Relaciones secuencia-estructura-función en glicosiltransferasas con plegamiento GTA: una aproximación bioinformática

Javier Romero García

Universitat Ramon Llull

TESIS DOCTORAL

Título Relaciones secuencia-estructura-función en glicosiltransferasas con plegamiento GTA: una aproximación bioinformática

Realizada por Javier Romero-García

en el Centro Institut Químic de Sarrià

y en el Departamento de Bioenginyeria

Dirigida por los doctores Xevi Biarnés Fontal y Antoni Planas Sauter

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

A mis peques: Gael y Dante.

Esta tesis ha sido difícil, como todas, y emocionante, como muchas también. Aquí están resumidas horas y horas de trabajo, de lecturas, de aprendizaje continuo y de creatividad, mucha creatividad. Ha habido momentos de frustración, sí que los ha habido, pero también de una enorme satisfacción, la de ver mi proyecto crecer y la de culminar una meta tras otra hasta llegar aquí, el instante de clavar la bandera, alzar la vista, admirar el camino recorrido y dar gracias a todos los que, de una forma u otra, han hecho posible que esto sucediese. Los primeros, los que me dieron la vida, mis padres:

Gracias mamá, por haberme hecho responsable, coherente y consecuente. Por hacerme entender que mis éxitos o fracasos, son eso, mis éxitos y mis fracasos, y es mi decisión saber qué hacer con ellos, tomando las riendas de mi vida y de mi destino. Gracias por enseñarme a respetar, a escuchar y a valorar sin juzgar, por hacerlo siempre con tu ejemplo y por hacer de mí la persona que soy. Y por quererme siempre, siempre.

Gracias papá, por poner el mundo al alcance de mi mano, por conseguir lo imposible, por mantener siempre vivo mi pensamiento crítico y alimentar la insaciable curiosidad, que siempre has fomentado. Gracias por estimar, apreciar y respetar mis decisiones, aunque no las compartas, y por enseñarme el valor de las cosas bien hechas, del esfuerzo, del trabajo y el poder del conocimiento. Y por quererme siempre, siempre.

Sin ellos no sería quien soy y por ellos estoy donde estoy. Pero hay alguien más, que ha sufrido muy de cerca los altibajos de esta tesis ciclotímica; a mi lado, día y noche, con el estoicismo y la paciencia de quien tiene mucho mucho amor que dar:

Gracias Natalia, gracias por todo. Sin ti, sin tu apoyo, tu comprensión y sacrificio, esto nunca se habría conseguido. Te conocí y no quise conocer nada más. Los mejores momentos de mi vida los recuerdo siempre contigo, porque tú haces que sean mejores, únicos. Contigo, todos los días son sábados. Dejaste mucho por estar a mi lado, porque estuviéramos juntos. Después, llegó esta tesis y llegaron más cosas, ya tú sabes, y empezamos a conocer los lunes, los martes, los miércoles … Cada minuto que he dedicado a mi trabajo, es un minuto que me has regalado tú. Gracias por estar a mi lado, por escucharme y entenderme. Gracias por tu visión del mundo, por tu amor a los más débiles, por tu empatía, tu sensibilidad; por las juergas que nos hemos dado y las que nos hemos perdido, qué pasa, ya las recuperaremos. Por tu profunda mirada, tu silenciosa risa y por hacerme feliz. Esto es tan tuyo como mío, gracias mi amor.

Además de ellos, hay otras dos personas que, literalmente, han hecho posible la realización de todo este proyecto científico, mis directores de tesis, los doctores Antoni Planas y Xevi Biarnés:

Gracias Toni, por tu confianza, por darme esta oportunidad, por creer que sería válido para ello y luchar por mí cuando lo he necesitado. Gracias por tu apoyo, pero también por los buenos ratos, nuestros debates sobre ciencia, libros, historia, hasta política. Gracias por elevar el listón al máximo.

A ti, Xevi, no sé si darte las gracias o escribirte un poema. Gracias, mil gracias, un millón de gracias. Detrás de cada hipótesis, resultado y conclusión de este trabajo están tus sólidos conocimientos y tu inteligente criterio. Cuanto más aprendía de ti, más mostrabas conocer. Me has dedicado tu tiempo, tu atención y siempre has estado disponible, siempre con una sonrisa, a pesar del ingente trabajo, presión y fechas límite con las que siempre, también, andas batallando. Me has enseñado y me has demostrado que la humildad, la naturalidad, el compañerismo y la buena actitud, son el mejor recipiente para la excelencia científica. Tú, eres excelente y no podría haber tenido un director mejor. Los dos nos hemos estrenado en estas

lides y aunque, seguro que detrás de mí vendrán muchos más (por el bien de la comunidad científica así lo espero), para el que escribe tú siempre serás el primero y el último. Gracias por ser como eres, gracias por conseguir esto conmigo.

Gracias a todos mis compañeros de máster, de doctorado. A todos y cada uno de vosotros, gracias; por las risas, las confidencias, los viajes, por los desayunos, las meriendas y éxitos que hemos celebrado juntos. Desde un simple saludo hasta una buena fiesta, todos habéis contribuido para que mi tiempo con vosotros haya sido fantástico e inolvidable. No me pidáis agradecimientos personalizados por que no acabaríamos nunca, ya sabéis cuánto hemos compartido y de todo eso está bañada esta tesis, en toda ella estáis también vosotros. Gracias chicos y chicas.

Gracias Magda, porque buena gente se escribe con M y porque me echaste un cable cuando más lo necesitaba.

Gracias Patri, por cederme espacio en la mesa, bolis, grapadora … y todo cuanto pudiera necesitar. Eres genial.

Gracias Andrea, por tu valía, por tu criterio, por tu sonrisa, por el tiempo que pasamos juntos y por ser mi amiga.

Merci au Dr. Henrissat ainsi qu'à l'ensemble du personnel de CAZy. Merci de m'avoir permis de rencontrer une équipe d'une si haute qualité scientifique et humaine, pour m'avoir enseigné, et pour avoir fait que, pendant trois mois, je me suis levé chaque jour excité et heureux, pour avoir appris et pour avoir été avec vous. Il n'y aura pas un seul verre de pastis que je ne boirai sans trinquer à CAZy.

Gracias a todos los miembros de mi tribunal: el Dr. Oliva, el Dr. Estrada, la Dra. Domene, la Dra. Faijes, el Dr. Guerin, el Dr. Fita, el Dr. Texidó y también a la Dra. André.

También doy, con todo mi cariño, un enorme gracias a todos mis compañeros de la Facultad de Biología de la Universidad de Málaga, en especial al Dr. Thode:

Gracias Guille, contigo empezó todo y mira a dónde hemos llegado, gracias por ser tan cercano y tan especial, y gracias a ti también Antonio, por lo que compartimos y lo que aprendí de ti.

Gracias a mis retoños, por no dejarme dormir, por no dejarme escribir, por no dejarme comer ni trabajar, por interrumpirme, descentrarme, por volverme loco y por hacer de mí el hombre más feliz que haya pisado la Tierra.

Quiero dar las gracias a la CIENCIA, así en mayúsculas, y al Método Científico y a los bomberos y peluqueros y surfistas y bailaores. A los quinceañeros, a los jubilados, a los jardineros, a las suegras y a los cuñaos … a todos los que no se conforman, a los que buscan aunque no encuentren, a todos los que se preguntan: ¿Por qué? ¿Cómo? o ¿Para qué?

Gracias al ser humano y a su incontenible curiosidad.

Merci a tots!

"Quizá la conciencia surja cuando la simulación cerebral del mundo llega a ser tan compleja, que debe incluir un modelo de sí misma."

Richard Dawkins – El gen egoísta.

RESUMEN DE LA TESIS

En esta tesis se ha realizado un estudio global, acerca de las relaciones secuencia-estructura-función de proteínas glicosiltransferasa con plegamiento GTA, mediante un enfoque bioinformático y de biología computacional. La tesis está estructurada en cuatro capítulos principales. En los dos primeros se aborda este estudio para una enzima esencial de *Mycoplasma genitalium* involucrada en la síntesis de glicolípidos de membrana (MG517). En el tercer capítulo se estudian los cambios conformacionales de un bucle catalítico, en el mecanismo de una proteína de *Micobacterium tuberculosis* (GpgS), que inicia la ruta biosintética de los lipopolisacáridos 6-O-metilglucosa (MGLPs) en este organismo. Por último, en el capítulo 4 se aborda la relación entre una región específica de las glicosiltransferasas con plegamiento GTA, y la especificidad por el sustrato.

Comenzando por el estudio de proteínas específicas, como la proteína MG517 de *Mycoplasma genitalium* o GpgS de *Micobacterium tuberculosis*, pasando por el de las estructuras de todas las proteínas cristalizadas con plegamiento GTA, hasta la utilización de todas las secuencias existentes para esta superfamilia de proteínas (más de 100000), se han encontrado características comunes a todas ellas que relacionan la secuencia, con la estructura y la función de cada proteína. Todo ello se ha conseguido utilizando técnicas y métodos bioinformáticos y computacionales, entre los que destacan: el modelado por homología, las simulaciones de Dinámica Molecular y Metadinámica, el *Docking* de ligandos, la superposición de estructuras tridimensionales, alineamientos múltiples y la construcción de árboles filogenéticos.

Gracias a este estudio, se ha podido identificar una topología consenso común a todas las GTAs, con la que se ha construido un modelo tridimensional de la región N-terminal de la proteína MG517, de la que hasta ahora no existía estructura conocida. El modelo tridimensional ha sido validado por dinámica molecular y experimentalmente, lo cual ha permitido identificar posiciones catalíticas clave en MG517 como E193 base general, D40, Y126, Y169, I170 y Y218 de unión a sustratos. Además, para MG517 se ha propuesto un modelo de interacción monotópica con la membrana, mediante una hélice anfipática en su región C-terminal.

La misma topología consenso, ha permitido el refinamiento de un alineamiento múltiple de GTAs, con el que se ha generado un perfil *Hidden Makov Model* (HMM) para la región N-terminal de este grupo de proteínas. Este perfil facilita el alineamiento de nuevas proteínas GTAs y la identificación de sus estructuras secundarias. También ha permitido identificar, dentro de la topología consenso, una región de secuencia y estructura muy variable, incluso para proteínas de la misma familia, donde se posiciona el aceptor específico de cada proteína y que hemos denominado "Región Variable".

Se han descrito los cambios conformacionales del bucle catalítico en la proteína GpgS mediante simulaciones de dinámica molecular de larga duración y distintos cálculos de metadinámica utilizando multitud de variables colectivas. Se ha demostrado que las distintas conformaciones son una propiedad intrínseca de la proteína, pero están desplazadas hacia su forma inactiva en ausencia de ligandos. La presencia del sustrato dador más el metal en el centro activo, promueve el movimiento de las cadenas laterales de dos residuos del bucle, Arg256 e His258, que desplaza el equilibrio hacia la forma activa y la estabiliza mediante la interacción de His258 con el metal, proponiéndose un mecanismo de ajuste inducido para esta proteína.

Completando estos resultados con cálculos de *docking* de ligandos, se ha podido proponer el orden de entrada de los ligandos, siendo el aceptor el primero en llegar al centro activo, seguido por el dador, momento en que sucede el cambio conformacional en el bucle. A raíz de estas simulaciones, se ha observado que la interacción del metal con un residuo de histidina en el bucle catalítico, es una característica común a la gran mayoría de familias GTAs, junto a las ya conocidas del motivo DXD o la tétrada de aspartatos propuesta en literatura D-DXD-D, que se propone cambiar a D-DXD-D-H.

Se ha estudiado la evolución de la superfamilia GTAs y su relación con los sustratos, encontrando que el plegamiento global del dominio GTA, define la especificidad del aceptor y que la homología entre secuencias está más influida por esta molécula aceptora del azúcar que por la molécula dadora. La región variable parece sufrir una presión evolutiva menor que el resto de la secuencia, lo que explica su mayor variabilidad, sin embargo, contiene residuos altamente conservados que interaccionan con el aceptor específico de cada proteína. Se ha utilizado esta región para la generación de perfiles HMM específicos para cada aceptor y familia de proteínas. Estos perfiles se han utilizado con éxito para el *screening* de proteínas GTA de función desconocida y la predicción de su aceptor.

ABSTRACT

This thesis has conducted a comprehensive study on the sequence-structure-function relations for glycosyltransferase proteins with GTA fold, through a bioinformatic computational biology approach and. The thesis is divided into four main chapters. In the first two, is approached this study by an essential enzyme in *Mycoplasma genitalium* involved in synthesis membrane glycolipids (MG517). In the third chapter, conformational changes of a catalytic loop are studied in the mechanism of a protein of *Mycobacterium tuberculosis* (GpgS), which starts the 6-O-methyl glucose biosynthetic pathway of lipopolysaccharide (MGLPs) in this organism. Finally, in chapter 4, the relationship between a specific region of the glycosyltransferase with GTA folding and substrate specificity is addressed.

Starting with the study of specific proteins, such as *Mycoplasma genitalium* MG517 protein or *Mycobacterium tuberculosis* GpgS, through the structures of all proteins crystallized with GTA folding, and the use of all existing sequences for this protein superfamily (over 100000), it has found common characteristics to them all linking sequence, structure and function of each protein. All this, has been achieved using bioinformatics and computational techniques and methods, among which are: homology modeling, simulations of Molecular Dynamics and Metadinámica, the Docking of ligands, overlapping three-dimensional structures, multiple alignments and phylogenetic tree building.

Thanks to this study, it has been possible to identify a consensus topology common to all GTAs, with which it has built a three-dimensional model of the N-terminus of the protein MG517, region which hitherto was known structure. The three-dimensional model has been validated experimentally by molecular and dynamic, which has identified key catalytic MG517 positions as general base E193, D40, Y126, Y169, Y218 I170 and substrate binding. Furthermore, for it has been proposed a MG517 monotopic membrane interaction model by an amphipathic helix in the C-terminal region.

The same topology consensus has permitted the refinement of a multiple alignment of GTAs, with which we generate a profile Hidden Makov Model (HMM) for the N-terminal region of this group of proteins. This profile facilitates the alignment of GTAs new proteins and identifying their secondary structures. It has also enabled us to identify, within the consensus topology, a region of highly variable sequence and structure, even for proteins of the same family, where each protein specific acceptor is positioned that we have called "Variable Region".

Have been described conformational changes in the catalytic loop GpgS protein, by long duration Molecular Dynamics simulations and different calculations of metadinámica using many collective variables. It has been shown that different conformations, are an intrinsic property of the protein, but are displaced towards its inactive form in the absence of ligands. The presence of the donor substrate plus metal in the active site, promotes the movement of the side chains of two residues of the loop, Arg256 and His258, which shifts the equilibrium to the active form and stabilizes by the interaction of His258 with metal, proposing an induced fit mechanism for this protein. Completing these results with docking of ligands calculations, it has been possible to propose the order of entry of the ligands, being the acceptor the first to reach the active site, followed by the donor, when that happens the conformation loop changes. Following these simulations, it has been observed that the interaction of the metal with a histidine residue in the catalytic loop, is a feature

common to the vast majority of GTAs families, with the already known motif DXD or tetrad aspartates proposal in literature D-DXD-D, proposed switch to D-DXD-DH.

We have studied the evolution of GTAs superfamily and their relationship with the substrates, finding that the overall folding of GTA domain, defines the acceptor specificity and the homology between sequences is more influenced by the acceptor molecule of sugar than the molecule donor. The variable region seems to suffer less evolutionary pressure than the rest of the sequence, which explains its greater variability, however, contain highly conserved residues that interact with the specific acceptor for each protein. This region has been used for profiling specific HMM for each acceptor and protein family. These profiles have been successfully used for screening GTA proteins of unknown function and predicting their acceptor.

ÍNDICE

ÍNDICE DE FIGURAS 13

ABREVIATURAS 20

INTRODUCCIÓN A LAS GLICOSILTRANSFERASAS Y BIOLOGÍA COMPUTACIONAL 21

1. GLICOSILTRANSFERASAS 23
1.2 Transferencia del azúcar: Mecanismos. 25
1.3 Plegamiento 27
1.4 Clasificación 28
2. BIOLOGÍA COMPUTACIONAL 31
2.1 Modelado por homología 33
2.2 Docking (Acoplamiento molecular) 35
2.3 Dinámica Molecular 37
2.3.1 Diseño de la simulación 38
2.3.2 Campo de fuerzas 39
2.3.3 Algoritmo básico de una simulación MD 40
2.4 Metadinámica 43
2.5 BIAS-Exchange 47

MARCO DE LA TESIS 49

1. PROTEÍNA MG517 53
2. PROTEÍNA GPGS 55

OBJETIVOS 57

IDENTIFICACIÓN DE RESIDUOS CATALÍTICOS EN LA PROTEÍNA MG517, A PARTIR DE LA GENERACIÓN DE UN MODELO TRIDIMENSIONAL. 63

1. INTRODUCCIÓN 65
2. SELECCIÓN DE LA REGIÓN A MODELAR (N-TERMINAL) 66
3. MODELADO 74
4. SITIO DE UNIÓN DEL DADOR 77
5. EVALUACIÓN FUNCIONAL POR MUTAGÉNESIS DIRIGIDA 80
6. MEJORA DE LOS MODELOS POR DINÁMICA MOLECULAR DE LARGA DURACIÓN 84
7. SITIO DE UNIÓN DEL ACEPTOR 87
8. DISCUSIÓN 89

INTERACCIÓN CON LA MEMBRANA DE LA PROTEÍNA MG517, A TRAVÉS DE LA REGIÓN C-TERMINAL. 91

1. INTRODUCCIÓN 93
2. IDENTIFICACIÓN DE ZONAS DE INTERACCIÓN 96
2.1. Evaluación del perfil hidrofóbico 97
2.2 Evaluación del perfil anfipático 100
3. MODELADO DE HÉLICES HIDROFÓBICAS Y ANFIPÁTICAS 101
4. ESTABILIDAD DE LOS MODELOS GENERADOS 102
4.1 Hélices en solvente acuoso 102
4.2 Hélice 4 inserta en membrana 103
5. ENERGÍA DE UNIÓN A LA MEMBRANA. 113
6. DISCUSIÓN. 117

CAMBIOS CONFORMACIONALES DEL BUCLE CATALÍTICO EN EL MECANISMO DE LA PROTEÍNA GPGS. 119

1. INTRODUCCIÓN 121
2. CONSTRUCCIÓN DE MODELOS A PARTIR DE ESTRUCTURAS CRISTALOGRÁFICAS 125
3. ESTUDIO DE LOS MODELOS Y BÚSQUEDA DE DESCRIPTORES DEL BUCLE. 127
4. SIMULACIONES DE MODELOS DE GPGS EN AUSENCIA DE LIGANDOS 134
4.1 Simulaciones de modelos basados en la estructura 4DDZ: MD4DDZ 134
4.2 Simulaciones de modelos basados en la estructura 4Y6N: MD4Y6NApo 139

4.3 Energías libres asociadas a cambios conformacionales de GpgS apo 142
5. Simulaciones del complejo ternario y modelos con ligandos 150
5.1 Simulaciones del complejo ternario de GpgS basadas en los modelos de 4Y6N 150
5.1.1. Cambio conformacional del UDPGlc: 153
5.2 Interacción de los ligandos con la proteína 156
5.3 Energías libres de los cambios conformacionales en presencia de ligandos (I). CV: Diedro Ala257-Distancia 159
5.4 Energías libres de cambios conformacionales en presencia de ligandos (II): Otras CVs. 163
6. Estudio hidropático del centro activo 169
7. Discusión 171

LA REGIÓN VARIABLE COMO PREDICTOR DE ESPECIFICIDAD DE SUSTRATO EN GLICOSILTRANSFERASAS GTA. 173

1. Introducción 175
2. Análisis estructural de GTAs 177
3. Evolución de la región variable respecto al resto de la secuencia 181
3.1 Familia GT2 181
3.2 Familia GT7 188
3.3 Familia GT8 193
3.4 Familias GT55, GT78 y GT81 196
4. Región variable como herramienta predictiva 202
4.1 Perfiles de Secuencias 202
4.2 Predicciones basadas en los perfiles HMM 206
5. Discusión 209

CONCLUSIONES 211

MÉTODOS 217

1. Modelado de la proteína MG517 de Mycoplasma genitalium 219
2. Estudio de la región C-terminal de la proteína MG517 224
3. Estudio del bucle catalítico en la proteína GpgS 228
4. Uso de la Región Variable como herramienta predictiva 233

BIBLIOGRAFÍA 235

ANEXOS 245

Anexo 1. Árbol filogenético de los micoplasmas y su relación con clostridium. 247
Anexo 2. Alineamiento de GTAs cristalizadas 1 248
Anexo 2. Alineamiento de GTAs cristalizadas 2 249
Anexo 3. Características de los cristales de proteínas GTA. 250
Anexo 4. Árbol estructural generado por POSA. 251
Anexo 5. Alineamientos para Modeller, para el modelado de MG517. 252
Anexo 6. Superposición por PROCHECK de los ángulos Ψ (phi) y Φ (psi) de los diferentes modelos en el diagrama de Ramachandran. 254
Anexo 7. Modelado de las estructuras 1 a 4. 255
Anexo 7. Modelado de las estructuras 1 a 4, continuación. 256
Anexo 8. Tipos de átomos, cargas y masas del UDPGlc. 257
Anexo 9. Gráfico de evolución DSSP de cada estructura a lo largo de la simulación MD. 258
Anexo 10. Representación SAP de las estructuras seleccionadas tras la MD. 259
Anexo 11. Resultados de I-TASSER para la secuencia completa de MG517. 260
Anexo 12. Comparativa entre el valor de los diedros Φ y Ψ para el cristal y el modelo generado. 261
Anexo 13. Comparación de los diedros del bucle RAHRN y residuos anejos entre el cristal, modelo, minimización de energía durante el equilibrado de la MD y MD. 262
Anexo 14: Meta4Y6Napo4. Proyecciones de energía: 264
Anexo 15: Meta4Y6NUDPGlc4. Proyecciones de energía: 266
Anexo 16: Meta4Y6NUDPGlc5. Proyecciones de energía: 268
Anexo 17. Poblaciones de diedros Arg256-His258 en las simulaciones metadinámicas. 270

ANEXO 18. ESTRUCTURAS GTA ANALIZADAS. 274
ANEXO 19. METODOLOGÍA DE IDENTIFICACIÓN DE LAS DIFERENTES RV. 278
ANEXO 20. LONGITUD DE CADA PERFIL, ANTES Y DESPUÉS DE LA BÚSQUEDA CONTRA CAZY. 281
ANEXO 21. OPTIMIZACIÓN DE RESULTADOS TRAS SER SOMETIDOS AL *PIPELINE* DE CAZY 282
ANEXO 22. GENERACIÓN DE ARCHIVOS GROMACS DESDE PDB 285

Scripts **287**

***SCRIPT* PHYLIP.SCR** 289
***SCRIPT* PHYLIP1.SCR** 290
***SCRIPT* ARBOL.SCR** 291
PARÁMETROS PHYLIP 292
***SCRIPT* UPDATE_SS.SH** 293
SCRIPT MONTAJENTER.SCR 294
SCRIPT MONTAJEHELICE.SCR 296
SCRIPT PRODUCCIÓN.SCR 297
SCRIPT MOLECULARDYNAMICS.SCR MD MODELOS MG517 299
SCRIPT METADINÁMICA ENLAZADO DE EJECUCIONES. HÉLICE ANFIPÁTICA INSERTA EN MEMBRANA. 300
***SCRIPT* EQUILIBRADO.SCR. MD CRISTAL 4DDZ** 301
SCRIPT EQUILIBRADO.SCR. MD MODELO 4Y6N 303
SCRIPT EXTRACCIÓN DE CÓDIGOS UNIPROT DE CAZY GTAS 304
SCRIPT EXTRACCIÓN DE SECUENCIAS FASTA UNIPROT DE CAZY GTAS 306
SCRIPT EXTRACCIÓN DE LA INFORMACIÓN DE TEXTO PARA CADA ENTRADA UNIPROT DE CAZY GTAS 308
SCRIPT EXTRACCIÓN DEL Nº EC Y OTRA INFORMACIÓN DE CADA ENTRADA UNIPROT DE CAZY GTAS 309
SCRIP PARA REALIZAR LA "PRUEBA DE SOLIDEZ" DE LOS PERFILES HMM 310

Inputs **311**

MODELADO MG517. 313
MODELADO HÉLICES 314
MODELADO GPGS. CRISTAL 4DDZ, GENERACIÓN DEL .INI 315
MODELADO GPGS. CRISTAL 4DDZ 316
MODELADO GPGS: CRISTAL 4DDZ. ALINEAMIENTO DE ENTRADA 317
MODELADO GPGS. CRISTAL 4Y6N 318
MODELADO DEL CRISTAL 4Y6N. ALINEAMIENTO DE ENTRADA 319
ENERGÍA DE MINIMIZACIÓN: NTER MG517 Y HÉLICES EN SOLVENTE ACUOSO 320
EQUILIBRADO AGUAS CON POSITION RESTRAINS: NTER MG517 Y HÉLICES EN SOLVENTE ACUOSO 321
EQUILIBRADO DE LA TEMPERATURA: NTER MG517 Y HÉLICES EN SOLVENTE ACUOSO 322
EQUILIBRADO DE LA PRESIÓN: NTER MG517 Y HÉLICES EN SOLVENTE ACUOSO 323
PRODUCCIÓN: NTER MG517 Y HÉLICES EN SOLVENTE ACUOSO 324
EQUILIBRADO DE LA TEMPERATURA: HÉLICE INSERTA EN MEMBRANA 325
EQUILIBRADO DE LA PRESIÓN1: HÉLICE INSERTA EN MEMBRANA 326
EQUILIBRADO DE LA PRESIÓN2: HÉLICE INSERTA EN MEMBRANA 327
PRODUCCIÓN: HÉLICE INSERTA EN MEMBRANA 328
ENERGÍA DE MINIMIZACIÓN. MODELOS 4DDZ 329
EQUILIBRADO DE LAS AGUAS. MODELO 4DDZ 330
EQUILIBRADO DE LA TEMPERATURA. MODELO 4DDZ 331
EQUILIBRADO DE LA PRESIÓN. MODELO 4DDZ 332
ENERGÍA DE MINIMIZACIÓN. MD MODELO 4Y6N. *CONJUGATED GRADIENTS* 333
ENERGÍA DE MINIMIZACIÓN. MD MODELO 4Y6N. *STEEPEST DESCENT* 334
EQUILIBRADO DE LAS AGUAS. MD MODELO 4Y6N. PROTEÍNA Y LIGANDOS FIJOS 335
EQUILIBRADO DE LAS AGUAS Y PROTEÍNA. MD MODELO 4Y6N. PROTEÍNA LIBRE, LIGANDOS FIJADOS 336
EQUILIBRADO DE LA TEMPERATURA. MD MODELO 4Y6N. 337
EQUILIBRADO DE LA PRESIÓN. MD MODELO 4Y6N 339
PRODUCCIÓN. MD MODELO 4Y6N 340
HÉLICE ANFIPÁTICA INSERTA EN MEMBRANA 341
HÉLICE ANFIPÁTICA NO-APOLAR INSERTA EN MEMBRANA 343
***CONSTRAINS* ÁTOMOS QUE INTERACCIONAN CON EL MG EN EL MODELO 4Y6N (GPGS TERNARIO)** 345
METAD4DDZAPOLA 346

METAD4Y6NAPO **347**
HILLS METAD4Y6N FES RESTRINGIDO **348**
META4Y6NAPO3 **349**
META4Y6NAPO4 **350**
***DOCKING* DAG SOBRE MODELOS DE MG517** **358**
***DOCKING* DAG SOBRE MODELOS DE MG517** **359**

ÍNDICE DE FIGURAS Y TABLAS

INTRODUCCIÓN

Figura 1. Polisacáridos y glicoconjugados. 21
Figura 2. Grupos aceptores. 22
Figura 3. Mecanismos de transferencia del azúcar. 23
Figura 4. Mecanismo *inverting*. Desplazamiento tipo SN2. 23
Figura 5. Mecanismo *retaining*. 24
Figura 6. Mecanismo *retaining*. Desplazamiento tipo S_Ni. 24
Figura 7. Plegamiento GTAs. 25
Figura 8. Modelado por homología. Zona segura. 31
Figura 9. Fuerzas de un campo de fuerzas empírico. 37
Figura 10. Ecuación para el cálculo de la energía potencial. 38
Figura 11. Pasos en una simulación MD. 38
Figura 12. Nodos de la RES. 39
Figura 13. Simulación Metadinámica. 42
Figura 14. Sesgo potencial sin BIAS-Exchange. 45
Figura 15. Sesgo potencial con BIAS-Exchange. 46
Tabla 1. Métodos de muestreo usados en docking. 33
Tabla 2. Función de puntuación de AutoDock. 33

MARCO DE LA TESIS

Figura 16. *Mycoplasma genitalium*. 51
Figura 17. Reacción catalizada por MG517. 52
Figura 18. *Mycobacterium tuberculosis*. 53
Figura 19. Biosíntesis de MGLP. 54

Figura 20. Filogenia de Mollicutes. 63
Figura 21. Filogenia de GT2 en Firmicutes y Tenericutes. 64
Figura 22. Superposición de GTAs. 67
Figura 23. Regiones N y C terminal de MG517. 67
Figura 24. Topología consenso de las GTAs. 68
Figura 25. Alineamientos de GTAs cristalizadas. 70
Figura 26. Dendrograma de GTAs cristalizadas. 71
Figura 27. Modelos de MG517. 74
Figura 28. Ubicación del dador en los modelos. 75
Figura 29. TLC de mutantes. 78
Figura 30. Convergencia de la región variable 83
Figura 31. Eventos de docking. 86
Tabla 3. GTAs cristalizadas (junio 2013). 65
Tabla 4. Servidores automáticos de modelado. 72
Tabla 5. Residuos cercanos al dador. 76
Tabla 6. Medida de la actividad de los mutantes. 79
Tabla 7. Resumen de las simulaciones. 83
Tabla 8. Fluctuación RMS durante la MD. 84
Tabla 9. Docking aceptor. 86

INTERACCIÓN CON LA MEMBRANA DE LA PROTEÍNA MG517, A TRAVÉS DE LA REGIÓN C-TERMINAL

Figura 32. Proteínas de membrana. 91
Figura 33. PsiPred MG517. 93
Figura 34. Estructuras Robetta. 96
Figura 35. Hidropatía MPEx MG517. 97
Figura 36. Combinación de predictores. 98
Figura 37. Hélices modeladas. 99
Figura 38. Modelado Hélice 4. 99
Figura 39. Evolución DSSP, hélices en solvente acuoso. 100
Figura 40. Sistemas con membrana. 102
Figura 41. Sistemas 1, 2 y 3. 103
Figura 42. Sistema 4 y 5. 104
Figura 43. Sistema 6. 105
Figura 44. Comparación hélice 4 en solvente y membrana. 106
Figura 45. Interacciones hélice-membrana. 106
Figura 46. Sistemas 6, 7 y 8. Orientación en membrana. 107
Figura 47. Rotación de la hélice. 108
Figura 48. Núcleo hidrofóbico de la hélice en el sistema 6. 108
Figura 49. Hélices mutantes. 109
Figura 50. Rotación en mutantes. 110
Figura 51. Hélice truncada. 110
Figura 52. CVs simulación Metadinámica. 111
Figura 53. Evolución de las CVs. 112
Figura 54. FES de las simulaciones. 113
Figura 55. Posiciones de la hélice no-apolar. 114

Figura 56. Bucle RAHRN en el cristal 4DDZ. 120
Figura 57. Cadenas laterales del bucle RAHRN en el cristal 4DDZ. 120
Figura 58. Bucle RAHRN en el cristal 4Y6N. 121
Figura 59. Modelo del dímero GpgS a partir del cristal 4Y6N. 124
Figura 60. Conformaciones del bucle en el cristal 4DDZ. 125
Figura 61. Diedros φ y ψ del bucle RAHRN y homólogos. 128
Figura 62. CVs del bucle RAHRN. 129
Figura 63. Diedro Ala257-Distancia. 130
Figura 64. Diedro His258-Distancia. 130
Figura 65. Diedro Arg256-Distancia. 131
Figura 66. Distancia MD4DDZLA. 133
Figura 67. Distancia-Diedro Ala257 MD4DDZApoLA. 133
Figura 68. Distancia MD4DDZApoLI. 134
Figura 69. Distancia-Diedro Ala257 MD4DDZApoLI. 134
Figura 70. Evolución parámetros estructurales MD4DDZApoLA. 136
Figura 71. Evolución parámetros estructurales MD4DDZApoLI. 136
Figura 72. Distancia MD4Y6NApo1. 137
Figura 73. Evolución parámetros estructurales MD4Y6NApo1. 138
Figura 74. Evolución parámetros estructurales MD4Y6NApo2-10, monómero A. 139
Figura 75. Evolución parámetros estructurales MD4Y6NApo2-10, monómero B. 139
Figura 76. FES MetaD4DDZ. 141
Figura 77. FES MetaD4Y6NApo1. 142
Figura 78. FES MetaD4Y6NApo2. 143
Figura 79. FES MetaD4Y6NApo3. 144
Figura 80. Proyección FES CVs Ala257-Distancia. 146
Figura 81. Proyección FES CVs His258-Arg256. 147
Figura 82. Distancia MD4Y6NUDPGlc·PGA. 148
Figura 83. Distancia MD4Y6NUDPGlc·PGA, monómero B. Distancia His258-Mg. 150
Figura 84. Diedro His258 y distancia MD4Y6NUDPGlc·PGA, monómero B. 150
Figura 85. Interacciones UDPGlc. 151
Figura 86. Conformaciones del UDPGlc. 152
Figura 87. Distancia UDPGlc-Asp134. 153
Figura 88. Energía de afinidad entre UDPGlc y GpgS en la simulación. 156
Figura 89. FES MetaD4Y6NUDPGlc·PGA. 157
Figura 90. FES MetaD4Y6NUDPGlcFree. 158
Figura 91. FES MetaD4Y6NUDPGlcFix. 159
Figura 92. FES MetaD4Y6NPGA. 160
Figura 93. FES MetaD4Y6NUDPGlc3. 162
Figura 94. Comparación FES MetaD4Y6NApo3 y MetaD4Y6NUDPGlc3. 162
Figura 95. Comparación proyección FES CVs His258-Arg256 BIAS-Exchange Apo y con UDPGlc. 162
Figura 96. Proyección FES Distancia-UDPGlc diedro. 166

Figura 97. Esfera de hidratación entre ligandos. 166
Figura 98. Moléculas de agua durante la simulación. 167
Tabla 10. Cristales de GpgS. 120
Tabla 11. Parámetros de GpgS y proteínas homólogas. 127
Tabla 12. Metaestados BIAS-Exchange Apo. 145
Tabla 13. Distancia de interacción con el Mg. 149
Tabla 14. Resultados del docking. 154
Tabla 15. Metaestados BIAS-Exchange MetaD4Y6NUDPGlc4 164
Tabla 16. Metaestados BIAS-Exchange MetaD4Y6NUDPGlc5 165

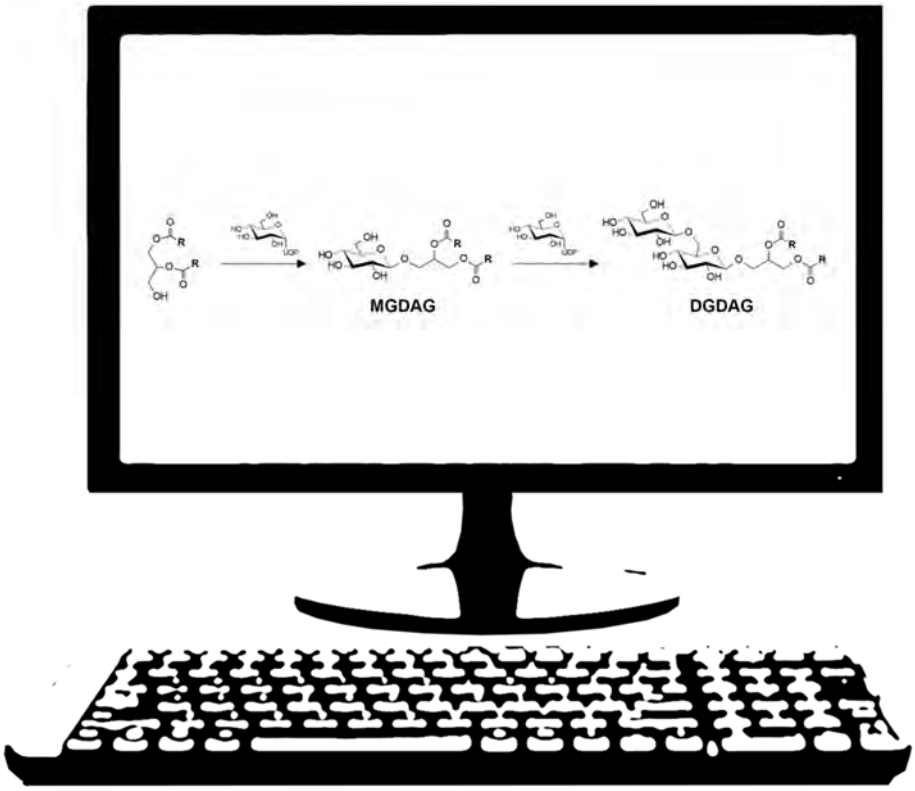
Figura 99. Topología consenso en las GTAs. 175
Figura 100 Superposición estructuras GTA. 176
Figura 101. Árbol, familia GT2_P. 180
Figura 102. Alineamiento de la RV de hialurano sintasa. 181
Figura 103. Logo de la RV de hialurano sintasa. 181
Figura 104. Alineamiento de la RV de celulosa sintasa. 182
Figura 105. Logo de la RV de celulosa sintasa. 182
Figura 106. Logos para la mananosintasa/glucomanano sintasa y xilano sintasa. 183
Figura 107. Árbol, familia GT2_P, sin RV y solo la RV. 184
Figura 108. Árbol, familia GT2_P, diferentes segmentos. 185
Figura 109. Alineamiento de GT7. 187
Figura 110. Superposición de tres estructuras de la familia GT7. 188
Figura 111. Árbol, familia GT7. 189
Figura 112. Árbol GT7 sin RV. 189
Figura 113. Árbol GT7, solo RV y fragmento. 190
Figura 114. Alineamiento de GT8. 192
Figura 115. Árbol, familia GT8. 193
Figura 116. Superposición de GT55, GT78 y GT81. 195
Figura 117. Alineamiento de GT7. 196
Figura 118. Árbol familias GT55, GT78 y GT81. 197
Figura 119. Árbol familias GT55, GT78 y GT81, sin RV. 198
Figura 120. Árbol familias GT55, GT78 y GT81, solo la RV. 198
Figura 121. Histograma de precisión de los perfiles. 205
Tabla 17. Estructuras con ligando y relación con RV. 178
Tabla 18. Grupos iniciales según el EC. 200
Tabla 19. Grupos finales de perfiles HMM. 203
Tabla 20. PerfilesHMM válidos como predictores 206

MÉTODOS

Figura 122. Preparación de archivos para UDPGlc. 219
Figura 123. DAG y DPG. Aceptores de MG517. 221
Tabla 21. Agrupamiento y selección de estructuras MG517. 220

ABREVIATURAS

CV	Variable colectiva
DAG	Diacilglicerol
DolP	Dolicol fosfato
Fuc	Fucosa
GA	Glicerato
Gal	Galactosa
GalA	Galacturónico
GalNAc	N-Acetil Galactosa
Glc	Glucosa
Glc_lipopolis	Glucosa unida a un lipopolisacárido
GlcA	Glucurónico
Glc-EGFlike	Glucosa unida a repeticiones EGF
GlcNAc	N-Acetil Glucosa
Glgn	Glicogenina
GTA	Glicosiltransferasa con plegamiento GTA
HMM	Hidden Markov Model
Ines	Inespecífico
Man	Manosa
MD	Dinámica Molecular
MetaD	Metadinámica
MioIno	Mioinositol
NeuNAc	N-Acetil Neuramínico
PAG	Fosfoglicerato
Rha	Ramnosa
SphngNAc	N-Acetil Esfingosina
UndP	Undecaprenol
Xyl	Xilosa



INTRODUCCIÓN A LAS GLICOSILTRANSFERASAS Y BIOLOGÍA COMPUTACIONAL

1. Glicosiltransferasas

Las glicosiltransferasas (GTs) son proteínas que transfieren azúcares, más concretamente, catalizan la formación del enlace glicosídico entre un azúcar y una molécula aceptora, y forman uno de los grupos de proteínas más numerosos de la naturaleza. Son enzimas ubicuas, responsables de toda la diversidad y complejidad de oligosacáridos y glicoconjugados encontrados en el mundo vivo, por lo que sus funciones pueden considerarse fundamentales para todos los organismos.

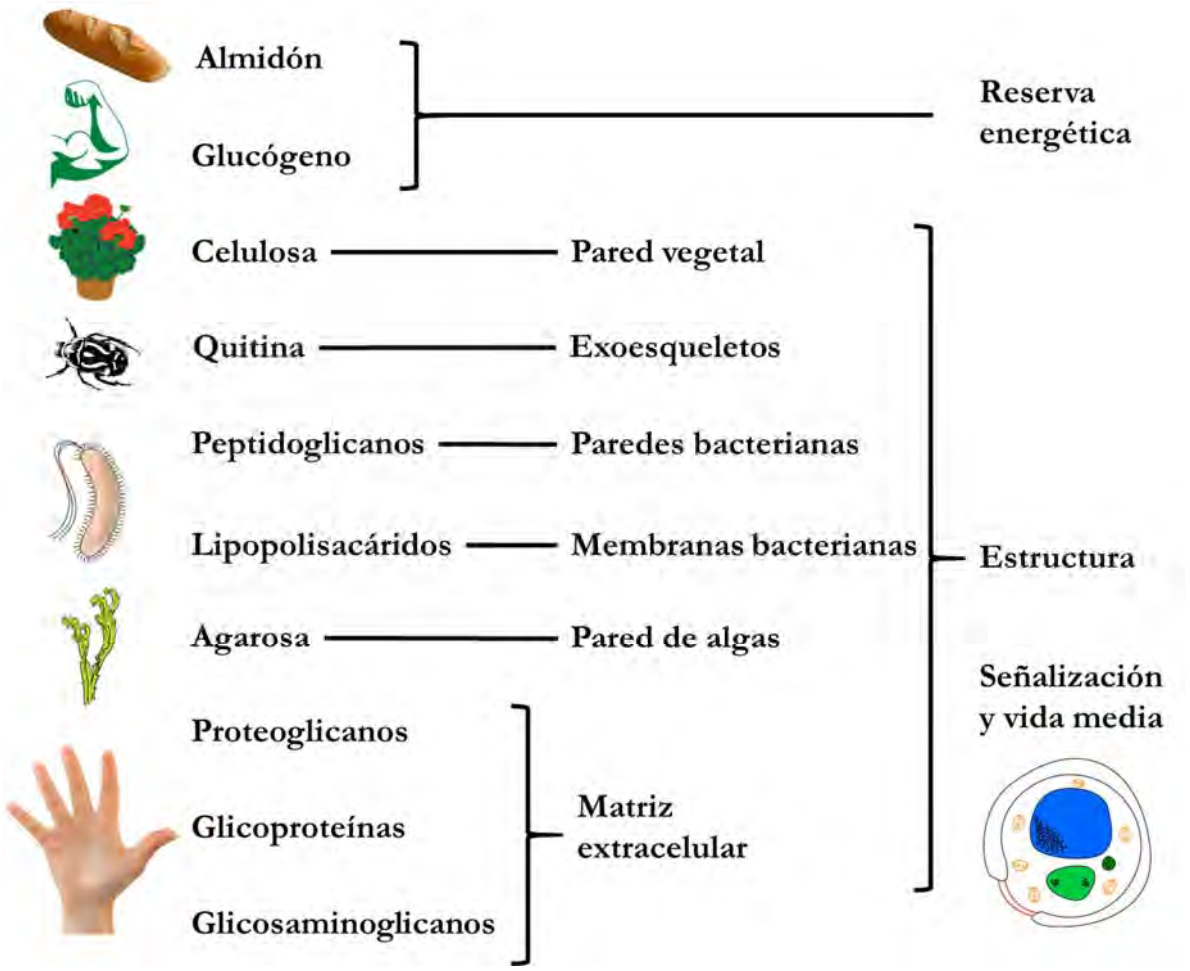
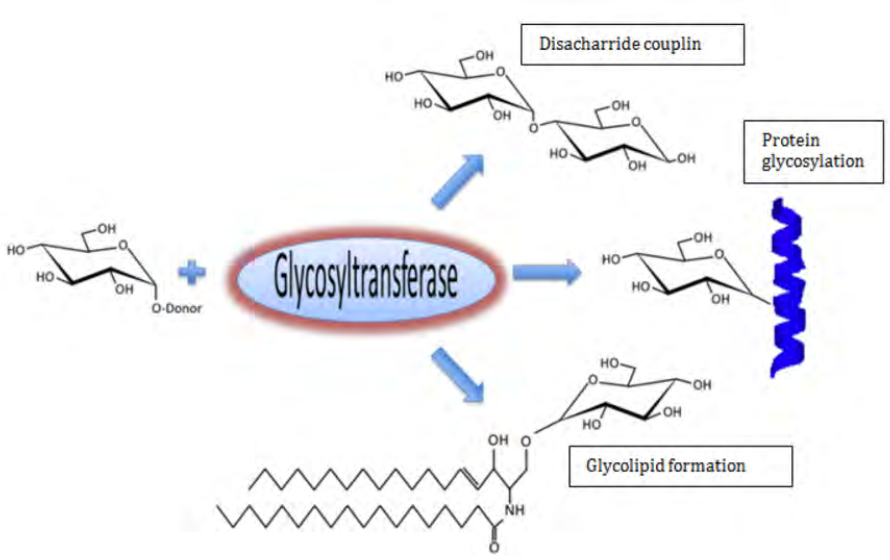


Figura 1. Algunos de los polisacáridos (homopolisacáridos y heteropolisacáridos) y glicoconjugados presentes en la naturaleza que deben su existencia a las glicosiltransferasas.

Esta diversidad química requiere que las enzimas que catalizan su síntesis, degradación y modificación sean muy específicos. La molécula aceptora del azúcar en GTs puede ser un gran número de biomoléculas diferentes, por ejemplo, otro azúcar, proteínas, lípidos y otras moléculas pequeñas, como ácidos nucleicos o antibióticos (figura 2).



Fuente: http://www.sbhsciences.com/Glycosyltransferase.asp

Figura 2. Diferentes grupos aceptores del azúcar.

Los azúcares transferidos están generalmente en forma de nucleósidos activados (dador), aunque también pueden llegar a ser fosfolípidos y fosfatos no sustituidos[1]. A las GTs y otras enzimas dependientes de nucleótidos de azúcar, se les suele llamar enzimas de tipo Leloir, en honor al premio Nobel en Química (1970), Luis F. Leloir quien descubrió la primera de estas moléculas.

"Las GTs bacterianas están vinculadas a la biosíntesis de paredes celulares y de importantes azúcares presentes en sus superficies, como los lipopolisacáridos (LPS) de las bacterias Gram negativas, polisacáridos capsulares (CPS) o los lipoarabinomananos de micobacterias. En el caso de las bacterias patógenas, muchos de estos azúcares son los responsables de la virulencia[2]. Algunas glicosiltransferasas bacterianas se han utilizado para la síntesis a gran escala de oligosacáridos que podrían tener interés farmacológico e industrial[3].

Las GTs de plantas participan en la biosíntesis de las complejas paredes celulares[4], además, en la glicosilación de pequeñas moléculas con un rol esencial en la modulación de la germinación, inactivación de fitohormonas y otros procesos fisiológicos[5].

Las GTs de animales se vinculan con la biosíntesis de azúcares de superficie que envuelven a la célula formando una capa llamada glicocálix. Estos azúcares sirven de sitios específicos de unión con otros receptores celulares, bacterias[6], virus[7], parásitos, toxinas y hormonas, además de definir los tipos sanguíneos[8]. Por otra parte, son importantes en las interacciones celulares durante la fertilización[9], desarrollo, diferenciación, transformación oncogénica[10], inflamación[11], defensa inmunológica[12] así como muchas enfermedades."[a]

[a] Tesis doctoral: Javier Antonino Linares Pastén.

1.2 Transferencia del azúcar: Mecanismos.

En todas las GTs, la formación del enlace glicosídico entre dador y aceptor puede ocurrir con retención o inversión de la configuración del carbono anomérico del dador (figura 3).

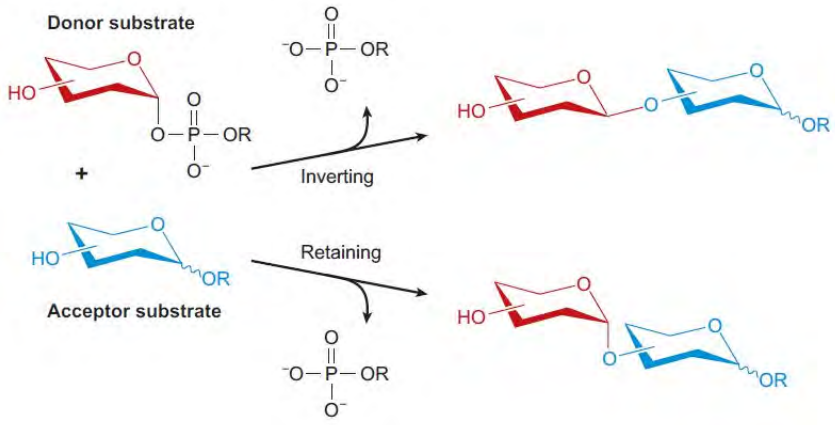


Figura 3. El azúcar se transfiere al oxígeno nucleofílico de un hidroxilo sustituyente en el aceptor, formando un enlace *O*-glicosídico, también puede hacerlo a un nitrógeno nucleofílico (enlace *N*-glicosídico) o sulfuro nucleofílico (enlace *S*-glicosídico, con la formación de compuestos tioglicósidos en plantas) y carbono nucleofílico (antibióticos *C*-glicosídicos).

El mecanismo tipo *inverting*, con inversión de la configuración, consiste en un desplazamiento directo del tipo S_N2: un residuo del centro activo hace de base catalítica, desprotonando el grupo nucleófilo del aceptor y facilitando el desplazamiento tipo S_N2 (figura 4). Sin embargo, la ausencia de esa base catalítica en algunas familias evidencia que deben existir otros mecanismos alternativos.

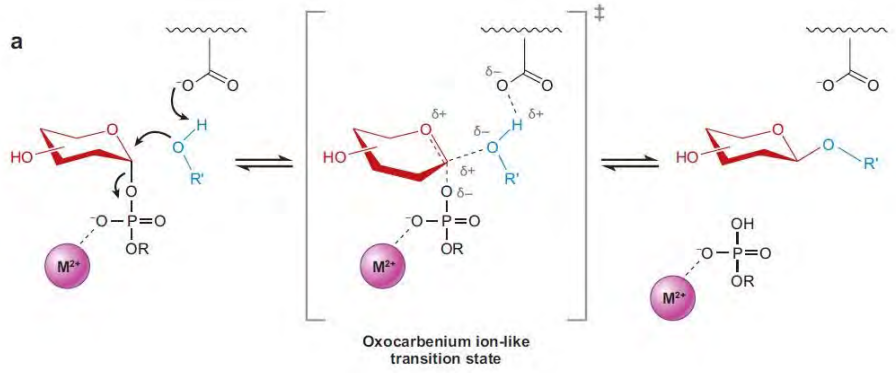


Figura 4. Desplazamiento tipo S_N2, configuración anomérica invertida, vía un único estado de transición del tipo oxocarbocatión.

Para las GTs tipo *retaining*, con retención de la configuración, el mecanismo ha sido motivo de controversia. Por analogía a las glicosilhidrolasas tipo *retaining* se propuso un mecanismo de doble desplazamiento y un intermediario glicosil-enzima, unido covalentemente, requiriéndose un nucleófilo adecuadamente colocado dentro del centro activo. El grupo saliente difosfato probablemente actúa como base catalítica, activando el grupo hidroxilo del aceptor por el ataque nucleofílico[13] (figura 5).

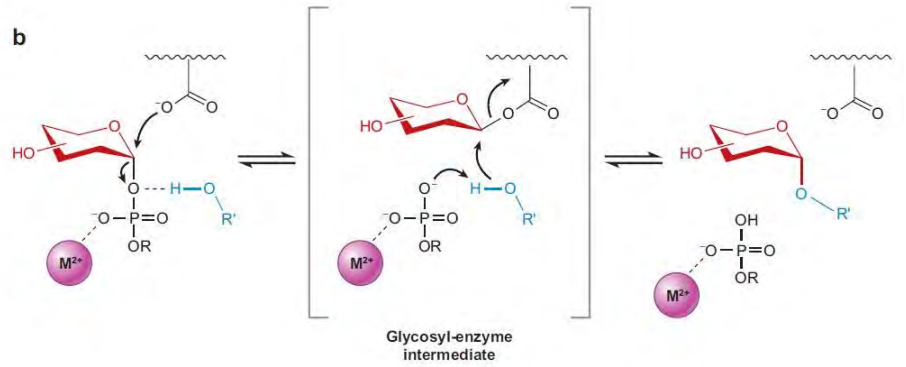


Figura 5. Mecanismo de doble desplazamiento propuesto para GTs con retención de la configuración que requieren la formación de un enlace covalente glicosil-enzima intermedio.

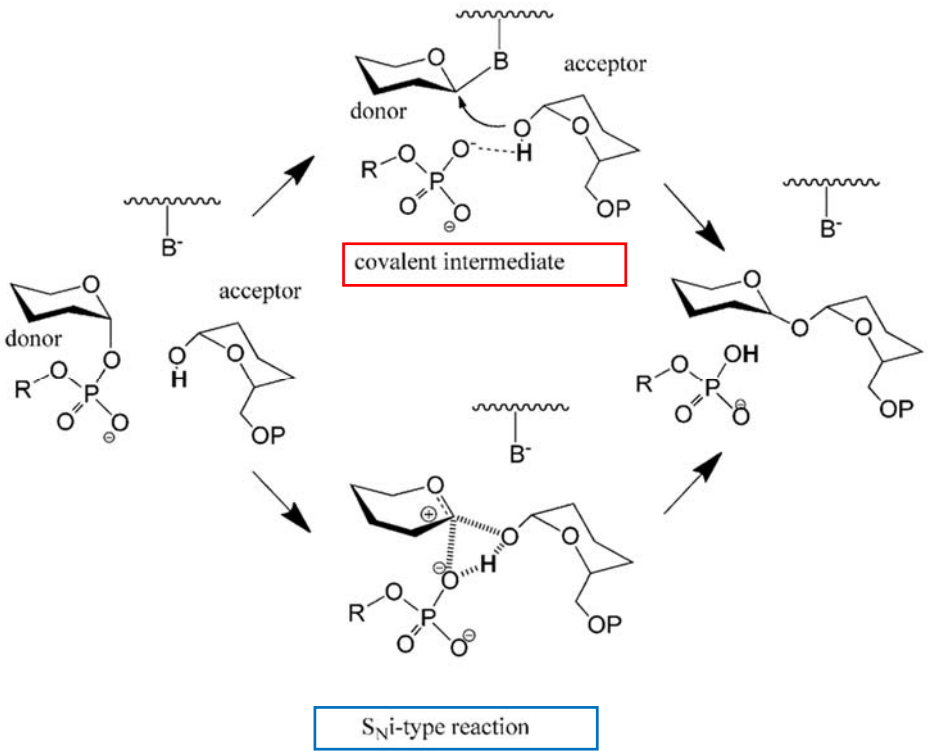


Figura 6. Posibles mecanismos de reacción propuestos para GTs tipo *retaining*. Los enlaces parcialmente rotos o formados vienen indicados por una línea discontinua[14].

Cuando se resolvió la primera estructura tridimensional de una GT con retención de la configuración, LgtC de *Neiseria meningitidis*, no se pudo observar ningún tipo de aminoácido con un grupo nucleófilo (Asp o Glu) cerca del carbono anomérico. La falta de evidencias sobre el mecanismo de doble desplazamiento, impulsó a diversos autores a proponer un nuevo mecanismo, por el que la reacción tiene lugar vía un único desplazamiento por una cara, y al que se llama **mecanismo tipo S_Ni**. En este mecanismo el hidroxilo nucleofílico del aceptor ataca el carbono anomérico, por la misma cara por la que sale el grupo saliente, nucleósido, dando lugar a un estado de transición del tipo oxocarbocatiónico (figura 6). **Este último mecanismo, sin la existencia de un intermediario oxocarbónico ha sido confirmado, primero por técnicas computacionales[14] y recientemente por métodos también experimentales[15]**.

1.3 Plegamiento

La enorme mayoría de GTs con estructura resuelta hasta el momento, presentan dos arquitecturas generales, llamadas **GT-A** y **GT-B**, aunque al menos otros dos tipos han sido también identificados. Además, diversos análisis han desvelado que muchas de las familias que aún no han sido caracterizadas, tendrán alguno de esos dos plegamientos mayoritarios[13] (figura 7).

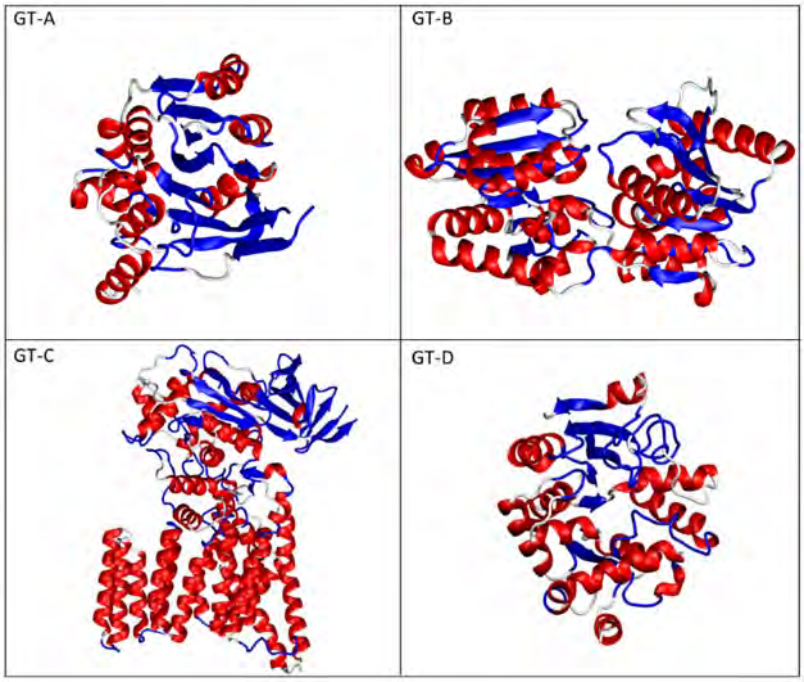


Figura 7. Diferentes tipos de plegamiento para las GTs. La mayoría de ellas se agrupan en las superfamilias GTA y GTB. GTA: SpsA de *Bacillus subtilis*. PDB 1QG8. GTB: WaaG de *Escherichia coli*. PDB 2IV7. También se han encontrado otras arquitecturas muy minoritarias como GTC y GTD. GTC: PglB de *Campylobacter lari*. PDB 3RCE. GTD: Transferasa de función desconocida de *Streptococcus parasanguinis*. PDB 4PFX.

El plegamiento GTA se describió por primera vez en el enzima con retención de la configuración SpsA, de *Bacillus subtilis*, del que se resolvió la estructura tridimensional[16]. Constituida por una lámina β cercada por hélices α en los costados, la arquitectura global de del plegamiento (GT-A) es similar a la de dos dominios tipo Rossmann adyacentes. Dos dominios α/ β /α (estrictamente el dominio Rossmann es β /α/ β) fuertemente asociados, con una hoja β central y continua. Por ello, a veces se describe al plegamiento GTA, como un plegamiento con un único dominio.

La mayoría de enzimas GTAs poseen el motivo Asp-X-Asp (conocido como DXD) en el cual los carboxilatos se coordinan con un catión divalente. Aunque este motivo está descrito frecuentemente como una característica determinante de las GTAs existen ejemplos actuales de enzimas con este plegamiento, que no tienen el motivo DXD, si bien siempre existe un aminoácido aniónico (Glu o Asp) en su lugar.

La arquitectura de los enzimas con plegamiento GTB consiste también de dos dominios tipo Rossmann. Pero en este caso los dominios están menos estrechamente relacionados y se asocian a través del centro activo, que se encuentra en el interior del hueco resultante (figura 7).

Existen al menos dos tipos más de plegamiento para las GTs; uno es el de las proteínas de la superfamilia **GTC**, que son integrales de membrana y presentan un motivo DXD modificado. La mayoría además utilizan fosfolípidos azucarados como dadores[17]. A esta superfamilia pertenecen las oligosacariltransferasas (OST), proteínas que realizan la trasferencia del azúcar, mediante enlace *N*-glicosídico, a la Asn de otras proteínas con el motivo Asn-X-Ser/Thr y que están relacionadas con el plegamiento y control de calidad de estas proteínas, así como el desarrollo del organismo o interacciones con patógenos[18]. Por último, el plegamiento **GTD** ha sido propuesto para una proteína relacionada con la glicosidación de adhesinas en *Streptococcus parasanguinis*[19] que difiere estructuralmente de los anteriores y cuya función es todavía desconocida.

1.4 Clasificación

Las glicosiltransferasas han sido clasificadas por homología de secuencia en 98 familias (Abril 2016), por la *Carbohydrate Active enZyme database* (CAZy)[20], base de datos que mantiene una clasificación de enzimas activos en carbohidratos (glicosil hidrolasas, glicosil transferasas, carbohidrato estearasas, polisacárido liasas, módulos de unión a carbohidrato y otros funcionalidades auxiliares). CAZy provee de una poderosa herramienta predictiva, ya que la estructura y el mecanismo de acción son invariables dentro de la mayoría de familias. Por tanto, donde esta estructura y mecanismo han sido anotados, muchas asunciones son aplicables para otros miembros de la familia. La especificidad de sustrato, sin embargo, es más difícil de predecir y requiere la caracterización experimental de GTs individualmente. La determinación tanto del azúcar dador como del aceptor de una GT con función desconocida, puede ser un reto y es una de las razones por las que existe un número significativamente menor de GTs bien caracterizadas que, por ejemplo, para las glicosil hidrolasas.

La superfamilia de proteínas con plegamiento GTA está formada actualmente por casi 100000 secuencias diferentes divididas en 17 familias (GT2, GT6, GT7, GT8, GT12, GT13, GT15, GT21, GT24, GT27, GT43, GT55, GT64, GT78, GT81, GT82 y GT84). Los mecanismos tipo *inverting* o *retaining* se reparten entre estas familias, si bien cada familia está integrada por proteínas con un solo tipo de mecanismo. La gran conservación estructural del plegamiento GTA contrasta con una homología de secuencia muy baja, que se traduce en gran variabilidad de la secuencia de aminoácidos entre las diferentes familias. Esta variabilidad es acorde con la diversidad y complejidad de oligosacáridos y glicoconjugados que estos enzimas son capaces de generar.

La superfamilia de proteínas con plegamiento GTA se ha estudiado en este trabajo, utilizando un enfoque teórico, también denominado *in silico*, ya que está basado íntegramente en simulaciones y herramientas computacionales, enmarcadas en un área relativamente nueva de la Ciencia denominada "Biología Computacional". Los principios fundamentales en los que se basan estas herramientas se desarrollan en el siguiente capítulo.

2. Biología computacional

El trabajo en un laboratorio requiere de tiempo y recursos –humanos y tecnológicos–, con un coste económico, que en la mayoría de las ocasiones supone el cuello de botella de cualquier proyecto de investigación. Por ello es importante a la hora de iniciar uno, conocer las posibles vías muertas –que no rendirán resultados positivos e implicarán un derroche de recursos– o la ruta más directa hacia un determinado objetivo, con el consecuente ahorro de tiempo y coste para el laboratorio. Es decir, obtener una predicción del resultado antes de realizar el experimento.

"**Uno de los métodos más interesantes para predecir el futuro es el empleo de la simulación**: realizamos un modelo que no abarque la totalidad de la realidad, sino un restringido juego de entidades que pensamos pueden ser relevantes para el resultado final"[b].

Robert E. Shannon define la simulación como: "**El proceso de diseñar un modelo de un sistema real y llevar a término experiencias con él, con la finalidad de comprender el comportamiento del sistema o evaluar nuevas estrategias para su funcionamiento**"[c]. Por supuesto, ninguna cantidad de estos ejercicios de simulación puede predecir exactamente lo que sucederá en realidad, pero una buena simulación es enormemente preferible a un ensayo efectuado a ciegas. Por todo esto la simulación podría ser calificada "como un procedimiento experimental indirecto" [b].

Las simulaciones informáticas comenzaron a desarrollarse a la vez que lo hacían los ordenadores y la tecnología informática. El primer despliegue a gran escala de una simulación informática fue en el Proyecto Manhattan, durante la Segunda Guerra Mundial, para recrear una detonación nuclear. Desde entonces, junto con el desarrollo de mayor potencia computacional, las simulaciones han ido evolucionando desde útiles a indispensables, en ramas del conocimiento como la Biología, la Química, la Medicina, la Economía o el Medio ambiente.

En el campo de la Bioingeniería, el uso de simulaciones informáticas está ampliamente extendido; desde el mundo de los bioprocesos, donde se utilizan modelos cinéticos de consumo y obtención de energía, que se traducen en la producción eficiente de biomasa y metabolitos de interés por parte de organismos vivos, hasta el universo de la química cuántica, en donde se predice la rotura y formación de enlaces entre átomos a nivel cuántico, estableciendo las rutas termodinámicas de una reacción química. En algunos casos, las simulaciones suponen la estrategia más eficiente en el abordaje de un problema, en otros el único posible.

Esta tesis se sitúa en el nivel proteico. En las simulaciones realizadas no se llega nunca a descender al campo cuántico y por tanto no se contempla la rotura y formación de enlaces covalentes entre átomos. Las relaciones entre metabolitos y orgánulos intracelulares no es tratada de forma global y holística en la simulación, sino uno a uno, entre macromoléculas, como proteína y ligando o proteína y membrana, encuadrando este trabajo en el campo de la Biología estructural de proteínas y Biofísica de proteínas, al que el uso combinado de técnicas computacionales y de gestión de la

[b] "El gen egoísta". Richard Dawkins.
[c] "Simulation Modelling: The Art and Science". Robert E. Shannon

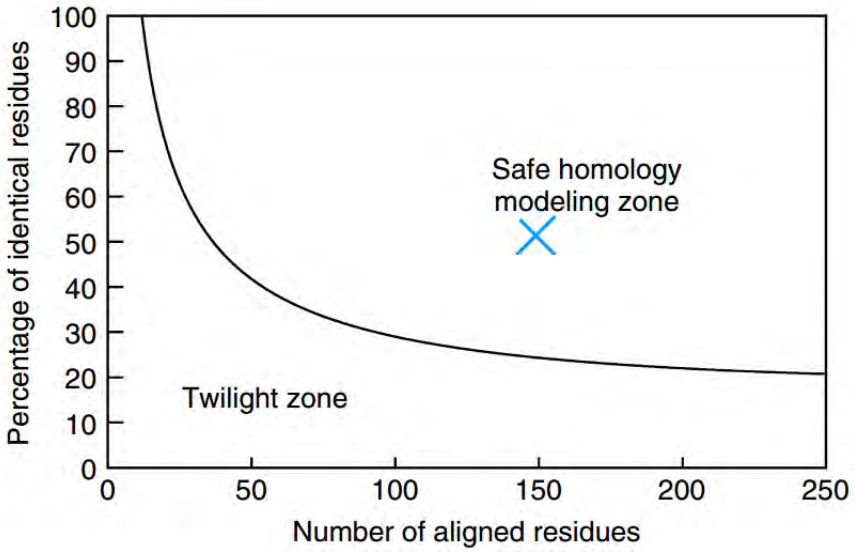
información lo sitúa también y de forma más certera en la Biología computacional, donde se relacionan, siguiendo el título de esta tesis, los tres niveles en el que se ha trabajado: **Secuencia** (estructura primaria), **Estructura** (estructuras secundaria a cuaternaria) y **Función** (información experimental e inferida).

Nota: Aprovecho para incluir el debate aun no concluso entre los términos Bioinformática y Biología computacional. Una distinción simple de ambas es que los bioinformáticos desarrollan *software* que luego los biólogos computacionales utilizan. Pero la diferencia es más sutil a la par que amplia. Para otros autores, ambas son áreas de estudio interdisciplinarias con una estrecha relación con la Biología. La Bioinformática sería el uso y aplicación de la tecnología de la información y ciencias de la computación al campo de la Biología Molecular, una disciplina muy práctica cuyo propósito es el desarrollo de soluciones informáticas de diversas clases para biólogos. La Biología Computacional sería una disciplina más teórica y mucho más basada en el uso de las Matemática y Estadísticas, con un grado mayor de entendimiento de los escenarios y procesos biológicos que también implica el desarrollo de algoritmos y modelos matemáticos. Como vemos, en realidad ambas son áreas muy interconectadas con un alto grado de solapamiento entre ellas. Basado en estas definiciones, este trabajo se sitúa en un área extensa (aunque enmarcada en el terreno estructural) que relaciona ambas disciplinas, utilizando la estadística, pero sin desarrollar nuevos algoritmos, aplicando soluciones informáticas, pero con un amplio conocimiento de los procesos biológicos, en un campo que bien podríamos llamar Bioinformática computacional.

Este trabajo se ubica íntegramente en el ámbito computacional debido al uso intensivo de varios métodos de simulación que requieren gran capacidad de cálculo, como son la "Dinámica Molecular" y "Metadinámica". Otras técnicas que se han empleado, también recurren al cálculo computacional pero no necesitan tantos recursos como las anteriores, es el caso del "Modelado por homología" y el "*Docking*". Los fundamentos de cada uno de estos métodos van a ser desarrollados a continuación.

2.1 Modelado por homología

El esclarecimiento de la estructura de una proteína a veces puede ser muy difícil; debido a complicaciones en la obtención de suficiente cantidad de proteína (por el clonaje, expresión y purificación de cantidades en torno al miligramo), dificultades asociadas al proceso de cristalización o también de la resolución de la proteína ya cristalizada. En este contexto, métodos que proponen la predicción de la estructura han ganado mucho terreno. **El modelado por homología**, a veces también denominado "modelado comparado", **es la construcción de un modelo 3D a resolución atómica de una proteína "diana", utilizando su secuencia de aminoácidos y la estructura tridimensional de una proteína homóloga, la "plantilla"**. Esta técnica se basa en la identificación de una o más estructuras proteicas conocidas, que probablemente se parecen a la estructura de la secuencia diana (*query)* y la realización de un alineamiento que superponga los residuos de la secuencia diana con los de la plantilla (*template*).



Fuente: "HOMOLOGY MODELING". Elmar Krieger, Sander B. Nabuurs, and Gert Vriend

Figura 8. Zonas en el alineamiento de secuencias. Se garantiza que dos secuencias contendrán la misma estructura si su longitud y el porcentaje de identidad de secuencia están dentro de la región marcada como segura (*safe*). El ejemplo de dos secuencias con 150 $\overline{aa}$ y 50 % de identidad se muestra con una X azul.

Se ha comprobado que la estructura está más conservada que la secuencia entre proteínas homólogas, aunque cuando la identidad de secuencia cae por debajo del 20 % la estructura puede ser muy diferente[21,22] (figura 8). Por tanto, proteínas evolutivamente relacionadas tienen secuencias similares y estructuras similares. También se ha comprobado que la estructura proteica está evolutivamente más conservada de lo que se esperaría basándola solo en la conservación de secuencia[23]. El alineamiento de secuencias y la estructura de la "plantilla" son entonces usados para generar un modelo estructural de la "diana".

La calidad del modelado por homología es dependiente de la calidad del alineamiento de secuencia y de la estructura plantilla. El abordaje puede ser complicado por la presencia de huecos (*gaps*) en el alineamiento, que indican una región estructural presente en la diana, pero no en la plantilla, también por la existencia de *gaps* en la plantilla, que provienen de zonas de pobre resolución en el procedimiento experimental para resolver la estructura (generalmente cristalografía por rayos X). La calidad del modelo se reduce cuando baja la identidad de secuencia: un modelo típico tiene ≈ 1-2 Å de RMSD entre los Cα, con una identidad de secuencia de un 70 % pero puede llegar a ser de ≈ 2-4 Å de RMSD con una identidad de solo el 25 %. Los errores son significativamente más altos para los bucles, donde la secuencia entre la diana y la plantilla podría ser completamente diferente. Al disminuir la identidad, se incrementan los errores en la posición y empaquetamiento de las cadenas laterales en la estructura modelada, cuya variación se sugiere como la principal razón para un modelado de pobre calidad[24]. Tomados en conjunto, todos estos "problemas en el modelado" son significativos e impiden el uso de modelos realizados por homología, para propósitos que requieren resolución atómica, como el diseño de fármacos o la predicción de interacciones proteína-proteína; incluso la estructura ternaria de una proteína puede ser difícil de predecir a partir del modelado por homología de sus subunidades. Sin embargo, esta técnica puede ser muy útil para obtener conclusiones "cualitativas" sobre la bioquímica de la secuencia diana; el modelado por homología dota al investigador de estructuras de "baja resolución" con las que formular hipótesis, sobre por qué ciertos residuos se hallan conservados, especialmente útil para experimentos de mutagénesis dirigida. En resumen, los problemas de modelado se reducen en realidad a encontrar la plantilla adecuada.

Los pasos a realizar en el modelado por homología son los siguientes:

1. Identificación de la plantilla.
2. Alineamiento de secuencia y corrección del alineamiento.
3. Generación de la cadena principal.
4. Modelado de bucles.
5. Modelado de las cadenas laterales y optimización.
6. Optimización del modelo general.
7. Validación del modelo.

Dependiendo del grado de conservación de secuencia, el modelado puede ser más o menos fácil, a veces incluso es necesario utilizar más de una plantilla para un mismo modelo.

La dificultad en la realización de los modelos en esta tesis, será determinante para la elección de la herramienta de modelado. Existen numerosos servidores automáticos para esto: Phyre[25], I-tasser[26], ROBETTA[27], Swiss Model Server[28], HHPred[29] … **Aquí hemos trabajado con una aplicación instalada en local: Modeller v.9.8[30], ya que permite una elección manual de todos y cada uno de los pasos del modelado y es nuestro interés mantener el control sobre el mismo.**

2.2 Docking (Acoplamiento molecular)

En el campo del Modelado molecular, el *Docking* es un método que predice la conformación preferida de una molécula al unirse a otra, para formar un complejo estable[31]. El conocimiento de esta orientación puede ser utilizado para predecir la fuerza de asociación o afinidad entre estas dos moléculas.

Con esta técnica se puede modelar la interacción entre una molécula pequeña y una proteína a nivel atómico, lo que permite caracterizar el comportamiento de pequeñas moléculas en la zona del centro activo de proteínas diana, así como elucidar los procesos bioquímicos fundamentales[32]. El proceso de *docking* involucra dos pasos básicos:

1. Las conformaciones del ligando, su posición y orientación dentro de cada sitio (a lo que usualmente se refiere como *pose*).

2. La clasificación de cada conformación mediante la medida de la afinidad de unión.

Estos dos pasos están separados y relacionados a métodos de muestreo (tabla 1) y tablas de puntuación respectivamente (tabla 2).

Algoritmo	**Característica**
Algoritmos de emparejamiento.	Basados en geometría, adecuado para *virtual screening* y base de datos enriquecidas por su alta velocidad.
Construcción creciente.	Basada en fragmentos y *docking* incremental.
MCSS.	Método basado en fragmentos para el diseño *de novo*.
LUDI.	Método basado en fragmentos para el diseño *de novo*.
Monte Carlo.	Búsqueda estocástica.
Algoritmo genético.	Búsqueda estocástica.
Dinámica Molecular.	Para un mejor refinado después del *docking*.

Tabla 1. Algunos métodos de muestreo usados en el *docking*[33].

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + W_{s}$$
$$\sum_{i,j} (S_i V_j + S_j V_i) \, e^{\left(-r_{ij}^2 / 2\sigma^2\right)}$$

Para dos átomos i, j, la energía atómica por parejas se evalúa por la suma de van der Waals, puentes de hidrógeno, energía electrostática y desolvatación. W son los factores ponderados para calibrar la energía libre empírica.

Tabla 2. Fórmula utilizada en la función de puntuación de AutoDock[34].

Conocer la localización del sitio de unión antes de realizar el *docking* incrementa significativamente la eficiencia del método, como de hecho ocurre en la mayoría de casos. Esta información es posible obtenerla por comparación de la proteína diana, con otras proteínas que comparten una función similar o cocristalizadas con otros ligandos. En ausencia de este conocimiento, existen programas de detección de cavidades o servidores *online*, como GRID[35,36], POCKET[37], SurfNet[38,39], PASS[40] y MMC[41]. Aquel proceso de *docking* realizado con ninguna asunción con respecto al centro activo es llamado "*docking* ciego".

Los primeros mecanismos de unión ligando-receptor se basaban en la teoría "llave-cerradura" propuesta por Fischer[42], en la que el ligando "encaja" en el receptor como una llave en una cerradura. Los métodos de *docking* iniciales se basaban en esta teoría y tanto el ligando como el receptor se trataban como cuerpos rígidos. Entonces la teoría del "ajuste inducido" propuesta por Koshland[43,44] llevó el modelo "llave-cerradura" un paso más allá, con la idea de que el centro activo de la proteína está continuamente reajustándose al ligando, mediante interacciones con este. Esta teoría sugiere que tanto el ligando como el aceptor han de ser tratados como elementos flexibles durante el *docking*, describiendo los eventos de unión de forma más precisa que cuando se tratan como rígidos.

Considerando las limitaciones de cálculo computacional, durante mucho tiempo los métodos de *docking* han considerado al ligando flexible y al receptor como rígido, y todavía sigue siendo la forma usual de trabajo[45]. No hace mucho que se llevan a cabo esfuerzos para utilizar la flexibilidad del receptor[46], sin embargo, este tipo de *docking*, especialmente cuando la flexibilidad alcanza la cadena principal del receptor, supone todavía un enorme reto para los métodos disponibles.

Existen numerosas soluciones informáticas para realizar *docking*, cada uno fundamentalmente basado en una de las metodologías propuestas en la tabla 1.

Hemos utilizado AutoDock 4.0[47] y AutoDock Vina, que utilizan el algoritmo genético para generar las *pose* (posiciones dentro del receptor), un método estocástico bien conocido y también rápido para los recursos computacionales locales de los que disponemos para este trabajo, sobre todo en el caso de AutoDock Vina. Además, este *software* permite seleccionar como flexibles ciertas posiciones del receptor, de gran utilidad en estructuras que han sido cristalizadas sin ligando.

2.3 Dinámica Molecular

Las dos principales herramientas en el estudio teórico de la dinámica de moléculas biológicas, son el método de MonteCarlo (MC) y la Dinámica Molecular (MD)[48]. **La MD es un método de simulación computacional que estudia el movimiento físico de átomos y moléculas**, a los que se permite interactuar durante un periodo de tiempo fijado, obteniendo una visión de la evolución dinámica del sistema y teniendo la ventaja, frente al método de MC, de presentar una ruta a las propiedades dinámicas del sistema: coeficientes de transporte, respuestas tiempo-dependientes a perturbaciones, etc. En la versión más común, las trayectorias de átomos y moléculas se determinan mediante la solución de ecuaciones de movimiento de Newton, donde las fuerzas entre las partículas y su energía potencial se calculan utilizando potenciales interatómicos o campos de fuerza de mecánica molecular.

Debido al enorme número de partículas que suelen contener los sistemas moleculares, es imposible determinar analíticamente las propiedades de sistemas tan complejos; la MD circunvala este problema usando métodos numéricos. Eso sí, por la misma razón, las simulaciones de MD son muy dependientes del estado inicial, pudiéndose generar errores acumulativos en la integración numérica, que pueden ser minimizados con la adecuada selección de algoritmos y parámetros, aunque no totalmente eliminados.

Las simulaciones MD generan información detallada sobre las fluctuaciones y cambios conformacionales en proteínas y ácidos nucleicos. Estos métodos actualmente se usan de rutina en la investigación de la estructura, dinámica y termodinámica de moléculas biológicas y sus complejos, ya que para sistemas que obedecen la "hipótesis ergódica"[d], la evolución temporal de una única molécula en una simulación MD, puede ser usada para determinar propiedades termodinámicas macroscópicas del sistema –el valor medio del movimiento de las partículas en un sistema ergódico corresponde a las medias de un *colectivo microcanónico*–. En física estadística, un "colectivo" es una colección de todos los posibles sistemas que tienen diferentes estados microscópicos, pero idénticos estados macroscópicos o termodinámicos. Existen diferentes colectivos de diferentes características:

- Colectivo microcanónico (NVE): El estado termodinámico se caracteriza por un número fijo de átomos (N), un volumen fijo (V) y una energía fija (E). Corresponde a sistemas aislados.
- Colectivo canónico (NVT): El estado termodinámico se caracteriza por un número fijo de átomos (N), un volumen fijo (V) y una temperatura fija (T).

[d] En la hipótesis ergódica, un conjunto de microestados, de un sistema de N partículas en un instante determinado, se relaciona con el tiempo necesario para que una partícula de ese sistema alcance todos esos microestados. Si medimos en un momento dado un aspecto concreto del movimiento de todas y cada una de las partículas del sistema, como su velocidad, y hacemos la media; en un sistema ergódico obtendremos el mismo resultado que si seguimos una partícula durante un período largo de tiempo, medimos repetidamente su velocidad y hallamos el promedio de estos valores.

- Colectivo isobárico-isotérmico (NPT): Se caracteriza por un número fijo de átomos (N), una presión fija (P) y una temperatura fija (T) **y ha sido el conjunto usado en este trabajo.**
- Gran colectivo canónico (μVT): El estado termodinámico se caracteriza por un potencial químico fijo (μ), un volumen fijo (V) y una temperatura fija (T).

2.3.1 Diseño de la simulación

El diseño de una simulación MD tiene que tener en cuenta el poder de cálculo computacional disponible. El tamaño de la simulación (N=nº de partículas), el paso de tiempo y la duración de la misma ha de ser cuidadosamente seleccionado, para que pueda terminar en un plazo de tiempo razonable. Al mismo tiempo, han de ser lo suficientemente largas para ser relevantes en la escala de tiempo del proceso natural estudiado:

- Movimientos locales (0.01 a 5 Å, 10^{-15} [fs] a 10^{-1} s).
    - Fluctuaciones atómicas.
    - Movimientos de las cadenas laterales.
    - Movimiento de bucles.
- Movimiento de cuerpos rígidos (1 a 10Å, 10^{-9} [ns] a 1s).
    - Movimiento de hélices.
    - Movimiento de dominios (movimiento bisagra).
    - Movimiento de subunidades.
- Movimientos a gran escala (> 5Å, 10^{-7} [μs] a 10^{4} s).
    - Transiciones hélice-no estructura.
    - Disociación/Asociación
    - Plegamiento y despliegue

El uso más intensivo de CPU en una MD clásica es la evaluación de la energía potencial, como función de las coordenadas internas de las partículas, cuya evaluación más costosa en términos computacionales es la de las fuerzas no enlazantes, como las interacciones electrostáticas entre pares y las fuerzas de van der Waals. Estas pueden ser tenidas en cuenta de forma explícita –como se ha hecho en nuestras simulaciones– o mediante aproximaciones, reduciendo el coste computacional.

Otro factor importante para el uso de CPU es el tamaño del paso de integración, que es el lapso de tiempo entre evaluaciones del potencial y utilizado para propagar las coordenadas del sistema, según las ecuaciones de movimiento. Este paso de tiempo debe ser lo suficientemente pequeño para evitar errores discretos (p. ej. más pequeño que la frecuencia de vibración más rápida del sistema). Pasos de integración típicos en MD están en el orden del femtosegundo (10^{-15} s).

Al simular moléculas en solución se ha de elegir entre solvente explícito o implícito. En solvente explícito el potencial para cada partícula debe ser calculado por el campo de fuerzas, con un alto

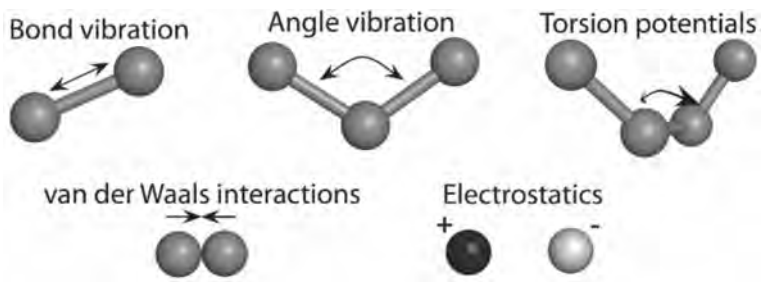
coste computacional, esto se puede reducir mediante aproximaciones como en el solvente implícito, pero la granulosidad y viscosidad del explícito **–seleccionado para nuestras simulaciones–**, es esencial para reproducir ciertas propiedades de los solutos y especialmente importantes para reproducir cinéticas.

Por último, el tamaño de la caja de simulación debe ser lo suficientemente grande para evitar artefactos debido a las condiciones límites o frontera o bien emplear "condiciones límite periódicas o *periodic boundary conditions* (PBC)", donde un lado de la simulación sale por el lado opuesto, mimetizando el espacio de fase que hemos usado en las simulaciones.

2.3.2 Campo de fuerzas

Una simulación molecular requiere definir la función de energía potencial: una descripción de los términos con los que las partículas de la simulación interactúan, que en Química y Biología se denomina "campo de fuerzas". Estos campos se pueden definir con muchos niveles de precisión física, en química los más comunes se basan en mecánica molecular y un tratamiento de las interacciones partícula-partícula que puede reproducir cambios estructurales y conformacionales, pero no reacciones químicas.

El campo de fuerzas usado en nuestras simulaciones se trata de un "campo de fuerzas empírico", consistente en la suma de fuerzas asociadas a enlaces químicos, ángulos de enlace y diedros, y fuerzas no enlazantes como las cargas electrostáticas y las fuerzas de van der Waals (figura 9). Estas fuerzas se calculan de forma empírica para cada átomo (figura 10) y luego se parametrizan en tablas que incluyen dicho campo de fuerzas y, debido a esto, este tipo de simulaciones no pueden modelar el proceso de formación y rotura de enlaces de forma explícita. Algunos de los campos de fuerzas clásicos más comunes son *AMBER*[49], *CHARMM*[50], *GROMOS*[51] y *OPLS*[52]. Cuando es requerido un nivel de detalle más fino, se usan potenciales basados en mecánica cuántica, algunas técnicas utilizan potenciales híbridos clásico/cuántico donde el sistema en general es tratado de forma clásica pero una pequeña región lo hace como un sistema cuántico, generalmente para simular una transformación química.



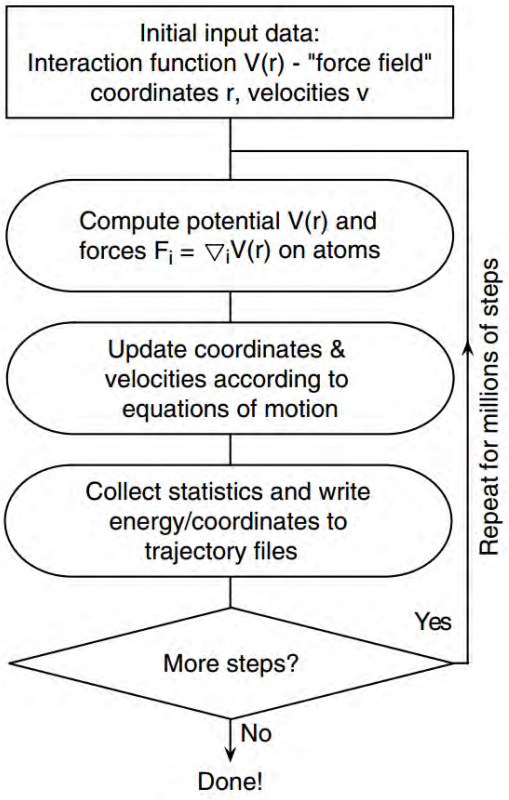
Fuente: "Molecular modeling of proteins". Methods in Molecular Biology 443. Various authors. Edited by Andreas Kukol. Human Presss.

Figura 9. Fuerzas enlazantes (arriba) y no enlazantes (abajo) utilizadas en un campo de fuerzas empírico.

$$E = \sum_{\text{bonds}} K_l(l-l_e)^2 + \sum_{\text{angles}} K_a(\theta-\theta_e)^2 + \sum_{\text{torsions}} \frac{V_n}{2}[1+\cos(n\phi-\gamma)] + \sum_{i<j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \text{conv}\frac{q_i q_j}{r_{ij}}$$

Figura 10. Cálculo de la energía potencial a partir de la resolución de las ecuaciones de movimiento de Newton. La energía potencial es una función de las posiciones atómicas $\{r_i\}$. La aceleración es por tanto una función también de las posiciones atómicas, sin embargo, estas posiciones cambian a lo largo del tiempo $\{r_i(t)\}$ por lo que la aceleración también cambia a lo largo del tiempo.

2.3.3 Algoritmo básico de una simulación MD



Fuente: "Molecular modeling of proteins". Methods in Molecular Biology 443. Various authors. Edited by Andreas Kukol. Human Presss.

Figura 11. Diagrama simplificado de una simulación típica de MD. La idea básica es generar estructuras de un conjunto natural calculando funciones de potencial e integrando ecuaciones de movimiento de Newton; estas estructuras se usan entonces para evaluar las propiedades de equilibrio del sistema. Un paso de tiempo típico está en el orden de 1-2 fs.

Una vez seleccionada la estructura de partida y definida la caja de simulación, el sistema se solvata y equilibra mediante la adición de iones. A continuación, mediante un proceso de minimización de energía, se relaja la estructura hasta dejarla en un estado estable de energía local, cuanto más bajo

mejor (optimización) y posteriormente el sistema se nivela a la temperatura y presión constantes requeridas para la simulación, en un paso previo a la producción denominado "equilibrado". Una vez que el sistema está listo se inicia el proceso de simulación MD propiamente dicho, cuyo esquema básico es el que se muestra en la figura 11. Cada paso de integración calcula el resultado de resolver las ecuaciones de movimiento de Newton para cada una de las fuerzas implicadas (figura 10).

La mayoría de publicaciones sobre dinámica de proteínas y ADN utilizan datos de simulaciones que rondan el ns (10^{-9} s) y el μs (10^{-6} s), para obtener estos tiempos son necesarios días, meses e incluso años en tiempo de CPU (horas de cómputo) y se hace necesaria la computación distribuida. En este trabajo hemos utilizado los recursos de la Red Española de Supercomputación (RES), que posee nodos distribuidos por la geografía española (figura 12), donde cada uno posee arquitectura de cálculo en paralelo por lo que es posible usar numerosos CPUs que reducen considerablemente el tiempo de cálculo para cada simulación, que en nuestro caso estaban entre los 500 y 2000 ns. De los diferentes nodos de la RES hemos realizado cálculos en Picasso (Málaga), Tirant (Valencia) y Magerit (Madrid).

Fuente: http://www.iac.es/

Figura 12. Nodos de la RES. Año 2015.

Existen diversos paquetes informáticos para realizar MD. Los más usuales en el campo de las biomoléculas son Amber, NAMD, DESMOND y **GROMACS.**

En las simulaciones de este trabajo se han usado los campos de fuerzas de AMBER y GROMOS. AMBER por su amplia utilización en moléculas biológicas, debido a su correcto tratamiento de las interacciones a larga distancia y GROMOS, para el caso en el que incorporamos membranas en las simulaciones, por su adecuada parametrización de hidrocarburos y cadenas alifáticas. Ambos son campos de fuerzas incorporados en el paquete informático GROMACS, que hemos utilizado para todas las simulaciones, por su versatilidad y número de herramientas para el análisis que ofrece y también, porque muestra un rendimiento computacional óptimo en uso local (muchas CPU en un mismo nodo) y un rendimiento aceptable en infraestructuras de *High performance computing* (HPC) (muchas CPUs distribuidas en distintos nodos).

2.4 Metadinámica

Entre las dos principales herramientas para el estudio teórico de moléculas biológicas, la MD presenta la ventaja frente al método de MC, de mostrar una ruta hacia las propiedades dinámicas del sistema, sin embargo, la MD adolece de un problema: el espacio muestral. Al tener que recorrer este espacio de forma dinámica, mediante la solución de las ecuaciones de movimiento de Newton, ciertas zonas de este espacio conformacional no serán nunca visitadas o lo harán en un intervalo de tiempo demasiado largo, algo que ocurre en menor frecuencia con el método estocástico de MC. Sin embargo, existe una solución a este problema: **la Metadinámica (MetaD)**. Fue sugerida por Laio y Parrinello en 2002[53] y **se presenta** esta **como una poderosa técnica para maximizar el muestreo en MD y reconstruir superficies de energía libre, como función de unos pocos grados de libertad seleccionados, a los que nos referimos normalmente como "Variables colectivas (CVs)".** En la MetaD, el muestreo se acelera por un sesgo potencial histórico-dependiente, que es construido de forma adaptable al espacio de las CVs[54]. Existen otros métodos que pueden ser incluidos en esta clase, en las que se utiliza este sesgo potencial (o fuerza) adicional, como el *umbrela sampling*[55], *local elevation*[56], *conformational flooding*[57], *adaptative force bias*[58], *steered MD*[59] y *self-healing umbrella sampling*[60].

Teoría

La MetaD consiste en una simulación MD convencional en la que un sesgo potencial histórico-dependiente (V_G), que es función de las CVs, es añadido al hamiltoniano[e] del sistema. Este potencial se puede escribir como una suma de Gaussianas depositadas a todo lo largo de la trayectoria del sistema, sobre el espacio de las CVs, para evitar que éste revisite configuraciones que ya han sido muestreadas. Este sesgo de energía se aplica continuamente durante la MD actuando directamente sobre las coordenadas microscópicas del sistema:

Siendo S un conjunto de d funciones de las coordenadas microscópicas R del sistema:

$$(R) = (S_1(R), \dots, S_d(R)).$$

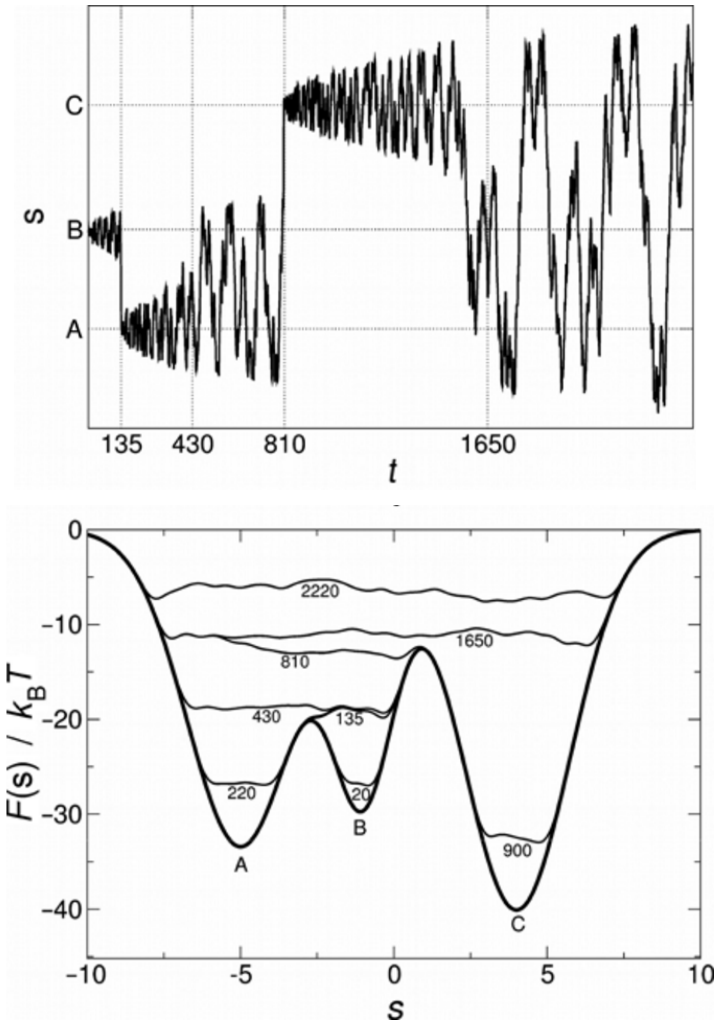
A tiempo t el sesgo potencial de la MetaD se puede escribir así:

$$V_G(S,t) = \int_0^t dt' \omega \exp\left(-\sum_{i=1}^{d} \frac{(S_i(R) - S_i(R(t')))^2}{2\sigma_i^2}\right)$$

Donde ω es una tasa de energía y σ_i es la anchura de la Gaussiana para la CV i. La tasa de energía es constante y expresada generalmente como la altura W de una Gaussiana y un paso τ_G. $\omega = \frac{W}{\tau G}$

[e] Energía total del sistema.

