraborty, S., Chauhan, S. C., Bafna, S., Meza, J. L., Hollingsworth, M. A., Mehta, P. P. & Batra, S. K. (2008). MUC4 mucin interacts with and stabilizes the HER2 oncoprotein in human pancreatic cancer cells. *Cancer research*, *68*(7), 2065-2070.

Cheng, L., Tachibana, K., Zhang, Y., Guo, J. M., Kahori Tachibana, K., Kameyama, A., Wang, H., Hiruma, T., Iwasaki, H., Togayachi, A., Kudo, T., & Narimatsu, H. (2002). Characterization of a novel human UDP-GalNAc transferase, pp-GalNAc-T101. *FEBS letters*, *531*(2), 115-121.

Chia, J., Goh, G., & Bard, F. (2016). Short *O*-GalNAc glycans: regulation and role in tumor development and clinical perspectives. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1860*(8), 1623-1639.

Chugh, S., Gnanapragassam, V. S., Jain, M., Rachagani, S., Ponnusamy, M. P. & Batra, S. K. (2015). Pathobiological implications of mucin glycans in cancer: Sweet poison and novel targets. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, *1856*(2), 211-225.

Clarke, A. J., Hurtado-Guerrero, R., Pathak, S., Schüttelkopf, A. W., Borodkin, V., Shepherd, S. M., Ibrahim, A. F. M. & Van Aalten, D. M. (2008). Structural insights into mechanism and specificity of O-GlcNAc transferase. *The EMBO journal*, *27*(20), 2780-2788.

Clausen, H., & Bennett, E. P. (1996). A family of UDP-GalNAc: polypeptide *N*-acetylgalactosaminyl-transferases control the initiation of mucin-type *O*-linked glycosylation. *Glycobiology*, *6*(6), 635-646.

Creighton, T. E. (Ed.). (1997). *Protein structure: a practical approach*. (No. 174). Oxford university press.

Cuatrecasas, P. (1970). Protein purification by affinity chromatography derivatizations of agarose and polyacrylamide beads. *Journal of Biological Chemistry*, *245*(12), 3059-3065.

Dam, T. K., Gerken, T. A., & Brewer, C. F. (2009). Thermodynamics of multivalent carbohydrate−lectin cross-linking interactions: Importance of entropy in the bind and jump mechanism. *Biochemistry*, *48*(18), 3822-3827.

Dam, T. K., Gerken, T. A., Cavada, B. S., Nascimento, K. S., Moura, T. R., & Brewer, C. F. (2007). Binding studies of α-GalNAc-specific lectins to the α-GalNAc (Tn-antigen) form of porcine submaxillary mucin and its smaller fragments. *Journal of Biological Chemistry*, *282*(38), 28256-28263.

Damasceno, L. M., Huang, C. J. & Batt, C. A. (2012). Protein secretion in Pichia pastoris and advances in protein production. *Applied microbiology and biotechnology*, *93*(1), 31-39.

Davies, G. J. (2001). Sweet secrets of synthesis. *Nature Structural and Molecular Biology*, *8*(2), 98.

Davies, D. R., & Segal, D. M. (1971). [25] Protein crystallization: micro techniques involving vapor diffusion. In *Methods in enzymology* (Vol. 22, pp. 266-269). Academic Press.

Deng, J., Wang, L., Chen, H., Li, L., Ma, Y., Ni, J. & Li, Y. (2013). The role of tumour-associated MUC1 in epithelial ovarian cancer metastasis and progression. *Cancer and Metastasis Reviews*, *32*(3-4), 535-551.

Dessau, M. A., & Modis, Y. (2011). Protein crystallization for X-ray crystallography. *Journal of visualized experiments: JoVE*, (47).

Doerr, A. (2015). Structural biology: Peering inside protein complexes with AFM. *Nature methods*, *12*(4), 284.

Duncan, E. L., Danoy, P., Kemp, J. P., Leo, P. J., McCloskey, E., Nicholson, G. C., ... & Brown, M. A. (2011). Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. *PLoS genetics*, *7*(4), e1001372.

Eichwald, E. (1865). Beiträge zur Chemie der gewebbildenden Substanzen und ihrer Abkömmlinge. I. Ueber das Mucin, besonders der Weinbergschnecke. *Ann. Chem. Pharm*, 134, pp. 177-211.

Emsley, P., & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, *60*(12), 2126-2132.

Errey, J. C., Lee, S. S., Gibson, R. P., Martinez Fleites, C., Barry, C. S., Jung, P. M., O'Sullivan, A. C., Davis, B. G. & Davies, G. J. (2010). Mechanistic insight into enzymatic glycosyl transfer with retention of configuration through analysis of glycomimetic inhibitors. *Angewandte Chemie*, *122*(7), 1256-1259.

Ferentz, A. E., & Wagner, G. (2000). NMR spectroscopy: a multifaceted approach to macromolecular structure. *Quarterly reviews of biophysics*, *33*(1), 29-65.

Festari, M. F., Trajtenberg, F., Berois, N., Pantano, S., Revoredo, L., Kong, Y., Solari-Saquieres, P., Narimatsu, Y., Freire, T., Bay, S., Robello, C., Bénard, J., Gerken, T., Clausen, H. & Osinaga, E. (2017). Revisiting the human polypeptide GalNAc-T1 and T13 paralogs. *Glycobiology*, *27*(2), 140-153.

Fishman, M., & Moyer, L. S. (1942). Electrophoresis of the Chlorophyll-Protein Complex. *Science*, *95*(2457), 128-129.

Folsom, L. J., & Imel, E. A. (2015). Hyperphosphatemic familial tumoral calcinosis: genetic models of deficient FGF23 action. *Current osteoporosis reports*, *13*(2), 78-87

Fritz, T. A., Hurley, J. H., Trinh, L. B., Shiloach, J., & Tabak, L. A. (2004). The beginnings of mucin biosynthesis: The crystal structure of UDP-

GalNAc: polypeptide α-*N*-acetylgalactosaminyltransferase-T1. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(43), 15307-15312.

Fritz, T. A., Raman, J., & Tabak, L. A. (2006). Dynamic association between the catalytic and lectin domains of human UDP-GalNAc: polypeptide α-*N*-acetylgalactosaminyltransferase-2. *Journal of Biological Chemistry*, *281*(13), 8613-8619.

Fu, C., Zhao, H., Wang, Y., Cai, H., Xiao, Y., Zeng, Y., & Chen, H. (2016). Tumor-associated antigens: Tn antigen, sTn antigen, and T antigen. *Hla*, *88*(6), 275-286.

Gerken, T. A., Butenhof, K. J., & Shogren, R. (1989). Effects of glycosylation on the conformation and dynamics of *O*-linked glycoproteins: carbon-13 NMR studies of ovine submaxillary mucin. *Biochemistry*, *28*(13), 5536-5543.

Gerken, T. A., Gilmore, M., & Zhang, J. (2002). Determination of the Site-specific Oligosaccharide Distribution of the *O*-Glycans Attached to the Porcine Submaxillary Mucin Tandem Repeat. *Journal of Biological Chemistry*, *277*(10), 7736-7751.

Gerken, T. A., Jamison, O., Perrine, C. L., Collette, J. C., Moinova, H., Ravi, L., Markowitz, S. D., Shen, W., Patel, H. & Tabak, L. A. (2011). Emerging paradigms for the initiation of mucin-type protein *O*-glycosylation by the polypeptide GalNAc transferase family of glycosyltransferases. *Journal of Biological Chemistry*, *286*(16), 14493-14507.

Gerken, T. A. (2004). Kinetic modeling confirms the biosynthesis of mucin core 1 (β-Gal (1− 3) α-GalNAc-*O*-Ser/Thr) *O*-glycan structures are modulated by neighboring glycosylation effects. *Biochemistry*, *43*(14), 4137-4142.

Gerken, T. A., Owens, C. L., & Pasumarthy, M. (1997). Determination of the Site-specific *O*-Glycosylation Pattern of the Porcine Submaxillary Mucin Tandem Repeat Glycopeptide. Model proposed for the polypeptide: GalNAc-transferase peptide binding site. *Journal of Biological Chemistry*, *272*(15), 9709-9719.

Gerken, T. A., Raman, J., Fritz, T. A., & Jamison, O. (2006). Identification of common and unique peptide substrate preferences for the UDP-GalNAc: polypeptide α-*N*-acetylgalactosaminyltransferases T1 and T2

derived from oriented random peptide substrates. *Journal of Biological Chemistry*, *281*(43), 32403-32416.

Gerken, T. A., Revoredo, L., Thome, J. J., Tabak, L. A., Vester-Christensen, M. B., Clausen, H., Gahlay, G. K., Jarvis, D. L., Johnson, R. W., Moniz, H. A. & Moremen, K. (2013). 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. *Journal of Biological Chemistry*, *288*(27), 19900-19914.

Gerken, T. A., Ten Hagen, K. G., & Jamison, O. (2008). Conservation of peptide acceptor preferences between Drosophila and mammalian polypeptide-GalNAc transferase ortholog pairs. *Glycobiology*, *18*(11), 861-870.

Gerken, T. A., Tep, C., & Rarick, J. (2004). Role of Peptide Sequence and Neighboring Residue Glycosylation on the Substrate Specificity of the Uridine 5'-Diphosphate− α-*N*-acetylgalactosamine: Polypeptide *N*-acetylgalactosaminyl Transferases T1 and T2: Kinetic Modeling of the Porcine and Canine Submaxillary Gland Mucin Tandem Repeats. *Biochemistry*, *43*(30), 9888-9900.

Gerken, T. A., Zhang, J., Levine, J., & Elhammer, A. (2002). Mucin core *O*-glycosylation is modulated by neighboring residue glycosylation status: Kinetic modeling of the site specific glycosylation of the apo-porcine submaxillary mucin tandem repeat by UDP-GalNAc: Polypeptide *N*-acetylgalactosaminyltransferases T1 and T2. *Journal of Biological Chemistry*.

Gill, D. J., Chia, J., Senewiratne, J., & Bard, F. (2010). Regulation of *O*-glycosylation through Golgi-to-ER relocation of initiation enzymes. *The Journal of cell biology*, *189*, 843-858.