Para entender el efecto de V_G (sesgo potencial) sobre la evolución del sistema, consideremos el caso simple de un potencial unidimensional (figura 13), donde existen tres mínimos locales A, B y C.



Fuente: Barducci A., Bonomi M., Parrinello M.. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2011. Vol. 1, 5. Pg 826-843

Figura 13. Ejemplo de simulación MetaD de una dimensión. El tiempo t se mide contando el número de Gaussianas depositadas. Arriba: Evolución temporal de las CVs durante la simulación. Abajo: Representación esquemática del "rellenado" del potencial subyacente (línea gruesa), mediante la deposición de Gaussianas a lo largo de la trayectoria.

La suma del potencial subyacente y el sesgo de la MetaD se muestran a diferentes tiempos (líneas finas). La estructura comienza en el mínimo local B. En una simulación MD estándar, el sistema permanecería "atascado" en este mínimo porque las barreras a superar son mayores que las fluctuaciones térmicas. En una simulación MetaD, se depositan Gaussianas a lo largo del tiempo, provocando que el potencial subyacente "crezca". En este caso, a t = 135 el sistema es empujado del mínimo local B a un nuevo mínimo. La ruta de escape más conveniente y natural es pasar la

barrera más baja y caer en el mínimo local A. Entonces la acumulación de Gaussianas comienza de nuevo. El sistema está atrapado en A hasta que la energía libre de este mínimo es completamente rellenada (t = 430). En este punto el sistema difunde entre los mínimos B y A. En el tiempo t = 810, el sistema puede acceder fácilmente a la región C. Al final, cuando este mínimo está también compensado por el sesgo potencial (t = 1650), la evolución del sistema se asemeja a un paseo aleatorio por este FES[f] aplanado.

Con este ejemplo se ilustran los beneficios de la MetaD:

1. Acelera el muestreo de eventos raros alejando el sistema de mínimos de energía libre.
2. Permite explorar nuevas rutas de reacción, ya que el sistema tiende a escapar de los mínimos de energía, superando las barreras menores en la dirección explorada.
3. No se requiere ningún conocimiento previo del FES.
4. Pasado un tiempo, el sesgo potencial (V_G) da una estimación no sesgada de la energía libre subyacente.

$$V_G(S,t \to \infty) = -F(S) + C$$

Donde C es una constante aditiva y la energía libre $F(S)$ que de manera analítica se define como:

$$F(S) = -\frac{1}{\beta} \ln \left(\int dR \ \delta(S - S(R))\, e^{-\beta U(R)} \right)$$

$\beta = (k_B T)^{-1}$, k_B es la constante de Boltzmann, T es la temperatura del sistema y $U(R)$ la función de energía potencial.

Pero al mismo tiempo la MetaD presenta dos contratiempos importantes:

1. En una única ejecución, la V_G no converge a una energía libre constante, sino que oscila alrededor de esta, lo que tiene dos consecuencias:
    a. El sesgo potencial "sobrellena" el FES y empuja el sistema hacia regiones de alta energía del espacio de las CVs.
    b. No es sencillo decidir cuándo parar la simulación.
2. El proceso para definir el conjunto de CVs adecuados para describir sistemas complejos es también una tarea difícil y requiere habitualmente del método de "ensayo y error", hasta dar con las CVs que expliquen y describan el proceso que se quiere estudiar.

Variables colectivas (CVs)

Los sistemas químicos contienen un enorme número de átomos que, en muchos casos, hacen imposible entender nada mediante la monitorización directa de sus posiciones. Por ello, se introduce el concepto de variable colectiva. Una CV es una función que representa las coordenadas

[f] FES (Free Energy Surface). Superficie de energía libre.

microscópicas del sistema[54], es decir, variables que describen los procesos dinámicos en los que estamos interesados y que son las que únicamente se monitorizan.

Para garantizar una aplicación efectiva de la MetaD, las CVs deben respetar una serie de requisitos:

1. Deben distinguir entre el estado inicial y final, así como describir todos los pasos intermedios relevantes.
2. Deben incluir todos los modos "lentos" del sistema; Se definen como "lentas" aquellas variables que no pueden ser adecuadamente muestreadas en la escala de tiempo de la simulación, esperando que las variables "rápidas" se ajusten rápidamente a la evolución de las "lentas.
3. Deben ser un número limitado por motivos de rendimiento computacional. Normalmente limitado a dos como máximo en una metadinámica convencional.

Ejemplo de CVs pueden ser la distancia interatómica o entre centros de masas, el ángulo de torsión de un diedro, el nº de coordinación o nº de interacciones entre varios átomos o cualquier otra variable que pueda medirse mediante una función.

Cada una de las funciones que explican el cambio de una CV es utilizada para el cálculo de V_G (la cantidad de energía asociada al cambio de estado/posición en la CV). **Estas CVs y sus funciones están implementadas en el *plugin* PLUMED[61] v1.3, que hemos utilizado con el paquete GROMACS**.

2.5 BIAS-Exchange

***BIAS-Exchange* es una técnica de MetaD[62], una aproximación específicamente diseñada para acelerar la medida de la energía de eventos raros en casos muy complejos, en los que las variables relevantes para el proceso son más de 2**.

El método trabaja de la siguiente manera:

- Primeramente, se listan todas las CVs que se consideran pueden ser relevantes para el proceso que se investiga. No hay límites para el número de variables.
- Se ejecutan en paralelo múltiples simulaciones MetaD, cada una con una CV de la lista activada, actuando sobre ella el sesgo potencial como en una MetaD normal.
- A intervalos de tiempo fijados, las configuraciones entre pares de réplicas se intentan intercambiar. El intercambio se acepta de acuerdo con el criterio de Metropolis[63] aplicado a la diferencia de energía acumulada en cada configuración.

De este modo la dinámica de cada réplica es llevada en una dirección que cambia estocásticamente con el tiempo. Lo que permite al sistema explorar un espacio de energía libre complejo (multidimensional) con gran eficiencia, aunque el sesgo potencial cada vez sea unidimensional[g].

Nada mejor que explicar el método y su importancia a través de un ejemplo:

Considérese una dinámica con un potencial bidimensional como el de la figura 14. Si se desarrolla la metadinámica sobre el eje x, se obtiene una estimación de la energía libre que está afectada por un gran error: el sistema muy raramente salta entre los pozos superiores e inferiores, solo debido a fluctuaciones térmicas (muy escasas), obtener el correcto valor de energía libre requeriría tomar la media de muchas transiciones a lo largo del eje y.

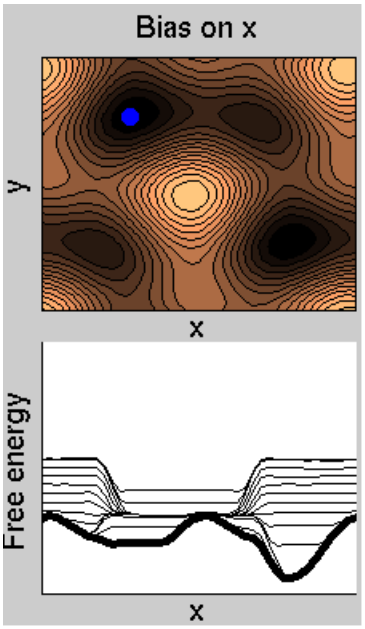


Figura 14: El sesgo potencial de energía está activado solo sobre el eje x. Sin intercambios acelerados por BIAS-Exchange, el cambio a los pozos en el eje y obedecen a las fluctuaciones propias de una dinámica molécula clásica

[g] Home Page of Alessandro Laio. Research - Bias Exchange. http://people.sissa.it/~laio/Research/Res_BE.php

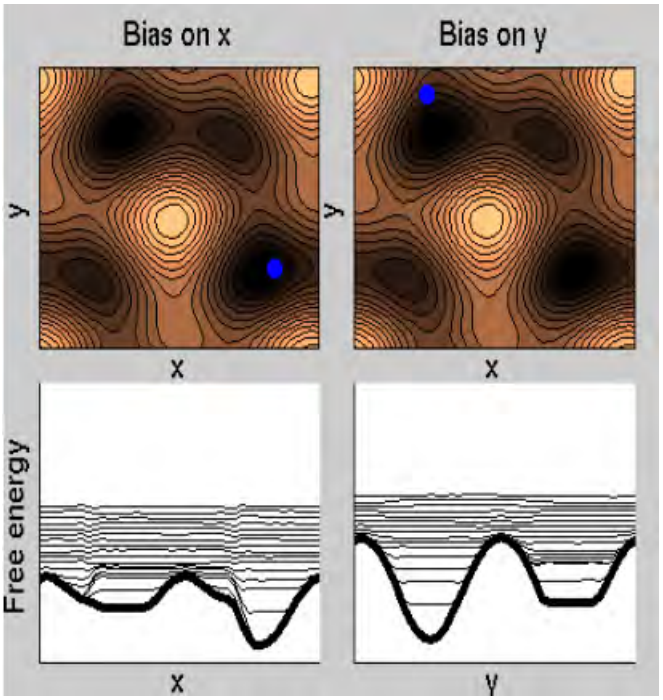


Figura 15: Simulaciones metadinámicas donde en cada una se aplica el sesgo potencial sobre x o y. El intercambio entre sus configuraciones permite una mejor exploración del FES y un mejor cálculo de la energía libre.

Ahora se ejecutan dos simulaciones metadinámicas sobre dos réplicas; una aplicando el sesgo potencial sobre x, y otra sobre y. A intervalos programados, permitimos que las dos réplicas intercambien las configuraciones, aceptando el intercambio de acuerdo a criterio Metrópolis[62], aun cuando el coste computacional se ha doblado respecto a la simulación anterior, se observa un reducción significativa de la histéresis[h].

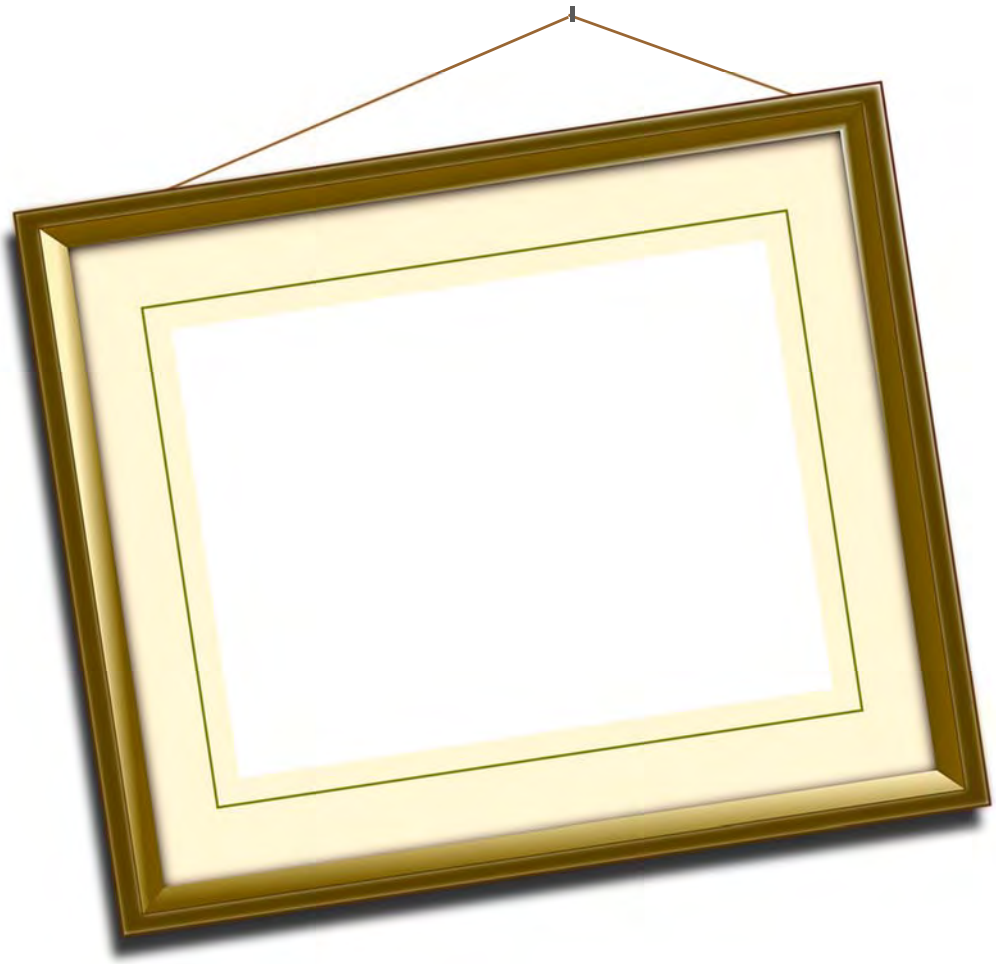
Ahora el potencial de la metadinámica casi compensa exactamente la energía libre, ambos como función de x e y: el perfil genera prácticamente líneas planas. La histéresis es mucho más reducida y el potencial metadinámico genera una buena estimación de energía libre.

Esta aproximación permite una reconstrucción paralela de la energía libre como función de muchas variables. El coste computacional se incrementa solo linealmente con el número de variables.

Cuantas más variables se añadan, mejor será la convergencia. Sin embargo, esta será lenta si hemos olvidado una variable importante.

Usando N variables, el resultado de la simulación no es la energía libre como función de las N variables, sino N proyecciones unidimensionales de la energía libre. Obtener la energía libre como función de todas las variables, requiere un postprocesado de los resultados. **Recientemente se ha publicado la herramienta METAGUI[64] que agiliza el postprocesado y análisis de los resultados de simulaciones de metadinàmica y en BIAS-Exchange Metadynamics, herramienta que hemos utilizado en esta tesis.**

[h] Histéresis: tendencia de un material a conservar una de sus propiedades, en ausencia del estímulo que la ha generado. Aquí el término histéresis hace referencia a la tendencia de una MetaD a permanecer más tiempo en un estado debido a la no incorporación de una variable colectiva lenta adicional.



MARCO DE LA TESIS

Creemos posible encontrar características estructurales y funcionales comunes a todas las proteínas glicosiltransferasa con plegamiento GTA, a través el estudio comparado de secuencias, estructuras y funciones de diferentes miembros de esta superfamilia.

Todas las proteínas de la superclase GTA comparten un mismo tipo de plegamiento, basado en el plegamiento tipo Rossmann. Sin embargo, los estudios realizados hasta la fecha revelan zonas desordenadas dentro de esta estructura, para las que no han podido establecerse relaciones entre diferentes proteínas, considerándose topologías específicas para cada tipo de GTA. A su vez, a pesar de que las GTAs son consideradas proteínas con un solo dominio, este plegamiento solo cubre una parte de la secuencia en muchas de las proteínas conocidas, generalmente en la parte inicial, y a la que se denomina región N-terminal; para el resto de la secuencia, la diversidad estructural y aminoacídica es mucho mayor, sin que existan hasta hoy evidencias de ninguna relación estructura-función común entre las diferentes GTAs. A esta región se le denomina región C-terminal (también llamada "extensión C-terminal", para evitar la confusión a veces observada entre "regiones" y "dominios", ya que no necesariamente esta región contiene o corresponde a un dominio).

Aun con la enorme variedad de compuestos producidos por esta superfamilia de proteínas, los diferentes ligandos se ubican, según las estructuras existentes, en posiciones similares en todas ellas, con dos únicos mecanismos de transferencia conocidos, los de tipo *inverting* o *retaining*. En la literatura científica usualmente, cada reacción se ha estudiado de forma específica para cada proteína, pero en vista de las similitudes encontradas, podrían existir residuos conservados en posiciones comunes a todas las GTAs, y quizá también podría hallarse una relación secuencia-estructura entre las diferentes familias, para los dos grupos de mecanismos de transferencia.

El establecimiento de patrones de secuencia, estructurales o de función entre proteínas diferentes, resulta de una gran utilidad en investigación, pues permite inferir información entre estas y realizar predicciones para proteínas desconocidas, en niveles muy iniciales de estudio. Del mismo modo que actualmente es posible determinar, tan solo conociendo su secuencia, si una nueva proteína pertenece a la superfamilia de las GTAs e incluso su familia y estructura, con este estudio pretendemos predecir, además, otras características como la molécula dadora o aceptora del enzima, su mecanismo de acción, los residuos catalíticos o cualquier otra información relevante que puedan compartir las proteínas con plegamiento GTA.

Todo este trabajo se encuadra en un contexto teórico o de experimientación indirecta. La intención de esta tesis es encontrar respuesta a las preguntas que surjan sobre las posibles relaciones secuenci-estructura-función de las glicosiltransferasas con plegamiento GTA, utilizando herramientas computacionales, no experimentales, mediante el uso de la simulación y técnicas bioinformáticas. Pretendemos aquí explotar al máximo las posiblidades predictivas de estas herramientas computacionales, trabajando *in silico*.

Partiendo de la información topológica de estructuras cristalográficas reales, generaremos modelos informáticos tridimensionales de esas estructuras, con los que se desarrollarán simulaciones

computacionales en diferentes sistemas. A veces, los modelos se construirán *de novo*, sin estructura de partida, serán segmentos estructurales pequeños, como hélices y también se construirán estructuras mutantes, puentes disulfuro o simulaciones que incorporen membranas lipídicas, todo de forma virtual, utilizando únicamente la literatura científica y el poder computacional de nuestros ordenadores y servidores de la Red Española de Supercomputación. Utilizaremos técnicas como la Dinámica Molecular o la Metadinámica para el estudio termodinámico de proteínas, sus fluctuaciones, sus cambios conformacionales y sus estados más probables. Mediaremos la energía de interacción entre proteína y ligandos y su emplazamiento en la estructura mediante *Docking*. Usaremos bases de datos de secuencias y estructuras, para la realización de alineamientos múltiples y superposiciones, utilizando *software* instalado en local y servidores automáticos de la red. Además, aplicaremos métodos de *scripting*, minería de datos y algoritmos estadísticos, que extraigan información nueva y relevante sobre el conocimiento actual de las glicosiltranferasas con plegamiento GTA, y que sea aplicable a proteínas con función desconocida. En definitiva, sacaremos el máximo partido posible de estas herramientas computacionales y métodos bioinformáticos, aportando respuestas a cuestiones planteadas sobre las relaciones secuencia-estructura-función en glicosiltransferasas con plegamiento GTA, y generando conclusiones que definan un marco sólido sobre el que abordar las mismas cuestiones, esta vez y fuera de este trabajo, desde un enfoque clásico y experimental.

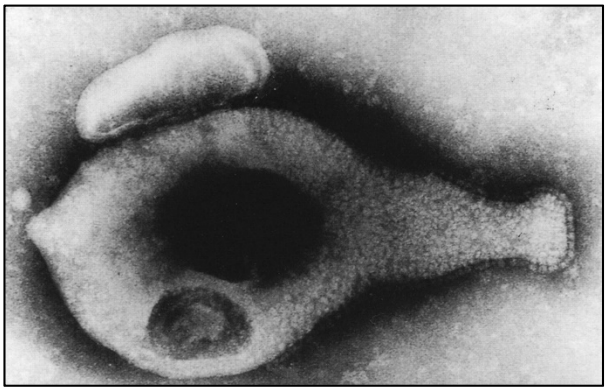
Para acometer este trabajo nos hemos valido, por un lado, de dos glicosiltransferasas modelo para el estudio particular de relaciones estructura-función: la proteína MG517 de *Mycoplasma genitalium* y la proteína GpgS de *Micobacterium tuberculosis*. Por otro lado, para la generalización de relaciones secuencia-estructura-función, hemos recurrido a toda la información de secuencia y estructural disponible para las proteínas GTA, de las bases de datos CAZy y PDB.

Nota: Puesto que el desarrollo del capítulo 1, sobre la identificación de residuos catalíticos en la proteína MG517, permitió su publicación como artículo científico, se ha incluido aquí la parte experimental del mismo, también incluida en el artículo. Esta parte experimental no fue desarrollada por mí, el autor de esta tesis, sino por mi compañero de doctorado Carles Francisco. Sin embargo, he querido conservar el formato con esta parte exeprimental, para evidenciar el poder de predicción y de dirección de estudios experimentales que poseen las aproximaciones computacionales y el uso de la simulación. Aunque otros capítulos también han culminado en la publicación de artículos o derivado en estudios experimentales (como el 3), ya no han sido incorporados en este trabajo.

1. Proteína MG517

La proteína MG517 (glicosildiacilglicerol sintasa) es el producto del gen mg517, de la bacteria patógena *Mycoplasma genitalium* (figura 16).

Los micoplasmas son parásitos obligados de células eucariotas que se caracterizan por su pequeño tamaño y por la **ausencia de pared celular**, algo que ha sido utilizado para separarlos taxonómicamente de otras procariotas en la clase Mollicutes[65,66,67] [mollis (flexible, fácil de quitar); cutis (piel)].



Fuente: David Taylor-Robinson and Jørgen Skov Jensen. Clin. Microbiol. Rev. July 2011 vol. 24 no. 3 498-514

Figura 16. MET de *M. genitalium* G-37. La célula muestra las características formas de botella y protuberancia u organelo terminal (truncada en la imagen).

El orden con mayor número de especies actualmente identificadas dentro de la clase Mollicutes es la de los Mycoplasmatales, cuyo género Mycoplasma posee más de 100 especies, muchos de ellos patógenos de la especie humana. El genoma secuenciado de micoplasmas tiene un reducido número de GTs anotadas (putativas), acorde con sus limitadas capacidades biosintéticas, como consecuencia de su reducido genoma, evolucionado por degeneración o evolución reductiva[65]. Muchas de estas GTs pertenecen a la familia de las GT2, donde tan solo dos enzimas ortólogos se han identificado experimentalmente, uno de *M. pneumonia*[68], agente causal de la neumonía atípica[69,70], y otro en nuestro laboratorio, de *M. genitalium*[71] quien provoca enfermedades urogenitales como uretritis y cervicitis no gonocócicas, agudas o crónicas e inflamación pélvica[72,73].

Puesto que *M. genitalium* no posee una pared celular de peptidoglicano, en ambientes hostiles esta bacteria posee muy poca estabilidad osmótica. Esta ausencia de pared celular además, es la responsable de su insensibilidad a los antibióticos más comunes, de la clase β-lactámicos, que actúan inhibiendo la síntesis de la pared celular[74]. Sin embargo, la membrana de micoplasmas contiene glicoglicerolípidos libres, que son cruciales para su estabilidad y están ausentes en las células animales.

Los principales glicoglicerolípidos en la membrana de los micoplasmas son el monoglicosildiacilglicerol (MGlcDAG) y el diglicosildiacilglicerol (DGlcDAG). El balance entre ellos es lo que define la propiedades de la membrana, como la curvatura, la fluidez y la estabilidad; esto ha sido demostrado en *Acholeplasma laidlawii*, una de las bacterias más investigadas, en cuanto a la función de los glicolípidos en las membranas biológicas, perteneciente a la clase Mollicutes[75,76,77].

GTs que sintetizan glicoglicerolípidos se han propuesto como potenciales dianas terapéuticas contras las infecciones producidas por micoplasmas[78].

En nuestro laboratorio se ha demostrado experimentalmente que la GT MG517 (glicosildiacilglicerol sintasa) es la proteína responsable de la biosíntesis de glicoglicerolípidos de membrana en *M. genitalium*, cuya función es esencial para la viabilidad del organismo[71]. La enzima está activada por fosfolípidos aniónicos y aunque es posible de purificar sin detergentes, hacerlo de este modo arrastra siempre con ella lípidos de membrana, por todo esto se considera que la GT MG517 está asociada a la membrana plasmática de *M. genitalium*, sin que se hayan identificado, sin embargo, ningún dominio transmembrana; por el perfil de hidrofobicidad de la proteína es posible que esta asociación se encuentre en la región C-terminal. Mediante ensayo con UDP-Glc marcado radiactivamente, se sabe que la síntesis de estos dos glucolípidos tiene lugar de forma secuencial en la membrana, obteniéndose MGlcDAG y a partir de este DGlcDAG, por un enlace β (1→6) entre las glucosas (figura 17). La misma enzima realiza esta transferencia de forma secuencial, algo que también sucede con la proteína ortóloga de *M. pneumoniae*, y aunque transfiere preferentemente glucosa también lo hace con la galactosa.

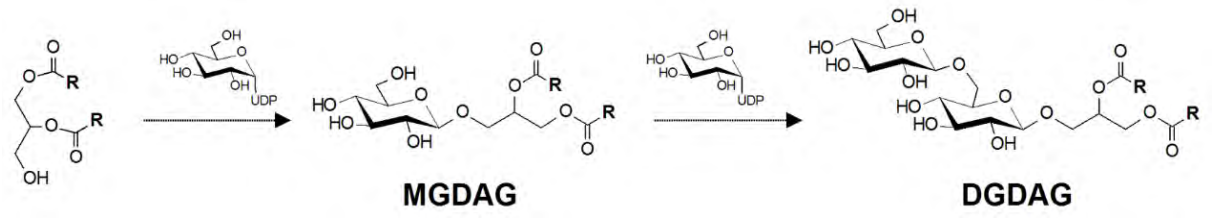


Figura 17. Reacción catalizada por la GT MG517 de *Mycoplasma genitalium*.

El estudio de formas truncadas de la GT MG517 reveló que la eliminación de la parte C-terminal de la proteína (134 últimos residuos), permite su purificación sin el uso detergentes, manteniendo el plegamiento, por lo que de nuevo se postula esta zona para la interacción con la membrana. La eliminación de tan solo 13 aminoácidos desde el extremo C-terminal, ocasiona la total pérdida de actividad del enzima[79], lo que indica que esta zona apical desarrolla un importante papel en la funcionalidad de MG517.

La ausencia de glicoglicerolípidos en las células hospedadoras (células animales) de micoplasma, hace de esta enzima una buena diana para la búsqueda de nuevos fármacos, mediante el diseño de inhibidores específicos. Sin embargo, su diseño racional se ha visto obstaculizado por la ausencia de información entre función y estructura para cualquier GT de micoplasmas, ya que no existe ninguna glicosildiacilglicerol sintasa con estructura GTA cristalizada.

2. Proteína GpgS

GpgS es el acrónimo del enzima glucosil-3-fosfoglicerato sintasa, del patógeno *Mycobacterium tuberculosis* (figura 18).

Este organismo es tristemente bien conocido por ser el causante de, posiblemente, la enfermedad infecciosa más prevalente del mundo, la tuberculosis. Antiguamente llamada *tisis*, la tuberculosis es una enfermedad insidiosa, ya que presenta signos muy leves en su fase inicial, con una variada sintomatología, siendo la neumonía una de las más graves y llegando a causar la muerte si no es tratada adecuadamente. Según datos de la OMS, en 2014, 9,6 millones de personas enfermaron de tuberculosis y 1,5 millones murieron por esta enfermedad. Además, es la principal causa de muerte de las personas infectadas por el VIH; en 2015, fue la causa de una de cada tres defunciones en este grupo. Se calcula que una tercera parte de la población mundial tiene tuberculosis latente; es decir, están infectadas por el microorganismo, pero aún no han enfermado ni pueden transmitir la infección.

Figura 18. *M. tuberculosis* (teñidos en rojo) en esputo.

M. tuberculosis es el principal responsable de los casos de tuberculosis en el mundo. Otras micobacterias, como *Mycobacterium bovis*, *Mycobacterium africanum*, *Mycobacterium canetti* y *Mycobacterium microti* pueden causar también la enfermedad, pero todas estas especies no lo suelen hacer en individuos sanos[80]. El único género de la familia de bacterias Mycobacteriaceae, es Mycobacterium. Esta familia está formada por bacilos aerobios inmóviles y no esporulados y entre sus patógenos, además de la tuberculosis está también el causante de la lepra: *Mycobacterium leprae*.

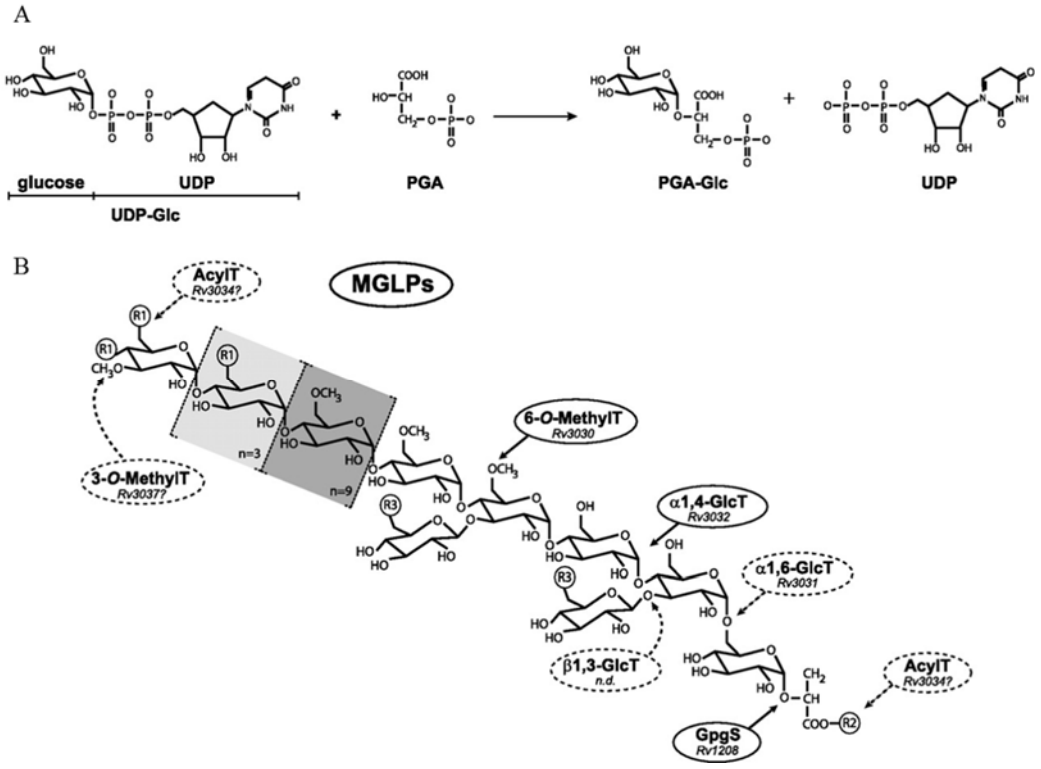
Las micobacterias comparten una pared celular compleja, más gruesa que la de muchas otras bacterias. Es rica en ácidos micólicos y en lípidos (más del 60 % del peso seco de la pared), lo que hace que su superficie sea hidrófoba y les confiere resistencia frente a muchos desinfectantes, detergentes y antibacterianos comunes. Esta es la principal responsable de la resistencia de este género de bacterias.

GpgS es una GTA que transfiere Glc desde el dador UDP-Glc a la posición 2 del aceptor fosfoglicerato (PGA), iniciando la ruta biosintética de los lipopolisacáridos 6-O-metilglucosa (MGLPs) en micobacteria (figura 19). Estos lipopolisacáridos se han propuesto como transportadores de ácidos grasos de cadena larga en el citosol de este patógeno, por lo que cualquier proteína que participe de forma esencial en esta ruta, como GpgS, es una buena candidata a tratar como diana terapéutica[81].

La estructura de GpgS ha sido resuelta para los organismos *M. tuberculosis* y *M. avium* subsp. paratuberculosis. Consiste en un homodímero formado por dos monómeros que interactúan a través de su región C-terminal y parte de la región N-terminal. Las estructuras cristalizadas muestran dos conformaciones diferentes para un bucle situado sobre la zona catalítica,

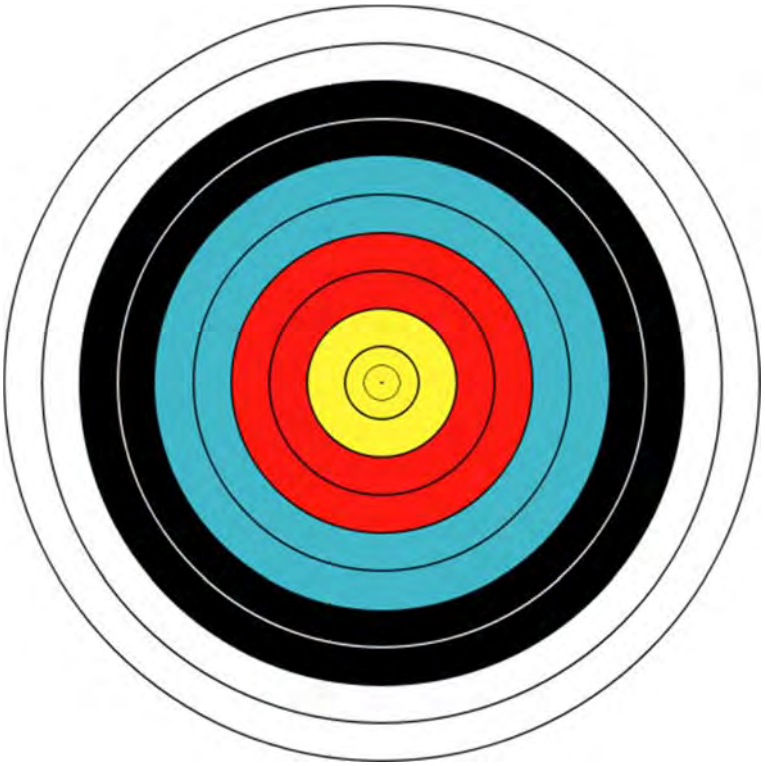
observándose una de ellas en los cristales con ligandos y la otra para las formas apo. Este bucle contiene, al menos, un residuo cuya interacción con el metal (que a su vez interacciona con el dador) parece depender precisamente de la conformación que el bucle adopte. Sin embargo, recientemente se ha obtenido un nuevo cristal de GpgS[81], que muestra ese bucle en ambas conformaciones en la misma estructura sin ligandos, lo que podría indicar un posible mecanismo interacción proteína-ligando de selección conformacional, al existir las dos formas en ausencia de ligandos, frente a otro posible mecanismo de ajuste inducido, donde son estos los que provocarían el cambio conformacional.

El mecanismo tipo *retaining* de este enzima ha sido estudiado experimentalmente[15], sin embargo, quedan todavía por determinar el orden de llegada de los ligandos o su relación con las conformaciones del bucle, elementos estos que ayuden a encontrar dianas en la proteína GpgS, para ser atacados por nuevos fármacos, en la defensa contra *M. tuberculosis*.



Fuente: Saioa Urresti et al. J. Biol. Chem. 2012;287:24649-24661

Figura 19. Biosíntesis de MGLP en micobacteria. A. Transferencia de Glc desde UDPGlc a la posición 2 de 3-fosfoglicerato para formar glucosil-3-fosfoglicerato. **B.** Estructura química de los MGLPs. R1, R2 y R3 son grupos acilos. R1: acetato, propionato o isobutirato. R2: octanoato. R3: succinato. Los MGLPs son una mezcla de 4 componentes principales que difieren en el contenido esterificado de succinato. Se muestran los nombres de los genes involucrados en los diferentes pasos de elongación y modificación.



OBJETIVOS

Puesto que hasta la fecha no ha sido posible determinar la estructura tridimensional de la glicosildiacilglicerol sintasa (MG517) de *Mycoplasma genitalium* (MG517) con la que abordar un diseño racional de fármacos:

1. **¿Podría realizarse el modelado estructural de la glicosildiacilglicerol sintasa (MG517) de *Mycoplasma genitalium* con la suficiente confianza como para asegurar que esa es realmente su estructura?**

2. **¿Es posible definir los residuos catalíticos de dicha enzima, y las posiciones que intervendrán en la afinidad por los ligandos?**

3. **¿Puede determinarse el lugar que ocuparán en la proteína, dador y aceptor?**

4. **En caso de existir, ¿se puede definir el tipo de asociación con la membrana?**

5. **¿Podría delimitarse con precisión los elementos estructurales que participan en la asociación con la membrana?**

Para dar respuesta a estas cuestiones, construiremos varios modelos de la proteína MG517, utilizando estructuras de proteínas filogenéticamente cercanas. Utilizaremos herramientas de alineamiento múltiple, superposición de estructuras y construcción de dendrogramas, con la información de proteínas GTA disponible. Los modelos serán validados por simulaciones de Dinámica Molecular clásica de larga duración y la posición de los ligandos determinada mediante *docking*. Estudiaremos las posiciones conservadas y estructuralmente cercanas a los ligandos, y señalaremos aquellos que mayor probabilidad tengan de afectar a la actividad del enzima. También estudiaremos el perfil hidropático de la proteína y de nuevo, por simulaciones de Dinámica Molecular clásica de larga duración validaremos, por separado, la estabilidad de estructuras particulares de la GT MG517 en asociación con una membrana fosfolipídica. Con la técnica de Metadinámica, mediremos la energía de asociación de estas estructuras a la membrana.

La relevancia que tengan los residuos señalados sobre la actividad del enzima, será estudiada experimentalmente mediante mutagénesis dirigida.

Mostraremos, en definitiva, un modelo estructural para la proteína MG517 en complejo con ligando, validado experimentalmente y una propuesta de unión a la membrana.

En relación a los cambios conformacionales del centro catalítico observados en las estructuras cristalográficas de la glucosil-3-fosfoglicerato sintasa de *Mycobacterium tuberculosis* (GpgS):

1. **¿Tiene alguna relación la conformación del bucle en el centro activo con la actividad catalítica de la glucosil-3-fosfoglicerato sintasa de *Mycobacterium tuberculosis* (GpgS)?**

2. **¿Son igualmente estables estas conformaciones en presencia o ausencia de ligandos?**

3. **¿Responde el mecanismo de esta proteína a un modelo de ajuste inducido o es más probable, por el contrario, el de selección conformacional?**

4. **¿Cuál es la relación entre el bucle y los ligandos? ¿Puede darse una explicación fenomenológica al cambio conformacional del bucle?**

5. **¿Son extrapolables a otras proteínas las respuestas a estas preguntas?**

En este caso serán modeladas varias estructuras de la proteína GpgS, a partir de cristales con diferentes conformaciones del bucle, y formas sin ligandos o en complejo con ellos. Con estos modelos se ejecutarán simulaciones de Dinámica Molecular clásica de larga duración, con la intención de monitorizar el comportamiento del bucle en distintas situaciones. Estas simulaciones se estudiarán en detalle, para definir las variables que más puedan afectar al cambio conformacional y las compararemos con los valores de otras proteínas homólogas también cristalizadas. Activaremos por separado esas variables en diferentes simulaciones de Metadinámica y analizaremos la estabilidad del bucle en presencia y ausencia de ligandos. Por último, mediante la técnica de BIAS-Exchange activaremos diversas variables a la vez y de nuevo analizaremos la estabilidad del bucle en sus diferentes conformaciones.

Plantearemos un modelo de entrada de los ligandos al centro activo y una explicación fenomenológica de su relación con las distintas conformaciones del bucle. Intentaremos definir el tipo de mecanismo de la proteína GpgS: ajuste inducido o selección conformacional.

Por último, ampliando el foco de estudio a todas las glicosiltransferasas con plegamiento GTA:

1. **¿Cuál es estrictamente la estructura compartida entre todas las GTAs? ¿Es posible definir una topología consenso?**

2. **¿Existen posiciones conservadas que puedan construir una "firma" para las GTAs, más allá del motivo "DXD"?**

3. **¿Cómo se delimitan exactamente las regiones N y C terminal en las GTAs? ¿Qué función tiene cada una? ¿Es compartida esta función por todas las proteínas de esta superclase?**

4. **¿Existe alguna relación común entre las GTAs para la secuencia-estructura-función que permita realizar algún tipo de predicción, como la posición o la naturaleza de los sustratos?**

5. **¿Toda la secuencia de las GTAs sufre la misma presión evolutiva o esta se acentúa o relaja en según qué zonas?**

El estudio de las estructuras existentes para todas las GTAs, que realizaremos para el modelado de la GT MG517, nos servirá para definir una topología consenso y construir un perfil Hidden Markov Model (HMM) para esta topología. Así las regiones N y C terminal quedarán separadas secuencial y estructuralmente por este perfil, con el que además será posible alinear nuevas GTAs o definir cada una de estas dos regiones en proteínas no cristalizadas. Podremos determinar la posición del bucle estudiado en GpgS, en las regiones N y C terminal y saber si esta es una estructura compartida por el resto de proteínas, por lo que quizás también podría extrapolarse su mecanismo. Estudiaremos la posición de los ligandos, en todas las estructuras GTA depositadas en la base de datos PDB y analizaremos las regiones de secuencia implicadas y sus residuos, buscando elementos comunes entre las diferentes familias GTA. Gracias a la información suministrada por la base de datos CAZy, haremos uso de nuevos perfiles HMM y árboles filogenéticos, con los que construir herramientas de predicción que relacionen secuencia-estructura-función entre las GTAs.

IDENTIFICACIÓN DE RESIDUOS CATALÍTICOS EN LA PROTEÍNA MG517, A PARTIR DE LA GENERACIÓN DE UN MODELO TRIDIMENSIONAL.

Romero-García, J., Francisco, C., Biarnés, X. & Planas, A. ***Structure-function features of a Mycoplasma glycolipid synthase derived from structural data integration, molecular simulations, and mutational analysis****. PLoS One 8, e81990 (2013).*

1. Introducción

La proteína MG517 (glicosildiacilglicerol sintasa) es el producto del gen mg517, de la bacteria patógena *Micoplasma genitalium*. La misma enzima realiza, de forma secuencial, la transferencia de un azúcar a los glicerolípidos de membrana: monoglicosildiacilglicerol (MGlcDAG) y diglicosildiacilglicerol (DGlcDAG). La inactivación de la proteína conlleva la muerte del microorganismo, por lo que ha sido propuesta como una posible diana terapéutica.

Hasta la fecha no existen estructuras resueltas para ninguna glicoglicerolípido (GGL) sintasa[i]. Aunque se han construido modelos tridimensionales para dos tipos de GGL sintasas con plegamiento GTB: la glucosildiacilglicerol sintasa de *Acholeplasma laidlawii* y *Streptococcus pneumonia*[82] que pertenecen a la familia GT4 y la monogalactosildiacilglicerol sintasa de *Spinacia aleracea*[83] perteneciente a la familia GT28. Todos estos modelos realizados por homología, tomaron como plantilla la estructura de *E. coli* MurG, que es la única estructura actualmente disponible para la familia GT28. Por otro lado, para el plegamiento GTA, ninguna estructura para GGL-GT ha sido publicada.

Los intentos por obtener una estructura cristalográfica de GT MG517 han sido infructuosos, por ello nos decidimos a construir un modelo tridimensional de su estructura, mediante modelado por homología y técnicas de dinámica molecular de larga duración.

Los micoplasmas podrían ser descendientes de bacterias Gram-positivas (figura 20 y anexo figura 1), probablemente de origen Chlostridium[84]. Se cree que esta transformación se llevó a cabo mediante un proceso de reducción genómica, concluyendo en *M. genitalium*, uno de los genomas más pequeños de un procariota autorreplicativo.

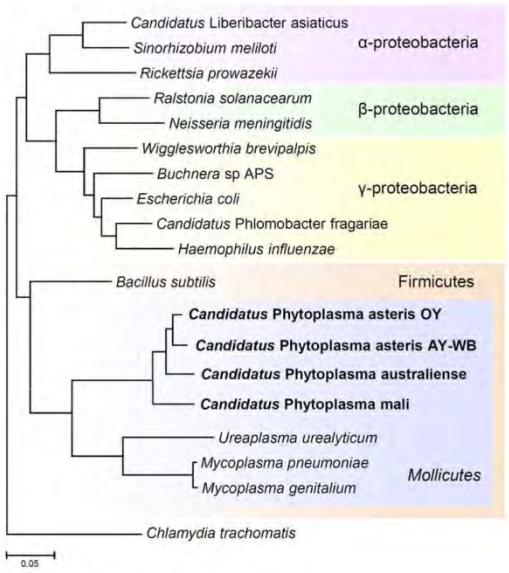


Figura20. Posición filogenética de Mollicutes entre las bacterias, usando el ARNr 16S.

Fuente: Genomic and evolutionary aspects of phytoplasmas. Kenro Oshima, Kensaku Maejima, and Shigetou Namba. Front Microbiol. 2013; 4: 230.

Puesto que la GT MG517 pertenece a la familia GT2, iniciamos la búsqueda de una plantilla con la que modelar esta proteína, entre las estructuras cristalizadas para esta familia de GTs.

[i] Tras la escritura de esta tesis se hizo pública la primera estructura tridimensional de una GGL. La Monogalactosyldiacylglycerol synthase 1 (MGD1) de *Arabidopsis thaliana*, de plegamiento GTB y perteneciente a la familia GT28. (PDB: 4WYI, 4X1T)

2. Selección de la región a modelar (N-terminal)

Se construyó un árbol filogenético (ver métodos) con las secuencias de las proteínas cristalizadas de la familia GT2 (2012): SpsA de *Bacillus subtilis* (UP P39621) (PDB 1H7Q), BF2801 de *Bacteroides fragilis* (UP Q5LBM4) (PDB 3BCV), Condroitin polimerasa de *Escherichia coli* (UP Q8L0V4) (PDB 2Z86) y Polimerasa de ácidos teicoicos de *Staphylococcus epidermidis* (Q5HLM5) (PDB 3L7I), con la intención de encontrar la secuencia (y su estructura) más cercana a la GT MG517. La proteína filogenéticamente más cercana era la de *Bacteroides fragilis,* pero el árbol carecía de la confianza suficiente como para aceptar este resultado.

Puesto que el género Micoplasma parece provenir de las bacterias Gram positivas, se realizó otro árbol con las secuencias de los Phyla Tenericutes y Firmicutes extraídas de UniProt, en total unas 18000 secuencias, de las cuales más de 1300 corresponden a la familia de las GT2, pero solo dos tienen estructuras cristalizadas. Estas se alinearon con el servidor PROMALS[85] y el árbol con el paquete PHYLIP[86] (ver métodos).

Tampoco el resultado mostraba la necesaria confianza (figura 21), evaluada mediante *bootstrap*.

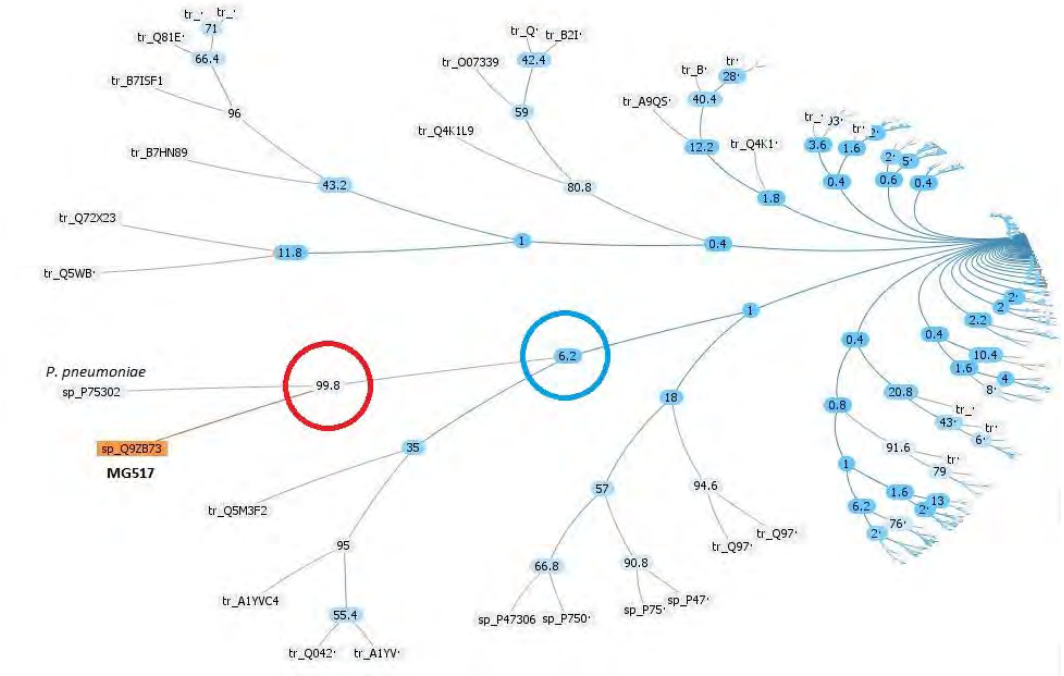


Figura 21. Representación hiperbólica del árbol filogenético para GT2 de Phylum Tenericutes y Firmicutes. El círculo rojo muestra el *bootstrap* para la relación filogenética entre *P. pneumoniae* y MG517, casi un 100%. El círculo azul y el resto de nodos en azul, muestra la relación filogenética entre los dos nodos que parten de él. Su bajo índice no permite confiar la relación encontrada.

No fue posible encontrar una proteína cristalizada cuya secuencia fuera lo suficientemente cercana a la de GT MG517. Con todo, se verificó que el plegamiento GTA era compartido por las

estructuras GT2 cristalizadas, por ello se comprobó hasta qué punto y cuánta estructura compartían las diferentes familias GT que se agrupan dentro de este tipo de plegamiento.

El listado de glicosiltransferasas GTA caracterizados hasta la fecha de este estudio (2013), con información estructural disponible, es un conjunto de secuencias/estructuras formado por 30 glicosiltransferasas de diferente origen y función (Tabla 3).

Tabla 3. Enzimas GTA con estructura cristalográfica resuelta en PDB. (Junio 2013)[a].

PDB	UniProt	Familia	Organismo / enzima	Secuencia (aa)	Dominio GTA (aa) [b]
1QG8	P39621	GT2	*Bacillus subtilis.* Spore coat polysaccharide biosynthesis protein (SpsA)	256	1-216
3L7J [c]	Q5HLM5	GT2	*Staphylococcus epidermidis.* Teichoic acid biosynthesis protein F	721	-
3BCV	Q5LBM4	GT2	*Bacteroides fragilis* (strain ATCC 25285 / NCTC 9343). Putative glycosyltransferase protein	342	3-226
2Z86	Q8L0V4	GT2	*Escherichia coli.* Chondroitin polymerase	686	148-388 431-630
4FIY	O53585	GT2	*Mycobacterium tuberculosis H37Rv* Galactofuranosyl transferase GlfT2 (GalTr)	637	158-398
4HG6	Q3J125	GT2	*Rhodobacter sphaeroides.* Possible Celulosa sintasa	788	139-375
1O7Q	P14769	GT6	*Bos taurus.* N-acetyllactosaminide alpha-1,3-galactosyltransferase Galactosyltransferase (α3GalT)	368	125-342
3IOH	P16442	GT6	*Homo sapiens.* Histo-blood group ABO system transferase	354	112-328
4AYL	A7LVT2	GT6	*Bacteroides ovatus.* Glycosyltransferase family 6	263	1-215
1NKH	P08037	GT7	*Bos taurus.* Beta-1,4-galactosyltransferase 1 (b4Gal-T1)	402	178-342
2FY7	P15291	GT7	*Homo sapiens.* Beta-1,4-galactosyltransferase 1 (b4Gal-T1)	398	174-338
3LW6	Q9VBZ9	GT7	*Drosophila melanogaster.* Beta-4-galactosyltransferase 7	322	73-235
1LL2	P13280	GT8	*Oryctolagus cuniculus.* Glycogenin-1	333	1-191
1G9R	P96945	GT8	*Neisseria meningitidis.* Glycosyl transferase	311	1-212
3TZT	C7RG54	GT8	*Anaerococcus prevotii* Glycosyl transferase family 8	273	1-214
3RMV	P46976	GT8	*Homo sapiens* Glycogenin-1	350	1-184
1FO8	P27115	GT13	*Oryctolagus cuniculus.* Alpha-1,3-mannosyl-glycoprotein 2-beta-Nacetylglucosaminyltransferase (GlcNAc-T I)	447	104-316
1S4N	P27809	GT15	*Saccharomyces cerevisiae.* Glycolipid 2-alpha-mannosyltransferase	442	119-390
1XHB	O08912	GT27	*Mus musculus.* Polypeptide N-acetylgalactosaminyltransferase 1 (ppGaNTase-T1)	559	114-346
2FFU	Q10471	GT27	*Homo sapiens.* Polypeptide N-acetylgalactosaminyltransferase 2 (ppGaNTase-T2)	571	134-361
2D7I	Q86SR1	GT27	*Homo sapiens.* Polypeptide N-acetylgalactosaminyltransferase 10 (ppGaNTase-T10)	603	143-372
3CU0	O94766	GT43	*Homo sapiens.* Galactosylgalactosylxylosylprotein 3-betaglucuronosyltransferase 3(GlcAT-I)	335	73-310
2D0J	Q9NPZ5	GT43	*Homo sapiens.* Galactosylgalactosylxylosylprotein 3-betaglucuronosyltransferase 2(GlcAT-S)	323	78-302
1V84	Q9P2W7	GT43	*Homo sapiens.* Galactosylgalactosylxylosylprotein 3-betaglucuronosyltransferase 1(GlcAT-P)	334	82-313

2ZU9	O58689	GT55	*Pyrococcus horikoshii.* Mannosyl-3-phosphoglycerate synthase (MPGS)	394	49-312
2WVL	Q72K30	GT55	*Thermus thermophilus.* Mannosyl-3-phosphoglycerate synthase (MpgS)	391	51-311
1OMZ	Q9ES89	GT64	*Mus musculus.* Exostosin-like 2	330	63-274
2BO4	Q9RFR0	GT78	*Rhodothermus marinus* (*Rhodothermus obamensis*). Mannosylglycerate synthase	397	1-218
3E26	O05309	GT81	*Mycobacterium tuberculosis.* Glycosyl-3-phosphoglycerate synthase (GpgS)	324	41-258
3CKJ	Q73WU1	GT81	*Mycobacterium avium* subsp. paratuberculosis K-10 Glycosyl-3-phosphoglycerate synthase (GpgS)	329	46-263
3O3P	B7SY86	GT81	*Rubrobacter xylanophilus* PRD-1 Mannosyl-3-phosphoglycerate synthase (MpgS)	387	40-256

[a] Para todos aquellos enzimas con más de un archivo PDB se escogió la estructura resuelta a mejor resolución.
[b] Residuos aminoacídicos correspondientes al dominio GTA y usados en el alineamiento.
[c] Los archivos PDB corresponden al dominio glicerofosfotransferasa de Q5HLM5, mientras que el dominio glicosiltransferasa no está resuelto.

De todos los cristales existentes para cada una de esas proteínas, se escogió el que presentaba mejor resolución. Se realizó una superposición rígida con estas estructuras utilizando el servidor POSA, con algunas modificaciones en ciertas proteínas:

- Se descarta la proteína F de *Staphylococcus epidermidis* (UP: Q5HLM5, PDB: 3L7J), ya que el fragmento cristalizado no incluye el dominio GT2.
- La condroitín polimerasa de *Escherichia coli* (UP: Q8L0V4, PDB:2Z86) posee dos dominios GT2 en su estructura, por lo que se separan por el bucle que los une (aa 390), utilizándose a partir de entonces dos secuencias y dos estructuras
- Las proteínas GALT1 (UP: O08912, PDB: 1XHB), GALT2 (PDB: Q10471, PDB: 2FFU) y GLT10 (UP: Q86SR1, PDB: 2D7I) poseen un dominio GT2 (PF00535) y otro tipo lectina (PF00652). Se elimina el dominio tipo lectina.
- La proteína β4GalT1 (UP P08037 PDB 1NKH), posee en su estructura la subunidad reguladora LA (cadena A, PF00062) que se elimina de la estructura.

POSA generó una superposición rígida con un núcleo estructural común de 59 aminoácidos y un RMSD de 2,67 Å. Cuando se visualiza esta superposición es patente el alto grado de conservación estructural existente entre las GTAs (figura 22).

La topología consenso para la estructura secundaria (figura 24) está formada por 7 hojas β que forman una lámina beta extendida y torsionada flanqueada por tres hélices α a cada lado de la plataforma que genera la lámina β. El motivo conservado DXD se localiza en el centro, entre la hoja β4 y la hélice α4, formando siempre un pequeño arco. Las hojas β5 y β7 se cruzan en la estructura, permitiendo la formación de una plataforma antiparalela de hojas β que se extiende más allá del motivo DXD. Aguas abajo de la hoja β7 existe una hélice α con un grado de conservación menor que el resto. Esta hélice α7 se haya ubicada en la misma posición, bajo la hélice α6, en la

mitad de las estructuras, mientras que para las otras o bien existe, pero se haya en otro lugar o no ha sido resuelta por cristalografía.

Figura 22. Superposición de las 30 estructuras estudiadas. Solo las estructuras que pertenecen a la topología consenso son mostradas (a excepción de la región variable mostrada en la figura 24).

Solo hasta la hoja β7 se puede afirmar la completa conservación de esta topología en todas las GTA, con la curiosa excepción de las GT7, quienes carecen de la hélice α2 y hoja β3, que son sin embargo reemplazadas por elementos estructuralmente homólogos situados en la extensión C-terminal, manteniéndose así el patrón estructural.

De este modo se decidió que solo sería posible modelar la GT MG517 hasta esta hoja β7, la que definimos como "región N-terminal".

Figura 23. Extensión de las regiones N y C terminal en la GT MG517.

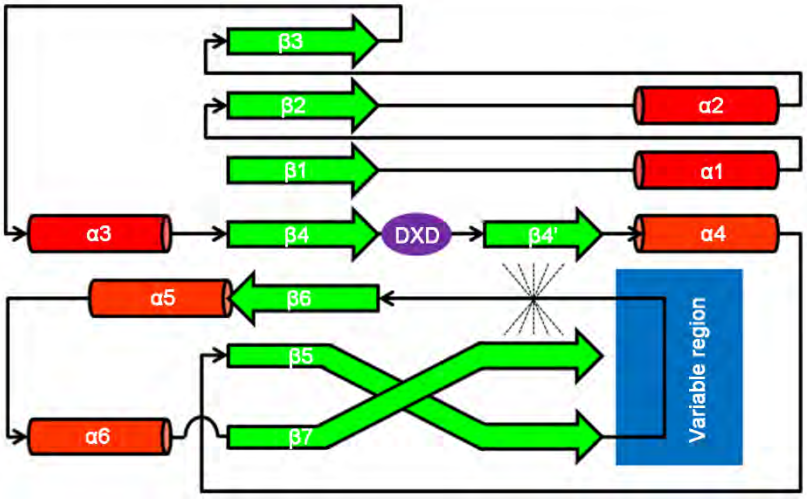


Figura 24. Topología consenso propuesta para el grupo de proteínas GTA.

Especialmente interesante en esta topología consenso es la región observable entre las hojas β5 y β6, donde existe una zona altamente variable para la que ningún patrón estructural pudo ser asignado. De hecho, esta zona es la razón de no haber podido escoger ninguna estructura como plantilla dentro de la familia GT2, ya que todos los cristales presentan una configuración diferente aquí. Algunos autores han considerado esta región como una posible zona de dimerización[87], pero no se ha ido más allá de esta idea.

Nota: Este estudio estructural se amplió posteriormente al conjunto de todos los cristales existentes para proteínas GTAs, hasta la fecha de presentación de este manuscrito y cuyos resultados se muestran en el capítulo 4: "La región variable como predictor de la especificidad por sustrato en proteínas GTA".

Conocida la topología consenso fue fácil seleccionar una plantilla con la que modelar la región conservada, ya que todas las estructuras la comparten, sin embargo, la región variable iba a dificultar este modelado, al no haber elementos que relacionen ninguna de las estructuras existentes con la GT MG517.

Llegado este punto, el modelado se abordó con un nuevo enfoque, **se elegiría una única plantilla para modelar la región conservada y cuatro diferentes plantillas para la región variable (RV), generando así cuatro diferentes modelos híbridos, que compartirían el mismo patrón estructural consenso con diferentes regiones variables. Cada uno de estos modelos se someterá después a MD, con la esperanza de que las diferentes RV converjan a una misma estructura común.**

El modelado por homología es muy dependiente del alineamiento de partida, entre la secuencia a modelar y la de su plantilla, por ello se requería un alineamiento preciso de la GT MG517 con la proteína que se usaría para modelar la región conservada y las que se usarían para la región variable, ninguna de las cuales había sido todavía seleccionada. Este alineamiento se realizó con el servidor T-coffee/Expresso que incorpora información estructural para el alineamiento y en el que incluso se puede seleccionar la estructura específica para cada secuencia.

El resultado de T-coffee, a pesar de ser bastante satisfactorio, no cumplía los requerimientos de precisión necesarios ya que era posible comprobar cómo ciertos residuos que se encontraban en posiciones homólogas en la superposición estructural, se encontraban alineados con residuos diferentes en el alineamiento de secuencias (figura 25). Se refinó entonces a mano el alineamiento utilizando la información estructural disponible por la superposición realizada, obteniendo un resultado que correlacionaba el alineamiento de residuos con la superposición estructural en todos los casos menos, en la región variable (Imagen del alineamiento aumentada en el anexo 2).

Para esta región variable, no existía homología de secuencia y de estructura entre diferentes familias, pero sí para proteínas de la misma familia (con la excepción de GT2, una familia muy heterogénea).

Nota: ¿Podría establecerse una relación aislada entre esta parte de la secuencia y la familia GT a la que pertenece la proteína? Esta idea, aún embrionaria, fue tomando forma a lo largo de la tesis y su estudio sistemático y conclusiones serán presentadas en el capítulo 4.

Con este alineamiento múltiple refinado a mano se construyó un perfil HMM para la familia GTA. Aunque no es posible incorporar información relevante para la región variable en este perfil HMM, este incluye las dos regiones conservadas que la flanquean. De este modo el perfil permite detectar y alinear adecuadamente estas regiones, para cualquier miembro del grupo de glicosiltransferasas con plegamiento GTA, aunque con diferentes resultados.

Este perfil para proteínas GTA se extiende más allá de cualquiera de los existentes en la base de datos PFAM, cuya información suele perderse más allá de la región variable y no incluye en ninguno de sus perfiles para GTAs, la hoja β7 ni parte de la hélice α6.

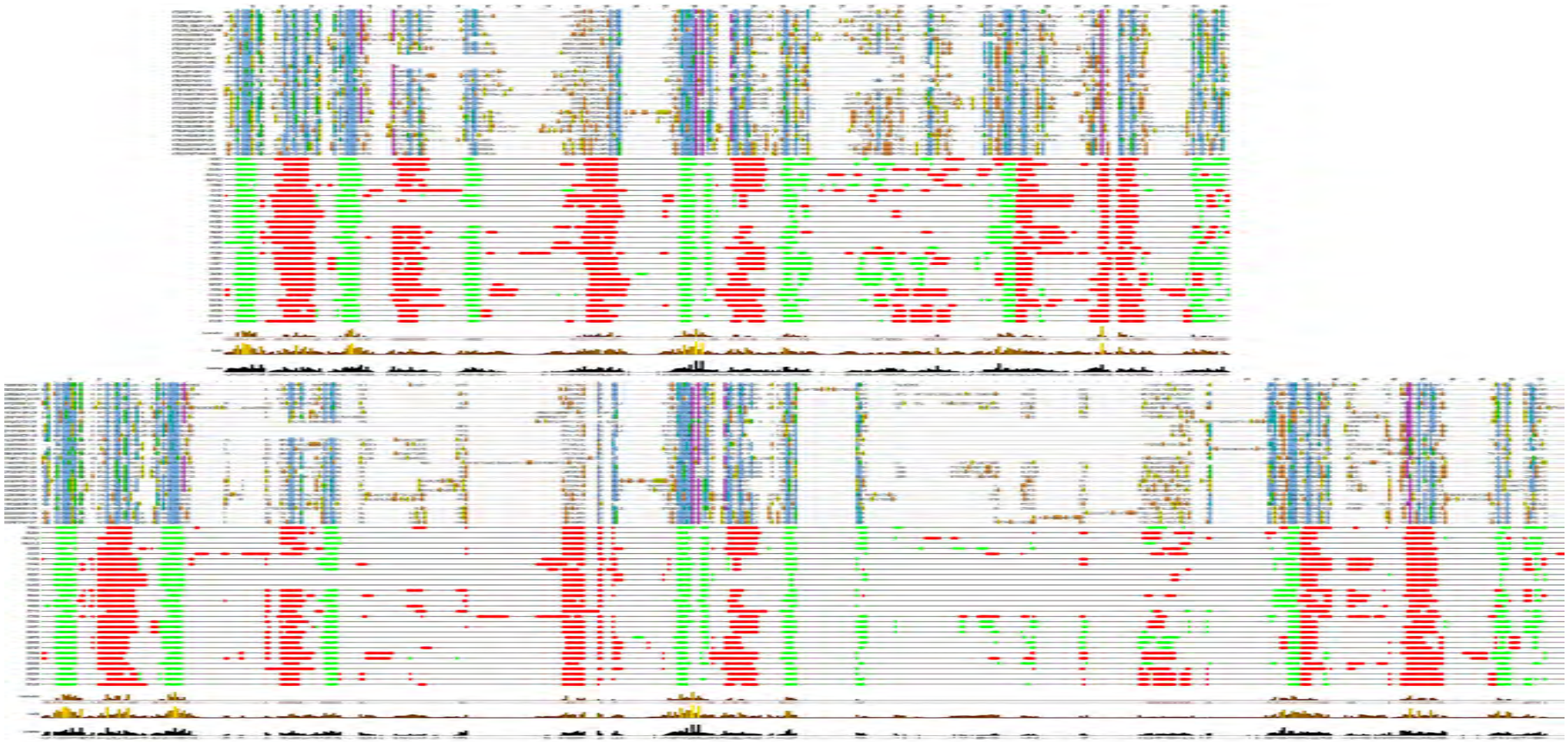


Figura 25. Alineamientos de la **región N terminal** de las proteínas mostradas en la tabla 3:

Arriba: Alineamiento refinado a mano corrigiendo las posiciones de los residuos según la superposición resultante del servidor POSA.

Abajo: Alineamiento resultante del servidor Tcoffee-Expressso

El refinamiento manual del alineamiento fue validado mediante la creación de un dendrograma (ver métodos). Con este alineamiento, las diferentes familias de proteínas son agrupadas juntas en sus respectivas ramas, con las excepciones de las proteínas O53585 (GlfT de *M. tuberculosis*) y Q3J125 (una posible celulosa sintasa de *R. sphaeroides*), asignadas a la familia GT2 por CAZy y que son agrupadas en una única rama, junto a las familias GT13 y GT64. La secuencia diana MG517 se localiza en la rama de las GT2, familia a la que pertenece según CAZy.

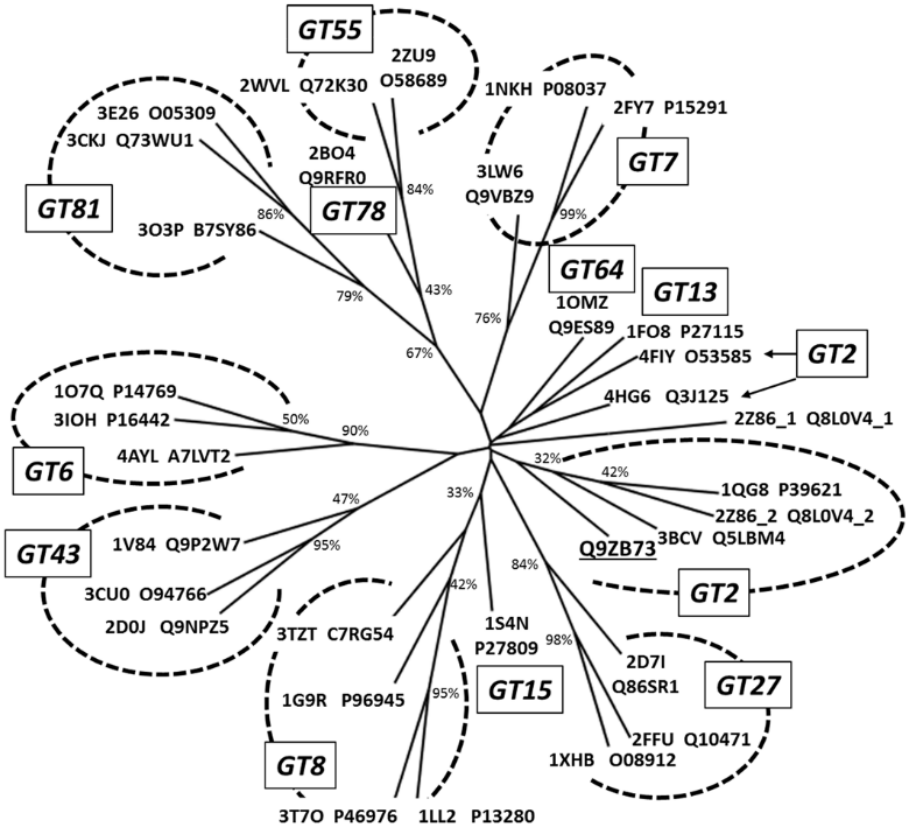


Figura 26. Dendrograma de proteínas GTA con estructura 3D conocida. Valores de *bootstrap* superiores al 30 % son señalados en los nodos.

Es interesante observar como ambos dominios GT de la condroitín polimerasa de *E. coli* son asignados a la familia GT2 por CAZy, pero según el dendrograma el dominio 1 (2Z86_1) estaría situado en una rama única, aislada del resto de familias, mientras que el dominio 2 (2Z86_2) sí que estaría con el resto de proteínas de la familia GT2.

3. Modelado

La región N-terminal de la GT MG517 (aa 1-220) muestra similitud de secuencia con la familia GTA, mientras que para la extensión C-terminal ha sido imposible encontrar homología con ninguna otra proteína. Por ello se decidió modelar la región GT N-terminal de MG517, que incluye la región variable.

Inicialmente se probaron diferentes servidores de modelado automático como HHPred, I-TASSER y LOMETS. Debido a la fuerte dependencia de los modelos respecto a sus plantillas, fundamentalmente en la región variable y a la imposibilidad de seleccionar un modelo "verdadero" entre ellos, por la ausencia de estructura homóloga en esta región, se descartó el uso de estos servidores, planteando una estrategia personal y diferente para el modelado de la proteína. La tabla 4 muestra las principales características de cada servidor y las plantillas seleccionadas para el modelado por cada servidor.

Tabla 4. Distintos servidores automáticos de modelado por homología.

Servidor	**HHPRED**	**I-TASSER**	**LOMETS**
Plantilla[a]	2Z86/3BCV/1QG8	3BCV/1XHB/2D7I/2Z86	2Z86
Selección[b]	Plantillas óptimas	Automático	Automático
Base de datos[c]	PDB	No redundante (automático)	No redundante (automático)
Modelo[d]	Estructuras fusionadas	*Threading*/Fragmentos combinados	*Threading*

[a] Plantillas usadas por cada servidor para crear el modelo.
[b] Selección de la plantilla. Solo HHPRED permite algún tipo de selección de la plantilla ("*by user*").
[c] Base de datos utilizada por el servidor para encontrar la plantilla. Solo HHPRED permite seleccionar la base de datos.
[d] Forma de crear el modelo final. HHPRED fusiona las plantillas si se han seleccionado varias. I-TASSER combina diferentes fragmentos de diferentes plantillas por *threading*.

Por esta dependencia de los modelos, con respecto a las plantillas para la región variable, no es posible seleccionar un modelo "correcto" entre ellos. Por tanto, **la estrategia elegida para el modelado de MG517 fue la construcción de modelos híbridos, mediante una combinación de plantillas para las regiones conservada y variable.**

La secuencia homóloga más cercana a la de MG517 es, según el árbol de agrupamiento realizado (figura 26), la de la GT2 de *Bacteroides fragilis*, de modo que se seleccionó para el modelado de la región conservada. Pero una primera ronda de generación de modelos usando esta estructura como plantilla, mostró que todos ellos perdían parte de la topología consenso. Esto sucedía principalmente porque la hoja β7 no está resuelta en el cristal de *Bacteroides fragilis* y por tanto los modelos carecen de información para esta región. Entonces una segunda ronda de modelos fue construida utilizando la siguiente estructura GT2 más cercana, según el árbol de agrupamientos, que corresponde al segundo dominio GTA de la condroitín polimerasa de *Escherichia coli* (2Z86_2) que tiene toda la hoja β7 completamente resuelta (y un 23% de identidad de secuencia con MG517, estando en el límite de seguridad para un modelado por homología). Así, la estructura tridimensional de la región conservada de MG517 (residuos 1 a 121 y 174 a 220) se modeló usando esta estructura como plantilla.

Puesto que no es posible asignar ninguna estructura consenso a la región variable en las GTA, el modelado de esta región (residuos 122 a 173) para MG517 se realizó utilizando diferentes plantillas de estructuras GTA, seleccionadas según los siguientes criterios:

- Una estructura para cada familia GTA.
- Longitud de secuencia similar a la región variable de MG517.
- Estructura resuelta en complejo con algún ligando.

Fueron seleccionadas cuatro estructuras para ser usadas como plantillas para la región variable. (Características de los cristales: Anexo 3):

1. α3GalT, una GT6 de *Bos Taurus* (1O7Q, 13% de identidad).
2. ppGaNAcT, una GT27 de *Homo sapiens* (2FFU, 8% de identidad).
3. Condroitín polimerasa, una GT2 de *Escherichia coli* (2Z86_1, 4% de identidad).
4. GlcAT-I, una GT43 de *Homo sapiens* (3CU0, 4% de identidad).

En oposición a un modelado *de novo* para esta región, esta forma de proceder reduciría el espacio conformacional de la región variable a geometrías ya identificadas para el grupo de proteínas con plegamiento GTA.

Se construyeron entonces cuatro modelos estructurales diferentes de MG517 (residuos 1 a 220), con una misma plantilla para la región conservada en todos los modelos, la de la estructura 2Z86_2, más una de las cuatro plantillas elegidas para la región variable. Cada modelo también contenía los ligandos de las estructuras plantilla para la región variable. Estas nuevas rondas de modelado generaron estructuras que sí contenían la topología consenso de las GTA en todas ellas, seis α hélices y siete hojas β conservadas en la misma posición de la plantilla original.

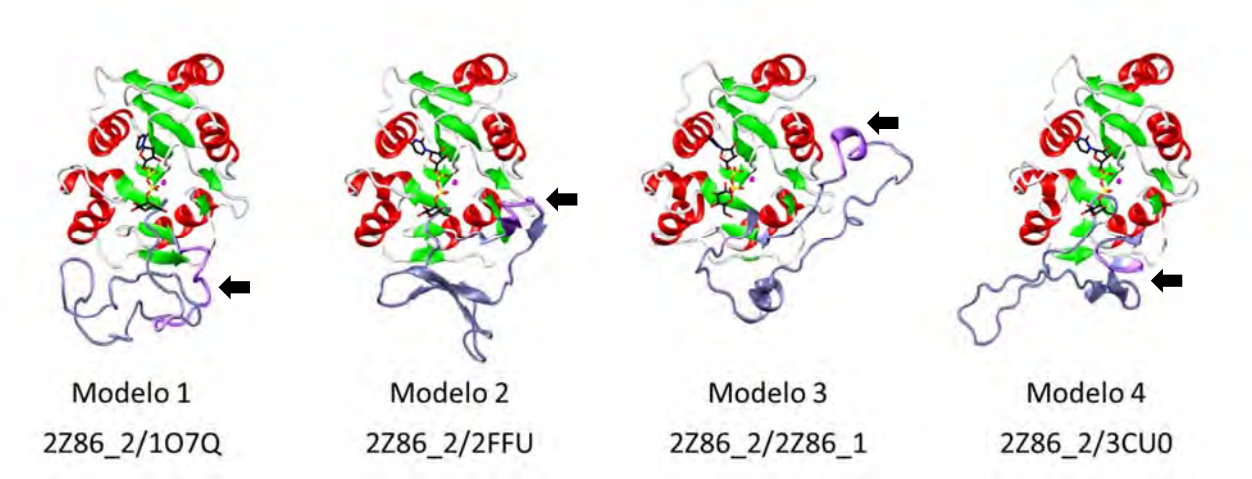


Figura 27. Modelos seleccionados para los diferentes híbridos generados con Modeller. La región variable está coloreada en azul; el segmento que tras la MD convergerá a una hélice α se muestra en un tono morado (señalado con una flecha). Figuras aumentadas en el anexo 7.

Solo la hoja β6 es antiparalela a las demás, interaccionando con la hoja β4 que precede al motivo DXD. Justo a continuación de este motivo se encuentra la pequeña hoja β4', formando una lámina β con las hojas β7 y β5 en dos de los modelos (1 y 3) y solo con la hoja β7 en los otros dos modelos (modelos 2 y 4). La región variable se localiza entre las hojas β5 y β6 y cada estructura mantiene el plegamiento de su propia plantilla: El modelo 1 (1O7Q/2Z86_2) muestra una larga cadena desestructurada, el modelo 2 (2FFU/2Z86_2) posee cuatro hojas β, el modelo 3 (2Z86_1/2Z86_2) presenta solo dos hélices, en lugar de las tres existentes en la plantilla y el modelo 4 (3CU0/2Z86_2) 2 de las 4 estructuras secundarias de la plantilla (todo hojas β) (Anexo 7). En este momento es evidente que no existe ninguna topología consenso entre los modelos para esta región, la variable. Sin embargo, para la región conservada, las principales diferencias entre los distintos modelos se encuentran solo en los bucles, además todos los ángulos de la cadena principal en estos modelos están localizados en regiones permitidas del diagrama de Ramachandran (Anexo 6) y un *DOPE Z-score* normalizado de -0.3.

4. Sitio de unión del dador

Se aprovechó la generación de modelos para estudiar la posición de los residuos de MG517, cercanos al ligando dador, ya que cada ligando pertenecía a una plantilla distinta (proteína distinta, la correspondiente a la región variable) y sería posible observar el grado de interacción entre zona conservada modelada y ligando, y compararla después con el resultado de las dinámicas.

Para cada uno de los cuatro diferentes modelos se generaron 100 estructuras diferentes (ver métodos) y todas ellas contenían el sustrato dador UDPGlc de la plantilla original para la RV (con la excepción del modelo 3, cuyo sustrato es el UDPGlcA y que fue modificado a UDPGlc).

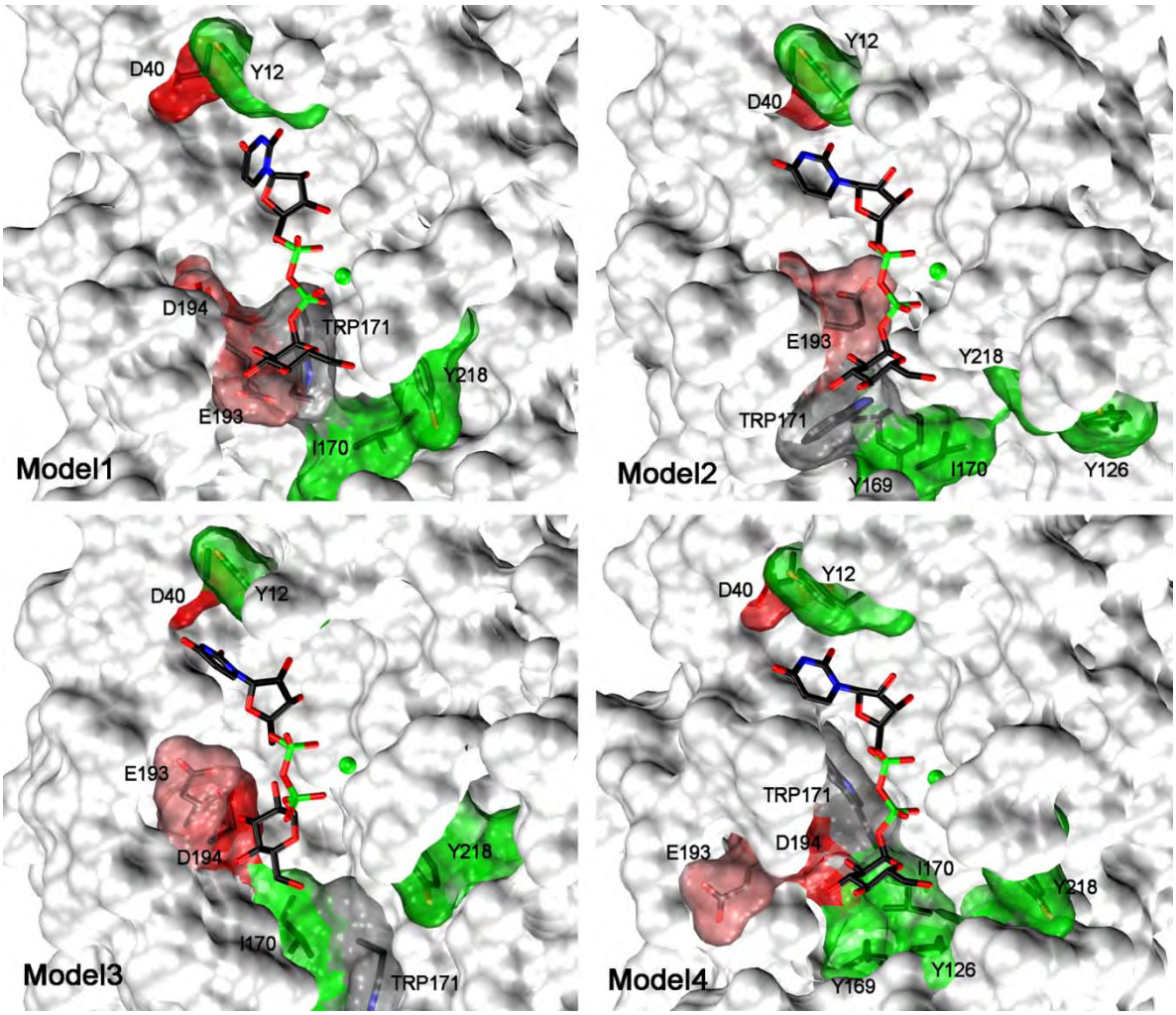


Figura 28. Ubicación del dador en las estructuras seleccionadas para los modelos 1 a 4.

Todo el conjunto de estructuras (las 400) se usó para identificar los residuos que se encuentran a una distancia ≤ 4 Å del dador. Se obtuvo una lista de 36 residuos (tabla 5) donde 9 de ellos, situados

en la región conservada, se seleccionaron por su potencial implicación en la unión con el ligando y en la catálisis, para ser estudiados mediante experimentos de mutagénesis dirigida.

Tabla 5. Residuos de la GT MG517 (Nt dominio GT, aa 1-220) localizados a ≤ 4Å del dador UDPGlc en los cuatro modelos. Se muestra el número de estructuras de cada modelo cuya posición del residuo se haya a la distancia analizada. Se han resaltado los residuos que fueron seleccionados para experimentos de mutagénesis. R: Residuo. Modelos: 1, 2, 3, y 4.

R	1	2	3	4	R	1	2	3	4	R	1	2	3	4
10	100	100	100	100	**77**	20				**169**	3			
11	100	80		50	**93**	100	100	100	100	**170**				10
12	100	100	100	100	**94**	100	100	100	100	**171**	40	70		
13			10		**95**	90	100	40	100	**172**		70		
14				10	**125**				50	**190**	6	7	14	7
40	10	40	100	10	**126**				60	**191**		8	22	8
68	3	6	3	4	**127**				20	**192**	58	68	81	63
69	4	5	8	2	**128**				20	**193**	69	56	99	65
70			53		**137**	6				**194**	16	2	100	8
72			100		**138**	11				**218**			10	
73	100	100	100	100	**139**	10				**219**	10	20		30
76		30	100	50	**168**	2		23		**220**	80	80		50

Estos 9 residuos (más la posición 138) se seleccionaron entre los demás por encontrarse conservados en el alineamiento de secuencia y /o haberse informado de mutaciones en posiciones equivalentes en otras proteínas. Las razones para cada mutante son las siguientes:

Y12 parece mantener una interacción por apilamiento (*stacking*) con el anillo de uracilo del UDP.

- **Y12M**: Conservando el tamaño se anula el *stacking* por eliminación del anillo. Se espera que se reduzca la actividad, al disminuir la estabilidad del dador.
- **Y12A**: Se anula el *stacking* por eliminación del anillo y se modifica el tamaño de la cadena lateral. Una disminución de la actividad aquí y no en el mutante anterior, se debería a una desestructuración de la zona y la hipótesis del apantallamiento se desmontaría.

D40 podría estabilizar al UDP por interacciones electrostáticas. La mutación de D40 debería verse amortiguada por las cargas de alrededor y no tendría efecto[88,89]. Se comprobará.

- **D40A**: Se anula la carga del residuo manteniendo su tamaño. Se espera que no tenga efecto.
- **D40K**: Se intercambia la carga y se modifica el tamaño. Con esto se intenta modificar lo máximo posible el papel de este residuo.

Y126 e Y169 son los residuos que flanquean la RV y además se encuentran cercanos tanto el uno al otro como a la Glc del dador en dos de los modelos (1 y 4). La mutación de un residuo en este entorno podría reducir la actividad catalítica[90].

- **Y126A:** Se espera reducción de la actividad enzimática.
- **Y126F**: Se espera que la mutación no tenga efecto.
- **Y169A:** Se espera reducción de la actividad enzimática.
- **Y169F**: Se espera que la mutación no tenga efecto.

I170 se encuentra también cerca del azúcar del dador y es clave para la especificidad del azúcar transferido en algunas enzimas[90].

- **I170A:** Se espera reducción de la actividad enzimática.

W171 se coloca preferentemente entre el UDP y la Glc. La gran mayoría de las proteínas cristalizadas poseen una Gly en este lugar, si bien existen un gran número de proteínas que tienen Trp. La gran diferencia entre ambos residuos indica una estrategia evolutiva seleccionada y diferente para las distintas proteínas. Se propone mutarlo para ver qué ocurre.

- **W171G:** La actividad debería verse drásticamente disminuida.
- **W171A:** Quizás alguna reducción de la actividad, pero no tanto como con el anterior mutante.

E193 o D194 podrían ser la base catalítica.

- **E193A:** Se espera reducción o anulación de la actividad enzimática si es ésta la base o ningún efecto si no lo es.
- **D194A**: Como en el caso anterior, se espera reducción o anulación de la actividad enzimática si es ésta la base o ningún efecto si no lo es.

Y218 sustituye a una histidina enormemente conservada en las enzimas GTA.

- **Y218A**: Se espera reducción de la actividad.

F138 también se consideró para el experimento de mutagénesis.

- **F138A**: Se localiza en la RV, cerca del posible sitio de unión del aceptor en dos de los modelos mientras que para los otros dos su orientación es diferente y alejada de este sitio, así con él va a ser posible discriminar entre modelos.

Los residuos del motivo DXD (D93, P94, D95) no fueron seleccionados ya que su papel es bien conocido en las enzimas GTA.

5. Evaluación funcional por mutagénesis dirigida

Estos experimentos fueron realizados por mi compañero de departamento y doctorando, Carles Francisco. Los mutantes se prepararon mediante mutagénesis dirigida y su actividad glicosiltransferasa medida por dos ensayos complementarios. En el primero se midió la expresión de glicoglicerolípidos (GGL) en células recombinantes de *E. coli*, ya que esta bacteria no los produce de forma natural, por medio de capas finas (figura 29).

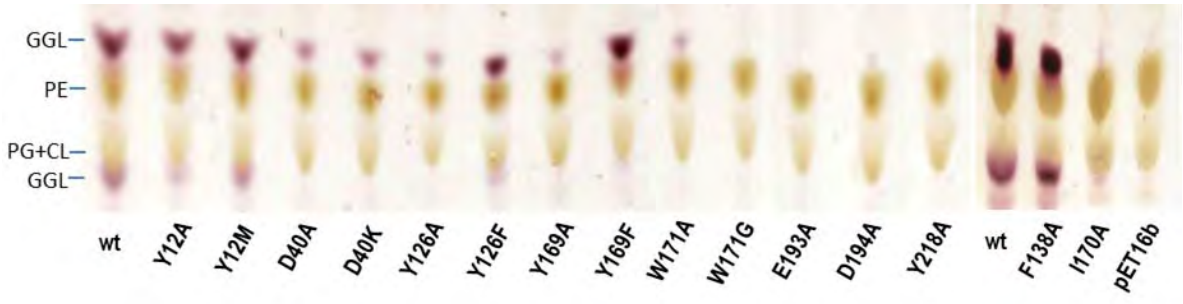


Figura 29. Cromatografía en capa fina de la extracción lipídica de los diversos mutantes.

Se pueden observar tres grupos de mutantes: los que tienen una producción similar de GGL que la enzima nativa (Y12A, Y12M, Y126F, F138A, y Y169F), aquellos muestran una actividad reducida (D40A, D40K, Y126A, Y169A, y W171A) y mutantes donde no se detecta ninguna producción de GGL (I170, W171G, E193A, D194A, y Y218A).

A continuación, la actividad específica se determinó en extractos celulares solubilizados (ensayos *in vitro*) del siguiente modo: La concentración de proteína total se determina por ensayo BCA. La actividad se determinó a concentración 1 mM de UDPGal, 100 μM Cer-NBD solubilizado con BSA (25μM), en 10 mM HEPES pH 8.0, 10 mM CHAPS, 10% glicerol), 10 mM $MgCl_2$ a 25ºC. La actividad específica bajo estas condiciones (v_0/[prot]) se expresó como la tasa inicial de formación de producto por miligramo de proteína total en el extracto (v_0 ($\mu M \cdot min^{-1}$)).

Los resultados se resumen en la tabla 6:

Tabla 6. Actividad enzimática de mutantes GT MG517.

Mutante	Vo/[prot] (μM/mg.min)	actividad %
MG517 wt	14.5	100
Y12A	7.05	49
Y12M	8.91	61
D40A	0.17	1,2
D40K	0.25	1,7
Y126A	2.9	20
Y126F	5.55	38
F138A	15.9	109
Y169A	2.56	18
Y169F	**54.0**	**372**
I170A	0.32	2.2
W171A	6.25	43
W171G	0.02	0,13
E193A	**0**	**0.00**
D194A	0.0035	0.02
Y218A	0.47	3.2

El papel que desempeña cada residuo analizado se basó en su localización estructural, según nuestros modelos (figura 28) y la comparación con otras enzimas GTA homólogas, como ya se avanzó en la explicación de cada mutante:

- La Y12 está bastante bien conservada en el alineamiento de secuencia, con un 33 % de identidad entre las proteínas GTA que tienen su estructura tridimensional resuelta. Forma parte de un bolsillo hidrofóbico, constituido por residuos al final de la hoja β1, que acomoda el anillo de uracilo del dador UDPGlc. Por ejemplo, *Mycobacterium* MAP2569c y GpgS (PDB 3CKJ y 3E26)[87,91] tienen una Leu (L57 y L52, respectivamente) en la posición estructuralmente equivalente que realiza *stacking* con el anillo de uracilo. Los mutantes Y12A e Y12M conservan el 49 % y el 61 % de la actividad nativa de la enzima, lo cual indica que no es un residuo esencial y que la posible interacción por *stacking* no es crítica para su actividad.

- El D40 es otro residuo bien conservado, con un 50 % de identidad. Este forma parte de una tétrada de aspartatos propuestos como elementos de reconocimiento y catálisis en las familias GT2, GT7, GT13 y GT43, según la estructura por rayos X de SpsA (D39-D98-D99-D191) (PDB 1QG8)[92]. El D40 al final de la hoja β2 es equivalente al D39 en SpsA,

que coordina el N3 del uracilo en el UDP. Los mutantes D40A y D40K tienen una actividad fuertemente reducida (<2 % que la GT MG517 nativa), de acuerdo a su papel propuesto. Sin embargo, mutaciones similares en proteínas homólogas tienen diferentes efectos. El mutante D44A en la ExoM de *Sinorhizobium meliloti* da como resultado la total pérdida de actividad[93], así como el mutante D156Q en la murina ppGaNTase-T1 (PDB 1XHB) que retiene solo el 0.1 % de actividad de la proteína nativa, por el contrario el mutante D41A de la proteína WbbE en *Salmonella* no parece afectar a su actividad[89].

- La Y126 se localiza al final de la hoja β5, justo al comienzo de la RV y cercana a la unidad de glucosa del dador UDPGlc en los modelos 1 y 4, y donde el carbono Cα está cercano (≈ 9 Å) al carbono Cα de la Y169, al final de la RV. Los mutantes Y26A y Y126F conservan en torno a un 20 % y un 38 % de la actividad de la GT MG517 nativa, sugiriendo que Y126 se haya involucrado en la unión del sustrato. En la misma región, al final de la hoja β5, la enzima α3GalT (PDB 1O7Q) tiene una Q247 que forma múltiples interacciones con el sustrato aceptor y la mutación Q247E reduce significativamente la actividad transferasa[90].

- La F138 se ubica en la RV, con su cadena lateral cercana a la zona putativa de unión del aceptor en los modelos 1 y 2. La mutación a Ala (F138A) no tiene efecto sobre la actividad, sugiriendo que este residuo no interacciona con los sustratos. Por tanto, los modelos 1 y 2 parece ser menos apropiados que los modelos 3 y 4 para describir la región variable.

- La Y169, I170 y W171 se encuentran situados al comienzo de hoja β6, después de la RV, pero no son residuos conservados. Y169 se haya cercana a Y126 al comienzo de la RV en los modelos 1 y 4 así como del sitio de unión del aceptor. El mutante Y169A mantiene un 18 % de actividad, más o menos acorde con el papel propuesto. Una sorpresa ha sido el mutante Y169F, el cual aumenta la actividad de la proteína a un 370 % respecto a la GT MG517 nativa. Además, este mutante muestra un perfil de productos diferente (en el ensayo de actividad *in vitro*), generando principalmente productos monoglicosilados (MGDAG) y prácticamente nada de diglicosilados (DGDAG). Una posible interpretación es que la eliminación del grupo OH de la Y169 incrementa la hidrofobicidad del sitio de unión del aceptor, favoreciendo la unión del lípido y provocando el efecto contrario en la unión del primer producto glicosilado, que coloca una Glc, más polar, en la misma posición para la segunda transferencia de azúcar. Este es desde luego un mutante interesante que merece una especial atención para futuros estudios que caractericen sus propiedades bioquímicas.