Gill, D. J., Clausen, H., & Bard, F. (2011). Location, location, location: new insights into *O*-GalNAc protein glycosylation. *Trends in cell biology,* 21(3), 149-158.

Glatter, O., & Kratky, O. (1982). Editors. *Small-Angle X-ray Scattering*. New York: Academic Press.

Gleeson, M. A., White, C. E., Meininger, D. P. & Komives, E. A. (1998). Generation of protease-deficient strains and their use in heterologous protein expression. *Methods in Molecular Biology,103*, 81–94.

Gomes, J., Marcos, N. T., Berois, N., Osinaga, E., Magalhães, A., Pinto-de-Sousa, J., Almeida, R., Gärtner, F. & Reis, C. A. (2008). Expression of UDP-*N*-acetyl-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferase-6 in gastric mucosa, intestinal metaplasia, and gastric carcinoma. *Journal of Histochemistry & Cytochemistry*, *57*(1), 79-86.

Gómez, H., Rojas, R., Patel, D., Tabak, L. A., Lluch, J. M., & Masgrau, L. (2014). A computational and experimental study of *O*-glycosylation. Catalysis by human UDP-GalNAc polypeptide: GalNAc transferase-T2. *Organic & biomolecular chemistry*, *12*(17), 2645-2655.

Gubbels, J. A., Felder, M., Horibata, S., Belisle, J. A., Kapur, A., Holden, H., Petrie, S., Migneault, M., Rancourt, C., Connor, J. P. & Patankar, M. S. (2010). MUC16 provides immune protection by inhibiting synapse formation between NK and ovarian tumor cells. *Molecular cancer*, *9*(1), 11.

Hagen, F. K., Hazes, B., Raffo, R., & Tabak, L. A. (1999). Structure-Function Analysis of the UDP-N-acetyl-d-galactosamine: Polypeptide *N*-acetylgalactosaminyltransferase essential residues lie in a predicted active site cleft resembling a lactose repressor fold. *Journal of Biological Chemistry*, *274*(10), 6797-6803.

Halim, A., Brinkmalm, G., Rüetschi, U., Westman-Brinkmalm, A., Portelius, E., Zetterberg, H., Blennow, K., Larson, G. & Nilsson, J. (2011). Site-specific characterization of threonine, serine, and tyrosine glycosylations of amyloid precursor protein/amyloid β-peptides in human cerebrospinal fluid. *Proceedings of the National Academy of Sciences*, *108*(29), 11848-11853.

Hang, H. C. & Bertozzi, C. R. (2005). The chemistry and biology of mucin-type *O*-linked glycosylation. *Bioorganic & medicinal chemistry*, *13*(17), 5021-5034.

Hanisch, F. G., Reis, C. A., Clausen, H., & Paulsen, H. (2001). Evidence for glycosylation-dependent activities of polypeptide *N*-acetylgalactosaminyltransferases GalNAc-T2 and-T4 on mucin glycopeptides. *Glycobiology*, *11*(9), 731-740.

Hanukoglu, I. (2015). Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites. *Biochemistry and Molecular Biology*

Education, *43*(3), 206-209.

Harris, G. K., Qian, Y., Leonard, S. S., Sbarra, D. C., & Shi, X. (2006). Luteolin and chrysin differentially inhibit cyclooxygenase-2 expression and scavenge reactive oxygen species but similarly inhibit prostaglandin-E2 formation in RAW 264.7 cells. *The Journal of nutrition*, 136(6), 1517-1521.

Hassan, H., Bennett, E. P., Mandel, V., Hollingsworth, M. A., Clausen, H., Ernst, B., Hart, G & Sinay, P. (2000a). Control of Mucin-Type O-Glycosylation: O-Glycan Occupancy is Directed by Substrate Specificities of Polypeptide GalNAc-Transferases. *Carbohydrates in chemistry and biology*, 273-292.

Hassan, H., Reis, C. A., Bennett, E. P., Mirgorodskaya, E., Roepstorff, P., Hollingsworth, M. A., Burchell, J., Taylor-Papadimitriou, J. & Clausen, H. (2000b). The lectin domain of UDP-*N*-acetyl-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferase-T4 directs its glycopeptide specificities. *Journal of Biological Chemistry*, *275*(49), 38197-38205.

Haugstad, K. E., Gerken, T. A., Stokke, B. T., Dam, T. K., Brewer, C. F., & Sletmoen, M. (2012). Enhanced self-association of mucins possessing the T and Tn carbohydrate cancer antigens at the single-molecule level. *Biomacromolecules*, *13*(5), 1400-1409.

Haugstad, K. E., Hadjialirezaei, S., Stokke, B. T., Brewer, C. F., Gerken, T. A., Burchell, J., Picco, G. & Sletmoen, M. (2016). Interactions of mucins with the Tn or Sialyl Tn cancer antigens including MUC1 are due to GalNAc–GalNAc interactions. *Glycobiology*, *26*(12), 1338-1350.

Häuselmann, I. & Borsig, L. (2014). Altered tumor-cell glycosylation promotes metastasis. *Frontiers in oncology*, *4*, 28.

Hazes, B. (1996). The (QxW) 3 domain: a flexible lectin scaffold. *Protein Science*, *5*(8), 1490-1501.

Hinoda, Y., Ikematsu, Y., Horinochi, M., Sato, S., Yamamoto, K., Nakano, T., Fukui, M., Suehiro, Y., Hamanaka, Y., Nishikawa, Y., Hida, H., Waki, S., Oka, M., Imai, K. & Yonezawa, S. (2003). Increased expression of MUC1 in advanced pancreatic cancer. *Journal of gastroenterology*, *38*(12), 1162-1166.

Hollingsworth, M. A. & Swanson, B. J. (2004). Mucins in cancer: protection and control of the cell surface. *Nature Reviews Cancer*, *4*(1), 45-60.

Holt, G. D., Haltiwanger, R. S., Torres, C. R., & Hart, G. W. (1987). Erythrocytes contain cytoplasmic glycoproteins. *O*-linked GlcNAc on Band 4.1. *Journal of Biological Chemistry*, *262*(31), 14847-14850.

Huang, Y. T., Hwang, J. J., Lee, P. P., Ke, F. C., Huang, J. H., Huang, C. J. & Lee, M. T. (1999). Effects of luteolin and quercetin, inhibitors of tyrosine kinase, on cell growth and metastasis‐associated properties in A431 cells overexpressing epidermal growth factor receptor. *British journal of pharmacology*, 128(5), 999-1010.

Hurtado-Guerrero, R., & Davies, G. J. (2012). Recent structural and mechanistic insights into post-translational enzymatic glycosylation. *Current opinion in chemical biology*, *16*(5-6), 479-487.

Hurtado-Guerrero, R. (2016). Recent structural and mechanistic insights into protein *O*-GalNAc glycosylation. *Biochemical Society transactions*, *44*(1), 61-67.

Ichikawa, S., Sorenson, A. H., Austin, A. M., Mackenzie, D. S., Fritz, T. A., Moh, A., Hui, S. L. & Econs, M. J. (2009). Ablation of the Galnt3 gene leads to low-circulating intact fibroblast growth factor 23 (Fgf23) concentrations and hyperphosphatemia despite increased Fgf23 expression. *Endocrinology*, *150*(6), 2543-2550.

Invitrogen, by LifeTechnologies. (2010). EasySelect ™ Pichia Expression Kit. For Expression of Recombinant Proteins Using pPICZ and pPICZα in Pichia pastoris. User Manual. Recuperado de: https://tools.thermofisher.com/content/sfs/manuals/easyselect_man.pdf.

Imberty, A., Piller, V., Piller, F., & Breton, C. (1997). Fold recognition and molecular modeling of a lectin-like domain in UDP-GalNac: polypeptide *N*-acetylgalactosaminyltransferases. *Protein engineering*, *10*(12), 1353-1356.

Irimura, T., Denda, K., Iida, S. I., Takeuchi, H., & Kato, K. (1999). Diverse glycosylation of MUC1 and MUC2: potential significance in tumor immunity. *The Journal of Biochemistry*, *126*(6), 975-986.

Jentoft, N. (1990). Why are proteins *O*-glycosylated?. *Trends in biochemical sciences*, *15*(8), 291-294.

Joshi, H. J., Hansen, L., Narimatsu, Y., Freeze, H. H., Henrissat, B., Bennett, E., Wandall, H. H., Clausen, H. & Schjoldager, K. T. (2018). Glycosyltransferase genes that cause monogenic congenital disorders of glycosylation are distinct from glycosyltransferase genes associated with complex diseases. *Glycobiology*, *28*(5), 284-294.

Joshi, H. J., Narimatsu, Y., Schjoldager, K. T., Tytgat, H. L., Aebi, M., Clausen, H., & Halim, A. (2018). SnapShot: *O*-glycosylation pathways across kingdoms. *Cell*, *172*(3), 632-632.

Ju, T., Lanneau, G. S., Gautam, T., Wang, Y., Xia, B., Stowell, S. R., Willard, M. T., Wang, W., Xia, J. Y., Zuna, R. E., Laszik, Z., Benbrook, D. M., Hanigan, M. H. & Cummings, R. D. (2008). Human tumor antigens Tn and sialyl Tn arise from mutations in Cosmc. *Cancer research*, *68*(6), 1636-1646.

Julien, S., Krzewinski-Recchi, M. A., Harduin-Lepers, A., Gouyer, V., Huet, G., Le Bourhis, X., & Delannoy, P. (2001). Expression of sialyl-Tn antigen in breast cancer cells transfected with the human CMP-Neu5Ac: GalNAc α2, 6-sialyltransferase (ST6GalNAc I) cDNA. *Glycoconjugate journal*, *18*(11-12), 883-893.

Julenius, K., Mølgaard, A., Gupta, R., & Brunak, S. (2005). Prediction, conservation analysis, and structural characterization of mammalian mucin-type *O*-glycosylation sites. *Glycobiology*, *15*(2), 153-164.

Kabsch, W. (2010). Xds. *Acta Crystallographica Section D: Biological Crystallography*, *66*(2), 125-132.