- I170 se haya ubicado en una región que alinea con la del residuo H280 en α3GalT[90] y con las posiciones 266 y 268 de las transferasas que definen los grupos sanguíneos AB0 (p. ej. PDB 3IOH), residuos de conocida función en la especificidad del nucleósido dador con Gal o N-acetil-Gal como azúcares a transferir[94]. El mutante I170A en la GT MG517 tiene una actividad fuertemente reducida (<2 % que en la enzima nativa), lo que podría suponer

que este residuo tiene también un importante papel en la unión de los ligandos, quizá en la especificidad del sustrato dador.

- El residuo W171 de la GT MG517 está conservado en un 20 % de las secuencias GTA (6158 secuencias de la base datos CAZy, alineadas con nuestro perfil HMM para todas las GTA), mientras que para el resto el residuo mayoritariamente encontrado es la Gly. Las proteínas GTA cristalizadas pertenecen a este segundo grupo, donde cerca del 70 % de las estructuras tienen una G y ninguna, salvo MG517 y la posible excepción de la proteína GlfT2 (PDB 4FIY), tiene un W en esta posición. El mutante W171G conserva tan solo un 0,1 % de la actividad nativa. Este resultado parece indicar que estos dos grupos de GTA podrían haber evolucionado de forma separada para acomodar o bien el W o bien la G, no siendo intercambiables los residuos, de acuerdo con lo esperado. Es interesante sin embargo observar que el mutante W171A mantiene un 43 % de la actividad nativa; aquí una cadena lateral mayor que la de la de G recupera parte de la actividad.

- E193 o D194 son los principales candidatos para actuar como base general en el mecanismo catalítico de la GT MG517. Uno de ellos, decíamos, podría ser parte de la tétrada de Asp propuesta como elemento de reconocimiento y catálisis[92] junto con el D40, D93 y D95 (los dos últimos formando parte del motivo DXD). En este caso el residuo D194 sería la primera opción, sin embargo, el mutante D194A todavía mantiene actividad detectable (0,02 %) mientras que el mutante E193A es completamente inactivo (confirmado por ensayos de actividad con la enzima purificada en alta concentración), por tanto, se propone al residuo E193 como base general para desprotonar el aceptor hidroxilo en el mecanismo de la GT MG517.

- Por último, Y218 se localiza en un bucle justo a continuación de la hoja β7 al final del dominio conservado GTA. Este residuo alinea con H258 en GpgS, que corresponde con una His altamente conservada en las enzimas GTA y que juega un importante papel en la unión con el metal[87],[91],[81]. Y así parece ser aquí, el mutante Y218A conserva solo el 3 % de la actividad nativa. Este residuo podría, en la GT MG517, coordinar el catión divalente junto al residuo D95 y los dos fosfatos del UDP.

6. Mejora de los modelos por Dinámica Molecular de larga duración

Los cuatro modelos estructurales realizados mediante modelado por homología con plantillas híbridas, mantienen la topología consenso para las proteínas GTA. Sin embargo, no es posible asignar ninguna estructura consenso para la región variable ya que esta es muy dependiente de la estructura para la plantilla elegida. Por ello se desarrollaron una serie de simulaciones de larga duración por medio de Dinámicas Moleculares (MD) para cada modelo, con la intención de que la RV recuperase parte de la estructura nativa en la GT MG517.

La baja identidad de secuencia entre la RV de MG517 y las plantillas correspondientes podrían haber introducido artefactos en las estructuras generadas, así que solo fue seleccionada la mejor estructura, según una puntuación DOPE normalizada (que proporciona *Modeller*) y mejor distribución de ángulos para la cadena principal (a través de *PROCHECK*), en cada ronda de modelos para ejecutar las simulaciones. Las simulaciones, tras un proceso de equillibrado (ver métodos), alcanzaron el estado estacionario en cuanto a conformación entre los 600 y 850 ns (dependiendo del modelo) con un RMSD para los átomos de la cadena principal de unos 3,5 Å de media y no fue detectado ningún cambio en los elementos que forman la topología consenso para las GTAs.

Al final de las simulaciones, las cuatro estructuras híbridas mantuvieron el plegamiento global y, de forma general, todos los elementos de estructura secundaria para la región conservada por el modelado por homología se mantuvieron también. Ningún cambio conformacional importante sucedió en esta región (tabla 7).

Cosa diferente es lo que ocurrió en la región variable. Las cuatro simulaciones sufrieron grandes cambios conformacionales y aunque no se encontró ningún plegamiento global común, este sí que se dio en un elemento concreto de esta zona, una hélice α al comienzo de la RV (figura 30). Es importante resaltar que esta hélice α no existía en las estructuras de partida para las simulaciones en tres de los 4 modelos. De hecho, las RV de los modelos 2 y 4 estaban formadas principalmente por hojas β. Para estos modelos, a la finalización de las simulaciones las primeras dos hojas β se convirtieron en una hélice α localizada en la misma posición, mientras que el resto de hojas β mantuvieron su conformación. Así mismo, la RV del modelo 1 tenía solo una pequeña hélice 3_{10} de tan solo 3 residuos; no solo se mantuvo esta hélice, sino que se extendió como hélice α reconformando parte de la zona desestructurada posterior. Por último, la RV del modelo 3 ya presentaba 2 hélices α, en lugar de hojas β. Las dos hélices se mantuvieron al finalizar la simulación, aunque cambiaron su conformación espacial. Comparado con los otros modelos, la posición inicial de la hélice α al inicio de la RV está desplazada unos 5 residuos aguas abajo.

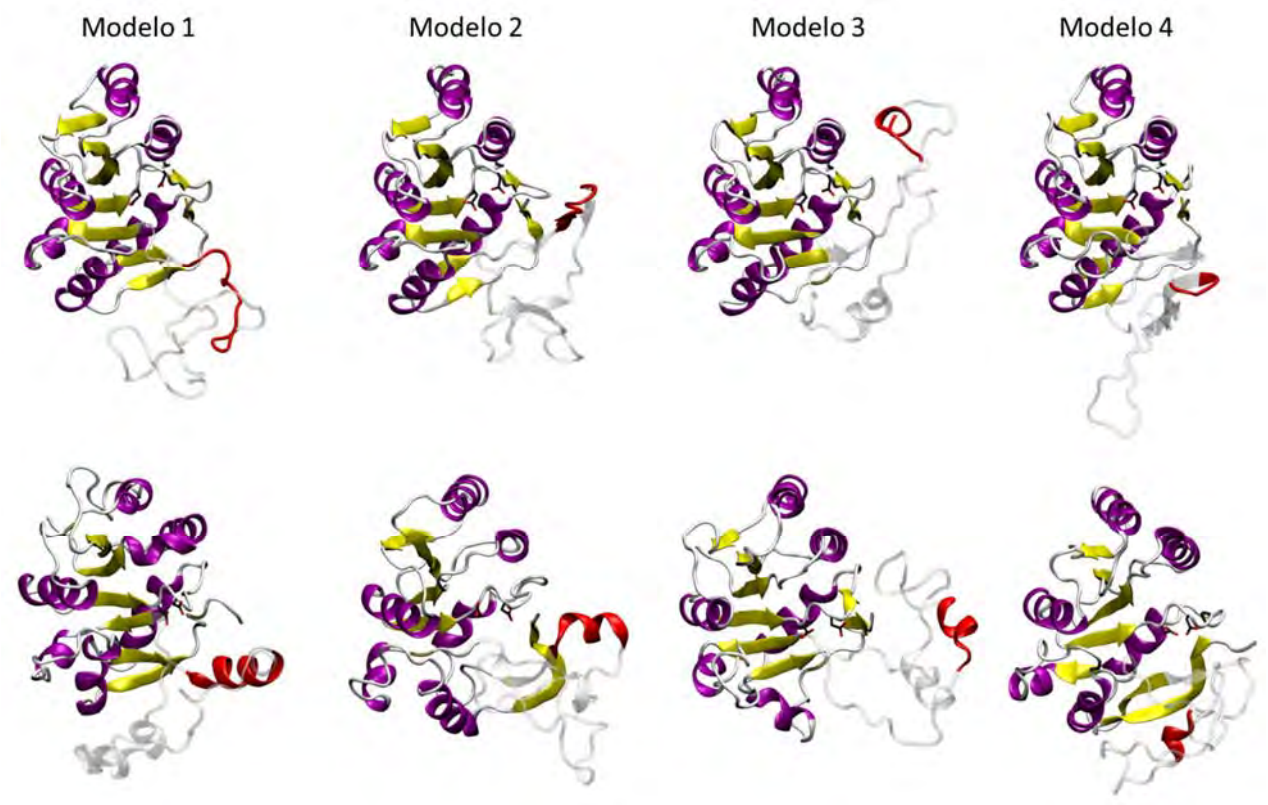


Figura 30. Estructura inicial de los modelos (arriba) y una vez finalizada la Dinámica Molecular (abajo). La estructura de la región consevada se muestra coloreada (amarillo: hojas β, magenta: hélices α). La RV se muestra traslúcida, salvo la región que ha convergido a una hélice α (en rojo).

Tabla 7. Resumen de las Dinámicas Moleculares y *clustering*.

Estructuras híbridas	Estado estacionario (ns) [a]	RMSD *cutoff* (nm) [b]	RMSD rangos (nm) [b]	Media RMSD (nm) [b]	RMSD *Cluster* seleccionado (nm) [b]	RMSD *Frame* seleccionado (nm) [b]	Estructura representativa [c]	Hélice formada [d]
Modelo1	850 – 1000	0.12	0.061 - 0.33	0.181	0.127	0.112	986.4 ns	S123-D133
Modelo2	800 – 1000	0.12	0.051 - 0.326	0.151	0.123	0.105	879.2 ns	Y126-D133
Modelo3	600 – 1000	0.13	0.061 - 0.700	0.265	0.136	0.116	646.9 ns	K131-K137
Modelo4	750 – 1000	0.14	0.061 - 0.568	0.135	0.135	0.114	960 ns	C128-K131

[a] Trayectoria usada para el análisis de agrupamiento (*Cluster annalysis)*. El momento de inicio fue cuando la fluctuación RMSD de la cadena principal era <0.5 Å.
[b] Análisis de agrupamiento.
[c] Estructuras seleccionadas como modelos representativos por MD.
[d] Posición de la hélice α común, generada por convergencia durante la MD.

Tabla 8. Media RMSD y fluctuación RMS (en paréntesis) para la estructura completa y los residuos seleccionados.

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**
RMSD (Å)				
Proteína completa	3.5	4.25	4.3	3.15
Y12	4.8 (1)	2.7 (1)	3.8 (1)	2.4 (1)
D40	3.5 (1)	6.2 (1)	3.7 (1.2)	2.4 (1)
Y126	4.9 (0.6)	3.2 (0.5)	10.7 (0.7)	10 (1.7)
F138	**7.5**	**7.5**	**17.7**	**18.2**
Y169	1.7 (0.6)	4 (0.6)	3.4 (1.4)	2.3 (2)
I170	2.4 (0.6)	5.8 (0.6)	3.1 (0.8)	2.7 (2.4)
W171	3.2 (0.6)	6.1 (0.7)	1.3 (0.7)	3.4 (2.2)
E193	3.1 (1.4)	8.8 (1)	9.7 (1.6)	2.3 (1.4)
D194	3 (1)	4.2 (1)	9.1 (1.5)	1.6 (1.2)
Distancia (Å)				
Y126-Y169 (Modelo)	9.6	15.9	18.9	9.5
Y126-Y169 (MD)	9.8	13.8	11.8	13.6

Los residuos seleccionados para la mutagénesis dirigida se analizaron por RMSD en las estructuras finales –después de las simulaciones por MD–, para comprobar el desvío individual de posición respecto a la de los modelos iniciales. El residuo que más se desvía es la F138, debido a la reestructuración de la región variable durante las simulaciones (en negrita en la tabla 8). El resto de los residuos permanecen en el valor RMSD medio de 4.3 Å. Valores mayores se vieron para E193 y D194 en el modelo3, seguramente por su localización en una región no estructurada de la hélice 6. Otro desplazamiento visible es el de la Y126 en los modelos 3 y 4, probablemente como consecuencia de las reestructuraciones de la región variable. La fluctuación durante la fase estacionaria fue pequeña (en paréntesis). La distancia entre la Y126 y la Y169, notablemente diferente entre modelos, convergió a un valor de 12 ± 3 Å en las estructuras finales.

Movimiento del dador.

Como resultado de las simulaciones, se observó que el dador en todas las trayectorias, se desplazaba del sitio activo, sin llegar a salir de él, ya que la interacción de los fosfatos con el metal (interacción electrostática) se mantenía, pero ocupando posiciones no catalíticas.

La rotura de la interacción entre el N3 del anillo de uracilo del UDP y el residuo D40 permite una oscilación del ligando que termina sacando a este de su adecuada posición en el sitio de unión, tal como se encuentran en los cristales en complejo con el dador. Se intentó volver a colocar el ligando en el centro activo mediante *dockings*, tanto rígidos como flexibles, sobre las estructuras finales de las simulaciones. No se consiguió recuperar la posición original. Un análisis más exhaustivo mostró que el sitio activo sufre un colapso que impide volver a colocar el UDPGlc en el modo inicial de la simulación, probablemente debido a la ausencia en los modelos de la parte C-terminal de la

proteína. El dador en los modelos es más accesible al solvente y es bastante probable que, como en las otras estructuras resueltas por cristalografía, la GT MG517 tenga cerrado su centro activo por otros elementos estructurales más allá de la topología consenso, como un bucle flexible y una posible hélice α7 que cubre el parche hidrofóbico expuesto en los modelos (Anexo 10), elementos que no fueron modelados por no contar con la suficiente homología ni en secuencia ni estructura entre las diferentes GTAs cristalizadas.

Con estas simulaciones de larga duración hechas por MD hemos podido fusionar toda la información estructural que teníamos de los cuatro modelos en una topología unificada, donde la región conservada mantiene la topología del plegamiento GTA, mientras que para la región variable sabemos que al menos su inicio está constituido por una hélice α de aproximadamente 10 residuos de longitud. Sin embargo, su exacta orientación espacial y el resto de la estructura para esta región variable siguen siendo desconocidas. Por último, las estructuras finales observadas en las simulaciones confirmaron que los modelos 3 y 4 eran los que mejor se adaptaban a los resultados experimentales de mutagénesis dirigida.

7. Sitio de unión del aceptor

Tres de las cuatro plantillas estructurales de GTAs usadas para modelar la GT MG517 incluían también sus respectivas moléculas aceptoras en la estructura (GalNAcβ4Glc en *Bos Taurus* 1O7Q, un octapéptido en *Homo sapiens* 2FFU y Galβ3Gal(6-SO_4) en *Homo sapiens* 3CU0). **Es importante observar que todas las moléculas aceptoras se ubicaban en la zona correspondiente a la RV en cada plantilla, reafirmando la idea de que la diversidad, tanto estructural como de secuencia en esta zona, podría ser la responsable de la especificidad por el sustrato aceptor.** Para confirmar esta hipótesis en la GT MG517, la localización preferente de su aceptor natural, el diacilglicerol (DAG), fue predicha mediante *docking* (ver métodos).

El resultado fue que el DAG se une preferentemente a la región variable (figura 31). La localización del aceptor en la estructura de MG517 es similar a la posición del aceptor natural para la estructura usada como plantilla y precisamente alrededor de la RV. Las energías de interacción entre DAG y MG517 predichas por docking son del mismo rango que para cada sustrato aceptor en las plantillas originales (tabla 9). También se realizaron *dockings* con la molécula DAG sobre las estructuras-plantilla, sin detectar ninguna unión entre el ligando y la proteína. Estos resultados refuerzan la idea de que diversidad estructural y de secuencia en la región variable de las proteínas GTA podría dirigir la especificidad de sustrato para cada ligando.

Tabla 9. Estimación de las energías de interacción del substrato aceptor para las estructuras modeladas y sus plantillas.

Estructura	Energía de interacción (kcal/mol) Aceptor original[b]	Energía de interacción (kcal/mol)[a] Aceptor DAG[c]
Modelo 1 (2Z86_2/1O7Q)		**-5.3 / -5.0**
Plantilla 1O7Q	-6,5	n.d.
Modelo 2 (2Z86_2/2FFU)		**-5.8 / -5.4**
Plantilla 2FFU	-5.3	n.d.
Modelo 3 (2Z86_2/2Z86_1)		**-5.8 / -5.4**
Plantilla 2Z86_1	---	n.d.
Modelo 4 (2Z86_2/3CU0)		**-5.7 / -5.4**
Plantilla 3CU0	-5.6	n.d.

[a] Energías de interacción estimadas para los eventos de unión más probables calculadas por AutoDock. Se muestran los intervalos estimados.
[b] Energías de interacción para el aceptor original en su estructura de rayos X, la usada para el modelado de la RV: GalNAcβ4Glc en *Bos Taurus* 1O7Q, un octapéptido en *Homo sapiens* 2FFU y Galβ3Gal(6-SO_4) en *Homo sapiens* 3CU0. No hay ningún ligando presente en la estructura 2Z86_1.
[c] Energías de interacción estimadas para el aceptor DAG (dipropionilglicerol) en la estructura modelada. No se detecta (n.d.) la unión de DAG en las estructuras originales de rayos X, usadas como plantillas para la RV.

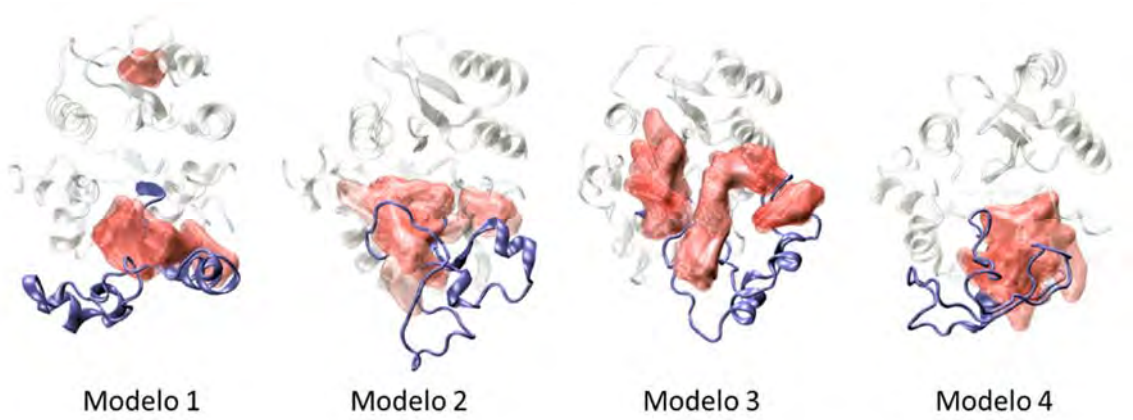


Figura 31. Eventos de *docking* sobre los modelos tras la simulación de MD señalados en rojo. RV en azul.

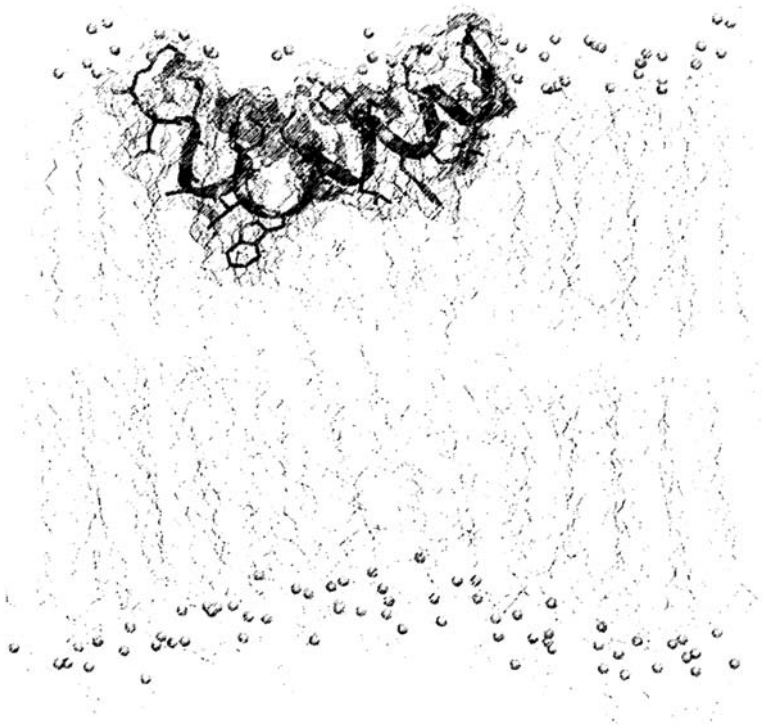
8. Discusión

Presentamos un modelo tridimensional del producto de un gen esencial para *Micoplasma genitalium*, la glicosiltransferasa MG517 que sintetiza mono y diglicosilgliceroles. Hemos construido cuatro modelos diferentes combinando la información estructural de diferentes cristales de proteínas GTA. Estos modelos han posibilitado el diseño de mutantes cuyas medidas de actividad *in vitro* permiten seleccionar los más representativos de la estructura real de MG517.

Proponemos el residuo E193 como la base catalítica de esta proteína de mecanismo *inverting*, siendo el modelo 3 quién mejor representa la posible posición de este aminoácido, mientras que el resto de las interacciones proteína-ligando se hayan mejor representadas por el modelo 4. Mutaciones en los residuos D40, Y126, Y169, I170, e Y218, quienes definen la unión del dador glicosilado en el modelo 4, alteran notablemente la actividad de la enzima.

Nuestros modelos también arrojan algo de luz sobre el sitio de unión del aceptor, que se localiza en la RV. **Creemos que la diversidad estructural y de secuencia de esta región para cada enzima GTA puede ser la responsable de la especificidad por el aceptor**. **La promiscuidad que muestra la GT MG517 para aceptar lípidos no glicosilados como monoglicosilados podría estar en parte controlada por el residuo no conservado Y169**. Sin embargo, **no hemos podido describir con nuestros modelos la geometría exacta del sitio de unión del aceptor, aunque sí proponemos que el inicio de esta región está formado por una hélice α de entre 4 y 10 residuos de longitud**. Además, los modelos presentan un importante parche hidrofóbico expuesto al solvente, que podría ser usado para la dimerización como es sugerido en otras proteínas o como es más probable, con la unión de la extensión C-terminal que no está representada en nuestros modelos (Anexo 10).

Hasta aquí hemos establecido la topología consenso y estructural de la familia de enzimas GTA, que hasta donde nosotros conocemos no había sido estudiado con esta profundidad, para toda la superfamilia. **El alineamiento múltiple, refinado gracias a la información estructural, nos ha permitido generar un perfil HMM representativo del clan GTA**. Este nuevo perfil tiene muchas aplicaciones potenciales como la detección de secuencias homólogas en genomas recién secuenciados o para guiar el alineamiento de secuencias GTAs. **El protocolo de modelado que hemos aplicado aquí debería ser considerado cuando se intente modelar otras estructuras GTA, por la ausencia de estructuras homólogas claras que puedan ser usadas como plantilla, debido a la región variable que existe en todas ellas**.



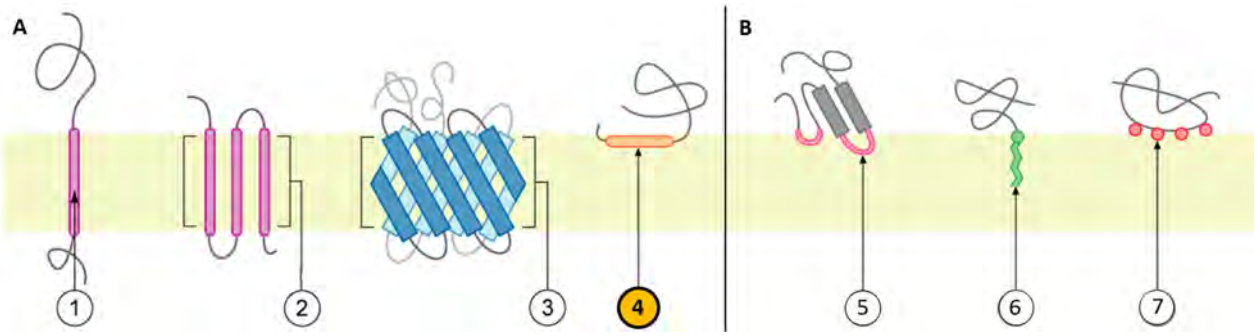
INTERACCIÓN CON LA MEMBRANA, DE LA PROTEÍNA MG517, A TRAVÉS DE LA REGIÓN C-TERMINAL

1. Introducción

La metodología utilizada para el modelado de la GT MG517, dejó patente la imposibilidad de encontrar estructuras homólogas para la extensión C-terminal. Ni tan siquiera a nivel de secuencia fue posible encontrar proteínas homólogas, a excepción de la muy relacionada filogenéticamente diacilglicerol beta-glicosiltransferasa de *M. pneumoniae*, de la que no se conoce su estructura.

En nuestro laboratorio se había demostrado la activación de la GT MG517 por fosfolípidos aniónicos y su más que probable interacción con la membrana de *M. genitalium*. Por el perfil hidrofóbico de la secuencia, la región más probable para esta unión era la C-terminal, que una vez eliminada permitía la purificación del enzima sin el uso de detergentes, manteniendo el plegamiento. Además, la eliminación de los últimos 13 residuos de esta región conlleva la total inactivación de la enzima.

En el marco de mi tesis, y con el objetivo de identificar zonas funcionales en la región C-terminal de MG517 relacionadas con la unión a membrana, el estudio de esta región se ha abordado redirigiendo la atención al ámbito de las proteínas de membrana.



Fuente: Imagen reconstruida desde la original. Wikipedida: Membrane protein.

Figura 32. Diferentes formas de asociación en las proteínas de membrana. A: Integrales (1, Bitópica a través de una α hélice. 2, politópica a través de α hélices. 3, politópica mediante láminas β. **4, monotópica mediante una hélice anfipática**, este tipo de asociación permanente no desorganiza la membrana al extraer la hélice. B: Periféricas (5, mediante bucles hidrofóbicos. 6, por unión covalente a un lípido. 7, por medio de interacciones electrostáticas).

"Las proteínas de membrana pueden ser clasificadas como "periféricas" o "integrales" según su grado de asociación con la membrana lipídica (Luckey 2008) (Figura 32). Las proteínas periféricas de membrana se unen temporalmente a una de las hemimembranas de la bicapa lipídica o a otras

proteínas de membrana. Interaccionan débilmente con la membrana, principalmente mediante interacciones no covalentes como las electrostáticas y los puentes de hidrógeno. Por ello, las proteínas periféricas de membrana pueden ser extraídas por métodos relativamente suaves como elevar la fuerza iónica y uso de tampones alcalinos, dejando intacta la membrana lipídica (Luckey 2008). Por el contrario, las proteínas integrales de membrana se unen estrecha y permanentemente a la membrana, y solo pueden ser extraídas mediante tratamientos con detergentes y solventes orgánicos que desorganizan la bicapa lipídica (White and Wimley 1999; Andersen and Koeppe 2007). Además, basándonos en su modo de inserción, las proteínas integrales también pueden ser clasificadas como monotópicas, bitópicas o politópicas (Blobel 1980). Las proteínas monotópicas se asocian firmemente a solo un lado de la bicapa lipídica, mientras que las bitópicas y politópicas se extienden por la membrana por medio de uno o más segmentos transmembrana (Elofsson and von Heijne 2007)."[j]

La 1,2-diacilglicerol 3-glucosiltransferasa de *Acholeplasma laidlawii* (AlMGS) es una glicoglicerolípido sintasa con plegamiento GTB. De la clase Mollicutes, este organismo carece de pared celular como *M. genitalium*. La densidad de carga y propiedades de curvatura de su membrana se controlan mediante la adición de azúcares a glicerolípidos, como también lo hace MG517 en *M. genitalium* (aunque en el caso de *A. laidlawii* no de forma secuencial sino a través de dos enzimas: AlMGS y AlDGS). Otra característica compartida entre las proteínas de estos dos organismos es la ausencia de segmentos transmembrana y sin embargo la necesidad de detergentes para su solubilización. En el dominio N-terminal de *A. laidlawii* (las GTBs sí poseen dominios N y C terminal) se halló una hélice de carácter anfipático, que contenía una mezcla de residuos hidrofóbicos y básicos, principalmente Lys y Arg y que se adhiere de forma permanente y monotópica a la membrana lipídica[95]. La proteína WaaG de *Escherichia coli* muestra una interacción con membrana similar; aunque no pertenece a la clase Mollicutes, se trata también de una glicosiltransferasa con plegamiento GTB, involucrada en la síntesis de lipopolisacáridos. En ella una hélice anfipática de unos 30 residuos de longitud es la responsable, en parte, de la interacción con membrana de esta proteína[96].

Esta forma de interacción monotópica, a través de una hélice anfipática, está asociada a proteínas cuya actividad se relaciona con el remodelado de las características físico-químicas de la membrana[97], ya sea por una acción directa de la hélice sobre ella[98,99] o bien por la adición de moléculas a la membrana, donde la hélice actúa como sensor de curvatura[100]. **Por las características compartidas con *A. laidlawii*, este tipo de interacción podría ser el utilizado por la proteína MG517.**

En esta parte de la tesis encontraremos evidencias computacionales sobre la existencia de una hélice anfipática en la parte apical de la extensión C-terminal de MG517, además de aportar información sobre las posibles estructuras y funciones del resto de la secuencia C-terminal.

[j] Albesa-Jové, D., Giganti, D., Jackson, M., Alzari, P. M. & Guerin, M. E. Structure-function relationships of membrane-associated GT-B glycosyltransferases. *Glycobiology* **24,** 108–124 (2014).

La simulación computacional para el estudio de proteínas de membrana, es una herramienta validada desde hace largo tiempo[101,102,103] y existe una extensa literatura acerca de esta cuestión[104,105]. El diseño de péptidos antibacterianos, aunque se trate de un tema diferente al aquí estudiado, ha genenerado grandes avances gracias a la simulación de péptidos en membranas[106]. Las diferentes estructuras encontradas serán estudiadas mediante simulaciones de Dinámica Molecular clásica (MD) en solvente acuoso explícito[107]. La interacción con membrana de los elementos seleccionados y su orientación sobre la misma, también se estudiará mediante MD y diferentes condiciones iniciales de posicionamiento[108,109]. La energía de interacción con la membrana también será estudiada mediante métodos computaciones, siguiendo una metodología similar a la de otros trabajos previos[110], pero utilizando Metadinámica clásica.

2. Identificación de zonas de interacción

En el capítulo anterior, para el modelado de la región N-terminal de MG517 se realizó una predicción de estructuras secundarias de toda la secuencia con PsiPred[111]. Con ello se pretendía comparar el resultado de la predicción, con la topología consenso propuesta para todas las GTAs, la cual se cumplió salvo para la hoja β6, donde PsiPred predijo una hélice para MG517 y que se consideró una predicción errónea en este caso. Una vez establecida la topología consenso, fue posible determinar la región de la secuencia que corresponde a la extensión C-terminal, que, tras el resultado de los alineamientos múltiples y superposición estructural, se vio que no guarda homología ni de secuencia ni de estructura con el resto de proteínas cristalizadas. **La predicción realizada por PsiPred (figura 33) indicaba que esta región está formada solo por hélices α**, tres en concreto, de gran longitud y variable puntuación de confianza en la predicción de la estructura.

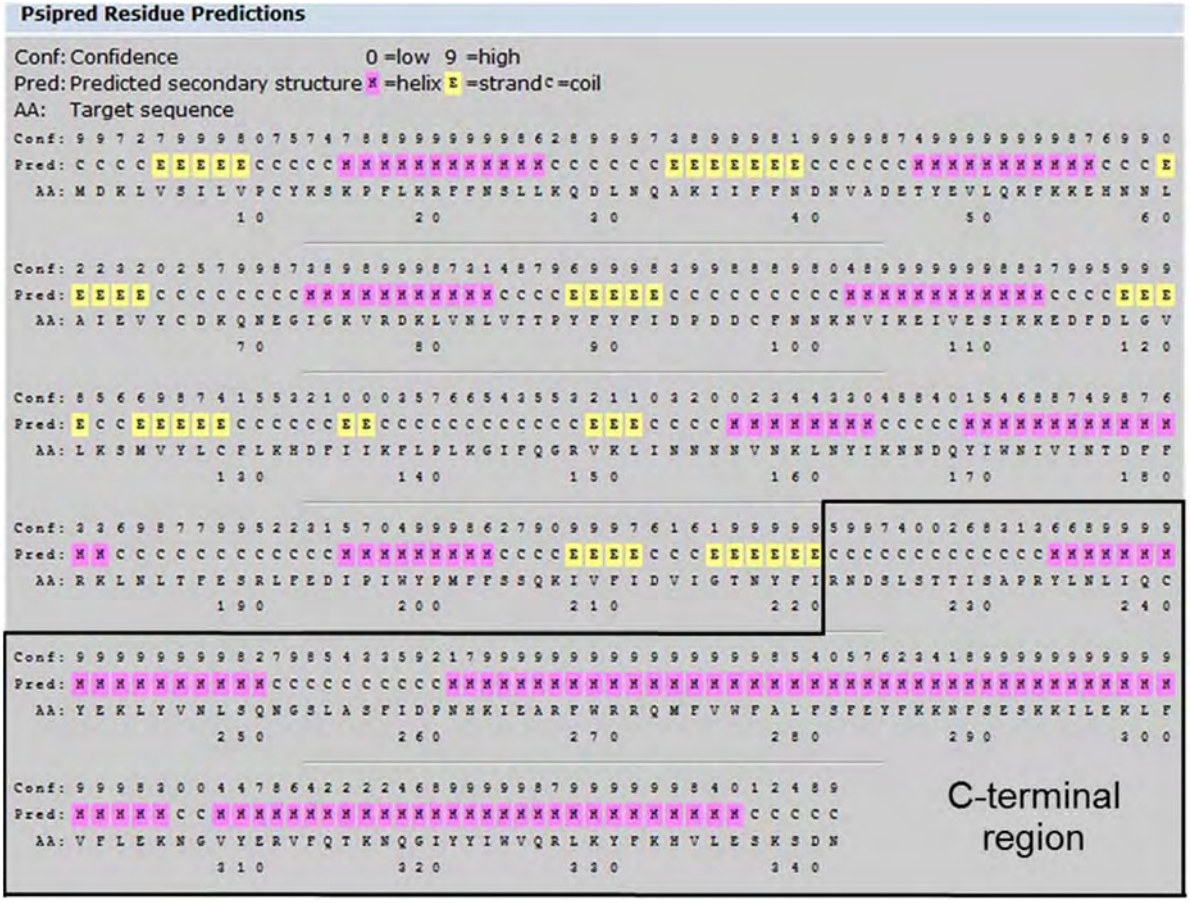


Figura 33. Predicción de la estructura secundaria de MG517 realizado por PsiPred. La región C-terminal (encuadrada en negro) estaría compuesta solo por hélices α.

A partir de la asignación de estructura secundaria, se llevaron a cabo distintas predicciones de propiedades hidrofóbicas y anfipáticas a lo largo de la secuencia completa de MG517.

Una búsqueda de hélices transmembrana con el servidor TMHMM[112] mostró que tal tipo de hélices no existen en toda la secuencia de MG517 (resultados no mostrados). La región C-terminal también se estudió con los servidores I-TASSER para la predicción de estructura terciaria mediante homología y con Robetta, para la predicción *ab initio* de la estructura de esta región. I-TASSER generó estructuras muy dispares y desordenadas (Anexo 11), mientras que Robetta, al utilizar solo la región C-terminal (por ser demasiado grande la secuencia completa), no podemos identificar las estructuras compatibles con una estructura globular, como se supone tiene MG517.

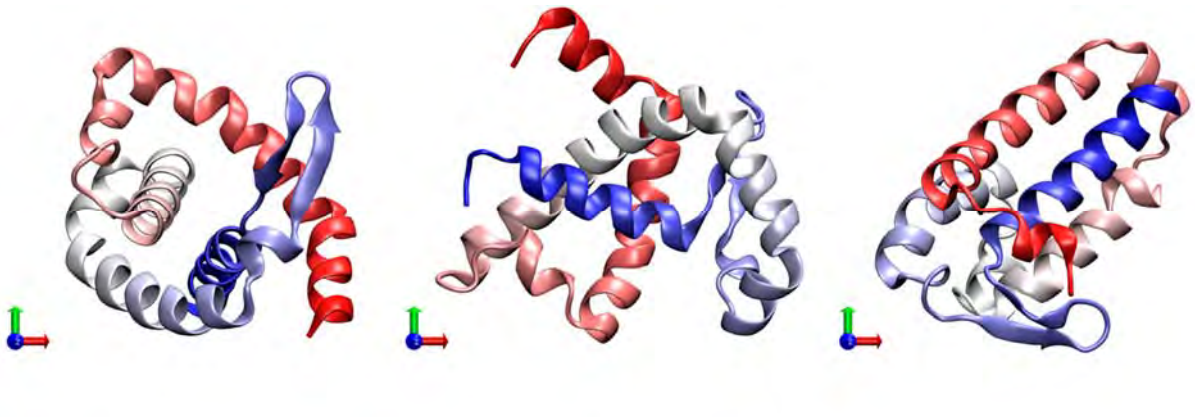


Figura 34. Ejemplos de la estructura de la región C-terminal según los resultados de Robetta. Coloreados desde el extremo N-terminal (rojo) al C-terminal (azul).

Aún así, la estructura central de la figura 34, podría ser la más apropiada para contener una hélice en el extremo C-terminal que pueda interaccionar con membrana, permitiendo al resto hacerlo con la proteína. Sin embargo, dado que el interés de este capítulo era y es estudiar esta interacción con membrana y no tanto, la estructura tridimensional de la región C-terminal, se dejó de lado esta vía.

2.1. Evaluación del perfil hidrofóbico

Se evaluó el perfil hidropático de la secuencia completa de MG517 con el programa MPEx[113] , que se basa exclusivamente en la composición de aminoácidos y no tiene en cuenta la estructura. Valores positivos en el perfil hidropático (Figura 35) indican elevada hidrofobicidad, mientras que valores negativos se asocian a elevada afinidad por agua. Esta predicción permitió proponer posibles zonas de interacción con la membrana en la región Cter de estructura desconocida, y así mismo, ubicar regiones de elevada hidrofobicidad en la estructura tridimensional de la región Nter, modelada en el capítulo anterior. A partir de este análisis, la interpretación de las regiones con elevada hidrofobicidad y de estas, aquellas con probable interacción con la membrana, para toda la secuencia de MG17 es la siguiente:

1. **Estructuras β1-α1-β2**: Se sitúan en la zona interior de la proteína, sobre todo las hojas β, la hélice α1 está flanqueada por la hélice α2 y parte de las estructuras hoja β4' y hélice α4. El acceso al solvente de las mismas ha de ser menor que para otras zonas de la proteína. Se supone que el perfil hidrofóbico de esta región está sobre todo relacionado con el plegamiento de la estructura y no con ninguna interacción con la membrana, además la topología en esta zona hace muy difícil una interacción de este tipo.
2. **Estructuras α3-β4 y motivo DXD**: Estas, fundamentalmente el motivo DXD, se encuentran en el núcleo catalítico de la enzima que necesita un entorno libre de aguas para poder transferir eficientemente el azúcar. De nuevo el perfil hidrofóbico de esta región puede no estar relacionado con la interacción con la membrana, sino con la actividad catalítica; tampoco la topología de esta zona hace posible esta interacción.
3. **Inicio de la región variable (residuos 120 a 150)**: Se vio en el estudio de la región N-terminal (capítulo anterior) que es posible que esta primera zona de la RV esté constituida por una hélice α. Además, en todas las RV estudiadas para las proteínas cristalizadas, estos primeros residuos se encuentran siempre en una posición exterior, accesible al solvente. Puesto que el aceptor diacilglicerol se encuentra entre los fosfolípidos de membrana y éste se coloca según los resultados del *docking* también en la RV, es probable que exista una zona de interacción de ésta con la membrana. Puesto que el perfil hidrofóbico indica que la zona más probable es esta hélice, **se propone este elemento de la RV como una de las estructuras que podrían interactuar con la membrana**.
4. **Estructuras β6-α5**: Al igual que la zona α3-β4 y motivo DXD, ésta es una posición interior de la proteína, contigua a la RV, y que también forma parte de la zona catalítica. La hoja β6 forma un codo donde se apoya el difosfato y el azúcar del aceptor, donde los residuos apolares son abundantes. De nuevo el perfil hidrofóbico de esta región puede no estar relacionado con una interacción con la membrana sino con la actividad catalítica. Tampoco la topología de esta zona haría posible esta interacción.
5. **Estructuras α6-β7 y una larga cadena de residuos en la extensión C-terminal**: Esta cadena de residuos, más allá de la zona modelada en el capítulo anterior (residuos 221 a 236) podría constituir la hélice α7 de la región Nter en la GT MG517 y cubrir el parche hidrofóbico al descubierto en nuestros modelos, bajo la hélice α6 (Anexo 10). La hidrofobicidad no es muy alta para esta zona y es una posición bastante exterior en proteínas homólogas. Sin descartar una interacción con la membrana, es probable que el perfil indique una interacción hidrofóbica entre las hélices α6 y una posible α7.
6. **Residuos 270 a 288**: Zona muy hidrofóbica (máximo índice de hidropatía de toda la secuencia). Al tratarse de una zona de estructura desconocida, su ubicación en el plegamiento de MG517 no puede predecirse. **La elevada hidrofobicidad podría indicar unión a membrana, pero de ser así, tendría que estar inserta en ella**. La ausencia de hélices transmembrana apunta más a un elemento estructural interno de la proteína.
7. **Residuos 314 a 333**: Zona de menor hidrofobicidad que la anterior pero aun así con índice de hidropatía positivo, lo cual **puede indicar asociación con membrana**.

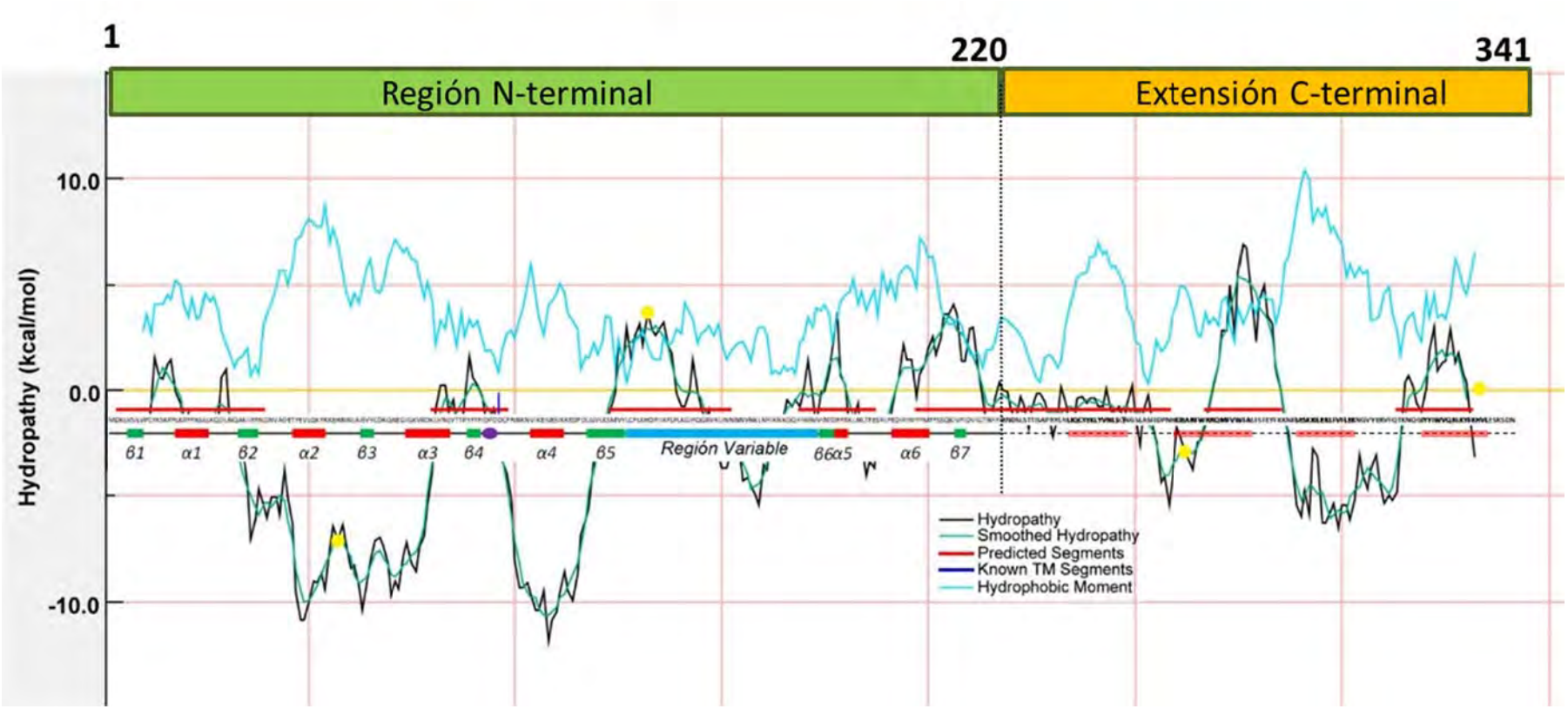


Figura 35. Resultado de aplicar el software MPEx a la secuencia de MG 517. En la figura se muestran diferenciadas las regiones N y C terminal. Sobre el perfil hidropático, cuya leyenda está impresa en el gráfico, se han incluido la secuencia completa de MG517 y bajo esta su estructura secundaria, validada por MD.

Las líneas horizontales rojas sobre la secuencia, corresponden a segmentos con cierta probabilidad de interacción con la membrana debido a su momento hidrofóbico.

Para la región Nter, al tener un modelo de estructura tridimensional, es posible descartar aquellos segmentos que por su disposición espacial no pueden acceder a la membrana. **Es interesante la zona inicial de la RV, donde las simulaciones por MD han convergido a una hélice α, que se encuentra accesible al solvente y que aquí se le asigna cierta probabilidad de interacción con la membrana**.

Para la región Cter, existen tres zonas de probable interacción con la membrana, que más o menos coinciden con las hélices α predichas por PsiPred para esta región.

2.2 Evaluación del perfil anfipático

Con la herramienta Helical Wheel[k,114], se buscó a lo largo de toda la secuencia de MG517 la presencia de hélices anfipáticas. Se detectó un segmento de 19 residuos en la parte apical de la extensión C terminal (G318-S337), cuya disposición de residuos es compatible con una hélice anfipática. Esta característica no se halló en ninguna otra zona de la secuencia (figura 36 D). Esta región coincide con el último segmento de secuencia para el cual el índice de hidropatía es positivo (residuos 314 a 333).

Combinando los resultados de predicción de estructura secundaria, con los análisis de hidropatía y anfipatía, se definen aquí los elementos de estructura secundaria que pueden estar involucrados con la unión a membrana de MG517 (Figura 36). En esta asignación, solo se tienen en cuenta, por un lado, aquellas regiones cuya puntuación para la predicción de estructura secundaria es 6 o superior y por otro aquellas en que el índice de hidropatía es positivo (y por tanto hidrofóbicas). Resultado de la asignación (Figura 36) son 5 hélices α que corresponden a las siguientes secuencias:

- Hélice 1: L_{237}LIQCYEKLYVNLS$_{249}$
- Hélice 2: H_{262}KIEARFWRRQMFVWFA$_{278}$
- Hélice 2.1: R_{271}QMFVWFALFSFEYFKK$_{287}$
- Hélice 3: F_{289}SESKKILEKLFVFLE$_{304}$
- Hélice 4: K_{316}NQGIYYIWVQRLKYFKHVLESK$_{338}$

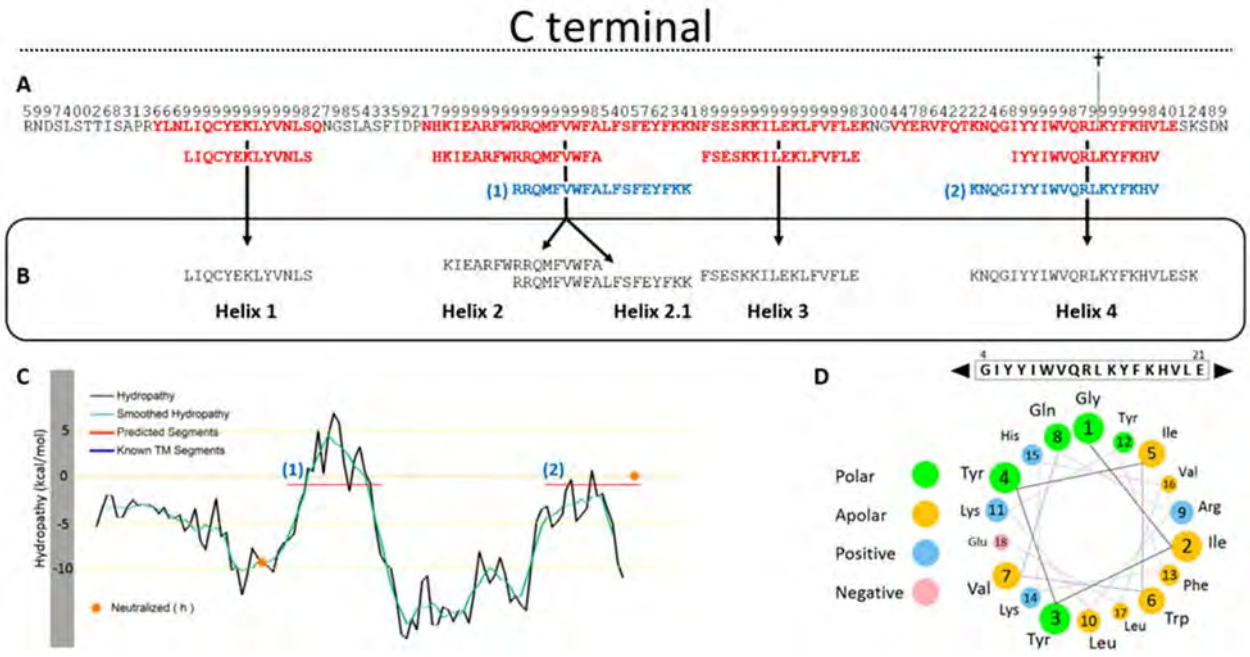


Figura 36. **A**. Secuencia C-ter de MG517 enviada a PsiPred. Los residuos coloreados en rojo corresponden a la predicción de una hélice α mientras que el resto corresponden a zonas desordenadas. Cada residuo tiene asignada una puntuación de confianza para la predicción de la estructura (sobre ellos en la secuencia). Solo

[k] http://cti.itc.virginia.edu/~cmg/Demo/wheel/wheelApp.html

las puntuaciones superiores a 6 las consideramos como hélices. (1) y (2): Segmentos de secuencia con probable interacción con la membrana según MPEx. **B.** Hélices seleccionadas para modelar combinando la información de los tres predictores (PsiPred, MPEx y Helical wheel), como el segmento (1) es una extensión de la hélice predicha 2 la denominamos "helix 2.1". **C**. Perfil MPEx para la secuencia C-ter, la alternancia de hidropatía positiva y negativa resultante designa dos segmentos de secuencia con probable unión a membrana: (1) que solapa con la predicción de PsiPred y (2) que cubre la última hélice predicha completamente. **D**. Representación *Helical wheel* para una ventana de 17 residuos del segmento apical de la región C-ter. La hélice 4 muestra una clara disposición anfipática de sus aminoácidos a lo largo de 23 residuos.

3. Modelado de hélices hidrofóbicas y anfipáticas

Las hélices se modelaron *de novo*, con el plegamiento de los residuos dirigido por *constrains* hacia una hélice α, con cuatro parámetros diferentes, que generan para cada una, cuatro modelos de hélices, entre las que se escoge la mejor modelada (ver métodos).

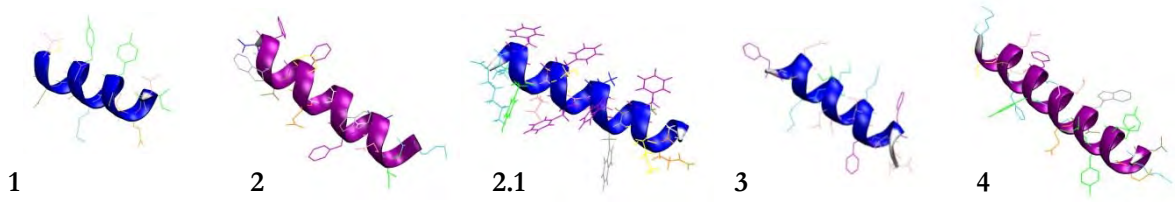


Figura 37. Hélices modeladas a través de *Modeller*.

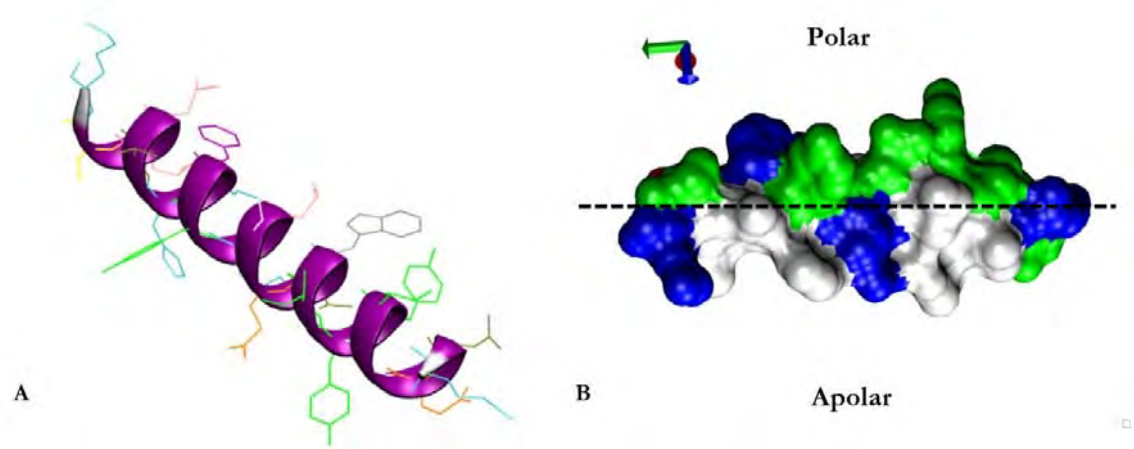


Figura 38. **A**, modelado de la hélice 4. **B**, superficie de interacción de los residuos. Se observa a la hélice dividida en dos zonas (línea de puntos), la superior comprende aminoácidos polares (color verde y azul) y la parte inferior aminoácidos apolares (color blanco). En color azul se representan los aminoácidos con carga positiva.

4. Estabilidad de los modelos generados

4.1 Hélices en solvente acuoso

La estabilidad de las 5 hélices identificadas como posibles zonas de interacción de MG517 a la membrana fue analizada mediante dinámica molecular. Cada una de las hélices modeladas fue solvatada por separado, con agua e iones en cajas de simulación cúbicas de 3 nm (ver métodos). Se realizaron a continuación varias simulaciones de larga duración (1-2 μs) en solvente acuoso. La estabilidad del plegamiento para cada hélice a lo largo de la simulación se monitorizó con gráficos DSSP (ver métodos) cuyo resultado se resumen en la figura 39.

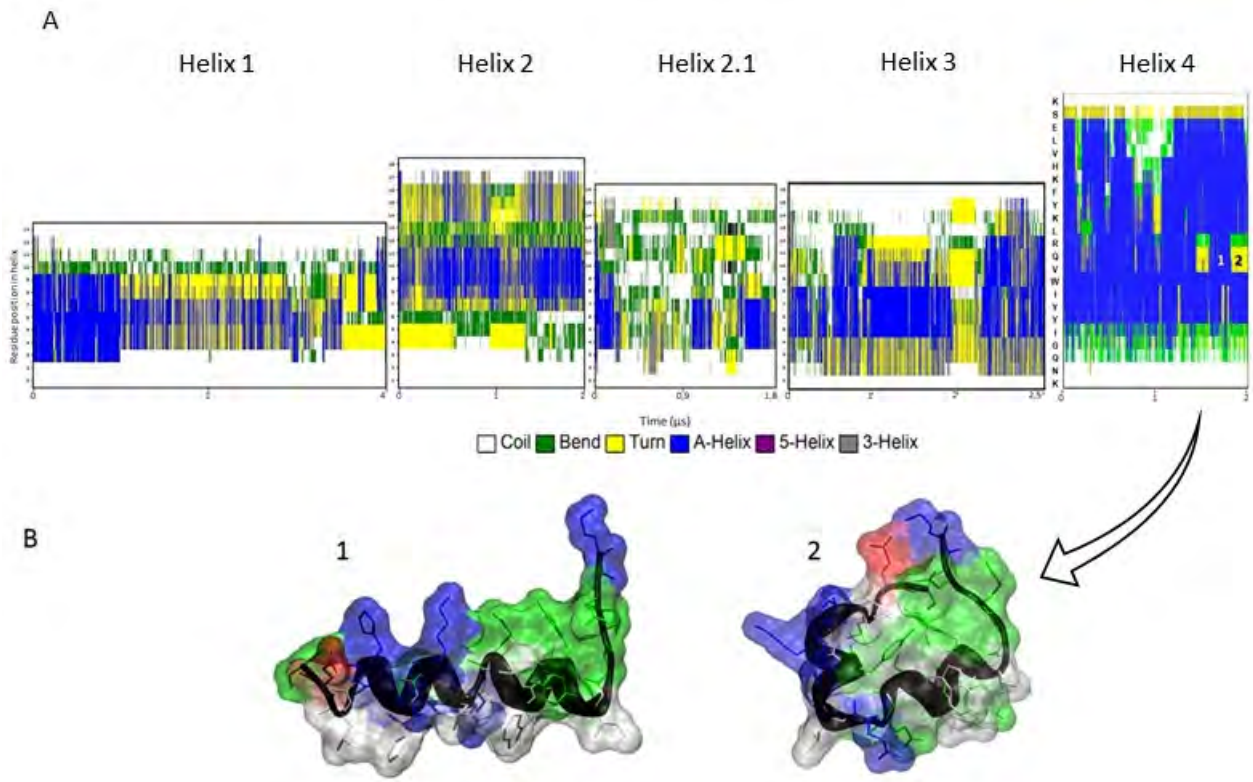


Figura 39 A. Gráfico de evolución DSSP para el plegamiento de la hélice en diferentes simulaciones de MD. El color del gráfico identifica la estructura en ese momento tal como está descrito en la leyenda. **B.** Conformaciones de la hélice 4: 1. Extendida, residuos apolares en una cara (blanco) y los polares en la opuesta (coloreados), y 2. Doblada, hélice colapsada.

Las hélices 1, 2, 2.1 y 3 cuyo tiempo de simulación fue 4, 2, 1,8 y 2,5 μs respectivamente, pierden a lo largo de la dinámica su conformación inicial. La hélice 1 mantiene su conformación para el 70 % de la secuencia durante el primer μs, por lo que en caso de ser realmente una hélice en la estructura de MG517, un entorno totalmente acuoso la termina desplegando y quizás necesite de

otros elementos para mantener su conformación. Esta hélice 1 en las simulaciones, correspondería a la hélice 7 de MG517, que según se vio en el capítulo 1, habría de cubrir el parche hidrofóbico observado en los modelos (Anexo 10) quedando expuesta al solvente, algo que concuerda con su perfil hidropático y el resultado de esta simulación; descartándose entonces como zona de unión a la membrana. Las otras tres hélices (2, 2.1 y 3) se despliegan desde tiempos muy tempranos de la simulación, manteniendo y solo puntualmente, menos del 20 % de su secuencia en forma de hélice. También se descartan estas tres hélices como zonas de unión a la membrana, mediante una interacción monotópica. Por el contrario, la hélice 4, simulada durante 2 μs, mantuvo el 75 % de la secuencia con la misma conformación de partida, estando los residuos desplegados, repartidos entre los extremos del péptido. Durante el último μs de simulación de la hélice 4, su conformación se dividió en dos microestados, uno con la estructura extendida y otra doblada sobre sí misma (figura 39). Se puede observar la disposición anfipática de los residuos en la conformación extendida y cómo la hélice colapsa en la conformación doblada, ocultando los residuos apolares del solvente acuoso y utilizando como bisagra los residuos centrales Val, Gln y Arg.

Este resultado apoya la hipótesis de la hélice anfipática en esta zona, continuando con ella los estudios de estabilidad en membrana.

4.2 Hélice 4 inserta en membrana

Para validar las posibles características anfipáticas de la hélice 4, se evaluó la estabilidad de dicha hélice en distintos modelos de membrana mediante simulaciones de dinámica molecular. Se pretende averiguar la movilidad de la hélice en la membrana, para encontrar la posición anfipática más favorable.

Construcción de los sistemas

Se construyeron 6 sistemas formados a partir de una bicapa lipídica periódica, descargada del *Biocomputing Group* de la Universidad de Calgary[l], que contiene 64 fosfolípidos de dipalmitoil fosfatidil colina (DPPC) en cada hemimembrana. Esta membrana se modificó manualmente para cada sistema, eliminando diferentes grupos colina de una de las hemimembranas, convirtiéndolos en grupos metilo, con la intención de alterar la carga de la membrana en los distintos sistemas (Figura 40 A). **Sistema 1**: sin modificar, todo dipalmitoil fosfatidilcolina. **Sistema 2**: ratio 1:1 de fosfatidil colina y fosfatidil metilo para una de las dos hemimembranas, dejando sin modificar la otra. **Sistemas 3, 4, 5 y 6**: todo fosfatidil metilo en una de las dos hemimembranas, sin modificar la otra. Se escogió la estructura representativa de la hélice 4, de la simulación en solvente acuoso, mediante un análisis de agrupamientos, estando en conformación extendida. Dicha estructura se colocó manualmente en diferentes posiciones respecto a la membrana, en estos 6 nuevos sistemas

[l] http://wcm.ucalgary.ca/tieleman/downloads

siguiendo el protocolo de Justin A. Lemkul[m]. Este protocolo inserta verticalmente el péptido $KALP_{15}$ en una membrana DPPC, mediante un proceso de expansión y contracción de la membrana sobre el péptido que se mantiene fijado en la posición deseada.

En los sistemas 1, 2 y 3 la hélice 4 se situó en el centro de la membrana (figura 40), estando la hélice completamente rodeada de cadenas lipídicas. En el sistema 4, la hélice se situó al mismo nivel que las cabezas polares de los fosfolípidos de la hemimembrana superior, y la membrana fue forzada a cerrarse sobre el péptido fijado, quedando así algunos lípidos cabalgados sobre ella. En el sistema 5 la hélice 4 se colocó fuera de la membrana, pero a una distancia que permitiese una interacción electrostática entre sus residuos polares y los fosfolípidos de la membrana, aquí no fue necesario utilizar el protocolo de expansión y contracción. Para el sistema 6 la hélice se colocó al mismo nivel que las cabezas polares de los fosfolípidos de la hemimembrana superior, como en el sistema 4, pero la membrana se cerró suavemente sobre la hélice, siguiendo el protocolo modificado de $KALP_{15}$, al mismo nivel que en el sistema 4 pero con la hélice completamente accesible al solvente en su parte superior (ver métodos). Los seis sistemas contenían la hélice anfipática en la misma orientación espacial, con la zona apolar en dirección -z y la polar en dirección z. La hélice en los sistemas 1, 2 y 3 no tenía acceso al solvente; en el sistema 4 sí, pero solo parcialmente ya que algunos lípidos se encontraban sobre ella, en el sistema 5 toda la hélice se encontraba en el solvente acuoso y en el 6 solo la parte polar de la misma. Los lípidos se reorganizan en paralelo y la hemimembrana inferior de los sistemas 4 y 6 forma un valle hacia dentro, que cubre el espacio dejado por los lípidos eliminados durante la inserción de la hélice.

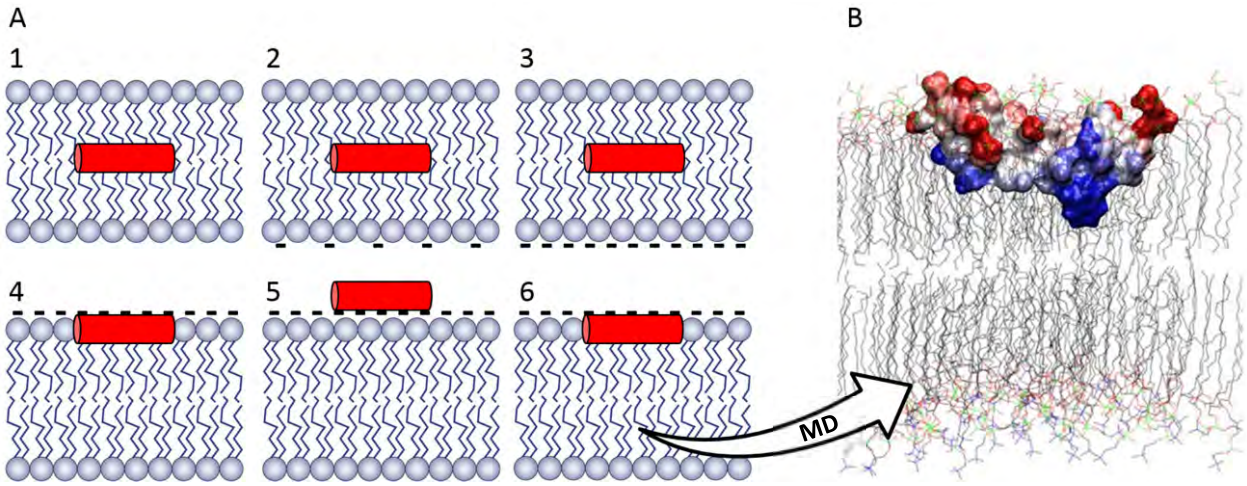


Figura 40. A. Representación esquemática de los sistemas Hélice4-membrana simulados. En gris la bicapa lipídica en distintas composiciones de lípido neutro DPPC y lípido aniónico DPPM (signo negativo) B. Representación del sistema 6 a tiempo final de la MD. La hélice es coloreada por software SAP[115] donde cuanto más rojo más polar y más azul más apolar, mostrando una clara orientación anfipática en la membrana.

[m] www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/membrane_protein/

Estabilidad por dinámica molecular

Se iniciaron estas simulaciones y pudimos comprobar como las hélices en los sistemas 1, 2 y 3 se desplegaban dentro de la membrana y ninguna se desplazaba hacia el solvente.

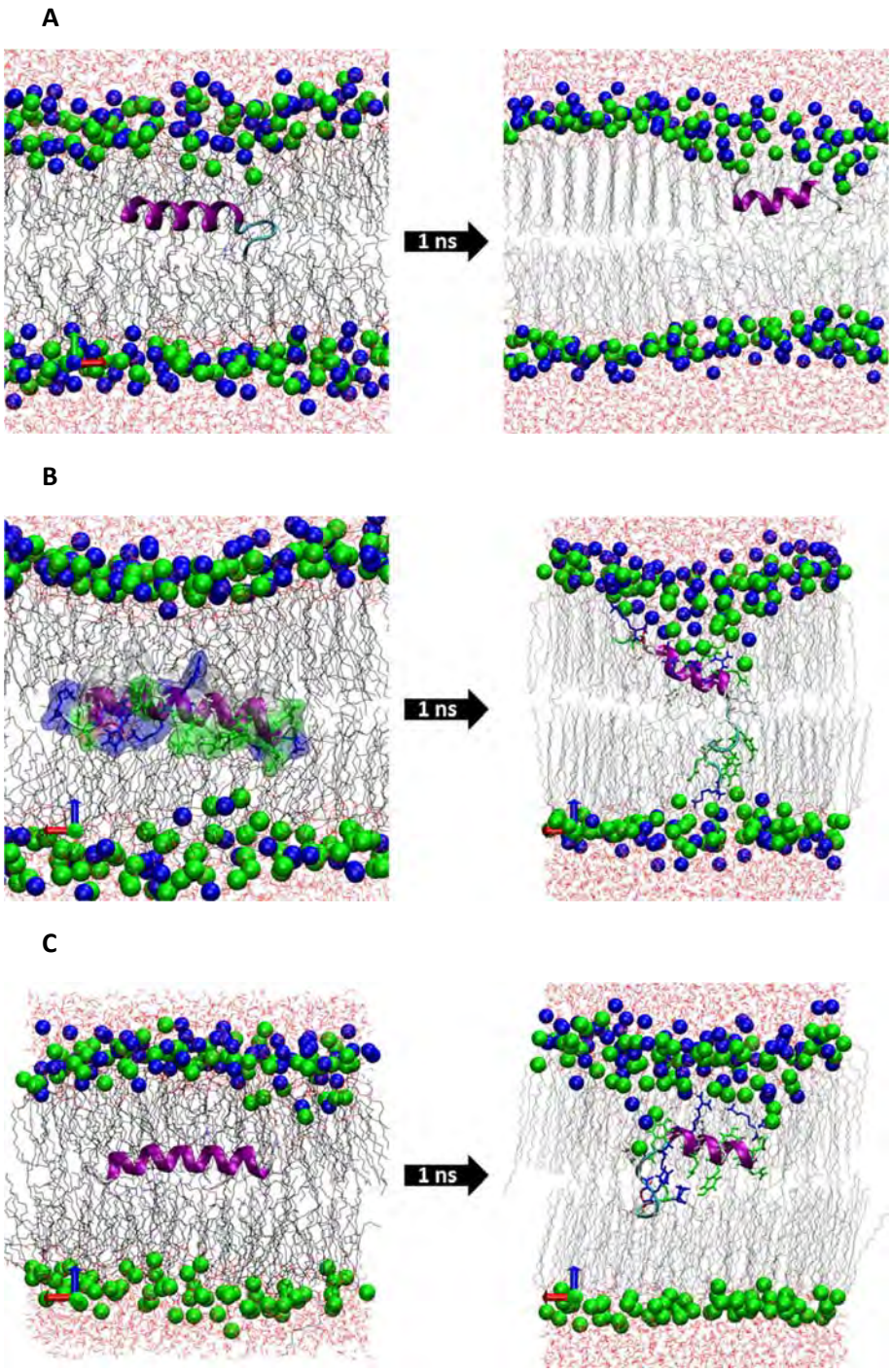


Figura 41. A: Sistema1. B: Sistema 2. C: Sistema 3. (Esferas azules DPPC, esferas verdes DPPM, aguas en rojo).

Para la hélice en el sistema 1, Transcurrido 1 ns de simulación la hélice no ha perdido del todo su conformación, pero permanece en el interior de la membrana sin acceso al solvente. (Figura 41A). En el sistema 2 la hélice sufre una importante desestructuración. Transcurrido 1 ns de simulación el movimiento del péptido entre las cadenas hidrocarbonadas ha permitido que interaccione con las cabezas polares por ambos extremos, desplegándose y fijándose en esta posición sin acceso al solvente (Figura 41B). El sistema 3 tampoco es estable en la posición inicial y tras 1 ns de simulación, el péptido interacciona con las cabezas polares de algunos fosfolípidos desde el interior de la membrana. Mediante esta interacción algunas de estas cabezas polares son arrastradas hacia el interior de la membrana. La hélice se despliega y nunca llega a acceder a la zona del solvente (Figura 41C).

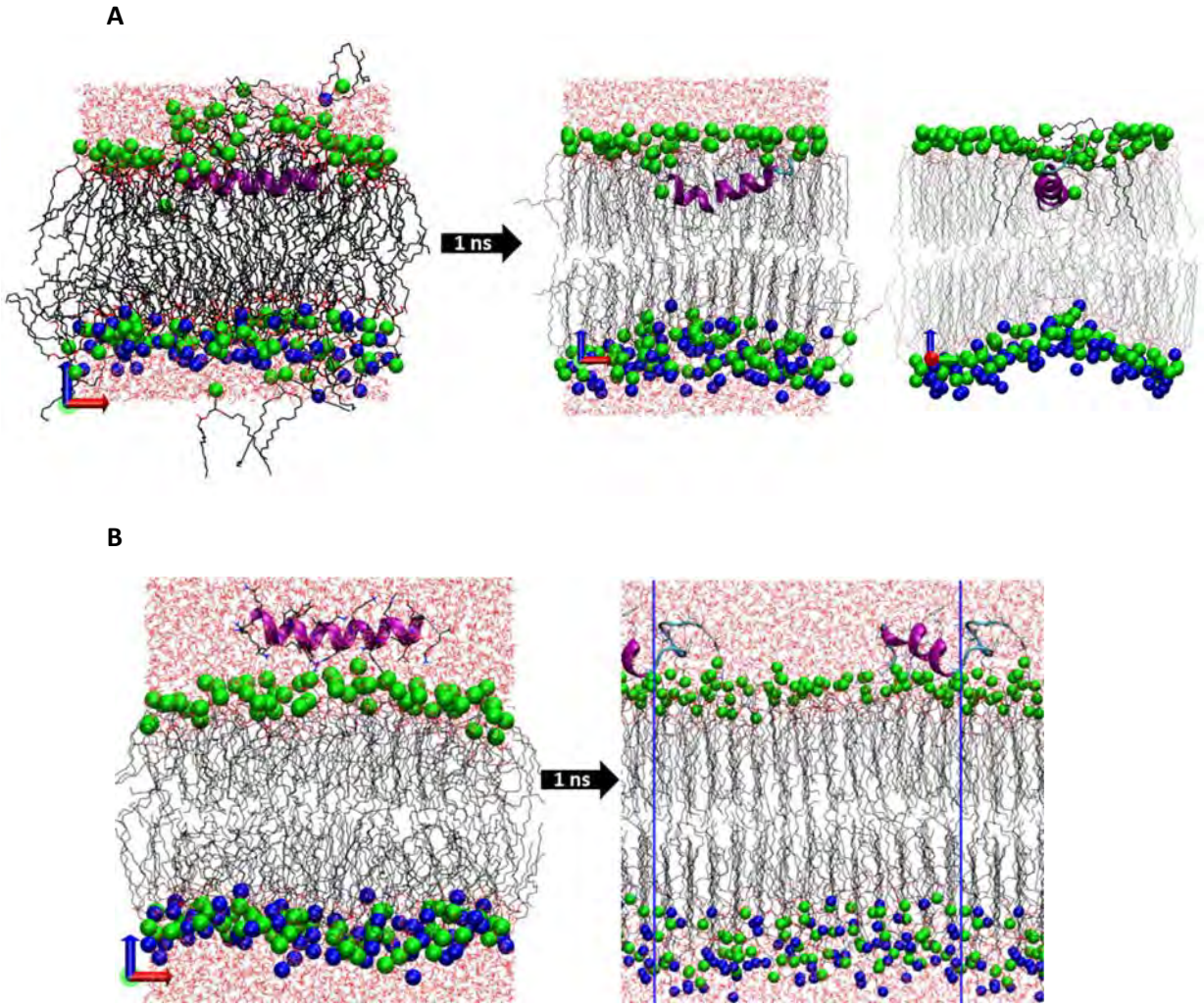


Figura 42. A: Sistema 4. Se observan lípidos sobre la hélice a tiempo final de la simulación B: Sistema 5. La hélice no llega a integrarse en la membrana. (Esferas azules DPPC, esferas verdes DPPM, aguas en rojo, las líneas verticales azules representan los límites de la caja y la condición periódica).

La hélice en el sistema 4 no ha perdido su conformación, pero tampoco puede quitarse de encima los lípidos cabalgados desde el inicio de la simulación (figura 42 A). El sistema 5 se despliega y no se inserta en la membrana (figura 42 B).

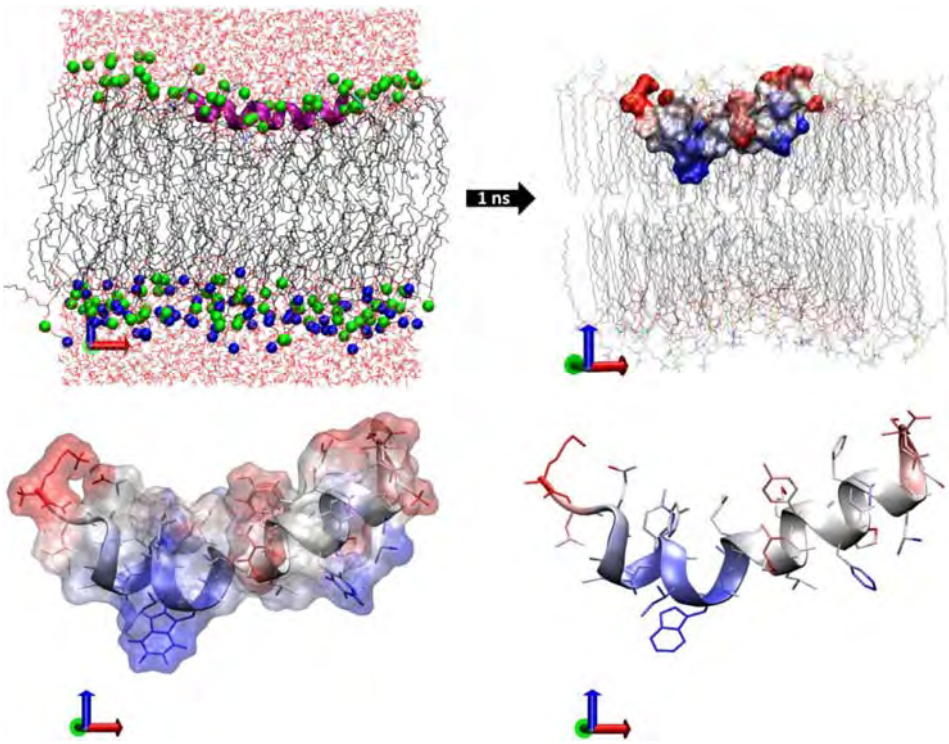
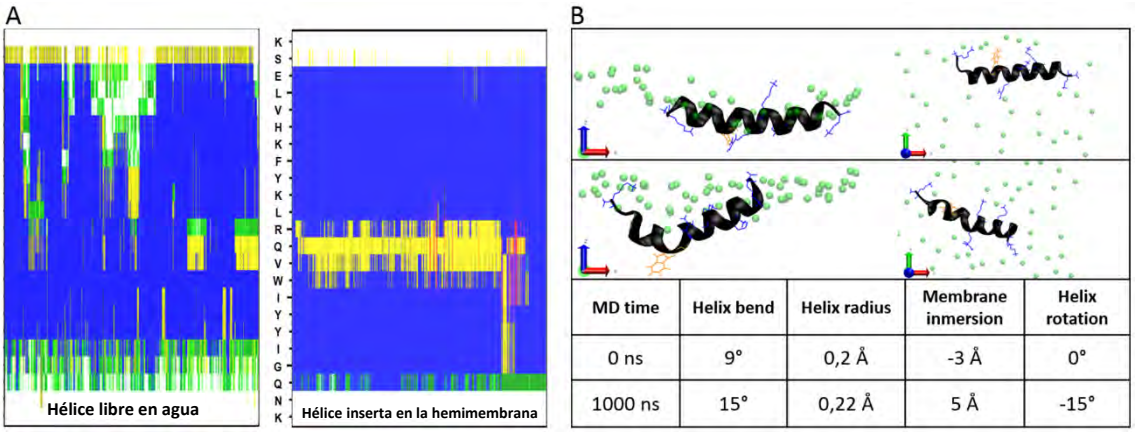


Figura 43. Sistema 6. Al finalizar la simulación la hélice continúa plegada en la misma posición, pero ha modificado la orientación de los residuos polares enterrándolos en la membrana. (Esferas azules DPPC, esferas verdes DPPM, aguas en rojo). El péptido ha sido además coloreado por la hidrofobicidad de sus residuos con el programa *Measure-SAP*. (Cuanto más rojo más hidrofílico, cuanto más azul más hidrofóbico).

Solo en el sistema 6, con el péptido inicialmente colocado en la interficie de la hemimembrana aniónica, se mantuvo durante toda la simulación la conformación en hélice α (figuras 43 y 44 A). Además, durante la dinámica molecular, las cadenas laterales de los residuos positivos del péptido modificaron su posición para aumentar la interacción con las cabezas polares (aniónicas) del lípido y la propia hélice giró 15º hasta adoptar una posición respecto a la membrana claramente anfipática, con los residuos polares orientados hacia el solvente acuoso y los apolares hacia el interior de la membrana. El único triptófano presente en esta hélice parece ser el principal responsable de la interacción hidrofóbica con la membrana, ya que es el residuo más enterrado entre las cadenas lipídicas provocando además un doblado de la hélice de 6º respecto a la conformación de partida completamente plana (figura 44 B). Al mismo tiempo, existen claras interacciones electrostáticas que se forman entre las cadenas laterales de la arginina y las lisinas presentes en la hélice y los oxígenos cargados negativamente de los fosfolípidos (figura 45).



MD time	Helix bend	Helix radius	Membrane inmersion	Helix rotation
0 ns	9°	0,2 Å	-3 Å	0°
1000 ns	15°	0,22 Å	5 Å	-15°

Figura 44. A. Comparación entre el gráfico de evolución DSSP de la hélice libre en agua e insertada en una hemimembrana. **B.** Cuadro superior: posición inicial de la hélice respecto a la hemimembrana (izquierda). Visión cenital (derecha), no hay fosfolípidos sobre la hélice. Cuadro medio: posición final de la hélice, (izquierda). Visión cenital (derecha), ningún fosfolípido se encuentra sobre la hélice. Tabla inferior: comportamiento de la hélice. "Helix bend": combado de la hélice. "Helix radius": amplitud del radio de la hélice. "Membrane inmersion": soterramiento de la hélice medida respecto las cabezas polares de la membrana. "Helix rotation": rotación de la hélice respecto ella misma en la orientación inicial. Las cabezas fosfatadas de fosfolípidos se representan como bolas verdes. Las cadenas laterales de Lys y Arg están coloreadas en azul, la del Trp en naranja.

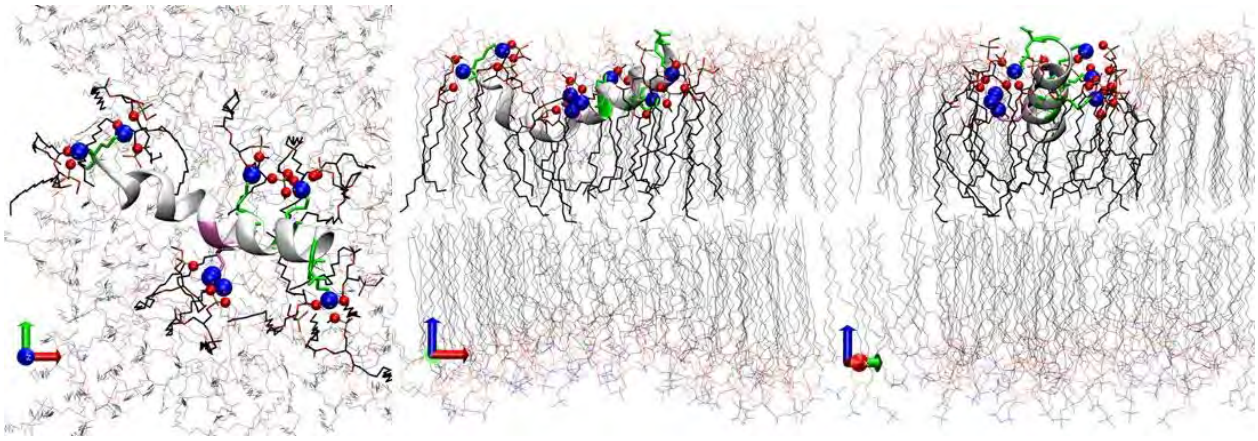


Figura 45. Vista cenital (izquierda), lateral y frontal (centro y derecha) de la hélice 4 en el sistema 6 a tiempo final de la MD (instantánea 931 ns). Los residuos K316, R327, K329 y K332 interaccionan con los oxígenos de los fosfolípidos a través de sus nitrógenos. La cadena lateral de las Lys se muestra en color verde y de la Arg en malva. Los fosfolípidos (cadena hidrocarbonada en negro) muestran los oxígenos situados a 3 Å o menos de los nitrógenos de estos residuos (esferas rojas), en este determinado instante. En la "hélice no-polar" estos residuos han sido sustituidos por alaninas y ninguna de las interacciones electrostáticas mostradas podrá formarse.

Estas interacciones no son fijas, la membrana es fluida y los fosfolípidos se mueven. Se monitorizó el nº de interacciones que se producían entre los oxígenos de los fosfolípidos y los nitrógenos de los residuos Arg y Lys de la hélice, para la MD del sistema 6 (figura 45) (ver métodos), quedando un valor medio de unas 12 interacciones entre el péptido y la membrana. Las interacciones monitorizadas se dan entre todos los oxígenos de los fosfolípidos de la hemimembrana DPM, y los nitrógenos de los residuos K316, R327, K329, K332 y K337, que rinden una media de unos 2-3 oxígenos interaccionando con cada residuo.

La hélice 4 en esta orientación, muestra una mejor conservación de la estructura en hélice α para el péptido inserto en la membrana, que la obtenida en solvente acuoso y se dispone de manera claramente amfipática en la membrana.

Validación de la orientación anfipática

Se preparó un nuevo experimento de simulación para validar la orientación anfipática en el sistema 6. Se construyeron dos nuevos sistemas con membranas, con igual metodología que el sistema 6 pero con la hélice 4 rotada 90º (sistema 7) y -90º (sistema 8) respecto a la hélice en el sistema 6, a lo largo de su eje longitudinal (figura 45). En estas orientaciones, los residuos polares de la hélice ya no están situados longitudinalmente paralelos a las cabezas polares de los lípidos si no que se sitúan perpendiculares al lípido. Con estos sistemas 7 y 8 se llevaron a cabo nuevas simulaciones de MD de larga duración en las mismas condiciones que el sistema 6. **Sorprendentemente, a lo largo de la simulación las hélices giraron espontáneamente hacia la misma posición recuperando una orientación respecto a la membrana equivalente a la del sistema 6.**

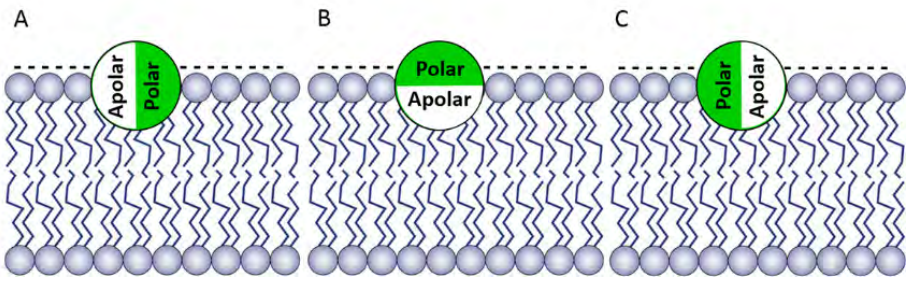


Figura 46. Sistemas 6 (B), 7 (A) y 8 (C). El sistema B es el sistema 6 original, la hélice en los sistemas 7 (A) y 8 (C), ha sido rotada y colocada en la membrana siguiendo el mismo protocolo que para el sistema 6. Los residuos polares (verde) y apolares (blanco) en los sistemas 7 (A) y 8 (C) ya no forman un constructo paralelo a la membrana sino perpendicular a ella.

Durante los 100 primeros ns de dinámica, las hélices de los 3 sistemas rotan optimizando la posición anfipática respecto a la membrana y el solvente. La hélice nativa rota -15º respecto a la orientación inicial.

Dado que la posición inicial de la hélice en el sistema 6 se hizo a mano, estos 15º grados suponen el error sistemático que tendrán los sistemas 7 y 8. Así la hélice rotada -90º del sistema 7 solo necesita rotar 75º para alcanzar la posición final, equivalente a la del sistema 6. Por el contrario, la hélice del sistema 8, rotada 90º, tiene que girar -115º para conseguir esta misma posición de los sistemas 6 y 7; sin embargo, parece que solo poder rotar -75º siendo este es el máximo giro que le permiten las interacciones electrostáticas formadas al inicio de la dinámica molecular (figura 47). Los residuos $Y_{320}YIW_{323}$ forman un núcleo apolar que atrae al péptido hacia el interior de la membrana (figura 48).

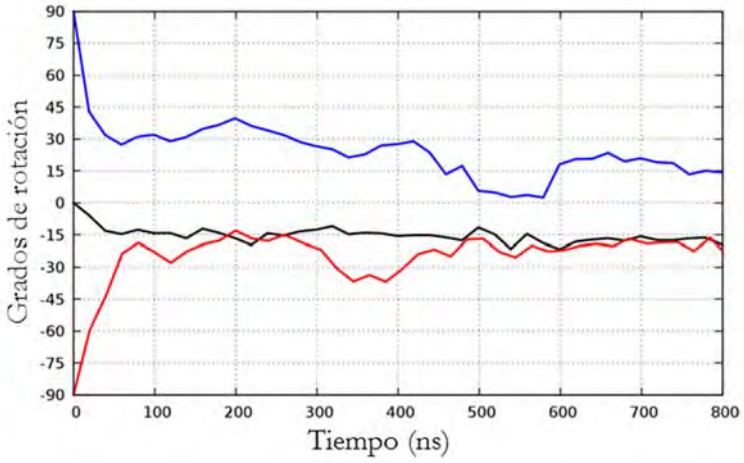


Figura 47. Rotación de la hélice en la membrana en los sistemas 6 (negro), 7 (rojo) y 8 (azul).

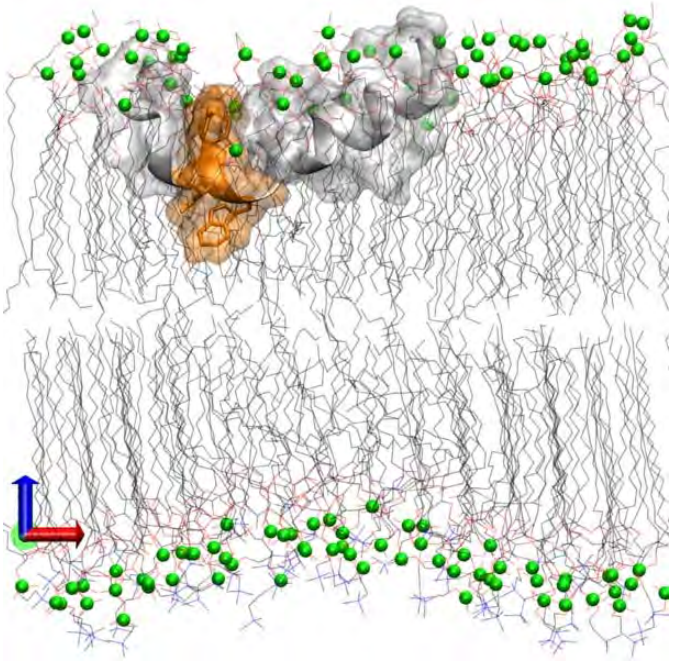


Figura 48. Visión lateral de la hélice 4 en el sistema 6 a tiempo final de la MD (instantánea a 931 ns). La hélice muestra el volumen de las fuerzas de Van der Waals, en naranja para el núcleo hidrofóbico. Los fosfolípidos (cadena hidrocarbonada en negro) muestran las cabezas polares de los fosfatos (en verde).

Estabilidad en membrana de variantes de la Hélice 4.

¿Qué residuos definen el carácter anfipático de la Hélice 4? Para responder a esta pregunta, y a partir de las interacciones observadas entre la hélice y la membrana en el sistema 6, se llevaron a cabo nuevas simulaciones con hélices mutadas, cuyas particularidades hidrofóbicas y electrostáticas estaban modificadas respecto al péptido nativo. Para ello se modelaron dos nuevas hélices, sustituyendo en una los residuos K316A, R327A, K329A y K332A y eliminando así 4 de los 5 residuos que establecen interacción electrostática con la membrana, esta hélice es llamada "hélice no-polar". Para la segunda hélice se sustituyeron los residuos Y320S, Y321Q, I322A y W323S por lo que el núcleo más hidrofóbico de la hélice 4 queda eliminado, llamando a esta "hélice no-apolar".

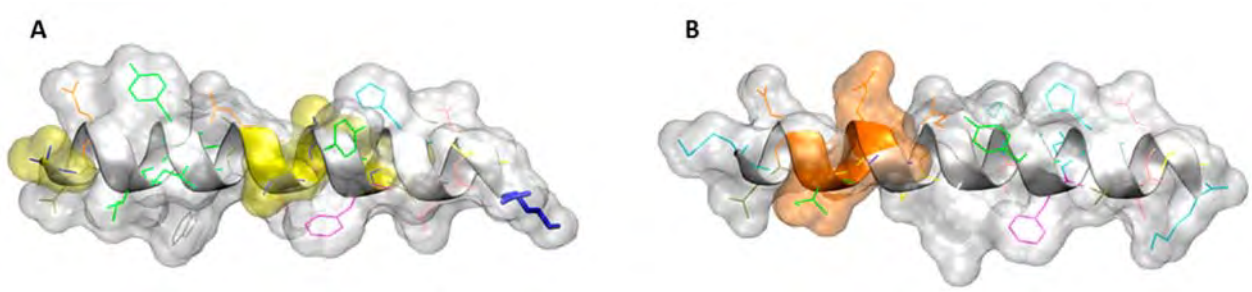


Figura 49. A: Hélice no-polar. En amarillo los residuos mutados K316A, R327A, K329A y K332A, todas las interacciones de estos residuos en la hélice nativa (entre 2 y 4 por residuo) no son aquí posibles, en azul la única interacción posible, K338. B: Hélice no-polar. En naranja, residuos que sustituyen el núcleo hidrofóbico Y320S, Y321Q, I322A y W323S, presente en la hélice nativa.

Se construyeron tres nuevos sistemas para cada una de estas hélices mutadas, idénticos a los sistemas 6, 7 y 8, es decir, misma orientación que la hélice nativa y rotada 90º y -90º, y se inició una nueva ronda de MD de 400 ns de duración (figura 50). La hélice no-polar gira rápidamente (durante los primeros 25 ns) hacia la orientación más estable, que coincide con la del sistema 6. En estas simulaciones la hélice no-polar parece tener un comportamiento más fluido en la membrana que la hélice nativa, pudiendo realizar rotaciones más rápidas y amplias, quizás debido a la pérdida de las limitaciones estéricas que imponen los residuos polares (arginina y lisinas) y a la pérdida de anclaje dirigido por las interacciones electrostáticas. La hélice no-apolar, donde estas interacciones electrostáticas se mantienen y se ha eliminado el núcleo hidrofóbico, nunca rota más de 45º y la posición final del sistema 6 nunca es alcanzado, aunque para esta hélice no-apolar las interacciones electrostáticas son las mismas que las del sistema 6.

Esta diferencia de comportamiento podría deberse a la ausencia del triptófano y del núcleo hidrofóbico en estas simulaciones, que en el resto de sistemas fuerzan a la hélice a girar hasta dirigirlos hacia el interior de la membrana.

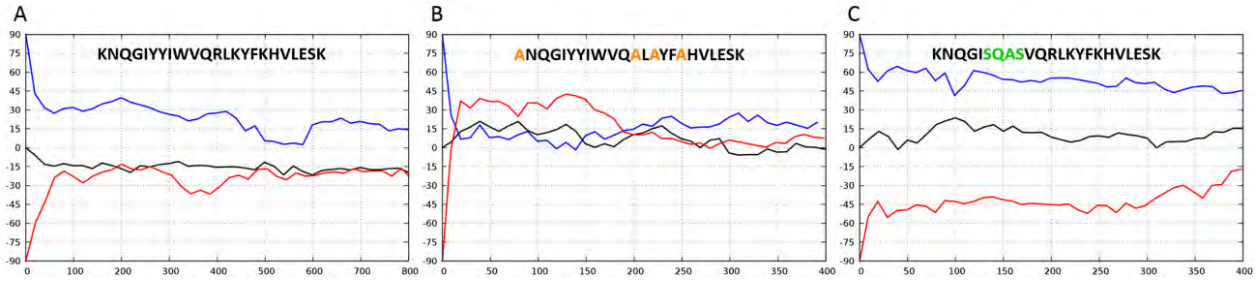


Figura 50. Evolución de la rotación de las hélices wt y mutantes. A. Hélice 4 (hélice wt, sistema 6). B. Hélice no polar. C. Hélice no apolar. Rotación 0º: Trazo negro. 90º: Trazo azul. -90º: Trazo rojo.