Kato, K., Jeanneau, C., Tarp, M. A., Benet-Pagès, A., Lorenz-Depiereux, B., Bennett, E. P., Mandel, U., Strom, T. M. & Clausen, H. (2006). Polypeptide GalNAc-transferase T3 and familial tumoral calcinosis Secretion of fibroblast growth factor 23 requires *O*-glycosylation. *Journal of Biological Chemistry*, *281*(27), 18370-18377.

Kato, K., Takeuchi, H., Miyahara, N., Kanoh, A., Hassan, H., Clausen, H., & Irimura, T. (2001). Distinct orders of GalNAc incorporation into a peptide with consecutive threonines. *Biochemical and biophysical research communications*, *287*(1), 110-115.

Khetarpal, S. A., Schjoldager, K. T., Christoffersen, C., Raghavan, A., Edmondson, A. C., Reutter, H. M., Ahmed, B., Ouazzani, R., Peloso, G. M., Vitali, C, Zhao, W., Somasundara, A. V. H., Millar, J. S., Park, Y., Fernando, G., Livanov, V., Choi, S., Noé, E., Patel, P., Ho, S. P., Kirchgessner, t. G., Wandall, H. H., Hansen, L., Bennett, E. P., Vakhrushev, S., Saleheen, D., Kathiresan, S., Brown, C. D., Jamra, R. A., LeGuern, E., Clausen, H., & Rader, D. J. (2016). Loss of function of GALNT2 lowers high-density lipoproteins in humans, nonhuman primates, and rodents. *Cell metabolism*, *24*(2), 234-245.

Kim, J. S., Kwon, C. S., & SoN, K. H. (2000). Inhibition of alpha-glucosidase and amylase by luteolin, a flavonoid. *Bioscience, biotechnology, and biochemistry*, *64(11),* 2458-2461.

Kong, Y., Joshi, H. J., Schjoldager, K. T. B. G., Madsen, T. D., Gerken, T. A., Vester-Christensen, M. B., Wandall, H. H., Bennett, E. P., Levery, S. B., Vakhrushev, S. Y. & Clausen, H. (2014). Probing polypeptide GalNAc-transferase isoform substrate specificities by in vitro analysis. *Glycobiology*, *25*(1), 55-65.

Kubota, T., Shiba, T., Sugioka, S., Furukawa, S., Sawaki, H., Kato, R., Wakatsuki, S., & Narimatsu, H. (2006). Structural basis of carbohydrate transfer activity by human UDP-GalNAc: polypeptide α-*N*-acetylgalactosaminyltransferase (pp-GalNAc-T10). *Journal of molecular biology*, *359*(3), 708-727.

Kufe, D. W. (2013). MUC1-C oncoprotein as a target in breast cancer: activation of signaling pathways and therapeutic approaches. *Oncogene*, *32*(9), 1073-1081.

Kufe, D. W. (2009). Mucins in cancer: function, prognosis and therapy. *Nature Reviews Cancer*, *9*(12), 874.

La, I. J., Karim, M., & Yoon, M. Y. (2008). Structure and functional effect of tryptophan mutants of Nicotiana tabacum acetohydroxyacid synthase. *Bulletin of the Korean Chemical Society*, *29*(9), 1823-1826.

Lairson, L. L., Chiu, C. P., Ly, H. D., He, S., Wakarchuk, W. W., Strynadka, N. C., & Withers, S. G. (2004). Intermediate trapping on a mutant retaining α-galactosyltransferase identifies an unexpected aspartate residue. *Journal of Biological Chemistry*, *279*(27), 28339-28344.

Lairson, L. L., Henrissat, B., Davies, G. J., & Withers, S. G. (2008). Glycosyltransferases: structures, functions, and mechanisms. *Annual review of biochemistry*, *77*.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography*, *26*(2), 283-291.

Lee, S. S., Hong, S. Y., Errey, J. C., Izumi, A., Davies, G. J., & Davis, B. G. (2011). Mechanistic evidence for a front-side, SNi-type reaction in a retaining glycosyltransferase. *Nature chemical biology*, *7*(9), 631-638.

Lengauer, T., & Rarey, M. (1996). Computational methods for biomolecular docking. *Current opinion in structural biology*, *6*(3), 402-406.

Li, X., Wang, J., Li, W., Xu, Y., Shao, D., Xie, Y., Kubota, T., Narimatsu, H. & Zhang, Y. (2011). Characterization of ppGalNAc-T18, a member of the vertebrate-specific Y subfamily of UDP-*N*-acetyl-α-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferases. *Glycobiology*, *22*(5), 602-615.

Lira Navarrete, E. (2015). Estudio de los mecanismos moleculares de enzimas de unión a nucleótidos. (Tesis Doctoral). Universidad de Zaragoza, Zaragoza.

Lira-Navarrete, E., Iglesias-Fernández, J., Zandberg, W. F., Compañón, I., Kong, Y., Corzana, F., .Mario Pinto, B., Clausen, H., Peregrina, J. M., Vocadlo, D. J., Rovira, C. & Hurtado-Guerrero, R. (2014). Substrate-Guided Front-Face Reaction Revealed by Combined Structural Snapshots and Metadynamics for the Polypeptide N-Acetylgalactosaminyltransferase 2. *Angewandte Chemie International Edition*, *53*(31), 8206-8210

Lo, M. C., Aulabaugh, A., Jin, G., Cowling, R., Bard, J., Malamas, M., & Ellestad, G. (2004). Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Analytical biochemistry*, *332*(1), 153-159.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2013). The carbohydrate-active enzymes database (CAZy) in 2013.*Nucleic acids research*, *42*(D1), D490-D495.

Lorenz, V., Ditamo, Y., Cejas, R. B., Carrizo, M. E., Bennett, E. P., Clausen, H., Nores, G. A. & Irazoqui, F. J. (2016). Extrinsic functions of lectin domains in *O*-*N*-acetylgalactosamine glycan biosynthesis. *Journal of Biological Chemistry*, *291*(49), 25339-25350.

Lowe, J.B. and Marth, J.D. (2003) A genetic approach to mammalian glycan function. *Annual Review Biochemistry*, 72, 643–691

Mac Sweeney, A., & D'Arcy, A. (2003). A simple and rapid method for mounting protein crystals at room temperature. *Journal of applied crystallography*, *36*(1), 165-166.

Madariaga, D., Martínez‐Sáez, N., Somovilla, V. J., García‐García, L., Berbis, M. Á., Valero‐Gónzalez, J., Martín- Santamaría, S., Hurtado-Guerrero, R., Asensio, J. L., Jiménez-Barberó, J. J., Avenoza, A., Busto, J. H., Corzana, F. & Peregrina, J. M. (2014). Serine versus Threonine Glycosylation with α‐O‐GalNAc: Unexpected Selectivity in Their Molecular Recognition with Lectins. *Chemistry-A European Journal*, *20*(39), 12616-12627.

Madsen CB, Lavrsen K, Steentoft C, Vester-Christensen MB, Clausen H, Wandall HH, Pedersen AE. (2013). Glycan elongation beyond the mucin associated Tn antigen protects tumor cells from immune-mediated killing. PLoS. One. 2013; 8:e72413.

Mandel, U., Hassan, H., Therkildsen, M. H., Rygaard, J., Jakobsen, M. H., Juhl, B. R., Dabelsteen, E., & Clausen, H. (1999). 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.

Martínez Oliván, J. (2014) Mecanismo de incorporación celular de LDL y β-VLDL: análisis estructural y funcional de la región extracelular del receptor de LDL. (Tesis Doctoral). Universidad de Zaragoza, Zaragoza.

Martínez‐Sáez, N., Castro‐López, J., Valero‐González, J., Madariaga, D., Compañón, I., Somovilla, V. J., Salvadó, m., Asensio, J., L., Jiménez-Barberó, J. J., Avenoza, A., Busto, J. H., Bernadés, G. J., Peregrina, J. M., Hurtado-Guerrero, R., & Corzana, F. (2015). Deciphering the Non‐Equivalence of Serine and Threonine O‐Glycosylation Points:

Implications for Molecular Recognition of the Tn Antigen by an anti-MUC1 Antibody. *Angewandte Chemie International Edition*, *54*(34), 9830-9834.

Martínez-Sáez, N., Supekar, N. T., Wolfert, M. A., Bermejo, I. A., Hurtado-Guerrero, R., Asensio, J. L., Jiménez-Barbero, J., Busto, J. H., Avenoza, A., Booms, G. J., Peregrina, J. M. & Corzana, F. (2016). Mucin architecture behind the immune response: design, evaluation and conformational analysis of an antitumor vaccine derived from an unnatural MUC1 fragment. *Chemical Science*, *7*(3), 2294-2301.

Matsukita, S., Nomoto, M., Kitajima, S., Tanaka, S., Goto, M., Irimura, T., Kim, Y. S., Sato, E. & Yonezawa, S. (2003). Expression of mucins (MUC1, MUC2, MUC5AC and MUC6) in mucinous carcinoma of the breast: comparison with invasive ductal carcinoma. *Histopathology*, *42*(1), 26-36.

Miwa, H. E., Gerken, T. A., Jamison, O., & Tabak, L. A. (2010). Isoform-specific *O*-glycosylation of osteopontin and bone sialoprotein by polypeptide *N*-acetylgalactosaminyltransferase-1. *Journal of Biological Chemistry*, *285*(2), 1208-1219.

Mosi, R. M., & Withers, S. G. (2002). Trapping of alpha-glycosidase intermediates. *Methods in enzymology*, *354*, 64.

Mukhopadhyay, P., Lakshmanan, I., Ponnusamy, M. P., Chakraborty, S., Jain, M., Pai, P., Smith, L. M., Lele, S. M. & Batra, S. K. (2013). MUC4 overexpression augments cell migration and metastasis through EGFR family proteins in triple negative breast cancer cells. *PLoS One*, *8*(2), e54455.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, *67*(4), 355-367.

Myszka, D. G., & Rich, R. L. (2000). Implementing surface plasmon resonance biosensors in drug discovery. *Pharmaceutical science & technology today*, *3*(9), 310-317.