La orientación de la hélice respecto a la membrana podría estar gobernada por los residuos apolares mientras que la interacción con ella se debería a los residuos cargados positivamente como la arginina y las lisinas. También es reseñable que la hélice no-polar, que ha perdido prácticamente todas las interacciones electrostáticas, es estable en la membrana por lo que la interacción hidrofóbica es lo suficientemente fuerte como para mantener a la hélice en la membrana.

Está comprobado que la deleción de los últimos 13 aminoácidos de la GT MG517, ocasiona una completa pérdida de actividad para la proteína[79]. Esta forma trucada supone la pérdida de los últimos 10 residuos de la hélice 4. Se modeló la hélice resultante (K_{316}NQGIYYIWVQRL_{328}) de la misma forma que las hélices anteriores (ver métodos) y se llevó a cabo una nueva dinámica molecular, con las mismas condiciones de simulación que para el sistema 6, con el péptido situado en la hemimembrana superior de una membrana lipídica y el cierre de la membrana por el protocolo de $KALP_{15}$. Esta hélice mostró un comportamiento similar al de la hélice 4: conformación estable en membrana y disposición anfipática de sus residuos (figura 51). No se puede por tanto concluir, que la pérdida de actividad de la proteína, por la deleción de los 13 últimos residuos de MG517 (10 últimos de la hélice 4) se deba a la pérdida de interacción de la hélice con la membrana.

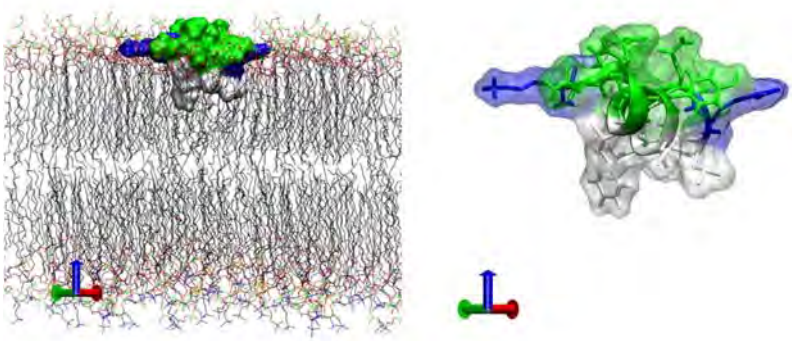


Figura 51. Hélice truncada K_{316}NQGIYYIWVQRL_{328} La hélice se mantiene plegada y estable en la membrana, con una disposición de las cadenas laterales de sus residuos claramente anfipática (verde y azul, residuos polares; blanco, residuos apolares).

5. Energía de unión a la membrana.

Hasta aquí, prácticamente ha quedado establecido ya que la hélice 4 es una hélice de carácter anfipático, estable en la membrana mediante una interacción monotópica cuyos residuos polares se orientan hacia el solvente y establecen interacciones electrostáticas con las cabezas polares de los lípidos, mientras que los hidrofóbicos se entierran en la membrana. Queda por conocer la fuerza de esta interacción, de modo que se permita responder a la pregunta de si esta unión es permanente o bien el péptido puede desprenderse fácilmente. Para ello se calculó la energía libre asociada al proceso de unión a membrana mediante simulaciones de Metadinámica (MetaD). Estos cálculos se llevaron a cabo con la hélice nativa y la hélice no-apolar por separado a fin de comparar la distinta estabilidad termodinámica de estas hélices en membrana. El mapa de energía libre de interacción se evaluó en base a dos variables colectivas: (i) el número de coordinación entre los residuos cargados positivamente de la hélice y los grupos negativos de la hemimembrana superior y (ii) la distancia entre el centro de masas de la hélice y el de la membrana (figura 52) (ver métodos).

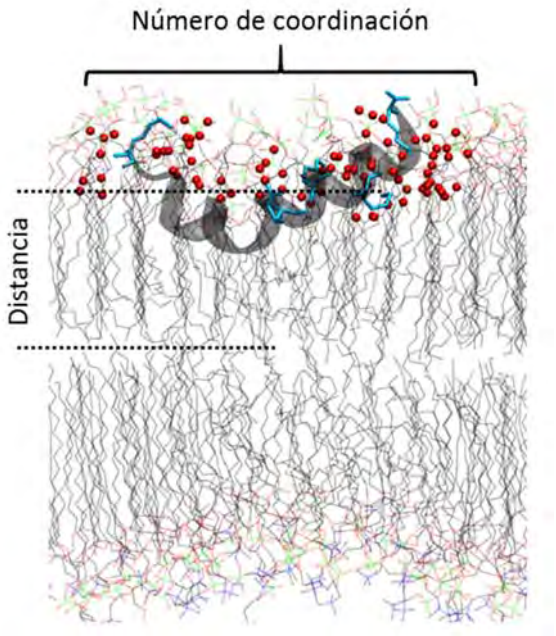


Figura 52. CVs utilizadas en la MetaD: **Nº de coordinación** entre los residuos Lys y Arg de la hélice y todos los átomos de oxígeno cargados de la hemimembrana superior a distancia de interacción (3 Å del N de interacción de los aminoácidos) (esferas rojas). **Distancia** entre el centro de masas del péptido y el centro de masas de la membrana (~1,8 nm). El FES de la simulación se restringió, para que la hélice no fuese más allá del centro de masas de la membrana ni se alejase de ella más de 6 nm.

De este modo durante las simulaciones se dirige a la hélice para que forme o rompa las interacciones con la membrana y para que se separe o se introduzca en la misma, permitiendo reconstruir el mapa energético asociado al proceso de unión en función de estas dos coordenadas geométricas. Las

simulaciones MetaD comienzan con la estructura a tiempo final de las simulaciones MD para los sistemas 6 y 8 (hélice nativa en membrana y hélice no-apolar respectivamente), que contienen el péptido inserto en la membrana en una posición y orientación estables. Estas dos simulaciones, una con la hélice nativa y otra con la hélice no-apolar, se extendieron durante 1 μs. donde se produjeron dos eventos de extracción e inserción del péptido en la membrana, que se analizaron con METAGUI, con un tiempo de equilibrado de 84 ns.

Aunque no se obtuvieron muchos eventos de salida y entrada del péptido en membrana, se puede ver a través de la monitorización de las variables, que la extracción completa de la hélice siempre va acompañada de la pérdida de interacciones con los fosfolípidos de la membrana (figura 53).

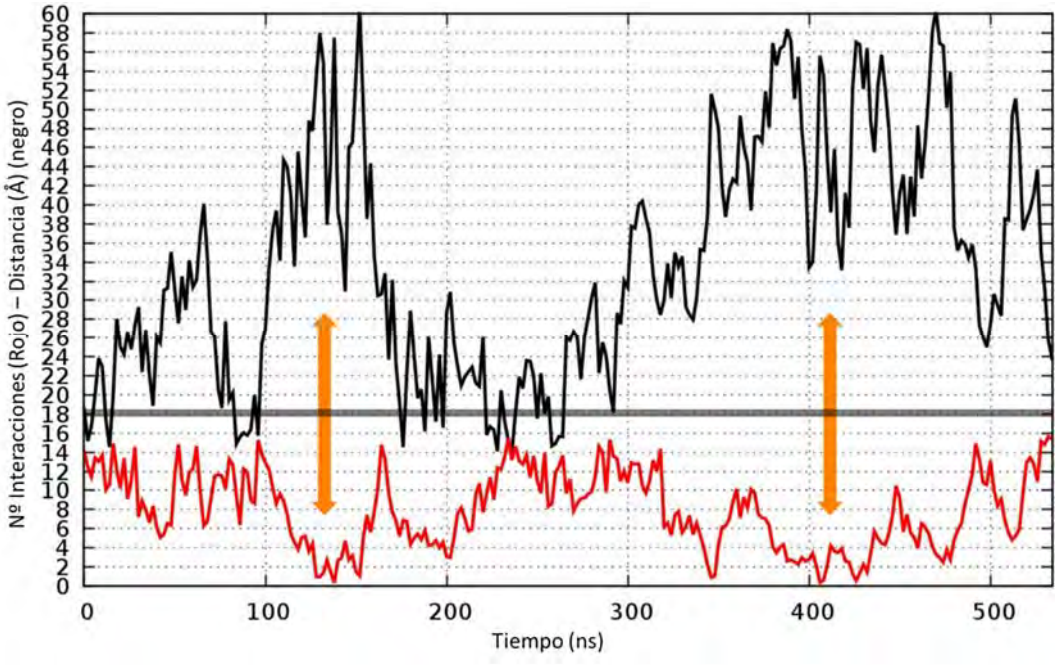


Figura 53. VC de la simulación metadinámica para la hélice 4 nativa. Trazo negro: Distancia entre el centro de masas del péptido y el centro de masas de la membrana. Trazo rojo: Nº de coordinación entre los residuos Lys, Arg del péptido y los oxígenos de la hemimembrana superior. Las flechas indican los momentos de nº de coordinación 0, que corresponden a la extracción completa de la hélice respecto a la membrana. La línea horizontal delimita la membrana superior.

Los mapas energéticos resultantes muestran que, la posición más estable de la hélice 4 nativa es, unida a la membrana (mínimo "in" en figura 54). En realidad existen dos metaestados muy cercanos de misma energía en que la hélice está ubicada alrededor de 20 Å de distancia del centro de masas de la membrana (21 y 18 Å exactamente) con la misma posición y orientación de la hélice vista en la MD del sistema 6 y el mismo número medio de interacciones, 12. La completa pérdida de interacciones entre el péptido y la membrana se consigue a una distancia entre ambos de 50-55 Å (mínimo "out" en figura 54); a una distancia menor, posiciones verticales de la hélice todavía

pueden interaccionar con los fosfolípidos. La energía libre de asociación de la hélice a la membrana se estima como la diferencia de energía libre entre los metaestados "in" y "out" del mapa energético. La alta energía libre de interacción de la hélice nativa hace de esta una interacción irreversible.

Por otro lado, el mapa energético calculado para la hélice no-apolar muestra que la posición más estable de la hélice corresponde a un metaestado en que el péptido se sitúa un poco más enterrado en la membrana que la hélice nativa (mínimo "in" en figura 54 y 55) a una distancia entre centros de masas de 18 Å, así como entre 2 y 4 interacciones más (quizás debido a los residuos mutados). Existe un segundo metaestado en esta zona de similar energía ("lie", figura 54 y 55) que corresponde a una estructura de la hélice con el mismo número de interacciones electrostáticas que la hélice nativa (12 interacciones como número de coordinación) pero que sitúa la hélice no-apolar a una distancia de 27 Å entre los centros de masa de la membrana y el péptido. Este metaestado indica que la pérdida del núcleo hidrofóbico de la hélice mutada elimina también la fuerza de arrastre existente en la hélice nativa hacia el interior de la membrana. Del mismo modo que para la hélice nativa, la completa disociación de la hélice no-apolar ocurre a una distancia de 53 Å (mínimo "out" en figura 54 y 55). La energía libre de unión a membrana de la hélice no-apolar es también muy alta.

Si comparamos las energías relativas de los metaestados dentro y fuera de la membrana para ambas metadinámicas se puede comprobar que se necesita menos energía (~ 5 kcal/mol) para extraer la hélice no-apolar de la membrana que la hélice nativa, por lo que la contribución del núcleo hidrofóbico no sería solo a la orientación de la hélice sino también a su estabilidad en la membrana.

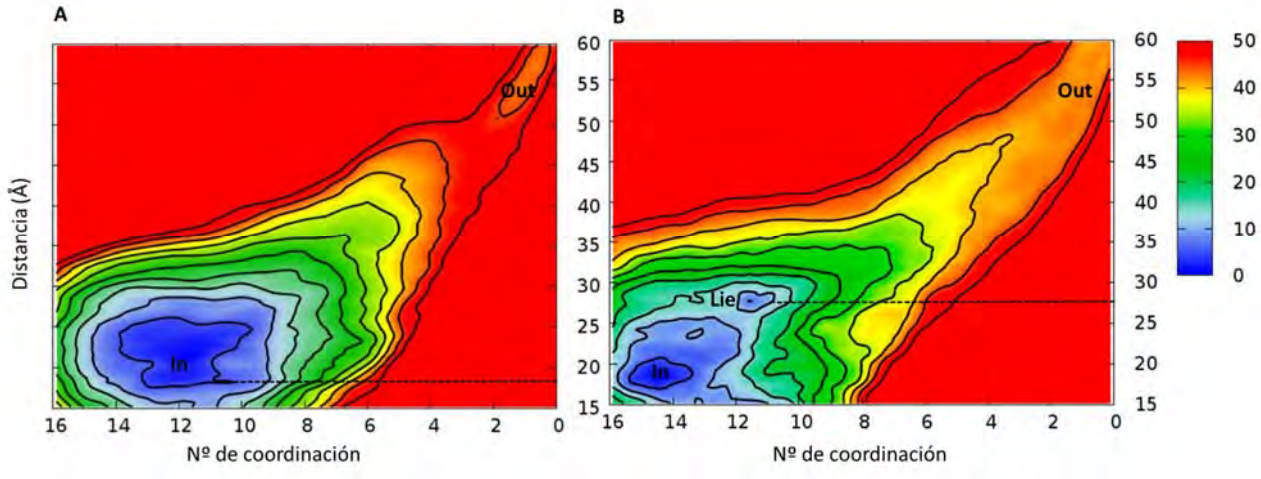


Figura 54. Superficie de energía libre (FES) de la MetaD para la hélice 4 wt (A) y hélice 4 no-apolar (B). Isotermas 5 kcal/mol ± 2.5 kcal/mol. La línea punteada indica la distancia entre hélice y membrana (A) parcialmente soterrada y (B) con total acceso al solvente.

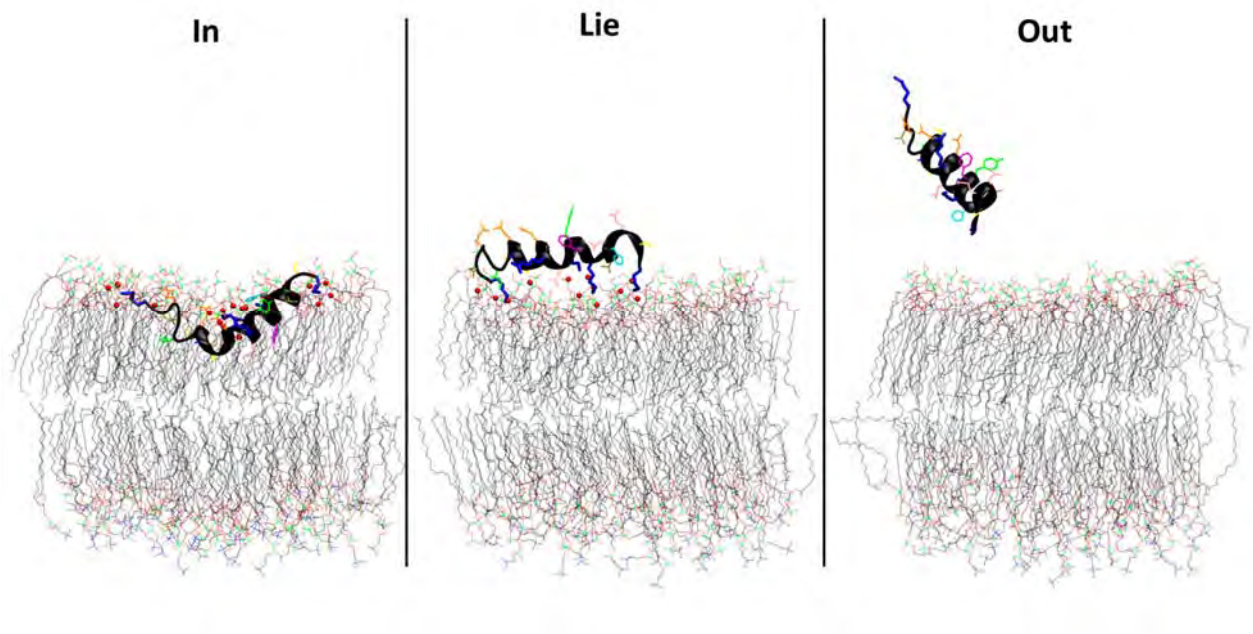


Figura 55. Posiciones de la hélice mutada "no-apolar" en la simulación MetaD. Las esferas rojas son oxígenos de los fosfolípidos a menos de 3 Å de los nitrógenos de los residuos Arg y Lys. "In", la hélice se encuentra inserta en la membrana, unos 16 oxígenos interaccionan con los residuos Arg y Lys. "Lie", la hélice se haya sobre la membrana, unos 11 oxígenos interaccionan ahora con los residuos Arg y Lys. "Out", la hélice se encuentra fuera de la membrana, sin ninguna interacción con los fosfolípidos.

6. Discusión.

Hemos encontrado evidencias computacionales que confirman la existencia de una hélice anfipática localizada en el extremo C-terminal de la GT MG517 que se adhiere monotópicamente a una bicapa lipídica de composición mixta DPPC/DPPM. Esta hélice puede ser uno de los elementos de interacción, quizá el más importante, de la GT MG517 con la membrana celular. Por la energía necesaria para su disociación, la unión de esta hélice a la membrana es permanente y una vez la proteína se encuentra en la membrana, ya no se separa de ella debido al menos esta unión. La extensión C-terminal de la proteína GT MG517 podría estar formada tan solo por hélices alfa, existiendo evidencias por el perfil hidropático de que la primera de ellas se situaría bajo la hélice 6 de la topología consenso de la región Nter, cubriendo el parche hidrofóbico observado en las estructuras modeladas del capítulo 1 (Anexo 10). Entre ésta primera hélice y la última (hélice anfipática) podrían existir otras dos hélices para las que no se ha podido determinar su ubicación en la estructura. La elevada hidrofobicidad de la primera de estas dos hélices sugiere que más que un elemento de interacción con la membrana, se trata de un elemento estructural interno de la proteína, debido a la ausencia de hélices transmembrana.

Puesto que la deleción de parte de la hélice anfipática no supone la pérdida de asociación con la membrana pero sí provoca la inactivación completa de la enzima, tal como se ha comprobado experimentalmente[116],[79] y también computacionalmente respecto a la interacción, especulamos con una posible pérdida de la función sensora como causa de la pérdida de actividad[97]. Este rol, por tanto, podría ser aplicado a la hélice anfipática según esta evidencia o quizás también, a la pérdida de comunicación con la enzima para trasladar información sobre la membrana, cuyo papel como sensor de la membrana para el péptido vuelve a ser necesario. En este contexto, es interesante volver a los resultados del servidor Robetta mostrados al inicio de los resultados de este capítulo (Figura 34), que muestra un contenido estructural rico en hélices en la región C-terminal, coincidiendo con los resultados de PsiPred. Para la estructura central de la figura 34, una hoja β se forma entre los extremos de las hélices última (posible hélice anfipática,) y penúltima, adentrándose hacia la primera hélice (posible hélice α7 en MG517, cubriendo el parche hidrofóbico, Anexo 10). En la estrutura completa de MG517, esta conformación situaría el giro de dicha hoja β cerca del centro activo (zona del aceptor) y la RV. Esta conformación podría ser compatible con la transmisión de información al centro activo a través de un elemento remoto, como sucede en otras proteínas que actúan sobre carbohidratos[117], siendo esta hipótesis, meramente especulativa.

Dado el reducido genoma de *M. genitalium*, no deben ser muy numerosas las proteínas que se encuentren de algún modo asociadas a la membrana. Este organismo es capaz de dividirse en ausencia de la proteína FtsZ[118], indispensable para la formación del *septum* en la mayoría de bacterias. Se ha propuesto para *M. genitalium* que una compleja red de proteínas, asociadas con su biología podría tomar la función de genes asociados a la división[118], como proteínas *moonlight* entre las cuales, por su relación con las propiedades físico-químicas de la membrana, podría estar la GT

MG517. El software MPEx encontró una posible secuencia señal para el traslocón, que toma parte de las hélices 2 y 2.1 (secuencia R_{267}FWRRQMFVWFALFSFEYF$_{285}$). *M. genitalium* posee en su genoma genes para las subunidades secY y secA de la traslocasa, careciendo del resto de las proteínas necesarias para formar el traslocón. Si se observase que *M. genitalium* es capaz de adquirir proteínas del huésped, algo no comprobado pero siendo este un claro ejemplo de reducción genómica y de su parasitismo obligado, podría adquirir el resto de subunidades necesarias para el traslocón y ser un mecanismo activo en este organismo. Aunque todavía no se ha identificado si la GT MG517 desarrolla su función en una o las dos caras de su membrana, el mecanismo de paso a través de esta podría estar mediado por un traslocón.

Con este trabajo, creemos haber contextualizado suficientemente la interacción de MG517 con la membrana y definido los lugares de interés de la región C-terminal. De este modo dejamos allanado el camino para un abordaje experimental directo, habiendo propuesto incluso los mutantes a desarrollar, su comportamiento en membrana, los residuos que intervendrán en la interacción y su forma de hacerlo. Hemos encontrado y/o propuesto una explicación razonada y computacionalmente demostrada, de los resultados experimentales de la proteína MG517 sobre su purificación y activación por fosfolípidos.

También hemos relacionado la región variable y la posición del aceptor en ella, con una posible interacción con la membrana en su parte inicial. Y aportamos indicios de otras funciones de esta proteína que pueden afectar al desarrollo de *M. genitalium*.

En resumen, hemos generado conocimiento, y lo hemos hecho observando, planteando hipótesis, realizando experimentos (indirectos o computacionales) y demostrando la validez de nuestras conclusiones. El diseño racional de fármacos contra *M. genitalium*, dispone ahora de nueva información y de nuevas zonas que explorar en este organismo, obtenidas mediante una metodología teórica y lista para ser confirmada experimentalmente.

CAMBIOS CONFORMACIONALES DEL BUCLE CATALÍTICO EN EL MECANISMO DE LA PROTEÍNA GPGS.

Albesa-Jové, D. *et al.* A Native Ternary Complex Trapped in a Crystal Reveals the Catalytic Mechanism of a Retaining Glycosyltransferase. *Angew. Chem. Int. Ed. Engl.* **54,** 9898–902 (2015).

1. Introducción

La proteína GpgS es una glicosiltransferasa del organismo patógeno *Mycobacterium tuberculosis*. Cataliza la transferencia de una molécula de glucosa a otra de fosfoglicerato (PGA). Se trata de una proteína que se encuentra libre en el citosol de este microorganismo y que inicia la ruta biosintética de los lipopolisacáridos 6-O-metilglucosa (MGLPs) (figura 19). Estos lipopolisacáridos se han propuesto como transportadores de ácidos grasos de cadena larga, en el citosol de este patógeno y cualquier proteína que participe en esta ruta, sería una buena candidata a tratar como diana terapéutica. De este modo, todos los estudios que ayuden a desentrañar el mecanismo de acción de GpgS o cualquier otra característica asociada a esta proteína, pueden también ayudar a encontrar un mecanismo de inhibición de su actividad, la alteración de la ruta biosintética de los MGLPs y un posible fármaco contra la tuberculosis.

En este capítulo se ha abordado un aspecto fundamental de cualquier enzima, como son los cambios conformacionales asociados a la unión de sustratos en el centro catalítico. Este aspecto es de particular relevancia en GTAs, para las que se han descrito la existencia de un bucle flexible que parece tener alguna función catalítica y relacionada además con la presencia de ligandos.

En las estructuras GTAs disponibles en literatura, entre la región Nter y Cter suele existir una zona desestructurada, flexible, ya que existen muchos cristales en donde no ha podido ser resuelta. Cuando es posible ver esta pequeña región suele ser formando una lámina β con el elemento β4' de la topología consenso (figura 24), este último justo a continuación del motivo DXD. En la β4GalT1 de la familia GT7, comparaciones estructurales entre las formas apo y en complejo con ligandos muestran que la unión del ligando dador podría inducir un cambio conformacional en este bucle, que estabiliza el aceptor y crea el sitio de entrada para el dador[119]. Las glicosiltransferasas que generan los grupos sanguíneos A y B, la GalNAc-transferasa (GTA) y Gal-transferasa (GTB) de la familia GT6 también contienen este bucle, estando aquí en la zona final de la región C-terminal y mostrando conformaciones abierta o cerrada también en función de la unión del dador[120]. En la GT27 GalNAc-T2, ese mismo bucle existe de nuevo en las conformaciones abierta y cerrada, abierta en la forma apo y cerrada cuando está en complejo con el dador[121]. Aquí además la presencia del azúcar en el dador, estabiliza la forma cerrada y su pérdida tras la catálisis facilita la apertura del mismo y la salida del dador. El mismo bucle flexible y la misma relación con la unión de los ligandos la encontramos en la GT78 MpgS[122].

Recientemente ha sido posible resolver este bucle en GpgS ($R_{256}AHRN_{260}$) en dos diferentes conformaciones en el mismo cristal (PDB 4DDZ)[81], una estructura apo, en una proporción 44:56, que se han denominado conformación LA, visible en proteínas cristalizadas con ligando y conformación LI, observada en estructuras apo, sin ligando (Tabla 10). Estas conformaciones según la presencia o no del ligando, sugiere que este podría provocar el cambio conformacional (mecanismo de ajuste inducido), como es común en otras glicosiltransferasas[123,124,125]. Sin embargo, la presencia de ambas conformaciones en una misma estructura apo en GpgS, indica que el cambio conformacional es una propiedad intrínseca de esta proteína, aludiendo a un posible mecanismo de selección conformacional.

	PDB	Ligando	Conformación del bucle
M. tuberculosis	3E25	UDP, 3PG, Mg	No resuelto
	3E26	Apo	LI
	4DDZ	**Apo**	**LI/LA**
	4DE7	UDP	LA
	4DEC	UDP, 3PG, Mn	LA
	4Y6N	**UDPGlc, 3PG, Mn**	**LA**
M. paratuberculosis	3CKJ	Apo (MRD) 4	LA
	3CKN	UDP, MN	LA
	3CKO	UDP	LA
	3CKQ	UDP-Glc, MN	LA
	3CKV	UDP	LA

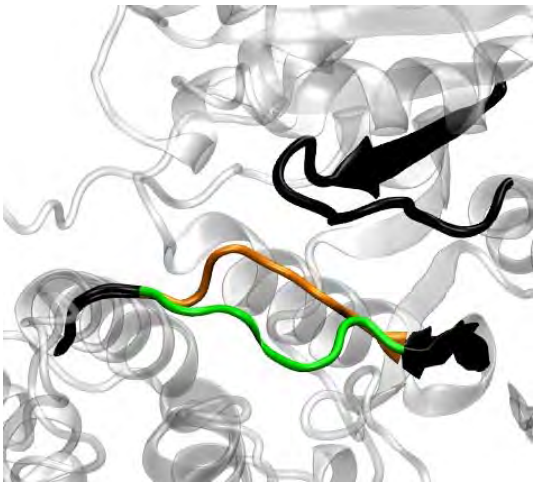


Tabla 10. Cristales de la proteína GpgS. y **Figura 56**. Ampliación de la zona del bucle RAHRN en el cristal 4DDZ para un homodímero. La conformación LA se muestra en color naranja, la conformación LI en verde. El motivo DXD queda sobre estas, coloreado en negro. Ambas conformaciones coexisten en el cristal.

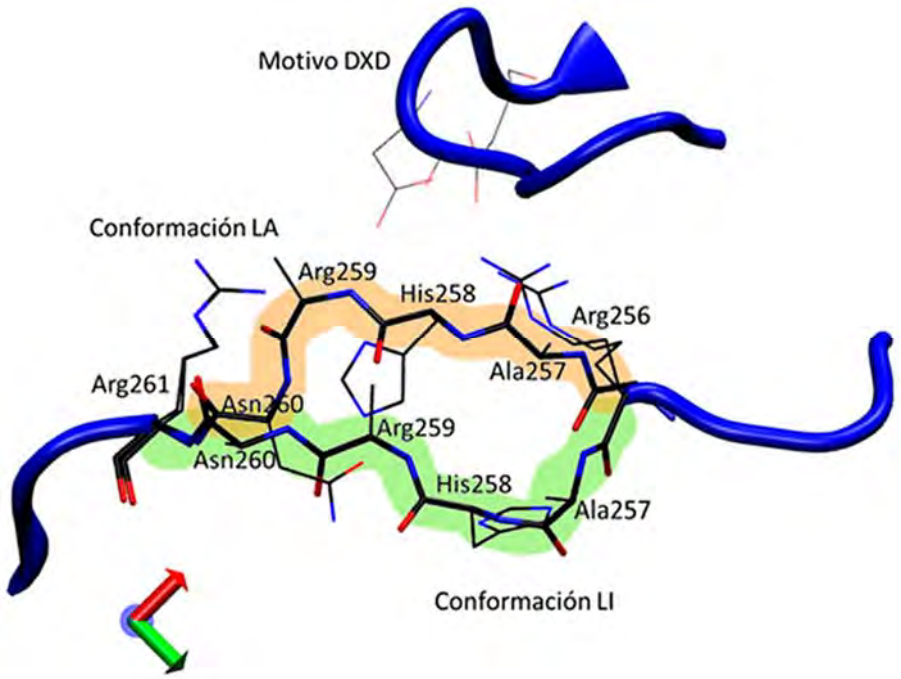


Figura 57. Conformaciones LA (sombreado naranja), LI (sombreado verde) y cadenas laterales de los residuos del cristal 4DDZ, algunas no están resueltas (Arg259 en la conformación LA y LI y Arg260 en LI).

La proteína GpgS se halla en complejo ternario en la estructura 4Y6N, esto es, la proteína junto con el ligando dador (UDP-Glc) metal (Mn) y también el ligando aceptor: fosfoglicerato (PGA). En esta estructura el bucle se muestra en una única conformación, la forma LA.

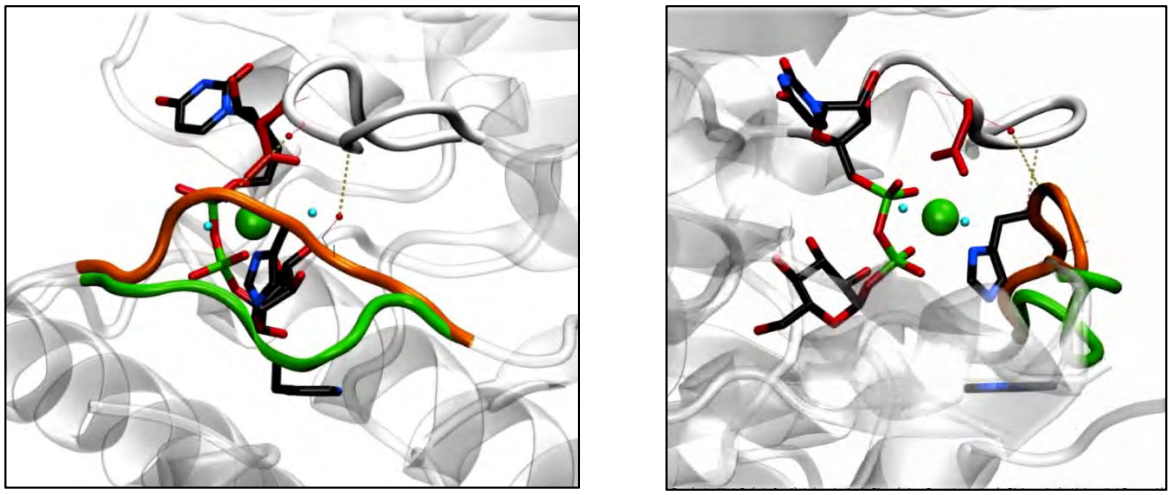


Figura 58. Imagen del centro activo de la estructura ternaria 4Y6N con sus ligandos. El bucle RAHRN se muestra en naranja, **el bucle en conformación LI (verde) de la estructura 4DDZ se ha superpuesto en la imagen, ya que no está presente en el cristal 4Y6N**. Puentes de hidrógeno que estabilizan la conformación LA están marcados por líneas puenteadas. El metal interacciona con el ligando dador, aguas y la histidina 258 del bucle RAHRN. La única interacción posible entre His258-Metal se da en la conformación LA.

Estudios cinéticos y estructurales han revelado que muchas enzimas GTA, siguen un mecanismo ordenado en el cual el catión divalente y el dador se unen primero al centro activo, seguidos por el aceptor[126,127,128] y que se ha visto confirmado experimentalmente, concretamente en GpgS[81]. Sin embargo, estudios experimentales posteriores con esta misma proteína, apuntan al orden inverso, siendo el aceptor PGA quien une primero entrando después el dador[129].

Mediante simulaciones de dinámica molecular y metadinámica, nos hemos propuesto estudiar este bucle flexible en la proteína GpgS y cuantificar su equilibrio en presencia y ausencia de ligandos, para así determinar si sus diferentes conformaciones obedecen a un mecanismo de ajuste inducido o de selección conformacional, así como el posible orden de unión de los ligandos.

Este capítulo, algo más largo y complejo que los anteriores, está estructurado de la siguiente manera:

A. Modelado de estructuras y análisis de los modelos.
B. Análisis del bucle en las diferentes estructuras para encontrar descriptores de las conformaciones del bucle. Comparación de los descriptores con proteínas homólogas.
C. Simulaciones de la **estructura apo**, Dinámica Molecular clásica, Metadinámicas bidimensionales y por método de BIAS-Exchange. Monitorización y uso de los descriptores.
D. Simulaciones de las **estructuras con ligandos**, Dinámica Molecular clásica, Metadinámicas bidimensionales y por método de BIAS-Exchange. Monitorización y uso de los descriptores. *Docking* de ligandos.
E. Estudio dinámico de la hidropatía del CA.
F. Discusión.

Las simulaciones realizadas y su nomenclatura en el trabajo son las siguientes:

PDB	**Ligandos**	**Método de simulación**	**Nombre**
4DDZ	Apo	Dinámica Molecular	MD4DDZApoLI
		Dinámica Molecular	MD4DDZApoLA
		Metadinámica	MetaD4DDZApoLA
4Y6N	Apo	Dinámica Molecular	MD4Y6NApo1
		Dinámica Molecular	MD4Y6NApo2-10
		Metadinámica	MetaD4Y6NApo1
		Metadinámica	MetaD4Y6NApo2
		Metadinámica	MetaD4Y6NApo3
		BIAS-Exchange	MetaD4Y6NApo4
	Complejo ternario UDP-Glc + PGA + metal	Dinámica Molecular	MD4Y6N
		Metadinámica	MetaD4Y6NUDPGlc·PGA
	Complejo binario UDP-Glc + metal	Metadinámica	MetaD4Y6NUDPGlcFree
		Metadinámica	MetaD4Y6NUDPGlcFix
		Metadinámica	MetaD4Y6NUDPGlc3
		BIAS-Exchange	MetaD4Y6NUDPGlc4
		BIAS-Exchange	MetaD4Y6NUDPGlc5
	Complejo PGA	Metadinámica	MetaD4Y6NPGA

Las simulaciones en las que se halla presente el metal, no contemplan la coordinación His258-Metal –visible en la estructura del complejo ternario–, más que en su componente electrostática. Dado que el fin de este capítulo es estudiar la cinética del bucle RAHRN y no la coordinación del metal, consideramos acertadas y suficientes las metodologías utilizadas para ello, aún con este *handicap*, tal y como se ha hecho ya en otros estudios, también de glicosiltransferasas[130].

2. Construcción de modelos a partir de estructuras cristalográficas

Modelado del cristal 4DDZ

Esta estructura es un monómero que no tiene resueltos los residuos 167 a 182, de la RV, y 295 a 301 de un bucle externo, ni está tampoco resuelta la cadena lateral de la Arg259, así como algunos átomos de la Asn260 en la conformación LI. Los residuos Gly254 a Pro262 tienen una ocupación 56:44, que corresponden a las conformaciones LA:LI, estos se separan a mano, creándose dos archivos pdb, donde cada uno contiene el bucle en una sola conformación.

Por simetría se obtuvo el homodímero y se modelaron *de novo* los átomos que faltaban con *Modeller* v9.8. Para cada conformación se realizan 10 modelos, cada uno con 10 bucles generados por *Modeller*, en total 100 estructuras, donde en todas se forzó la generación de un puente disulfuro entre la Cys178 de ambos monómeros, puente que se ha comprobado experimentalmente existe y une covalentemente ambos monómeros. El modelado se realiza fijando todos los residuos de la proteína para que el modelo sea lo más parecido posible al cristal de partida, con la excepción de los residuos 167 a 185 y Arg259 en la conformación LA y LI y Asn260 en la conformación LI (ver *Inputs*).

El modelado presentó problemas, ya que enlazaba los bucles de la RV de un monómero con el bucle del otro monómero. Solo cuatro modelos de los 100 de la conformación LA presentaban una estructura donde los bucles no se hubieran enlazado, además nunca se consiguen modelar correctamente los dos monómeros a la vez, por lo que se toma el monómero A de uno de estos 4 modelos y se superpone al B, construyendo una estructura dimérica que servirá de estructura molde para un segundo modelado con las mismas características del primero.

Esta misma estructura molde fue la que se utilizó para el modelado de las estructuras con conformación LI. Antes del modelado, los residuos del bucle LI más uno extra a cada extremo (253 a 263) fueron transferidos manualmente desde la estructura original a la estructura molde. Esta elaboración de los bucles, generó dos *clashes* entre los residuos Arg167 de la RV y la His258 del bucle, y también entre la Pro168 de la RV y la Ala257 del bucle, sin embargo, estos fueron subsanados por la minimización del modelado.

La selección del modelo final se hizo mediante un *cluster analysis* realizado con GROMACs (ver métodos).

Modelado del cristal 4Y6N

Se utilizó la estructura molde que se acaba de describir para el modelado de esta segunda estructura. Aquí los residuos no resueltos en el cristal son: 167 a 179, de la RV y 295 a 302, del bucle externo. También se forzó la creación del puente disulfuro Cys179. En este caso los modelos se realizaron con los ligandos en el interior: Metal, UDPGlc y PGA. El Mn se sustituyó por Mg.

La selección del modelo final se hizo mediante un *cluster analysis* realizado con GROMACs, (ver métodos).

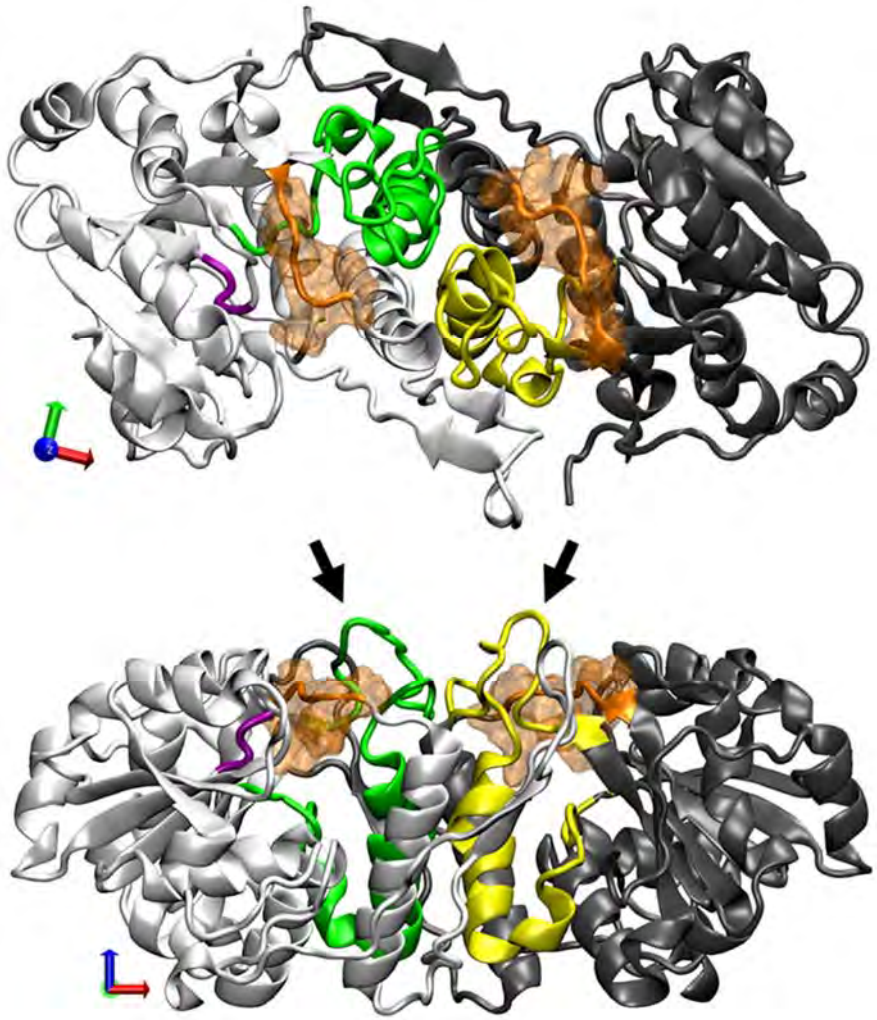


Figura 59. Modelo del dímero de GpgS en el cristal 4Y6N. Representación frontal (arriba) y lateral (abajo). Monómero A representado en color blanco con la RV en color verde y Monómero B en color gris con la RV en color amarillo. La zona del bucle, que se muestra en conformación LA se ha coloreado en naranja para los dos monómeros, así como el motivo DXD en púrpura. Las flechas apuntan a las zonas de la RV modeladas *de novo*. Puede observarse como estas se extienden hacia el exterior de la proteína. Los ligandos no se han representado.

Las formas apo del cristal 4Y6N provienen de la eliminación de los ligandos en los modelos generados para esta estructura (ver métodos), tras la realización de una dinámica molecular del complejo ternario, (ver métodos) y cuya dinámica se desarrollará en el punto 4 de este capítulo.

3. Estudio de los modelos y búsqueda de descriptores del bucle.

Se comparó el valor de los diedros φ y ψ de los residuos del bucle y de sus extremos, para las dos conformaciones del cristal 4DDZ.

Ala257 y Arg259 tienen una variación de más de un 30 % en sus diedros φ para cada conformación. El diedro ψ de la Ala257 presenta casi un 75 % de variación entre las formas LA y LI.

	LA	LI		LA	LI	
Residuo	**φ**	**φ**	Var φ %	**ψ**	**ψ**	Var ψ %
G254	1,97	1,90	2,2	-1,78	-1,63	4,8
V255	-2,46	-2,74	15,3	2,47	2,21	8,3
R256	**-1,90**	**-1,88**	0,6	**2,20**	**2,31**	**3,5**
A257	**-1,94**	**-0,92**	**32,5**	**2,42**	**-1,52**	**74,5**
H258	**-2,77**	**-3,06**	9,2	**2,80**	**2,08**	**2,29**
R259	**-1,33**	**-1,17**	5,1	**1,68**	**2,99**	**41,7**
N260	**-1,03**	**-1,86**	**3**	**2,91**	**2,23**	**4**
R261	-2,46	-2,42	26,4	2,62	2,59	0,95
P262	-1,22	-1,04	5,7	2,59	2,46	4,1
L263	-1,11	-1,03	2,5	-0,66	-0,65	0,3

Valor de los diedros φ y ψ de los residuos del bucle (negrita) y colindantes expresados en radianes.

Se compararon entonces estos valores del cristal con los modelos generados verificando si la diferencia entre las conformaciones seguía conservándose (Anexo 12). Las estructuras modeladas mantienen el valor de los diedros de las plantillas. La principal variación entre el cristal y los modelos se da en los residuos Arg259 y Asn260, residuos que están parcialmente resueltos en el cristal y aquí han sido modelados *de novo*; en los modelos la variación del diedro de Arg259 entre LA y LI se ha reducido respecto al cristal.

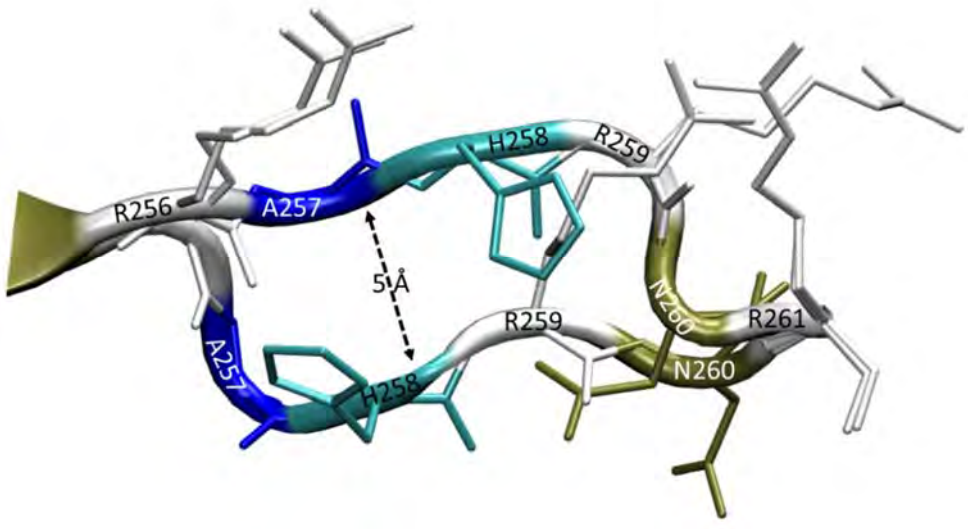


Figura 60. Conformación LA y conformación LI superpuesta del bucle RAHRN de modelado del cristal 4DDZ. La imagen se muestra desde el lado opuesto a la de las figuras 56 y 57. Todas las cadenas laterales están ahora representadas.

Con el fin de profundizar en las diferencias entre conformación LA y LI del bucle en distintas estructuras cristalográficas, y poder monitorizar las conformaciones del bucle a lo largo de las simulaciones, se escogieron los siguientes parámetros estructurales para describir la conformación del bucle:

1. Distancia entre los residuos del bucle y la hoja β4', definidos como la distancia Asp136(O)-Arg259(N) e Ile138(N)-Ala257(O), ya que estas últimas distancias representan la formación o rotura de puentes de hidrógeno entre el bucle y la hoja β4', que parecen apropiadas para explicar la estabilidad de las distintas conformaciones del bucle. No se utilizaron las distancias con los hidrógenos pues las estructuras PDB no las incorporan.

2. La posición de la cadena lateral del residuo His258. En la forma LA, el residuo His258 y se posiciona de tal modo, que el nitrógeno ND1 (ha de estar desprotonado) estabiliza el metal necesario en la reacción. En la forma LI, la cadena del residuo His258 no puede estabilizar el metal.

3. Diedros de Ramachandran de todos los residuos que forman el bucle y anejos (Gly254-Leu263). En total 20 ángulos diedros (φ y ψ) de la cadena lateral, de los cuales se escogieron los que presentaban mayor variabilidad (Tabla 11)

Además de los dos cristales mencionados hasta ahora, existen otros para la proteína GpgS, así como para la proteína homóloga MpgS (manosil glicerato sintasa), pertenecientes a las familias GTA: GT81 y GT78 respectivamente. Los parámetros anteriores se analizaron en los cristales de estas proteínas homólogas; comparando a la vez si la presencia del ligando afectaba de algún modo a estos valores.

Existen 6 estructuras GpgS de *M. tuberculosis*, entre las que están los cristales 4DDZ y 4Y6N y 5 estructuras de *M. paratuberculosis*, además también se compararon 6 estructuras de MpgS de *R. xylanophilus*. En cada una se apuntó la presencia de ligandos, la conformación (por inspección visual) del bucle en caso de estar resuelto y los parámetros estructurales que se acaban de definir.

En la Tabla 10 se recogen, para cada uno de los cristales de estas proteínas, los diferentes ligandos y la conformación del bucle, los valores de la distancia de los puentes de hidrógeno, que forman la lámina β entre el bucle y la hoja β4', y el diedro CG-CB-CA-C de la His258:

	PDB	Ligandos	Bucle resuelto	Conformación	Distancia PdH (Å)		Diedro His(rad)
	M. tuberculosis				D136(O)-R259(N)	I138(N)-A257(O)	His258
GpgS O05309	3E25	UDP, 3PG, Mg	No	?	-	-	-
	3E26	Apo	Sí	LI	7,33	6,73	2,98
	4DDZ	**Apo**	**Sí**	**LI/LA**	**8,23/3,96**	**10,73/2,96**	**2,22/-0,38**
	4DE7	UDP	Sí	LA	4,29	3,03	-0,97
	4DEC	UDP, 3PG, Mn	Sí	LA	4,18	2,71	-1,02
	4Y6N	**UDPGlc, 3PG, Mn**	**Sí**	**LA**	**3,9**	**2,82**	**-0,95**
	M. paratuberculosis				D141(O)-R264(N)	I143(N)-G262(O)	His263
GpgS Q73WU1	3CKJ	Apo (MRD) 4	Sí	LA	3,06	2,82	-0,81
	3CKN	UDP, MN	Sí	LA	3,83	3,01	-0,59
	3CKO	UDP	Sí	LA	3,08	2,79	-0,86
	3CKQ	UDP-Glc, MN	Sí	LA	4,04	2,96	-1,23
	3CKV	UDP	Sí	LA	3,57	2,91	-0,81
	R. xylanophilus				D137(O)-R257(N)	R139(N)-Q255(O)	His258
MpgS Q1ATN7	3F1Y_A	apo (Mg not in DXD)	Sí	LI	7,37	5,89	-1,84
	3F1Y_C	apo (Mg not in DXD)	Sí	LI	7,42	7,07	1,42
	3KIA_A	5GP, MG	Sí	LI	6,21	5,38	0,87
	3KIA_C	5GP, MG	No	?			
	3O3P_A	MG, GDD	Sí	Intermedia	5,62	4,35	1,86
	3O3P_B	MG, GDD	Sí	LI	7,42	6,03	1,08

Tabla 11. Parámetros entre el cristal 4DDZ de GpgS y otros cristales y proteínas homólogas. 4DDZ presenta valores de distancia y ángulo más extremos que el resto de proteínas en sus conformaciones LA y LI.

La distancia para el puente de hidrógeno Asp136-Arg259 en la conformación LI en el cristal 4DDZ, es mayor que en el resto de estructuras con la misma conformación, mientras que para LA se sitúa en valores medios. La distancia del segundo puente de hidrógeno Ile138-Ala257 es muy variable. Respecto a los ligandos, para la proteína GpgS la conformación LI parece solo observarse en las estructuras apo, mientras que la conformación LA está asociada estructuras con ligandos, salvo en un caso, el cristal 3CKJ de *M. paratuberculosis*, que tiene un metilpentanodiol en el lugar del aceptor. En MpgS, sin embargo, solo se observa la forma LI; existen estructuras con ligandos, pero estos son moléculas análogas al ligando original y no el ligando en sí.

La Tabla 10 muestra que los residuos Ala257 y Arg259 tienen más de un 30 % de variación (Ala257 un 75 %) entre las conformaciones LA y LI del cristal para el diedro ψ. Se compararon entonces los diedros φ y ψ de cada uno de los residuos del bucle RAHRN en los distintos cristales homólogos (figura 61).

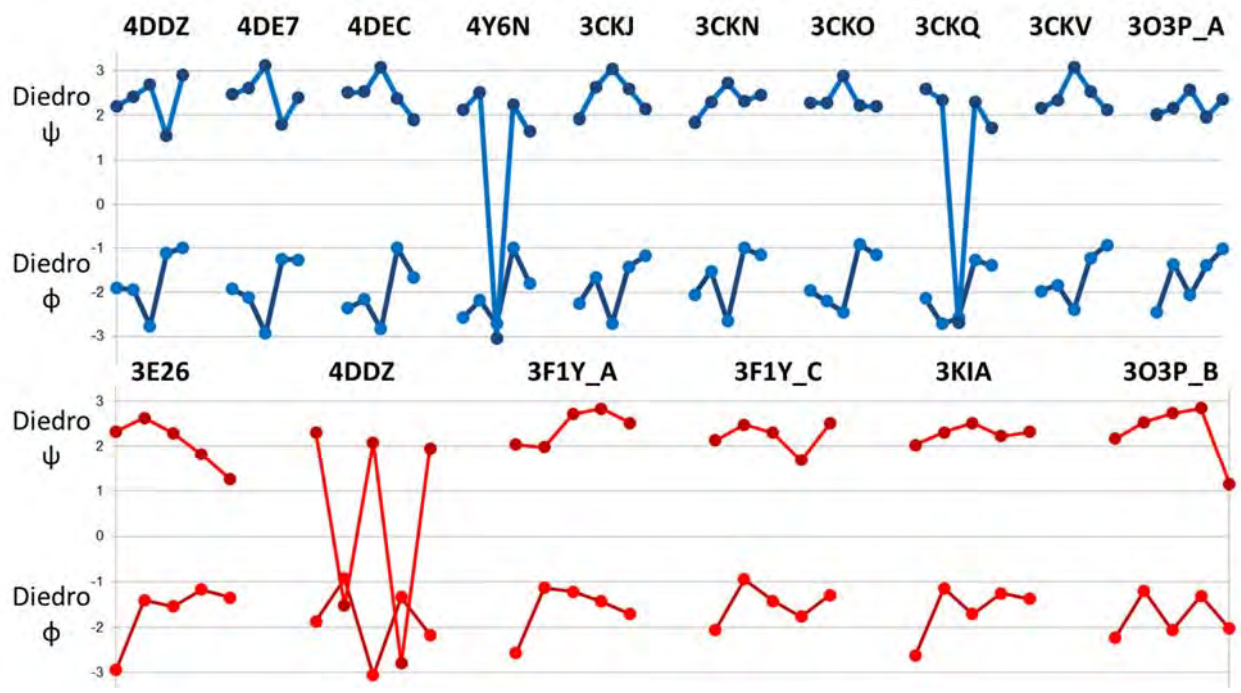


Figura 61. Diedros ψ y φ del bucle RAHRN en la familia GT81 y GT78 para las conformaciones LA (azul) y LI (roja). **Los residuos RAHRN y homólogos están representados por puntos en el mismo orden**.

Los diedros ψ de Ala257 y Arg259, en la forma LA del cristal 4DDZ, no solo presentan gran variación entre las conformaciones LA y LI, sino que también parecen encontrarse en conformaciones aberrantes respecto al resto de proteínas; además, se sitúan en valores prohibidos en el diagrama de Ramachandran. El diedro ψ de la His258 para el cristal 4Y6N también se desvía del resto, como el cristal 3CKQ, pero no están situados en valores prohibidos en el diagrama de Ramachandran.

Así pues, de todos los parámetros estructurales que se han analizado, los seleccionados por presentar mayores diferencias entre conformaciones distintas del bucle son:

1. **Distancia**: Como expresión de la formación y rotura de los puentes de hidrógeno entre Asp136(O)-Arg259(N) e Ile138(N)-Asp257(O).
2. **Diedro ψ del residuo Ala257**: Como reflejo de la forma del bucle.
3. **Diedro (C-CA-CB-CG) del residuo His258**: Como representación de la orientación óptima para coordinación con el metal del centro catalítico.

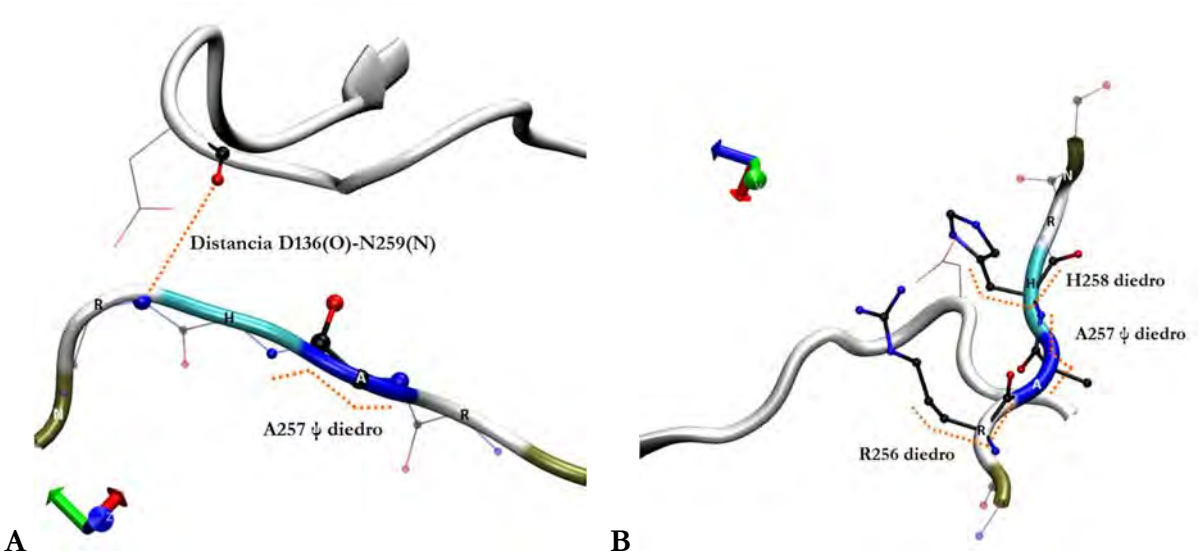


Figura 62. Parámetros estructurales que permiten monitorizar el cambio conformacional del bucle en GpgS: A: 1. Distancia de puente de hidrógeno entre Asp136 y Asn259, 2. diedro ψ de la cadena principal de Ala257, B: 3. diedro CG-CB-CA-C de His258 y 4. diedro CG-CB-CA-C de Arg256.

Nota: Se prescinde de estudiar la Arg259, porque simulaciones posteriores muestran que una simple minimización de energía del modelo reduce la variación LA/LI de este residuo a poco más de un 2%.

Se utilizarán combinaciones de estas variables para la representación, en gráficos bi-dimensionales y de manera inequívoca, de la conformación del bucle en las estructuras de GpgS y proteínas homólogas (ver a continuación Figuras 63, 64 y 65). Así mismo, se utilizarán dichas variables para monitorizar el progreso de las simulaciones de dinámica molecular y como coordenadas de reacción en los cálculos de energía libre basados en metadinámica (ver siguientes apartados).

Figuras 63, 64 y 65: Comparativas entre los diferentes cristales estudiados. Contorno de figuras: conformaciones LI. Figuras opacas: Conformaciones LA. Círculos GpgS de *Mycobacterium tuberculosis* (GT81): rojo (cristal 4DDZ), verde (cristal 4Y6N). Triángulos azules: MpgS *Rubrobacter xylanophilus* (GT78). Triángulos amarillos invertidos: GpgS de *Mycobacterium paratuberculosis* (GT81). (Los triángulos invertidos [contorno y opacos] hacen referencia a cristales con ligandos en su interior, los no invertidos son estructuras apo).

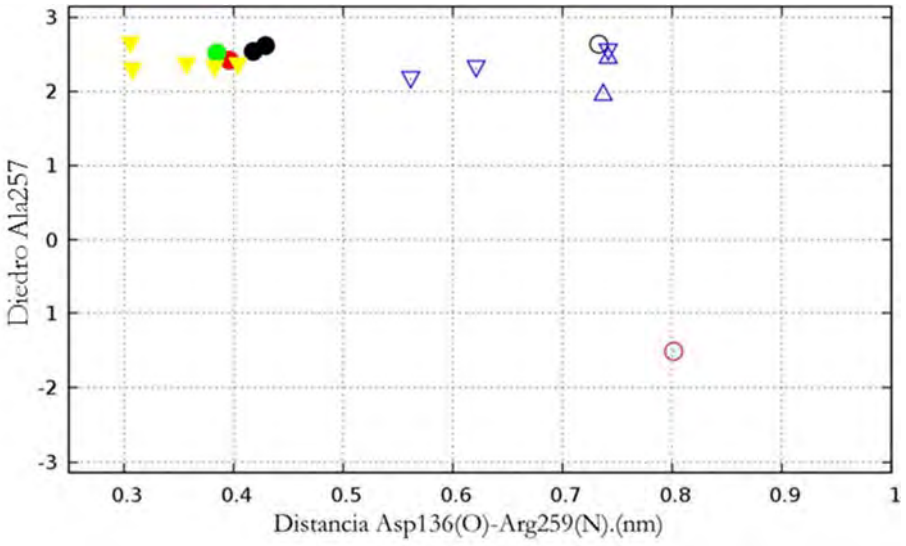


Figura 63. Diedro Ala257-Distancia. La distancia 0,5 parece delimitar las conformaciones LA/LI, el valor del diedro para ambas está en torno a 2 rad, menos para el cristal 4DDZ en conformación LI. Es interesante ver estructuras con ligandos que parecen estar en conformación LA (triángulos azules invertidos).

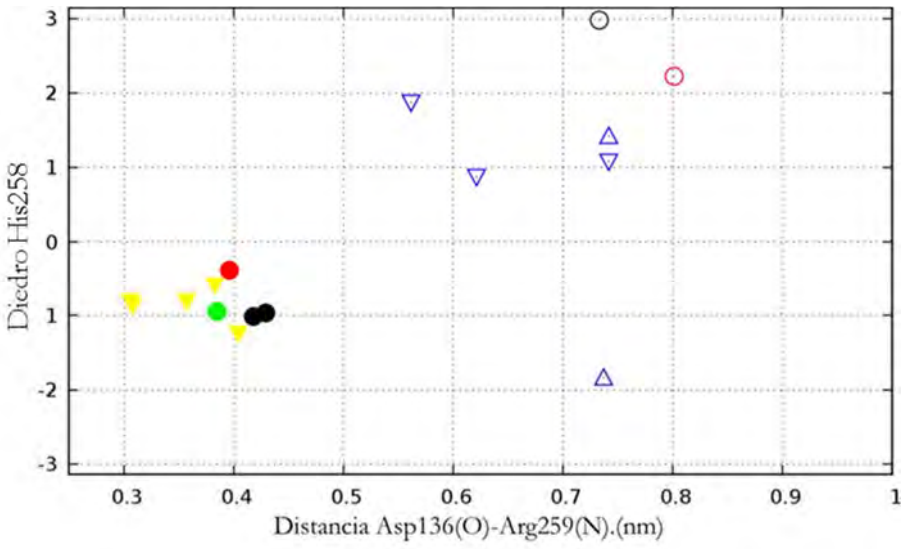


Figura 64. Diedro His258-Distancia. Las conformaciones LA se encuentran en un valor de diedro His258 de -1 rad mientras que las LI ocupan el espacio de 1 a 3, con un representante en -2 rads.

Se incluye a continuación la posición del diedro (C-CA-CB-CG) del residuo Arg256 para los cristales de 4DDZ, 4Y6N y proteínas homólogas. **Este diedro, como se verá en el apartado "Simulaciones con ligandos", parece cobrar relevancia en el movimiento del bucle y será también monitorizado en estas dinámicas apo.**

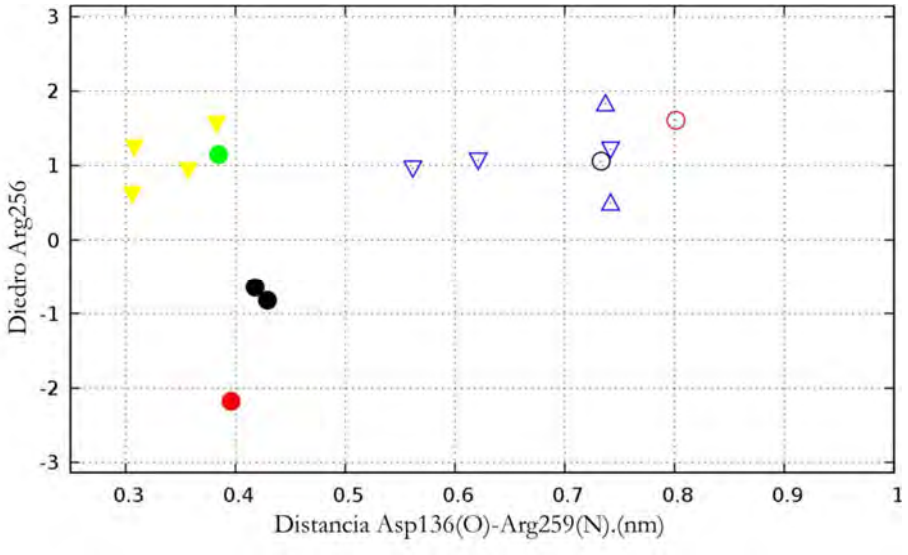


Figura 65. Diedro Arg256-Distancia. Las conformaciones LI se encuentran en un valor de diedro Arg256 de 1 rad mientras que las LA ocupan los diedros -1 y 1, con un representante en -2 rads. La conformación LI parece presentar un valor más estable de este diedro que la LA, al contrario que el diedro His258.

4. Simulaciones de modelos de GpgS en ausencia de ligandos

Se prepararon diferentes simulaciones de dinámica molecular clásica para distintas estructuras de la forma apo de GpgS. Por un lado, el modelado del cristal 4DDZ se separó en dos estructuras diferentes las conformaciones LA y LI, por el otro, del modelado y dinámica molecular del complejo ternario, cristal 4Y6N, (dinámica que se desarrollará en el punto 4.3.4), se obtuvieron 10 estructuras, todas con la conformación inicial LA, a las que se extrajeron los ligandos, obteniendo así otras 10 estructuras apo (ver métodos). Resultando un total de 12 estructuras de partida apo, 11 en conformación LA y 1 en conformación LI. Con cada una de estas se inició una simulación de dinámica molecular con la intención de estudiar la dinámica del bucle y de los residuos que lo forman. Hay que recordar que cada estructura de partida está compuesta de dos monómeros de GpgS, estando cada monómero en la misma conformación.

¿Serían estables cada una de las conformaciones? ¿Podrá observarse cambio conformacional en el bucle? ¿Se mantendrán los residuos en los parámetros estructurales encontrados en cristales y modelos?

4.1 Simulaciones de modelos basados en la estructura 4DDZ: MD4DDZ

Conformación inicial LA (MD4DDZApoLA): En el monómero A, durante el equilibrado, en el primer paso de minimización de energía los átomos Asp136(O)-Arg259 (NH) se separan. En el cristal, estos átomos parecen formar uno de los puentes de hidrógeno que estabilizan la forma LA, generando una lámina β con la hoja β4'. Aunque esta separación es pequeña –menor de 0,2 Å–, como los átomos Asp136(O)-His258(CAH) se encuentran más cerca, también en el cristal, este último sustituye a Arg259(NH), desestabilizando la lámina β y causando que el bucle se abra durante el equilibrado y tienda a cambiar rápidamente de conformación a LI, en los primeros ns de MD (figura 66 A). Además, con solo 6 ns de dinámica propiamente dicha, parte de la región variable del monómero B se aproxima al bucle abierto, formando con él una hoja β antiparalela, entre los residuos Ser171(O)-Asn260(NH) y Val174(NH)-His258(O), estable durante toda la MD y que mantiene al bucle en forma LI, pero al mismo tiempo le resta la flexibilidad que caracteriza a esta forma. El bucle en el monómero B se mantiene en la conformación inicial LA durante toda la dinámica (figura 66 B).

Conformación inicial LI (MD4DDZApoLI): Lo más destacable aquí es el rápido cambio del diedro ψ de la Ala257, desde el valor inicial del cristal, -1,5 rad a 0 rad, lo que sucede en ambos monómeros de la estructura en los primeros pasos de minimización de energía (flechas en la figura 69). Más tarde durante el proceso de calentamiento este valor vuelve a cambiar, de 0 a 2,5 rads. El

bucle del monómero A nunca llega a cambiar de conformación manteniéndose en la forma LI durante toda la dinámica (figura 68 A). El bucle del monómero B sin embargo parece cambiar de la forma LI a LA, tras 6 ns de ejecución manteniéndose así 325 ns, donde vuelve a cambiar a conformación LI (figura 68 B), sin embargo, esta forma LA, medida por la distancia, se debe más a la aproximación del motivo DXD que a un cambio conformacional del bucle.

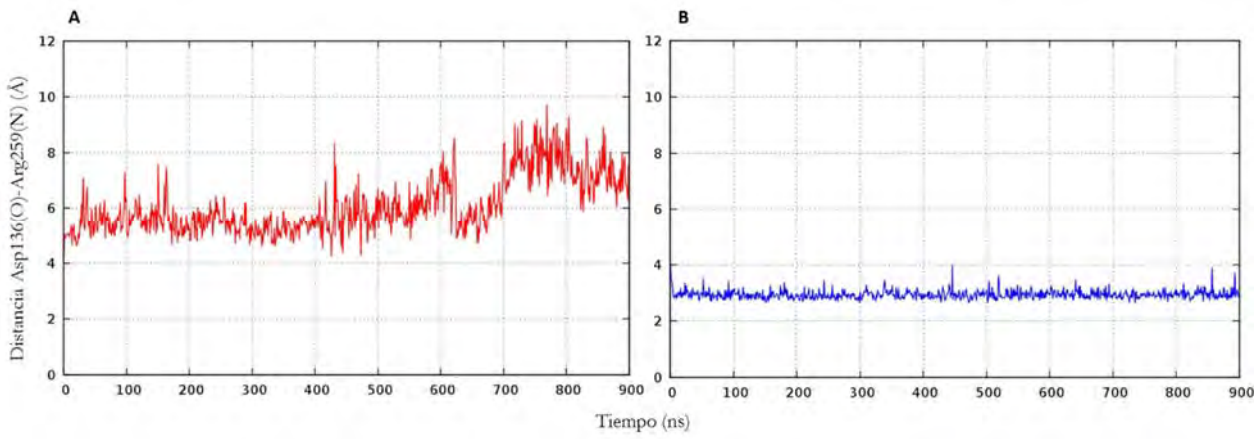


Figura 66. Conformación inicial LA: Medidas de la distancia entre el bucle y el motivo DXD. A, monómero A. B, monómero B.

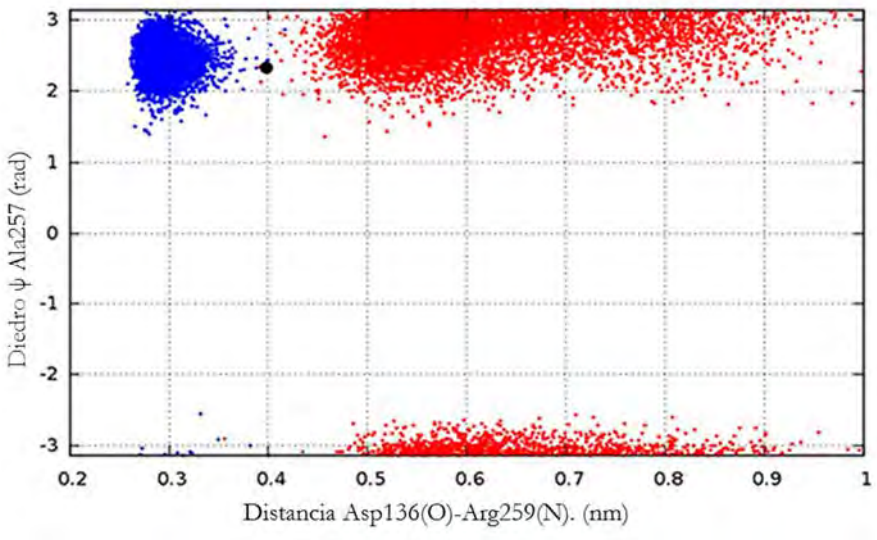


Figura 67 MD4DDZApoLA. Simulación iniciada a partir del modelo LA. Rojo: Monómero A. Azul: Monómero B. El círculo negro marca el valor de estas CVs en el cristal 4DDZ. El monómero A ha cambiado de LA a LI al inicio de la MD. El monómero B se mantiene estable en la conformación LA.

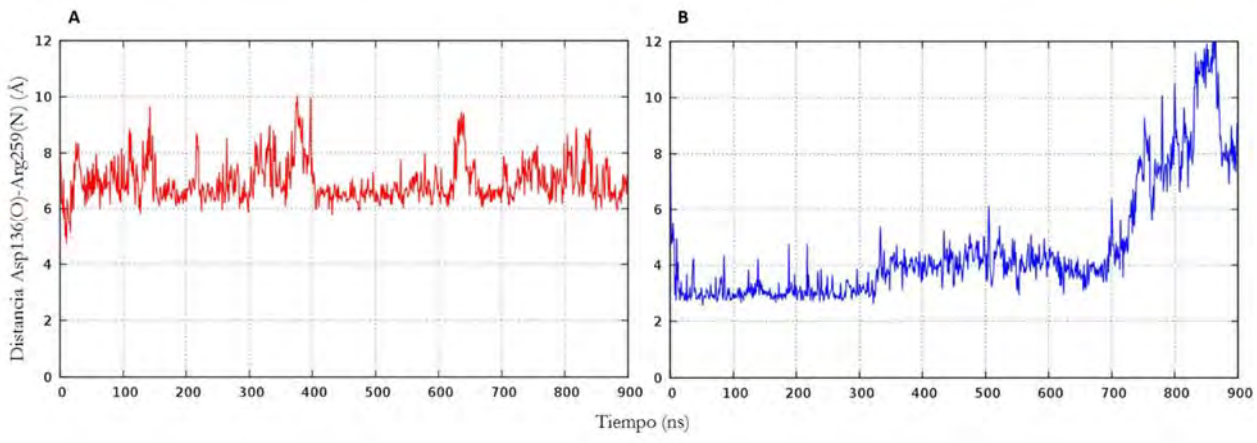


Figura 68. Conformación inicial LI: Medidas de la distancia entre el bucle y el motivo DXD. A, monómero A. B, monómero B.

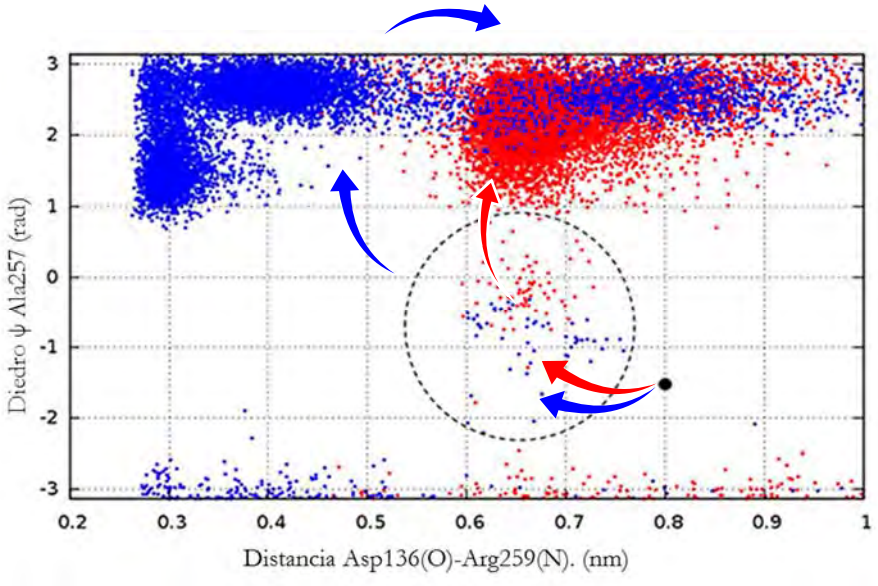


Figura 69. **MD4DDZApoLI. Simulación iniciada a partir del modelo LI**. Rojo: Monómero A. Azul: Monómero B. El círculo negro marca el valor de estas CVs en el cristal 4DDZ. **Ambos monómeros cambian la posición del diedro desde -1,5 a 0 durante el equilibrado para, durante el calentamiento, cambiar de nuevo hasta el valor estable entre 1 y 3 rad**. El monómero A ha permanecido toda la simulación en la conformación LI, sin embargo, el monómero B ha podido alcanzar valores que mimetizan la conformación LA, no tanto por el movimiento del bucle en sí, sino por la aproximación del motivo DXD.

Los cambios de conformación parecen por tanto posibles. Resultaba extraño el movimiento del diedro ψ de la Ala257 en la conformación LI, por ello se monitorizó de nuevo cada uno de los diedros estudiados para compararlos durante el proceso de equilibrado y dinámica (Anexo 12).

El puente de hidrógeno representado por Ile138-Ala257 se mostró bastante más inestable que Asp136-Arg259, en concordancia con lo observado en el grupo de proteínas homólogas. Cuando el bucle se mantiene estable en su forma LA, la distancia es de unos 3 Å, cercano a los valores de *M. paratuberculosis* mientras que la forma LI es bastante más flexible, moviéndose entre valores de 4 y 8 Å, en este caso próximo a los valores de *R. xylanophilus*.

Así pues, las simulaciones de GpgS apo partiendo de la estructura 4DDZ mostraron que ambas conformaciones del bucle (LA y LI) parecen alcanzables por las fluctuaciones térmicas, aunque el número de eventos de transición de una conformación a la otra es muy bajo. La distancia Asp136-Arg259 y los diedros de la cadena principal del cristal 4DDZ para cada conformación, se suavizan durante la dinámica y toman valores similares a los del resto de proteínas homólogas.

A continuación, se analizó la evolución de las conformaciones (definida por la distancia Asp136-Arg259) junto a la de los diedros de las cadenas laterales de Arg256 (cuya importancia se verá en el capítulo 4.3.4) y His258, para comprobar si existe algún tipo de correlación entre la orientación de las cadenas laterales de estos residuos y el cambio conformacional (figuras 70 y 71), en las simulaciones de GpgS apo partiendo del cristal 4DDZ.

Para la estructura de partida en conformación LA, el bucle del monómero A (Figura 70 A) se abre durante la fase de equilibrado y permanece abierto hasta el final de la simulación. El valor de la His258 cambia a 1 y se alterna con algunas visitas al valor π, no volviendo a visitar el estado -1. El diedro de la Arg256 cambia -1 y se mantiene en este valor con muy pocos cambios a 1 o π. En el monómero B (Figura 70 B) no se ha producido ningún cambio conformacional y el bucle permanece siempre en posición LA, el diedro His258 cambia a 1 a pesar de estar siempre en conformación LA (distancias menores de 5 Å) pero el diedro Arg256 permanece en 1 prácticamente durante toda la simulación.

En la estructura de partida en conformación LI, el bucle en el monómero A (Figura 71 A), per manece estable en conformación LI, con los diedros Arg256 y His258 en valor 1, valor al que cambia Arg256 al inicio de la simulación, con muy pocos estados en -1. En el monómero B (Figura 71 A) sí se han producido cambios en la distancia y el diedro. Aquí, estos cambios de distancia no obedecen tanto al cambio de conformación del diedro, como a un acercamiento del bucle DXD al bucle RAHRN. Los cambios de diedro no parecen tener una correlación clara con los estados del bucle.

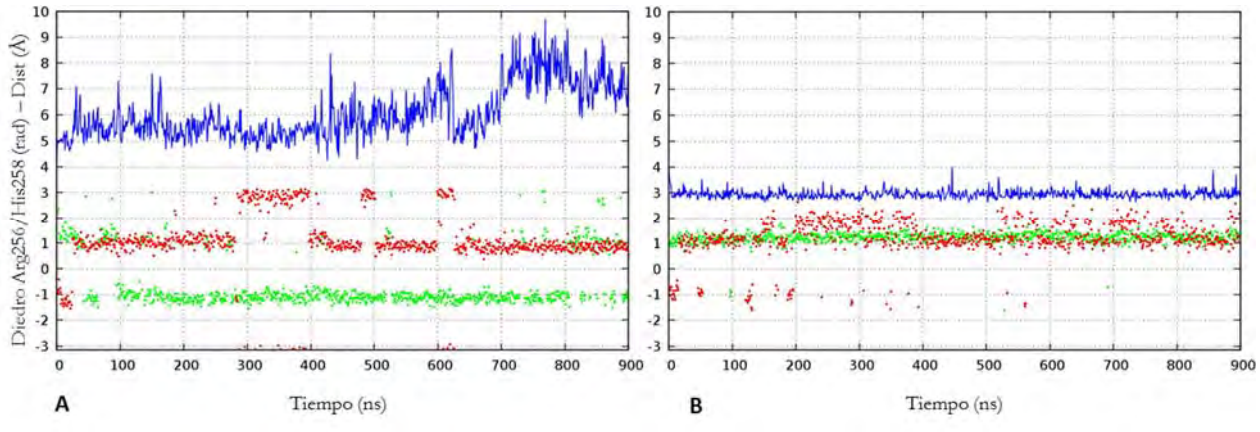


Figura 70. Simulación MD4DDZApoLA. Evolución de los parámetros estructurales del bucle: (i) distancia Asp136(O)-Arg259(N) (azul), (ii) diedro cadena lateral His258 (rojo) y (iii) diedro cadena lateral Arg256 (verde), durante la simulación de GpgSApo partiendo del cristal 4DDZ en conformación LA en el monómero A (izquierda) y monómero B (derecha).

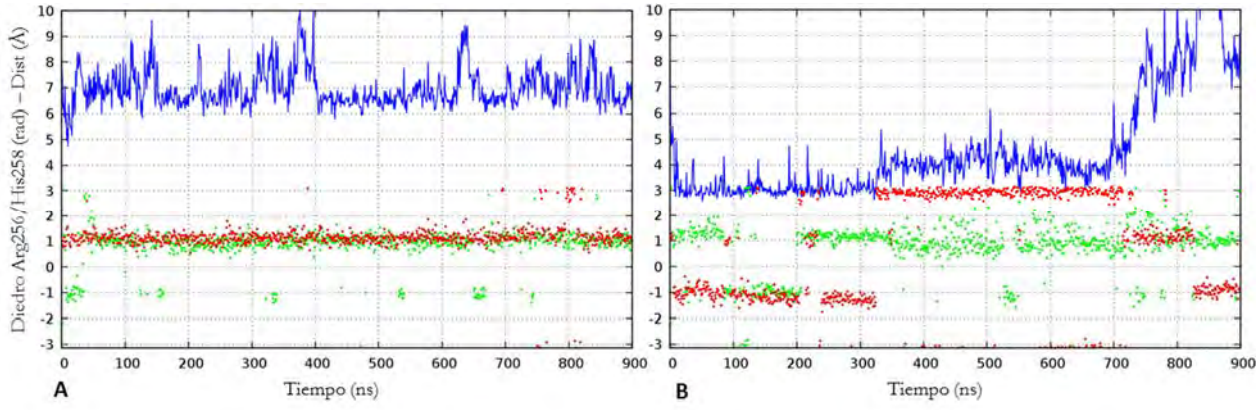


Figura 71. Simulación MD4DDZApoLI. (i) distancia Asp136(O)-Arg259(N) (azul), (ii) diedro cadena lateral His258 (rojo) y (iii) diedro cadena lateral Arg256 (verde)El monómero A se mantiene abierto durante toda la trayectoria (LI). El valor de la His258 cambia a 1 para no volver a visitar el estado -1. El diedro de la Arg256 se mantiene en 1 toda la simulación. En el monómero B, donde se han producido varios cambios conformacionales, His258(-1) es el valor visitado en conformación LA (distancias menores de 5 Å) pero el diedro Arg256 permanece en 1 prácticamente durante toda la simulación.

Puesto que todas las combinaciones parecían posibles no se pudo establecer aquí ninguna correlación, pero sí determinar que el cambio de valor en los diedros tenía un grado de ocurrencia muy bajo, menor en el residuo His258 que en la Arg256.

4.2 Simulaciones de modelos basados en la estructura 4Y6N: MD4Y6NApo

De una dinámica molecular clásica del complejo ternario (punto 3) se extrajeron 10 instantáneas de los 100 primeros ns de simulación, una cada 10 ns, a cuyas estructuras se les eliminaron los ligandos (ver métodos). Las 10 estructuras resultantes de GpgS apo disponían el bucle en conformación LA. Con cada una de ellas se inició una simulación de dinámica molecular (nombradas MD4Y6N-Apo1 a MD4Y6N-Apo10) con la intención de compararlas con las anteriores MD4DDZ-LA y MD4DDZ-LI y ver si los resultados partiendo de estructuras distintas eran equiparables. La simulación MD4Y6N-Apo1 se extendió hasta 1 μs mientras que las demás tuvieron una ejecución de 100 ns, cada una.

MD4Y6NApo1: El monómero A permanece siempre en conformación LA, pero el monómero B cambia a LI pocos ps después de iniciarse la simulación (figura 72). Tras 300 ns de MD el bucle vuelve a la conformación LA durante pocos ns, cambia rápidamente de LA a LI y de LI a LA de nuevo y permanece en LA 400 ns más para volver a cambiar a LI hasta el final de la simulación (Figura 72). Estos resultados confirman los observados en las simulaciones MD4DDZ, que el cambio espontáneo de LI a LA en la forma apo es posible, pero a una frecuencia menor que la de LA a LI, en la escala temporal del sub-microsegundo.

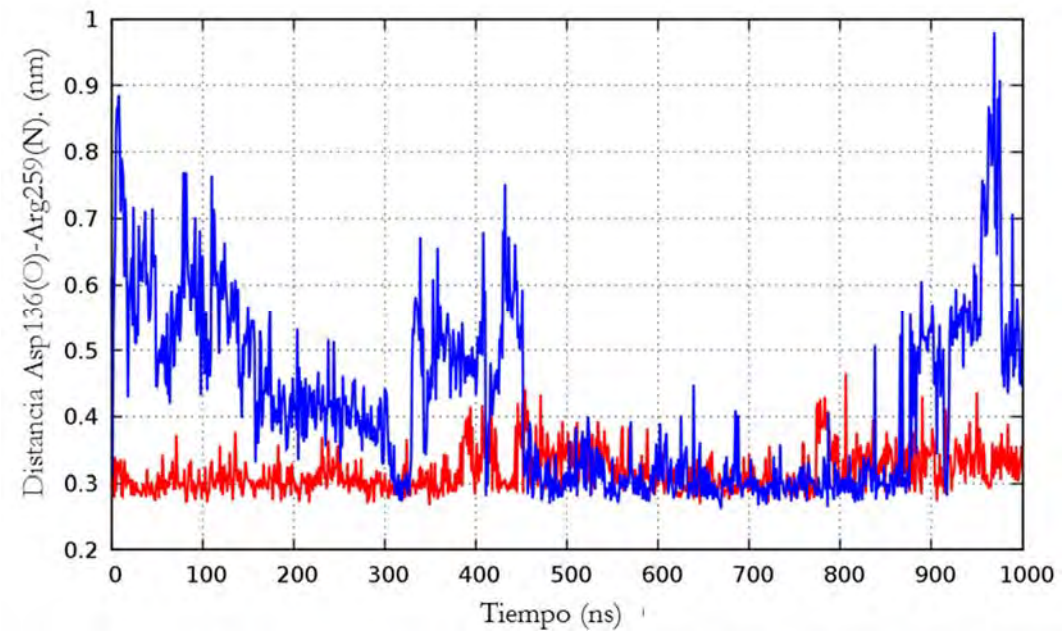


Figura 72. Simulación MD4Y6NApo1. Evolución de la distancia de puente de hidrógeno Asp136-Arg259 a lo largo de la simulación de GpgS apo partiendo de la estructura 4DDZ para el monómero A (rojo) y el monómero B (azul).

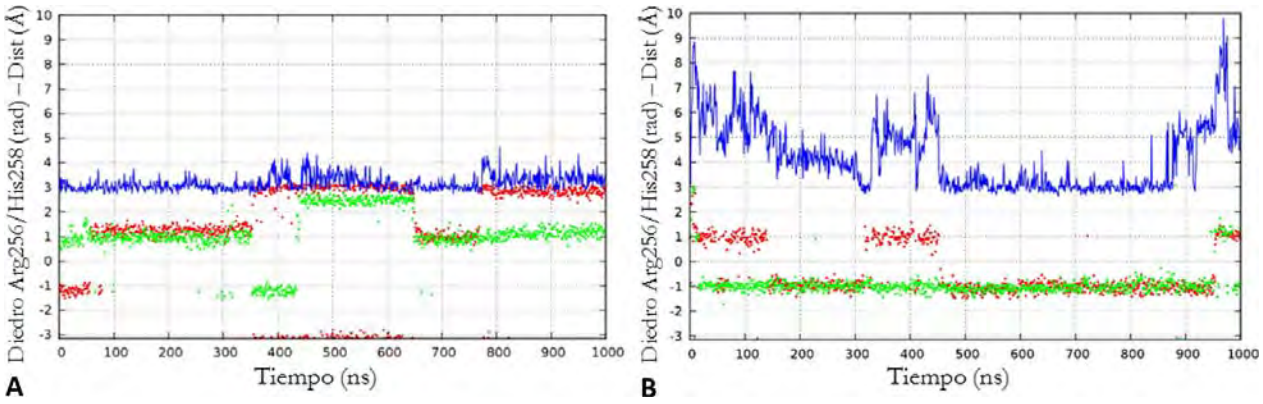


Figura 73. Simulación MD4Y6N-Apo1. **A**: Monómero A, **B**: Monómero B. Azul: Distancia del bucle RAHRN (Å), Verde: diedro Arg256, Rojo: diedro His258.

En cuanto a la posible correlación entre las torsiones de las cadenas laterales de Arg256 y His258 con la abertura y cierre del bucle para esta simulación MD4Y6NApo1 se observó lo siguiente (figura 73): en el monómero A, que no cambia de conformación, el diedro His258 se coloca en las posiciones -1, 1 y π rad con prevalencia por los valores 1 y π, mientras que la Arg256 presenta valores de -1, 1 y π con prevalencia sobre todo de 1. En el monómero B los valores visitados para ambos diedros son -1 y 1, estando ausente el valor π. Aquí el bucle cambia de conformación varias veces y en cada cambio se produce la correlación de conformación/diedros: **LA: Arg256(1) /His258(-1)**, con una zona gris entre la distancia Asp136(O)-Arg259(N) de 5 Å que parece delimitar la frontera entre las conformaciones LA-LI y donde el diedro His258 puede tomar ambas conformaciones; la Arg256 toma siempre el valor -1 (salvo unas decenas de ns al final de la simulación, en conformación LI). En todas las simulaciones el diedro ψ de A257 (no mostrado) permanece entre valores de 2 y 3 radianes en ambos monómeros, como ocurría en las simulaciones MD4DDZ.

MD4Y6NApo2-10: Estas simulaciones cortas (100 ns) ofrecen 18 oportunidades de estudiar el comportamiento del bucle, ya que cada monómero se estudia por separado. Considerando el cambio conformacional como la superación estable de 5 Å de la variable distancia Asp136-Arg256, el bucle cambió a la conformación LI en 8 trayectorias (4 por monómero), ya que todas comenzaban en LA y ninguna revertió a LA. Cuando la fluctuación del bucle se mantuvo por debajo de los 5 Å, la distancia tiende a estabilizarse en 3 Å, lo que ocurre en 5 simulaciones (2 para el monómero A y 3 en el monómero B). El resto de trayectorias no cambió su comportamiento para el bucle, manteniéndose estables en una distancia de 3Å o fluctuando por debajo de 5Å. Lo más interesante es una correlación sutil entre los diedros de los residuos Arg256 y His258 con las conformaciones del bucle: el diedro His258(-1) parece asociado a la conformación LA en 10 de las simulaciones y muestra mayor estabilidad combinado con el diedro Arg256(1), lo que ocurre en 7 simulaciones. Sin embargo, 3 de las ejecuciones muestran la combinación opuesta para la forma LA; por lo que no se puede asumir para la proteína GpgS apo que la orientación de las cadenas laterales de Arg256 y His258 influya directamente sobre los cambios de conformación del bucle de manera absoluta. El valor del diedro ψAla257 (se mantiene siempre en torno a 2-3 rads.

Figuras 74 y 75. Evolución temporal de las simulaciones MD4Y6NApo2 a 10. Azul: Distancia del bucle RAHRN (Å), Verde: diedro de la Arg256, Rojo: diedro His258. Los círculos negros: la relación conformación LA-diedros Arg256(1)/His258(-1) se cumple en todo o parte de la simulación.

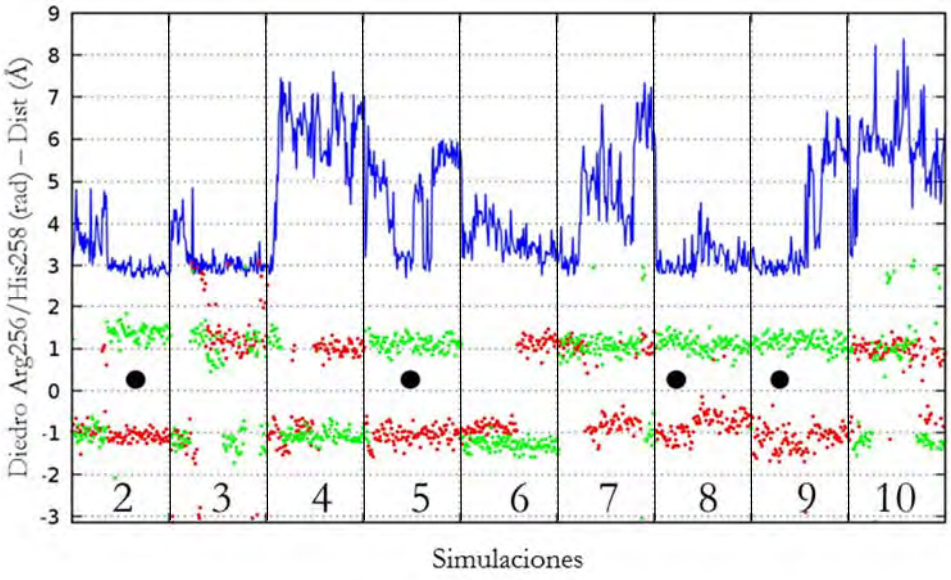


Figura 74. Monómero A En la forma apo el diedro Arg256 no parece estar tan relacionado con las conformaciones del bucle como His258.

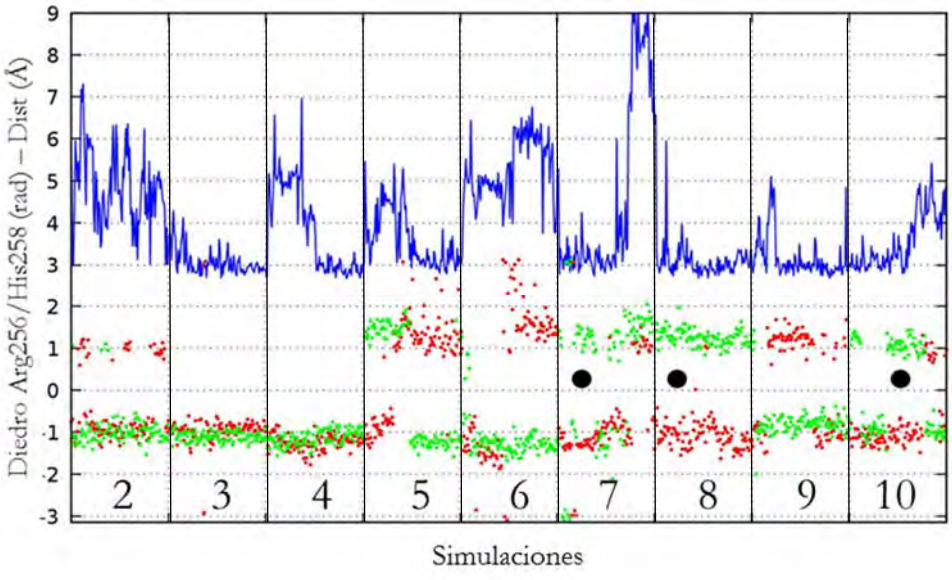


Figura 75. Monómero B. La correlación LA-His(-1)/Arg(1) parece menor que en el monómero A.

En conjunto estos resultados indican que ambas conformaciones LA y LI del bucle son alcanzables en GpgS en ausencia de ligandos, siendo más favorable la transición de LA a LI. Asociado al cambio conformacional del bucle, existe cierta variabilidad en las orientaciones de las cadenas laterales que forman parte de este, en particular de His258 y Arg256. Sin embargo, se aprecia en las simulaciones que el movimiento de estas cadenas laterales no correlaciona directamente con el cambio conformacional del bucle en la forma apo, pudiendo tratarse de movimientos lentos en comparación con la abertura y cierre del bucle.

4.3 Energías libres asociadas a cambios conformacionales de GpgS apo

Se ha visto como las conformaciones LA/LI son observables en las formas apo, sin embargo, LI parece más estable que LA. Llegados a este punto quisimos cuantificar la energía libre asociada a este cambio conformacional y estudiar también los diferentes estados que para el diedro Ala257 aparecían en las simulaciones MD4DDZ. Además, era interesante averiguar el papel que los residuos Arg256 y His258, junto con el mencionado Ala257 podrían tener en el cambio conformacional del bucle. Los cálculos de energía libre se llevaron a cabo mediante simulaciones de metadinámica. Para ello es necesario definir un conjunto de variables colectivas como coordenadas de reacción (ver fundamentos teóricos de la introducción, apartado "Metadynámica"). En este y siguientes apartados, de manera general se han utilizado como variables colectivas distintas combinaciones de los 4 parámetros estructurales definidos en el capítulo 3 (Figura 62). Las estructuras de partida para el estudio de la termodinámica del bucle en GpgS forma apo, fueron seleccionadas de las simulaciones de dinámica molecular descritas en el capítulo anterior en las que se partía de distintos cristales (MD4DDZLA y MD4Y6N).

MetaD4DDZ. Variables colectivas: (Distancia, Torsión Ala257)

Se realizó una primera simulación de metadinámica bidimensional, con la forma apo de la proteína GpgS provinente de la simulación MD4DDZApoLA.

Se utilizaron dos variables colectivas (CV): La **distancia** Asp136(O)-Arg259(N) y el **diedro ψ** de la cadena principal del residuo **Ala257**, para explorar el espacio conformacional y determinar el mapa energético. Con anterioridad se había visto que la combinación de estas dos variables representaba correctamente y de manera diferenciada las dos conformaciones del bucle LA y LI en los cristales de proteínas homólogas (figura 63) y las simulaciones MD de GpgS (figuras 67 y 69).

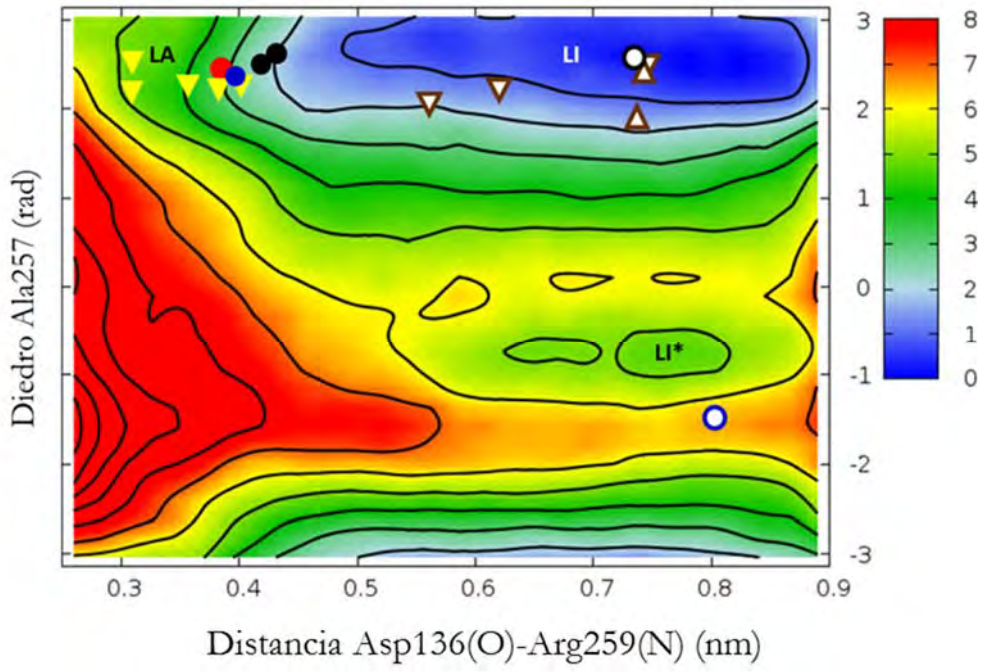


Figura 76. Las líneas de contorno delimitan áreas de energía de 1 kcal/mol ± 0,5 kcal/mol. Las conformaciones del bucle en estructuras homólogas a GpgS, se representan mediante símbolos vacíos para formas apo, y mediante símbolos rellenos para formas en complejo con ligandos (ver leyenda Figura 61).

El mapa de energía libre resultante (Figura 76) muestra que existe un pozo de energía principal, a un valor de distancia Ala136-Arg259 entre 0,5-0,9 nm y diedro Ala257ψ 2,5 rad, que lo sitúa en la zona de conformación LI. Con una considerable barrera de unas 6 kcal/mol, existe otro metaestado LI*, más o menos al mismo valor de distancia, pero con un diedro entre 0 y -1 rad. Este segundo metaestado es hacia el que tiende la simulación MD-4DDZ-LI que parte de la estructura de GpgS del cristal 4DDZ en conformación LI. Durante la minimización de energía y equilibrado de dicha simulación se mantiene en la zona LI* del que sale durante el calentamiento hacia la zona del metaestado LI. En la zona correspondiente a la conformación LA del bucle, no existe un mínimo local energético propiamente dicho, aunque con la misma energía que LI*. Ahora bien, al carecer de barrera energética que lo separe, esta zona del mapa energético es accesible, por pertenecer al mismo pozo cinético que LA.

Las conformaciones del bucle en estructuras de proteínas homólogas a GpgS en su forma apo, se sitúan en la zona del metaestado LI, zona más estable para esta conformación del bucle. Solo la conformación del bucle en la estructura 4DDZ de GpgS apo se sitúa en el metaestado LI*, de 1 kcal/mol de estabilidad y 6 kcal/mol de barrera con LI. Por otro lado, las conformaciones del bucle en estructuras de proteínas homólogas a GpgS en complejo con ligandos, se sitúan en un espacio en principio no estable de la superficie de energía libre, pero si accesible. Debe notarse que, como se demostrará más adelante, la energía libre asociada al cambio conformacional del bucle se verá alterada precisamente por la presencia de los ligandos en estas simulaciones.

MetaD4Y6NApo1. Variables colectivas: (Distancia, Torsión Ala257)

Con el mismo conjunto de variables colectivas, se calculó el mapa energético conformacional utilizando ahora una estructura de partida diferente de GpgS en forma apo, la obtenida en la simulación MD4Y6NApo1 (Sección "Simulaciones con ligandos").

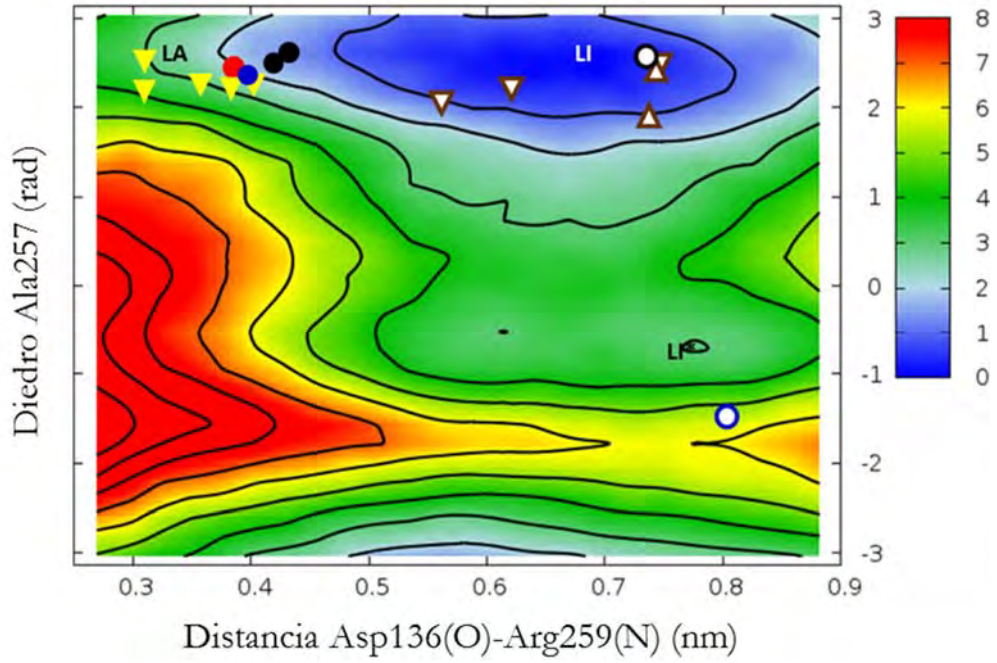


Figura 77. Isotermas: 1 kcal/mol ± 0,5 kcal/mol. Como en la simulación metadinámica anterior, las conformaciones LI de proteínas homólogas se sitúan en la zona del metaestado LI. El metaestado LI*, de 1 kcal/mol de estabilidad reduce la barrera con LI (3 kcal/mol) a la mitad. Los cristales en conformación LA, se sitúan en un espacio en principio no estable del FES, pero sí visitable para formas apo.

El nuevo mapa energético es muy similar y comparable al anterior (Figura 76). La configuración más estable está en la conformación LI, con mínimos similares a los de la metadinámica con el modelo 4DDZ, el principal de ellos a un valor de distancia Ala136-Arg259 entre 0,5-0,8 nm y de diedro Ala257ψ de 2,5 rad y otro mínimo (LI*) a un valor de distancia 0,6-0,8 nm pero con valor de diedro Ala257ψ -0,8. Tampoco se aprecia mínimo en la zona correspondiente a conformación LA, y su energía en esa conformación sigue siendo la misma que LI*, el segundo mínimo estable de la conformación LI. La diferencia energética entre LA-LI sería de entre 2-3 kcal/mol ± 0,5 kcal/mol a favor de LI, 1 kcal/mol menor que con el modelo 4DDZ.

MetaD4Y6NApo2. Variables colectivas: (Distancia, Torsión Ala257)

Se repitió la simulación metadinámica MetaD4Y6NApo1, con las mismas condiciones que la anterior, pero en este caso se delimitó el espacio conformacional, para que la exploración quedara restringida entre los diedros 1 y π (ver métodos), ya que el metaestado LI* de elevada energía no es alcanzable mediante fluctuaciones térmicas, según los resultados de las simulaciones de dinámica molecular clásica, acelerando así la resolución de la superficie de energía libre en esta metadinámica.

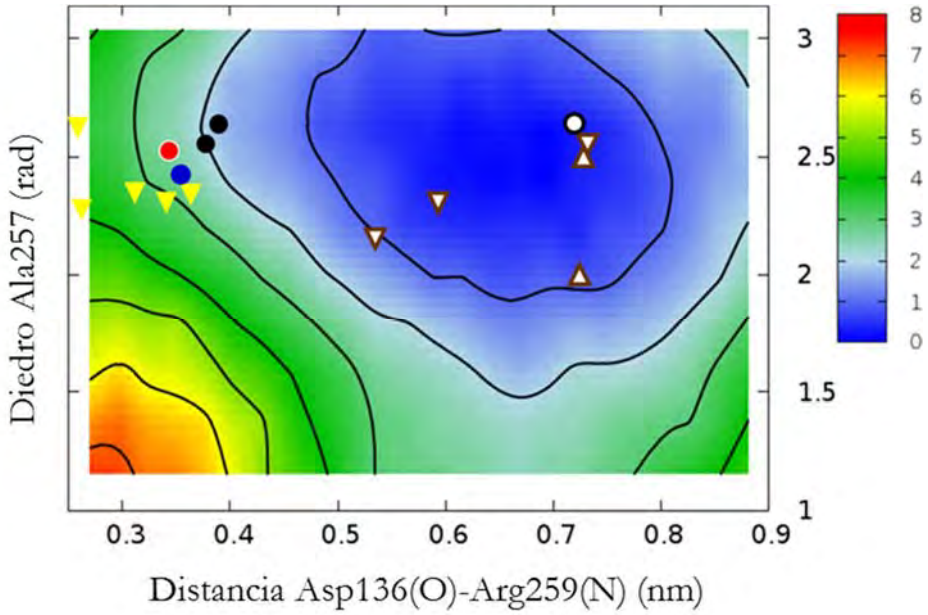


Figura 77. Isotermas: 1 kcal/mol ± 0,5 kcal/mol. Como en la simulación metadinámica anterior, las conformaciones LI de proteínas homólogas se sitúan en la zona del metaestado LI, zona más estable para esta conformación. Los cristales en conformación LA, se sitúan en un espacio en principio no estable del FES, pero sí visitable, unas 2-3 kcal/mol por encima del metaestado LI. Todas las estructuras LA tenderán hacia la conformación LI.

Los resultados de estas simulaciones metadinámicas para los cristales son equiparables. La conformación más estable es LI, unas 2-4 kcal/mol ± 0,5 kcal/mol más estable que la zona correspondiente a LA. Según estos estudios energéticos, la conformación LA no parece tener una gran estabilidad en la forma apo, ya que no se encuentra ningún pozo energético en ella, aunque forma parte del mismo pozo cinético que LI. De hecho, la rampa de 2-4 kcal/mol ± 0,5 kcal/mol entre el mínimo LI y la zona LA es, a la vista de las simulaciones de dinámica molecular, superable por las fluctuaciones térmicas, aunque poco frecuente y en general, en la forma apo las conformaciones LA tienden a cambiar y mantenerse en LI, conformación predominante para la proteína GpgS sin ligandos.

El hecho de que la conformación LA solo sea vista en proteínas que incluyen ligandos (salvo en la estructura 4DDZ) y LI mayoritariamente observada en proteínas apo, así como las poblaciones visitadas en las simulaciones de MD y la energía obtenida en las simulaciones metadinámicas, apoya entonces la hipótesis de que la conformación más probable para GpgS en forma apo es LI. Precisamente esa conformación está ausente en el cristal 4DDZ, donde las representadas son LA y LI*, dos conformaciones cuyos metaestados presentan igual energía. Así pues, las dos conformaciones de este cristal no corresponderían a un equilibrio conformacional, sino a haber atrapado dos metaestados de energía similar debido a las condiciones de cristalización. No se trataría de un artefacto, ya que las poblaciones existen y corresponden a mínimos de energía, pero no son las conformaciones más estables y sería necesario superar una gran barrera de energía para poder visitarlos (entre 4-6 kcal/mol ± 0,5 kcal/mol para LI*), por lo que raramente serían visitados en la forma apo de GpgS en disolución.

MetaD4Y6NApo3. Variables colectivas: (Distancia, Torsión His258)

De las simulaciones de dinámica molecular presentadas en la sección anterior se desprendió que la orientación de las cadenas laterales del bucle podía tener cierta influencia en su conformación. Se presenta aquí el cálculo del mapa energético asociado al cambio conformacional del bucle en función de las variables colectivas (i) distancia de puente de hidrógeno Asp136(O)-Arg259(N) y (ii) torsión del diedro His258. La estructura de partida sigue siendo la misma MD4Y6NApo correspondiente a la forma apo de GpgS.

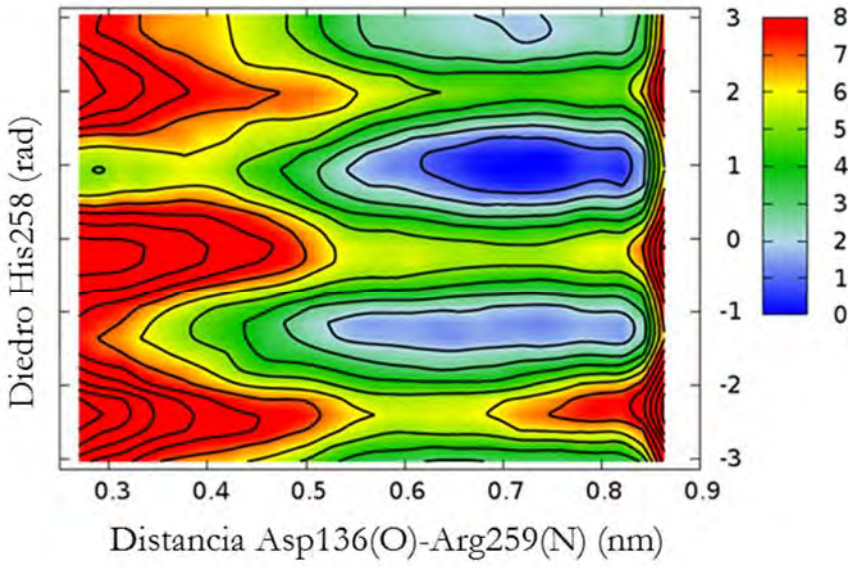


Figura 79. Representación energética de las CVs Distancia e His258 Isotermas: 1 kcal/mol ± 0,5 kcal/mol..

El mapa de energía libre resultante muestra que, en concordancia con las simulaciones de dinámica molecular: 1, -1 y π rads siguen siendo los valores más visitados del diedro de la cadena lateral His258, pero siempre en la conformación LI. La orientación del diedro de His258 en 1 rad es la configuración más estable (1 kcal/mol ± 0,5 kcal/mol más estable que los otros valores del diedro). No se ha conseguido desplazar el equilibrio hacia LA, aunque se aprecia un mínimo en esta conformación cuando la cadena lateral de la His258 está orientada con un valor de diedro de 1 rad.

Así pues, la activación del diedro His258 no altera el equilibrio conformacional de la forma apo de GpgS respecto los cálculos anteriores; la conformación más estable sigue siendo la forma LI, aunque es posible ver cierta estabilidad para la conformación LA.

MetaD4Y6NApo4. BIAS-Exchange. Variables colectivas: (Torsión Arg256, Torsión Ala257, Torsión His258)

Se completaron los estudios de las formas apo con una nueva metadinámica, implementando esta vez la técnica del BIAS-Exchange. Para comprobar si las conformaciones LA:LI eran alcanzables mediante la rotación de los diedros considerados ahora más interesantes, es decir cadena lateral de Arg256, cadena principal de Ala257 y cadena lateral de His258.

Siguiendo un esquema BIAS-EXCHANGE (ver fundamentos teóricos), se lanzaron 4 simulaciones de metadinámica que intercambiaron cada cierto tiempo sus estados conformacionales. Cada una tenía activada una CV y una de ellas era una simulación sin ninguna variable activada. Además, se monitorizó la distancia del bucle respecto al motivo DXD (mediante la distancia de puente de hidrógeno entre Asp136 y Arg259) para poder definir el estado del bucle en los distintos metaestados resultantes.

Como resultado de la simulación se obtiene un mapa energético en tres dimensiones (3 CV). Dada la dificultad en visualizar estos mapas energéticos en gráficos multidimensionales, se procedió a la identificación de metaestados y sus energías mediante la herramienta de análisis METAGUI (ver métodos). Los resultados se muestran en la tabla 12.

	Distancia	Diedro H258	Diedro A257	Diedro R256	Energía kcal/mol
M1	2,8	-1,3 **(-1)**	2,4	1,3 **(1)**	0
M2	2,8	2,8 **(π)**	2,6	0,9 **(1)**	0,5
M3	2,8	-1,1 **(-1)**	2,6	-1,1 **(-1)**	0,5
M4	6,4	1,1 **(1)**	2,6	-1,1 **(-1)**	0,7
M5	2,8	1,1 **(1)**	2,6	-1,4 **(-1)**	0,7

Tabla 12. Energía de los diferentes metaestados. Entre paréntesis, valores redondeados.

El análisis de los resultados de la simulación BIAS-EXCHANGE muestra la existencia de cinco metaestados para los tres diedros del bucle que han sido analizados (tabla 12). La forma más estable (M1) está localizada en los diedros Arg256 y His258 en los valores 1 rad y-1 rad respectivamente. Esta conformación del bucle corresponde a la forma LA puesto que la distancia ente Asp136 y Ag259 toma valores en torno a 2,8 Å en este metaestado M1. Por otro lado, el metaestado M4 corresponde con una conformación LI del bucle. La conformación LI tiene una diferencia mínima de energía con LA (diferencia de energía libre entre M4 y M1 de 0,7 kcal/mol), aunque con valores de diedros para las cadenas laterales de Arg259 y His258 diferentes entre sí. **Otros valores de diedro en las conformaciones LA:LI tienen energías similares por lo que podemos concluir en esta metadinámica que tanto una conformación del bucle como la otra tienen la misma estabilidad, con cierta preferencia por la conformación LA (Figura 80).** Si bien la simulación se ha realizado sobre la misma estructura de la proteína GpgS en forma apo, este resultado es distinto al obtenido anteriormente cuando la energía libre asociada al cambio conformacional se evaluaba sin tener en cuenta la rotación de las cadenas laterales de los residuos que forman el bucle.

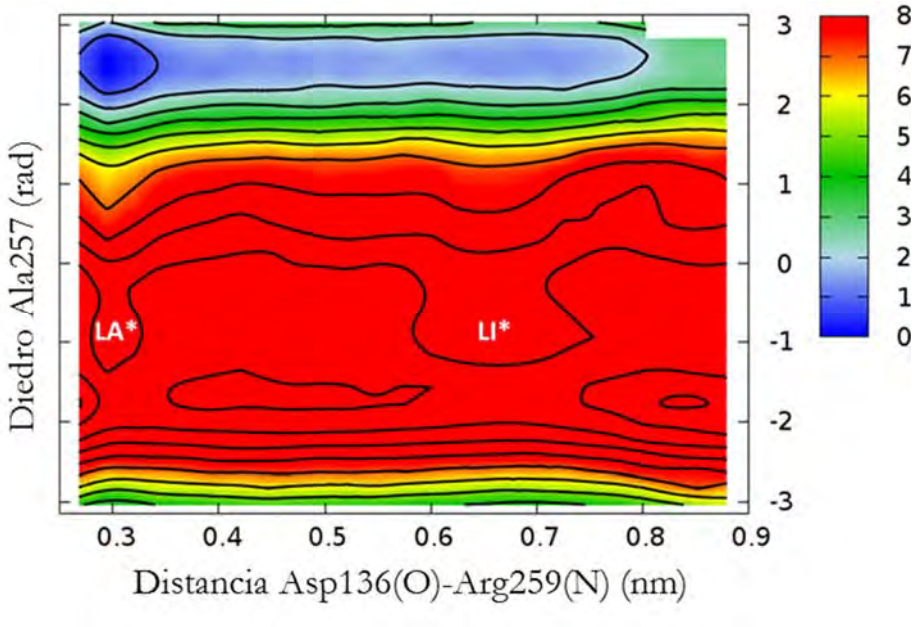


Figura 80. Proyección de la energía para el diedro de la Ala257 sobre la distancia. Isotermas: 1 kcal/mol ± 0,5 kcal/mol.**.**

El metaestado más estable es la conformación LA, con un valor de diedro ψ para Ala257 entre 2 y 3 rad. Comienza a observarse el metaestado LI*, aunque con una barrera muy alta y otro metaestado LA* no visto anteriormente, también con una gran barrera

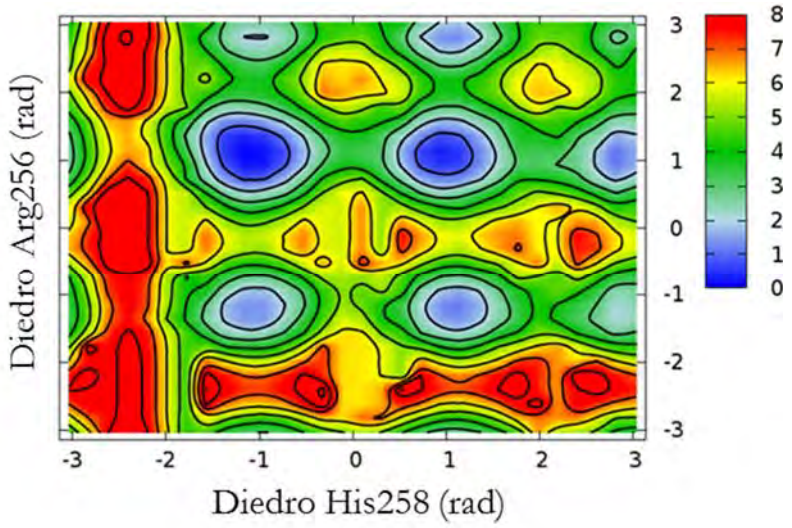


Figura 81. Representación energética de las CVs Arg256 e His258. Isotermas: 1 kcal/mol ± 0,5 kcal/mol..

Puede observarse que en el estado Arg256(-1), los diedros -1 y 1 de la His258 son menos accesibles pues poseen barreras mayores entre ellos que para el estado Arg256(1), donde también son más accesibles los estados de diedro π para las dos CVs.

Otra forma manera cuantitativa de comprobar cómo la activación de estos diedros promueve el cambio conformacional, es mediante el número de transiciones LA/LI observado en las simulaciones; mientras que en las anteriores simulaciones el cambio observado entre las conformaciones LA/LI ocurría, de media, cada 100 ns, en esta ocurre cada 10 ns, aún y sin activar la distancia de abertura y cierre del bucle. El cambio conformacional es pues, con la activación de los diedros, 1 orden de magnitud más accesible (datos no mostrados).

La activación de los diedros Arg256, Ala257, His258 desplaza el equilibrio conformacional hacia LA, aunque superando solo ligeramente la estabilidad de la conformación LI. Puesto que la activación del diedro Ala257 en las simulaciones MetaD4DDZLA y MetaD4Y6NApo no consigue alterar este equilibrio, siendo LI más estable siempre, la clave del cambio conformacional del bucle podría estar en la movilidad de las cadenas laterales de los residuos Arg256 o His258.

Nota: El siguiente paso lógico hubiese sido realizar una simulación metadinámica activando también el diedro de Arg256, sin embargo, la presentación de los resultados en este trabajo no sigue necesariamente el orden cronológico en el que fueron realizados y esta no se pudo ejecutar. Ideamos otra forma de hacerlo, una activación indirecta gracias al estudio de las estructuras con ligandos, que se explicará en el siguiente apartado.

5. Simulaciones del complejo ternario y modelos con ligandos

5.1 Simulaciones del complejo ternario de GpgS basadas en los modelos de 4Y6N

MD4Y6NUDPGlc·PGA

Con la estructura seleccionada del modelado de la proteína GpgS en complejo con UDP-Glc y PGA (basados en el cristal 4Y6N) se inició el proceso de ejecución de una dinámica molecular clásica. Para esta simulación el proceso de equilibrado fue más suave que en los anteriores (ver métodos) y se fijaron las distancias entre los átomos que interaccionan con el metal en el monómero B. Todo esto para intentar alterar lo menos posible la posición de los ligandos en el interior de centro activo, desde la posición inicial del cristal, ya que durante los primeros 80 ns de simulación se extrajeron estructuras para estudios QM/MM que sirvieron para esclarecer el mecanismo tipo S_Ni de la proteína GpgS[15].

Aquí, el objetivo fue el análisis del equilibrio conformacional del bucle catalítico de GpgS en presencia de ligandos. Se ejecutó una MD de 860 ns de duración y se estudió cada monómero por separado. A modo de resumen, en el monómero A la conformación LA fue estable durante toda la simulación mientras que para el monómero B a mitad de la simulación el bucle cambia a la conformación LI, permaneciendo así hasta el final de la simulación (Figura 82).

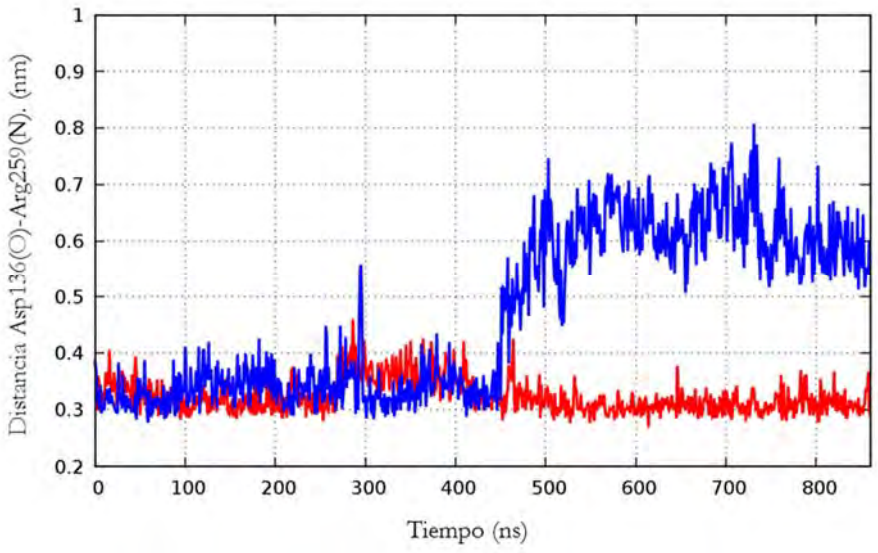


Figura 82. Evolución de la distancia del bucle catalítico al motivo DXD a lo largo de la simulación MD del complejo ternario (MD4YSN). **Monómero A**: Línea roja, siempre en conformación LA. **Monómero B**: Línea azul, el bucle cambia de conformación de LA a LI a los 450 ns. En este monómero B, la distancia de los átomos que interaccionan con el metal se ha fijado mediante *constrains*.

Monómero A: No se aplicó ningún tipo de restricción al movimiento de los residuos, durante el equilibrado ni en la MD. El ligando UDPGlc permanece en la misma conformación inicial; la His258 se aleja del metal, manteniendo estable una distancia de ≈ 4 Å entre su Nδ y el Mg durante toda la simulación. Este alejamiento del metal tiene lugar durante la fase de equilibrado donde la coordinación His258(Nδ)-Metal es sustituida por Asp136(O1)-Metal. Desde entonces y hasta el final de la simulación ambos oxígenos de la cadena lateral del residuo Asp136 formarán enlaces con el Mg. El resto de las interacciones con el metal se mantiene durante la MD: un oxígeno de cada fosfato del UDP, las mismas dos moléculas de agua desde el inicio y los oxígenos de Asp136, que suman los seis átomos que coordinan el metal, con una distancia entre 1,8-2 Å, algo más compacta que la distancia inicial (Tabla 13).

Átomos que interaccionan con el metal	Distancia en el cristal 4Y6N (Å)	Distancia media en la simulación MD MonA
H258(Nδ)	2,13	3,96
UDP(O3)	2,51	1,88
UDP(O6)	2,13	1,91
D136(O2)	2,26	1,95(O2)/1,98(O1)
H_2O(O)	1,98	2,03
H_2O(O)	2,67	2,03

Tabla 13. Distancias de interacción entre el metal (Mn) y los átomos coordinados en el cristal 4Y6N.

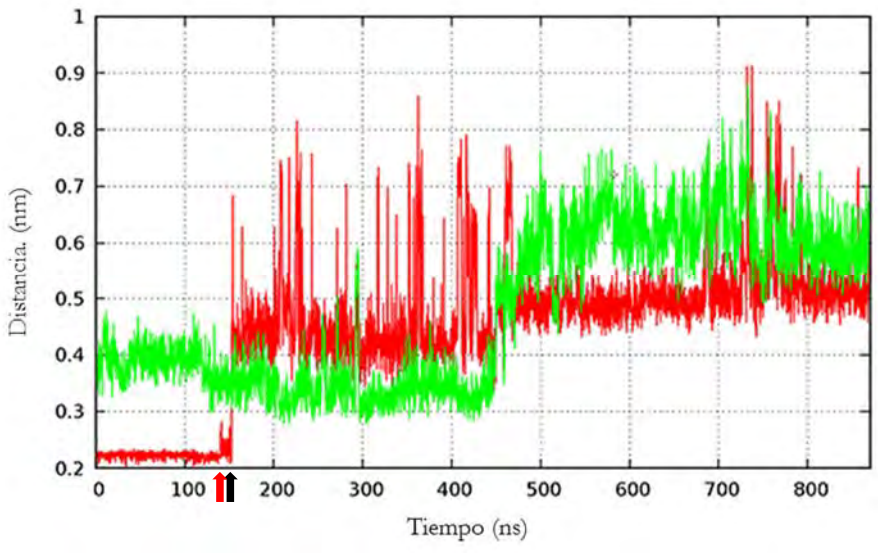


Figura 83. M**onómero B del complejo ternario (MD4Y6N)**. La distancia Asp136(O)-Arg259(N) (línea verde) cambia a los 450 ns, considerada como el cambio conformacional de LA a LI. La distancia His258-Metal (línea roja) está constreñida hasta los 140 ns y entonces liberada (flecha roja). A los 156 ns la His258 pierde la distancia de coordinación con el metal (flecha negra), y se produce el cambio conformacional del UDPGlc.

Monómero B: Las distancias de la tabla 13 se fijaron mediante *constrains* para mantener las del cristal/modelo (ver métodos). Tras 140 ns de simulación la *constrain* a la interacción His258(Nδ)-Metal se elimina (figura 83). Parecido a como ocurre en el monómero A, la interacción His258-Metal se pierde inmediatamente y es sustituida en este caso por el Asp134(OD1)-Metal, cuyas distancias con el metal pasan a ser 4,5 y 1,9 Å respectivamente. A los 450 ns de MD el bucle cambia de conformación, de LA a LI (la distancia Asp136-Arg256 pasa de 0,5 a 0,8 nm) y permanece así hasta el final de la simulación. Parece que el uso de estas constricciones ocasiona una desestabilización general de esta zona y la pérdida de la interacción His258-Metal. La distancia His258-Metal cambia rápidamente de 2,1 a 4,5 Å pocos ns después de liberar este residuo, sin embargo, el cambio conformacional posterior de LA a LI incrementa esta distancia en solo 0,5 Å (Figura 83), por lo que cierto movimiento de la cadena lateral de His258 puede neutralizar el aumento de distancia del cambio conformacional.

En cuanto a las torsiones del bucle y sus cadenas laterales, estas son más estables que en las simulaciones GpgS en forma apo anteriores. El diedro Ala257ψ permanece entre valores de 2 y 3 radianes en ambos monómeros. En el monómero B presenta alguna inestabilidad desde el momento en que cambia el bucle de conformación hasta estabilizarse en 3 rads. El diedro CG-CB-CA-C de la cadena lateral de la His258 mantiene el valor inicial del cristal, ≈ -1 rad hasta que sucede el cambio conformacional LA-LI (figura 84) que cambia a 1 rad. El diedro de la Arg256 se mantiene siempre en 1. Así pues, para estas simulaciones parece que el cambio de conformación del bucle LA a LI y de la cadena lateral de la His258 parecen sucesos acoplados.

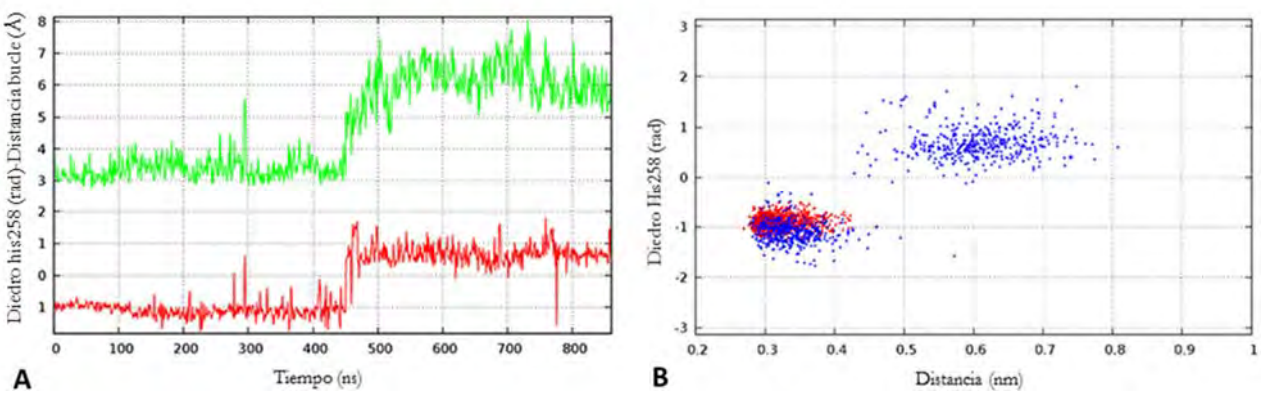


Figura 84. Simulación MD del complejo ternario de GpgS (MD4Y6NUDPGlc·PGA) A. Evolución de la torsión de la cadena lateral de la His258 (línea roja) y de la distancia Asp136(O)-Arg259(N) (línea verde) en el monómero B que sufre un cambio conformacional del bucle. **B**. Movimiento del bucle y de la cadena lateral de la His258 en el Monómero A (rojo) que no sufre cambio conformacional del bucle y monómero B (azul)

5.1.1. Cambio conformacional del UDPGlc:

La simulación anterior revela además un cambio conformacional importante en los sustratos, en particular del UDPGlc.

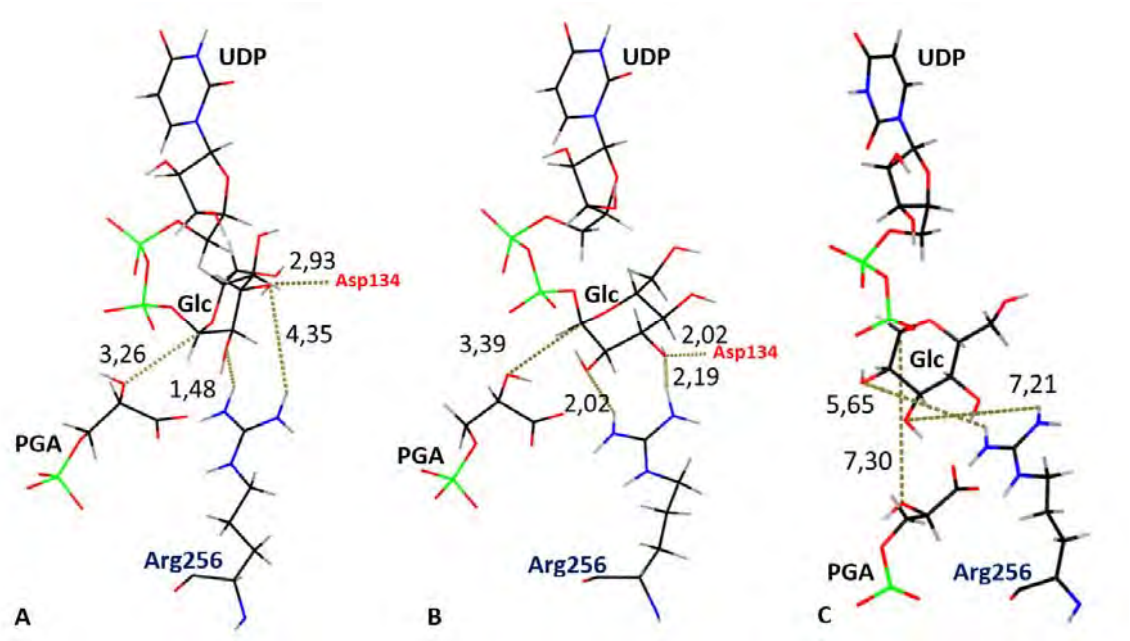


Figura 85. **A**. Ligandos UDPGlc y PGA en el cristal 4Y6N e interacciones entre ellos y los residuos de la proteína Asp134 y Arg256. **B**. Los mismos elementos en el monómero A tras la MD: El diedro de la Glc ha modificado su conformación estabilizándose con los nitrógenos de la Arg256 mientras se mantiene la interacción de la Glc con Asp134 existente en el cristal. **C**. Monómero B de la MD. Perdida la interacción de la Glc con el Asp134 y Arg256, el UDPGlc sufre un cambio conformacional que aleja el carbono anomérico de la Glc con el hidroxilo del PGA imposibilitando la reacción catalítica.

El UDPGlc, en el cristal 4Y6N, presenta el diedro Op4-P2-O1r-C1r (denominado como "c" en la Figura 86), que conecta la Glc con los fosfatos, a un valor de 2,3 rad y los elementos de interacción Glc(O2')-Arg256(NH2), Glc(O2')-Asp134(OD2), Glc(O3')-Asp134(OD2) y Glc(O3')-Lys114(NZ) estabilizan la configuración del UDPGlc (figura 85, A).

En la simulación MD4Y6NUDPGlc·PGA, para el monómero A, el átomo Glc(O2') solo forma puente de hidrógeno con Arg256(NH2); la interacción Glc O3'- Lys114(NZ) se pierde y en su lugar Glc(O3') forma un puente de hidrógeno con Arg256(NH1) (figura 85, B). Estos puentes de hidrógeno son estables durante toda la dinámica y modifican el valor del diedro "c" del UDPGlc a -2.93 rad (que por periodicidad del ángulo diedro es cercano al valor del cristal 2,34 rad) (figura 85, B).

Para el monómero B, la interacción con Asp134 es inestable desde el principio probablemente debido a las constricciones usadas en los aminoácidos del entorno y comienza a perderse a los 75 ns, momento en que el UDPGlc empieza a cambiar de conformación (diedro "c" = -2,93 rad) y se pierde completamente a los 150 ns, con el cambio definitivo del diedro "c" de la Glc a -1,5 rad (figura 82). Con este cambio además se pierden todas las interacciones iniciales de la Glc con el centro activo y la distancia con el PGA aumenta hasta hacer imposible la catálisis (figura 83, C). Puesto que el residuo Asp136, segundo aspartato del motivo DXD, tiene constreñida inicialmente su distancia con el metal en el monómero B, probablemente este hecho restrinja la flexibilidad del residuo Asp134 y sea el motivo de la pérdida de interacción inicial con la Glc.

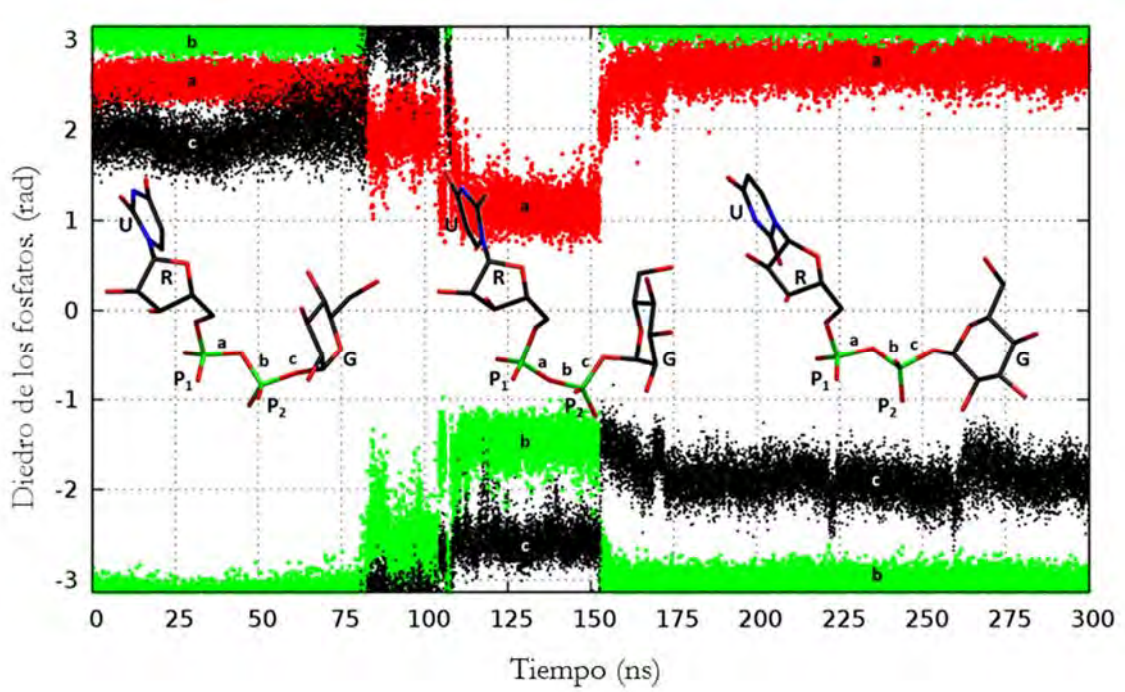


Figura 86. Movimiento del UDPGlc. U: Uracilo. R: Ribosa. P1 & P2: Fosfatos 1 y 2. G: Glucosa. Diedro a: Op1-P1-Op4-P2, b: P1-Op4-P2-O1r, c: Op4-P2-O1r-C1r. Los diedros a, b y c acumulan algún tipo de tensión desde el momento en que se pierde la interacción Glc-Asp134. A los 100 ns el UDPGlc toma la conformación estable del monómero A. A los 140 ns se libera la constricción de la His258 con el metal, cuya interacción es sustituida por la Arg234 y pocos ns después (a los 156 ns de simulación), la Glc cambia de conformación permaneciendo en esta hasta el final de la MD.

El UDPGlc es capaz de adoptar dos conformaciones diferentes dentro del centro activo de GpgS: una "doblada" que es la observable en el cristal y en el monómero A de la simulación MD4Y6N (aquella que no ha cambiado su conformación) y otra "extendida" en la que no es posible la transferencia del azúcar al PGA y observada en el monómero B de la misma simulación. La conformación "extendida" del UDPGlc puede darse tanto en la conformación LA del bucle como LI, sin poder determinar todavía la afinidad y estabilidad relativas.

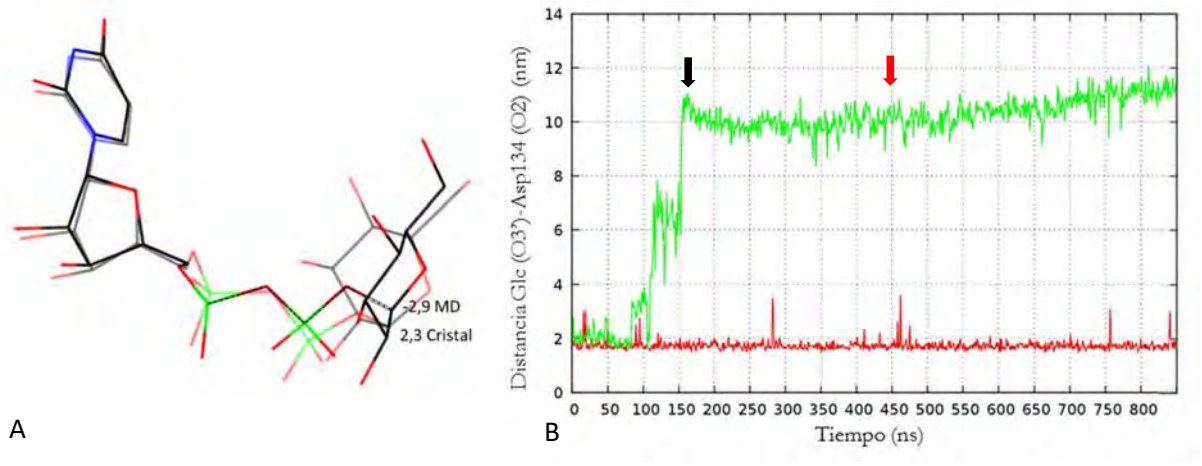


Figura 87. **A**. Superposición del ligando UDPGlc del cristal 4Y6N (sombreado) cuyo diedro Glc es 2,3 y de monómero A de la MD del complejo ternario cuyo diedro estable es -2,9. **B**. Distancia de interacción en los monómeros A (rojo) y B (verde) entre la Glc y el Asp134. En el monómero B esta distancia toma su mayor valor con el cambio conformacional del UDPGlc (flecha negra), el cambio conformacional del bucle (flecha roja) no afecta ni a la conformación del UDPGlc ni a la distancia entre Asp134 y Glc.

La interacción con el Asp134 por parte de la Glc parece importante para mantener la conformación "doblada" y catalítica del UDPGlc. En cuanto al bucle, las interacciones formadas entre la Glc y el residuo Arg256 durante la simulación para el monómero A (que no cambia de conformación), también contribuyen a la estabilidad de la conformación "doblada" del UDPGlc, sin embargo, no existe una gran diferencia entre los diedros 1 y -1 de la Arg, y en principio ambos pueden interaccionar con la Glc en posiciones similares. Así que el papel de la Arg256 para la estabilidad de la conformación catalítica del UDPGlc parece confirmado, aunque todavía no, la relevancia de las distintas orientaciones de su cadena.

5.2 Interacción de los ligandos con la proteína

Los resultados de las simulaciones apo reportaron que la conformación LA es visitable sin ligandos, así estos podrían estabilizar esta conformación del bucle en lugar de inducir a un cambio, sin embargo, una selección conformacional propiamente dicha sería harto ineficiente puesto que el equilibrio está, en la forma apo, muy desplazado hacia la conformación LI y sería muy infrecuente que los ligandos encontrasen la conformación activa. Se realizaron una serie de estudios de *docking* con los ligandos UDPGlc y PGA y diferentes estructuras de la proteína GpgS, provenientes de cristales e instantáneas de las simulaciones de dinámica molecular. De este modo podría comprobarse si existe mayor afinidad entre los ligandos y alguna de las conformaciones y no podría entonces descartarse una selección conformacional. Se utilizaron ligandos rígidos y flexibles, restringiendo el espacio de búsqueda a la zona del centro activo de ambos monómeros.

Los experimentos se hicieron sobre los modelos de los cristales 4DDZ (cada modelo en una sola conformación, LA o LI) y sobre el modelo para el cristal 4Y6N. De las simulaciones cortas MD4Y6NApo2, 3, 4 y 7 se extrajeron cuatro instantáneas a diferentes tiempos (70 ns, 90 ns, 90 ns y 90 ns respectivamente) que presentaban diferentes conformaciones del bucle según el monómero y sobre estas también se hizo *docking*. La proteína puede contener en cada experimento diferentes ligandos que están indicados en la tabla 14.

El resultado, en síntesis, es que el PGA se coloca en todos los casos, en forma y lugar similares al cristal 4Y6N (nunca en la posición del dador, como se ha visto en otros estudios[129]; el UDPGlc flexible también se posiciona en el centro activo, en posición similar al cristal, pero con la conformación "extendida" en todos los casos menos en aquel *docking* en el que se encontraba además posicionado el PGA. El UDPGlc rígido solo encuentra la posición exacta del cristal con la estructura 4Y6N, que es además el evento de *docking* de energía libre más baja.

Con el ligando UDPGlc flexible solo se consigue la posición y conformación del cristal original cuando está también presente el PGA (evento señalado en rojo en la tabla 14). La conformación rígida del ligando UDPGlc (del cristal 4Y6N), no tiene cabida en la conformación LA del cristal 4DDZ, por impedimentos estéricos entre la cadena lateral del residuo Arg261 y los fosfatos del UDPGlc; para mover esta cadena lateral también es necesario mover la del residuo Glu265. Se aprecia una menor afinidad del PGA por la proteína con el ligando UDPGlc en el interior (asterisco en la tabla 14), lo que podría indicar el orden de entrada de estos ligandos (primero el PGA, después UDPGlc), pero el mismo experimento sobre las estructuras de la simulación MD4Y6N1, no destaca diferencias de afinidad del PGA en presencia o ausencia del dador.

Tabla 14. Resultado de los diferentes experimentos de *docking*. Las estructuras corresponden al modelado de cada cristal o a instantáneas de las diferentes simulaciones de MD. En la columna "posición": "No", implica que no ha habido eventos de *docking* que superpongan con el ligando del cristal 4Y6N. "Sí", que el ligando sí superpone con el de 4Y6N, pero no en la misma conformación y "Cristal" que tanto la posición como la conformación coinciden con 4Y6N.

Ligand	Ligand restrictions	Structure	Loop conformation	Other ligands	*Docking* CA	Interaction energy (kcal/mol)
UDPGlc	Flexible	4DDZ	LA	Mg	No	-
				Mg, PGA	No	-
			LI	Mg	Yes	-10
				Mg, PGA	Yes	-10
UDPGlc	Rigid	4DDZ	LA	Mg	No	-
			LI	Mg	No	-
UDPGlc	Flexible	4Y6N	LA	Mg	Yes	-10
				Mg, PGA	**Crystal**	**-10**
UDPGlc	Rigid	4Y6N	LA	Mg, PGA	Crystal	-12
UDPGlc	Flexible	MD4Y6NApo				
		2 Mon A	LA	Mg	No	-10
		Mon B	LI	Mg	No	-10
		3 Mon A	LA	Mg	Yes	-8
		Mon B	LA	Mg	No	-8
		4 Mon A	LI	Mg	Yes	-9
		Mon B	LA	Mg	Yes	-9
		7 Mon A	LI	Mg	Yes	-8
		Mon B	LI	Mg	Yes	-8
PGA	Flexible	4DDZ	LA		Yes	-6
			LI		Yes	-6
		4Y6N	LA	Mg	Yes	-7
			LA	Mg, UDPGlc	Yes	-5*
		MD4Y6N				
		Mon A	LA	Mg	Crystal	-7
		Mon B	LI	Mg	Crystal	-7
		Mon A	LA	Mg, UDPGlc	Crystal	-7
		Mon B	LI	Mg, UDPGlc	Crystal	-7

No hay una diferencia significativa ni para el UDPGlc ni para el PGA entre las conformaciones LA o LI, por lo que se descarta un posible mecanismo de selección conformacional para la proteína GpgS. La conformación del UDPGlc en el cristal solo pudo conseguirse manteniendo rígida la molécula o incluyendo en la proteína el PGA. Con el ligando UDPGlc flexible, la ausencia del dador tiende a colocar la Glc en el espacio que ocupa el PGA, en conformación extendida, la presencia de PGA parece evitar que esto ocurra y ayuda a mantener el diedro de la Glc en una posición catalítica. Esto sugiere que quizás el aceptor PGA se ubica en la proteína antes de la llegada del UDPGlc, favoreciendo la conformación catalítica del dador. Los residuos Arg261 y Glu265 pueden impedir la unión del UDPGlc a la conformación LA, ya que sus cadenas laterales pueden ser impedimentos estéricos para los fosfatos del ligando.

Otra forma de estimar las diferencias de energía de unión entre el UDPGlc y GpgS es teniendo en cuenta la flexibilidad de la proteína y monitorizando la conformación del UDPGlc y del bucle RAHRN, evaluando la puntuación del *docking* para cada paso de la trayectoria obtenida durante la simulación del complejo ternario. En esta simulación el monómero A permaneció en conformación LA y el UDPGlc en el diedro -2,9 que se considera estable; el monómero B cambió a conformación

LI, así como el diedro de la Glc, de doblada a extendida, por lo que se midió la energía de interacción a cada paso de cada monómero por separado (Figura 88).

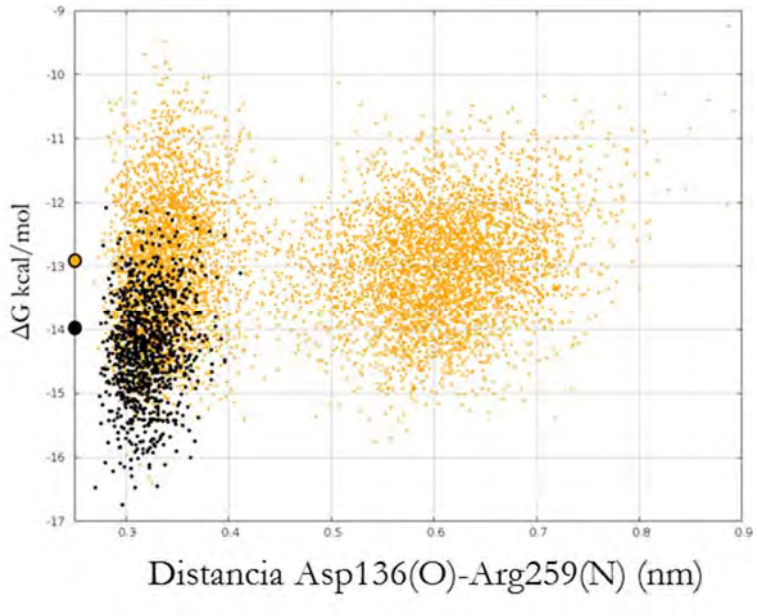


Figura 88. Energía de afinidad entre UDPGlc y GpgS en el monómero B. Punteado negro: primeros 150 ns, misma conformación que el cristal 4Y6N. Punteado naranja: el UDPGlc pasa a conformación "extendida" y no vuelve a recuperar la conformación de partida. A los 450 ns el bucle cambia de conformación a LI sin cambio en la energía de afinidad para el ligando UDPGlc.

En el monómero B, los primeros 150 ns, allí donde el UDPGlc mantiene la conformación "doblada" ya sea en la conformación del cristal (valor del diedro 2,3) o la estable en MD (valor del diedro -2,93), la afinidad de unión es de media 1 kcal/mol mayor que para la conformación extendida de la Glc, para la que es indiferente la conformación del bucle RAHRN ya que ambas rinden la misma energía de interacción. El monómero A, que mantiene el bucle cerrado y la conformación "doblada" del UDPGlc tiene una energía similar a la del monómero B con la Glc "doblada".

El mismo estudio se realizó con el ligando PGA sin que hubiera diferencias de afinidad con la proteína para las diferentes conformaciones del bucle.

Por todo lo expuesto, se descarta un mecanismo puro de selección conformacional de los ligandos para las conformaciones LA/LI de la proteína GpgS. Dador y aceptor presentan la misma energía de afinidad por la proteína en cualquier conformación del bucle. Sin embargo, la conformación del dador sí es importante para la conformación LA del bucle, ya que la forma "doblada" del UDPGlc presenta una afinidad mayor por el CA de la proteína, que la forma "extendida".

5.3 Energías libres de los cambios conformacionales en presencia de ligandos (I). CV: Diedro Ala257-Distancia

Las simulaciones de dinámica molecular sugieren que el cambio conformacional del bucle catalítico de GpgS en una estructura sin ligandos es posible de forma espontánea, si bien parece que no de forma equitativa, siendo más probable el cambio de LA a LI que el paso contrario. El cambio conformacional del bucle parece estar influenciado por la movilidad de las cadenas laterales de los residuos que forman el bucle. Estas observaciones también han sido confirmadas al evaluar las energías libres asociadas al cambio conformacional por metadinámica. Por otro lado, simulaciones equivalentes de dinámica molecular del complejo ternario de GpgS, indican que la presencia de los ligandos parece bloquear los diedros Arg256/His258 en un determinado valor. El diedro His258 parece también alterarse por el cambio conformacional y estar afectado por la presencia o no de ligandos en la proteína. Del mismo modo, la activación de los diedros Arg256, Ala257 y His258 podía alterar el equilibrio conformacional hacia la forma LA. Ahora se repetirán algunas de las simulaciones metadinámicas realizadas para las formas apo, pero en presencia de diferentes ligandos y se compararán los resultados.

Todas las simulaciones tienen activas las CVs: Distancia de puente de hidrógeno entre Asp136 yArg259 y la Torsión de la cadena lateral del residuo Ala257 (con el espacio conformacional restringido a valores de diedro entre 1 y π).

MetaD4Y6NUDPGlc·PGA: En esta simulación la estructura modelada del cristal 4Y6N, con todos sus ligandos en el interior (UDPGlc y PGA) es sometida a una MetaD en las mismas condiciones que las anteriores. (ver métodos)

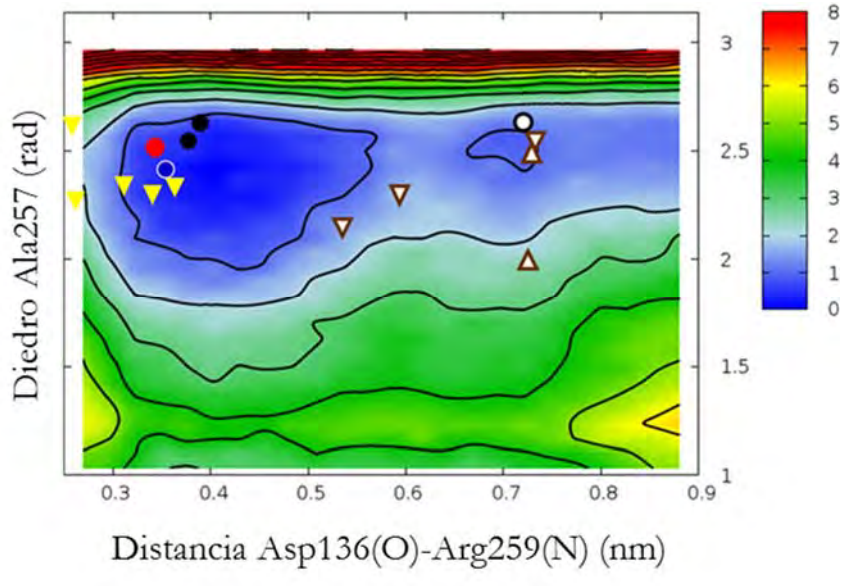


Figura 89. Superficie de Energía Libre asociada al cambio conformacional del bucle en GpgS en forma complejo ternario. Isotermas 1 kcal/mol. Error ± 0,5 kcal/mol.

La superficie de energía libre muestra la presencia de dos mínimos de energía. El más estable, a una distancia Asp136-Arg259 de 0,4 nm y diedro Ala257ψ de 2,4 rad, corresponde con una conformación LA del bucle catalítico. En este metaestado, las orientaciones de las cadenas laterales de los residuos del bucle se conservan como en el cristal: diedro His258(-1 rad) y diedro Arg256(1 rad). El segundo metaestado está situado ligeramente por encima en energía libre (0,5 kcal/mol) y corresponde a una conformación LI del bucle (distanciaAsp136-Arg259 de 0,7 nm y diedro Ala256 2,6 rad). Así pues, se observa en este caso una inversión del equilibrio conformacional donde la forma LA es la más estable con una barrera de 0,5-1 kcal/mol por lo que en realidad ambas conformaciones son prácticamente equiprobables. Se puede observar que los cristales con ligando (formas opacas en la figura 89) se agolpan en el metaestado más estable, conformación LA, aunque algunos (triángulos invertidos) se encuentran entre esta y el metaestado LI. Los cristales apo presentan los valores de distancia y diedro en el metaestado LI (salvo el cristal 3F1Y, cadena A cuyo diedro Ala256 es inferior al resto).

La presencia de todos los ligandos (UDPGlc y PGA) consigue alterar el equilibrio conformacional haciendo que la conformación LA del bucle sea la más estable y reduciendo la barrera entre ambas. El mínimo LA contiene estructuras del bucle con parámetros estructurales similares a los del cristal (torsiones de la cadena principal y laterales de los residuos).

MetaD4Y6NUDPGlcFree: Ligandos en el interior, UDP-Glc y metal (sin PGA).

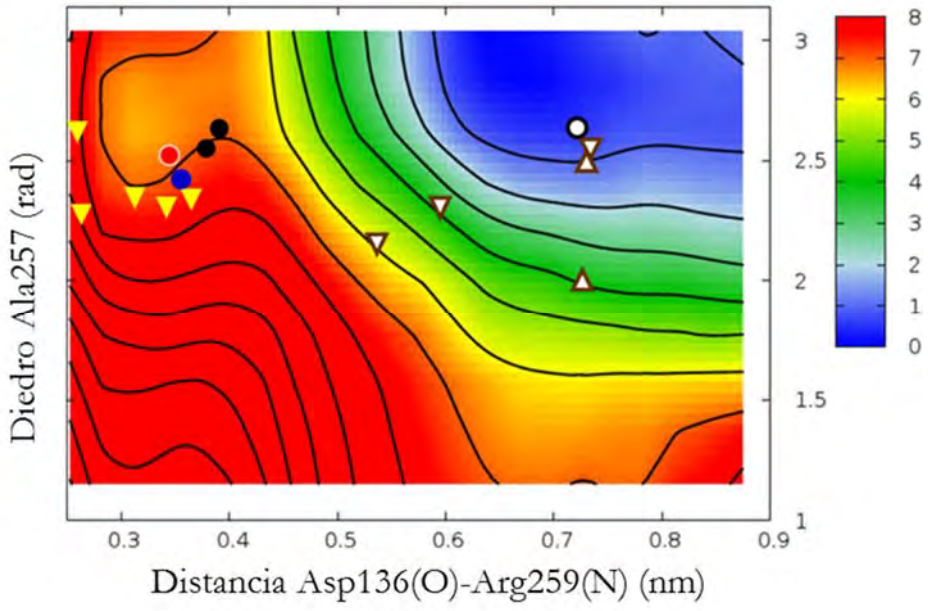


Figura 90. Superficie de Energía Libre asociada al cambio conformacional del bucle en GpgS en forma complejo binario UDP-Glc + metal. Isotermas 1 kcal/mol. Error ± 0,5 kcal/mol.

LI sigue siendo el mínimo más estable, aquí entre 0,6-0,9 y diedro 2,8. En esta metadinámica sí es posible observar un metaestado en LA, si bien su barrera es muy pequeña de alrededor de 0,5 kcal/mol. La diferencia energética LA-LI es en esta ocasión de 6,5 kcal/mol, 1,5 kcal/mol más que en la forma apo.

El ligando UDPGlc adquiere la conformación extendida (diedro -1,5 rad) al poco de iniciar la simulación, entonces la Arg256 pierde la interacción con la Glc y cambia su diedro de 1 a -1. No hay inversión de equilibrio conformacional en el bucle.

MetaD4Y6NUDPGlcFix: En el interior UDP-Glc y metal (NO PGA). **Diedro UDPGlc fijo**.

En la simulación del complejo ternario se pudo observar que diferentes conformaciones del UDPGlc eran posibles dentro del centro activo, la catalítica o "doblada" que interacciona con la Arg256 y la no catalítica o "extendida" que pierde esa interacción. Como por las simulaciones apo se creyó que en el diedro de la Arg256 podría estar la clave del cambio conformacional, y a fin de evitar la pérdida de interacción con el UDPGlc, en esta simulación se fijó el diedro del UDPGlc en el valor estable de la MD y conformación "doblada", -2,9 rad (ver métodos).

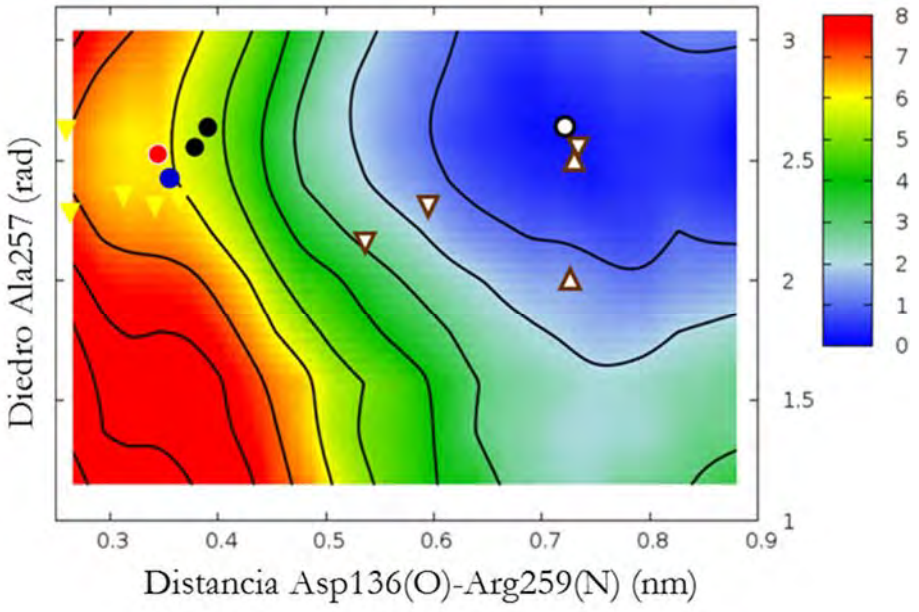


Figura 91. **Superficie de Energía Libre asociada al cambio conformacional del bucle en GpgS en forma complejo binario UDPGlc (diedro fijado) + metal.** Isotermas 1 kcal/mol. Error ± 0,5 kcal/mol.

LI sigue siendo el mínimo más estable, aquí entre 0,6 y 0,9, con diedro 2,8. El metaestado en LA no es tan claro como en la anterior simulación, pero sí se aprecia un codo de energía estable en él. No se observa barrera entre LA-LI y la diferencia energética sigue siendo de 6,5 kcal/mol.

No se encuentran diferencias entre las conformaciones de la Glc, doblada o extendida y su relación con las formas LA-LI.

MetaD4Y6NPGA: Ligandos en el interior **PGA** (NO UDPGlc ni metal).

Aquí se retiran de la estructura UDPGlc y metal, dejando solo la molécula aceptora PGA.

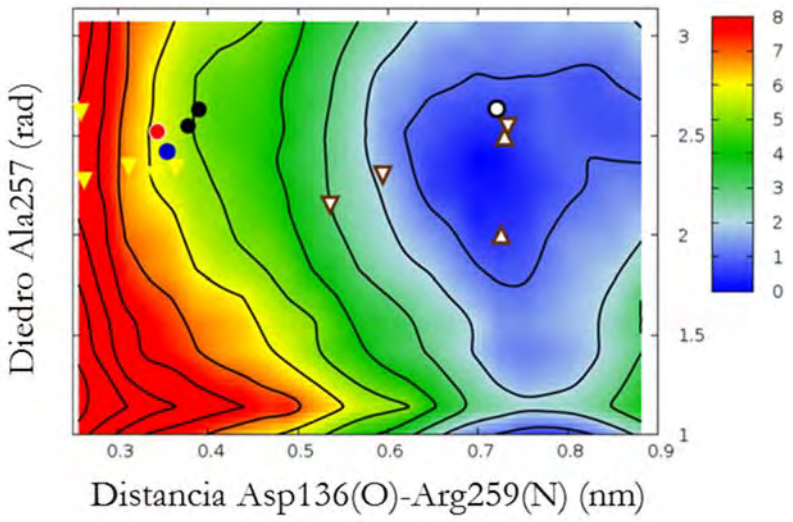


Figura 92. Superficie de Energía Libre asociada al cambio conformacional del bucle en GpgS en forma complejo unario PGA. Isotermas 1 kcal/mol. Error ± 0,5 kcal/mol.

Como hasta ahora, el mínimo sigue estando en la conformación LI a una distancia de 0,65-0,9. Sin embargo el diedro Ala257ψ parece haber cambiado un poco, bajando su valor a alrededor de 2,2. El mínimo LA parece haber desaparecido si bien la diferencia energética LA-LI es similar a las dos metadinámicas anteriores, 6,5 kcal/mol.

El resultado de todas estas metadinámicas (complejo ternario y complejos binario y apo por separado) confirma que, en la forma apo, el equilibrio conformacional está desplazado hacia la conformación LI y que ésta será la forma más probable para la proteína sin ligandos. También indica que la presencia de los distintos ligandos, por separado, no altera este equilibrio por lo que habría de descartarse un mecanismo de ajuste inducido por alguno de ellos. Sin embargo, el resultado de la MetaD4Y6NUDPGlc·PGA (metadinámica del complejo ternario) muestra una inversión del equilibrio conformacional del bucle de LI a LA, cuya diferencia con la forma apo es precisamente la presencia de todos los ligandos. Ello indica claramente que la unión de UDPGlc y PGA al centro activo de GpgS tienen un efecto directo sobre las conformaciones del bucle.

Ahora bien, sería necesario tener en cuenta además el orden de llegada y el camino de entrada que pueden seguir los ligandos en función de la conformación del bucle, o bien todos a la vez o bien uno detrás de otro, aunque en este último caso, solo cuando el segundo se posicione ocurrirá el cambio conformacional. Estas últimas consideraciones se han desarrollado en una nueva serie de cálculos de energía libre por metadinámica recogidos en la siguiente sección.

5.4 Energías libres de cambios conformacionales en presencia de ligandos (II): Otras CVs.

La monitorización del PGA en la simulación MD4Y6N mostró que, aunque son fundamentalmente los residuos de la RV los que contribuyen a la estabilidad de esta molécula en el CA, los residuos del bucle RAHRN: His258 y Arg260 también lo hacen, sobre todo en la conformación LA; en la conformación LI se mantiene presente la interacción con His258 y es mucho menor la de la Arg260, aunque no se puede descartar que sea debido al movimiento del PGA en esta conformación. En un entorno dinámico esta molécula soluble tiene que acceder al centro activo desde el exterior. Una posible ruta podría ser a través de la región variable, ya que es por aquí por donde también parece salir una vez glicosilada[15], sin embargo el bucle RAHRN en conformación LI no facilita la entrada del PGA debido a impedimentos estéricos que no parecen existir en la conformación LA; así o bien el PGA encuentra el bucle en conformación LA en la forma apo, algo que se ha demostrado muy infrecuente o bien el PGA induce el cierre del bucle, que por las simulaciones metadinámicas se ha visto que tampoco es así.

Por todo esto pensamos que el cambio conformacional ha de ser debido a la unión del UDPGlc; entre otras cosas porque la presencia del dador junto con el metal, parece ser la única interacción directa y necesaria (interacción metal-His258) para la estabilización del bucle, en la que participa un ligando y solo es posible en una determinada conformación, LA; además de la interacción Arg259-Glc, que parece mantener al UDPGlc en una conformación catalítica. Así que realizamos otra tanda de simulaciones metadinámicas, para tratar de medir la estabilidad del bucle en sus diferentes formas, en función de la orientación de la cadena lateral de la His258 y su proximidad al metal.

Nota: Como ya se comentó en la introducción de este capítulo, la coordinación His258-Metal no está parametrizada en nuestros cálculos. Únicamente la posible interacción electrostática que se produzca entre ambos será considerada. Una medida correcta de su energía de coordinación requeriría de cálculos cuánticos, como se ha hecho para la explicación del mecanismo S_Ni[15] de esta proteína. Pero no ha sido nunca nuestra inteción medir esta interacción, sino la dinámica del bucle RAHRN y consideramos que, si bien es probable que en presencia de metal, la conformación LA tendrá infraestimada su energía, globlamente el FES de las metadinámicas puede darnos un valor cualitativo sobre si realmente el ligando modifica o no el equilibrio conformacional.

MetaD4Y6NUDPGlc3. Variables colectivas: (Distancia, Torsión His258)

Como en la simulación MetaD4Y6N4Apo3, se activó el diedro de la cadena lateral de la His258 respecto a la distancia Asp136-Arg259, en este caso con el ligando UDPGlc+metal en su interior. De modo parecido a como se describió anteriormente, el UDPGlc cambia a conformación

"extendida" al poco de iniciarse la simulación y la orientación de la cadena lateral de la Arg259 se mantiene en valores de diedro de -1 rad.

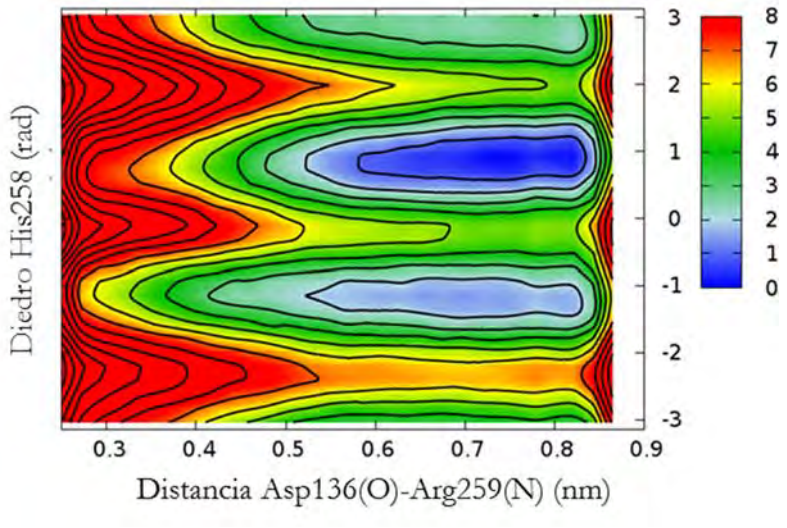


Figura 93. Representación energética de las CVs Distancia e His258. Isotermas 1 kcal/mol, error ± 0,5 kcal/mol.

1, -1 y π siguen siendo las poblaciones visitadas para el diedro His258, pero siempre en la conformación LI. His258(1) es 1 kcal/mol más estable que el diedro -1 y 2 kcal/mol más que el diedro π. No se ha conseguido desplazar el equilibrio hacia LA. Aunque los resultados parecen similares a la simulación MetD4Y6N4Apo3 (Figura 79) y no existe un desplazamiento del equilibrio conformacional hacia LA, un estudio más detallado de la zona LA de las dos simulaciones, ofrece un cambio de estabilidad respecto a la orientación de la cadena lateral de la His258 (figura 95).

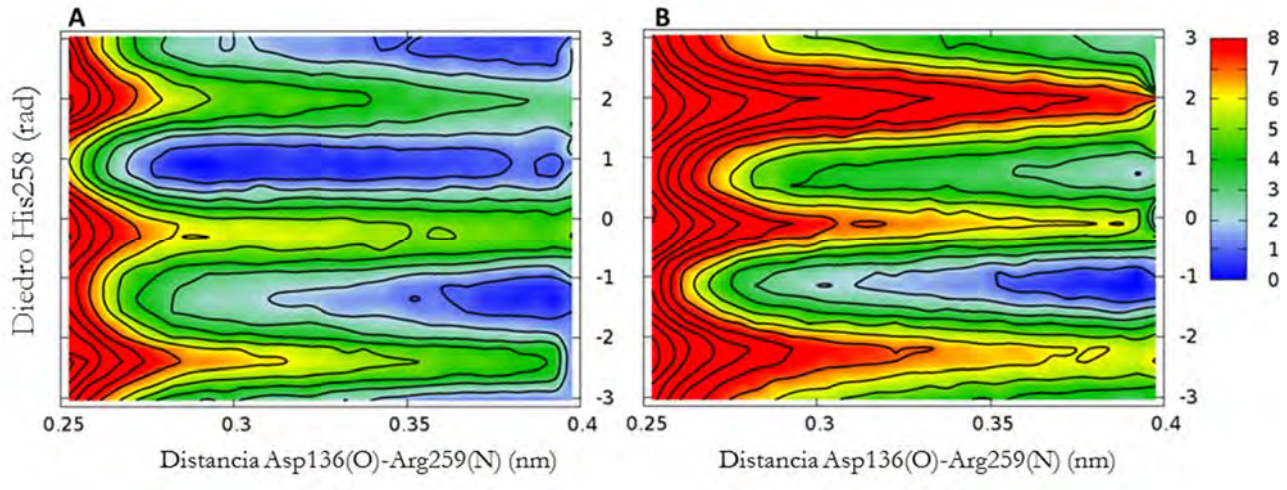


Figura 94. Superficies de energía libre para el cambio conformacional del bucle en función de la Distancia y diedro His258 en la zona de conformación LA. Isotermas 1 kcal/mol, error ± 0,5 kcal/mol. **A.** Simulación Apo, **B.** Simulación con UDPGlc+metal.

En la estructura apo (figura 79 y figura 94 A), la zona LA (alrededor de 0,3 nm) muestra un mínimo energético para valores de diedro His258 centrados en 1 rad. La presencia del UDPGlc (Figura 94 B) sí desplaza el equilibrio hacia valores del diedro His258 en -1 rad cuando el bucle se haya en conformación LA.,

Así pues, tanto en presencia como ausencia de ligando UDPGlc + Metal, la activación del diedro de la His258 junto a la distancia no altera el equilibrio LA:LI del bucle catalítico en GpgS. Sin embargo, **la presencia del ligando sí que parece aumentar, en cierta medida, la estabilidad de la cadena lateral de la His258 en la misma orientación que la observada en el cristal ternario (4Y6N) cuando el bucle se halla en conformación LA.**

MetaD4Y6NUDPGlc4. BIAS-Exchange. Variables colectivas: (Torsión Arg256, Torsión Ala257, Torsión His258)

Como ya se hizo para la forma apo de GpgS (MetaD-4Y6N-Apo4), se repitió la simulación metadinámica por el método de BIAS-Exchange con las CVs: diedros de los residuos Arg256, Ala257 y His258, pero en esta ocasión solo con el metal+UDPGlc en el centro activo (y diedro "c" de la UDPGlc fijado en -2,9 rad).

	Distancia	Diedro H258	Diedro A257	Diedro R256	Energía kcal/mol
M1 (LI)	6,9	1,26 **(1)**	2,5	0,84 **(1)**	0
M2 (LA)	2,8	-1,26 **(-1)**	2,9	-1,26 **(-1)**	0,5
M3 (LI)	6,9	0,84 **(1)**	2,5	-1,26 **(-1)**	0,5
M4 (LA)	2,8	-1,25 **(-1)**	2,9	2,9 **(π)**	0,5
M5 (LI)	7,7	1,25 **(-1)**	1	1,6 **(1)**	3

Tabla 15. Energía de los diferentes metaestados. Entre paréntesis, valores redondeados.

Los resultados son energéticamente similares a los de la simulación apo (MetaD4Y6NApo4) pero con sutiles diferencias. El metaestado más estable (M1) corresponde a la conformación del bucle LI, con los valores de diedro Arg256/His258 en 1/1. Para el resto de metaestados, el diedro Arg256 puede tomar diferentes valores, pero His258 siempre está en -1 rad cuando el bucle adopta conformación LA (mínimos M2 y M4) y en 1 rad cuando la conformación es LI (mínimos M3 y M5). Únicamente se observa el diedro His258 en valor -1 y conformación LI en un metaestado de energía superior al resto (mínimo M5, + 2,5 kcal/mol), cuyo diedro Ala257 no corresponde con el canónico para MD. Éste último mínimo (M5) corresponde en realidad a la conformación del bucle LI* observada en el cristal 4DDZ de GpgS en forma apo y que también había sido identificado en metadinámicas anteriores (MetaD4DDZApo y MetaD4Y6NApo1).

En cuanto a la correlación entre diedros His258 y Arg256, al proyectar la energía libre aquí calculada sobre estas dos coordenadas y compararlas con la simulación apo (MetaD4Y6NApo2) se observa que la presencia de UDPGlc bloquea el estado Arg259(1)/His258(-1): Aunque el estado Arg259(1)/His258(-1) parece no ser el más estable energéticamente, el UDPGlc provocará que salir de este sea más costoso que en la forma apo, al aumentar su barrera (figura 95).

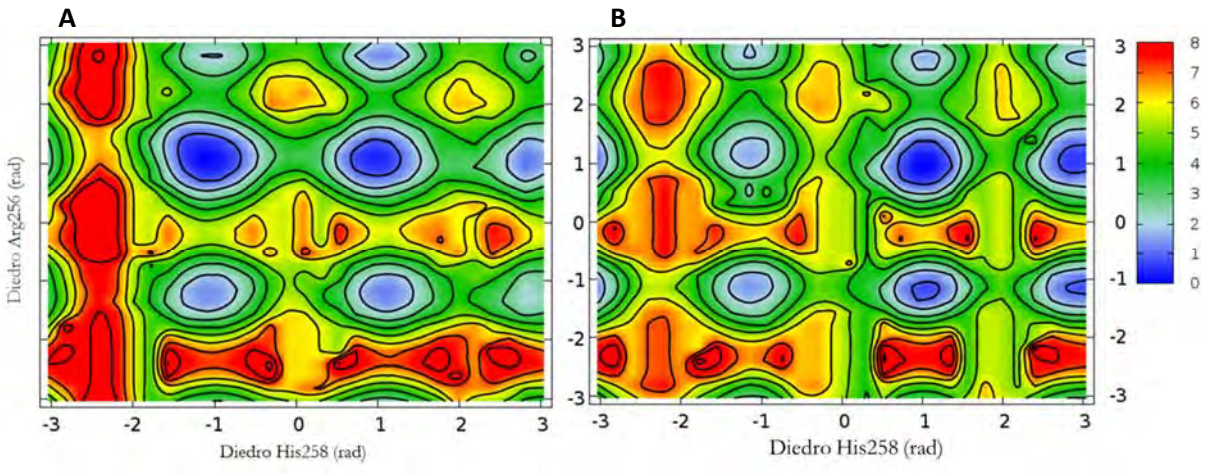


Figura 95. Energía libre proyectada sobre las variables colectivas diedro Arg256 y diedro His258 para las simulaciones: A apo (MetaD4Y6NApo2) y B complejo con UDPGlc+Metal (esta simulación).

MetaD4Y6NUDPGlc5. BIAS-Exchange. Variables colectivas: (Distancia, Torsión Arg256, Torsión His258, Distancia metal-His258, Diedro UDPGlc)

Se realizó una última simulación metadinámica por BIAS Exchange. En este caso se incluyeron todas las CVs utilizadas en las anteriores simulaciones: (i)distancia Asp136(O)-Arg259(N), (ii) torsión Ala257 (variables necesarias para explorar el metaestado LI* cristalizado en la estructura 4DDZ); (iii)torsión del diedro His258, (iv) distancia metal-His258 (para comprobar si puede existir estabilización del diedro His258 en -1 rad por interacción de la His258 con el metal en la conformación LA del bucle); (v)diedro "c" del ligando UDPGlc (para valorar la relación entre la conformación del ligando y la orientación de la Arg259, teniendo en cuenta la interacción observada entre ellos en la dinámica molecular del complejo ternario, MD4Y6NUDPGlc·PGA). **Mediante el uso de esta última variable se espera una activación indirecta del diedro Arg259 mediante el movimiento de la Glc, por medio de la activación directa del diedro UDPGlc**. La hipersuperfície de energía libre obtenida (en función de 5 variables colectivas) fue analizada mediante la herramienta de análisis METAGUI (ver métodos). Los metaestados identificados se recogen en la Tabla 16.

	Distancia (Å)	Diedro A257 (rad)	DiedroH258 (rad)	H258-Metal (Å)	Diedro Glc (rad)	Energía (kcal/mol)
M1 (LA)	3,2	2,5	-0,9 **(-1)**	4	-2,9	0
M2 (LI)	8,4	2,9	2,9 **(π)**	4	-1,3	0,7
M3 (LA)	3,6	2,9	2,9 **(π)**	7	-0,8	0,7
M4 (LI)	5,8	2,9	0,9 **(1)**	4	1,7	0,8
M5 (LI)	7,9	2,5	0,8 **(1)**	4	-2,5	0,9
M6 (LA)	3,2	2,5	2,9 **(π)**	4	-1,6	1,8
M7 (LA)	3,6	2,9	-0,8 **(-1)**	4	0,8	2,6

Tabla 16. Energía libre de los diferentes metaestados y sus coordenadas en función de las 5 variables colectivas utilizadas para el cálculo. Entre paréntesis, valores redondeados.

El metaestado más estable (M1) se corresponde a la conformación LA del bucle (distancia 3,2 Å), con el diedro His258 en valor -1 y el diedro del UDPGlc en -2,9 rads (conformación "doblada") que coinciden con los valores de la dinámica molecular MD4Y6NUDPGlc·PGA y equivalentes a la configuración del cristal 4Y6N. El segundo metaestado (M2), 0,7 kcal/mol más alto en energía corresponde a la conformación LI (distancia 8,4 Å) con el ligando UDPGlc en conformación "extendida" (diedro "c" UDPGlc en -1,3 rad) y diedro His258 en valor π.

Existen otros dos metaestados correspondientes a una conformación LI (M4 y M5, situados 0,8 y 0,9 kcal/mol por encima en energía libre). La diferencia de estos dos estados respecto a M2 es que la His258 adopta una configuración distinta (diedro His258 en 1 rad) En ambos metaestados, el UDPGlc adopta conformaciones tipo "doblada" (valores del diedro "c" de UDPGlc en 1,7 rad y -2,5 rad respectivamente).

La CV distancia metal-His258 no mostró ninguna preferencia energética entre las conformaciones LA/LI; son visibles dos metaestados de distancias 4 y 7 Å entre metal e His258, compartidos en las dos conformaciones del bucle. La interacción entre estos dos elementos, clave durante la catálisis, está infraestimada en las simulaciones basadas en dinámica molecular clásica, puesto que no describen bien las energías de coordinación con metales. Probablemente, el cambio conformacional observado de LI a LA, con el metal en el interior, desplazará el equilibrio hacia la forma LA estabilizándola mucho más en condiciones naturales, al tener en cuenta esta interacción. (Anexo 16)

El diedro del UDPGlc muestra en la proyección (figura 96) tres metaestados de idéntica energía. El más amplio corresponde a la conformación extendida, que cubre toda la extensión de las conformaciones LA/LI del bucle, con un valor entre -2, -1 rads. El segundo metaestado es el de la conformación doblada observada en la estructura del cristal (entre 1 y 2 rads), que aquí curiosamente se distribuye por la conformación LI del bucle. El tercer y último metaestado es el de la conformación doblada estable en la dinámica molecular clásica (-2,9 rads), que solo es visible en la conformación LA.

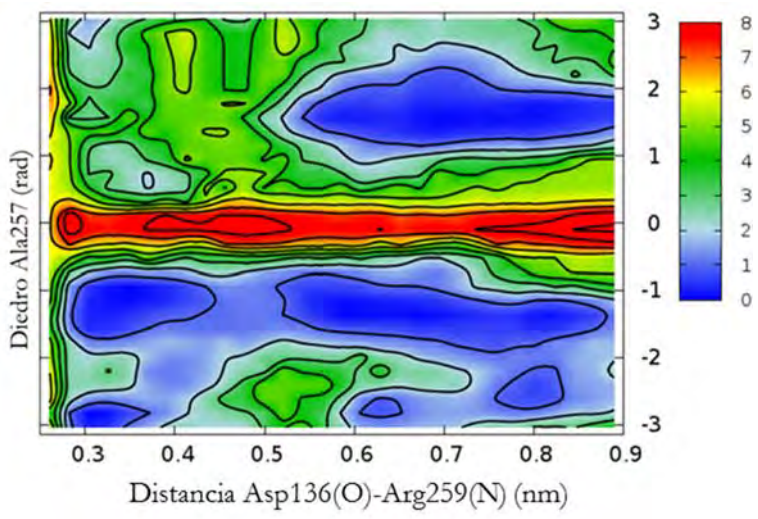


Figura 96. Proyección de la energía para el diedro UDPGlc sobre la distancia. Se observan los metaestados -2,93 y 1,5, conformaciones dobladas de la MD y cristal 4Y6N, y el metaestado -1,5 para la conformación extendida, todos ellos de igual energía.

Respecto al diedro Arg256, no activado como variable en esta simulación, sus tres típicos valores de diedro -1, 1 y π, se distribuyen equitativamente por las conformaciones LA, LI y metaestados de His258 o UDPGlc diedro, por lo que no se puede asignar un determinado valor de diedro para ellos, pero ocurre un hecho relevante en estas simulaciones, que la equipara a la simulación metadinámica anterior, por BIAS-Exchange: El diedro de la Arg256 que siempre se había manifestado como una variable lenta, cambia rápidamente su valor al tiempo que lo hace el diedro UDPGlc en las simulaciones en las que está activado, de igual modo que lo hacía en la simulación MetaD4Y6NApo4, donde sí estaba activado el diedro Arg256. El movimiento de los diedros de ambas moléculas (UDPGlc y Arg259) está claramente relacionado. Esto cobra especial relevancia si como hasta ahora se ha ido considerando, la activación del diedro Arg256 y su movimiento es la variable más influyente en la explicación del cambio conformacional, no es el valor que adopte este diedro, sino su movimiento, su cambio y las diferentes conformaciones del UDPGlc quienes promoverían este cambio, afectando indirectamente a la conformación del bucle en un claro mecanismo de ajuste inducido.

El movimiento del diedro UDPGlc induce el movimiento del diedro Arg256, que reduce la barrera para el cambio conformacional, desplazando el equilibrio hacia la forma LA. El diedro His258(-1), es la forma estable en el bucle en la conformación LA, no por su interacción con el metal, que no es reproducible computacionalmente, sino por una propiedad intrínseca del diedro; esta configuración con toda probabilidad se verá aumentada con la interacción metal-His258. La conformación de diedro del UDPGlc más estable, es la doblada en conformación del bucle LA, otras conformaciones doblada (valor del cristal) tienen un mismo valor energético en cualquier conformación.

6. Estudio hidropático del centro activo

Se ha sugerido que las diferentes conformaciones de bucles adyacentes al centro activo de glicosiltransferasas, restringen la entrada de moléculas de agua y juegan un papel crucial, no solo en la unión de los ligandos sino en la propia catálisis, incluyendo enzimas tanto tipo *inverting* como tipo *retaining*[81]. Se quiso comprobar si las diferentes conformaciones del bucle RAHRN en GpgS, también afectaban a la entrada de aguas en su centro activo. Para ello se midió el número de moléculas de agua que se encontraban en una esfera de 6 Å de los residuos del UDPGlc (O5r, O2r y C1r) y del PGA (O3, O2 y O1), a lo largo de la MD del complejo ternario (figura 97). Se escogió una zona entre los 200 y 600 ns donde la distancia UDPGlc se mantiene estable y el bucle cambia claramente de conformación de LA a LI (Figura 98).

El número de aguas se ve muy influenciado por la distancia entre los ligandos; esta se mantiene estable en el monómero A, con 4-7 moléculas de agua en su interior, pero sufre grandes cambios en el monómero B (figura 98 A).

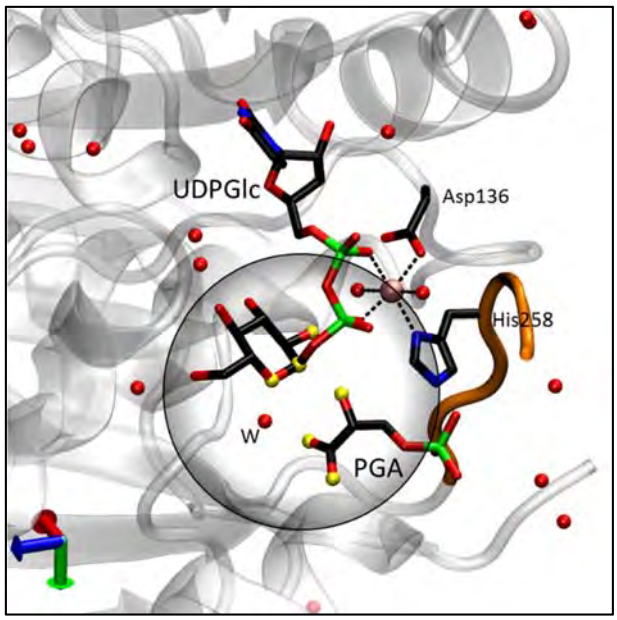


Figura 97. Esfera aproximada de 6 Å de diámetro, utilizada para contar el nº de moléculas de agua a lo largo de la simulación MD4Y6N (complejo ternario). La imagen muestra el CA del complejo ternario (cristal 4Y6N), con los ligandos en su interior, (aguas en rojo). Los átomos O5r, O2r, C1r del UDPGlc y O3, O2 y O1 del PGA se representan en amarillo, estos delimitan el centro de la esfera. El metal, (en rosa) interacciona con 2 oxígenos de los fosfatos del UDPGlc, uno del Asp134, el Nδ de la His258 y dos moléculas de agua formando la coordinación octaédrica del ión. En el cristal, una sola molécula de agua se encuentra en la esfera de solvatación (W3). El bucle RAHRN, en conformación LA, se muestra en color naranja.