Noble, J. E., & Bailey, M. J. (2009). Quantitation of protein. *Methods in enzymology, 463,* 73-95.

Ohtsubo, K., & Marth, J. D. (2006). Glycosylation in cellular mechanisms of health and disease. *Cell*, *126*(5), 855-867.

Ozaki, H., Matsuzaki, H., Ando, H., Kaji, H., Nakanishi, H., Ikehara, Y., & Narimatsu, H. (2012). Enhancement of metastatic ability by ectopic expression of ST6GalNAcI on a gastric cancer cell line in a mouse model. *Clinical & experimental metastasis*, *29*(3), 229-238.

Padler-Karavani, V. (2014). Aiming at the sweet side of cancer: aberrant glycosylation as possible target for personalized-medicine. *Cancer letters*, *352*(1), 102-112.

Park, J. H., Katagiri, T., Chung, S., Kijima, K., & Nakamura, Y. (2011). Polypeptide *N*-acetylgalactosaminyltransferase 6 disrupts mammary acinar morphogenesis through *O*-glycosylation of fibronectin. *Neoplasia*, *13*(4), 320-IN10.

Párraga Martínez, J., Delgado Calvo-Flores, G., Delgado Calvo-Flores, R., Martín García, J. M., & García Ruiz, J. M. (2010). Práctica de cristalización de proteínas por técnicas de contradifusión en geles, para estudiantes de Farmacia. *ARS Pharmaceutica, 51(2*): 316-323.

Paulson, J. C., & Colley, K. J. (1989). Glycosyltransferases. Structure, localization, and control of cell type-specific glycosylation. *Journal of Biological Chemistry*, *264*(30), 17615-17618.

Pedersen, J. W., Bennett, E. P., Katrine, T. B. S., Meldal, M., Holmér, A. P., Blixt, O., Cló, E., Levery, S. B., Clausen, H. & Wandall, H. H. (2011). Lectin domains of polypeptide GalNAc transferases exhibit glycopeptide binding specificity. *Journal of Biological Chemistry*, *286*(37), 32684-32696.

Pedersen, N. B., Wang, S., Narimatsu, Y., Yang, Z., Halim, A., Schjoldager, K. T. B. G., Madsen, T. D., Seidah, N. G., Bennett, E. P., Levery, S. B. & Clausen, H. (2014). Low density lipoprotein receptor class A repeats are *O*-glycosylated in linker regions. *Journal of Biological Chemistry*, *289*(25), 17312-17324.

Pelaseyed, T., Zäch, M., Petersson, Å. C., Svensson, F., Johansson, D. G. & Hansson, G. C. (2013). Unfolding dynamics of the mucin SEA domain probed by force spectroscopy suggest that it acts as a cell-protective device. *The FEBS journal*, *280*(6), 1491-1501.

Peng, C., Togayachi, A., Kwon, Y. D., Xie, C., Wu, G., Zou, X., Sato, T., Ito, H., Tachibana, K., Kubota, T., Noce, T., Narimatsu, H. & Zhang, Y. (2010). Identification of a novel human UDP-GalNAc transferase with unique catalytic activity and expression profile. *Biochemical and biophysical research communications*, *402*(4), 680-686.

Perez-Vilar, J. & Hill, R. L. (1999). The structure and assembly of secreted mucins. *Journal of Biological Chemistry*, *274*(45), 31751-31754.

Perrine, C., Ju, T., Cummings, R. D., & Gerken, T. A. (2008). Systematic determination of the peptide acceptor preferences for the human UDP-Gal: glycoprotein-α-GalNAc β3 galactosyltranferase (T-synthase). *Glycobiology*, *19*(3), 321-328.

Perrine, C. L., Ganguli, A., Wu, P., Bertozzi, C. R., Fritz, T. A., Raman, J., Tabak, L. A. & Gerken, T. A. (2009). Glycopeptide-preferring polypeptide GalNAc transferase 10 (ppGalNAc T10), involved in mucin-type *O*-glycosylation, has a unique GalNAc-*O*-Ser/Thr-binding site in its catalytic domain not found in ppGalNAc T1 or T2. *Journal of Biological Chemistry*, *284*(30), 20387-20397.

Ponnusamy, M. P., Seshacharyulu, P., Vaz, A., Dey, P. & Batra, S. K. (2011). MUC4 stabilizes HER2 expression and maintains the cancer stem cell population in ovarian cancer cells. *Journal of ovarian research*, 4(1), 7.

Pratt, M. R., Hang, H. C., Ten Hagen, K. G., Rarick, J., Gerken, T. A., Tabak, L. A., & Bertozzi, C. R. (2004). Deconvoluting the functions of polypeptide *N*-α-acetylgalactosaminyltransferase family members by glycopeptide substrate profiling. *Chemistry & biology*, *11*(7), 1009-1016.

Prokop, O., & Uhlenbruck, G. (1969). *N*-acetyl-D-galactosamine in tumor cell membranes: demonstration by means of Helix agglutinins. *Die Medizinische Welt*, *46*, 2515.

Rachagani, S., Torres, M. P., Moniaux, N. & Batra, S. K. (2009). Current status of mucins in the diagnosis and therapy of cancer. *Biofactors*, *35*(6), 509-527.

Radhakrishnan, P., Dabelsteen, S., Madsen, F. B., Francavilla, C., Kopp, K. L., Steentoft, C., Vakhrushev, S. Y., Olsen, J. V., Hansen, L., Bennett, E. P., Woetmann, A., Yin, G., Chen, L., Song, H., Bak, M., Hlady, R. A., Peters, S. L., Opavski, R., Thode, C., Qvortrup, K., Schjoldager, K. T. B. G., Clausen, H., Hollingsworth, M. A. & Wandall, H. H. (2014).

Immature truncated *O*-glycophenotype of cancer directly induces oncogenic features. *Proceedings of the National Academy of Sciences*, *111*(39), E4066-E4075.

Rafaelsen, S., Johansson, S., Ræder, H., & Bjerknes, R. (2014). Long-term clinical outcome and phenotypic variability in hyperphosphatemic familial tumoral calcinosis and hyperphosphatemic hyperostosis syndrome caused by a novel GALNT3 mutation; case report and review of the literature. *BMC genetics*, *15*(1), 98.

Raman, J., Fritz, T. A., Gerken, T. A., Jamison, O., Live, D., Liu, M., & Tabak, L. A. (2008). The catalytic and lectin domains of UDP-GalNAc: polypeptide α-*N*-acetylgalactosaminyltransferase function in concert to direct glycosylation site selection. *Journal of Biological Chemistry*, *283*(34), 22942-22951.

Raman, J., Guan, Y., Perrine, C. L., Gerken, T. A., & Tabak, L. A. (2011). UDP-*N*-acetyl-α-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferases: completion of the family tree. *Glycobiology*, *22*(6), 768-777.

Rao, S. T., & Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *Journal of molecular biology*, *76*(2), 241-256.

Reddy, A., Grimwood, B. G., Plummer, T. H., & Tarentino, A. L. (1998). High-level expression of the endo-β-N-acetylglucosaminidase F2 gene in *E. coli*: one step purification to homogeneity. *Glycobiology*, *8*(6), 633-636.

Reis, C. A., David, L., Correa, P., Carneiro, F., de Bolós, C., Garcia, E., Mandel, U., Clausen, H. & Sobrinho-Simões, M. (1999). Intestinal metaplasia of human stomach displays distinct patterns of mucin (MUC1, MUC2, MUC5AC, and MUC6) expression. *Cancer Research*, *59*(5), 1003-1007.

Revoredo, L., Wang, S., Bennett, E. P., Clausen, H., Moremen, K. W., Jarvis, D. L., Ten Hagen, K. G., Tabak, L. A. & Gerken, T. A. (2015). 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*(4), 360-376.

Riniker, S., Allison, J. R., & van Gunsteren, W. F. (2012). On developing coarse-grained models for biomolecular simulation: a review. *Physical Chemistry Chemical Physics*, *14*(36), 12423-12430.

Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., & Xu, S. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic acids research*, *31*(7), 1805-1812.

Rottger, S., White, J., Wandall, H. H., Olivo, J. C., Stark, A., Bennett, E. P., Whitehouse, C., Berger, E. G., Clausen, H. & Nilsson, T. (1998). Localization of three human polypeptide GalNAc-transferases in HeLa cells suggests initiation of *O*-linked glycosylation throughout the Golgi apparatus. *Journal of cell science*, *111*(1), 45-60.

Ruan, H. B., Singh, J. P., Li, M. D., Wu, J., & Yang, X. (2013). Cracking the *O*-GlcNAc code in metabolism. *Trends in Endocrinology & Metabolism*, *24*(6), 301-309.

Rydzik, A. M., Brem, J., van Berkel, S. S., Pfeffer, I., Makena, A., Claridge, T. D., & Schofield, C. J. (2014). Monitoring Conformational Changes in the NDM-1 Metallo-β-lactamase by 19F NMR Spectroscopy. *Angewandte Chemie International Edition*, *53*(12), 3129-3133.

Sancho, J. (2013). The stability of 2-state, 3-state and more-state proteins from simple spectroscopic techniques... plus the structure of the equilibrium intermediates at the same time. *Archives of biochemistry and biophysics*, *531*(1-2), 4-13.

Semisotnov, G. V., Rodionova, N., Razgulyaev, O. I., Uversky, V. N., Gripas', A. F., & Gilmanshin, R. I. (1991). Study of the “molten globule” intermediate state in protein folding by a hydrophobic fluorescent probe. Biopolymers: Original Research on Biomolecules, 31(1), 119-128.

Schedin-Weiss, S., Winblad, B., & Tjernberg, L. O. (2014). The role of protein glycosylation in Alzheimer disease. *The FEBS journal*, *281*(1), 46-62.

Schjoldaher, K. T. B. G., & Clausen, H. (2012). Site-specific protein *O*-glycosylation modulates proprotein processing—deciphering specific functions of the large polypeptide GalNAc-transferase gene family. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1820*(12), 2079-2094.

Schjoldager, K. T. B., Joshi, H. J., Kong, Y., Goth, C. K., King, S. L., Wandall, H. H., Bennett, E. P., Vakhrushev, S. Y. & Clausen, H. (2015).