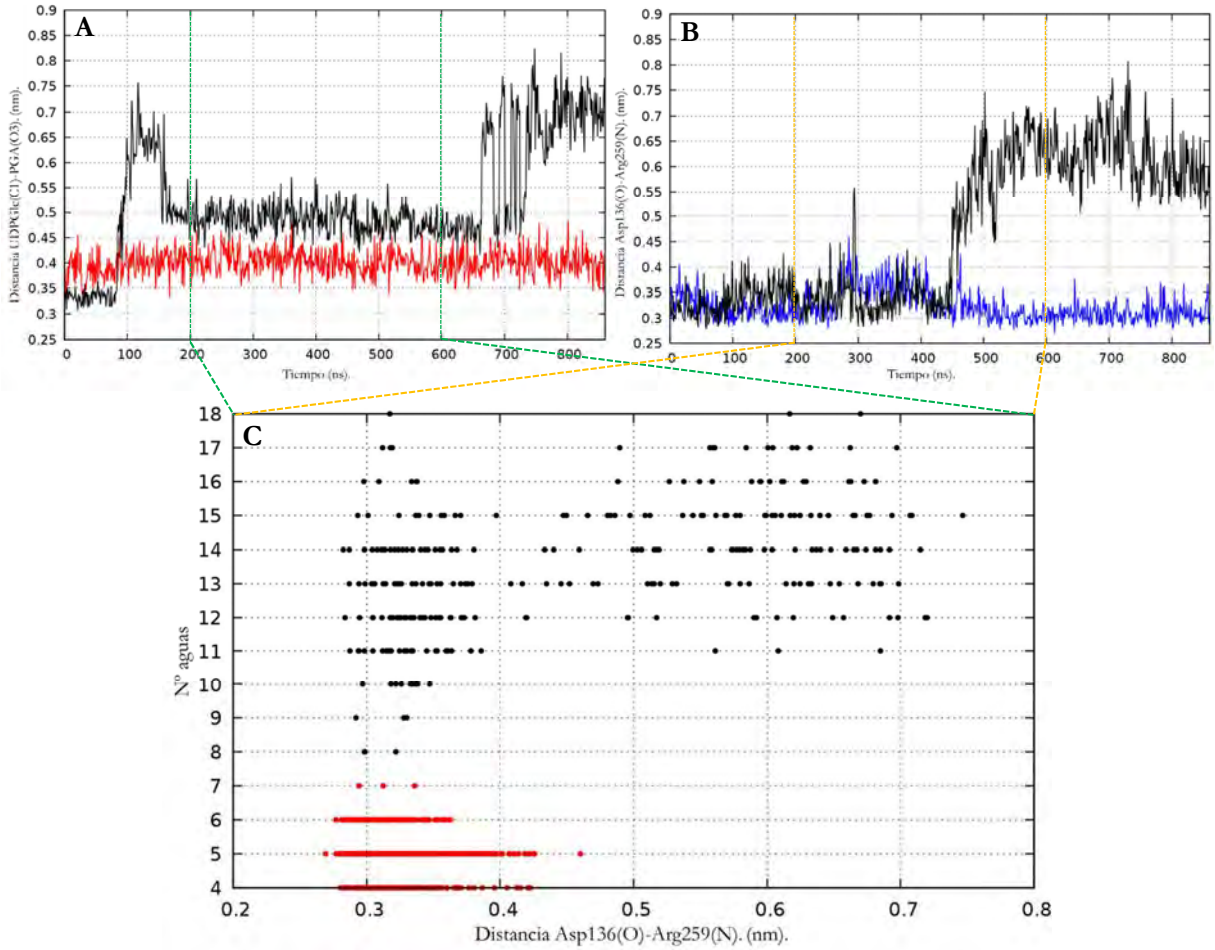


Figura 96. A. Distancia entre ligandos: Rojo, monómero A. La distancia se mantiene constante en torno a 4 Å. Negro, monómero B. Los ligandos se distancian y acercan sin alcanzar nunca el valor inicial. Entre los 200 y 600 ns hay una zona de valor estable que se utiliza para medir el número de moléculas de agua. **B**. Distancia bucle-DXD: Mientras el monómero A mantiene estable la distancia en 0,35 nm (azul), en el monómero B (negro) se produce un aumento que implica el cambio de conformación del bucle, entre los 200 y 600 ns. **C**. Una esfera de 6 Å para el monA (rojo) contiene un valor estable de 4-7 moléculas de agua. En monB (negro), la distancia entre ligandos es mayor y el conjunto de moléculas de agua alrededor de ellos oscila entre 8 y 18. Cuando en este monómero el bucle cambia de conformación de LA a LI, esta fluctuación se reduce a un número entre 11 y 18, aumentando el número mínimo de moléculas de agua, por lo que la conformación LA posibilita la salida de hasta 3 moléculas de agua.

La conformación LA permite al centro activo alcanzar estados de solvatación de unas 3 moléculas menos de agua que la conformación LI. Esto querría decir que el cambio de conformación de LI a LA y el cierre del bucle, podría conllevar la expulsión de un porcentaje alto de moléculas de agua aumentando considerablemente la hidrofobicidad del entorno, que tomando como referencia el monómero A podría ir de 4-7 moléculas en la conformación LA, a 7-10 en la conformación LI.

7. Discusión

El uso combinado de simulaciones de dinámica molecular clásica y metadinámicas, nos ha permitido encontrar un equilibrio conformacional entre dos estados: LA y LI, para el bucle $R_{256}AHRN_{260}$ de la proteína GpgS sin ligandos. La diferencia energética entre esos dos estados se sitúa entorno 2,5 kcal/mol, siendo la conformación LI la más estable. La forma LA no parece ser estable, puesto que no existe barrera entre ambas conformaciones sino más bien una rampa energética entre LA-LI, que es superable por las fluctuaciones térmicas del sistema. Se ha visto por dinámica molecular clásica que la conformación LA es visitable, aunque en menor medida que LI. Del mismo modo, la conformación LA puede estabilizarse en la forma apo, mediante la formación de dos puentes de hidrógeno entre los residuos Ala257, Arg259 del bucle y Asp136, Ile138 del motivo DXD y principalmente por el diedro de la cadena lateral del residuo His258, encallado en los valores 1 y π, sin visitar el valor -1.

La presencia del UDPGlc+metal en el centro activo de la proteína, estabiliza el diedro His258 en valor -1, cuando el bucle se encuentra en conformación LA. Solo en esta forma de bucle y diedro es posible la interacción metal-His258 que completa la coordinación octaédrica del ión, junto al Asp136, los fosfatos del UDP y dos aguas estructurales, tal y como se observa en los cristales. Aunque no se ha reproducido esta interacción, para la que se necesitarían estudios QM/MM[15], si hemos podido medir cierta estabilidad de la forma LA por ella. La estabilización energética real de esta interacción, sumada a la producida por los puentes de hidrógeno antes mencionados de la forma LA, probablemente dotará a esta conformación de una estabilidad mucho mayor en presencia de UDPGlc que la reportada en nuestros estudios. Aun así, la barrera entre las conformaciones LA/LI, no se reduce por la presencia del UDPGlc+metal, es más, aumenta al doble al encontrarse este en el centro activo. Aunque el ligando dador estabilice la conformación LA, es un impedimento para el cambio de conformación del bucle desde la forma LI.

Los experimentos de *docking* con los ligandos dador UDPGlc y aceptor PGA, no mostraron ninguna preferencia de estos por las conformaciones LA o LI de la proteína GpgS, con o sin metal. En una situación como esta, donde queda descartado un mecanismo de selección conformacional, la llegada al CA del dador en primer lugar encontraría el bucle en conformación LI preferentemente e impediría el cambio conformacional (la barrera aumenta) y la estabilización del metal y el ligando (la His258 no estaría en el diedro adecuado ni a la distancia correcta), además los resultados indican una disminución de la afinidad por la proteína del PGA, cuando el UDPGlc se haya presente.

No hemos encontrado ninguna relación directa entre el PGA y las conformaciones del bucle RAHRN. El UDPGlc presenta el diedro entre la Glc y el UDP en tres estados principales, uno en conformación extendida, no catalítico y estable –cuando el PGA no está en el CA o se encuentran alejados los ligandos– y otros dos en conformación doblada, catalíticos y solo estables cuando se encuentra el PGA en el interior de GpgS. Estas conformaciones dobladas del UDPGlc muestran un grado de afinidad 1 kcal/mol mayor por el CA que la extendida y son las únicas que posibilitan la transferencia del azúcar.

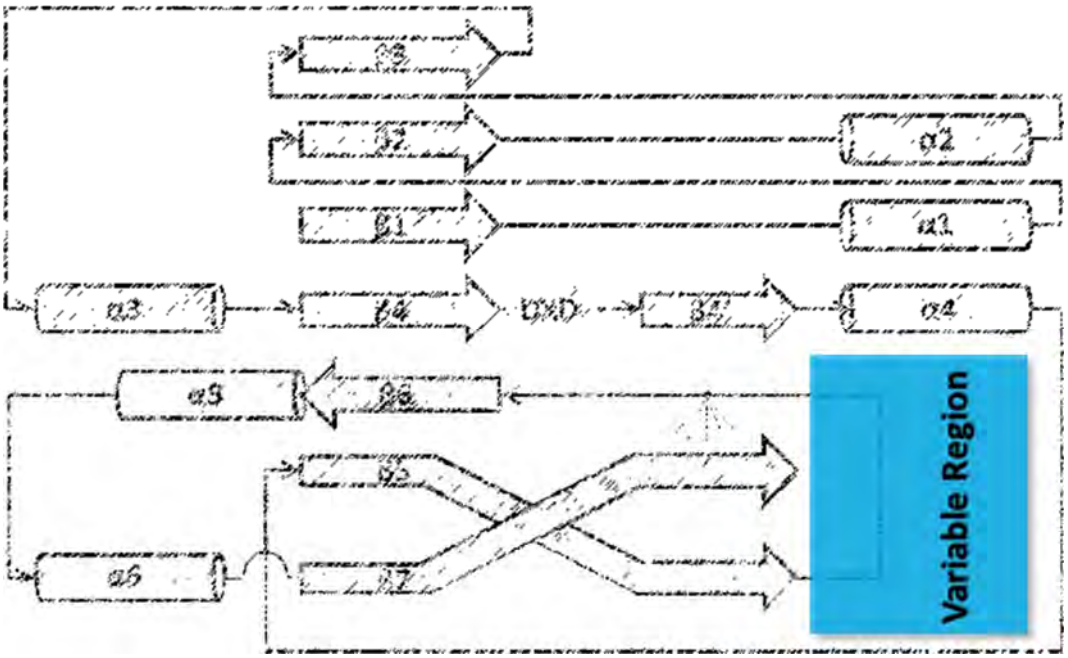
Por todo esto consideramos que la llegada del aceptor PGA al centro activo de la proteína se produce antes que la del UDPGlc. Siendo así, la conformación del bucle que encontraría el PGA sería LI; si como suponemos, en esta conformación existen impedimentos estéricos para que el PGA acceda al sitio catalítico a través de la RV, probablemente lo hará entonces directamente por el bolsillo del centro activo, por el mismo lugar que el UDPGlc y quizás una de las razones por las que el PGA parece reducir su afinidad por la proteína cuando el UDPGlc se encuentra ya posicionado. Estos resultados, el orden de entrada y la posible inhibición competitiva del PGA al acceder al CA por el mismo sitio que el dador, los consideramos compatibles con los presentados por Kumar *et al*[129].

Pero **¿Qué reduce la barrera entre las conformaciones LA/LI?** Hemos comprobado que en la forma apo de GpgS, la activación del movimiento de un diedro de la cadena lateral de la Arg256, produce una bajada de esta barrera hasta dejarla en 1 kcal/mol y una ligera estabilización de LA frente a LI. Sin activar este movimiento, obtenemos el mismo resultado con el complejo ternario y nunca con los ligandos por separado. Las conformaciones dobladas del UDPGlc se estabilizan mediante la interacción, entre otras, de la Glc con la cadena lateral de esta Arg256, cuyo diedro interactúa de igual manera en los valores 1 y -1 con la Glc. El cambio entre estos valores de diedro en la Arg256 es lento cuando solo actúan fluctuaciones térmicas, pero es facilitado cuando lo promueve el movimiento de la Glc, tal y como se ha demostrado en nuestras simulaciones.

Por todo esto creemos que con el PGA en el sitio catalítico y el bucle en conformación LI, el UDPGlc llega a la proteína en conformación extendida y la presencia del PGA provoca su cambio a la conformación doblada y la interacción con la Arg256. Este movimiento del diedro UDPGlc promovido por el PGA, provoca a su vez el movimiento del diedro Arg256 que facilita el cambio conformacional del bucle de LI a LA, mediante la reducción de la barrera energética entre estas dos conformaciones. Una vez el bucle se encuentra en la forma LA, el UDPGlc+metal estabiliza esta forma mediante el bloqueo de los diedros Arg256(1) y His258(-1) y la interacción electrostática entre el metal y la His258.

Nota: Aunque encontramos en esta descripción fenomenológica, el mecanismo de ajuste inducido más plausible para el cambio conformacional del bucle, no podemos omitir que en la simulación metadinámica del complejo ternario, donde la presencia de todos los ligandos parece promover la conformación LA, el diedro del UDPGlc se haya fijado en conformación doblada y aunque estabiliza el diedro de la Arg256 en valor 1 para la forma LA, (como en el cristal) no puede facilitar su movimiento al estar fijado, fenómeno que nosotros proponemos como elemento clave para la reducción de la barrera entre LA/LI. Por ello, aunque insistimos en la importancia de la Arg256 y el movimiento de su diedro como explicación para el cambio conformacional, no podemos descartar otras variables no descritas.

Por último, el cambio conformacional del bucle hacia la forma LA no contribuye solo a la estabilidad de los ligandos en el CA, sino creando además un entorno más hidrofóbico, al permitir la expulsión de 3 moléculas de agua, favoreciendo la reacción.



LA REGIÓN VARIABLE COMO PREDICTOR DE ESPECIFICIDAD DE SUSTRATO EN GLICOSILTRANSFERASAS GTA

1. Introducción

Durante el modelado de la GT MG517 se definió un patrón estructural consenso para todas las GTA (capítulo 1), con una región conservada, común en lo esencial para todas ellas y dentro de esta, una pequeña región, en comparación con el tamaño de la conservada, que era variable tanto en secuencia como en estructura y a la que llamamos consecuentemente, región variable (RV). De la observación del alineamiento, las estructuras y sus ligandos en las proteínas cristalizadas, llegamos a la conclusión de que, para las proteínas con plegamiento GTA, los residuos que interactúan con el ligando dador, se hallan repartidos por toda la región conservada, ya sea por una función catalítica o estabilizadora de la molécula en el centro activo. Aunque cada proteína logra la estabilidad del dador mediante diferentes estrategias, debido también a la diversidad de moléculas dadoras, todos ellos se encuentran en "zonas calientes" de la región conservada, zonas a las que pertenecen los residuos utilizados para los experimentos de mutagénesis dirigida de la GT MG517 del capítulo 1. Del mismo modo observamos que la molécula aceptora se ubicaba siempre en la misma zona de las estructuras, la región variable. Todos los cristales que incluían la molécula aceptora, presentaban la molécula en el entorno de la región variable. El propio resultado del *docking* de DAG (la molécula aceptora de MG517, capítulo 1.7), indicaba que para esta proteína también el aceptor se posicionaba preferentemente en la RV.

La reacción de transferencia del azúcar a su molécula aceptora en las glicosiltransferasas es específica en cada proteína. Cada GT une específicamente un nucleósido y transfiere preferentemente un tipo de azúcar, según qué residuos contenga su secuencia en esas "zonas calientes" de las que hablamos para la región conservada. De la misma forma, el azúcar se transfiere a una molécula aceptora concreta y puesto que es en la RV en donde esta se ubica, podría ser esta zona la responsable de la especificidad para cada GT por un determinado aceptor.

Nos dispusimos a demostrar esta hipótesis, así como responder a otras tres grandes preguntas:

1. ¿Está la especificidad por el aceptor de las GTAs explicada en exclusiva por la región variable?
2. ¿Podemos predecir el aceptor de una GTA conociendo la secuencia de su región variable?
3. ¿Sigue la región variable un proceso de evolución divergente como la región conservada, o bien convergente, de modo que diferentes proteínas que comparten aceptor también compartirán región variable? Estos dos escenarios evolutivos se explicarían del siguiente modo:
    - Podemos pensar que un determinado aceptor es reconocido solo por una familia o un grupo particular de proteínas dentro de una misma familia, que estarían relacionadas filogenéticamente. En este caso las RV de estas proteínas serán homólogas y sería posible construir un perfil con ellas, con el que predecir el aceptor de una proteína desconocida, si parte de su secuencia alinea con este perfil; también estaríamos hablando de una evolución divergente de la RV, donde a partir de

secuencias ancestrales esta región habría ido divergiendo y variando su especificidad por nuevos aceptores.

- Por otro lado, que un mismo aceptor fuese compartido por diferentes familias de GTAs, filogenéticamente lejanas y con RV no homólogas, significaría que diferentes secuencias pueden tener especificidad por una misma molécula y que la explicación de esta especificidad está un nivel por encima de la secuencia, en la estructura, siendo esta necesaria para realizar una predicción de un posible aceptor. Un contexto como este implicaría una evolución convergente de la RV, donde desde diferentes familias y siguiendo una evolución acelerada en comparación con la región conservada, la RV ha podido converger hacia la especificidad por un mismo aceptor, por adquirir estructuras similares o posiciones clave similares de los residuos implicados en la especificidad.

Para responder a estas preguntas realicé una estancia de tres meses en la Unidad de Glicogenómica de AFMB - CNRS - Université d'Aix-Marseille, donde se ubica y gestiona la base de datos CAZy, la principal y mayor referencia documental de proteínas que degradan, modifican o crean enlaces glicosídicos. Allí se realizó un estudio estructural y filogenético de las diferentes familias GTA, centrándonos en aquellas que comparten un gran número de aceptores diferentes.

2. Análisis estructural de GTAs

Se amplió el estudio de todas las proteínas GTA cristalizadas hasta la fecha, a todas las estructuras existentes (abril 2015), ya que en el capítulo para el modelado de MG517, solo se usaron las que tenían mejor resolución y podían existir otras estructuras con diferentes zonas resueltas o conformaciones y también diferentes ligandos. Se estudiaron 498 estructuras GTA de 278 cristales, que corresponden a 34 proteínas de 13 diferentes familias GTA (Anexo 18). Todas las estructuras comparten los elementos del patrón estructural consenso propuesto para este grupo de GTs, con la reseñable excepción de la familia GT7 cuya hélice α4 está ausente en las estructuras cristalizadas y donde los elementos α2 y β3 se hallan presentes, pero provienen de la zona N-terminal y C-terminal respectivamente. La topología consenso comprende pues 8 hojas β conservadas y 6 hélices α (figura 99).

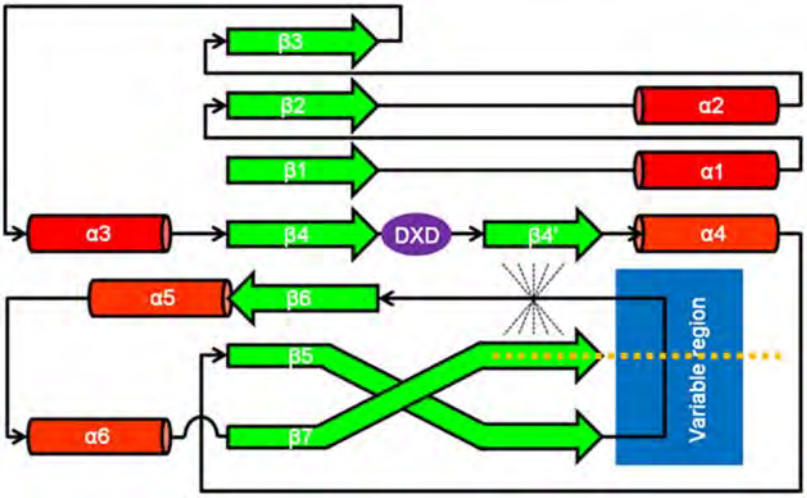


Figura 99. Topología consenso propuesta para el grupo de proteínas GTA. (La línea punteada amarilla representa el bucle desestructurado que conecta la región N-terminal con la extensión C-terminal, cuya longitud es muy variable).

Las hojas β1 a β7 forman la lámina β principal de las GTAs, las hélices α1, α2 y α4 se sitúan a un lado de la lámina β principal, generalmente frente a las hojas β1, β2 y β4' mientras que las hélices α3, α5 y α6 se encuentran en el lado opuesto, frente a las hojas β4, β6 y β7. Las hojas β5 y β7 suelen extenderse y cruzarse formando una segunda lámina β, perpendicular a la principal, con la hoja β4' en conformación antiparalela. La hoja β4' no siempre está presente y su formación depende de la extensión de la hoja β7 y de su conformación, que confiere a esta extensión de la hoja β7 capacidades catalíticas, a modo de tapadera, en algunas proteínas[127] como se ha visto para GpgS (capítulo 3) donde esta extensión la forma el bucle RAHRN.

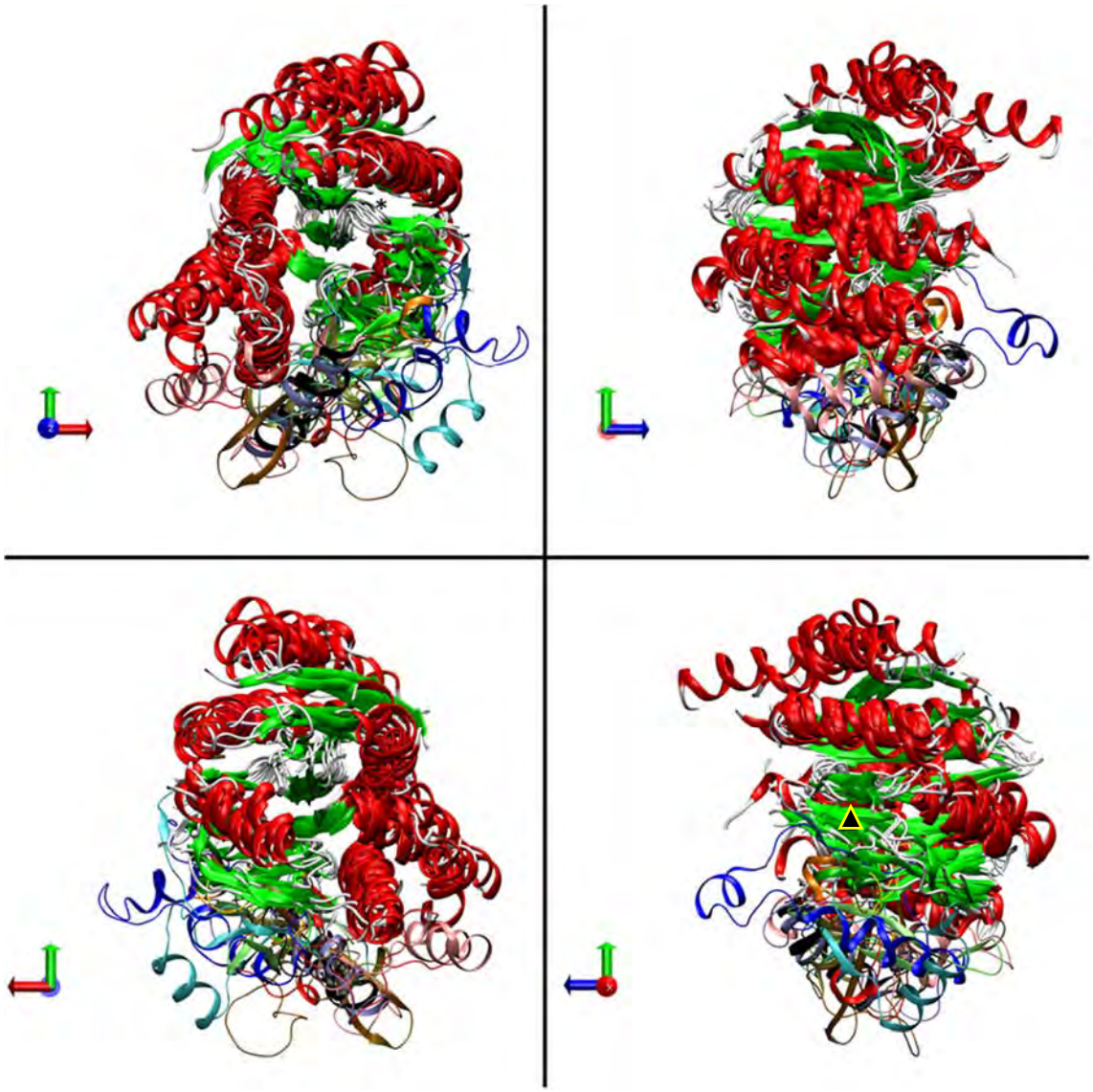


Figura 100. Superposición de estructuras GTA. En la imagen se muestran 16 estructuras cristalográficas superpuestas mostrando solo los elementos estructurales de la topología consenso más la RV. Se muestran 4 orientaciones distintas según los ejes de coordenadas. La topología consenso consta de un esqueleto de hojas β (en verde) rodeado por 6 hélices α (en rojo). El asterisco (primer cuadrante) identifica el motivo DXD y el triángulo negro (cuarto cuadrante) la extensión β7 bajo el elemento estructural β4'. En la zona baja de la topología consenso, pero aun dentro de la misma se encuentra la RV →Azul: 4HG6 (GT2). Rojo: 1O7Q (GT6). Naranja: 2AE7 (GT7). Amarillo: 4LW6 (GT7). Canela: 1GA8 (GT8). Verde: 3T7O (GT8). Rosa: 1S4P (GT15). Cian: 5AJP (GT27). Lima: 3CU0 (GT43). Ocre: 1ON8 (GT64). Azul acero: 2BO6 (GT78). Negro: 4DEC (GT81).

Entre las hojas β4 y β4' se encuentra el motivo DXD, común en las GTAs, formando un bucle. La cavidad formada por las hojas β1 y β2, el motivo DXD y las hélices α3 y α6 conforman el lugar en donde se aposenta el ligando dador, conteniendo la hélice α6 el residuo conservado que suele actuar

como base catalítica en la transferencia del azúcar. La extensión de la hoja β7 no está siempre resuelta pues a partir de esta y tras superar la hoja β4', se dan dos tipos de conformaciones diferentes en las GTAs y por ello no incluidas en la estructura consenso, o bien la hoja se curva hacia la hélice α1, dejando abierta la hendidura catalítica como ocurre en las familias GT8, GT13 y GT15 o bien continúa con un bucle flexible (y difícil de cristalizar) que finaliza en una hélice α7 (no incluida en la estructura consenso) situada bajo la hélice α6 y que cierra la hendidura catalítica, tal como se observa en el resto de familias, salvo GT6 y GT43 donde ha sido imposible determinar su posición a partir del cristal, aunque se hayan presentes ambos elementos. Esta extensión de la β7 –que puede contener otros elementos intermedios–, es especialmente interesante tras su cruce con la β5, bajo la β4', pues suele contener un residuo que interactúa directamente con el metal, generalmente una histidina (figura 99, línea punteada amarilla). Algunas familias, como GT27, GT78 o GT81, parecen poder modificar esta interacción gracias la flexibilidad conformacional de esta zona, que le confiere un papel clave en la estabilidad de los ligandos[131,91]. Por último, la hoja β6 y hélice α5 generan una estructura fusionada y continua fácilmente identificable, la hoja β6 además forma un codo en todos los cristales estudiados, es entre este codo y la hoja β5 donde se ubica la región variable. Cuando se compara la estructura para esta región entre los diferentes cristales, esta no superpone, (figura 100) y aunque dentro de la misma familia la superposición es mayor, esta difiere cuando el aceptor es diferente. Únicamente en la familia GT6 es posible encontrar la misma estructura de región variable para dos aceptores distintos: Gal para 1O7Q[90] y Fuc para 3V0O[132] aunque con sutiles diferencias en su secuencia. Del mismo modo las familias GT55 y GT81 presentan una gran homología en su estructura y comparten aceptor: fosfoglicerato (PGA)[133,81], mientras que GT78 comparte con estas dos familias parte de sus estructura y parte de su aceptor: glicerato (GA)[134].

En cada estructura se identificaron los residuos que se encontraban a no más de 4 Å del aceptor, cuando este estaba presente (tabla 17). En todos los casos siempre se encontraba uno o más residuos de la región variable y algunos de estos residuos estaban directamente implicados en el posicionamiento del aceptor como bGalT7 en GT7[135], GlcAT-I en GT43[136] o GpgS en GT81[81]. En otros casos estos residuos se encontraban en elementos alejados de la región variable, generalmente en la hélice 6 como BcsA en GT2[137] o GTB en GT6[94] aunque residuos de la RV se hallen también involucrados:

Tabla 17. Estructuras con ligando y su relación con la región variable.

Familia GT	Descripción CAZY	PDB	Uniprot ID	Ligando	RV	Residuos ligando*
GT2	Celulosa sintasa subunit A (BcsA;RSP_0333)	4HG6	Q3J125	GLC	274-321	H276, Y302, C318, **W383**
GT6	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	1O7Q	P14769	NAG	247-282	**N247, W249, W250**, T259, Y278, H280, W314, D317, W356, K359
GT6	B-specific a-1,3-galactosyltransferase (GTB)	3SXA	Q9UQ63	GAL	233-268	H233, F236, T245, Y264, **W300**, E303
GT6	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	3V0O	P16442	Accep. analog	233-268	H233, P234 F236, T245, Y264, **W300**, E303, D326 , L329, A343
GT6	UDP-GalNAc: 2'-fucosyl lactose a-N-acetylgalactosaminyltransferase (BoGT6a;BACOVA_01932)	4CJC	A7LVT2	AMINOGAL	122-157	H122, **W189**, E192
GT7	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	4KRV	P08037	Accep. analog	273-292	K279, F280, Y286, Y289, W314, D318, D319, R359
GT7	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	2AE7	P15291	NAGMAN	274-288	Y282, Y285, F286, W310, G311,G312, D314, D315, R355 F356, I359,
GT7	xylosylprotein b-4-galactosyltransferase I / 7 (bGalT7/CG11780)	4LW6	Q9VBZ9	XYL	168-185	**Y179**, D212
GT8	UDP-Gal: a-1,4-galactosyltransferase (LgtC)	1GA8	P96945	4-DEOXYLAC	129-155	I76, H78, F132, V133, N153, I79, Y186, Q189, P211, T212, C246, G247, P248
GT8	glycogenin (Gyg;Gyg1;GYG1)	3U2U	P46976	GLC	125-131	M75, R77, L80, D125, G127 N133, Q164, L214
GT15	GDP-Man: a-1,2-mannosyltransferase (Kre2;Mnt1;YDR483w;D8035.26)	1S4P	P27809	MAN	274-327	L213, Y214, Y220, L277, E279, Y280, H323, W325, R358
GT27	UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase 10 (GalNT10;ppGalNAcT10;pp-GalNAc-T10)	2D7I	Q86SR1	PEPTIDE	257-309	V255, A266, L270, Y284, F361, W282, E334, F361 to Y365
GT43	UDP-GlcA: b-1,4-galactosyl-xylosylprotein b-1,3-glucuronosyltransferase 3 (GlcAT-I;B3GAT3)	3CU0	O94766	GAL	217-255	V221, E227, G222, G223, **W243**, R247, V251, D252, R277, G278, G281
GT43	UDP-GlcA: galactosylgalactosylxylosylprotein 3-b-glucuronyltransferase 1 (GlcAT-P;B3gat1)	1V84	Q9P2W7	LACNA	223-257	G223, **F245**, G280, G281
GT64	a-N-acetylhexosaminyltransferase (Extl2)	1ON8	Q8C089	GAL	178-215	R181, **W284**, F290
GT78	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	2BO6	Q9RFR0	GLYCERATE	129-166	R131, A136, M137, I138, T139, L226, M229
GT81	UDP-Glc: glucosyl-3-phosphoglycerate synthase (Rv1208;GpgS;MtGpgS;MTCI364.20)	4DEC	P9WMW8	PGA	166-211	G184, **R185**, V186, **T187**, L209, N260

*Residuos no más lejos de 4 Å del ligando. Se encuentran subrayados aquellos residuos incluidos en la RV. En negrita los que están directamente implicados en el posicionamiento del aceptor.

3. Evolución de la región variable respecto al resto de la secuencia

En vista de estos resultados, se estudió la secuencia de esta región, ampliando el número de proteínas a todas las incluidas en la base de datos CAZy, para todas las familias GTA. Algunas de estas familias son más heterogéneas que otras, con el caso paradigmático de la familia GT2, el grupo más numeroso y diverso de todas las GTAs. Para reducir esa heterogeneidad, CAZy dispone de una subdivisión (no publicada) de grupos por homología dentro de cada familia, asignando una letra o un número a cada uno de esos grupos dentro de cada familia (ver métodos). Se estudiaron 6 familias GTAs, que fueron elegidas por la variedad de aceptores que contienen: GT2, GT7, GT8, GT55, GT78 y GT81. Otras familias como GT12, GT13, GT15, GT21, GT27, GT43 o GT64 poseen cada una un solo tipo de aceptor conocido.

La familia GT2 es muy grande y heterogénea para ser utilizada en conjunto, por ello se escogió un subgrupo concreto, el GT2_P, que incluye a un gran número de proteínas con la característica común de tener actividad procesiva y al menos tres aceptores diferentes; Glc, GlcNAc y GlcA. Las familias completas GT7 (4 aceptores diferentes conocidos) y GT8 (7 aceptores diferentes conocidos). También las familias GT55, GT78 y GT81; aunque GT55 y GT78 contiene un solo tipo de aceptor conocido y GT81 dos, por las características comunes que comparten región variable y aceptores. Para cada secuencia de cada familia estudiada, se anotó también la información disponible acerca de la naturaleza del aceptor.

Cada grupo se alineó con diversos métodos (TCoffee o Mafft, ver métodos) y se realizaron árboles filogenéticos con Fasttree. Los árboles filogenéticos se realizaron con el dominio GTA completo de las proteínas de cada grupo y también, con la RV extraída, únicamente con la RV, y con otros segmentos de la secuencia de longitud idéntica a la RV.

3.1 Familia GT2

Familia muy heterogénea de GTAs, con representantes en todos los reinos y funciones también muy diversas. Se realizó un árbol filogenético para el dominio GTA de todas las proteínas del grupo homólogo GT2_P (ver métodos), grupo de algo más de 6000 secuencias cuyas proteínas poseen actividad procesiva. Se identificó cada secuencia con su aceptor cuando era conocido y su taxón y se realizó el árbol filogenético (ver métodos) (figura 101).

Se observan dos grandes grupos, el de las proteínas que tienen como aceptor la GlcNAc (b-1,4-N-acetilglucosaminiltransferasa, AagC, quitina sintasa y hialuronano sintasa) y el de las proteínas cuyo aceptor es la Glc (celulosa sintasa, xiloglucano b-1,4-glucano sintasa, b-1-3-glucanosintasa cíclica, b-1,3-glucanosiltransglucosidasa, b-manosiltransferasa, b-1,4-manano sintasa y b-1,4-manano sintasa / glucomanano sintasa) con un representante de Man como aceptor, dentro del grupo de Glc, y teniendo en cuenta que la hialuronano sintasa puede aceptar también GlcA. Se puede ver en el árbol como estos dos grupos se hallan bien separados y las proteínas que forman cada uno,

relacionadas filogenéticamente entre ellas. Otras ramas fuera de estos grupos podrían contener aceptores distintos a GlcNAc o Glc, aunque no se puede afirmar con rotundidad con los presentes datos.

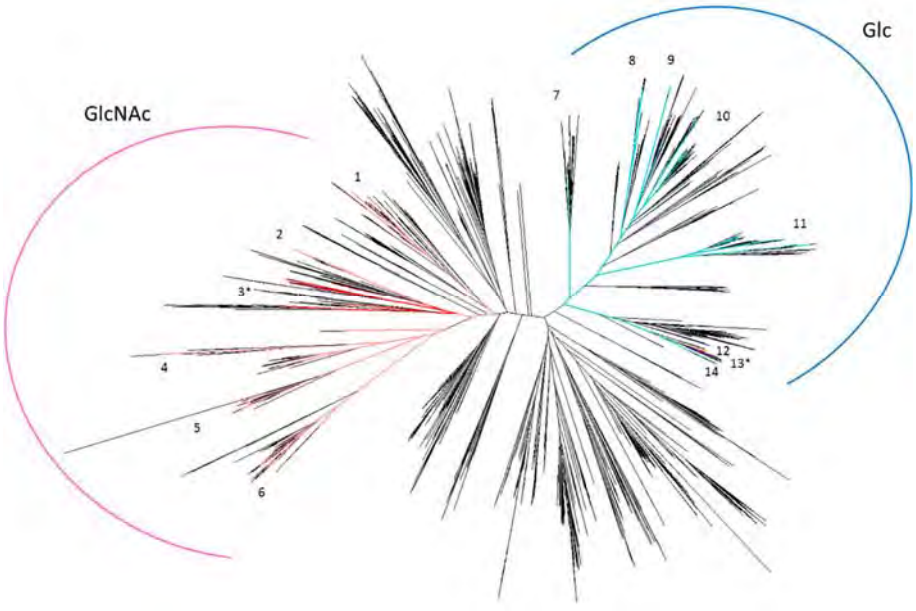


Figura 101. Árbol filogenético de la familia GT2, subgrupo P, proteínas con mecanismo catalítico procesivo, construido a partir del alineamiento del dominio GTA consenso entero. Las ramas coloreadas indican secuencias cuyo aceptor es conocido. Dos grandes grupos de aceptores: GlcNAc (rojo), Glc (azul). La actividad enzimática específica de estas viene identificada por los números:1: Poli-b-1,6-N-acetilglucosamina sintasa. (Bacteria). 2: N-acetilglucosaminiltransferasa. (Bacteria). 3: Hialurano sintasa. (Bacteria, Metazoa, Fungi). 4: Quitina sintasa. (Fungi). 5: Quitina sintasa. (Metazoa). 6: Quitina sintasa. (Fungi). 7: b-1,3-glucanosiltransglucosidasa. B-1-3-glucanosintasa cíclica. (Bacteria). 8: Celulosa sintasa. (Fungi). 9: Celulosa sintasa. (Bacteria). 10: Celulosa sintasa. (Bacteria). 11: Plantas. Celulosa sintasa. (Metaphyta). 12: b-1,4-manano sintasa. 13: Glucomanano sintasa. (Metaphyta). 14: Xiloglucano sintasa. (Metaphyta). *La familia 3 puede aceptar también GlcA. La familia 13 supuestamente solo acepta Man.

Comparando el alineamiento en la RV de algunas de las proteínas incluidas en varios taxones, como por ejemplo las hialuronano sintasa, se puede comprobar que guardan un alto grado de homología (figura 102), con posiciones bien conservadas aun cuando se comparan diferentes reinos. Esta homología va disminuyendo a medida que se incorporan todas las secuencias incluidas en la rama, pero lo hace de forma equitativa en todos los reinos del grupo y aun se observan ciertas posiciones conservadas, evidenciadas en la representación en forma de logo consenso del alineamiento (figura 103).

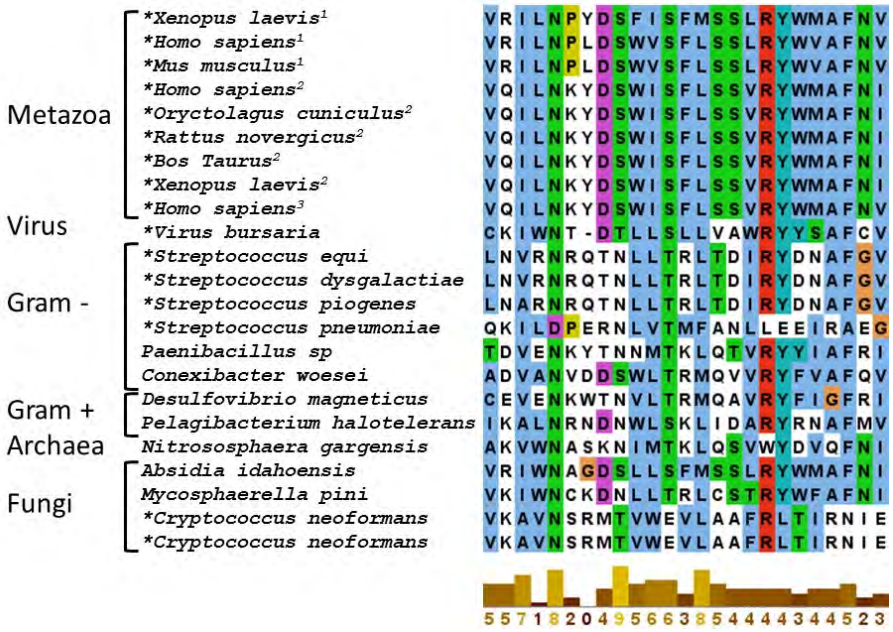


Figura 102. Alineamiento de la RV de proteínas representativas de la Hialurano sintasa. Se señalan con un * aquellas proteínas con aceptor conocido mediante evidencia experimental. Los superíndices identifican las isoformas 1, 2 y 3 de la Hialurano sintasa. Algunas posiciones están conservadas aún entre diferentes reinos.

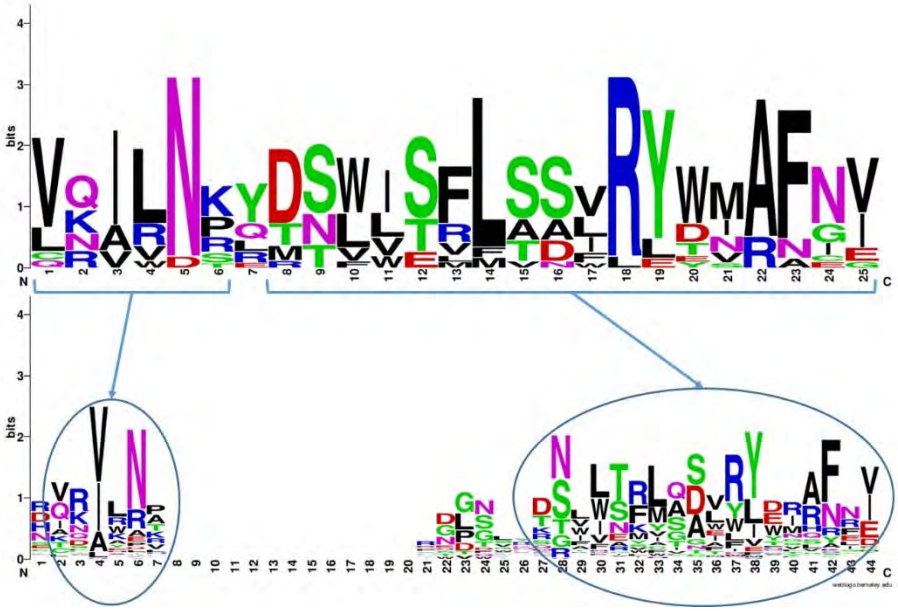


Figura 103. Representación en forma de Logo consenso del alineamiento de la RV de las proteínas Hialurano sintasa. Arriba, solo proteínas con aceptor conocido mediante evidencia experimental (16 secuencias). Abajo, todas las proteínas agrupadas en la misma rama por el árbol filogenético (97 secuencias). Son visibles dos regiones conservadas, a pesar de la adición de proteínas con aceptor desconocido (cuyas inserciones ocasionan el *gap* visible).

La presencia de patrones conservados de secuencia se puede observar también en otros grupos de proteínas como las celulosa sintasa (figura 104 y 105).

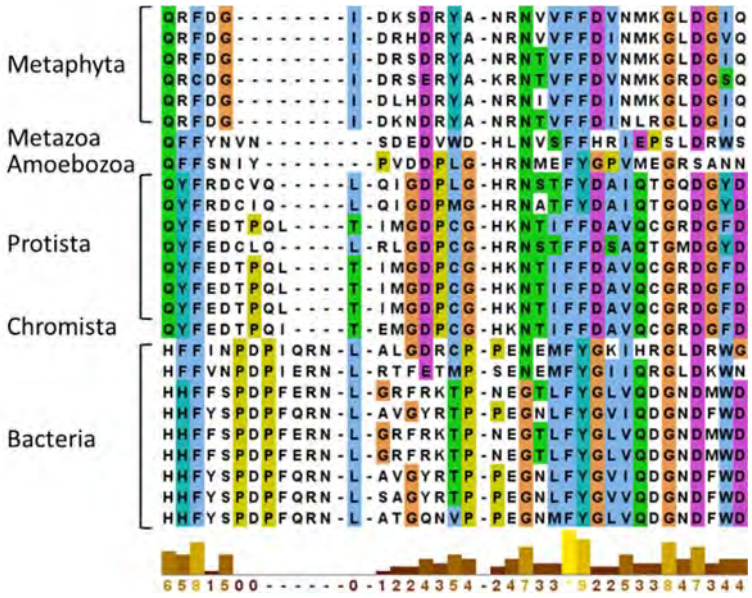


Figura 104. Alineamiento de la RV de proteínas con aceptor conocido de la celulosa sintasa.

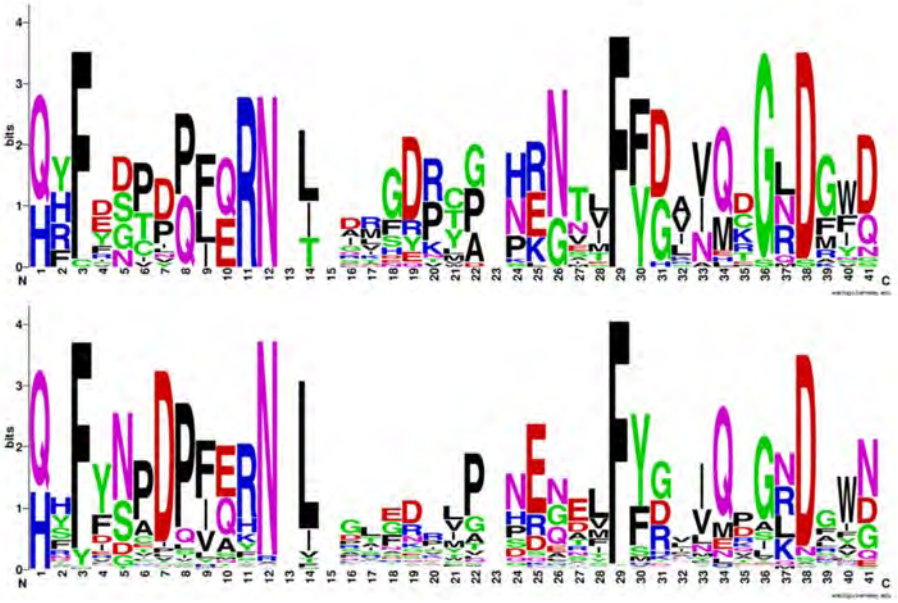


Figura 105. Logo para la RV de proteínas celulosa sintasa. Arriba, solo proteínas con aceptor conocido mediante evidencia experimental (25 secuencias, las de la figura 100). Abajo, todas las proteínas agrupadas en la misma rama por el árbol filogenético (254 secuencias). Son visibles regiones conservadas, a pesar de la adición de proteínas con aceptor desconocido.

El único aceptor dentro del grupo Glc que es diferente, el de la proteína b-1,4-manano sintasa, se encuentra dentro de un grupo de plantas cuya RV está muy relacionada con otras proteínas del mismo taxón, que también tienen Glc como aceptor (xiloglucano sintasa) o incluso son capaces de ceder el ligando a ambos aceptores Glc o Man (b-1,4-manano sintasa / glucomanano sintasa). La gran homología de la xiloglucano sintasa 4 (figura 102) con las otras dos proteínas y que todas tengan la Glc como aceptor, permite incluir estas 3 RV en un mismo grupo de Glc como aceptor, pero con mayor posibilidad de transferir a Man que el resto.

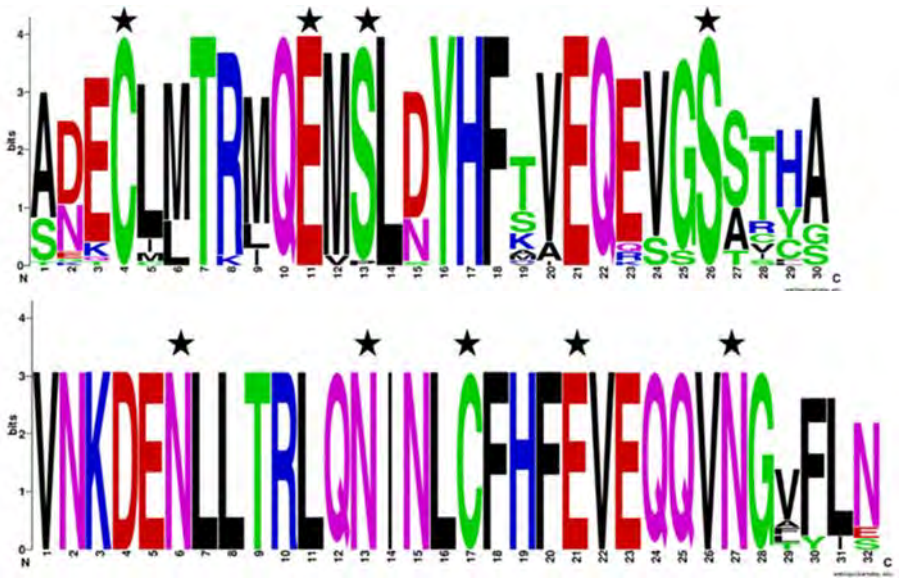


Figura 106. Arriba, logos para las proteínas b-1,4-manano sintasa/Glucomanano sintasa que transfieren a manosa o glucosa y b-1,4-manano sintasa que solo transfiere a Man (40 secuencias en total). Ambas proteínas tienen un alto grado de homología que las hace indistinguibles al nivel de la RV. **Abajo**, logo de la RV para proteínas Xiloglucano sintasa, que transfieren a Glc (11 secuencias). Por sus diferentes características en posiciones homólogas, las estrellas podrían identificar los residuos de cada grupo de proteínas que definen la especificidad para Glc o también Man como aceptor.

Existe pues una clara diferencia entre regiones variables de grupos de proteínas de la familia GT2 con actividad catalítica distinta (diferente aceptor). La RV contiene residuos muy conservados en grupos de proteínas que comparten aceptores similares. Sin embargo, como se vio en la tabla 16, los residuos responsables de la ubicación del aceptor, no están todos concentrados en la RV, por ejemplo, para la única GT2_P cristalizada, la estructura 4HG6 de una celulosa sintasa, el residuo que principalmente posiciona la Glc aceptora es el W383[137], situado en la hélice 6 de la topología consenso, aunque otros residuos alrededor del aceptor estén todos en la RV. No podemos concluir que la especificidad por el aceptor esté exclusivamente reducida a la RV.

Si toda la secuencia de la proteína se ve afectada de igual modo por la presión evolutiva, el alineamiento de diferentes fragmentos de la secuencia debería rendir árboles similares. Se comparó entonces la evolución de la RV con el resto de la secuencia, mediante la realización de varios árboles filogenéticos: con los alineamientos de las secuencias completas (dominio GTA consenso), sin incluir la RV y solo con la RV.

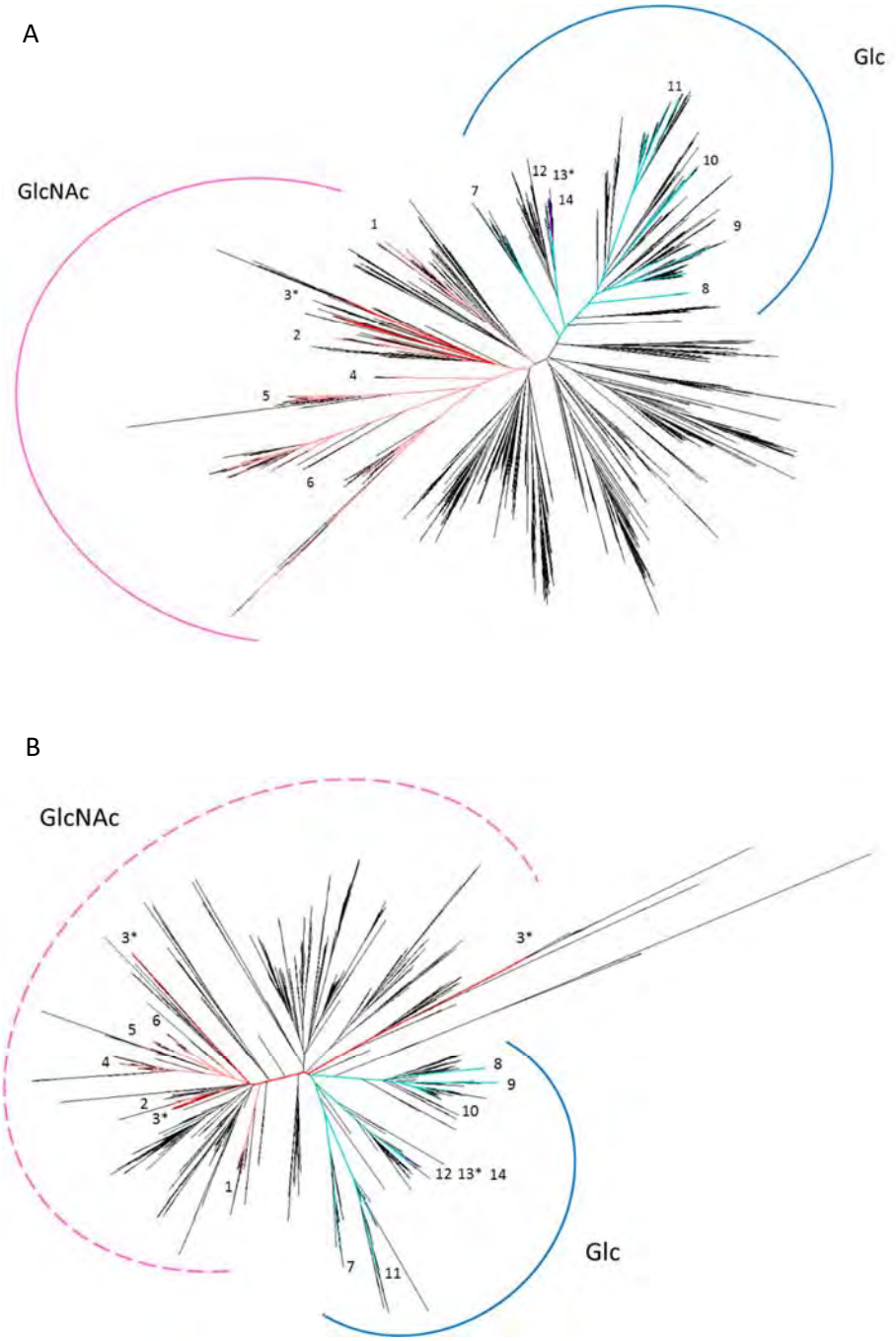


Figura 107. Árbol filogenético de la familia GT2, subgrupo P. **A. La RV ha sido eliminada. B. Solo la RV**. Mismo esquema de colores y numeración que en la figura 101.

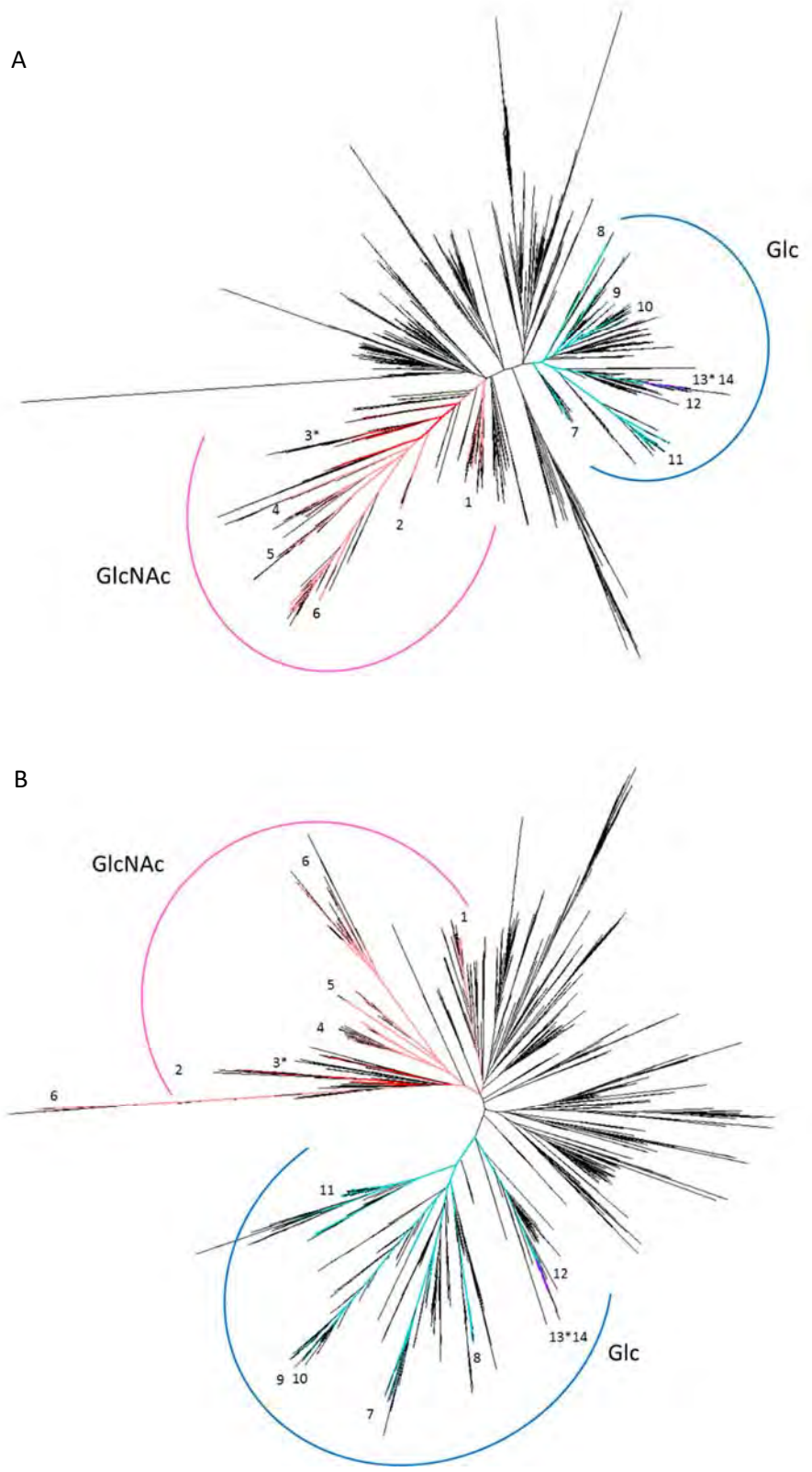


Figura 108. Árbol filogenético de la familia GT2, subgrupo P. **A.** Segmento α3β4DXDβ4'α4β5**. B.** Segmento β6α6β7. Mismo esquema de colores y numeración que en la figura 101.

Para la familia GT2_P, como ya se ha visto, los diferentes aceptores se encuentran bien separados al utilizar todo el dominio GTA (figura 101). Este resultado se repite al extraer la RV de la secuencia (figura 107 A), sin embargo, al usar solo la RV el resultado es más confuso (figura 107 B) y si bien los grupos (GlcNAc, Glc) siguen estando bien separados, dentro de estos se confunden algunos grupos de proteínas, como para la hialuronidasa, que no ocurren en los árboles previos. Se esperaba este resultado para la RV, ya que su secuencia es mucho más corta y, por tanto, sensible a diferencias que para las secuencias completas o sin RV, aun así, es posible ver que en líneas generales los grupos de aceptores siguen estando separados. Para comprobar cuánto de este último resultado era debido a la longitud de la secuencia utilizada en los alineamientos, se realizaron nuevos árboles tomando diferentes zonas de la secuencia, la comprendida entre los elementos conservados α3-β5 y entre β6-β7 y extrayendo una longitud de residuos idéntica a la de su RV correspondiente. El resultado es una agrupación de aceptores más consistente que la de la RV (figura 108 A y B).

Un resultado como este implica que la especificidad por el aceptor se encuentra distribuida en la región conservada de la topología consenso, y que pequeños segmentos de la misma todavía conservan esta información. La RV parece poseer una evolución similar al resto de la secuencia, con una cierta mayor variabilidad, que precisamente añade ruido al agrupamiento por aceptores, aunque ciertas posiciones se encuentran muy conservadas. La RV sería, por tanto, la región de la secuencia menos indicada para atribuir la especificidad por el aceptor. ¿Por qué entonces se encuentran precisamente aquí posicionados los aceptores en las proteínas estudiadas? ¿Por qué esa mayor variabilidad? Veamos qué sucede en las siguientes familias.

3.2 Familia GT7

La familia GT7 contiene proteínas fundamentalmente encontradas en el Reino Metazoa, aunque también existen representantes en otros dominios; se incluyen proteínas como condroitín sulfato sintasa, muy importante para el desarrollo de la matriz extracelular o β1-4-galactosiltransferasa (GalT) que produce proteoglicanos. En esta familia, (figura 109) los elementos estructurales de la región conservada que circundan la RV, poseen mayor homología entre las diferentes proteínas que la propia RV, donde esta homología solo está presente entre proteínas con un mismo aceptor. A pesar de esta variabilidad interproteica para la RV, los residuos que interactúan directamente con el aceptor suelen estar conservados, como aquí es el caso de una tirosina o fenilalanina en las posiciones marcadas en rojo en el alineamiento de la figura 109. En el alineamiento se puede observar que estos residuos se encuentran en la RV. La relación entre homología de secuencia y aceptor también es apreciable a nivel de estructura de la RV, todos los cristales para las proteínas que tienen Glc o GlcNAc como aceptor

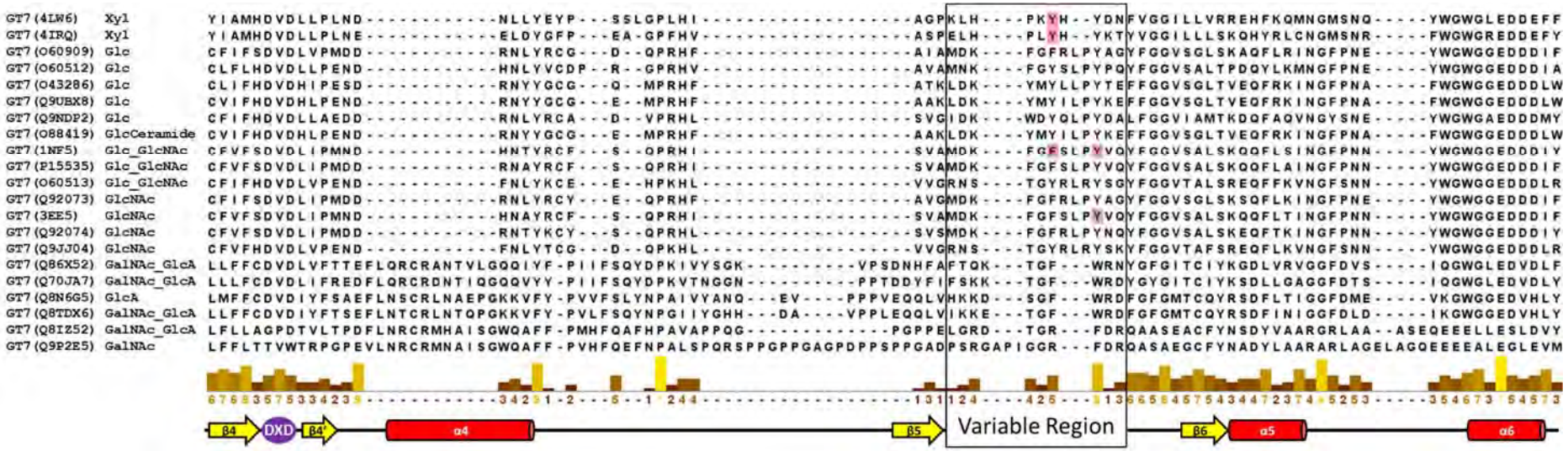


Figura 109. Alineamiento de proteínas GT7 con aceptor conocido mediante evidencia experimental. Se muestra el código UniProt o PDB cuando existe. Cada entrada muestra el aceptor. La RV está enmarcada por un recuadro, los residuos que en la estructura contactan directamente con el aceptor, están remarcados en rojo. Bajo el alineamiento se encuentra la puntuación por conservación y los elementos estructurales de la topología consenso que circundan la RV. Xyl: Xilosa, Glc: Glucosa, GlcCeramide: Glucosa unida a una ceramida, GlcNAc: N-Acetil-glucosamina, GalNAc: N-Acetil-galactosamina, GlcA: glucurónico.

tienen la misma estructura para la RV, sin embargo, esta es ligeramente diferente para la xilosa (figura 105). Posiblemente la estructura para la RV de aquellas que transfieren a GlcA o GalNAc sea similar entre ellas y diferente de las de Xyl o Glc/GlcNAc. Todas las proteínas que transfieren a xilosa, glucosa o N-acetilglucosamina tienen como dador UDP-α-D-galactosa (familia de las B4GALT), mientras que el resto tienen como dador y/o aceptor UDP-N-acetil-α-D-galactosamina y/o β-D-glucurónico (familia de las Condroitín sulfato sintasa).

Nota: A pesar de que por las estructuras cristalizadas parece que GT7 es la única familia que carece de la hélice α4 en la topología consenso, es interesante que, una predicción de PsiPred para proteínas no cristalizadas como Q9P2E5 o Q8IZ52 (últimas secuencias en la figura 105), asigne una alta probabilidad de α hélice para esa misma región de la secuencia, acorde con nuestra topología consenso.

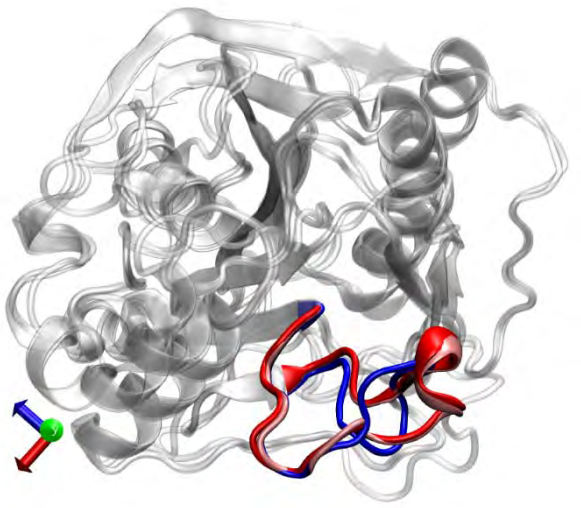


Figura 110. Superposición de tres estructuras de la familia GT7. Azul: 4LW6 (Aceptor: xilosa), Rojo: 1NF5 (Aceptor: Glc/GlcNAc), Rosa: 3EE5 (Aceptor: GlcNAc). 1NF5 y 3EE5 que comparten aceptor tienen la misma estructura de RV mientras que 4LW6 que transfiere a un aceptor diferente tiene una RV ligeramente diferente.

Esta familia está compuesta de 191 secuencias. Al realizar un árbol filogenético con ellas se encuentran de nuevo bien separados los aceptores (figura 110). Proteínas que transfieren solo a GlcA (condroitín sulfato N-acetilgalactosaminiltransferasa 2), o que pueden hacerlo a ambas (condroitín sulfato sintasa 1) se encuentran agrupadas. Las que transfieren a xilosa se hallan separadas en otra rama y por último y bien separado de estos está el grupo de las que tienen como aceptor Glc, GlcNAc o pueden transferir a ambas moléculas. En este caso los diferentes taxones no están tan segregados como en GT2_P y para cada grupo se mezclan representantes de cada uno (cordatha, artropoda, nematoda, rotifera …). Es interesante observar, como para el grupo de las B4GALT, los diferentes aceptores (Xyl, Glc y GlcNAc) son bien segregados, a pesar de que todos poseen el mismo dador UDP-α-D-galactosa, indicando que la secuencia y posiblemente su estructura ha evolucionado hacia la especificidad de un determinado aceptor, por encima del dador.

Cuando se realiza un nuevo árbol sin la RV, los grupos de aceptores siguen separándose bien (figura 112) rindiendo un resultado similar al del domino completo GTA (figura 111).

En un árbol realizado solo con la RV (figura 113 A), siguen viéndose los grupos separados pero algunas secuencias, como las que transfieren indistintamente a Glc/GlcNAc han sido desagrupadas. Si se utiliza el segmento α3-β5 para un nuevo árbol (figura 113 B), los grupos vuelven a estar bien definidos, con resultados similares a los de los árboles realizados con todo el dominio GTA o sin la RV. Puesto que la longitud de las secuencias en este árbol es idéntica al de la RV, esta diferencia ha de deberse a la mayor variabilidad de la RV que para otras zonas de la proteína, como ocurría en la familia GT2_P.

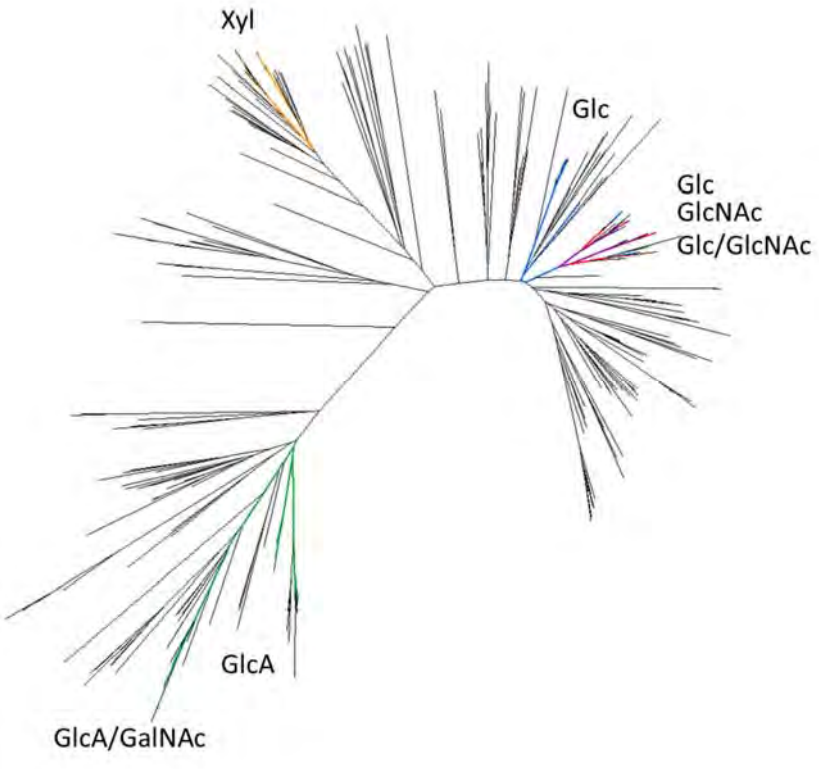


Figura 111. Árbol filogenético de la familia GT7. Construido a partir del alineamiento de secuencias del dominio GT7 consenso completo.

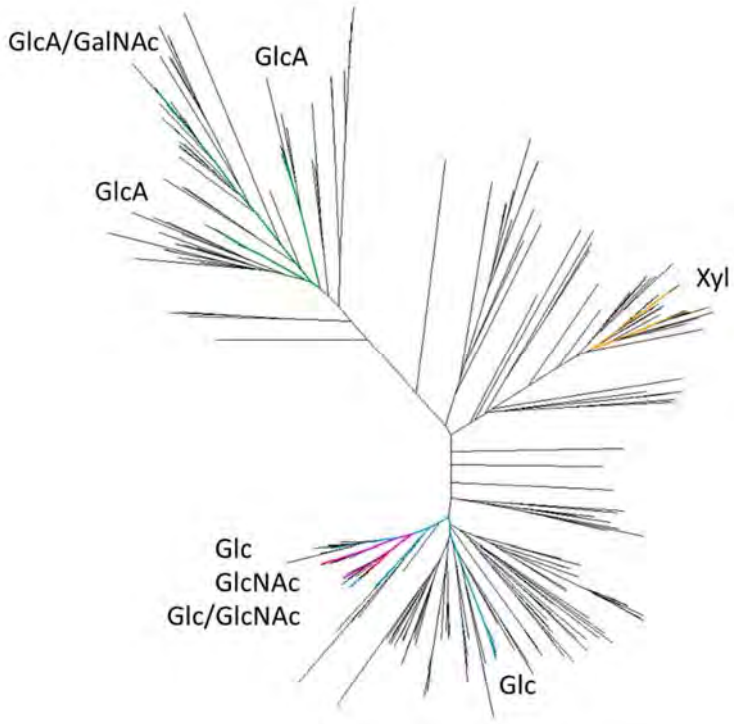


Figura 112. Árbol filogenético de la familia GT7 con la RV extraída.

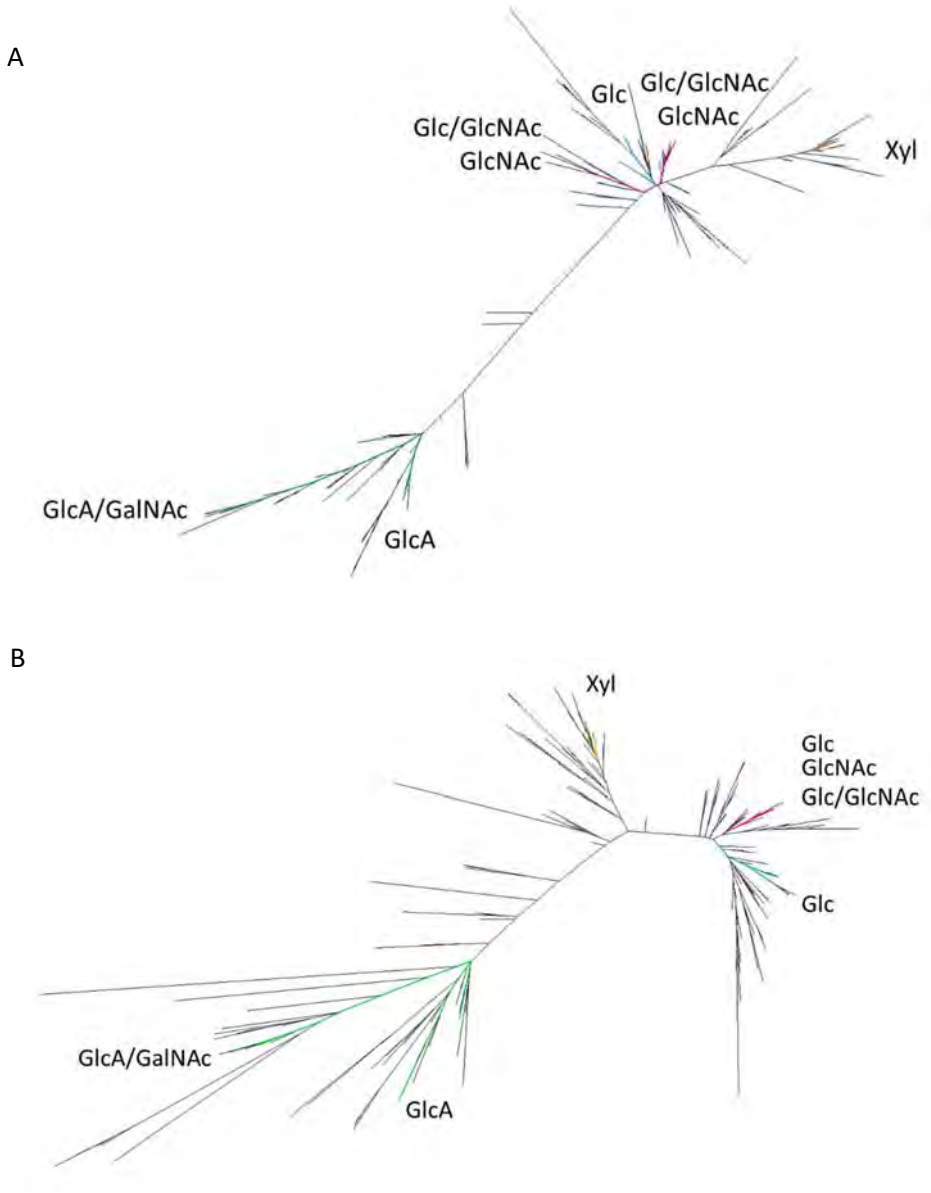


Figura 113. Árbol filogenético de la familia GT7. **A. Solo la RV. B. Segmento α3β4DXDβ4'α4β5.**

Parece confirmarse que la especificad por el aceptor viene definida en la zona conservada de la topología consenso. El plegamiento global de la proteína, determinada por la secuencia, establece el tipo de aceptor. La RV vuelve a mostrarse como una sección de la secuencia en donde la presión evolutiva es más laxa, permitiendo una mayor variabilidad que parece no correlacionar con la especificidad con el aceptor, como sí lo hace el resto de la secuencia aun cuando se haga por segmentos de la misma longitud de la RV. Sin embargo en esta región, se siguen observando posiciones clave, que se encuentran bien conservadas entre proteínas con aceptores similares y que difieren de proteínas con otros aceptores.

3.3 Familia GT8

La familia GT8 está formada por unas 6300 proteínas con diferentes funciones, con representantes en los diferentes dominios de la vida. Posee una estrategia catalítica diferente a las familias GT2_P y GT7. En las GT8 varios bucles externos a la RV ejercen la responsabilidad del posicionamiento del aceptor (ej. glicogenina), así la RV ha perdido parte de su relevancia en favor de otras regiones. Por esto, para algunas proteínas GT8 la RV casi ha desaparecido (como la glicogenina, xilano-glucuroniltransferasa 2 o la inositol fosforilceramida glucuroniltransferasa 1). Sin embargo, existen otras que dentro de esta familia sí que tienen una RV de considerable longitud, que de nuevo conservan cierta homología cuando comparten aceptor. Del mismo modo que sucede con la familia GT7, puede verse que la homología en los elementos de la región conservada es mayor entre las diferentes proteínas que para sus RV y que habiendo perdido esta cierta importancia, aún conserva residuos que interactúan con el aceptor (Pro, Asp en la glicogenina y Phe, Val en LgtC en las posiciones señaladas en rojo en la figura 114).

Al visualizar el alineamiento de algunos representantes de los grupos con aceptores conocidos (figura 114) se puede observar el pequeño tamaño de algunas de las RV de estas proteínas, así como grandes diferencias entre algunas RV que comparten un mismo aceptor, señal de la pérdida de relevancia catalítica de esta región en la familia GT8. Sin embargo, los residuos cercanos al aceptor en las proteínas cristalizadas todavía se encuentran en la RV. También es posible comparar el alto grado de conservación de los elementos colindantes a la RV, a pesar de que ninguno de ellos participa activamente en el reconocimiento del dador o aceptor, comparada con la propia RV donde esta homología solo se vislumbra entre proteínas con el mismo aceptor, aunque menor que en otras familias.

Esta reducción de la RV en favor de otras zonas de la región conservada parece no ser exclusiva de la familia GT8. Así, la hélice α que en la glicogenina contiene el residuo clave en el posicionamiento del aceptor (Tyr195)[138] superpone con la hélice, en GT13 (GnTI) que contiene un residuo con un rol similar (Ser233)[139]. Dicha posición también superpone con la estructura que, en GT15 (Kre2p/Mnt1p)[140], contiene a la Tyr391 que está relacionada con cambios conformacionales debido a la unión del aceptor. Aun así, es interesante decir que en esta familia GT15, en las proteínas cristalizadas, los residuos y posiciones conservadas claves para la unión del aceptor, Glu279 y Tyr280, se encuentran en la RV.

Debido al pequeño tamaño de algunas de las RV para esta familia, no se han realizado otros árboles filogenéticos para la RV o sin ella.

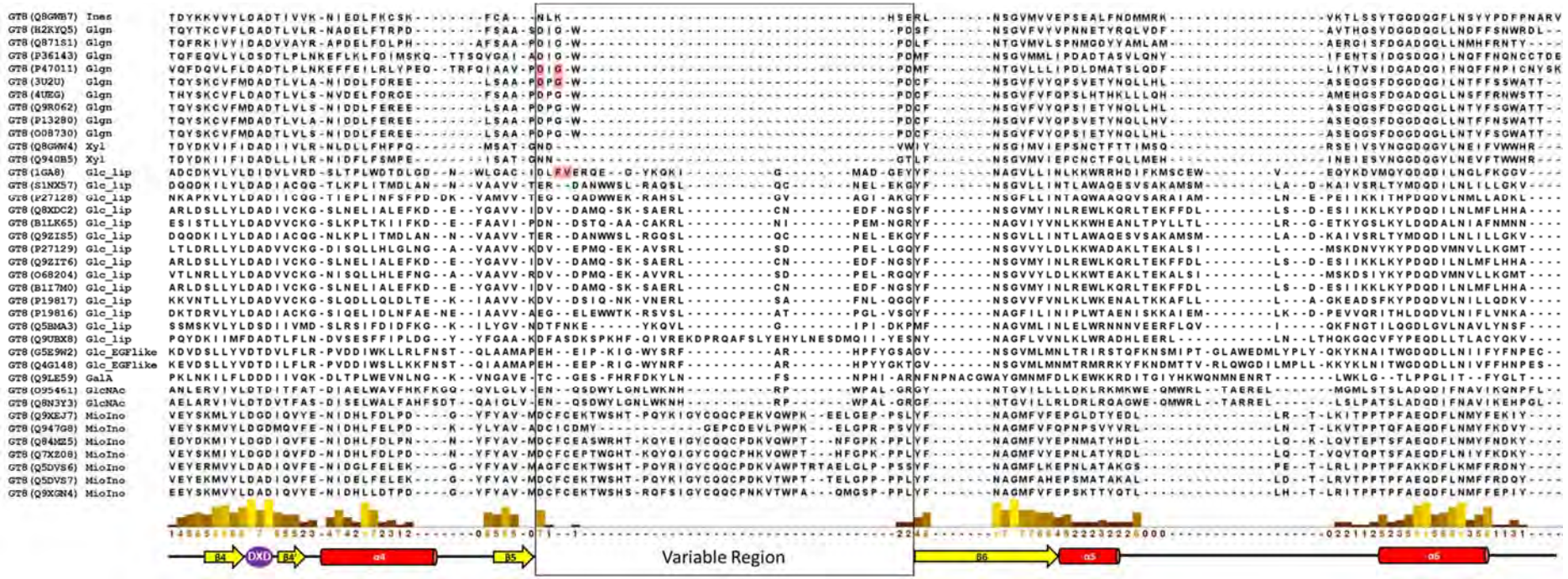


Figura 114. Alineamiento de proteínas GT8 con aceptor conocido mediante evidencia experimental. Se muestran los códigos PDB cuando existen, si no el código UniProt. Para cada entrada se muestra además el aceptor. La RV está enmarcada por un recuadro y los residuos en su interior que, en la estructura contactan directamente con el aceptor, remarcados en rojo. Bajo el alineamiento se encuentra la puntuación por conservación y los elementos estructurales de la topología consenso que circundan la RV.

Aceptores: Ines: Inespecífico, Glgn (Glicogenina): Glucosa y tirosina, Xyl: Xilosa, Glc_lip: Glucosa unida a un lípido, Glc_EGFlike: Glucosa unida a EGF_like, GalA: Glucurónico, GlcNAc: N-Acetilglucosamina, MioIno: Mioinositol.

La extensa familia GT8 está subdividida en numerosos subgrupos homólogos (A a M, datos no publicados facilitados por CAZY). Se observa en el árbol filogenético creado con las proteínas que forman esta familia, que los diferentes aceptores se encuentran también correctamente agrupados y separados (figura 115). No todos los subgrupos tienen identificados un aceptor, pero para los que sí lo tienen se puede encontrar cierta correlación entre el tipo de aceptor y el taxón al que pertenecen. El subgrupo H está formado por proteínas bacterianas que transfieren a Glc unida a un lipopolisacárido. Es importante señalar que todas las proteínas de este subgrupo H, comparten un mismo aceptor (unidad de glucosa, Glc) pero no un mismo dador; la lipopolisacárido 1,2-glucosiltransferasa (B1LK65) transfiere Glc y la lipopolisacárido 1,3-galactosiltransferasa (P27128) transfiere Gal, pero ambas lo hacen a una Glc unida a un lipopolisacárido[n].

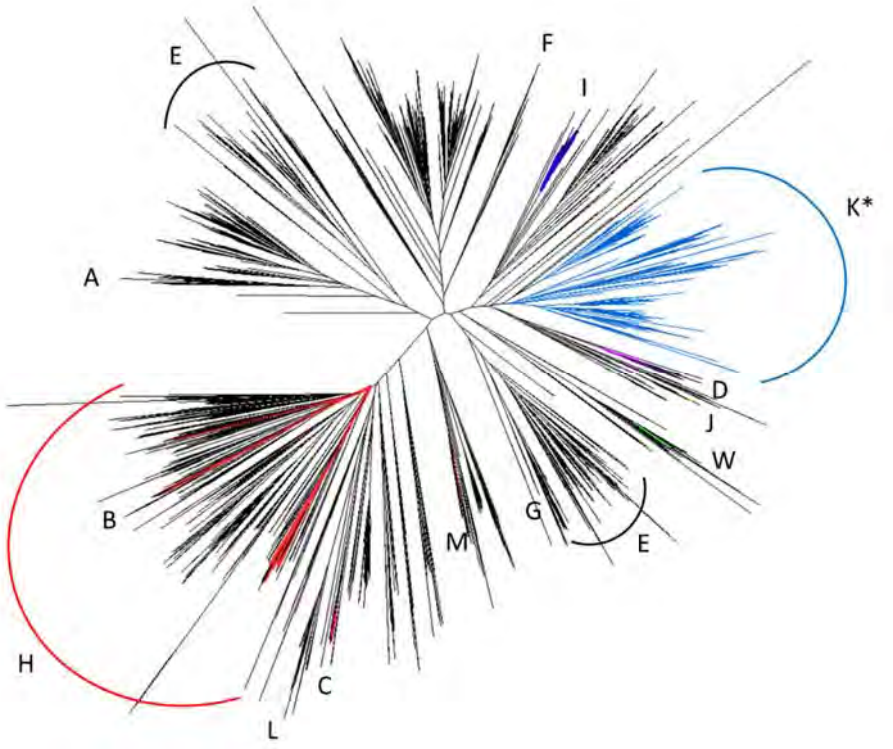


Figura 115. Árbol filogenético de la familia GT8. Dominio GTA.

A: ? (levaduras), B: Glc_lipopoli (Bacteria), C: GlcNAc (Metazoa), D: Xyl (Metaphyta), E: ? (Fungi), F: ? (Insecta), G: ? (Levadures), H: Glc_lipopoli (Bacteria), I: Mioino (Metaphyta), J: Ines (plantas), K: Glgn (Fungi, Metazoa), L: GalA (Metaphyta), M: Glc_EGFlike (Metazoa).

La familia B está formada solo por el género Helicobacter. La familia H contiene un grupo heterogéneo de bacterias.

El subgrupo M, con representantes Metazoa, transfiere también a Glc, pero esta ha de estar unida a repeticiones EGF-like y se encuentra bien separado del subgrupo B. Entre ellos están los

[n] En la misma rama que este existe otro subgrupo, el B, también de proteínas bacterianas y que transfiere a Glc unida a un lipopolisacárido pero cuyo mecanismo está clasificado como *inverting* en oposición al resto de GT8 que tienen un mecanismo tipo *retaining*.

subgrupos L (Metazoa) y C (Metaphyta) que transfieren a GalA y a GlcNAc respectivamente. Los subgrupos J y D incluyen proteínas de plantas; en el subgrupo J se encuentran proteínas que pueden aceptar múltiples aceptores mientras que el D transfiere aparentemente solo a Xyl. Ambos poseen las RV más cortas de las proteínas estudiadas (4-6 residuos). El subgrupo I, también de plantas, transfiere a mioinositol y en el subgrupo K se encuentran las glicogeninas que transfieren unidades de Glc, formando un polisacárido de 6-10 Glc[140,141] utilizándose a sí misma como aceptor inicial y con representantes de los reinos Fungi y Metazoa. Por último, no existen aceptores conocidos para los subgrupos E, F o G y que presumiblemente no serán ninguno de los anteriores citados por estar filogenéticamente separados del resto, si bien E y G podrían compartirlo dada su relación filogenética.

Como en las anteriores familias, los diferentes nodos parecen agrupar mejor los grupos por similitud del aceptor, antes incluso que por taxón o dador. Aunque esta familia evidencia que la RV puede llegar a perderse y ser sustituida por otros elementos, también confirma que es el aceptor quién más presiona en la evolución de la superfamilia de las GTAs.

3.4 Familias GT55, GT78 y GT81

Un ejemplo paradigmático de esta relación RV-aceptor son las familias GT55, GT78 y GT81. La familia GT81 se subdivide en los grupos A, B, C y otras secuencias que no han podido agruparse por homología (datos no publicados, facilitados por CAZY). GT55, GT81_A y GT81_B incluyen proteínas cuyo aceptor obligado es el fosfoglicerato y nunca el glicerato (EC: 2.4.1.217, 2.4.1.266), mientras que GT78 y GT81_C incluyen proteínas cuyo aceptor es el glicerato aunque se ha comprobado que bajo ciertas condiciones también podrían aceptar fosfoglicerato[81] (EC: 2.4.1.268). Si se comparan las estructuras cristalizadas existentes para cada una de estas familias (GT55, GT78 y GT81_B) puede verse como sus RV superponen perfectamente (figura 114), con la diferencia de un bucle no resuelto en GT55 y GT81 que no existe en GT78. En el alineamiento (figura 113) se puede observar como entre los residuos que interactúan con el aceptor se encuentra una Arg extraordinariamente conservada en GT55 y GT81_B, también en GT81_A, mientras que GT78 y GT81_C tienen en su lugar otro residuo en la misma posición.

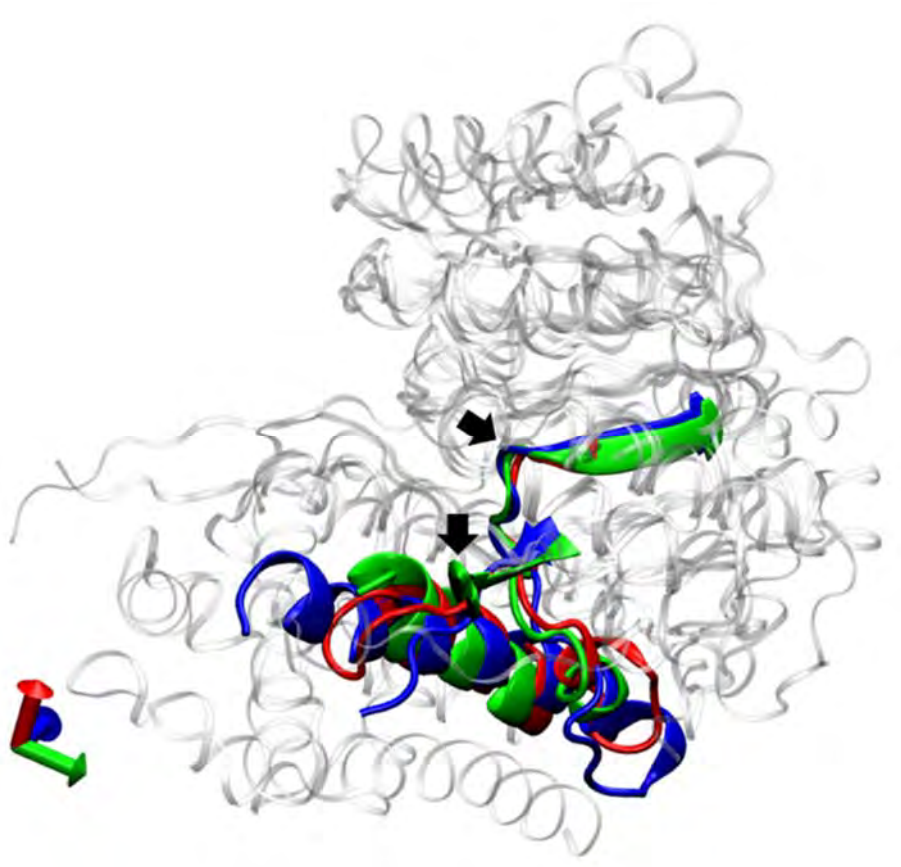


Figura 116. Superposición de las estructuras 2WVL (GT55) azul, 2BO6 (GT78) rojo y 4DEC (GT81) verde. Las flechas negras marcan el comienzo y final de la RV. GT55 y GT81 poseen un bucle no cristalizado que no existe en GT78, después de ese bucle las tres estructuras comparten una hélice α de longitud similar.

Por la similitud estructural de sus RV se han comparado estas 3 familias GTAs. Cuando se visualiza el alineamiento en la zona de la RV (figura 117) se puede comprobar que GT55 y GT81 (subgrupos A y B) comparten un núcleo central de la RV conservado, precisamente los que se encuentran interaccionando con el aceptor PGA, este es diferente en las familias GT78 y GT81 subgrupo C que transfieren a GA. El residuo que interactúa directamente con el PGA es la Arg de ese núcleo central, que ha sido sustituido en GT81_C por una Thr (polar no ionizable) y en GT78 por una Met (apolar), siendo esta la causa de la preferencia por GA en estos dos grupos de proteínas[81] y estableciendo una relación directa de la RV con el aceptor, a través de al menos una posición clave ubicada en esta región.

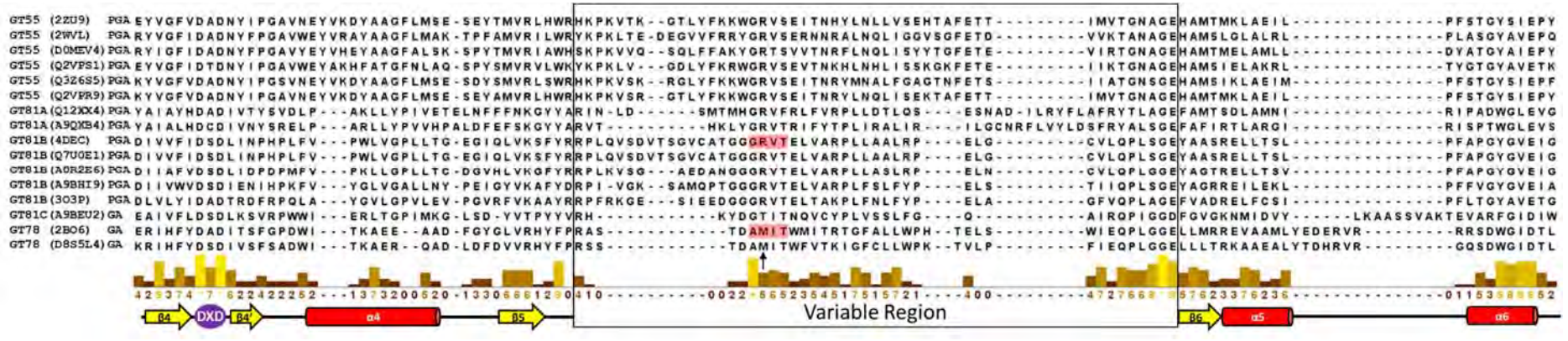


Figura 117. Alineamiento de proteínas GT55, GT78 y GT81 con aceptor conocido mediante evidencia experimental. Se muestran los códigos PDB cuando existen, si no el código UniProt. Para cada entrada se muestra además el aceptor. La RV está enmarcada por un recuadro y los residuos en su interior, que contactan con el aceptor, remarcados en rojo. Bajo el alineamiento se encuentra la puntuación por conservación y los elementos estructurales de la topología consenso que circundan la RV. La pequeña flecha negra señala la posición del residuo que discrimina entre los aceptores glicerato o fosfoglicerato.

Un árbol filogenético de estas tres familias de proteínas (figura 118) separa correctamente tanto las familias GT55, GT78 y GT81 como a los subgrupos de GT81, pero sitúa filogenéticamente relacionados de forma clara a las familias GT78 y el subgrupo C de la GT81, ambas con el aceptor GA, aunque con un valor de *bootstrap* bajo (0,35 en la figura 118).

Si se extrae de las secuencias la RV y vuelve a realizarse otro árbol filogenético, se obtiene el mismo resultado, pero elevando el valor de *bootstrap* que relaciona GT78 con GT81_C, a un valor de 0,97 (figura 119), que indica que la RV es ciertamente un generador de ruido y variabilidad, al menos entre estos dos grupos.

Un nuevo árbol utilizando solo la RV reporta de nuevo los grupos claramente separados (figura 120), formando GT78 un grupo aparte que sin embargo se encuentra más cercano a GT81_B que a GT81_C, lo que se interpreta como que más allá de los residuos directamente implicados en la interacción con el aceptor, el resto de los incluidos en la RV dan lugar a una variabilidad mayor, propia de esta región de las GTAs.

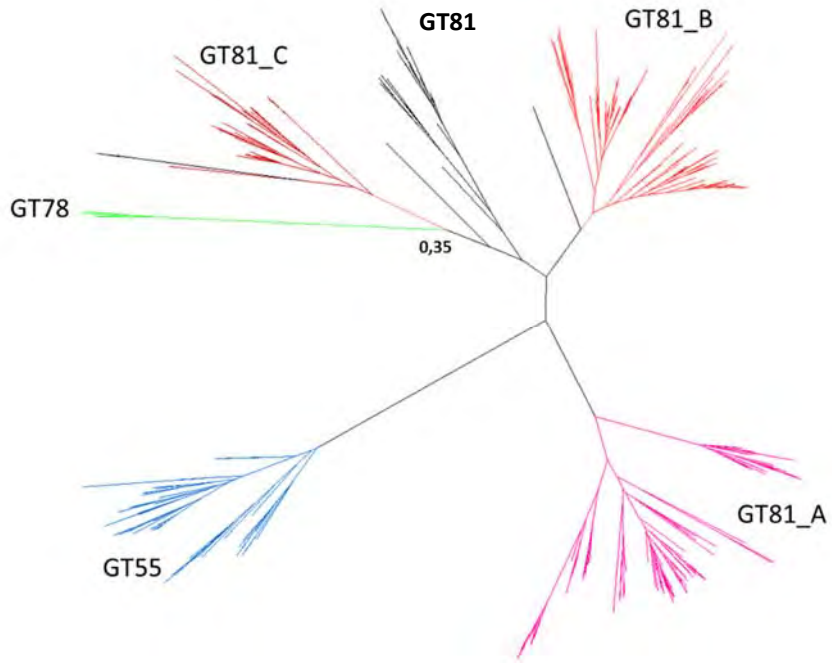


Figura 118. Árbol filogenético de las familias GT55, GT78 y GT81. Dominio GTA. Se muestra el valor de *bootstrap* para el nodo que enlaza las familias GT78 y GT81_C.

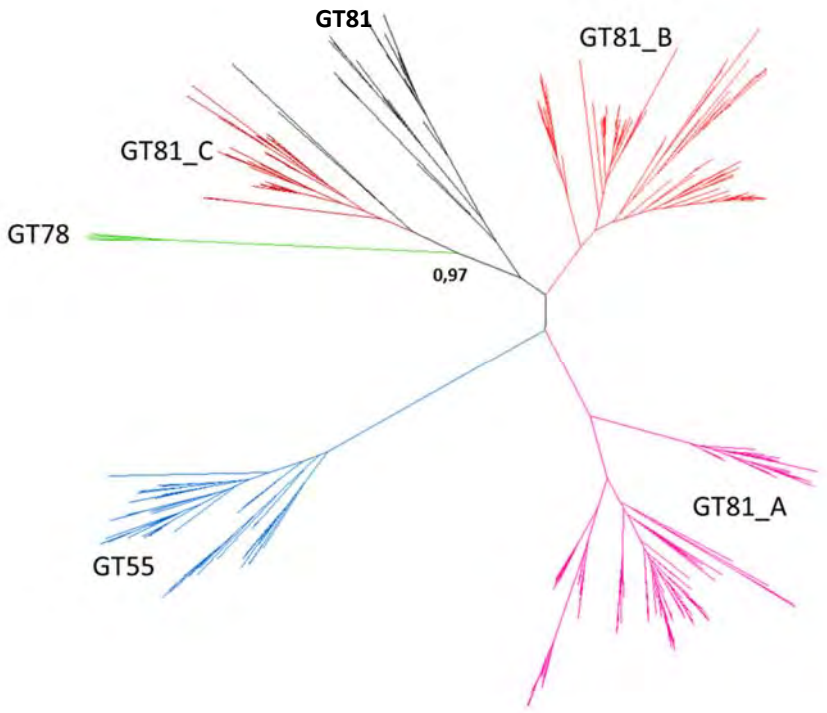


Figura 119. Árbol filogenético de las familias GT55, GT78 y GT81. **RV extraída**. El valor de *bootstrap* para el nodo que enlaza las familias GT78 y GT81_C aumenta considerablemente.

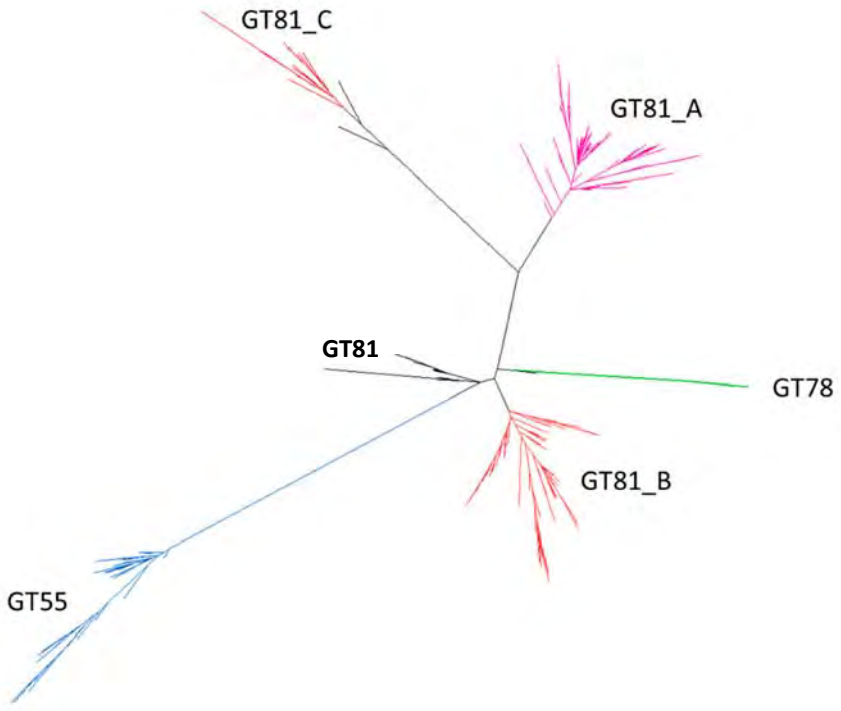


Figura 120. Árbol filogenético de las familias GT55, GT78 y GT81. **Solo la RV.**

Para este grupo de familias, la RV es un claro elemento de variabilidad entre ellas, que confirma las conclusiones en las familias anteriores, de que la RV sufre una presión evolutiva menor que el resto de la secuencia. Siempre que en esta región se encuentren conservados y bien posicionados los residuos que directamente interaccionan con el aceptor, el resto de la RV es susceptible de mayor índice de mutación que el resto de la proteína.

Hemos estudiado 5 grupos distintos de familias de GTAs que comprenden unas 13000 secuencias. Mediante el uso de alineamientos, árboles filogenéticos y la inspección visual de las estructuras cristalizadas encontradas en estas familias, concluimos que la especificidad por el aceptor está definida por la zona conservada de la topología consenso en las GTAs. El plegamiento de esta zona –toda la proteína menos la RV y que viene definido por la secuencia–, genera una estructura que es específica para un determinado aceptor.

La RV posee siempre algún residuo clave para la interacción con el aceptor, en algunos casos todos, aunque haya otros, como los observados en la familia GT8 en los que la RV ha perdido importancia al ser sustituida por otros elementos a lo largo de la secuencia. Se ha visto además que la RV es una zona en la que la presión evolutiva parece menor, permitiéndose mayor variabilidad que en el resto de la secuencia. Solo unos pocos residuos de esta región suelen ser los responsables de la interacción con el aceptor, permitiéndose el resto de la secuencia en esta zona más libertad de mutación que para el resto de la proteína.

¿Por qué esta mayor variabilidad?

Debido al inmenso número de posibles aceptores existentes para las glicosiltransferasas, pensamos que la RV genera flexibilidad para la inclusión de nuevos aceptores. Un lugar donde solo unos pocos residuos son suficientes para interactuar con un aceptor específico, porque la posición de estos residuos viene condicionada, no por la RV en sí, sino por su ubicación en el entorno del resto de la proteína, permitiéndose así mayor variabilidad en el resto de residuos que no interaccionan directamente con el aceptor.

Creemos, que estos residuos clave en la RV que interaccionan con el aceptor, generan una "huella", con la que es posible identificar tanto la familia de la proteína como su aceptor específico.

En la siguiente sección utilizaremos la generación de modelos ocultos de Markov (HMM), a partir de las diferentes RV, para obtener perfiles que incluyan esa "huella". Probaremos que esos perfiles son válidos para usarlos a modo de *screening* en proteínas desconocidas, permitiendo identificar en estas proteínas, su pertenencia al grupo de las GTAs, su familia y subgrupo y más relevante aún, su probable aceptor. Estos resultados serían posteriormente validados a través de un alineamiento con proteínas de la familia identificada.

4. Región variable como herramienta predictiva

4.1 Perfiles de Secuencias

Creímos posible recoger la variabilidad de la zona de la secuencia estudiada, en forma de perfiles HMM específicos de cada familia y aceptor. En estos, los residuos más relevantes deberían tener el peso suficiente para singularizar cada RV en relación con su aceptor y deberían diferenciarse de cualquier otra glicosiltransferasa que contuviese otro aceptor o perteneciese a otra familia.

Se escogió un conjunto de proteínas de la base de datos de CAZy, aquellas definidas como GTAs y que cuentan con un nº EC de al menos 2.4.-.-, por lo que existe evidencia experimental de que se trata de una glicosiltransferasa. Este no es más que un sistema para reducir el enorme grupo de secuencias de proteínas GTA que contiene la base de datos CAZy, pero además permite seleccionar a todas aquellas proteínas que tienen un aceptor conocido, con una evidencia experimental empírica, ya que todas tienen un nº EC asignado. Se usaron los subgrupos por homología de cada familia realizados por CAZy (ver métodos) y se anotó, cuando es conocido, el aceptor para cada uno de ellos, de tal modo que siempre estuviesen identificados la familia CAZY, el subgrupo cuando existiese y el aceptor cuando fuera conocido.

Algunas secuencias se consideraron huérfanas pues no fue posible agruparlas por homología ni ningún grupo, ni entre ellas; de estas por ejemplo existen 22 secuencias dentro de la familia GT2. Para muchas de ellas se desconoce el aceptor, pero dado que no guardan homología y que sus RV tampoco alineaban, se utilizaron por separado e identificaron por su código UniProt (tabla 18).

Tabla 18. Grupos iniciales de secuencias GTA con EC mínimo de 2.4 agrupados por homología aceptor cuando este es conocido.

FAM	Nº EC	DESCRIPCIÓN ACEPTOR	ACRÓNIMO ACEPTOR	GRUPO	NOMBRE HMM
GT2	2.4.1.-	?	?	Z	GT2_Z_A0Q5C5
					GT2_Z_Q8KZ90
GT2	2.4.1.-	?	?	Z	GT2_Z_UndP
GT2	2.4.1.199	(1->6)-alpha-D-mannosyloligosaccharide	Man	Z	GT2_Z_Man
GT2	2.4.2.53	ditrans,octacis-undecaprenylphosphate	UndP	Z	GT2_Z_UndP
GT2	2.4.1.-	?	?	I	GT2_I
GT2	2.4.1.-	?	?	A	GT2_A
GT2	2.4.1.-	?	?	F	GT2_F_O07340
					GT2_F_O87183
GT2	2.4.1.305	N-acetyl-alpha-D-glucosaminyl-diphospho-ditrans,octacis-undecaprenol	GlcNAc	a	GT2_a_GlcNAc
GT2	2.4.1.288	beta-D-galactofuranosyl-(1->5)-beta-D-galactofuranosyl-(1->4)-alpha-L-rhamnopyranosyl-(1->3)-N-acetyl-alpha-D-glucosaminyl-diphospho-trans,octacis-decaprenol	Gal	b	GT2_b_Gal
GT2	2.4.1.-	?	?	G	GT2_G
GT2	2.4.1.-	?	?	B	GT2_B
GT2	2.4.1.-	?	?	S	GT2_S_P33697
					GT2_S_P33700
					GT2_S_Q9XBL5
GT2	2.4.1.-	?	?	C	GT2_C

FAM	Nº EC	DESCRIPCIÓN ACEPTOR	ACRÓNIMO ACEPTOR	GRUPO	NOMBRE HMM
GT2	2.4.1.-	?	?	L	GT2_L
GT2	2.4.1.-	?	?	D	GT2_D
GT2	2.4.1.-	?	?	V	GT2_V
GT2	2.4.1.117	dolichyl phosphate	DolP	V	GT2_V_DolP
GT2	2.4.1.83	dolichyl phosphate	DolP	V	GT2_V_DolP
GT2	2.4.1.17	INESPECIFIC	Ines	Y	GT2_Y_Ines
GT2	2.4.1.-	?	?	U	GT2_U
GT2	2.4.1.-	?	?	E	GT2_E
GT2	2.4.1.157	1,2-diacyl-sn-glycerol	DAG	X	GT2_X_DAG
GT2	2.4.1.-	?	?	N	GT2_N
GT2	2.4.1.-	?	?	P	GT2_P
GT2	2.4.1.12	(1,4-beta-D-glucosyl)(n)	Glc1	P	GT2_P_Glc1
GT2	2.4.1.16	(1,4-(N-acetyl-beta-D-glucosaminyl))(n)	GlcNAc	P	GT2_P_GlcNAc
GT2	2.4.1.212	beta-D-glucuronosyl-(1->3)-N-acetyl-beta-D-glucosaminyl-(1->4)-(nascent hyaluronan)	GlcA	P	GT2_P_GlcA
GT2	2.4.1.34	((1->3)-beta-D-glucosyl)(n)	Glc2	P	GT2_P_Glc2
GT2	2.4.1.-	?	?	F	GT2_F
GT2	2.4.1.-	?	?	R	GT2_R
GT2	2.4.1.289	N-acetyl-alpha-D-glucosaminyl-diphospho-ditrans,octacis-undecaprenol	GlcNAc	R	GT2_R_GlcNAc
GT2	2.4.1.-	?	?	M	GT2_M_P37782
					GT2_M_Q9I4K5
GT2	2.4.1.-	?	?	O	GT2_O
GT2	2.4.1.-	?	?	Q	GT2_Q
GT2	2.4.1.-	?	?	H	GT2_H
GT2	2.4.1.-	?	?		GT2_A0PJ42
					GT2_A8E1V5
					GT2_B4ERA8
					GT2_O01346
					GT2_O07339
					GT2_O31986
					GT2_P33702
					GT2_P74820
					GT2_Q0P9C6
					GT2_Q1L2K4
					GT2_Q52256
					GT2_Q56914
					GT2_Q56915
					GT2_Q5LFK5
					GT2_Q8KB11
					GT2_Q8R632
					GT2_Q8Y949
					GT2_Q9A5M0
					GT2_Q9AQI9
					GT2_Q9S520
					GT2_Q9X9S1
					GT2_Q9XC63
GT2	2.4.1.157	1,2-diacyl-sn-glycerol	DAG		GT2_DAG
GT2	2.4.1.287	alpha-L-rhamnopyranosyl-(1->3)-N-acetyl-alpha-D-glucosaminyl-diphospho-trans,octacis-decaprenol	Rha		GT2_Rha
GT2	2.4.1.303	N-acetyl-alpha-D-glucosaminyl-diphospho-ditrans,octacis-undecaprenol	GlcNAc		GT2_GlcNAc
GT6	2.4.1.88	N-acetyl-D-galactosaminyl-(1->3)-N-acetyl-D-galactosaminyl-(1->3)-D-galactosyl-(1->4)-D-galactosyl-(1->4)-D-glucosyl-(1<->1)-ceramide	GalNAc	B	GT6_B_GalNAc
GT6	2.4.1.88	N-acetyl-D-galactosaminyl-(1->3)-N-acetyl-D-galactosaminyl-(1->3)-D-galactosyl-(1->4)-D-galactosyl-(1->4)-D-glucosyl-(1<->1)-ceramide	GalNAc		GT6_GalNac
GT6	2.4.1.309	alpha-L-Fuc-(1->2)-beta-D-Gal-(1->3)-alpha-D-GalNAc-(1->3)-alpha-D-GalNAc-diphospho-ditrans,octacis-undecaprenol	Fuc	A	GT6_A_Fuc_Q5JBG6

FAM	Nº EC	DESCRIPCIÓN ACEPTOR	ACRÓNIMO ACEPTOR	GRUPO	NOMBRE HMM
GT6	2.4.1.37	alpha-L-fucosyl-(1->2)-D-galactosyl-R	Fuc		GT6_Fuc
GT6	2.4.1.40	glycoprotein-alpha-L-fucosyl-(1->2)-D-galactose	Fuc		GT6_Fuc
GT6	2.4.1.87	alpha-D-galactosyl-(1->3)-beta-D-galactosyl-(1->4)-beta-N-acetylglucosaminyl-R	Gal		GT6_Gal
GT7	2.4.1.174	beta-D-glucuronyl-(1->3)-D-galactosyl-proteoglycan	GlcA	A	GT7_A_GlcA
GT7	2.4.1.175	beta-D-glucuronosyl-(1->3)-N-acetyl-beta-D-galactosaminyl-proteoglycan	GlcA/GalNAc	A	GT7_A_GlcA/GalNAc
GT7	2.4.1.244	N-acetyl-beta-D-glucosaminyl group	GlcNAc	A	GT7_A_GlcNAc
GT7	2.4.1.-		?		GT7_GlcNAc/Glc
GT7	2.4.1.133	O-beta-D-xylosyl-[protein]	Xyl		GT7_Xyl
GT7	2.4.1.22	D-glucose	Glc		GT7_GlcNAc/Glc
GT7	2.4.1.274	beta-D-glucosyl-(1<->1)-ceramide	Glc		GT7_GlcNAc/Glc
GT7	2.4.1.275	N-acetyl-D-glucosamine	GlcNAc		GT7_GlcNAc1
GT7	2.4.1.38	N-acetyl-beta-D-glucosaminylglycopeptide	GlcNAc		GT7_GlcNAc/Glc
GT7	2.4.1.90	N-acetyl-D-glucosamine	GlcNAc		GT7_GlcNAc/Glc
GT8	2.4.1.-		?	H	GT8_H_Glc_lip
GT8	2.4.1.-		?	H	GT8_H_Glc_lip
GT8	2.4.1.44	lipopolysaccharide	Glc_lip	H	GT8_H_Glc_lip
GT8	2.4.1.58	lipopolysaccharide	Sugar_lip	H	GT8_H_Glc_lip_P19817
					GT8_H_Glc_lip_Q9ZIS5
GT8	2.4.1.58	lipopolysaccharide	Sugar_lip	H	GT8_H_Glc_lip
GT8	2.4.1.-		?	A	GT8_A
GT8	2.4.1.-		?	G	GT8_G
GT8	2.4.1.186	glycogenin	Gly	K	
GT8	2.4.1.-		?	I	GT8_I
GT8	2.4.1.123	myo-inositol	MioIno	I	GT8_I_MioIno
GT8	2.4.1.43	(1->4)-alpha-D-galacturonosyl)	GalA	L	GT8_L_GalA
GT8	2.4.1.-		?	D	GT8_D
GT8	2.4.1.-		?	B	GT8_B
GT8	2.4.1.17	INESPECIFIC	Ines	J	GT8_J_Ines
GT8	2.4.2.-		Glc-EGFlike	M	GT8_M_Glc-EFGlike
GT12	2.4.1.92	O-(N-acetyl-alpha-neuraminyl)-(2->3)-O-beta-D-galactopyranosyl-(1->4)-beta-D-glucopyranosyl-(1<->1)-ceramide	NeuNAc		GT12_NeuNAc
GT13	2.4.1.-	?	?	A	GT13_Man
GT13	2.4.1.101	3-(alpha-D-mannosyl)-beta-D-mannosyl-R	Man	B	GT13_Man
GT13	2.4.1.101	3-(alpha-D-mannosyl)-beta-D-mannosyl-R	Man	C	GT13_Man
GT13	2.4.1.101	3-(alpha-D-mannosyl)-beta-D-mannosyl-R	Man		GT13_Man
GT15	2.4.1.131	3-(alpha-D-mannosyl)-beta-D-mannosyl-R	Man		GT15_Man
GT21	2.4.1.-		?		GT21_SphngNAc
GT21	2.4.1.80	N-acylsphingosine	SphngNAc		GT21_SphngNAc
GT24	2.4.1.-		?		GT24
GT27	2.4.1.41	polypeptide	Peptide		GT27_Pep
GT27	2.4.1.-		GalNAc		GT27_GalNAc
GT43	2.4.-.-		?		GT43_Gal
GT43	2.4.1.135	3-beta-D-galactosyl-4-beta-D-galactosyl-O-beta-D-xylosylprotein	Gal		GT43_Gal
GT43	2.4.2.-		?		GT43_Gal
GT55	2.4.1.217	3-phospho-D-glycerate	PAG		GT55_PAG
GT64	2.4.1.223	beta-D-glucuronosyl-(1->3)-beta-D-galactosyl-(1->3)-beta-D-galactosyl-(1->4)-beta-D-xylosyl-proteoglycan	GlcA		GT64_GlcA
GT64	2.4.1.224	beta-D-glucuronosyl-(1->4)-N-acetyl-alpha-D-glucosaminyl-proteoglycan	GlcA		GT64_GlcA
GT78	2.4.1.269	D-glycerate	GA		GT78_GA
GT81	2.4.1.-		?	A	GT81_PGA
GT81	2.4.1.266	3-phospho-D-glycerate	PGA	A	GT81_PGA
GT81	2.4.1.-		?	B	GT81_PGA
GT81	2.4.1.217	3-phospho-D-glycerate	PGA	B	GT81_PGA
GT81	2.4.1.266	3-phospho-D-glycerate	PGA	B	GT81_PGA
GT81	2.4.1.266	3-phospho-D-glycerate	PGA	B	GT81_PGA
GT81	2.4.1.268	D-glycerate	GA	C	GT81_GA

Mediante diferentes técnicas se identificó la RV de las secuencias en cada grupo (ver métodos y Anexo 19) y las secuencias se usaron para generar perfiles HMM que fueron sometidos a una prueba de solidez (ver métodos), cuya finalidad es averiguar si todas las secuencias que definen el perfil aportan la misma información o si estarían mejor ubicada en otro perfil. Los grupos que no superaban esta prueba fueron retirados o rehechos, de modo que finalmente se obtuvieron un total de 107 perfiles HMM (tabla 19) asociados a la RV de cada conjunto de proteínas de una misma subfamilia GTA con aceptor común. El número de secuencias que pueblan cada perfil puede verse en el anexo 20.

Tabla 19. Perfiles HMM asociados a la RV de una misma subfamilia GTA con aceptor común.

GT2 (69 HMM)	GT2_A GT2_a_GlcNAc GT2_A0PJ42 GT2_A8E1V5 GT2_B GT2_b_Gal GT2_B4ERA8 GT2_C GT2_c GT2_D GT2_DAG GT2_E GT2_F_O07340 GT2_F_O87183 GT2_G GT2_GlcNAc GT2_H_B1B4J9 GT2_H_Q5J7C7	GT2_I GT2_L GT2_L_Q8KWR0 GT2_M_P37782 GT2_M_Q9I4K5 GT2_N GT2_O GT2_O01346 GT2_O07339 GT2_O31986 GT2_P GT2_P_C3U576 GT2_P_Glc1 GT2_P_Glc1_Q6RCS2 GT2_P_Glc2 GT2_P_GlcA GT2_P_GlcNAc	GT2_P_GlcNAc_P78746 GT2_P_GlcNAc_Q8V735 GT2_P33702 GT2_P74820 GT2_Q GT2_Q0P9C6 GT2_Q1L2K4 GT2_Q52256 GT2_Q56914 GT2_Q56915 GT2_Q5LFK5 GT2_Q8KB11 GT2_Q8R632 GT2_Q8Y949 GT2_Q9A5M0 GT2_Q9AQI9 GT2_Q9S520	GT2_Q9X9S1 GT2_Q9XC63 GT2_R_GlcNAc GT2_R_P74817 GT2_Rha GT2_S_P33697 GT2_S_P33700 GT2_S_Q9XBL5 GT2_U GT2_V_B5U882 GT2_V_DolP GT2_X_DAG GT2_Y_Ines GT2_Z_A0Q5C5 GT2_Z_Man GT2_Z_Q8KZ90 GT2_Z_UndP
GT6 (4 HMM)	GT6_A_Fuc GT6_B_GalNAc GT6_Fuc GT6_Gal	**GT7 (6 HMM)**	GT7_A_GlcA GT7_GlcNAc1 GT7_A_GlcA/GalNAc GT7_Xyl	GT7_A_GlcNAc GT7_GlcNAc/Glc
GT8 (14 HMM)	GT8_A GT8_C GT8_D GT8_G	GT8_H_Glc_lip GT8_H_Glc_lip_P19817 GT8_H_Glc_lip_Q9ZIS5 GT8_H_O24967	GT8_H_O25962 GT8_H_Q48484 GT8_I_MioIno GT8_J_Ines	GT8_L_GalA GT8_M_Glc-EFGlike
GT12 (1 HMM)	GT12_NeuNAc	**GT13 (1 HMM)**	GT13_Man	
GT15 (1 HMM)	GT15_Man	**GT21 (1 HMM)**	GT21_SphngNAc	
GT24 (1 HMM)	GT24	**GT27 (2 HMM)**	GT27_GalNAc GT27_Pep	
GT43 (1 HMM)	GT43_Gal	**GT55 (1 HMM)**	GT55_PAG	
GT64 (1 HMM)	GT64_GlcA	**GT78 (1 HMM)**	GT78_GA	
GT81 (2 HMM)	GT81_GA GT81_PGA	**GT82 (1 HMM)**	GT82	

El conjunto de secuencias utilizado para obtener la primera generación de HMM es relativamente pequeño, debido a la restricción de utilizar únicamente secuencias con información funcional anotada. El número de secuencias de cada HMM se mueve en torno a 1 y 56. La población de secuencias de cada HMM se incrementó mediante una única búsqueda de la primera generación de perfiles contra la base de datos CAZy (Anexo 20). Los resultados se incorporaron a cada perfil, siendo estos últimos los perfiles HMM definitivos. El número de secuencias de cada HMM es ahora en torno a 1 y 2856. Esta segunda generación de HMM, son los perfiles presentados, que asocian patrones de secuencia en la RV con aceptor para cada conjunto de proteínas de una misma subfamilia GTA.

4.2 Predicciones basadas en los perfiles HMM

La gran utilidad de los perfiles HMM generados es la de poder identificar secuencias GTAs y predecir al mismo tiempo su aceptor. A fin de comprobar dicha utilidad, se escaneó la base de datos GenBank en búsqueda de secuencias GTAs a partir de solo la información contenida en nuestros perfiles.

La base de datos de GenBank fue descargada y realizada una búsqueda en esta base de datos con cada perfil por separado. Los resultados se trataron de forma independiente y fueron agrupados según el porcentaje de secuencia cubierto (60, 65, 70, 75 y 80 %) y el umbral de *evalue* donde se realizó el corte (de 1E-1 a 1E-20) generando un conjunto de secuencias que fueron incluidas en el *pipeline* de CAZy para comprobar cuántas de las secuencias encontradas eran realimente GTA y estaban correctamente asignados a su familia y subgrupo.

Del resultado de esta comprobación se seleccionaron los mejores parámetros para cada HMM por separado, ya que son muy heterogéneos, con la intención de optimizar los resultados (ver métodos), es decir encontrar aquellos parámetros que maximicen los aciertos (glicosiltransferasas cuyo grupo y familia corresponden al perfil HMM) y minimicen los fallos (proteínas que no son glicosiltransferasas o tienen la familia o el grupo equivocado). Los resultados son variables, pero globalmente satisfactorios (Anexo 21). Todos los perfiles encontraron aciertos en la búsqueda de GenBank y el 70 % de ellos tenían un 90 % o superior precisión (la precisión define el porcentaje de aciertos en el resultado), lo que hace un 10 % o menos de falsos positivos, es decir proteínas que no son CAZymes (figura 121).

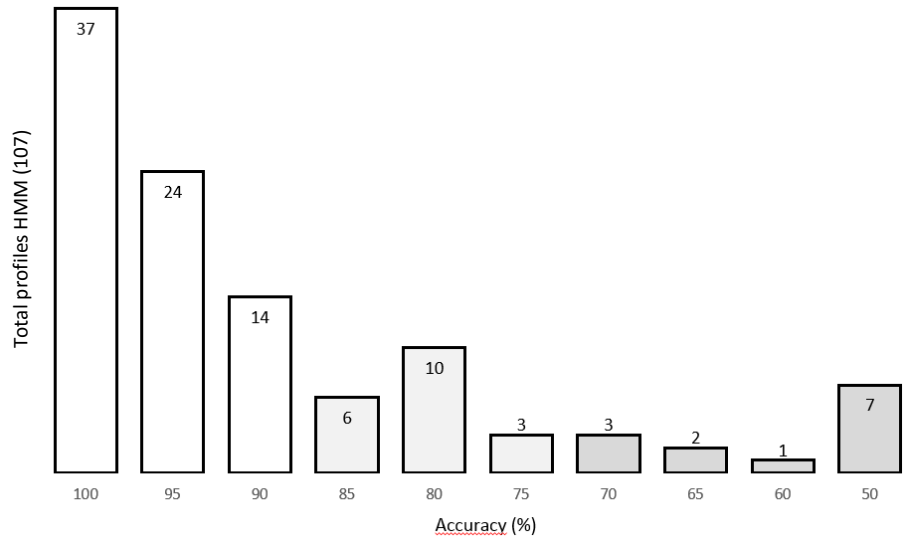


Figura 121. Nº de HMM por grupos de precisión. Los tres primeros grupos representan el 70 % de todos los perfiles, que tienen un 10 % o menos de falsos positivos.

Solo dos perfiles (GT2_Z_UndP y GT27_Pep) encontraron secuencias que pertenecían a un subgrupo o familia diferente al del HMM, pero en un porcentaje inferior al 0,3 % de los aciertos totales. El resto de los HMM siempre fueron capaces de discriminar entre las diferentes familias GTA y los subgrupos homólogos. Algunos perfiles tienen una precisión muy mala como GT2_Q52256 o GT8_G y no pueden ser usados como predictores, pero para el resto, la mayoría de ellos encontraron más aciertos que los utilizados para construir el HMM, con buena precisión para los parámetros seleccionados, donde generalmente es suficiente un recubrimiento del 60 % y un umbral de corte para el *evalue* relativamente bajo (1,00E-06 de media). El perfil HMM GT27_pep se construyó con tan solo 13 secuencias, encontrándose en la búsqueda, 4801 aciertos con solo un 5 % de falsos positivos; se presume que estas casi 5000 secuencias transferirán una molécula de azúcar a un péptido. De igual modo, el perfil HMM GT7_A_GlcNAc se construyó con 8 secuencias y encontró tras la búsqueda en GenBank, un total de 415 aciertos con una precisión del 86 %, se asume aquí que todas estas secuencias encontradas son proteínas GT7, que pertenecen al subgrupo A y transferirán un azúcar a una molécula de GlcNAc.

Por tanto, perfiles HMM de tan solo 7 columnas (GT7_A_GlcNAc), tienen sensibilidad suficiente para detectar proteínas de una superfamilia tan heterogénea como las GTAs e identificar a qué familia pertenece, incluso asignando el subgrupo homólogo y con gran probabilidad, su aceptor.

Para muchos de ellos, sobre todo en GT2, existen perfiles para grupos de proteínas de las que se desconoce el aceptor, pero una vez identificada su RV y construido el HMM, solo es necesaria la caracterización de una de ellas para identificar el aceptor y asignárselo a todo el grupo. Mostrando solo los perfiles donde se conoce el aceptor, se puede presentar un listado de 43 perfiles, preparados para ser asignados a las proteínas encontradas en cada búsqueda (tabla 20).

Nombre HMM	Evalue cutoff	Cobertura	Residuos cubiertos	Hits	Falsos positivos	Precisión (%)
GT2_a_GlcNAc	1,00E-07	0,6	14	1	0	100
GT2_b_Gal	1,00E-14	0,6	22	386	66	83
GT2_DAG	1,00E-07	0,65	22	4	0	100
GT2_P_Glc1	1,00E-07	0,75	29	3838	892	77
GT2_P_Glc1_Q6RCS2	1,00E-06	0,8	26	14	0	100
GT2_P_Glc2	1,00E-07	0,7	21	407	78	81
GT2_P_GlcA	1,00E-07	0,7	18	667	64	90
GT2_P_GlcNAc	1,00E-06	0,7	20	2834	124	96
GT2_P_GlcNAc_P78746	1,00E-09	0,6	15	190	9	95
GT2_P_GlcNAc_Q8V735	1,00E-06	0,7	21	19	0	100
GT2_R_GlcNAc	1,00E-06	0,8	42	543	155	71
GT2_Rha	1,00E-07	0,6	16	431	60	86
GT2_V_DolP	1,00E-05	0,75	24	4002	594	85
GT2_X_DAG	1,00E-14	0,6	17	235	32	86
GT2_Z_Man	1,00E-07	0,6	8	47	4	91
GT2_Z_UndP	1,00E-05	0,75	9	748	179	76
GT6_A_Fuc	1,00E-11	0,6	16	4	0	100
GT6_B_GalNAc	1,00E-08	0,6	16	372	18	95
GT6_Fuc	1,00E-12	0,7	19	674	36	95
GT6_Gal	1,00E-12	0,6	16	347	51	85
GT7_A_GlcA	1,00E-04	0,75	9	499	30	94
GT7_A_GlcA/GalNAc	1,00E-05	0,7	8	450	21	95
GT7_A_GlcNAc	1,00E-04	0,8	7	413	56	86
GT7_GlcNAc/Glc	1,00E-07	0,6	7	780	56	93
GT7_GlcNAc1	1,00E-06	0,6	7	162	9	94
GT7_Xyl	1,00E-01	0,7	7	300	10	97
GT8_H_Glc_lip	1,00E-09	0,6	9	108	13	88
GT8_H_Glc_lip_P19817	1,00E-05	0,8	11	104	14	87
GT8_H_Glc_lip_Q9ZIS5	1,00E-06	0,75	11	119	11	91
GT8_I_MioIno	1,00E-05	0,75	15	445	22	95
GT8_J_Ines	1,00E-05	0,8	6	110	16	85
GT8_L_GalA	1,00E-01	0,8	16	1206	68	94
GT8_M_Glc-EFGlike	1,00E-06	0,7	14	563	28	95
GT12_NeuNAc	1,00E-07	0,65	25	366	31	92
GT13_Man	1,00E-06	0,8	12	629	29	95
GT15_Man	1,00E-07	0,6	20	1163	53	95
GT21_SphngNAc	1,00E-08	0,75	21	773	52	93
GT27_GalNAc	1,00E-09	0,65	32	340	21	94
GT27_Pep	1,00E-06	0,65	29	4801	278	94
GT43_Gal	1,00E-06	0,7	22	1163	52	96
GT55_PAG	1,00E-08	0,6	18	63	14	78
GT64_GlcA	1,00E-05	0,75	17	1282	51	96
GT78_GA	1,00E-06	0,75	20	8	0	100
GT81_PGA	1,00E-08	0,7	15	1013	159	84

Tabla 20. Perfiles HMM utilizados como predictor de aceptores para proteínas GTAs.

5. Discusión

Hemos definido la especificidad por el aceptor en las GTAs, en la zona conservada de la topología consenso. Entre proteínas de la misma familia, hemos encontrado también indicios de que el tipo de aceptor establece niveles de homología por encima del dador y a veces incluso del taxón. Esto es así, puesto que árboles filogenéticos de diferentes aceptores y utilizando diferentes fragmentos de la secuencia, agrupan estos aceptores en los mismos nodos, independientemente del organismo o el dador, salvo cuando se utiliza la RV, donde debido a esta mayor variabilidad, los grupos son más difusos.

La RV es, efectivamente, más variable que el resto de la secuencia. Hemos hallado evidencias de una menor presión evolutiva sobre esta región, comparada con el resto de la secuencia en las GTAs (zona conservada). Todos los aceptores visualizados en proteínas GTA cristalizadas, se ubican en esta región, que contiene residuos que interaccionan directamente con el aceptor y que sí están relacionados con su especificidad, pero no el resto de la secuencia en esta zona. Creemos que esta menor presión evolutiva sobre la RV, se debe a que el responsable del posicionamiento de los residuos clave, para la interacción con el aceptor es el resto de la proteína sobre la RV y no la RV en sí misma, lo que permite mayor tasa de mutación en la última y que esta región actúe a modo de "laboratorio de pruebas" para nuevos aceptores.

A pesar de que la especificidad por el aceptor sea debida al plegamiento general del dominio GTA sobre la RV, esta última zona posee residuos muy conservados, que contribuyen a la interacción con el sustrato. Hemos utilizado la conservación de estos residuos como "huella" identificativa del aceptor en cada grupo de proteínas, mediante la generación de perfiles HMM de la RV, y hemos comprobado la viabilidad de estos perfiles como elementos de *screening*, con los que ha sido posible atribuir información a proteínas desconocidas, como su pertenencia al grupo de glicosiltransferasas, el tipo de plegamiento GTA, la familia y subgrupo homólogo al que pertenecen y, además, predecir el posible aceptor.

Encontramos de gran utilidad estos perfiles que, junto a una validación mediante árboles filogenéticos de sus resultados, pueden llegar a predecir el aceptor de una proteína desconocida.

CONCLUSIONES

Relaciones secuencia-estructura-función en glicosiltransferasas con plegamiento GTA:

1. Se han encontrado evidencias suficientes como para afirmar que **todas las proteínas incluidas dentro del plegamiento GTA siguen un patrón estructural consenso.** Se ha identificado en este patrón una región de secuencia y estructura variable, entre las hojas β5 y β6, a la que hemos denominado "región variable". Este patrón consenso se puede extender a partir de la hoja β7 en dos topologías bien diferenciadas según la familia: a) La hoja β7 se curva hacia la hélice α1, dejando abierta la hendidura catalítica como ocurre en las familias GT8, GT13 y GT15. b) Continúa con un bucle flexible que finaliza en una hélice α, situada siempre bajo la hélice α6 y cierra la hendidura catalítica, como se observa en el resto de familias.

2. El alineamiento múltiple de las proteínas cristalizadas y su refinamiento manual, según la información estructural, ha permitido la generación de un **perfil HMM global para la detección de nuevas GTAs, con el que además es posible identificar la región variable**.

3. El correcto **posicionamiento del sustrato dador de azúcar** en el centro activo es responsabilidad de diferentes residuos repartidos por la región conservada de la estructura consenso. La especificidad por el nucleósido quedaría definida por las estructuras previas al motivo DXD, mientras que la especificad por el azúcar transferido sería responsabilidad de residuos más allá de la región variable, dentro de la topología consenso.

4. Se propone que el **patrón de secuencias sobre el cual se asienta el dador** en todas las enzimas GTA quede definido por **D-DXD-D-H**. La primera posición se halla entre las estructuras β2 y α2 de la topología consenso y es extraordinariamente conservada en Asp o Glu y raramente en Asn o Arg. El motivo DXD se halla en medio de la lámina de hojas β. La tercera posición se halla en la hélice α6 de la topología consenso y es siempre un residuo Asp o Glu, que en las proteínas GTA estudiadas con inversión de la configuración, actúa como base catalítica. La cuarta posición se halla tras la hoja β7 y es siempre una His situada frente al motivo DXD, participando activamente en la estabilidad del dador mediante una interacción con el metal.

5. Por su lado, **el aceptor se posiciona siempre en la zona de la región variable**. Aún así, **es el plegamiento global el que define la especificidad por el aceptor**. La ubicación de la región variable en la estructura global de la proteína se relaciona con el destino final del aceptor: con acceso al solvente para moléculas solubles, inserta en membrana para moléculas de membrana o integrada en la región C-terminal en proteínas procesivas.

6. **La divergencia evolutiva entre proteínas de una misma familia** de glicosiltransferasas con plegamiento tipo GTA, **se produce por la diversidad de aceptores, antes que por la de sustratos dadores**.

7. El **patrón de secuencias específico de la región variable según el tipo de aceptor en cada proteína**, se ha usado para generar perfiles HMM específicos que permiten el *screening* de aceptores y familias GTA en proteínas de función desconocida. La relación entre región variable y aceptor ha quedado evidenciada y consideramos hacerla extensible a todas las proteínas con plegamiento GTA.

Identificación de residuos catalíticos en la proteína MG517:

8. Se ha generado un **modelo tridimensional para la región N-terminal de MG517**, validado por Dinámica Molecular, y compatible con la estructura consenso del plegamiento tipo GTA. No se ha podido determinar con exactitud la estructura de la región variable, aunque sí se ha identificado una hélice α en su inicio, de entre 4 y 10 residuos, que podría interactuar con la membrana.

9. Se identifican gran parte de los **residuos catalíticos o involucrados en reconocimiento del sustrato dador UDP-Glc de MG517**: el residuo E193, propuesto como base catalítica y D40, Y126, Y169, I170, e Y218 que alteran notablemente la actividad de la enzima.

10. El **sustrato aceptor DAG de MG517 se une preferentemente a la región variable**.

Interacción proteína MG517 con la membrana, a través de la región C-terminal:

11. Se ha descrito computacionalmente la **asociación a membrana de la enzima MG517, mediante una interacción monotópica irreversible por medio de una hélice anfipática**, situada en la parte apical de la región C-terminal. Sin descartar otros elementos de interacción.

12. La **región C-terminal de MG517 parece estar compuesta principalmente de hélices α**. La primera de estas hélices α podría situarse bajo la hélice 6 de la enzima MG517, como sucede en estructuras cristalizadas de otras glicosiltransferasas GTAs, cubriendo el parche hidrofóbico accesible al solvente en los modelos generados.

Cambios conformacionales del bucle catalítico en el mecanismo de la proteína GpgS:

13. **El bucle catalítico de la proteína GpgS adopta distintas conformaciones**. Esta es una propiedad intrínseca de la proteína tanto en presencia como en ausencia de ligandos. En ausencia de ligandos, el equilibrio conformacional está desplazado hacia la conformación inactiva del bucle, con una barrera de entre 2-3 kcal/mol y sin que se aprecie estabilidad para la conformación activa. En presencia de ambos ligandos el equilibrio se desplaza hacia la conformación activa del bucle. **Concluimos que las distintas conformaciones de este bucle obedecen a un mecanismo de ajuste inducido por ligandos.** La descripción fenomenológica de este ajuste inducido está relacionada con el movimiento de las cadenas laterales de Arg259 e His258, activado por la conformación del ligando UDPGlc. La forma activa del bucle catalítico estabiliza la unión del sustrato dador y al mismo tiempo genera un entorno más hidrofóbico en el centro activo que podría facilitar la catálisis.

14. Se realiza una propuesta para el **orden de entrada de los ligandos**, con el aceptor PGA en primer lugar y el dador UDPGlc+metal a continuación, momento en que el bucle cambiaría de conformación.

15. El mecanismo de regulación por ligando para este bucle se considera no extensible para todo el conjunto GTA, estando restringido a aquellas proteínas cuya extensión de la hoja β7 de la estructura consenso, no se halle firmemente fijada por la extensión de la hoja β5.

Structue-sequence-function relationships in glycosyltransferases with GTA fold:

1. It has been found enough evidence to say that **all proteins included within the GTA folding follow a structural consensus pattern**. It has been identified inside this pattern a region of variable sequence and structure , between the sheets β5 and β6, which we called "variable region". The consensus structure can be extended from the β7sheet in two distinct topologies well differentiated according to the family: a) The β7 sheet is bent towards the helix α1, leaving open the catalytic cleft as occurs in GT8, GT13 and GT15 families. b) Continue with a flexible loop ending in an α-helix, always beneath the propeller α6 and closes the catalytic cleft, as observed in other families.

2. The multiple alignment of crystallized protein sequences and its manual refinement, according to structural information, enabled the generation of a **global HMM profile for detecting new GTAs, with which it is also possible to identify the variable region**.

3. The correct **placing of the sugar donor substrate** in the active site is attributed to different residues along the conserved region of the consensus structure. Nucleoside specificity would be defined by the pre DXD motif structures while the transferred sugar specificity would be the responsibility of residues beyond the variable region, inside the consensus topology.

4. **The sequence pattern on which the donor** sits in all GTA enzymes, is now defined by **D-DXD-D-H**. The first position is between β2 and α2 estructures the consensus topology and is extraordinarily conserved in Asp or Glu or also, although rarely in Asn or Arg. The DXD motif is in the midst of the β sheets. The third position is in helix α6 of the consensus topology and is always an Asp or Glu residue, which in GTA proteins studied with inversion of configuration, acts as a catalytic base. Finally, the fourth position is behind the β7 sheet and is always a His placed opposite the to DXD motif, actively participating in the stability of the donor through an interaction with the metal.

5. **The acceptor is always placed in the area of the variable region**. Even so, **the overall protein folding defines the acceptor specificity**. The variable region location in the protein structure is also related to the final destination of the acceptor: accessible to the solvent if a soluble molecule, inserted into a membrane if it is a membrane molecule or integrated into the C-terminal region for processive proteins.

6. **The evolutionary divergence between proteins of the same family** of GTs with GTA fold, **is produced by the diversity of acceptors, rather than by the donor**.

7. The **specific sequence patterns of the variable region for each protein according to the type of acceptor**, has been used to generate HMM profiles for each family of GTA proteins and its acceptor. These can be used for screening GTA families and acceptors in unknown proteins. The relationship between variable region and acceptor has been evidenced and consider making it extensible to all proteins that fall in the GTA fold.

Catalitic residues identification in the MG517 protein, by means a 3D model generation.

8. A **3D model for the N-terminal MG517** has been generated, validated by Molecular Dynamics, compatible with the consensus structure of the GTA fold. It has been impossible to determine with precision, the structure of the variable region, although it has been identified an α helix in the beginning, between 4 and 10 residues, converged by Molecular Dynamics that could interact with the membrane.

9. The **catalytic residues and those involved in the donor substrate recognition, UDPGlc of MG517 have benn identified**: E193, proposed as catalytic base, D40, Y126, Y169, I170 and Y218 that alter significantly the enzyme activity.

10. The **acceptor substrate DAG MG517 preferentially binds to the variable region**.

MG517 membrane interaction, by means of the C-terminal región.

11. It has been computationally described the **MG517 enzyme membrane association by an irreversible monotopic interaction by means an amphipathic helix**, situated in the apical part of the C-terminal region. Other interaction elements are not discarded.

12. **The C-terminal región of MG517 seems to be composed mainly of α helices**. The first of these helices could be placed below the MG517 helix 6 of the consensus topology, covering the hydrophobic patch that is accesible to solvent in the generated models.

Conformational changes in a catalytic loop in the GpgS protein mechanism.

The GpgS catalytic loop adopts two different conformations. **This is an intrinsecal property of the protein, that shows them in the presence or absence of ligands**. In the apo form, the conformational equilibrium is displaced towards the inactive conformation, with a 2-3 kcal/mol barrier, showing no stability for the active conformation. With ligands, the equilibrium is displaced to the active loop conformation. We conclude that **the different conformations of this loop obbey to an induced fit mechanism by ligands**. The phenomenological description of this induced fit is related to the movement of the side chains of **Arg259** and **His258**, activated by the conformation of the ligand UDPGlc. When the PGA is present in the active site, Arg259 movement causes reduction of the barrier between the different conformations of the loop (within 0,5 kcal / mol). The active form is then stabilized loop by a change in the side chain of residue His258, allowing interaction with the metal. The active form of the catalytic loop stabilizes the binding of the donor substrate and simultaneously generates a more hydrophobic environment in the active site that could facilitate the catalysis.

13. The order of ligands binding has been proposed, with the acceptor PGA binding first and donor UDPGlc + metal later, when the loop conformational change takes place.

14. This loop mechanism seems to be shared only by those GTA proteins whose β7 sheet extension of the consensus structure is not firmly fixed by the β5 sheet extension, as GT27, GT78 or GT81 families.

MÉTODOS

1. Modelado de la proteína MG517 de Mycoplasma genitalium

La lista de enzimas GTA caracterizadas hasta la fecha se obtuvo desde la base de datos CAZy[142]. Las estructuras tridimensionales de estas enzimas se descargaron de Protein Data Bank (PDB) y su correspondiente secuencia de aminoácidos desde UniProt.

Las estructuras se superpusieron utilizando el servidor POSA[143]. Las anotaciones de la estructura secundaria para cada proteína se obtuvo con DSSP[22], modificando las anotaciones a mano para dejar solo dos tipos, hélice α (T, G y S a H) u hoja β (B a E). La predicción para la GT MG517 cuya estructura es desconocida se realizó con PsiPred[111]. Las secuencias se extrajeron de los archivos PDB y se alinearon con el servidor PROMALS[85], que utiliza un algoritmo que incluye perfiles basados en alineamientos múltiples de secuencias y que también incorpora información de la estructura secundaria. El refinado del alineamiento se hizo comparando visualmente la superposición de las estructuras con VMD[144], con el resultado del alineamiento con PROMALS. Aquellos aminoácidos de diferentes estructuras superpuestos en la misma región del espacio fueron colocados en la misma columna en el alineamiento; las anotaciones de la estructura secundaria se usaron también para guiar este refinado, utilizando el *script* "UPDATE_SS.SH" (Anexos *scripts*). Con HMMER[145] se construyó un perfil Hidden Markov Model (HMM) para el grupo de proteínas GTA, utilizando el alineamiento refinado y parámetros por defecto. Dos nuevos alineamientos múltiples de secuencias, esta vez con las secuencias completas extraídas desde UniProt, se generaron utilizando el perfil HMM por un lado y el servidor TCoffee Expresso, que incluye para su alineamiento las estructuras tridimensionales específicas de cada secuencia y se compararon. Con el último alineamiento múltiple obtenido con el perfil HMM, se generó y obtuvo con HMMER el último y definitivo perfil de HMM para el grupo de proteínas con plegamiento GTA.

Con el paquete de programas PHYLIP[86] se hizo un agrupamiento de las secuencias alineadas con el perfil HMM, utilizando el algoritmo Neighbor-joinning[146], la matriz BLOSUM62 y un remuestreo bootstrap de 1000 conjuntos de datos. Este bootstrap fue el paso final del programa CONSENSUS, incluido en el paquete PHYLIP, que dibuja un árbol consenso y asigna los valores finales a cada nodo del árbol. Como no estaba paralelizado el paquete PHYLIP, el proceso se automatizó mediante algunos *scripts* en *BASH*, para ello y ya que los computacionales se basaban en un servidor con 4 CPUs, se dividió el proceso en 4, con cuatro carpetas donde cada una contenía ¼ del conjunto de datos, que era asignado en un proceso diferente y que luego se volvían a juntar para CONSENSUS. (Anexos *scripts*)

Los modelos estructurales de la glicosiltransferasa de *Mycoplasma genitalium* MG517 (para la región N terminal, aa 1-220) se construyeron con el software MODELLER v.9.8[147], donde se usaron varias estructuras GTA como plantillas. Cuatro series diferentes de modelos se construyeron combinando la estructura del segundo dominio GT de la condroitín polimerasa de *E. coli* (PDB 2Z86, residuos 430 a 662) para la región conservada más una de las siguientes cuatro plantillas para la región variable:

1. La estructura de la GT6 bovina α3GalT (PDB 1O7Q, aa 242 a 287).
2. La estructura de la GT27 humana ppGaNTase-T1 (PDB 2FFU, aa 247 a 314).
3. El primer dominio GT de la condroitín polimerasa de *E. coli* (PDB 2Z86, residuos 263 a 335).
4. La estructura de la GT43 humana GlcAT-I (PDB 3CU0, residuos 213 a 259).

El último alineamiento múltiple obtenido con el perfil HMM para el grupo de proteínas GTA cristalizadas se usó para guiar el modelado de la estructura sobre las plantillas anteriormente citadas (Anexo 5). Modeller necesita que se le especifique todo aquello que no forma parte íntegra de la secuencia de la proteína, como los ligandos o el metal que los interpretará como bloques compactos, creando con ellos un modelo que pueda acomodarlos.

Cada bloque es una unidad que se especifica a Modeller como tal en el alineamiento, colocando un punto al final del mismo por cada unidad que el archivo PDB de la plantilla contenga. Las coordenadas tanto del UDP como del metal de cada plantilla para la RV (estructuras 1O7Q, 2FFU, 2Z86_1 y 3CU0) se han cortado del archivo PDB original y se han añadido al archivo PDB de la estructura para la región conservada (2Z86_2), por esa razón estas dos unidades (UDP y Mn) aparecen como puntos añadidos a la secuencia de 2Z86_2 en cada alineamiento. Para las plantillas usadas para la RV:

- 1O7Q: Que contiene β-D-Galactosa y N-acetil-D-glucosamina en el lugar aceptor (2 puntos).
- 2FFU: Que contiene un péptido de 9 aminoácidos en el lugar aceptor. 9 aminoácidos añadidos fuera de la proteína (9 puntos).
- 2Z86_1: Que no contiene ninguna molécula en el lugar del aceptor (0 puntos).
- 3CU0: Con una molécula digalactosil y el sulfato. (2 puntos).

La secuencia a modelar de MG517, puesto que se va a modelar con todos estos, incluye en la secuencia para cada modelo híbrido, tantos puntos como tengan una y otra plantilla. El formato es ".pir", obtenido a través del programa JalView y el alineamiento, con la única modificación de los encabezamientos de cada secuencia. Se generaron 10 modelos estructurales para cada combinación de plantillas (Anexos *scripts*), donde las restricciones al residuo K34 se eliminaron ya que la posición de su cadena lateral generaba problemas en el modelado. Cada modelo se refinó utilizando un protocolo simulado de calentamiento que viene implementado en MODELLER. Además, para cada uno de estos modelos, 10 nuevos modelos diferentes refinados se generaron para aquellas partes de la estructura que no estaban alineadas en las plantillas (zonas de gaps) también implementado en MODELLER. Los modelos incluían el sustrato dador (UDPGlc) en posición equivalente a la molécula de UDP en sus respectivas plantillas. La unidad completa de UDPGlc se modeló desde la de UDP-glucurónico presente en la estructura de la plantilla 2Z86 (dominio 2). Las 100 estructuras finales modeladas para cada combinación de plantillas se evaluaron por una asignación empírica de energías, como el DOPE normalizado[148], que también incluye MODELLER y se analizó la distribución de diedros de Ramachandran con PROCHECK[149] (Anexo 6).

Una estructura representativa de cada una de las cuatro series de modelos se escogió para iniciar una serie de simulaciones por Dinámica Molecular. Esta se escogió entre las 20 mejores de cada serie según la puntuación DOPE normalizada y los siguientes criterios:

- Residuos D40, D93, D95 y D194 no más lejos de 6 Å del ligando UDPGlc y residuos E193 o D194 orientados hacia el enlace de unión entre el difosfato del UDP y la glucosa.

El estado inicial de protonación de cada estructura se asignó con el servidor H++[150] (Salinidad 0,15 mM; Dieléctrica interna 6; Dieléctrica externa 80; pH: 6,5; Método de cálculo: Poisson-Boltzman); puesto que el estado de protonación de los residuos coincidía con el de GROMACs v4.5.3[151], se dejó el asignado por defecto por este programa, con el que se realizaron las simulaciones.

El ligando UDPGlc se incorporó posteriormente a las estructuras mediante superposición con el *plugin MultiSeq*, incorporado en VMD, en posición equivalente a la de las plantillas, ya que la parametrización de esta molécula no está incorporada en GROMACs.

La topología con los parámetros del campo de fuerzas y las coordenadas del UDP-Glc fueron generados con las herramientas del programa de modelado AMBER (Antechamber[152] y Xleap) y convertidas al formato GROMACS mediante la herramienta amb2gmx.pl[o], siguiendo este esquema.

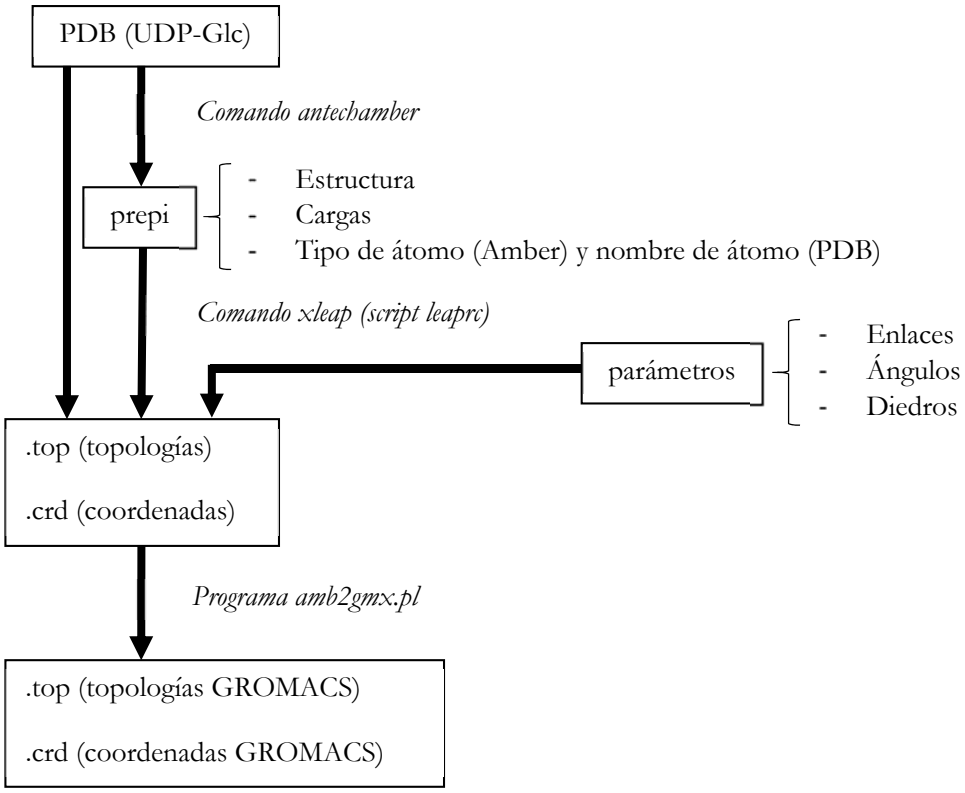


Figura 122. Esquema de preparación de los archivos para la inclusión de UDP-Glc.

[o] Disponible en: http://web.stanford.edu/group/pandegroup/folding/ffamber/amb2gmx.pl

La estructura del ligando UDPGlc fue obtenida de la proteína 2Z861_1 que contiene UDPGlcA, tras la eliminación manual del oxígeno sobrante. Los parámetros Amber para el ligando se tomaron de Petrová *et al* 1999[153] (Anexo 8). El metal Mg fue sustituido por Mn. Las simulaciones se realizaron con el campo de fuerzas Amber03, caja cúbica (-d 0,9), tratamiento de solvente explícito (tip3p) y pH neutro con iones Na^+ y Cl^- a una concentración final de 0,15 M. Equilibrado: cinco rondas de minimización de energía de 10000 pasos cada una, por el método *conjugated gradients*. 50000 pasos de *position restrains* para la hélice para equilibrar el solvente. 100000 pasos de calentamiento, con 6 puntos elevación de temperatura desde los 175 K hasta los 300 K para la proteína y directamente para las aguas, con un termostato "Nose-Hoover", por último 100000 pasos bajo un baróstato "Parrinello-Rahman" para equilibrar la presión. (Anexo *scripts*) (Anexo *inputs*).

Todas las simulaciones se ejecutaron usando el algoritmo LINCS para restringir la longitud de los enlaces y se aplicaron condiciones periódicas en todas direcciones. Las fuerzas electrostáticas de amplio rango se trataron con el método Fast Particle-Mesh Ewald (PME). Las fuerzas de Van der Waals y el potencial de Coulomb se trataron con un *cut-off* de 0,8 nm y el tiempo de simulación fue de 2 fs. No se generaron velocidades iniciales. Todas las simulaciones corrieron bajo un entorno NPT. (Anexo *inputs*)

Todas las simulaciones se ejecutaron en el supercomputador Picasso de la Universidad de Málaga. Mediante otro *script* se automatizó el lanzamiento de cada simulación, ya que no era posible simular más de 24 horas continuas (Anexo *scripts*).

Al final de cada simulación todas las estructuras generadas durante la trayectoria se agruparon, con el método "gromos" y RMSD como métrica[154]. Para elegir la estructura representativa de cada simulación se utilizó un consenso entre el tamaño del grupo (el mayor), su localización temporal (el más cercano al final de la simulación) y la ergodicidad encontrada (ver tabla a continuación).

Modelo	***Cut off* (nm)**	**Average RMSD (nm)**	**Nº de *clusters***	**Estructura seleccionada**	
1O7Q/2Z86_2	0,12	0,18	90	*Cluster* 3	986,4 ns
2FFU/2Z86_2	0,11	0,21	155	*Cluster* 1	879,2 ns
2Z86_1/2Z86_2	0,13	0,15	35	*Cluster* 1	646,9 ns
3CU0/2Z86_2	0,14	0,31	65	*Cluster* 1	960,0 ns

Tabla 21. Agrupamiento y selección de estructuras.

Para predecir el sitio de unión putativo de sustrato diacilglicerol (aceptor) y las estructuras de la GT MG517 se utilizó Autodock v4.2.3[155]. La estructura es la resultante de la selección anterior, incluidos los ligandos UDPGlc y metal. Se siguió una estrategia de *docking ciego* donde la estructura al completo se escaneó para la búsqueda de sitios de unión putativos. Como aceptor y sustrato de prueba se usó dipropionil glicerol (DPG) a partir del diacilglicerol (DAG) de la estructura 3A0B (figura 123).

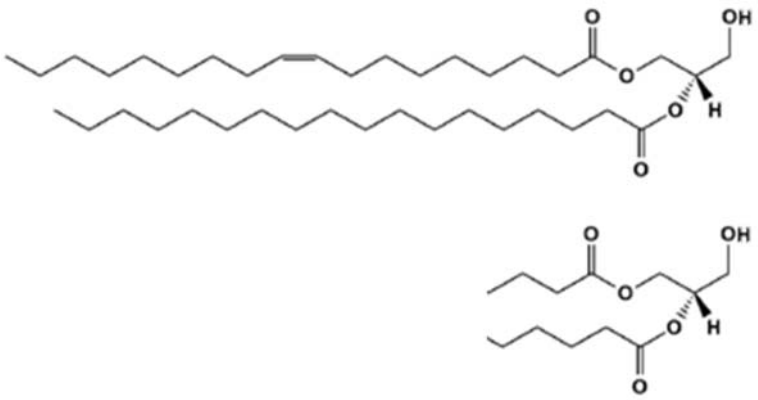


Figura 123. Estructuras del DAG, arriba y DPG, abajo. Se asume que las cadenas hidrocarbonadas del DAG estarán insertas en la membrana y no participarán en la unión al centro activo de la GT MG517 por lo que para el *docking* se usó DPG.

Se computó una parrilla de solvatación, electrostática y potencial de afinidad para cada tipo de átomo de la estructura completa, mediante Autogrid v4.2.3[155]. Cada enlace de la molécula sustrato se consideró flexible durante el proceso de *docking*, que constó de 100 rondas de algoritmo genético, donde en cada ronda se consideró una población inicial de 300 miembros, con coordenadas de orientación y posición inicial aleatorias, así como conformaciones aleatorias del sustrato flexible. El algoritmo genético se extendió durante 27000 generaciones, con un máximo de evaluaciones de energía de 25E6 (Anexos *inputs*).

2. Estudio de la región C-terminal de la proteína MG517

La secuencia de MG517 fue analizada con PsiPred v32 (23-mayo-2010) y MPEx v3.2.11[113] (28-enero-2014) utilizando los valores por defecto. El modelado se hizo con MODELLER v9.8. Todas las hélices se modelaron *de novo*, forzando el plegamiento de los residuos hacia una α hélice, con cuatro parámetros diferentes, que generan para cada una, cuatro modelos de hélices, entre las que se escoge la mejor modelada (mediante inspección visual) (Anexo *scripts).*

Hélices	**Secuencia**
Hélice 1	LIQCYEKLYVNLS
Hélice 2	KIEARFWRRQMFVWFA
Hélice 2.5	RRQMFVWFALFSFEYFKK
Hélice 3	FSESKKILEKLFVFLE
Hélice 4 (wt)	KNQGIYYIWVQRLKYFKHVLESK
Hélice 4 (no-polar)	ANQGIYYIWVQALAYFAHVLESK
Hélice 4 (no-apolar)	KNQGISQASVQRLKYFKHVLESK
Hélice 4 (truncada)	KNQGIYYIWVQRL

La Dinámica Molecular *all atom* se realizó con GROMACS v4.5.3 en los servidores de los nodos de supercomputación de la RES: Picasso, Tirant y Magerit, de Málaga, Valencia y Madrid respectivamente. Se construyeron principalmente dos sistemas: a) Péptido y solvente. b) Péptido, membrana y solvente.

A) **Péptido y solvente**.

Caja cúbica de 3 nm y campo de fuerzas Amber03. Agua como solvente explícito (tip3p) y pH neutro con iones Na^+ y Cl^- a una concentración final de 0,15 M. Equilibrado: cinco rondas de minimización de energía de 10000 pasos cada una, por el método *conjugated gradients*. 50000 pasos de *position restrains* para la hélice para equilibrar el solvente. 100000 pasos de calentamiento, con 6 puntos elevación de temperatura desde los 175 K hasta los 300 K para la proteína y directamente para las aguas, con un termostato "Nose-Hoover", por último 100000 pasos bajo un baróstato "Parrinello-Rahman" para equilibrar la presión (Anexo *scripts).*

Todas las simulaciones se ejecutaron usando el algoritmo LINCS para restringir la longitud de los enlaces y se aplicaron condiciones periódicas en todas direcciones. Las fuerzas electrostáticas de amplio rango se trataron con el método Fast Particle-Mesh Ewald (PME). Las fuerzas de Van der Waals y el potencial de Coulomb se trataron con un *cut-off* de 0,8 nm y el tiempo de simulación fue de 2 fs. No se generaron velocidades iniciales. Todas las simulaciones corrieron bajo un entorno NPT. (Anexo *inputs*).

El gráfico de evolución DSSP se hizo con la herramienta de GROMACs "*do_dssp*"[p]. Este comando lee un archivo de trayectoria y calcula la estructura secundaria de cada *frame* llamando al programa "dssp", obtenido de http://swift.cmbi.ru.nl/gv/dssp.

[p] Do_dssp asume que el ejecutable dssp se encuentra localizado en /usr/local/bin/dssp. Si no fuera el caso, habría que cargar entonces una variable de entorno DSSP, que apunte al ejecutable dssp.

B) **Péptido, membrana y solvente:**

Seguimos el protocolo para construir e insertar un péptido en una membrana disponible en: http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/membrane_protein/index.html. La membrana consiste en una bicapa de 128 moléculas de dipalmitoil fosfatidilcolina (DPPC), descargada desde la página http://wcm.ucalgary.ca/tieleman/downloads.

La estructura a insertar en la membrana se seleccionó mediante un análisis de agrupamientos con GROMACs de la simulación en solvente acuoso de la hélice 4, usando el método GROMOS y un *cutoff* de 0,3; con el que se obtuvo un RMSD medio de 0,48 y un total 161 grupos, escogiéndose la estructura media del grupo 1, a 1230 ns de simulación. La posición de la hélice y orientación respecto a la membrana se realizó manualmente mediante GROMACs "*editconv*", para cada sistema.

El empaquetado de lípidos alrededor de la hélice, se hizo con el programa "Inflategro.pl"[156] disponible desde la página web http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/membrane_protein/Files/inflategro.txt, que "infla" la membrana separando los lípidos en las 3 dimensiones del espacio y luego la cierra sobre la estructura mediante sucesivas rondas de minimización de energía, eliminando aquellos que chocan con la estructura, hasta un valor seleccionado de área por lípido de unos 70 $Å^2$.

Sistema1: Membrana de 128 DPPC. La hélice 4 se coloca en el centro de masas de la membrana perpendicular a los fosfolípidos. Cierre mediante "inflategro.pl"

Sistema2: La mitad de las colinas de la hemimembrana inferior son eliminadas manualmente, la membrana queda formada por 31 dipalmitoil fosfatidil metilo (DPPM) y 97 DPPC. La hélice 4 se coloca en la misma posición que en el sistema 1. Cierre de la bicapa forzado mediante presión durante el equilibrado.

Sistema3: La hemimembrana inferior contiene solo DPPM. Esta membrana está formada por 64 DPPM y 64 DPPC. La hélice 4 se coloca en el centro de masas de la bicapa, perpendicular a los fosfolípidos, como en los dos sistemas anteriores. Cierre de la bicapa forzado mediante presión durante el equilibrado.

Sistema 4: La hemimembrana superior contiene solo DPPM, teniendo carga negativa y formada por 64 DPPM y 64 DPPC. La hélice 4 se coloca al nivel de las cabezas fosfatadas de los fosfolípidos en la hemimembrana superior (≅1,85 nm entre los centros de masas del péptido y la membrana). Cierre de la bicapa forzado mediante presión durante el equilibrado, la hélice es fijada con *position restrains* durante el proceso. El péptido queda con fosfolípidos sobre él.

Sistema 5: La hemimembrana superior contiene solo DPPM, y formada por 64 DPPM y 64 DPPC. La hélice 4 se coloca sobre la hemimembrana superior, rotada 45° para favorecer la interacción entre los residuos polares y las cabezas negativamente cargadas de los fosfolípidos de membrana. Para permitir una interacción rápida y evitar que la hélice se separe de la membrana, la posición inicial permitía las siguientes interacciones:

- K316(HZ3)-DPM91(P8): Distancia 3,12 Å.
- R327(HH12)-DPM102(P8): Distancia 2,95 Å.
- R327(HH21)-DPM98(P8): Distancia 3,59 Å.
- K332(HZ2)-DPM112(P8): Distancia 2,75 Å

Sistema 6: La hemimembrana superior contiene solo DPPM y formada por 64 DPPM y 64 DPPC. La hélice 4 se coloca al nivel de las cabezas fosfatadas de los fosfolípidos en la hemimembrana superior, como en el sistema 4 (≅1,85 nm entre los centros de masas del péptido y la membrana) pero aquí sí se utiliza el empaquetado por InflateGRO y ningún fosfolípido cubre a la hélice, quedando esta con su parte apolar encarando el interior de la membrana y la polar completamente accesible al solvente.

Los sistemas son cajas triclínicas de 6,4x6,4x6,6 nm (tamaño inicial de la hemimembrana "dpp128.pdb") y campo de fuerzas "gromos53a6". Se utiliza agua como solvente explícito (tp3p).

Debido a que "inflategro.pl" puede dejar los fosfolípidos muy separados tras el cierre de la membrana, el agua puede intercalarse tanto entre ellos como en el centro de masas de la membrana, para evitarlo durante la solvatación se elevó el radio de Van der Waals de los carbonos a 0,375, mediante la modificación del archivo "vdwradii.dat" que incluimos en la carpeta de solvatación, las pocas aguas que todavía quedasen dentro de la membrana se eliminaron a mano. Se añadieron iones Na^+ y Cl^- a una concentración final de 0,15 M, hasta un pH neutro.

Equilibrado: Debido a que las hélices provenían de una MD previa y que la membrana se cerró sobre ella mediante rondas de minimización de energía, no fue necesario aquí comenzar el equilibrado con una nueva minimización. Tampoco se hizo ningún paso de equilibrado de aguas por separado. El calentamiento se hizo directamente a 323 K sobre los grupos "péptido", "membrana" y "solvente", manteniendo péptido y membrana fijos mediante *position restrains* durante 100000 pasos con un termostato "Nose-Hoover". El vacío generado en el interior de la membrana generó problemas al equilibrar la presión (las hemimembranas tendían a cerrarse sobre sí mismas en lugar de hacerlo hacia la opuesta), por ello este paso se hizo en dos tiempos. Durante el primero, de 50000 pasos, se fijaron los fosfatos de la membrana en el eje z por *position* restrains dejando libres x e y. El segundo, también de 50000 pasos, la membrana estaba libre, se utilizó un baróstato "Parrinello-Rahman" y el péptido fijado por *position* restrains. (Anexo *scripts).*

Todas las simulaciones se ejecutaron usando el algoritmo LINCS para restringir la longitud de los enlaces y se aplicaron condiciones periódicas en todas direcciones. Las fuerzas electrostáticas de amplio rango se trataron con el método Fast Particle-Mesh Ewald (PME). Las fuerzas de Van der Waals y el potencial de Coulomb se trataron con un *cut-off* de 1,2 nm y el tiempo de simulación fue de 2 fs. No se generaron velocidades iniciales. Todas las simulaciones corrieron bajo un entorno NPT. (Anexo *inputs*). La evolución de los sistemas y la conformación de las hélices se monitorización mediante inspección visual de las trayectorias a través de VMD.

Los sistemas 7 y 8 se montaron del mismo modo que el sistema 6, con una hélice rotada 90º y -90º respectivamente con "edit_conv" de GROMACs. También como el sistema 6 se hizo la simulación para la hélice 4 (anfipática) truncada.

Las simulaciones metadinámicas se prepararon monitorizando la MD del sistema 6. Mediante el *plugin* PLUMED para VMD y la variable colectiva "*Coordination*", se estudió el número de interacciones entre los nitrógenos de los residuos K316(N y NZ), R327(NH1, NH2 y NE), K329(NZ), K332(NZ) y K337(NZ) y todos los oxígenos de la hemimembrana DPM, así como la distancia entre los centros de masas del péptido y la membrana, con la variable colectiva "*Distance*". Se ajustaron los valores de la función de la CV "*Coordination*", para que el resultado fuera lo menos discreto posible y nunca llegase a cero cuando no se produjesen interacciones. También se monitorizó la variación de conformación de la hélice con la VC "*Alphabeta*". Se preparó un *input* para la Metadinámica; con la CV "*Coordination*" se exploraría la energía resultante de la formación y rotura de todas las interacciones posibles entre los residuos seleccionados y la membrana. La CV "*Distance*" introduciría el péptido en la membrana, hacia su centro de masas y lo alejaría del mismo hasta sacarlo de la membrana. Con estas dos variables se iniciaría una simulación metadinámica bidimensional. Se usó la CV "*Alphabeta*" para fijar la conformación de la hélice α, ya que sin esta la hélice se desplegaba. La distancia se también se restringió con "muros" entre 15 Å (límite inferior) y 60 Å (límite superior), de tal modo que el péptido nunca se introdujese en la membrana mucho más abajo de los fosfatos (no era nuestro interés) y tampoco superase la condición periódica al salir de la membrana (que podría adherir el péptido a la hemimembrana inferior). (Anexo *inputs*)

El lanzamiento de ejecuciones se automatizó con un *script* (Anexo *scripts*) Tras 55 ns de simulación de esta metadinámica, el algoritmo LINCS dio problemas entre algunos de los N activados para las CV y la simulación se detenía, por lo que se hizo un nuevo archivo .tpr, desactivando esta opción y se continuó la simulación.

El análisis de la simulación y los cálculos para la superficie de energía libre se realizaron con VMD y METAGUI.

3. Estudio del bucle catalítico en la proteína GpgS

Modelado del cristal 4DDZ

La estructura 4DDZ es un monómero al que le faltan los residuos 167 a 182, de la RV, y 295 a 301 de un bucle externo, además no está resuelta la cadena lateral de la Arg259, así como algunos átomos de la Asn260 en la conformación LI. Por simetría se obtuvo el homodímero y se modelaron *de novo* los átomos que faltaban con Modeller v9.8. Para cada conformación se realizan 100 modelos, cada uno con 10 bucles generados por Modeller, en total 1000 estructuras, donde en todas se forzó la generación de un puente disulfuro entre la Cys179 de ambos monómeros, puente que se ha comprobado experimentalmente existe y une covalentemente ambos monómeros.

Los residuos Gly254 a Pro262 tienen una ocupación 56:44, que corresponden a las conformaciones LA:LI, estos se separan a mano, creándose dos archivos pdb donde cada uno contiene el bucle en una sola conformación.

El modelado se realiza fijando todos los residuos de la proteína para que el modelo sea lo más parecido posible al cristal de partida, con la excepción de los residuos 167 a 185 y Arg259 en la conformación LA y LI y Asn260 en la conformación LI.

El modelado presentó problemas, ya que enlazaba los bucles de la RV de un monómero con el bucle del otro monómero. Solo cuatro modelos de los 1000 de la conformación LA presentaban una estructura donde los bucles no se hubieran enlazado, además nunca se consiguen modelar correctamente los dos monómeros a la vez, por lo que se toma el monómero A de uno de estos 4 modelos y se superpone al B, construyendo una estructura dimérica que servirá como .ini para Modeller en un segundo modelado con las mismas características del primero.

Este mismo .ini fue el que se utilizó para el modelado de las estructuras con conformación LI. Antes del modelado, los residuos del bucle LI más uno extra a cada extremo (253 a 263) fueron transferidos manualmente desde la estructura original al .ini. Este "corta y pega" de los bucles, generó dos *clashes* entre los residuos Arg167 de la RV y la His258 del bucle, y también entre la Pro168 de la RV y la Ala257 del bucle, sin embargo, estos fueron subsanados por la minimización del modelado.

La selección del modelo final se hizo mediante un *cluster analysis* realizado con GROMACs, con el método GROMOS y un *cut off* de 0,12. Se obtuvieron 470 grupos con un RMSD medio de 0,22, donde se seleccionó la primera estructura del grupo más poblado como modelo final (un RMSD medio de 0,11).

Modelado del cristal 4Y6N

Se utilizó el .ini del cristal 4DDZ para el modelado de esta segunda estructura. aquí los residuos no resueltos en el cristal son: 167 a 179, de la RV y 295 a 302, del bucle externo. También se forzó la creación del puente disulfuro Cys179. En este caso los modelos se realizaron con los ligandos en el interior: Metal, UDPGlc y PGA. El Mn se sustituyó por Mg.

La selección del modelo final se hizo mediante un *cluster analysis* realizado con GROMACs, con el método GROMOS y un *cut off* de 0,14. Se obtuvieron 300 grupos con un RMSD medio de 0,22, donde se seleccionó la primera estructura del grupo más poblado como modelo final (un RMSD medio de 0,13).

Dinámica Molecular

La Dinámica Molecular *all atom* se realizó con GROMACS v4.5.3 en nodo de supercomputación de la RES: Magerit de la Universidad Politécnica de Madrid.

El estado de protonación para todos los modelos se calculó con H++ server (Salinidad 0,15 mM; Dieléctrica interna 6; Dieléctrica externa 80; pH: 6,5; Método de cálculo: Poisson-Boltzman). Todos los residuos Glc, Asp, Lys y Arg están desprotonados. La His27 está totalmente protonada, el resto de His está protonadas en ε salvo His258 que está protonada en δ. El metal Mg fue sustituido por Mn.

Los modelos fueron solvatados en agua e iones en una caja triclínica de 10,6x7,4x6,2 nm a una concentración de iones final de 0,15 mM y pH 7.

A) Equilibrado de los modelos 4DDZ en conformación LA y LI:

Caja triclínica de 10,6x7,4x6,2 nm y campo de fuerzas Amber03. Agua como solvente explícito (tip3p) y pH neutro con iones Na+ y Cl- a una concentración final de 0,15 M. Equilibrado: cinco rondas de minimización de energía de 10000 pasos cada una, por el método conjugated gradients. 100000 pasos de position restrains para la proteína para equilibrar el solvente. 750000 pasos de calentamiento, con 10 puntos de elevación de temperatura, durante los 90000 primeros pasos, desde los 175 K hasta los 300 K para la proteína y directamente para las aguas, con un termostato "Nose-Hoover", por último 100000 pasos bajo un baróstato "Parrinello-Rahman" para equilibrar la presión (Anexo *scripts*).

B) Equilibrado de los modelos 4Y6N.

El equilibrado de esta estructura siguió un protocolo diferente, más suave, para no alterar la posición y distancia entre los ligandos dador y aceptor. Además, en el monómero A los ligandos Mn, UDPGlc y PGA se fijaron con *restrains* (1000 kJ/mol/nm2) solamente durante el equilibrado de las aguas y equilibrado de la proteína. Para el monómero B los ligandos Mn, UDPGlc y PGA se fijaron con *restrains* (4000 kJ/mol/nm2) en todos los pasos de equilibrado. Con la intención de conservar la distancia interatómica entre los átomos que interaccionan con el metal, en el

monómero B, estos también fueron fijados con *constrains* utilizando el *plugin* PLUMED y un archivo de entrada METAINP. (Anexo *inputs*)

Caja triclínica de 10,6x7,4x6,2 nm y campo de fuerzas Amber03. Agua como solvente explícito (tip3p) y pH neutro con iones Na+ y Cl- a una concentración final de 0,15 M. Equilibrado: 11 rondas de minimización de energía de 10000 pasos cada una, alternando los métodos de *steepest descend* y *conjugated gradients*. 100000 pasos de *position restrains* para la proteína para equilibrar el solvente. A continuación, otros 100000 pasos con el solvente y la proteína libres (pero no los ligandos). 1000000 pasos de calentamiento, elevando la temperatura de 15 en 15 K, desde la temperatura del paso anterior, hasta los 300 K para los grupos proteína, aguas, ligandos monómero A y ligandos monómero B, con un termostato "Nose-Hoover" y tantos puntos de *annealing* como sean necesarios, por último 100000 pasos bajo un baróstato "Parrinello-Rahman" para equilibrar la presión (Anexo *scripts*).

Tras el equilibrado comienza la MD. Cada 20 ns las *restrains* para los ligandos en el monómero B se reducen en 1000 unidades. Tras 80 ns de simulación el monómero B no tiene *restrains*. Después de 140 ns de simulación se elimina la *constrain* entre el metal y la His258.

Dinámica Molecular de 4Y6N apo

Utilizando los 100 primeros ns de la simulación del complejo ternario como intervalo, se tomó una estructura cada 10 ns y fueron eliminados el metal y los ligandos. Para cada estructura se simularon 100 ns de MD, en la misma caja de simulación, donde los primeros 20 ns la proteína estaba fijada con *position restrains* (1000 kJ/mol/nm^2) para equilibrar el solvente. Para la primera estructura (extraída a los 10 ns de la MD del complejo ternario) se extendió la simulación hasta 1 μs.

Simulaciones metadinámicas

Las Metadinámicas se realizaron con GROMACS v4.5.3 y el *plugin* PLUMED v1.3 en el nodo de supercomputación de la RES: Magerit de la Universidad Politécnica de Madrid.

Se probaron diferentes CVs en diversas rondas de simulación con la intención de seleccionar las más apropiadas para definir el cambio conformacional del bucle. Estos primeros intentos están basados en la información obtenida de las dinámicas moleculares.

- MetaD4DDZLA: Estructura equilibrada de la MD 4DDZ apo. (*Inputs* MetaD4DDZLA).
    - Duración 1,65 μs.
    - Ergodicidad: 400 ns.
- MetaD4Y6Napo1: Estructura MD4Y6NApo10 como estructura de partida. (*Inputs* MetaD4Y6Napo).
    - Duración 640 ns.
    - Ergodicidad: 120 ns.
- MetaD4Y6Napo2: Estructura MD4Y6NApo10 como estructura de partida. (*Inputs* MetaD4Y6Napo).

  - FES restringido: Se restringió el FES a la zona externa de los valores de diedro Ala257ψ -2 y 1, evitando que se explorase entre ellos para acelerar la simulación. La restricción se hizo manualmente, colocando en la parte inicial del archivo HILLS gaussianas de 20 kj de a lo largo de toda la distancia de exploración (1 gaussiana / Å) y en los valores de diedro -1,7 y 0,7. Utilizamos el mismo procedimiento en todas las simulaciones con FES restringido.
  - Duración 800 ns.
  - Ergodicidad: 450 ns.
- MetaD4Y6NApo3: Se extrajeron manualmente los ligandos UDPGlc, metal y PGA de la MD4Y6N (Complejo ternario) tras 100 ns de simulación y se inició la metadinámica. Metadinámicas realizadas por el método de BIAS-Exchange, con 3 CVs, una por simulación y una simulación 0 sin ninguna CV activada, se incluye el METAINP para la simulación 0, el resto tendrá activadas sus correspondientes CVs. (*Inputs* MetaD4Y6NApo3).
  - Duración 600 ns.
  - Ergodicidad: 100 ns.
- MetaD4Y6NApo4: Se utilizó la estructura MD4Y6NApo10 como estructura de partida. (*Inputs* MetaD4Y6NApo4).
  - Duración 660 ns.
  - Ergodicidad: 100 ns.
- MetaD4Y6NUDPGlc·PGA: Complejo ternario. FES restringido. (*Inputs* Meta4Y6N1).
  - Duración 500 ns.
  - Ergodicidad: 100 ns.
- MetaD4Y6NUDPGlcFree: Se extrajo manualmente el ligando PGA de la MD4Y6N (Complejo ternario) tras 100 ns de simulación y se inició la metadinámica. FES restringido. (*Inputs* Meta4Y6N2).
  - Duración 1,3 μs.
  - Ergodicidad: 250 ns.
- MetaD4Y6NUDPGlcFix: Se extrajo manualmente el ligandos PGA de la MD4Y6N (Complejo ternario) tras 100 ns de simulación y se inició la metadinámica. FES restringido. El diedro de la Glc se fijó con la CV "*alphabeta*". (*Inputs* Meta4Y6N3).
  - Duración 500 ns.
  - Ergodicidad: 100 ns.
- MetaD4Y6NPGA: Se extrajeron manualmente los ligandos: Mn y UDPGlc de la MD4Y6N (Complejo ternario) tras 100 ns de simulación y se inició la metadinámica. FES restringido. El diedro del PGA se fijó con la CV "*alphabeta*" (*Inputs* Meta4Y6N4).
  - Duración 500 ns.
  - Ergodicidad: 100 ns.
- MetaD4Y6NUDPGlc3: Se utilizó la estructura de la simulación MD4Y6N (Complejo ternario) tras 100 ns para iniciar la MetaD. FES restringido. El diedro de la Glc se fijó con la CV "*alphabeta*". (*Inputs* Meta4Y6N5).

    - Duración 660 ns.
    - Ergodicidad: 100 ns.
- MetaD4Y6NUDPGlc4: Se extrajeron manualmente los ligandos PGA de la MD4Y6N (Complejo ternario) tras 100 ns de simulación y se inició la metadinámica. Metadinámicas realizadas por el método de BIAS-Exchange, con 3 CVs, una por simulación y una simulación 0 sin ninguna CV activada, se incluye el METAINP para la simulación 0, el resto tendrá activadas sus correspondientes CVs. (*Inputs* Meta4Y6N7).
    - Duración 470 ns.
    - Ergodicidad: 100 ns.
- MetaD4Y6NUDPGlc5: Se extrajeron manualmente los ligandos PGA de la MD4Y6N (Complejo ternario) tras 100 ns de simulación y se inició la metadinámica. Metadinámicas realizadas por el método de BIAS-Exchange, con 5 CVs, una por simulación y una simulación 0 sin ninguna CV activada, se incluye el METAINP para la simulación 0, el resto tendrá activadas sus correspondientes CVs. (*Inputs* Meta4Y6N8).
    - Duración 480 ns.
    - Ergodicidad: 100 ns.

Algunas de las metadinámicas tienen un límite de 1 nm a la distancia máxima de la CV Asp236(O)-Arg259(N), establecida mediante "muros". El valor del diedro de la Glc se monitorizó durante la MD del complejo ternario, con la CV "alphabeta" y esta misma se utilizó para fijar este diedro, en las metadinámicas correspondientes, mediante "muro". (*Inputs* metadinámicas).

Cada metadinámica se analizó con la herramienta GRAF y METAGUI, las simulaciones con VMD y las gráficas han sido realizadas con GNUPLOT.

Para el cálculo de la energía de afinidad de los ligandos por la proteína GpgS, en sus diferentes cristales y conformaciones, se usó Autodock Vina utilizando los parámetros por defecto. La trayectoria de la MD del complejo ternario, con un dt100 generó 860 *frames* que fueron convertidos a .pdb con la herramienta "trjconv –dump", se extrajeron los ligandos de cada tiempo y se calculó de manera independiente para cada tiempo, la energía de afinidad entre el ligando y la proteína.

4. Uso de la Región Variable como herramienta predictiva

278 archivos de estructuras fueron descargados desde la base de datos PDB. Los archivos se dividieron según las diferentes cadenas y dominios GTA, obteniendo 498 estructuras con sus ligandos que se superpusieron con el *plugin* MultiSeq de la herramienta de visualización VMD, se anotaron las características de cada proteína y se comparó su topología mediante inspección visual.

Desde la base de datos CAZy, tuvimos acceso a secuencias de proteínas GTA que tenían asignado un nº EC de 2.4. de las que separamos el dominio GTA definido por CAZy e identificamos la RV de cada una. Previo a este acceso, construimos un *script* para poder descargar automáticamente desde el servidor público de CAZy, los códigos UniProt de proteínas GTA, y desde UniProt las secuencias y los nº EC de cada una. Sin embargo, esta última forma de obtener las secuencias es incompleta, ya que no todas las secuencias GTAs de CAZy son públicas, ni la información de UniProt está tan actualizada como la de CAZy, por lo que, aunque incluimos los *scripts* para extraer esta información, no se usó su resultado sino el cedido por el personal de CAZy.

Para identificar la RV nos valimos de diversos métodos, el principal nuestro perfil HMM para proteínas GTA que generamos en el apartado 4.1. Cada secuencia por separado fue alineada con este perfil HMM utilizando HMMER 3.0 (con –mapali para forzar los alineamientos), identificando la RV en la gran mayoría de ellas; para otras el alineamiento solo llegaba hasta el DXD o no había alineamiento. Con estas últimas secuencias usamos BLAST para encontrar proteínas con un grado de homología no superior al 30 % con el que usar el perfil HMM para proteínas GTA y encontrar en estas la posible RV, después se alineaban estas últimas con las problemáticas y se identificaba la RV. Cada una de las RV fue extraída de la secuencia generando 3 archivos, uno con el dominio GTA completo, otro solo con la RV y otro con el dominio GTA sin la RV.

El alineamiento de cada grupo se hizo a través de la herramienta para escritorio Jalview 2.9.0b2 y el método de T-coffee por defecto. A partir de estos alineamientos se extrajeron en bloque, las RV identificadas.

Agrupamos las secuencias que contenían solo la RV por familia y subgrupo y también por aceptor conocido y generamos un perfil HMM con cada grupo utilizando HMMER 3.0 y ninguna consideración especial. Estos perfiles fueron sometidos a una prueba de concepto para comprobar su validez, con las herramientas "hmmpress" y "hmmscan" de HMMER y que básicamente consiste en extraer una secuencia cada vez de cada perfil HMM, generar un nuevo perfil con las restantes y buscar en qué perfil HMM encaja mejor la secuencia extraída mediante puntuación evalue. Si el mejor perfil es el de partida, la secuencia se mantiene en él, si no, ha de extraerse para el perfil definitivo.

Los perfiles HMM ya validados se utilizaron frente a la base de datos CAZy con "hmmsearch" para incrementar el número de secuencias que lo forman. Utilizamos una cobertura de la RV del 70 % y un evalue del 0.001 como *cut off* para seleccionar los resultados.

Las secuencias se incorporaron a los perfiles y alinearon con MAFFT v7.221 y las condiciones: "--anysymbol --op 1.6 --ep 0.35 --maxiterate 10", después se generaron los nuevos y definitivos perfiles con HMMER y valores por defecto.

Se realizó una búsqueda con cada HMM frente a la base de datos de proteínas de GenBank (versión 25 de abril de 2015) y las condiciones "--tblout --domtblout -Z 1".

De las tablas obtenidas para cada resultado se filtraron los códigos según cobertura de la RV (60 %, 65 %, 70 %, 75 % y 80 %) y evalue (0.1 a 1E-20) y se obtuvieron las secuencias de cada conjunto de códigos, de la base de datos descargada de GenBank. Estas secuencias entraron en el *pipeline* de CAZy como secuencias desconocidas, donde las herramientas automáticas de CAZy determinan la clasificación de la proteína.

En una hoja de datos, para cada HMM, se listó el nº de aciertos y de falsos positivos para cada evalue (eje x) y cobertura de la RV (eje y). Se calculó la pendiente entre cada dos puntos y seleccionó como mejor condición la que presentaba el valor más cercano a uno, escogiendo el punto más bajo.

Para la realización de los árboles filogenéticos, las secuencias se descargaron de la base de datos de CAZy y filtradas con CD-HIT v4.6[157] con valores por defecto (90 % de identidad salvo para la familia GT8 que se utilizó un 85 %). El alineamiento se realizó con MAFFT v7.221[158] y las condiciones: "--anysymbol --op 1.6 --ep 0.35 --maxiterate 10". Los árboles se hicieron con Fasttree v2.1.8[159] y valores por defecto, para su visualización y análisis se utilizó Dendroscope v3.2.10[160].

BIBLIOGRAFÍA

1. Gloster, T. M. ScienceDirect Advances in understanding glycosyltransferases from a structural perspective. *Curr. Opin. Struct. Biol.* **28,** 131–141

2. Benz, I. & Schmidt, M. A. Never say never again: protein glycosylation in pathogenic bacteria. *Mol. Microbiol.* **45,** 267–76 (2002).

3. Endo, T. & Koizumi, S. Large-scale production of oligosaccharides using engineered bacteria. *Curr. Opin. Struct. Biol.* **10,** 536–541 (2000).

4. Scheible, W.-R. & Pauly, M. Glycosyltransferases and cell wall biosynthesis: novel players and insights. *Curr. Opin. Plant Biol.* **7,** 285–95 (2004).

5. Keegstra, K. & Raikhel, N. Plant glycosyltransferases. *Curr. Opin. Plant Biol.* **4,** 219–24 (2001).

6. De Angelis, E., Watkins, A., Schäfer, M., Brümmendorf, T. & Kenwrick, S. Disease-associated mutations in L1 CAM interfere with ligand interactions and cell-surface expression. *Hum. Mol. Genet.* **11,** 1–12 (2002).

7. Markine-Goriaynoff, N. *et al.* Glycosyltransferases encoded by viruses. *J. Gen. Virol.* **85,** 2741–54 (2004).

8. Yamamoto, F., Clausen, H., White, T., Marken, J. & Hakomori, S. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345,** 229–233 (1990).

9. Talbot, P. Cell Adhesion and Fertilization: Steps in Oocyte Transport, Sperm-Zona Pellucida Interactions, and Sperm-Egg Fusion. *Biol. Reprod.* **68,** 1–9 (2002).