Deconstruction of *O*-glycosylation—GalNAc-T isoforms direct distinct subsets of the *O*-glycoproteome. *EMBO reports*, e201540796.

Schjoldager, K. T. B., Vakhrushev, S. Y., Kong, Y., Steentoft, C., Nudelman, A. S., Pedersen, N. B., Wandall, H. H., Mandell, U., Bennett, E. P., Levery, S. B. & Clausen, H. (2012). Probing isoform-specific functions of polypeptide GalNAc-transferases using zinc finger nuclease glycoengineered SimpleCells. *Proceedings of the National Academy of Sciences*, *109*(25), 9893-9898.

Schjoldager, K. T. B., Vester-Christensen, M. B., Bennett, E. P., Levery, S. B., Schwientek, T., Yin, W., Blixt, O. & Clausen, H. (2010). *O*-Glycosylation Modulates Proprotein Convertase Activation of Angiopoietin-like Protein 3. Possible role of polypeptide GalNAc-transferase-2 in regulation of concentrations of plasma lipids. *Journal of Biological Chemistry*, *285*(47), 36293-36303.

Schmitt, J., Hess, H., & Stunnenberg, H. G. (1993). Affinity purification of histidine-tagged proteins. *Molecular biology reports*, *18*(3), 223-230.

Schwientek, T., Bennett, E. P., Flores, C., Thacker, J., Hollmann, M., Reis, C. A., Behrens, J., Mande, U., Keck, B., Schäfer, M. A., Haselmann, K., Zubarev, R., Roepstroff, P., Burchell, J. M., Taylor-Papadimitriou, J., Hollingsworth, M. A. & Clausen, H. (2002). Functional Conservation of Subfamilies of Putative UDP-*N*-acetylgalactosamine: Polypeptide Acetylgalactosaminyltransferases in *Drosophila*, *Caenorhabditis elegans*, and Mammals. *Journal of Biological Chemistry*, *277*(25), 22623-22638.

Semisotnov, G. V., Rodionova, N., Razgulyaev, O. I., Uversky, V. N., Gripas', A. F., & Gilmanshin, R. I. (1991). Study of the “molten globule” intermediate state in protein folding by a hydrophobic fluorescent probe. *Biopolymers: Original Research on Biomolecules*, *31*(1), 119-128.

Shogren, R., Gerken, T. A., & Jentoft, N. (1989). Role of glycosylation on the conformation and chain dimensions of *O*-linked glycoproteins: light-scattering studies of ovine submaxillary mucin. *Biochemistry*, *28*(13), 5525-5536.

Sletmoen, M., Gerken, T. A., Stokke, B. T., Burchell, J., & Brewer, C. F. (2018). Tn and STn are members of a family of carbohydrate tumor antigens that possess carbohydrate–carbohydrate interactions. *Glycobiology*, *1*, 6.

Song, X., Airan, R. D., Arifin, D. R., Bar-Shir, A., Kadayakkara, D. K., Liu, G., Gilad, A. A., Van Zjil, P. C. M., McMahon, M. T. & Bulte, J. W. M. (2015). Label-free in vivo molecular imaging of underglycosylated mucin-1 expression in tumour cells. *Nature communications*, *6*, 6719.

Sørensen, T., White, T., Wandall, H. H., Kristensen, A. K., Roepstorff, P., & Clausen, H. (1995). UDP-*N*-acetyl-α-D-galactosamine: polypeptide *N*-Acetylgalactosaminyltransferase. Identification and separation of two distinct transferase activities. *Journal of Biological Chemistry*, *270*(41), 24166-24173.

Stanley, P. (2011). Golgi glycosylation. *Cold Spring Harbor perspectives in biology*, *3*(4), a005199.

Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T. B. G., Lavrsen, K., Dabelsteen, S., Pedersen, N. B., Marcos-Silva, L., Gupta, R., Bennett, E. P., Mandel, U., Brunak, S., Wandall, H. H., Levery, S. B. & (2013). Precision mapping of the human *O*-GalNAc glycoproteome through SimpleCell technology. *The EMBO journal*, *32*(10), 1478-1488.

Steentoft, C., Vakhrushev, S. Y., Vester-Christensen, M. B., Schjoldager, K. T. G., Kong, Y., Bennett, E. P., Mendel, U., Brunak, S., Wandall, H. H., Levery, S. B. & Clausen, H. (2011). Mining the *O*-glycoproteome using zinc-finger nuclease–glycoengineered SimpleCell lines. *Nature methods*, *8*(11), 977.

Stura, E. A., Nemerow, G. R., & Wilson, I. A. (1992). Strategies in the crystallization of glycoproteins and protein complexes. *Journal of crystal growth*, *122*(1-4), 273-285.

Suga, M., Isobe, M., & Hatakeyama, T. (2000). Cryopreservation of competent intact yeast cells for efficient electroporation. *Yeast*, *16*(10), 889-896.

Sugiura, M., Kawasaki, T., & Yamashina, I. (1982). Purification and characterization of UDP-GalNAc: polypeptide *N*-acetylgalactosamine transferase from an ascites hepatoma, AH 66. *Journal of Biological Chemistry*, *257*(16), 9501-9507.

Supraha Goreta, S., Dabelic, S., & Dumic, J. (2012). Insights into complexity of congenital disorders of glycosylation. *Biochemia medica: Biochemia medica*, *22*(2), 156-170.

Tabak, L. A. (2010, August). The role of mucin-type *O*-glycans in eukaryotic development. In *Seminars in cell & developmental biology* (Vol. 21, No. 6, pp. 616-621). Academic Press.

Tarp, M. A., & Clausen, H. (2008). Mucin-type *O*-glycosylation and its potential use in drug and vaccine development. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1780*(3), 546-563.

Taylor-Papadimitriou, J., Burchell, J., Miles, D. W. & Dalziel, M. (1999). MUC1 and cancer. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1455*(2-3), 301-313.

Ten Hagen, K. G., Fritz, T. A. & Tabak, L. A. (2003). All in the family: the UDP-GalNAc: polypeptide *N*-acetylgalactosaminyltransferases. *Glycobiology*, *13*(1), 1R-16R.

Ten Hagen, K. G., Tran, D. T., Gerken, T. A., Stein, D. S., & Zhang, Z. (2003). Functional characterization and expression analysis of members of the UDP-GalNAc: polypeptide *N*-acetylgalactosaminyltransferase family from Drosophila melanogaster. *Journal of Biological Chemistry*, *278*(37), 35039-35048.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., … & Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, *466*(7307), 707.

Tetaert, D., Richet, C., Gagnon, J., Boersma, A., & Degand, P. (2001). Studies of acceptor site specificities for three members of UDP-GalNAc: *N*-acetylgalactosaminyltransferases by using a synthetic peptide mimicking the tandem repeat of MUC5AC. *Carbohydrate research*, *333*(2), 165-171.

Thériault, C., Pinard, M., Comamala, M., Migneault, M., Beaudin, J., Matte, I., Boivin, M., Piché, A. & Rancourt, C. (2011). MUC16 (CA125) regulates epithelial ovarian cancer cell growth, tumorigenesis and metastasis. Gynecologic oncology, 121(3), 434-443.Campbell, B. J., Finnie, I. A., Hounsell, E. F. & Rhodes, J. M. (1995). Direct demonstration of increased expression of Thomsen-Friedenreich (TF) antigen in colonic adenocarcinoma and ulcerative colitis mucin and its concealment in normal mucin. *The Journal of clinical investigation*, *95*(2), 571-576.

Tian, E., & Ten Hagen, K. G. (2007). A UDP-GalNAc: polypeptide *N*-acetylgalactosaminyltransferase is required for epithelial tube formation. *Journal of Biological Chemistry*, *282*(1), 606-614.

Topaz, O., Shurman, D. L., Bergman, R., Indelman, M., Ratajczak, P., Mizrachi, M., Khamaysi, Z., Behar, D., Petronius, D., Friedman, V., Zelikovic, I., Raimer, S., Metzker, A., Richard, G. & Sprecher, E. (2004). Mutations in GALNT3, encoding a protein involved in *O*-linked glycosylation, cause familial tumoral calcinosis. *Nature genetics*, *36*(6), 579.

Tózsér, J. (Ed.). (2011). *Protein Biotechnology.* Debrecen, Hungría: University of Debrecen.

Tran, D. T., & Ten Hagen, K. G. (2013). Mucin-type *O*-glycosylation during development. *Journal of Biological Chemistry*, *288*(10), 6921-6929.

Vagin, A., & Teplyakov, A. (2010). Molecular replacement with MOLREP. *Acta Crystallographica Section D: Biological Crystallography*, *66*(1), 22-25.

Van den Steen, P., Rudd, P. M., Dwek, R. A., & Opdenakker, G. (1998). Concepts and principles of *O*-linked glycosylation. *Critical reviews in biochemistry and molecular biology*, *33*(3), 151-208.

Van der Laan, J. M., Swarte, M. B., Groendijk, H., Hol, W. G., & Drenth, J. (1989). the influence of purification and protein heterogeneity on the crystallization of p-hydroxybenzoate hydroxylase. *European journal of biochemistry*, *179*(3), 715-724.

Van Putten, J. P. M. & Strijbis, K. (2017). Transmembrane Mucins: Signaling Receptors at the Intersection of Inflammation and Cancer. *Journal of Innate Immunity*, *9*(3), 281–299.

Varki, A., Cummings, R. D., Esko, J. D. Stanley, P., Hart, G. W., Aebi, M., Darvill, A. G., Kinoshita, T., Packer, N. H., Prestegard, J. H.,, Schnaar, R. L. & Seeberger, P. H. (2017) *Essentials of Glycobiology*, 3rd Edition. New York: Cold Spring Harbor.

Vosseller, K., Sakabe, K., Wells, L., & Hart, G. W. (2002). Diverse regulation of protein function by *O*-GlcNAc: a nuclear and cytoplasmic carbohydrate post-translational modification. *Current opinion in chemical biology*, *6*(6), 851-857.

Wagner, G. K., & Pesnot, T. (2010). Glycosyltransferases and their assays. *ChemBioChem*, *11*(14), 1939-1949.

Wandall, H. H., Hassan, H., Mirgorodskaya, E., Kristensen, A. K., Roepstorff, P., Bennett, E. P., & Clausen, H. (1997). Substrate specificities of three members of the human UDP-*N*-acetyl-α-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferase family, GalNAc-T1,-T2, and-T3. *Journal of Biological Chemistry*, *272*(38), 23503-23514.

Wandall, H. H., Irazoqui, F., Tarp, M. A., Bennett, E. P., Mandel, U., Takeuchi, H., Kato, K., Irimura, T., Suryanarayanan, G., Hollingsworth, M. A. & Clausen, H. (2007). 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*(4), 374-387

Wang, R. Q. & Fang, D. C. (2003). Alterations of MUC1 and MUC3 expression in gastric carcinoma: relevance to patient clinicopathological features. *Journal of clinical pathology*, *56*(5), 378-384.

Wang, S., Mao, Y., Narimatsu, Y., Ye, Z., Tian, W., Goth, C. K., E., Pedersen, N. B., Benito-Vicente, A., Martin, C., Uribe, K. B., Hurtado-Guerrero, R., Christoffersen, C., Seidah, N., Nielsen, R., Christensen, E. I., Hansen, L., Bennett, E. P., Vakhrushev, S. Y., Schjoldager, K. T. & Clausen, H. (2018). Site-specific *O*-glycosylation of members of the low-density lipoprotein receptor superfamily enhances ligand interactions. *Journal of Biological Chemistry*, *293*(19), 7408-7422.

Wei, X., Xu, H. & Kufe, D. (2005). Human MUC1 oncoprotein regulates p53-responsive gene transcription in the genotoxic stress response. *Cancer cell*, *7*(2), 167-178.

Weidemann, A., König, G., Bunke, D., Fischer, P., Salbaum, J. M., Masters, C. L., & Beyreuther, K. (1989). Identification, biogenesis, and localization of precursors of Alzheimer's disease A4 amyloid protein. *Cell*, *57*(1), 115-126.

White, T., Bennett, E. P., Takio, K., Sørensen, T., Bonding, N. & Clausen, H. (1995). Purification and cDNA cloning of a human UDP-*N*-acetyl-α-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferase. *Journal of Biological Chemistry*, *270*(41), 24156-24165.

Wildt, S., & Gerngross, T. U. (2005). The humanization of *N*-glycosylation pathways in yeast. *Nature Reviews Microbiology*, *3*(2), 119-123.

Wilson, J. E. (1976). Applications of blue dextran and Cibacron Blue F3GA in purification and structural studies of nucleotide-requiring enzymes. *Biochemical and biophysical research communications*, *72*(3), 816-823.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., McNicholas, S. J. & Wilson, S. K. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D*, *67*(4), 235-242.

Wu, S. & Letchworth, G. J. (2004). High efficiency transformation by electroporation of Pichia pastoris pretreated with lithium acetate and dithiothreitol. *Biotechniques*, *36*(1), 152-155.

Wu, Y., Ahmed, A., Waghmare, R., & KAHN, D. (2006). Cleaning polyethersulfone membranes after ultrafiltration-diafiltration in monoclonal antibody production. *Biopharm international*, *19*(4).

Wu, Y. M., Liu, C. H., Hu, R. H., Huang, M. J., Lee, J. J., Chen, C. H., & Huang, H. C. (2011). Mucin glycosylating enzyme GALNT2 regulates the malignant character of hepatocellular carcinoma by modifying the EGF receptor. *Cancer research*. doi: 10.1158/0008-5472.CAN-11-1161

Yang, Z., Halim, A., Narimatsu, Y., Joshi, H. J., Steentoft, C., Schjoldager, K. T. B. G., Schulz, M. A., Sealover, N. R., Kayser, K. J., Bennett, E. P., Levery, S. B., Vakhrushev, S. Y. & Clausen, H. (2014). The GalNAc-type *O*-glycoproteome of CHO cells characterized by the SimpleCell strategy. *Molecular & Cellular Proteomics*, *13*(12), 3224-3235.

CAPÍTULO 9- ABREVIATURAS

aa. Aminoácido
ADN. Ácido desoxirribonucleico
AFM. Microscopía de fuerza atómica
ANS. 1-anilino-8-naftleno sulfonato
ATP. Adenosín trifosfato
BMGY. Medio tamponado rico en glicerol (*Buffered Glycerol complex Medium*)
BMMY. Medio tamponado rico en metanoll (*Buffered Methanol complex Medium*)
CAZy. Enzimas activas hacia carbohidratos
CCI. Interacción carbohidrato-carbohidrato
DMSO. Dimetilsulfóxido
dNTPs. Desoxinucleótidos trifosfato.
DTT. Ditiotreitol
EDTA. Ácido etildiaminotetraacético
F. Forward.
GALA. Activación de las ppGalNAc-Ts.
GalNAc. *N*-Acetilgalactosamina
GalNAc-T. *N*-Acetilgalactosamina transferasas
GDP. Guanidín trifosfato
GlcNAc. *N*-Acetilglucosamina
GT. Glicosiltransferasa
HPLC. Cromatografía Líquida de Alta Presión
LB. Caldo de Lisogenia (*Lysogeny broth*)
LS. Bajo en sal (*Low Salt*)
MPTs. Modificaciones postraduccionales
OD. Densidad óptica
PAGE. Electroforesis en gel de poliacrilamida
PCR. Reacción en cadena de la polimerasa
PDB. Banco de Datos de Proteínas
PEG. Polietilenglicol
R. Reverse.
RMN. Resonancia Magnética Nuclear
SAXS. Dispersión de rayos X de ángulo pequeño
SDS. Dodecilsulfato sódico
S_N1. Sustitución nucleófila unimolecular
S_N2. Sustitución nucleófila bimolecular
S_Ni. Sustitución nucleófila intramolecular
STD. Diferencia de transferencia de saturación
S-Tn. Antígeno de Thomsen Friedenreich sialilado
TCEP. tris(2-carboxietil)fosfina
Tn. Antígeno de Thomsen-Friedenreich
TRIS. Tris(hidroxietil)aminometano
UDP. Uridina difosfato
UV. Ultravioleta

YPD. Extracto de levadura-peptona-dextrosa
YPDS. Extracto de levadura-peptona-dextrosa-sorbitol.

CAPÍTULO 10- APÉNDICES

10.1. MEDIOS DE CULTIVO

Low Salt LB-Agar

-10 g triptona

-5 g extracto de levadura

-5 g NaCl

-15 g agar

Low Salt LB

-10 g triptona

-5 g extracto de levadura

-5 g NaCl

YPD Agar

-20 g peptona

-10 g extracto de levadura

-20 g dextrosa

-15 g agar

YPDS Agar

-20 g peptona

-10 g extracto de levadura

-20 g dextrosa

-15 g agar

-183 g D-sorbitol

YPD

-20 g peptona

-10 g extracto de levadura

-20 g dextrosa

YPDS

-20 g peptona

-10 g extracto de levadura

-20 g dextrosa

-183 g D-sorbitol

BMGY

-20 g peptona

-10 g extracto de levadura

-13,4 g YNB (Yeast Nitrogen Base)

-4 x 10^{-3} g biotina

-10 ml glicerol

-100 mM fosfato potásico, pH 6.0

BMMY

-20 g peptona

-10 g extracto de levadura

-13,4 g YNB (Yeast Nitrogen Base)

-4 x 10^{-3} g biotina

-10 ml metanol

-100 mM fosfato potásico, pH 6.0

10.2. PROPIEDADES DE LAS PROTEÍNAS RECOMBINANTES

Se recogen a continuación las propiedades de las dos proteínas recombinantes silvestres con las que se ha trabajado a lo largo de esta tesis. Si bien a nivel experimental se tuvieron en cuenta, no se incluyen las de los respectivos mutantes y quimeras debido a que la diferencia es muy escasa y sólo conlleva sutiles variaciones en el valor de los diferentes parámetros.

10.2.1. *Hs*GalNAc-T2

KREAEAKVRWPDFNQEAYVGGTMVRSGQDPYARNKFNQVESDKLRMDRAIPDTRHDQCQRK
WRVDLPATSVVITFHNEARSALLRTVVSVLKKSPPHLIKEIILVDDYSNDPEDGALLGKIEKVRVLN
DRREGLMRSRVRGADAAQAKVLTFLDSHCECNEHWLEPLLERVAEDRTRVVSPIIDVINMDNFY
VGASADLKGGFDWNLVFKWDYMTPEQRRSRQGNPVAPIKTPMIAGGLFVMDKFYFEELGKYMM
MDVWGGENLEISFRVWQCGGSLEIIPCSRVGHVFRKQHPYTFPGGSGTVFARNTRRAAEVWDE
YKNFYYAAVPSARNVPYGNIQSRLELRKKLSCKPFKWYLENVYPELRVPDHQDIAFGALQQGTN
CLDTLGHFADGVVGVYECHNAGGNQEWALTKEKSVKHMDLCLTVVDRAPGSLIKLQGCREDSR
QKWEQIEGNSKLRHVGSNLCLDSRTAKSGGLSVEVCGPALSQQWKFTLNLQQ

En rojo, secuencia de exportación para *Pichia pastoris*.

Número de aminoácidos: 497
Peso molecular: 56576.44
pI teórico: 8.48
Número de residuos cargados positivamente (Asp + Glu): 68
Número de residuos cargados negativamente (Arg + Lys): 63
Coeficiente de extinción molar: 88350 $M^{-1}cm^{-1}$ (a 280 nm en agua, asumiendo que todas las cisteínas están reducidas)

Número de residuos cargados negativamente (Arg + Lys): 69
Coeficiente de extinción molar: 89270 $M^{-1}cm^{-1}$ (a 280 nm en agua, asumiendo que todas las cisteínas están reducidas)

10.2.2. *Hs*GalNAc-T4

KREAEAEAAGAGRARELGSRRLSDLQKNTEDLSRPLYKKPPADSRALGEWGKASKLQLNEDEL
KQQEELIERYAINIYLSDRISLHRHIEDKRMYECKSQKFNYRTLPTTSVIIAFYNEAWSTLLRTIHSV
LETSPAVLLKEIILVDDLSDRVYLKTQLETYISNLDRVRLIRTNKREGLVRARLIGATFATGDVLTFL
DCHCECNSGWLEPLLERIGRDETAVVCPVIDTIDWNTFEFYMQIGEPMIGGFDWRLTFQWHSVP
KQERDRRISRIDPIRSPTMAGGLFAVSKKYFQYLGTYDTGMEVWGGENLELSFRVWQCGGKLEI
HPCSHVGHVFPKRAPYARPNFLQNTARAAEVWMDEYKEHFYNRNPPARKEAYGDISERKLLRE
RLRCKSFDWYLKNVFPNLHVPEDRPGWHGAIRSRGISSECLDYNSPDNNPTGANLSLFGCHGQ
GGNQFFEYTSNKEIRFNSVTELCAEVPEQKNYVGMQNCPKDGFPVPANIIWHFKEDGTIFHPHS
GLCLSAYRTPEGRPDVQMRTCDALDKNQIWSFEKHHHHHH

En rojo, secuencia de exportación para *Pichia pastoris*. En azul, cola de Histidinas en el extremo C-terminal.

Número de aminoácidos: 557

Peso molecular: 64445.93
pI teórico: 7.12
Número de residuos cargados positivamente (Asp + Glu): 74
Número de residuos cargados negativamente (Arg + Lys): 75
Coeficiente de extinción molar: 102790 $M^{-1}cm^{-1}$ (a 280 nm en agua, asumiendo que todas las cisteínas están reducidas)

10.3. INFORMACIÓN GENERAL

10.3.1-El código genético

PRIMERA LETRA	SEGUNDA LETRA: U		SEGUNDA LETRA: C		SEGUNDA LETRA: A		SEGUNDA LETRA: G		TERCERA LETRA
U	UUU	Fenilalanina	UCU	Serina	UAU	Tirosina	UGU	Cisteína	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucina	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Triptófano	G
C	CUU	Leucina	CCU	Prolina	CAU	Histidina	CGU	Arginina	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamina	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucina	ACU	Treonina	AAU	Asparagina	AGU	Serina	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lisina	AGA	Arginina	A
	AUG	Metionina	ACG		AAG		AGG		G
G	GUU	Valina	GCU	Alanina	GAU	Aspártico	GGU	Glicina	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutámico	GGA		A
	GUG		GCG		GAG		GGG		G

Figura 10.1. Código genético. Tabla de correspondencias en la transcripción de los distintos codones a aminoácidos, en humanos. Existen algunas variaciones entre organismos.

10.3.2-Los aminoácidos

AMINOÁCIDO	ABREVIATURA	LETRA	Pka_1 (α- COOH)	Pka_2 α-NH_2	Pka_R R= cadena lateral	pI
Alanina	Ala	A	2.35	9.69		6.02
Arginina	Arg	R	2.17	9.04	12.48	10.76
Asparagina	Asg	N	2.02	8.80		5.41
Acido Aspártico	Asp	D	2.09	9.82	3.86	2.97
Cisteina	Cys	C	1.96	10.28	8.18	5.07
Fenilalanina	Phe	F	1.83	9.24		5.53
Glutamina	Glu	Q	2.17	9.13		5.65
Acido glutámico	Glu	E	2.19	9.67	4.25	3.22
Glicina	Gli	G	2.34	9.78		6.06
Histidina	His	H	1.82	9.17	6.00	7.58
Isoleucina	Ile	I	2.36	9.68		6.02
Leucina	Leu	L	2.36	9.64		6.00
Lisina	Lys	K	2.18	8.95	10.53	9.74
Metionina	Met	M	2.28	9.21		5.75
Prolina	Pro	P	1.99	10.6		6.30
Serina	Ser	S	2.21	9.15		5.68
Tirosina	Try	Y	2.20	9.11	10.07	5.65
Treonina	Tre	T	2.71	9.62		6.16
Triptófano	Trp	W	2.38	9.39		5.89
Valina	Val	V	2.32	9.62		5.97

Figura 10.2 Propiedades físicas de los aminoácidos. Tabla resumen de la nomenclatura y algunas propiedades físicas de los distintos aminoácidos. La abreviatura actualmente aceptada para el residuo Treonina en castellano es Thr, no Tre.

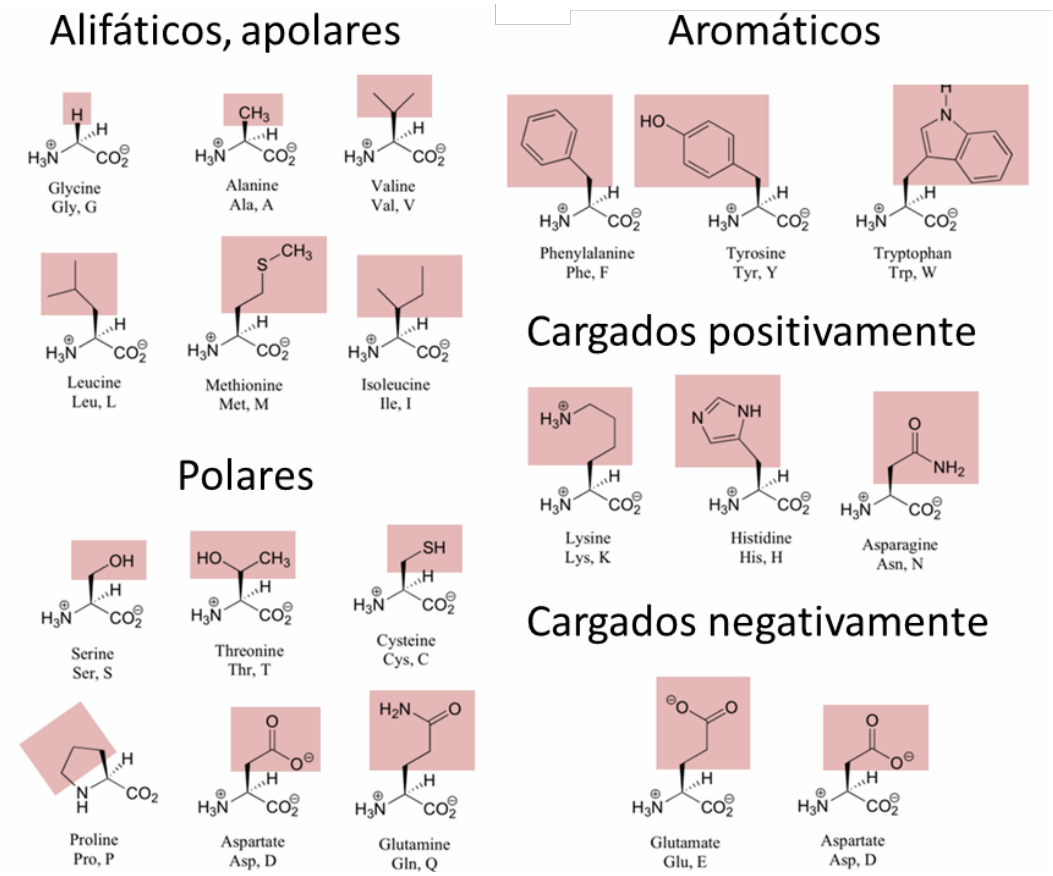


Figura 10.3. Estructura de los aminoácidos. Estructura y clasificación de los diferentes aminoácidos.

10.4. FACTOR DE IMPACTO Y ÁREA TEMÁTICA DE LAS REVISTAS

ARTÍCULO I

Lira-Navarrete, E., de las Rivas, M., Compañón, I., Carmen Pallarés, M. C., Kong, Y., Iglesias-Fernández, J., Bernades, G. J. L., Peregrina, J. M., Rovira, C., Bernadó, P., Bruscolini, P., Clausen, H., Lostao, A., Corzana, F. & Hurtado-Guerrero, R. (2015). Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein *O*-glycosylation. *Nature Communications*, 6(6937), DOI: 10.1038/ncomms7937.

Base de datos: Journal of Citations Reports (JCR)

Área temática: Multidisciplinary Sciences – Science

Índice de impacto (2016): 12.124

ARTÍCULO II

de las Rivas, M., Lira-Navarrete, E., James Paul Daniel, E., Compañón, I., Coelho, H., Diniz, A., Jiménez-Barbero, J., Peregrina, J. M., Clausen, H., Corzana, F., Marcelo, F., Jiménez-Osés, F., Gerken, . A. & Hurtado-Guerrero, R. (2017). The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. *Nature Communications, 8(1959),* DOI: 10.1038/s41467-017-02006-0.

Base de datos: Journal of Citations Reports (JCR)

Área temática: Multidisciplinary Sciences – Science

Índice de impacto (2016): 12.124

ARTÍCULO III

Ghirardello, M., **de las Rivas, M.**, Lacetera, A., Delso, I., Lira-Navarrete, E,. Tejero, T., Martín-Santamaría, S., Hurtado-Guerrero, R., & Merino, P. (2016). Glycomimetics Targeting Glycosyltransferases: Synthetic, Computational and Structural Studies of Less-Polar Conjugates, *Chemistry - A European Journal*, 22, 7215-7224.

Base de datos: Journal of Citations Reports (JCR)

Área temática: Chemistry, Multidisciplinary Science - Sciences

Índice de impacto (2016): 5.317

ARTÍCULO IV

Liu, F., Xu, K., Xu, Z., de las Rivas, M., Wang, C., Li, X., Lu, J., Zhou, Y., Delso, I., Merino, P., Hurtado-Guerrero, R., Zhang, Y. & Wu, F. (2017). The small molecule luteolin inhibits *N*-acetyl-alpha-galactosaminyltransferases and reduces mucin-type O-glycosylation of amyloid precursor protein. *Journal of Biological Chemistry*, 292(52), pp. 21304-21319.

Base de datos: Journal of Citations Reports (JCR)

Área temática: Biochemistry & Molecular Biology - Science

Índice de impacto (2016): 4.125

ARTÍCULO V

de las Rivas, M., Coelho, H., Diniz, A., Lira-Navarrete, E., Compañón, I., Jiménez-Barberó, J., Schjoldager, K., T., Bennett, E. P., Vakhrushev, S. Y., Clausen, H., Corzana, F., Marcelo, F. & Hurtado-Guerrero, R. (2018). Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts. *Chemistry - A European Journal.* DOI: 10.1002/chem.201800701.

Base de datos: Journal of Citations Reports (JCR)

Área temática: Chemistry, Multidisciplinary Science - Sciences

Índice de impacto (2016): 5.317

10.5. JUSTIFICACIÓN DE LA CONTRIBUCIÓN DEL DOCTORANDO

D. RAMÓN HURTADO-GUERRERO, investigador ARAID del Instituto de Biocomputación y Física de Sistemas Complejos de la Universidad de Zaragoza,

CERTIFICA:

que la contribución de la doctoranda Matilde Teresa de las Rivas González de Garay en los trabajos recogidos en esta tesis doctoral, titulada **Estudio de las preferencias de glicosilación de las ppGalNAc-Ts: implicación en el desarrollo de inhibidores**, y que ha sido dirigida por él y presentada en la modalidad de compendio de artículos, son los que se especifican antes de cada uno de los artículos que conforman la sección de Resultados de dicha tesis.

Que la participación y trabajo de la doctoranda a lo largo de su periodo de tesis doctoral ha sido requisito indispensable para la realización de los estudios reflejados en los artículos, así como lo ha sido la colaboración con otros grupos de investigación, dada la naturaleza multidisciplinar del acercamiento realizado.

Y para que así conste, firma en Zaragoza, a julio de 2018.

10.6. RENUNCIA DE LOS COAUTORES

Escuela de Doctorado
Universidad Zaragoza

RENUNCIA DE LOS COAUTORES DE LOS TRABAJOS PRESENTADOS COMO PARTE DE UNA TESIS DOCTORAL EN LA MODALIDAD DE COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Compañón Pérez	Nombre: Ismael	
DNI/Pasaporte/NIE: 16616491-A	Teléfono	Correo electrónico: ismael.companon@unirioja.es

2.- Tesis Doctoral
Título: Estudio estructural y mecanístico de las ppGalNAc-Ts.
Autor: Matilde de las Rivas González de Garay
Programa de doctorado: Doctorado en Bioquímica y Biología Molecular (RD 99/2011)

3.- Publicaciones que formarán parte de la tesis y de las que el firmante es coautor

1) Lira-Navarrete, E., de las Rivas, M., Compañón, I., Carmen Pallarés, M. C., Kong, Y., Iglesias-Fernández, J., Bernades, G. J. L., Peregrina, J. M., Rovira, C., Bernadó, P., Bruscolini, P., Clausen, H., Lostao, A., Corzana, F. y Hurtado-Guerrero, R. (2015). Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein O-glycosylation. Nature Communications, 6 (6937), DOI: 10.1038/ncomms7937.

2) de las Rivas, M., Lira-Navarrete, E., James Paul Daniel, E., Compañón, I., Coelho, H., Diniz, A., Jiménez-Barbero, J., Peregrina, J. M., Clausen, H., Corzana, F., Marcelo, F., Jiménez-Osés, F., Gerken, . A. y Hurtado-Guerrero, R. (2017). The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. Nature Communications, 8(1959), DOI: 10.1038/s41467-017-02006-0.

3) de las Rivas, M., Coelho, H., Diniz, A., Lira-Navarrete, E., Compañón, I., Jiménez-Barberó, J., Schjoldager, K., T., Bennett, E. P., Vakhrushev, S. Y., Clausen, H., Corzana, F., Marcelo, F. y Hurtado-Guerrero, R. (2018). Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts. Chemistry. A European Journal. DOI: 10.1002/chem.201800701.

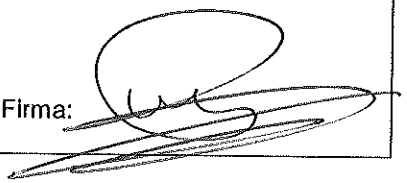
RENUNCIA:

Renuncio a que las publicaciones anteriores puedan ser presentadas como parte de otra tesis doctoral en la modalidad de compendio de publicaciones.

<lugar>,<fecha>

LOGROÑO, 23 de MAYO de 2018

Firma:



Escuela de Doctorado
Universidad Zaragoza

RENUNCIA DE LOS COAUTORES DE LOS TRABAJOS PRESENTADOS COMO PARTE DE UNA TESIS DOCTORAL EN LA MODALIDAD DE COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Daniel	Nombre: Earnest James Paul	
DNI/Pasaporte/NIE: H1807852	Teléfono 2163684556	Correo electrónico exp190@case.edu

2.- Tesis Doctoral
Título: Estudio estructural y mecanístico de las ppGalNAc-Ts.
Autor: Matilde de las Rivas González de Garay
Programa de doctorado: Doctorado en Bioquímica y Biología Molecular (RD 99/2011)

3.- Publicaciones que formarán parte de la tesis y de las que el firmante es coautor

1) de las Rivas, M., Lira-Navarrete, E., James Paul Daniel, E., Compañón, I., Coelho, H., Diniz, A., Jiménez-Barbero, J., Peregrina, J. M., Clausen, H., Corzana, F., Marcelo, F., Jiménez-Osés, F., Gerken, . A. y Hurtado-Guerrero, R. (2017). The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. Nature Communications, 8(1959), DOI: 10.1038/s41467-017-02006-0.

RENUNCIA:

Renuncio a que las publicaciones anteriores puedan ser presentadas como parte de otra tesis doctoral en la modalidad de compendio de publicaciones.

<lugar>,<fecha>
Cleveland, USA 6.12.2018

Firma: Earnest

Escuela de Doctorado
Universidad Zaragoza

RENUNCIA DE LOS COAUTORES DE LOS TRABAJOS PRESENTADOS COMO PARTE DE UNA TESIS DOCTORAL EN LA MODALIDAD DE COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Nobre Coelho	Nombre: Helena Maria	
DNI/Pasaporte/NIE: DNI:13791062 NIE:Y4157515A	Teléfono +351919006	Correo electrónico h.coelho@campus.fct.unl.pt

2.- Tesis Doctoral
Título: Estudio estructural y mecanístico de las ppGalNAc-Ts.
Autor: Matilde de las Rivas González de Garay.
Programa de doctorado: Doctorado en Bioquímica y Biología Molecular (RD 99/2011)

3.- Publicaciones que formarán parte de la tesis y de las que el firmante es coautor

1) de las Rivas, M., Lira-Navarrete, E., James Paul Daniel, E., Compañón, I., Coelho, H., Diniz, A., Jiménez-Barbero, J., Peregrina, J. M., Clausen, H., Corzana, F., Marcelo, F., Jiménez-Osés, F., Gerken, . A. y Hurtado-Guerrero, R. (2017). The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. Nature Communications, 8(1959), DOI: 10.1038/s41467-017-02006-0.

2) de las Rivas, M., Coelho, H., Diniz, A., Lira-Navarrete, E., Compañón, I., Jiménez-Barberó, J., Schjoldager, K., T., Bennett, E. P., Vakhrushev, S. Y., Clausen, H., Corzana, F., Marcelo, F. y Hurtado-Guerrero, R. (2018). Structural analysis of a GalNAc-T2 mutant reveals an induced-fit catalytic mechanism for GalNAc-Ts. Chemistry. A European Journal. DOI: 10.1002/chem.201800701.

RENUNCIA:

Renuncio a que las publicaciones anteriores puedan ser presentadas como parte de otra tesis doctoral en la modalidad de compendio de publicaciones.

<lugar>,<fecha>

Bilbao 31/05/2018

Helena Coelho

Firma: Helena Coelho

Escuela de Doctorado
Universidad Zaragoza

RENUNCIA DE LOS COAUTORES DE LOS TRABAJOS PRESENTADOS COMO PARTE DE UNA TESIS DOCTORAL EN LA MODALIDAD DE COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos: Liu	**Nombre:** Feng	
DNI/Pasaporte/NIE: E25439351	**Teléfono** 021-3420677	**Correo electrónico** liu89liu89liu@163.com

2.- Tesis Doctoral
Título: Estudio estructural y mecanístico de las ppGalNAc-Ts.
Autor: Matilde de las Rivas González de Garay
Programa de doctorado: Doctorado en Bioquímica y Biología Molecular (RD 99/2011)

3.- Publicaciones que formarán parte de la tesis y de las que el firmante es coautor

1) Liu, F., Xu, K., Xu, Z., de las Rivas, M., Wang, C., Li, X., Lu, J., Zhou, Y., Delso, I., Merino, P., Hurtado-Guerrero, R., Zhang, Y. and Wu, F. (2017). The small molecule luteolin inhibits N-acetyl-alpha-galactosaminyltransferases and reduces mucin-type O-glycosylation of amyloid precursor protein. Journal of Biological Chemistry, 292(52), pp. 21304-21319.

RENUNCIA:

Renuncio a que las publicaciones anteriores puedan ser presentadas como parte de otra tesis doctoral en la modalidad de compendio de publicaciones.

<lugar>,<fecha>
Shanghai, China, 25th.May.2018

Firma: Feng Liu

Escuela de Doctorado
Universidad Zaragoza

RENUNCIA DE LOS COAUTORES DE LOS TRABAJOS PRESENTADOS COMO PARTE DE UNA TESIS DOCTORAL EN LA MODALIDAD DE COMPENDIO DE PUBLICACIONES

1.- Datos personales del coautor		
Apellidos:	Nombre:	
DNI/Pasaporte/NIE:	Teléfono	Correo electrónico

2.- Tesis Doctoral
Título:
Autor:
Programa de doctorado:

3.- Publicaciones que formarán parte de la tesis y de las que el firmante es coautor

RENUNCIA:
Renuncio a que las publicaciones anteriores puedan ser presentadas como parte de otra tesis doctoral en la modalidad de compendio de publicaciones. <lugar>,<fecha> Firma: Ana Catarina Vitor Ferreira Diniz Ana Catarina Diniz