10. Hakomori, S. Glycosylation defining cancer malignancy: new wine in an old bottle. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 10231–3 (2002).

11. Becker, D. J. & Lowe, J. B. Fucose: biosynthesis and biological function in mammals. *Glycobiology* **13,** 41R–53R (2003).

12. Rudd, P. M. Glycosylation and the Immune System. *Science (80-.).* **291,** 2370–2376 (2001).

13. Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.* **77,** 521–55 (2008).

14. Ardèvol, A. & Rovira, C. The molecular mechanism of enzymatic glycosyl transfer with retention of configuration: Evidence for a short-lived oxocarbenium-like species. *Angew. Chemie - Int. Ed.* **50,** 10897–10901 (2011).

15. Albesa-Jové, D. *et al.* A Native Ternary Complex Trapped in a Crystal Reveals the Catalytic Mechanism of a Retaining Glycosyltransferase. *Angew. Chem. Int. Ed. Engl.* **54,** 9898–902 (2015).

16. Charnock, S. J. & Davies, G. J. Structure of the nucleotide-diphospho-sugar transferase, SpsA from Bacillus subtilis, in native and nucleotide-complexed forms. *Biochemistry* **38,** 6380–6385 (1999).

17. Liu, J. & Mushegian, A. Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci.* **12,** 1418–31 (2003).

18. Lizak, C., Gerber, S., Numao, S., Aebi, M. & Locher, K. P. X-ray structure of a bacterial oligosaccharyltransferase. *Nature* **474,** 350–5 (2011).

19. Zhang, H. *et al.* The highly conserved domain of unknown function 1792 has a distinct glycosyltransferase fold. *Nat. Commun.* **5,** 4339 (2014).

20. 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).

21. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–6 (1986).

22. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–637 (1983).

23. Kaczanowski, S. & Zielenkiewicz, P. Why similar protein sequences encode similar three-dimensional structures? *Theor. Chem. Acc.* **125,** 643–650 (2009).

24. Chung, S. Y. & Subbiah, S. A structural explanation for the twilight zone of protein sequence homology. *Structure* **4,** 1123–7 (1996).

25. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10,** 845–858 (2015).

26. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12,** 7–8 (2014).

27. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268,** 209–25 (1997).

28. Guex, N., Peitsch, M. C. & Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis* **30 Suppl 1,** S162–73 (2009).

29. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–8 (2005).

30. Eswar, N. *et al.* Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **Chapter 2,** Unit 2.9 (2007).

31. Lengauer, T. & Rarey, M. Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.* **6,** 402–406 (1996).

32. Mcconkey, B. J., Sobolev, V. & Edelman, M. The performance of current methods in ligand – protein docking. *October* **83,** 845–856 (2002).

33. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided. Drug Des.* **7,** 146–57 (2011).

34. Goodsell, D. S. & Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **8,** 195–202 (1990).

35. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28,** 849–57 (1985).

36. Kastenholz, M. A., Pastor, M., Cruciani, G., Haaksma, E. E. & Fox, T. GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.* **43,** 3033–44 (2000).

37. Levitt, D. G. & Banaszak, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **10,** 229–34 (1992).

38. Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and

intermolecular interactions. *J. Mol. Graph.* **13,** 323–30, 307–8 (1995).

39. Glaser, F., Morris, R. J., Najmanovich, R. J., Laskowski, R. A. & Thornton, J. M. A method for localizing ligand binding pockets in protein structures. *Proteins* **62,** 479–88 (2006).

40. Brady, G. P. & Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided. Mol. Des.* **14,** 383–401 (2000).

41. Mezei, M. A new method for mapping macromolecular topography. *J. Mol. Graph. Model.* **21,** 463–72 (2003).

42. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft* **27,** 2985–2993 (1894).

43. Koshland, D. E. Correlation of structure and function in enzyme action. *Science* **142,** 1533–41 (1963).

44. Hammes, G. G. Multiple conformational changes in enzyme catalysis. *Biochemistry* **41,** 8221–8 (2002).

45. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J. & Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **153 Suppl ,** S7–26 (2008).

46. Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A. & Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **49,** 534–53 (2006).

47. Morris, G. M. *et al.* Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19,** 1639–1662 (1998).

48. Allen, M. Introduction to molecular dynamics simulation. *Comput. Soft Matter From Synth. Polym. to …* **23,** 1–28 (2004).

49. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. (2002).

50. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4,** 187–217 (1983).

51. Christen, M. *et al.* The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **26,** 1719–51 (2005).

52. Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110,** 1657–1666 (1988).

53. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 12562–6 (2002).

54. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1,** 826–843 (2011).

55. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23,** 187–199 (1977).

56. Huber, T., Torda, A. E. & van Gunsteren, W. F. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J. Comput. Aided. Mol. Des.* **8,** 695–708 (1994).

57. Grubmüller, H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E. Stat. Phys. Plasmas. Fluids. Relat. Interdiscip. Topics* **52,** 2893–2906 (1995).

58. Darve, E. & Pohorille, A. Calculating free energies using average force. *J. Chem. Phys.* **115,** 9169–9183 (2001).

59. Park, S. & Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **120,** 5946–61 (2004).

60. Marsili, S., Barducci, A., Chelli, R., Procacci, P. & Schettino, V. Self-healing umbrella sampling: a non-equilibrium approach for quantitative free energy calculations. *J. Phys. Chem. B* **110,** 14011–3 (2006).

61. Barducci, A., Pfaendtner, J. & Bonomi, M. Tackling sampling challenges in biomolecular simulations. *Methods Mol. Biol.* **1215,** 151–71 (2015).

62. Piana, S. & Laio, A. A bias-exchange approach to protein folding. *J. Phys. Chem. B* **111,** 4553–9 (2007).

63. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21,** 1087 (1953).

64. Biarnés, X., Pietrucci, F., Marinelli, F. & Laio, A. METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. *Comput. Phys. Commun.* **183,** 203–211 (2012).

65. Shmuel Razin, R. H. *Molecular biology and pathogenicity of mycoplasmas*. (Springer, 2002).

66. Razin, S., Yogev, D. & Naot, Y. Molecular biology and pathogenicity of mycoplasmas. *Microbiol. Mol. Biol. Rev.* **62,** 1094–156 (1998).

67. Pollack, J. D., Williams, M. V & McElhaney, R. N. The comparative metabolism of the mollicutes (Mycoplasmas): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit. Rev. Microbiol.* **23,** 269–354 (1997).

68. Klement, M. L. R., Ojemyr, L., Tagscherer, K. E., Widmalm, G. & Wieslander, A. A processive lipid glycosyltransferase in the small human pathogen Mycoplasma pneumoniae: involvement in host immune response. *Mol. Microbiol.* **65,** 1444–57 (2007).

69. Cunha, B. A. The atypical pneumonias: clinical diagnosis and importance. *Clin. Microbiol. Infect.* **12 Suppl 3,** 12–24 (2006).

70. Marrie, T. J. *et al.* The role of atypical pathogens in community-acquired pneumonia. *Semin. Respir. Crit. Care Med.* **33,** 244–56 (2012).

71. Andrés, E., Martínez, N. & Planas, A. Expression and characterization of a Mycoplasma genitalium glycosyltransferase in membrane glycolipid biosynthesis: potential target against mycoplasma infections. *J. Biol. Chem.* **286,** 35367–79 (2011).

72. Falk, L., Fredlund, H. & Jensen, J. S. Signs and symptoms of urethritis and cervicitis among women with or without Mycoplasma genitalium or Chlamydia trachomatis infection. *Sex. Transm. Infect.* **81,** 73–8 (2005).

73. Horner, P. J., Gilroy, C. B., Thomas, B. J., Naidoo, R. O. & Taylor-Robinson, D. Association of Mycoplasma genitalium with acute non-gonococcal urethritis. *Lancet (London, England)* **342,** 582–5 (1993).

74. Taylor-Robinson, D. The Harrison Lecture. The history and role of Mycoplasma genitalium in sexually transmitted diseases. *Genitourin. Med.* **71,** 1–8 (1995).

75. Lindblom, G., Brentel, I., Sjölund, M., Wikander, G. & Wieslander, A. Phase equilibria of membrane lipids from Acholeplasma laidlawii: importance of a single lipid forming nonlamellar phases. *Biochemistry* **25,** 7502–10 (1986).

76. Dahlqvist, A. *et al.* Efficient modulation of glucolipid enzyme activities in membranes of Acholeplasma laidlawii by the type of lipids in the bilayer matrix. *Biochemistry* **34,** 13381–9 (1995).

77. Vikström, S., Li, L., Karlsson, O. P. & Wieslander, A. Key role of the diglucosyldiacylglycerol synthase for the nonbilayer-bilayer lipid balance of Acholeplasma laidlawii membranes. *Biochemistry* **38,** 5511–20 (1999).

78. Andrés, E., Biarnés, X., Faijes, M. & Planas, A. Bacterial glycoglycerolipid synthases: processive and non-processive glycosyltransferases in mycoplasma. *Biocatal. Biotransformation* (2012).

79. Canal, M. TFC Marcè Canal. *Tesis fin de carrera* (Institut Químic de Sarrià, 2009).

80. Malani, P. N. Harrison's Principles of Internal Medicine. *JAMA* **308,** 1813 (2012).

81. Urresti, S. *et al.* Mechanistic insights into the retaining glucosyl-3-phosphoglycerate synthase from mycobacteria. *J. Biol. Chem.* **287,** 24649–61 (2012).

82. Edman, M. *et al.* Structural features of glycosyltransferases synthesizing major bilayer and nonbilayer-prone membrane lipids in Acholeplasma laidlawii and Streptococcus pneumoniae. *J. Biol. Chem.* **278,** 8420–8 (2003).

83. Botté, C. *et al.* Molecular modeling and site-directed mutagenesis of plant chloroplast monogalactosyldiacylglycerol synthase reveal critical residues for activity. *J. Biol. Chem.* **280,** 34691–701 (2005).

84. Stein, M. & baseman, J. The evolving saga of Mycoplasma genitalium. *Clin. Microbiol. Newsl.* **28,** 41–48 (2006).

85. Pei, J. & Grishin, N. V. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* **23,** 802–8 (2007).

86. Felsenstein, J. 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* **5,** 164–166 (1989).

87. Pereira, P. J. B. *et al.* Mycobacterium tuberculosis glucosyl-3-phosphoglycerate synthase: structure of a key enzyme in methylglucose lipopolysaccharide biosynthesis. *PLoS One* **3,** e3748 (2008).

88. 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. U. S. A.* **101,** 15307–12 (2004).

89. Keenleyside, W. J., Clarke, A. J. & Whitfield, C. Identification of Residues Involved in Catalytic Activity of the Inverting Glycosyl Transferase WbbE from Salmonella enterica Serovar Borreze. *Society* **183,** 77–85 (2001).

90. Zhang, Y. *et al.* Roles of individual enzyme-substrate interactions by alpha-1,3-galactosyltransferase in catalysis and specificity. *Biochemistry* **42,** 13512–21 (2003).

91. Fulton, Z. *et al.* Crystal structure of a UDP-glucose-specific glycosyltransferase from a Mycobacterium species. *J. Biol. Chem.* **283,** 27881–90 (2008).

92. Tarbouriech, N., Charnock, S. J. & Davies, G. J. Three-dimensional structures of the Mn and Mg dTDP complexes of the family GT-2 glycosyltransferase SpsA: a comparison with related NDP-sugar glycosyltransferases. *J. Mol. Biol.* **314,** 655–61 (2001).

93. Garinot-Schneider, C., Lellouch, A. C. & Geremia, R. A. Identification of essential amino acid residues in the Sinorhizobium meliloti glucosyltransferase ExoM. *J. Biol. Chem.* **275,** 31407–13 (2000).

94. Patenaude, S. I. *et al.* The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nat. Struct. Biol.* **9,** 685–690 (2002).

95. Lind, J. *et al.* High Cationic Charge and Bilayer Interface-Binding Helices in a Regulatory Lipid Glycosyltransferase † , ‡. *Biochemistry* **46,** 5664–5677 (2007).

96. Liebau, J., Pettersson, P., Szpryngiel, S. & Mäler, L. Membrane Interaction of the Glycosyltransferase WaaG. *Biophys. J.* **109,** 552–563 (2015).

97. Drin, G. & Antonny, B. Amphipathic helices and membrane curvature. *FEBS Lett.* **584,** 1840–7 (2010).

98. Madsen, K. L., Bhatia, V. K., Gether, U. & Stamou, D. BAR domains, amphipathic helices and membrane-anchored proteins use the same mechanism to sense membrane curvature. *FEBS Lett.* **584,** 1848–1855 (2010).

99. Joanne, P. *et al.* Lipid reorganization induced by membrane-active peptides probed using differential scanning calorimetry. *Biochim. Biophys. Acta - Biomembr.* **1788,** 1772–1781 (2009).

100. Cui, H., Lyman, E. & Voth, G. A. Mechanism of membrane curvature sensing by amphipathic helix containing proteins. *Biophys. J.* **100,** 1271–9 (2011).

101. Ash, W. L., Zlomislic, M. R., Oloo, E. O. & Tieleman, D. P. Computer simulations of membrane proteins. *Biochim. Biophys. Acta - Biomembr.* **1666,** 158–189 (2004).

102. Daggett, V. & Levitt, M. Molecular dynamics simulations of helix denaturation. *J. Mol. Biol.* **223,** 1121–1138 (1992).

103. Biggin, P. C. & Sansom, M. S. . Interactions of α-helices with lipid bilayers: a review of simulation studies. *Biophys. Chem.* **76,** 161–183 (1999).

104. Lindahl, E. & Sansom, M. Membrane proteins: molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **18,** 425–431 (2008).

105. Vemparala, S., Domene, C. & Klein, M. L. Computational studies on the interactions of inhalational anesthetics with proteins. *Acc. Chem. Res.* **43,** 103–10 (2010).

106. Fjell, C. D., Hiss, J. A., Hancock, R. E. W. & Schneider, G. Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discov.* **11,** 37 (2011).

107. Nymeyer, H. & García, A. E. Simulation of the folding equilibrium of alpha-helical peptides: a comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 13934–9 (2003).

108. Kandasamy, S. K. & Larson, R. G. Molecular Dynamics Simulations of Model Trans-Membrane Peptides in Lipid Bilayers: A Systematic Investigation of Hydrophobic Mismatch. *Biophys. J.* **90,** 2326–2343 (2006).

109. Jensen, M. Ø., Mouritsen, O. G. & Peters, G. H. Simulations of a membrane-anchored peptide: structure, dynamics, and influence on bilayer properties. *Biophys. J.* **86,** 3556–75 (2004).

110. Deighan, M. & Pfaendtner, J. Exhaustively Sampling Peptide Adsorption with Metadynamics. *Langmuir* **29,** 7999–8009 (2013).

111. Bryson, K. *et al.* Protein structure prediction servers at University College London. *Nucleic Acids Res.* **33,** W36–8 (2005).

112. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–80 (2001).

113. Snider, C., Jayasinghe, S., Hristova, K. & White, S. H. MPEx: a tool for exploring membrane proteins. *Protein Sci.* **18,** 2624–8 (2009).

114. Schiffer, M. & Edmundson, A. B. Use of Helical Wheels to Represent the Structures of Proteins and to Identify Segments with Helical Potential. *Biophys. J.* **7,** 121–135 (1967).

115. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L. Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 11937–42 (2009).

116. Andrés, E., Martínez, N. & Planas, A. Glycolipid biosynthesis in mycoplasma genitalium: recombinant expression and characterization of a glycosyltransferase producing mono- and diglycosyldiacylglycerols in the plasma membrane. *J. Biol. Chem.* **286,** 35367–35379 (2011).

117. Val-Cid, C. *et al.* Structural-Functional Analysis Reveals a Specific Domain Organization in Family GH20 Hexosaminidases. *PLoS One* **10,** e0128075 (2015).

118. Lluch-Senar, M., Querol, E. & Piñol, J. Cell division in a minimal bacterium in the absence of ftsZ. *Mol. Microbiol.* **78,** 278–89 (2010).

119. Ramakrishnan, B., Balaji, P. . & Qasba, P. K. Crystal Structure of β1,4-Galactosyltransferase Complex with UDP-Gal Reveals an Oligosaccharide Acceptor Binding Site. *J. Mol. Biol.* **318,** 491–502 (2002).

120. Gastinel, L. N., Cambillau, C. & Bourne, Y. Crystal structures of the bovine beta4galactosyltransferase catalytic domain and its complex with uridine diphosphogalactose. *EMBO J.* **18,** 3546–57 (1999).

121. 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,** 8206–10 (2014).

122. Pesnot, T., Jørgensen, R., Palcic, M. M. & Wagner, G. K. Structural and mechanistic basis for a new mode of glycosyltransferase inhibition. *Nat. Chem. Biol.* **6,** 321–3 (2010).

123. Breton, C., Snajdrová, L., Jeanneau, C., Koca, J. & Imberty, A. Structures and mechanisms of glycosyltransferases. *Glycobiology* **16,** 29R–37R (2006).

124. Guerin, M. E. *et al.* Substrate-induced conformational changes in the essential peripheral membrane-associated mannosyltransferase PimA from mycobacteria: implications for catalysis. *J. Biol. Chem.* **284,** 21613–21625 (2009).

125. Guerin, M. E. *et al.* Molecular recognition and interfacial catalysis by the essential phosphatidylinositol mannosyltransferase PimA from mycobacteria. *J. Biol. Chem.* **282,** 20705–14 (2007).

126. Brew, K., Tumbale, P. & Acharya, K. R. Family 6 Glycosyltransferases in Vertebrates and Bacteria: Inactivation and Horizontal Gene Transfer May Enhance Mutualism between Vertebrates and Bacteria. *J. Biol. Chem.* **285,** 37121–37127 (2010).

127. Qasba, P. K., Ramakrishnan, B. & Boeggeman, E. Substrate-induced conformational changes in glycosyltransferases. *Trends Biochem. Sci.* **30,** 53–62 (2005).

128. Ly, H. D., Lougheed, B., Wakarchuk, W. W. & Withers, S. G. Mechanistic studies of a retaining alpha-galactosyltransferase from Neisseria meningitidis. *Biochemistry* **41,** 5075–85 (2002).

129. Kumar, G., Guan, S. & Frantom, P. A. Biochemical characterization of the retaining glycosyltransferase glucosyl-3-phosphoglycerate synthase from Mycobacterium tuberculosis. *Arch. Biochem. Biophys.* **564,** 120–127 (2014).

130. Snajdrová, L., Kulhánek, P., Imberty, A. & Koca, J. Molecular dynamics simulations of glycosyltransferase LgtC. *Carbohydr. Res.* **339,** 995–1006 (2004).

131. Guerin, M. E., Korduláková, J., Alzari, P. M., Brennan, P. J. & Jackson, M. Molecular basis of phosphatidyl-myo-inositol mannoside biosynthesis and regulation in mycobacteria. *J. Biol. Chem.* **285,** 33577–83 (2010).

132. Jørgensen, R., Pesnot, T., Lee, H. J., Palcic, M. M. & Wagner, G. K. Base-modified donor analogues reveal novel dynamic features of a glycosyltransferase. *J. Biol. Chem.* **288,** 26201–8 (2013).

133. Gonçalves, S. *et al.* Structural analysis of Thermus thermophilus HB27 mannosyl-3-phosphoglycerate synthase provides evidence for a second catalytic metal ion and new insight into the retaining mechanism of glycosyltransferases. *J. Biol. Chem.* **285,** 17857–68 (2010).

134. Flint, J. *et al.* Structural dissection and high-throughput screening of mannosylglycerate synthase. *Nat. Struct. Mol. Biol.* **12,** 608–14 (2005).

135. Ramakrishnan, B. & Qasba, P. K. Crystal structure of the catalytic domain of Drosophila beta1,4-Galactosyltransferase-7. *J. Biol. Chem.* **285,** 15619–26 (2010).

136. Pedersen, L. C. *et al.* Heparan/chondroitin sulfate biosynthesis. Structure and mechanism of human glucuronyltransferase I. *J. Biol. Chem.* **275,** 34580–5 (2000).

137. Morgan, J. L. W., Strumillo, J. & Zimmer, J. Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature* **493,** 181–6 (2013).

138. Chaikuad, A. *et al.* Conformational plasticity of glycogenin and its maltosaccharide substrate during glycogen biogenesis. 2–7 (2011). doi:10.1073/pnas.1113921108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1113921108

139. Lobsanov, Y. D. *et al.* Structure of Kre2p/Mnt1p: A yeast ??1,2-mannosyltransferase involved in mannoprotein biosynthesis. *J. Biol. Chem.* **279,** 17921–17931 (2004).

140. Lomako, J., Lomako, W. M. & Whelan, W. J. A self-glucosylating protein is the primer for rabbit muscle glycogen biosynthesis. *FASEB J.* **2,** 3097–103 (1988).

141. Smythe, C. & Cohen, P. The discovery of glycogenin and the priming mechanism for glycogen biogenesis. *Eur. J. Biochem.* **200,** 625–31 (1991).

142. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37,** D233–8 (2009).

143. Ye, Y. & Godzik, A. Multiple flexible structure alignment using partial order graphs. *Bioinformatics* **21,** 2362–9 (2005).

144. William Humphrey, A. D. and K. S. VMD: Visual molecular dynamics. *J. Mol. Graph.* Volume **14,** 33–38

145. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–37 (2011).

146. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–25 (1987).

147. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).

148. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15,** 2507–24 (2006).

149. Laskowski R A, MacArthur M W, Moss D S, T. J. M. 20.19. (1993). PROCHECK - a program to check the stereochemical quality of protein structures. J. App. Cryst., 26, . *J. App. Cryst* **26,** 283–291 (1993).

150. Gordon, J. C. *et al.* H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **33,** W368–71 (2005).

151. Berendsen, H. J. C., D. van der S. and R. van D. GROMACS: a message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **91,** 43–56 (1991).

152. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Antechamber, An Accessory Software Package For Molecular Mechanical Calculations. *J. Chem. Inf. Comput. Sci.*

153. Petrová, P., Koča, J. & Imberty, A. Potential Energy Hypersurfaces of Nucleotide Sugars: Ab Initio Calculations, Force-Field Parametrization, and Exploration of the Flexibility. *J. Am. Chem. Soc.* **121,** 5535–5547 (1999).

154. X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, A. E. M. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed* 38: 236–240 (1999). Available at: http://bioinf.uab.es/xavier/papers/angewcheminted_38-236.pdf. (Accessed: 26th April 2012)

155. Morris, G. M. *et al.* Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19,** 1639–1662 (1998).

156. Kandt, C., Ash, W. L. & Peter Tieleman, D. Setting up and running molecular dynamics simulations of membrane proteins. *Methods* **41,** 475–488 (2007).

157. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–2 (2012).

158. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–80 (2013).

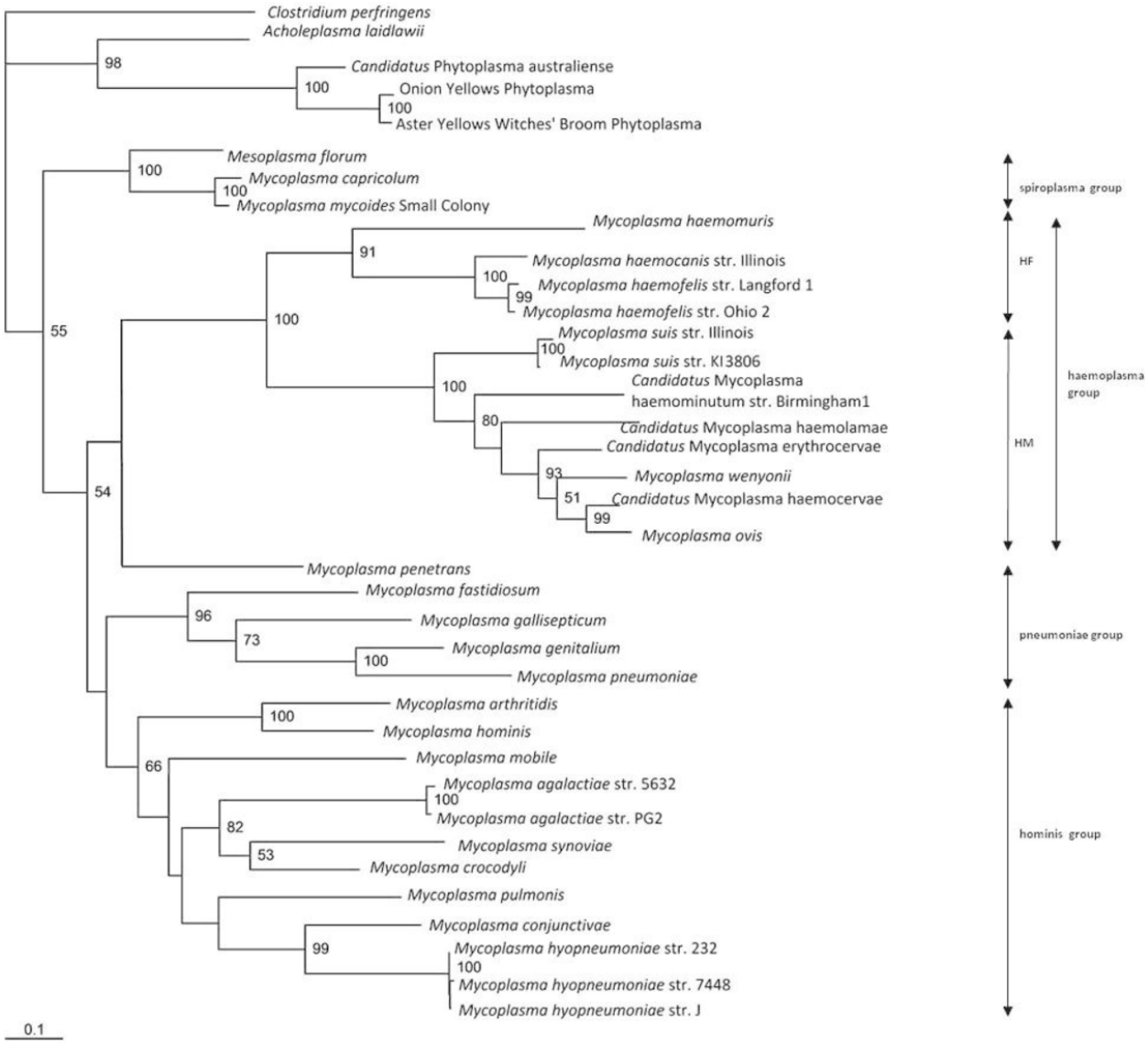
159. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26,** 1641–50 (2009).

160. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61,** 1061–7 (2012).

ANEXOS

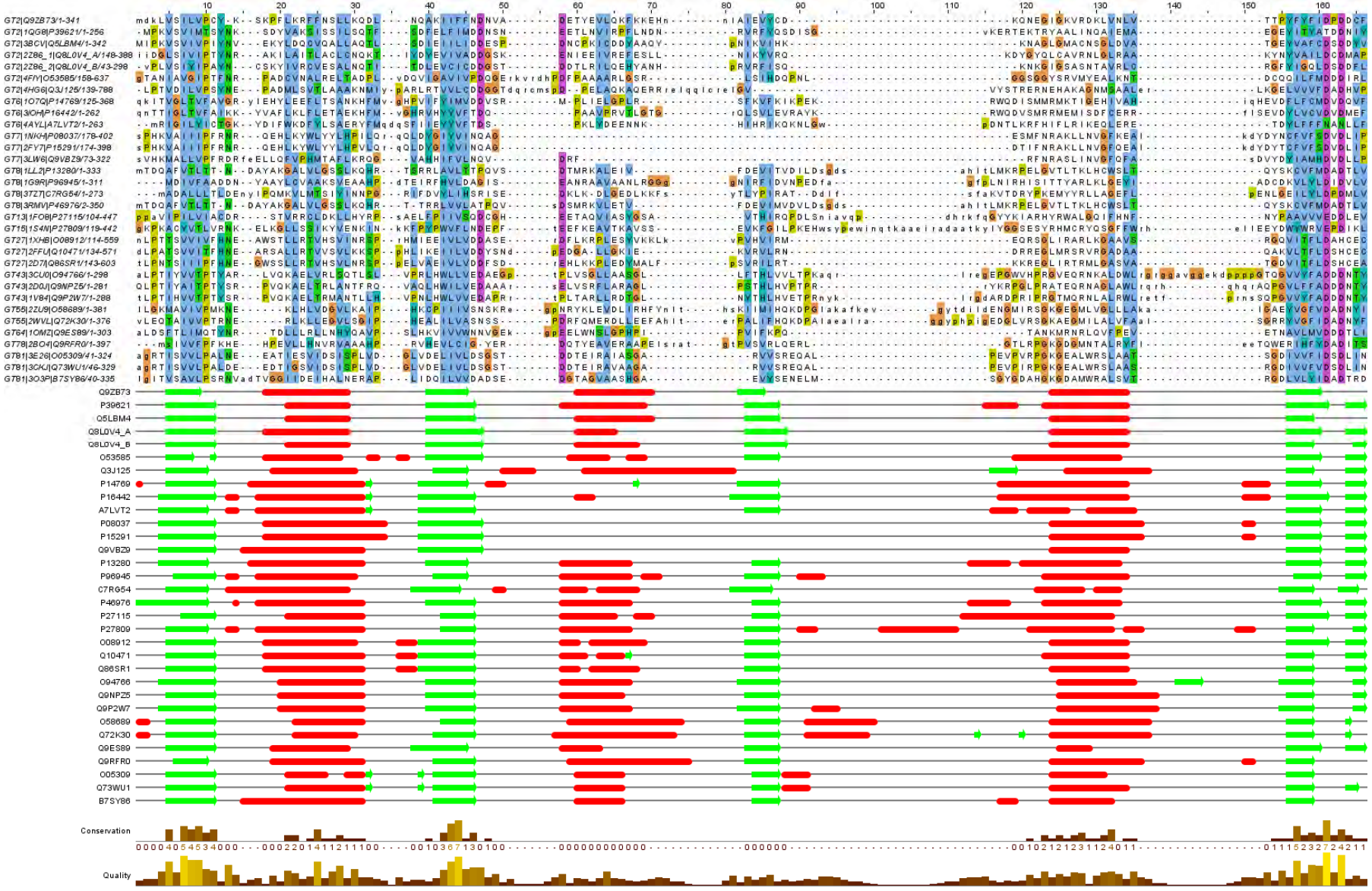
Anexo 1. Árbol filogenético de los micoplasmas y su relación con clostridium.

Árbol concatenado realizado por el método de *maximum likelihood* para los genes dnaK y gapA. El agrupamiento de Mycoplasma (hominis, spiroplasma y pneumoniae) está descrito por (Peters et al., 2008; Weisburg et al., 1989) así como haemoplasmas; HF indica el supgrupo haemofelis y HM el subgrupo haemominutum. Los datos se resmuestrearon 1000 veces y el resultado del *bootstrap* se muestra como porcentaje en los nodos (valores menores de 50 % no se muestran).

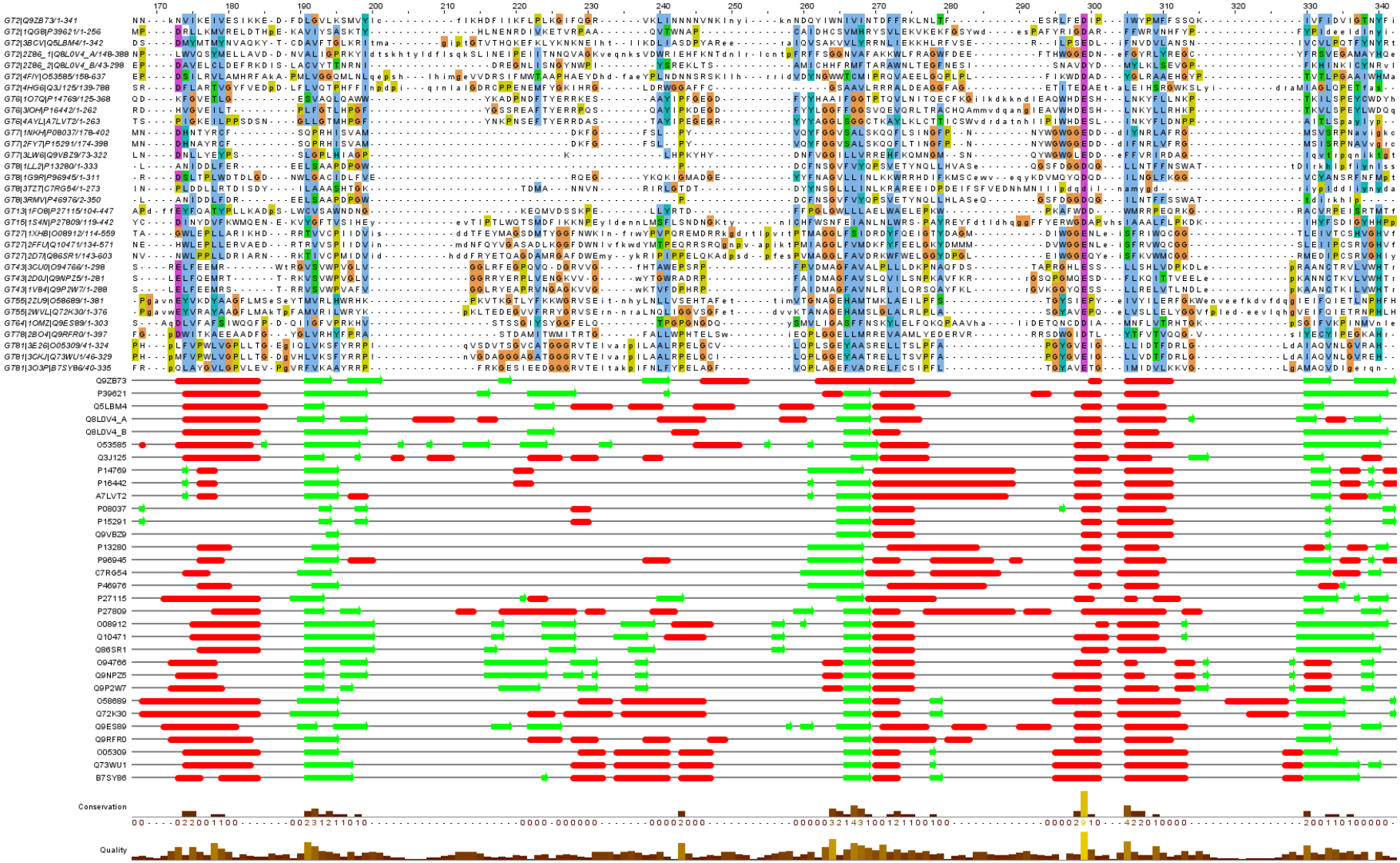


Fuente: C.A.E. Hicks, E.N. Barker, C. Brady, C.R. Stokes, C.R. Helps, and S. Tasker. Non-ribosomal phylogenetic exploration of Mollicute species: New insights into haemoplasma taxonomy. *Infect Genet Evol. 2014* Apr; 23(**100**): 99–105.

Anexo 2. Alineamiento de GTAs cristalizadas (arriba) mediante el perfil HMM construido y superposición de las estructuras secundarias (abajo).



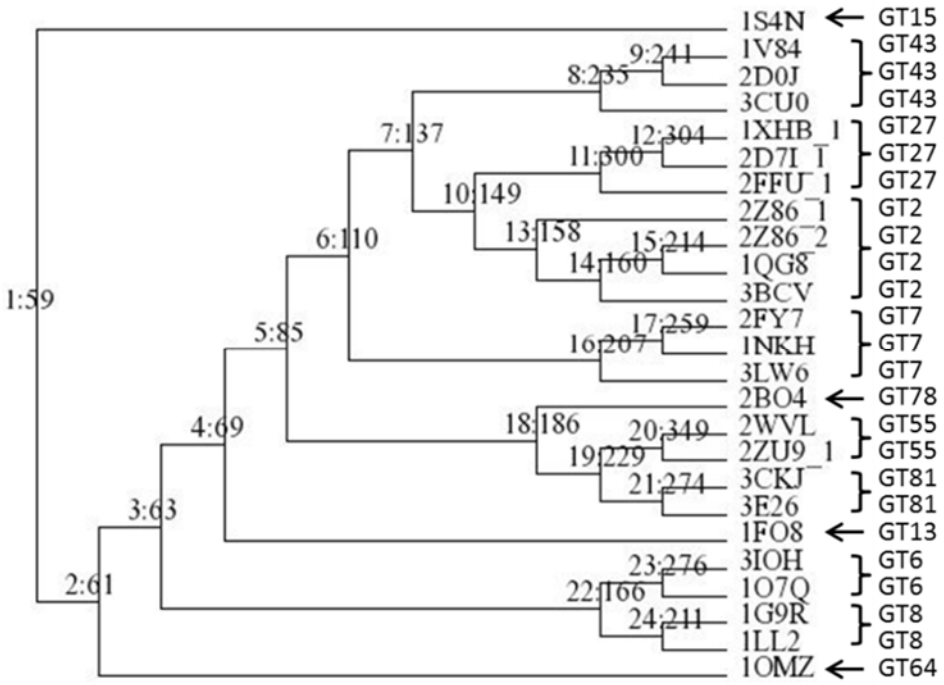
Anexo 2. Alineamiento de GTAs cristalizadas (arriba) mediante el perfil HMM construido y superposición de las estructuras secundarias (abajo), continuación.



Anexo 3. Características de los cristales de proteínas GTA.

PDB	Familia	RV	RV (aa)	Ligando	Resolución
1QG8	GT2	131-154	23	No	1.50
3BCV	GT2	125-168	43	No	2.35
2Z86_1	GT2	272-329	57	UGA	2.40
2Z86_2	GT2	553-575	22	UDP	2.40
4FIY	GT2	290-341	51	UDP	3.10
4HG6	GT2	280-318	38	UDP BGC	3.25
1O7Q	GT6	251-276	25	UDP GAL NAG	1.30
3IOH	GT6	237-262	25	No	1.25
4AYL	GT6	126-151	25	No	1.92
1NKH	GT7	278-286	8	UDP PG4	2.00
2FY7	GT7	274-282	8	No	1.70
3LW6	GT7	172-177	5	UDP	1.81
1LL2	GT8	129	1	UPG	1.90
1G9R	GT8	135-149	14	UPF	2.00
3TZT	GT8	136-152	16	No	2.10
3RMV	GT8	129	1	UDP	1.82
1FO8	GT13	248-264	16	No	1.40
1S4N	GT15	280-321	41	No	2.01
1XHB	GT27	241-288	47	No	2.50
2FFU	GT27	256-303	47	UDP	1.64
2D7I	GT27	269-314	45	UDP NGA	2.50
3CU0	GT43	222-248	26	UDP GAL	1.90
2D0J	GT43	213-239	26	No	2.00
1V84	GT43	223-250	27	UDP GAL	1.82
2ZU9	GT55	206-244	38	GDP	2.00
2WVL	GT55	205-244	39	GDD	2.81
1OMZ	GT64	187-209	22	UD2	2.10
2BO4	GT78	133-159	26	No	1.95
3E26	GT81	170-205	35	No	2.50
3CKJ	GT81	174-210	36	No	1.80
3O3P	GT81	170-202	32	GDP	2.53

Anexo 4. Árbol estructural generado por POSA.



Árbol generado por el servidor POSA, teniendo en cuenta solo la superposición estructural. El primer nº representa el índice y el segundo el nº de aminoácidos compartidos por el núcleo común. En este árbol, 2Z86_1 sí es agrupado junto a las GT2.

Anexo 5. Alineamientos para Modeller, para el modelado de MG517.

Subrayado en amarillo la región conservada, en azul la región variable para cada plantilla, de color verde las zonas solapantes:

Modelo híbrido 1O7Q/2Z86_2 (Modelo 1):

```
>P1;Q9ZB73
sequence:Q9ZB73:1::224::GT517:MycoplasmaGenitalium:0.00:0.00
MDKLVSILVPCYK--SKPFLKRFFNSLLKQDL---NQAKIIFFNDNVA-DETYEVLQKFKKEHNNL-AIEVY
CD------KQNEGIGKVRDKLVNLV----------------TTPYFYFIDPDDCFNNKNVIKEIVESIKKE
DFD-LGVLKSMVYLCFLKHDFIIKFLPLKGIFQGRVKLINNNNVNKLNYI------KNND--QYIWNIVINT
DFFRKLNLTF-----------ESR---L--FEDIPIWYPMFFSSQKIVFIDVIGTNYFI...*

>P1;2Z86_2
structureX:2Z86_2:430:A:632:A:GT2:Escherichiacoli:0.00:0.00
RVPLVSIYIPAYN--CSKYIVRCVESALNQTI---TDLEVCICDDGST-DDTLRILQEHYANH---PRVRFI
SQ------KNKGIGSASNTAVRLC----------------RGFYIGQLDSDDFLEP-DAVELCLDEFRKD
LSLACVYT------------------------------------------------------------RMFT
ARAWNLTEGF----------NESI-SNA---VDYDMYLKLSEVGP-FKHINKICYNRVL..--*

>P1;1O7Q
structureX:1O7Q::L::L:GT6:Bostaurus:0.00:0.00
------------------------------------------------------------------------
------------------------------------------------------------------------
----SVAQLQAWWYKADPNDF----TYERR--KESAAYI----PFGEGDFY-----------YHAAIFGGT
-------------------------------------------------------------..*
```

Modelo híbrido 2FFU/2Z86_2 (Modelo 2):

```
>P1;Q9ZB73
sequence:Q9ZB73:1::224::GT517:MycoplasmaGenitalium:0.00:0.00
MDKLVSILVPCYK--SKPFLKRFFNSLLKQDL---NQAKIIFFNDNVA-DETYEVLQKFKKEHNNL-AIEVY
CD------KQNEGIGKVRDKLVNLV----------------TTPYFYFIDPDDCFNNKNVIKEIVESIKKE
DFD-LGVLKSMVYLCFLKHDFIIKFLPLKGIFQGRVKLINNNNVNKLNYI------KNND--QYIWNIVINT
DFFRKLNLTF-----------ESR---L--FEDIPIWYPMFFSSQKIVFIDVIGTNYFI..........*

>P1;2Z86_2
structureX:2Z86_2:430:A:632:A:GT2:Escherichiacoli:0.00:0.00
RVPLVSIYIPAYN--CSKYIVRCVESALNQTI---TDLEVCICDDGST-DDTLRILQEHYANH---PRVRFI
SQ------KNKGIGSASNTAVRLC----------------RGFYIGQLDSDDFLEP-DAVELCLDEFRKD
LSLACVYT------------------------------------------------------------RMFT
ARAWNLTEGF----------NESI-SNA---VDYDMYLKLSEVGP-FKHINKICYNRVL..---------*

>P1;2FFU
structureX:2FFU::H::H:GT2:Homosapiens:0.00:0.00
------------------------------------------------------------------------
------------------------------------------------------------------------
----RVVSPIIDVINMDNFQYVGASADLKGGFDWNLVFKWDYMTPEQRRSRQGNPVAPIKTPMIAGGLFVMD
-------------------------------------------------------------.........*
```

Modelo híbrido 2Z86_1/2Z86_2 (Modelo 3):

```
>P1;Q9ZB73
sequence:Q9ZB73:1::224::GT517:MycoplasmaGenitalium:0.00:0.00
MDKLVSILVPCYK--SKPFLKRFFNSLLKQDL---NQAKIIFFNDNVA-DETYEVLQKFKKEHNNL-AIEVY
CD------KQNEGIGKVRDKLVNLV----------------TTPYFYFIDPDDCFNNKNVIKEIVESIKKE
DFD-LGVLKSMVYLCFLKHDFIIKFLPLKGIFQGRVKLINNNNVNKLNYI------KNND--QYIWNIVINT
DFFRKLNLTF-----------ESR---L--FEDIPIWYPMFFSSQKIVFIDVIGTNYFI..*


>P1;2Z86_2
structureX:2Z86_2:430:A:632:A:GT2:Escherichiacoli:0.00:0.00
```

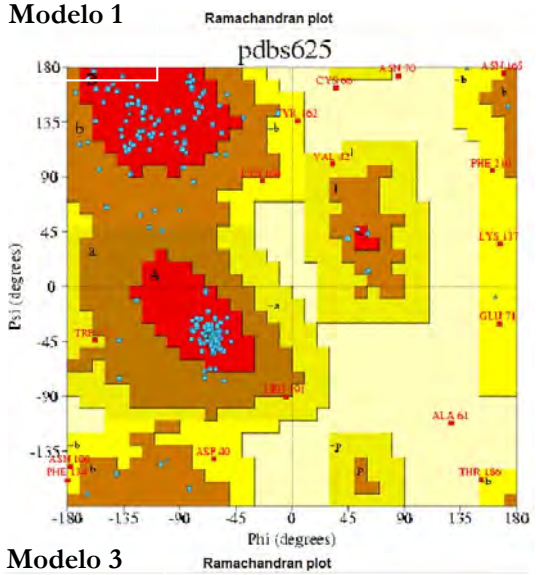
```
RVPLVSIYIPAYN--CSKYIVRCVESALNQTI---TDLEVCICDDGST-DDTLRILQEHYANH---PRVRFI
SQ------KNKGIGSASNTAVRLC----------------RGFYIGQLDSDDFLEP-DAVELCLDEFRKD
LSLACVYT-------------------------------------------------SAMICHHFRMFT
ARAWNLTEGF----------NESI-SNA---VDYDMYLKLSEVGP-FKHINKICYNRVL..*

>P1;2Z86_1
structureX:2Z86_1::O::O:GT2:Escherichiacoli:0.00:0.00
------------------------------------------------------------------------
------------------------------------------------------------------------
----ALIGPRKYIDTSKHTYLDFLSQ--KSLINEIPESVDWRIEHFKNTDNLRLCNTPFRF--FSG------
------------------------------------------------------------*
```

Modelo híbrido 3CU0/2Z86_2 (Modelo 4):

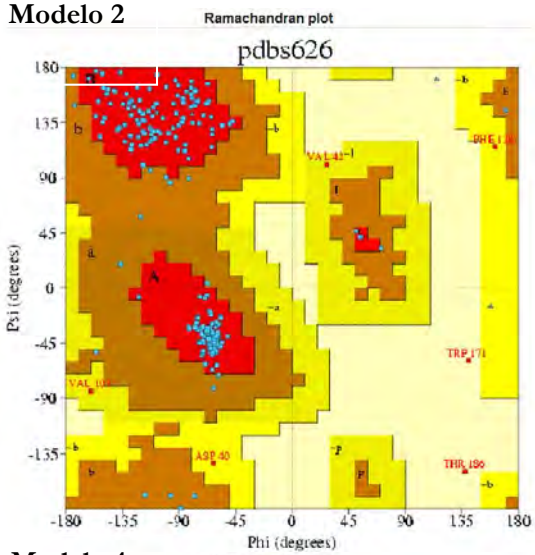
```
>P1;Q9ZB73
sequence:Q9ZB73:1::224::GT517:MycoplasmaGenitalium:0.00:0.00
MDKLVSILVPCYK--SKPFLKRFFNSLLKQDL---NQAKIIFFNDNVA-DETYEVLQKFKKEHNNL-AIEVY
CD------KQNEGIGKVRDKLVNLV----------------TTPYFYFIDPDDCFNNKNVIKEIVESIKKE
DFD-LGVLKSMVYLCFLKHDFIIKFLPLKGIFQGRVKLINNNNVNKLNYI------KNND--QYIWNIVINT
DFFRKLNLTF-----------ESR---L--FEDIPIWYPMFFSSQKIVFIDVIGTNYFI.....*

>P1;2Z86_2
structureX:2Z86_2:430:A:632:A:GT2:Escherichiacoli:0.00:0.00
RVPLVSIYIPAYN--CSKYIVRCVESALNQTI---TDLEVCICDDGST-DDTLRILQEHYANH---PRVRFI
SQ------KNKGIGSASNTAVRLC----------------RGFYIGQLDSDDFLEP-DAVELCLDEFRKD
LSLACVYT-------------------------------------------------------------RMFT
ARAWNLTEGF----------NESI-SNA---VDYDMYLKLSEVGP-FKHINKICYNRVL..---*

>P1;3CU0
structureX:3CU0::Y::Y:GT43:Homosapiens:0.00:0.00
------------------------------------------------------------------------
------------------------------------------------------------------------
----VSVWPVGLVGGLRFEGPQVQDG--RVVGFH--TAW---EPSRPFPV-------------DMAGFAVA
------------------------------------------------------------...*
```

Anexo 6. Superposición por PROCHECK de los ángulos ψ (phi) y ϕ (psi) de los diferentes modelos en el diagrama de Ramachandran.

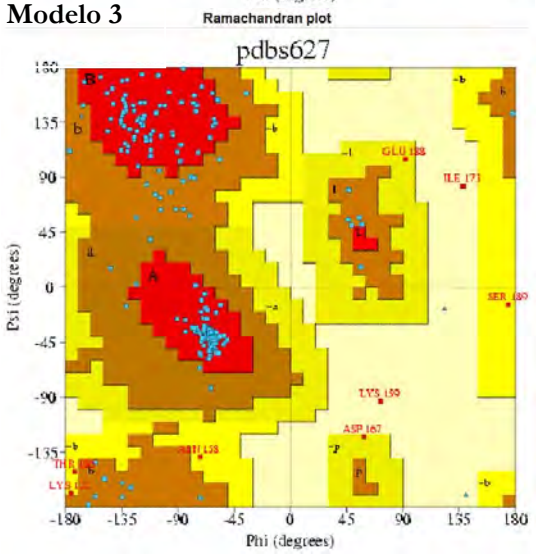
Modelo 1



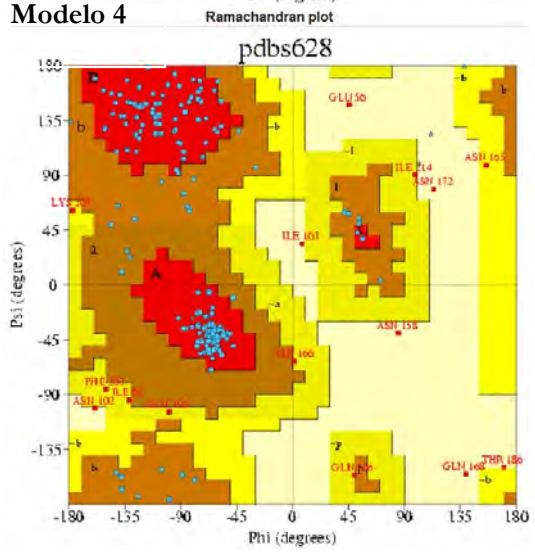
Modelo 2



Modelo 3



Modelo 4

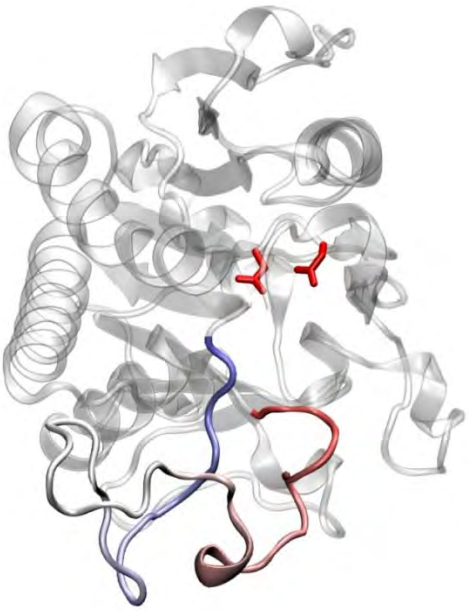


- Modelo 1: 3 prohibidos. 98,6 % dentro de los límites.
- Modelo 2: 2 prohibidos. 99,1 % dentro de los límites.
- Modelo 3: 2 prohibidos. 99,1 % dentro de los límites.
- Modelo 4: 4 prohibidos. 98,2 % dentro de los límites.

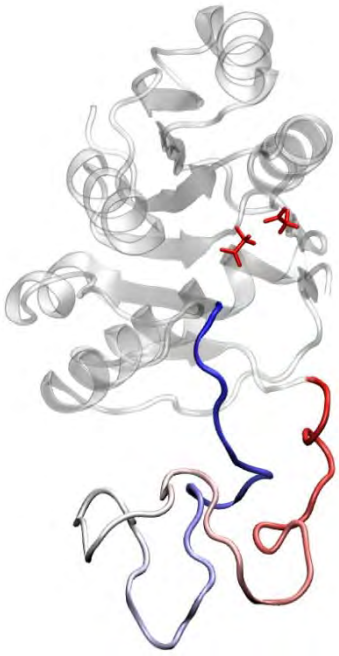
Anexo 7. Modelado de las estructuras 1 a 4.

A la izquierda la estructura original, a la derecha modelo híbrido generado por Modeller. Ambas estructuras tienen coloreadas la región variable, de rojo a azul, pasando por blanco y según la posición del aminoácido, por lo que igual color indica misma posición.

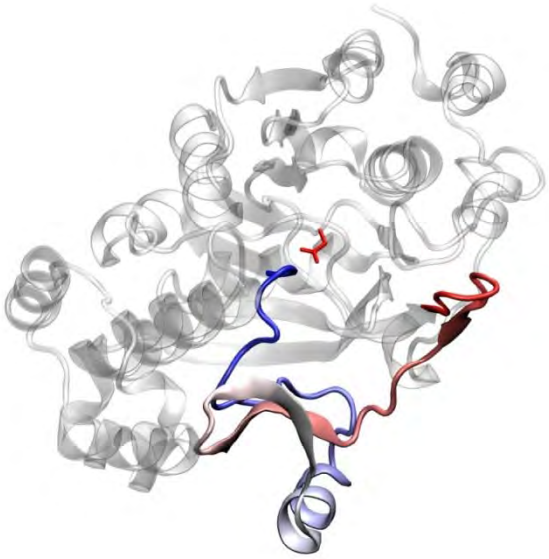
1O7Q



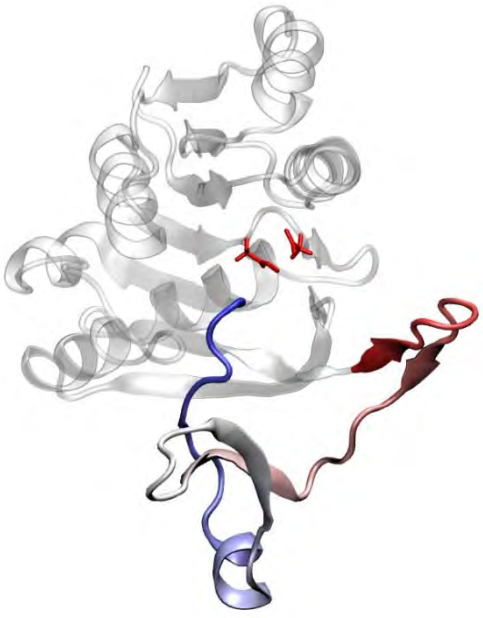
Modelo 1



2FFU

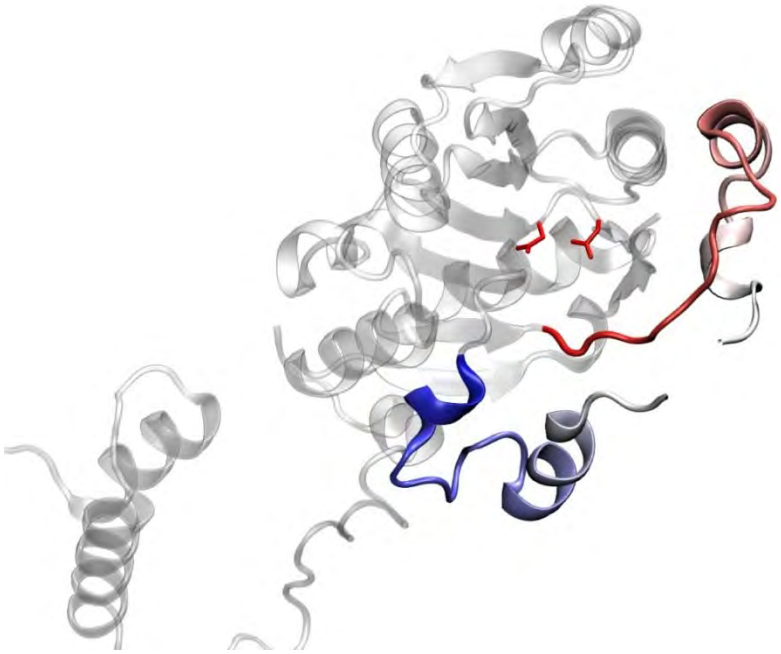


Modelo 2

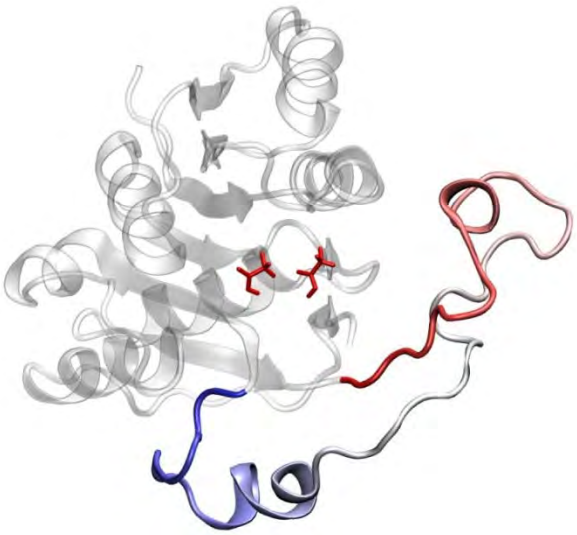


Anexo 7. Modelado de las estructuras 1 a 4, continuación.

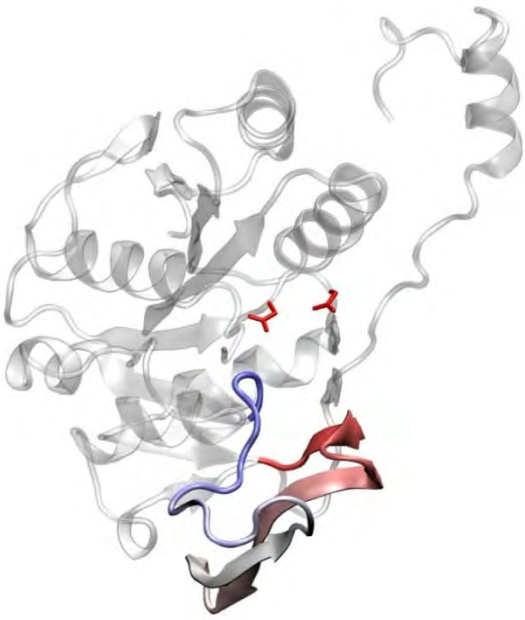
2Z86_1



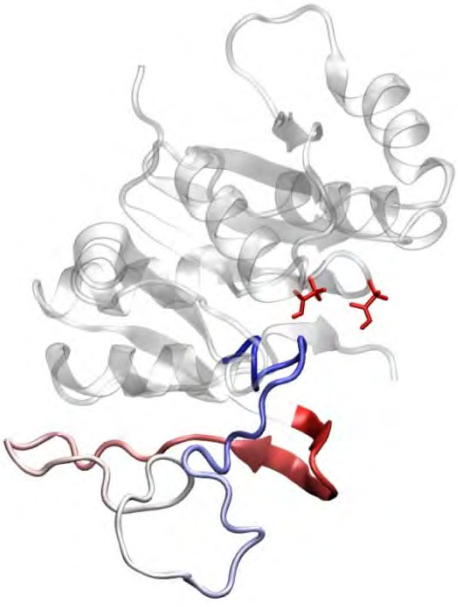
Modelo 3



3CU0



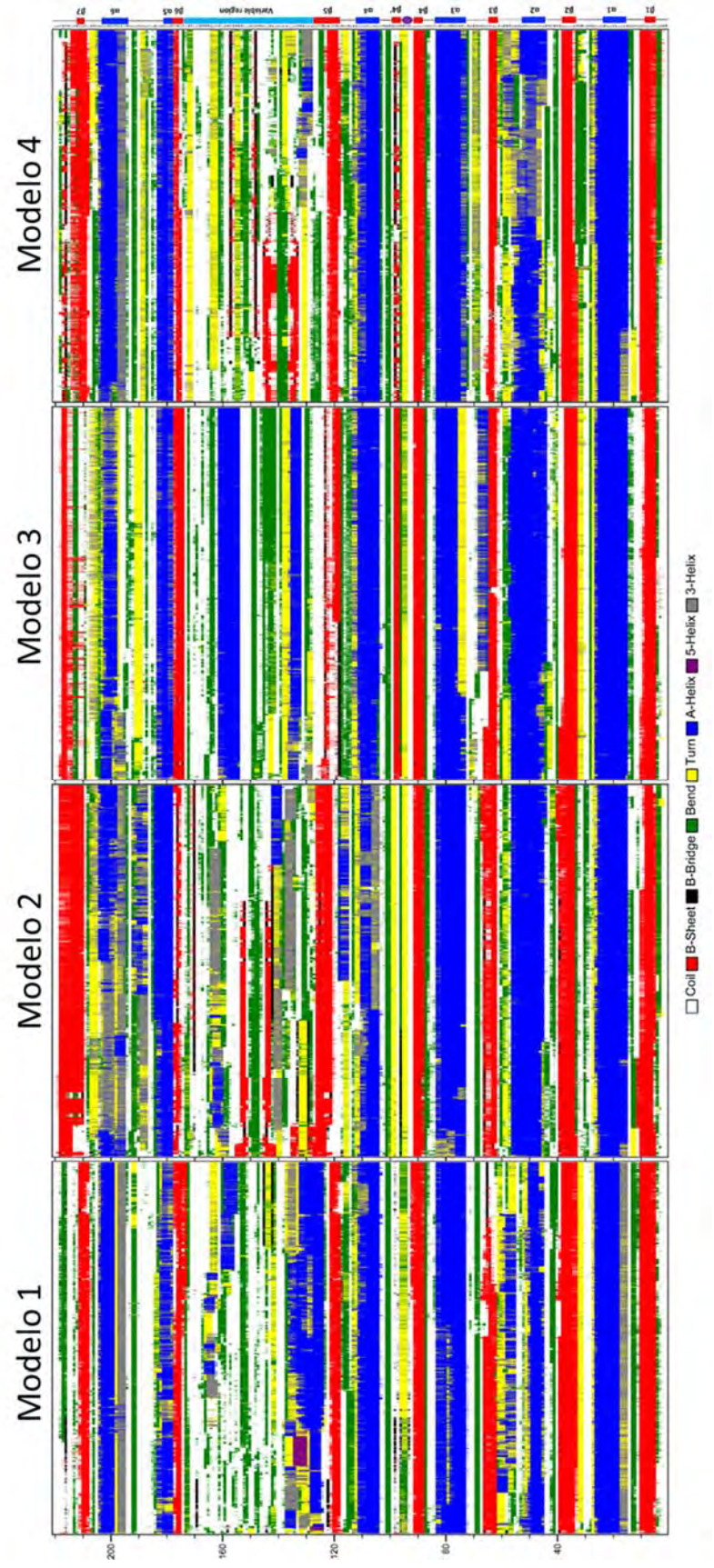
Modelo 4



Anexo 8. Tipos de átomos, cargas y masas del UDPGlc.

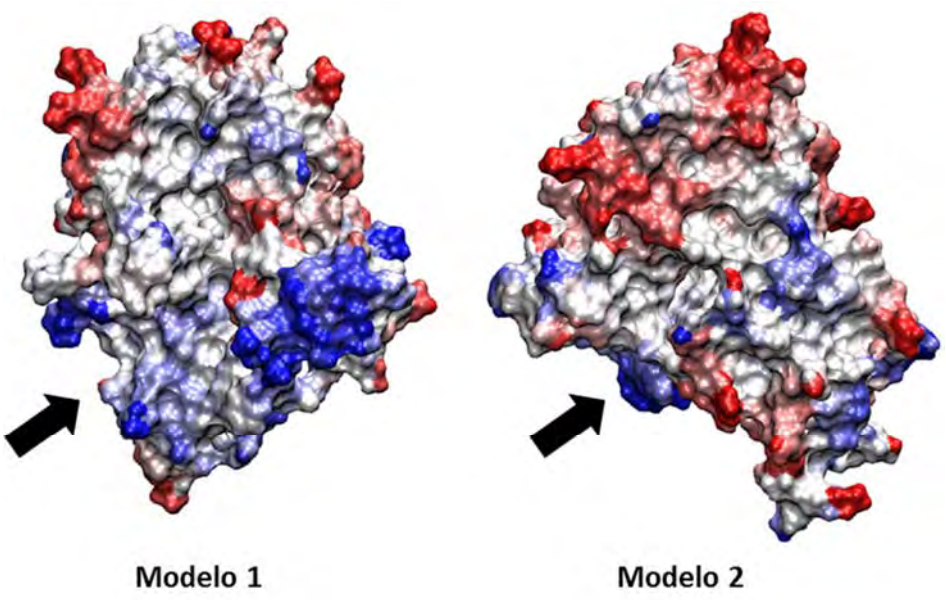
```
[ atoms ]
;   nr       type  resnr residue  atom   cgnr     charge       mass  typeB
chargeB
     1         OH      1    UGC    O6r      1   -0.74930  16.000000
     2         HO      1    UGC   H6Or      2    0.44910   1.000000
     3         CT      1    UGC    C6r      3    0.34650  12.000000
     4         H1      1    UGC    H61      4   -0.04780   1.000000
     5         H1      1    UGC    H62      5    0.04350   1.000000
     6         CT      1    UGC    C5r      6    0.40550  12.000000
     7         H1      1    UGC    H5r      7    0.05420   1.000000
     8         OS      1    UGC    O5r      8   -0.56740  16.000000
     9         CT      1    UGC    C4r      9   -0.09030  12.000000
    10         H1      1    UGC    H4r     10    0.07670   1.000000
    11         OH      1    UGC    O4r     11   -0.72140  16.000000
    12         HO      1    UGC   H4Or     12    0.47350   1.000000
    13         CT      1    UGC    C3r     13    0.46900  12.000000
    14         H1      1    UGC    H3r     14   -0.02180   1.000000
    15         OH      1    UGC    O3r     15   -0.73170  16.000000
    16         HO      1    UGC   H3Or     16    0.42950   1.000000
    17         CT      1    UGC    C2r     17    0.18730  12.000000
    18         H1      1    UGC    H2r     18    0.10560   1.000000
    19         OH      1    UGC    O2r     19   -0.67000  16.000000
    20         HO      1    UGC   H2Or     20    0.41140   1.000000
    21         CT      1    UGC    C1r     21    0.20050  12.000000
    22         H2      1    UGC    H1r     22    0.12740   1.000000
    23         OS      1    UGC    O1r     23   -0.47630  16.000000
    24          P      1    UGC     P2     24    1.09030  31.000000
    25         O2      1    UGC    Op6     25   -0.79280  16.000000
    26         O2      1    UGC    Op5     26   -0.79280  16.000000
    27         OS      1    UGC    Op4     27   -0.42330  16.000000
    28          P      1    UGC     P1     28    1.11290  31.000000
    29         O2      1    UGC    Op3     29   -0.79670  16.000000
    30         O2      1    UGC    Op2     30   -0.79670  16.000000
    31         OS      1    UGC    Op1     31   -0.49740  16.000000
    32         CT      1    UGC    C5m     32    0.05580  12.000000
    33         H1      1    UGC   H5m1     33    0.06790   1.000000
    34         H1      1    UGC   H5m2     34    0.06790   1.000000
    35         CT      1    UGC    C4m     35    0.10650  12.000000
    36         H1      1    UGC    H4m     36    0.11740   1.000000
    37         CT      1    UGC    C3m     37    0.20220  12.000000
    38         H1      1    UGC    H3m     38    0.06150   1.000000
    39         OH      1    UGC    O3m     39   -0.65410  16.000000
    40         HO      1    UGC   HOm3     40    0.43760   1.000000
    41         CT      1    UGC    C2m     41    0.06700  12.000000
    42         H1      1    UGC    H2m     42    0.09720   1.000000
    43         OH      1    UGC    O2m     43   -0.61390  16.000000
    44         HO      1    UGC   HOm2     44    0.41860   1.000000
    45         OS      1    UGC    O4m     45   -0.35480  16.000000
    46         CT      1    UGC    C1m     46    0.06740  12.000000
    47         H2      1    UGC    H1m     47    0.18240   1.000000
    48         N*      1    UGC     N1     48    0.04180  14.000000
    49         CM      1    UGC     C6     49   -0.11260  12.000000
    50         H4      1    UGC     H6     50    0.21880   1.000000
    51         CM      1    UGC     C5     51   -0.36350  12.000000
    52         HA      1    UGC     H5     52    0.18110   1.000000
    53          C      1    UGC     C4     53    0.59520  12.000000
    54          O      1    UGC     O4     54   -0.57610  16.000000
    55         NA      1    UGC     N3     55   -0.35490  14.000000
    56          H      1    UGC     H3     56    0.31540   1.000000
    57          C      1    UGC     C2     57    0.46870  12.000000
    58          O      1    UGC     O2     58   -0.54770  16.000000
```

Anexo 9. Gráfico de evolución DSSP de cada estructura a lo largo de la simulación MD.

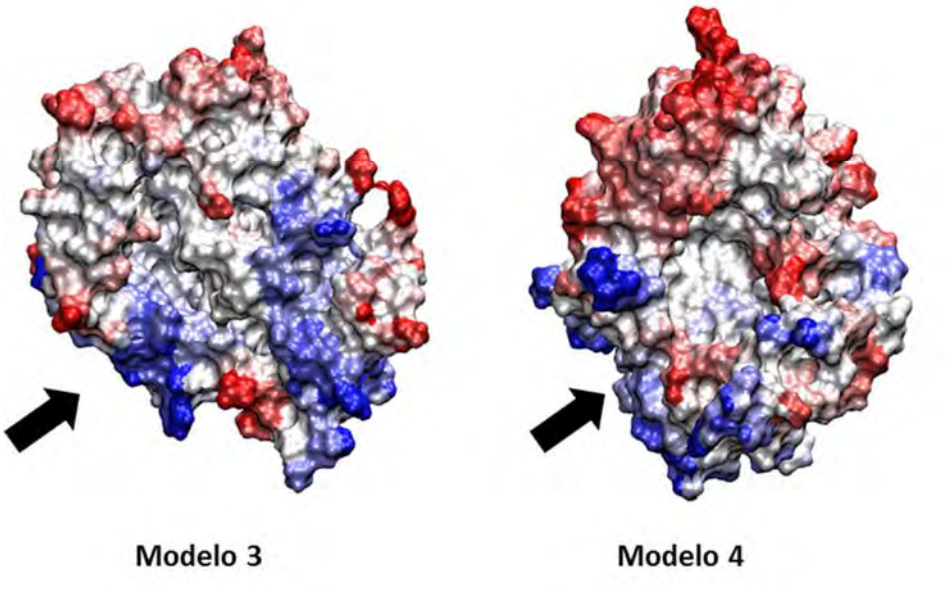


Anexo 10. Representación SAP de las estructuras seleccionadas tras la MD.

La flecha muestra el *patch* hidrofóbico donde podría estar ubicada la hélice 7 en la GT MG517. Cuanto más rojo más polar, cuanto más azul más apolar.

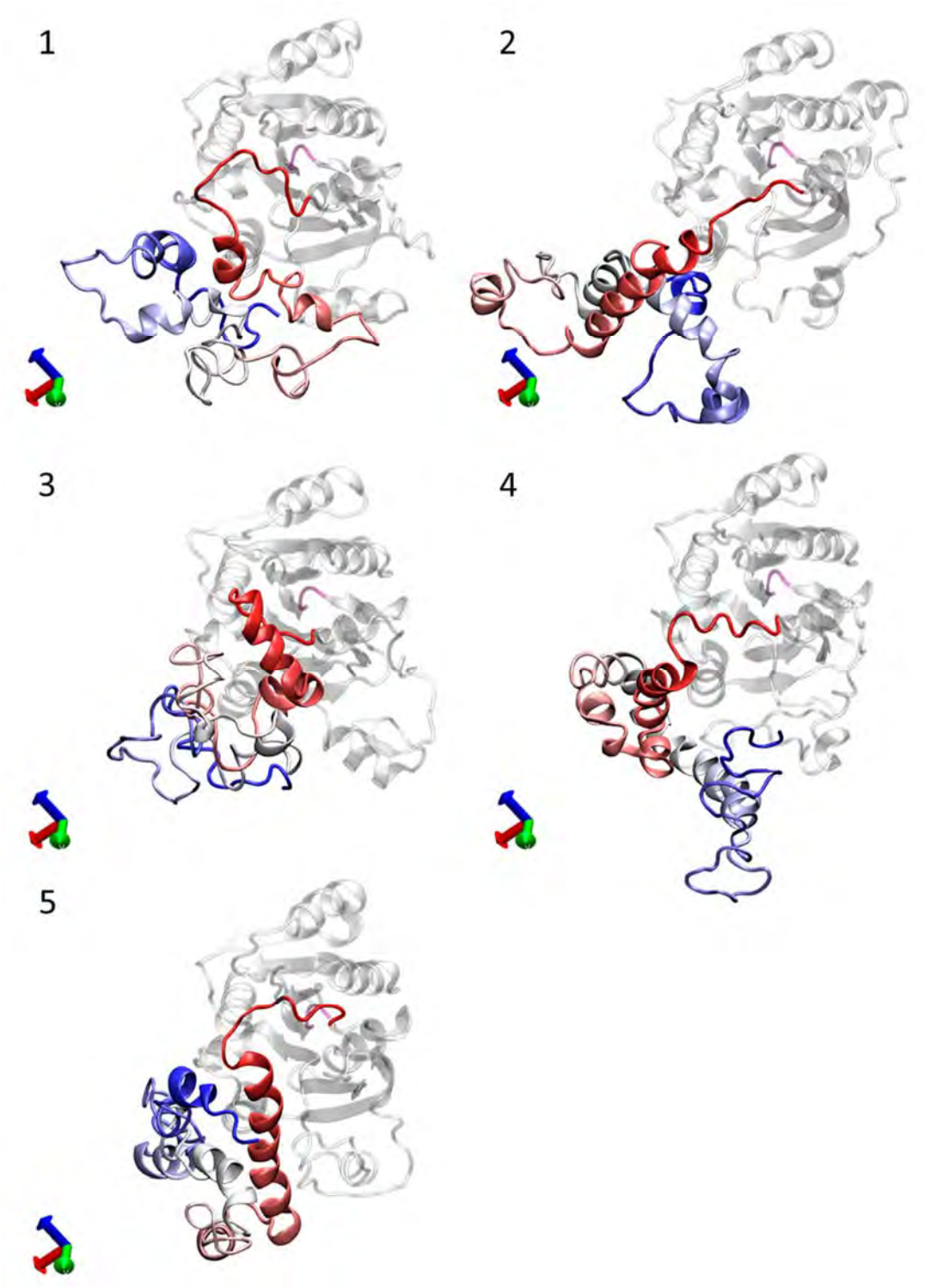


Modelo 1 Modelo 2



Modelo 3 Modelo 4

Anexo 11. Resultados de I-TASSER para la secuencia completa de MG517.



Estructuras generadas por I-TASSER: 1. C-score -2,58. 2. C-score -2,77. 3. C-score -2,82. 4. C-score-3,25. 5. C-score -3,68. Coloreada transparente la topología consenso (residuos 1 a 221), motivo DXD en magenta. Región C-terminal coloreada de rojo (residuo 222) a a azul (residuo 341).

Anexo 12. Comparativa entre el valor de los diedros φ y ψ para el cristal y el modelo generado.

Residuo		**LA** φ	**LI** φ	Var φ %	**LA** ψ	**LI** ψ	Var ψ %
G254	Cristal	1,96	1,89	2,40	-1,78	-1,62	4,94
	Modelo	1,97	1,90	2,22	-1,78	-1,63	4,80
V255	Cristal	-2,46	-2,75	8,99	2,47	2,20	8,74
	Modelo	-2,46	-2,74	9,02	2,47	2,21	8,21
R256	**Cristal**	**-1,90**	**-1,87**	**0,78**	**2,21**	**2,31**	**3,38**
	Modelo	**-1,90**	**-1,88**	**0,63**	**2,20**	**2,31**	**3,22**
A257	**Cristal**	**-1,95**	**-0,93**	**32,57**	**2,42**	**-1,52**	**74,39**
	Modelo	**-1,94**	**-0,92**	**32,49**	**2,42**	**-1,52**	**74,51**
H258	**Cristal**	**-2,77**	**-3,06**	**9,27**	**2,69**	**2,09**	**19,33**
	Modelo	**-2,77**	**-3,06**	**9,30**	**2,80**	**2,08**	**23,06**
R259	**Cristal**	**-1,12**	**-1,34**	**6,88**	**1,55**	**-2,80**	**61,41**
	Modelo	**-1,33**	**-1,17**	**5,07**	**1,68**	**2,99**	**41,66**
N260	**Cristal**	**-0,99**	**-2,17**	**37,57**	**2,91**	**1,96**	**30,37**
	Modelo	**-1,03**	**-1,86**	**26,48**	**2,91**	**2,23**	**21,57**
R261	Cristal	-2,46	-2,36	3,13	2,62	2,59	0,83
	Modelo	-2,46	-2,42	1,26	2,62	2,59	1,02
P262	Cristal	-1,22	-1,05	5,48	2,58	2,47	3,68
	Modelo	-1,22	-1,04	5,97	2,59	2,46	3,84
L263	Cristal	-1,11	-1,02	2,72	-0,66	-0,66	0,00
	Modelo	-1,11	-1,03	2,63	-0,66	-0,65	0,27

Anexo 13. Comparación de los diedros del bucle RAHRN y residuos anejos entre el cristal, modelo, minimización de energía durante el equilibrado de la MD y MD.

LA φ

Residuo	Cristal	Modelo	EM1	EM2	EM3	EM4	EM5	Inicio MD	MD(Mon1/Mon2)
G254	1,96	1,97	1,42	1,39	1,39	1,39	1,39	1,90	1,99 1,70
V255	-2,46	-2,46	-2,85	-2,86	-2,84	-2,84	-2,84	-2,17	-2,25 -1,86
R256	**-1,90**	**-1,90**	**-2,31**	**-2,24**	**-2,19**	**-2,16**	**-2,16**	**-1,33**	**-0,91** **-2,48**
A257	**-1,95**	**-1,94**	**-1,84**	**-1,85**	**-1,86**	**-1,84**	**-1,83**	**-1,98**	**-2,79** **-1,69**
H258	**-2,77**	**-2,77**	**-2,54**	**-2,61**	**-2,55**	**-2,58**	**-2,58**	**-2,99**	**-1,87** **-2,66**
R259	**-1,12**	**-1,33**	**-1,39**	**-1,35**	**-1,34**	**-1,33**	**-1,33**	**-1,14**	**-1,36** **-1,70**
N260	**-0,99**	**-1,03**	**-1,31**	**-1,34**	**-1,32**	**-1,33**	**-1,33**	**-1,48**	**-1,62** **-1,09**
R261	-2,46	-2,46	-2,31	-2,19	-2,17	-2,10	-2,09	-1,58	-2,01 -1,39
P262	-1,22	-1,22	-1,45	-1,40	-1,41	-1,40	-1,39	-1,33	-1,18 -1,13
L263	-1,11	-1,11	-1,28	-1,29	-1,29	-1,30	-1,29	-1,37	-1,17 -0,82

LA ψ

	Cristal	Modelo	EM1	EM2	EM3	EM4	EM5	Inicio MD	MD (Mon1/Mon2)
G254	-1,78	-1,78	-1,22	-1,13	-1,14	-1,13	-1,14	-1,99	-1,97 -1,55
V255	2,47	2,47	2,64	2,59	2,56	2,54	2,54	1,71	2,71 2,79
R256	**2,21**	**2,20**	**2,15**	**2,17**	**2,14**	**2,11**	**2,10**	**2,35**	**2,82** **2,09**
A257	**2,42**	**2,42**	**2,48**	**2,54**	**2,48**	**2,49**	**2,49**	**2,58**	**0,75** **2,22**
H258	**2,69**	**2,80**	**3,09**	**3,10**	**3,09**	**3,08**	**3,07**	**2,79**	**2,51** **2,82**
R259	**1,55**	**1,68**	**2,11**	**2,08**	**2,05**	**2,05**	**2,04**	**2,34**	**2,72** **1,95**
N260	**2,91**	**2,91**	**2,62**	**2,45**	**2,45**	**2,41**	**2,40**	**2,05**	**2,30** **2,22**
R261	2,62	2,62	2,67	2,75	2,76	2,77	2,76	2,75	2,85 2,52
P262	2,58	2,59	2,91	2,87	2,88	2,88	2,87	2,93	2,56 2,64
L263	-0,66	-0,66	-0,43	-0,45	-0,45	-0,44	-0,44	-0,72	-0,55 -0,78

LI φ

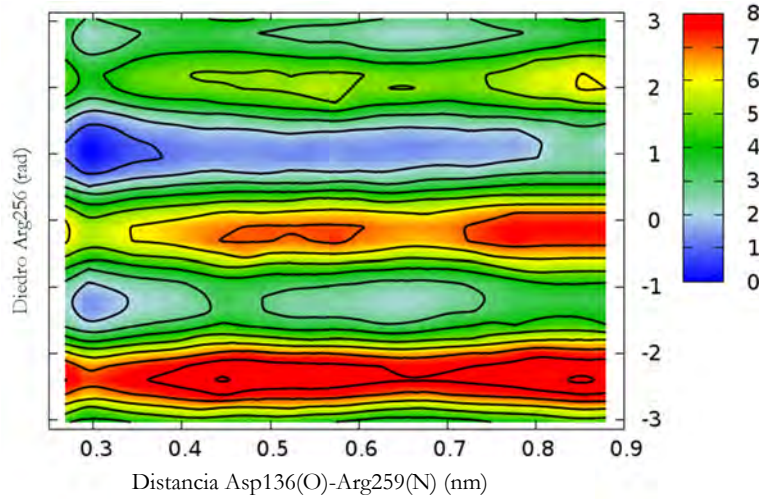
Residuo	Cristal	Modelo	EM1	EM2	EM3	EM4	EM5	Inicio MD	MD (Mon1/Mon2)
G254	1,89	1,90	1,38	1,37	1,38	1,39	1,40	1,27	2,52 -1,32
V255	-2,75	-2,74	-2,95	-2,96	-2,95	-2,97	-2,98	-1,43	-1,49 -1,05
R256	**-1,87**	**-1,88**	**-2,18**	**-2,15**	**-2,15**	**-2,15**	**-2,15**	**-1,80**	**-2,91** **-2,04**
A257	**-0,93**	**-0,92**	**-1,50**	**-1,52**	**-1,53**	**-1,51**	**-1,51**	**-2,17**	**-1,68** **-1,71**
H258	**-3,06**	**-3,06**	**1,19**	**1,19**	**1,17**	**1,17**	**1,16**	**-1,91**	**-1,55** **-2,78**
R259	**-1,34**	**-1,17**	**-0,64**	**-0,65**	**-0,66**	**-0,68**	**-0,68**	**-0,84**	**-1,02** **-1,29**
N260	**-2,17**	**-1,86**	**-1,17**	**-1,19**	**-1,20**	**-1,21**	**-1,22**	**-1,48**	**-1,17** **-0,93**
R261	-2,36	-2,42	-2,47	-2,48	-2,48	-2,51	-2,51	-2,82	-2,12 -1,93
P262	-1,05	-1,04	-1,11	-1,14	-1,16	-1,17	-1,18	-0,98	-1,04 -0,94
L263	-1,02	-1,03	-2,80	-2,73	-2,74	-2,73	-2,73	-0,93	-1,03 -0,95

LI ψ

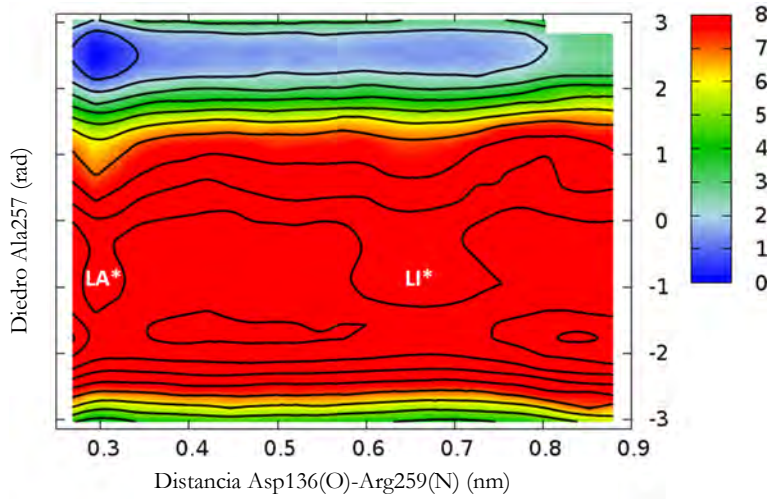
	Cristal	Modelo	EM1	EM2	EM3	EM4	EM5	Inicio MD	MD (Mon1/Mon2)
G254	-1,62	-1,63	-1,01	-0,99	-0,98	-0,93	-0,93	-2,50	3,05 2,79
V255	2,20	2,21	2,55	2,55	2,56	2,58	2,57	2,31	2,38 2,53
R256	**2,31**	**2,31**	**2,38**	**2,39**	**2,39**	**2,40**	**2,40**	**2,59**	**2,78** **2,82**
A257	**-1,52**	**-1,52**	**-0,01**	**0,00**	**0,02**	**0,02**	**0,03**	**2,46**	**2,60** **-2,86**
H258	**2,09**	**2,08**	**0,96**	**0,97**	**0,97**	**0,99**	**1,00**	**2,54**	**2,43** **2,32**
R259	**-2,80**	**2,99**	**2,12**	**2,13**	**2,13**	**2,15**	**2,16**	**2,65**	**2,31** **2,19**
N260	**1,96**	**2,23**	**2,46**	**2,48**	**2,49**	**2,50**	**2,50**	**2,68**	**2,28** **1,66**
R261	2,59	2,59	2,94	2,97	2,98	3,00	3,00	2,61	2,55 2,34
P262	2,47	2,46	-2,26	-2,36	-2,35	-2,40	2,40	2,01	2,63 2,67
L263	-0,66	-0,65	-0,61	-0,61	-0,60	-0,57	-0,57	-0,56	-0,51 -0,45

Anexo 14: Meta4Y6Napo4. Proyecciones de energía:

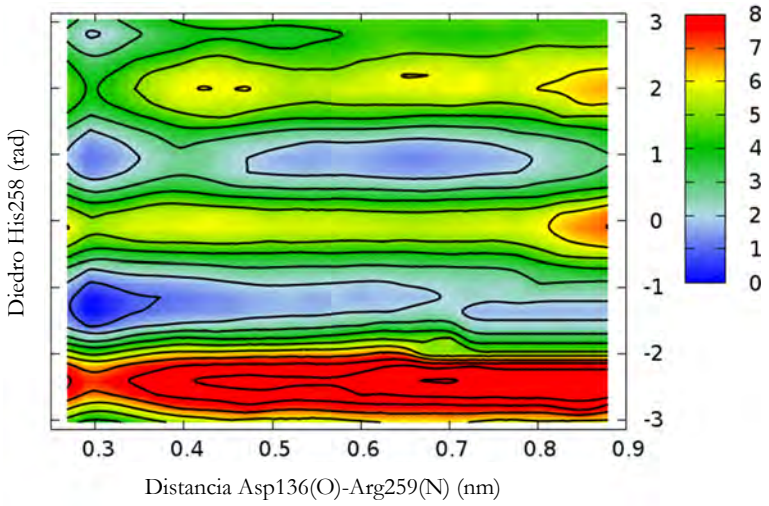
Isotermas: 1 kcal/mol ± 0,5 kcal/mol



Proyección de la energía para el diedro de la Arg256 sobre la distancia. El metaestado más estable se encuentra en la conformación LA, con un valor de diedro de 1, si bien existen otros dos metaestados de menor energía, pero igual entre ellos para el valor 1 en LI y -1 en LA.



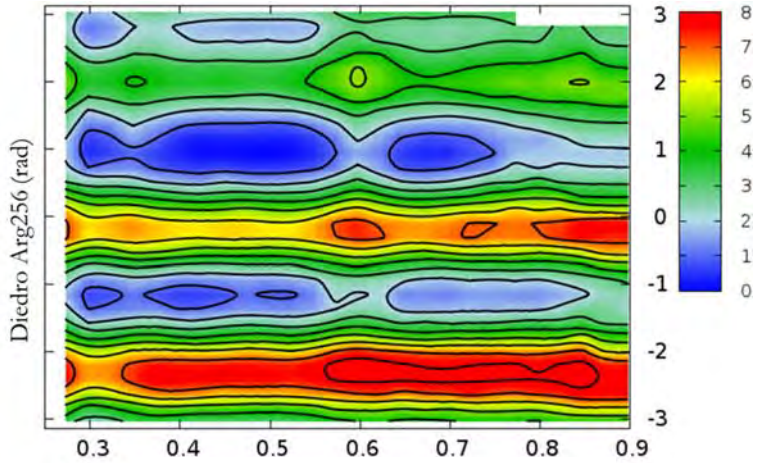
Proyección de la energía para el diedro de la Ala257 sobre la distancia. El metaestado más estable es la conformación LA, con un valor de diedro ψ para Ala257 entre 2 y 3 rad. Comienza a observarse el metaestado LI*, aunque con una barrera muy alta y otro metaestado LA* no visto anteriormente, también con una gran barrera



Proyección de la energía para el diedro de la His258 sobre la distancia. El metaestado más estable es para el valor -1 en conformación LA. El valor 1 tiene la misma energía en las conformaciones LA y LI.

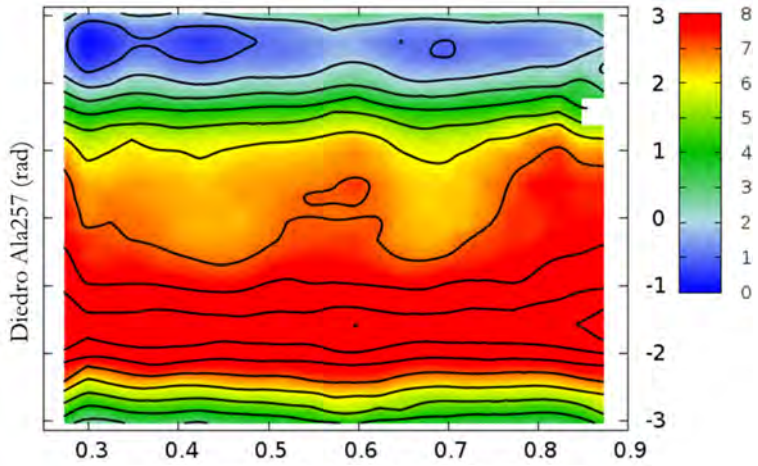
Anexo 15: Meta4Y6NUDPGlc4. Proyecciones de energía:

Isotermas: 1 kcal/mol ± 0,5 kcal/mol



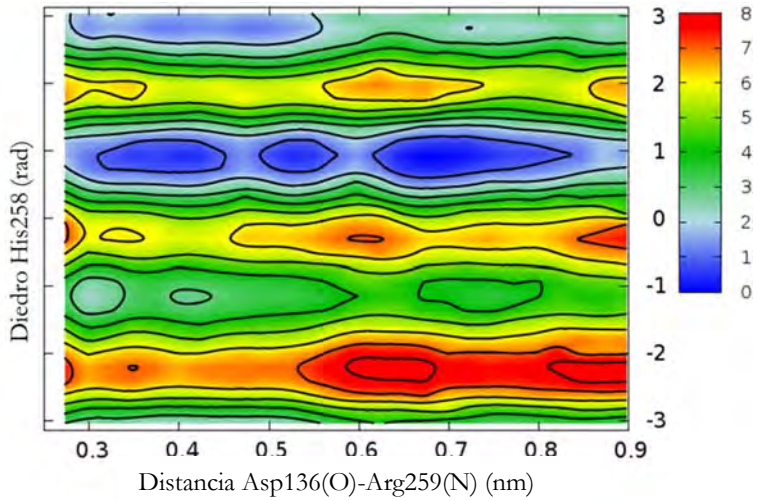
Distancia Asp136(O)-Arg259(N) (nm)

Proyección de la energía para el diedro de la Arg256 sobre la distancia. El metaestado más estable se encuentra en valor de diedro de 1, con la misma energía en cualquiera de las dos conformaciones, LA o LI.



Distancia Asp136(O)-Arg259(N) (nm)

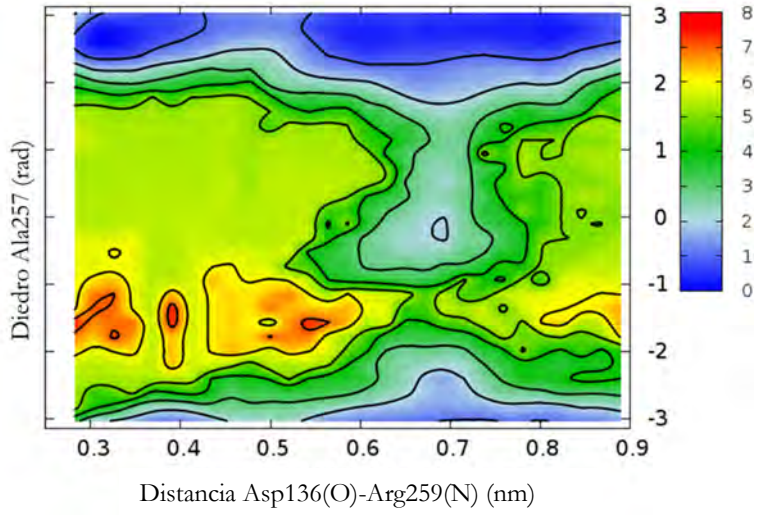
Proyección de la energía para el diedro de la Ala257 sobre la distancia. El metaestado más estable es la conformación LA, con un valor de diedro ψ para Ala257 entre 2 y 3 rad.



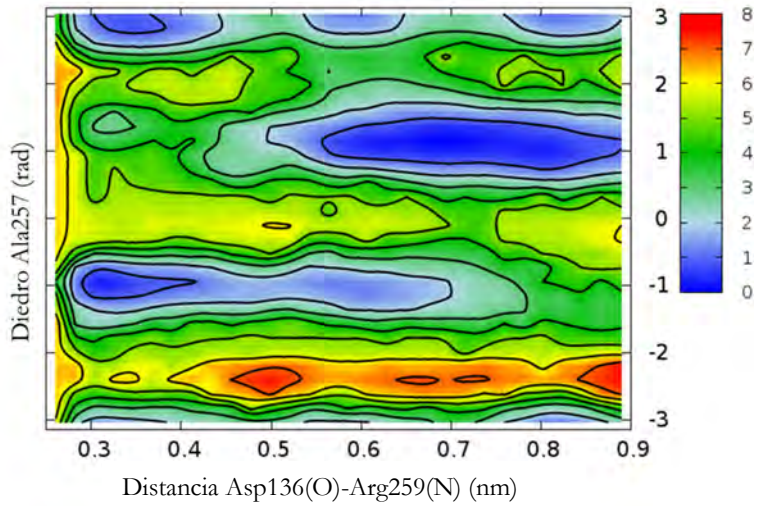
Proyección de la energía para el diedro de la His258 sobre la distancia. Extrañamente el metaestado más estable es para el valor 1 ya sea en conformación LA o LI. También aparece un mínimo en valor -1 y conformación LA, pero de energía mayor y menos estable que para el diedro en 1. No olvidemos que se trata de una proyección y los metaestados están influenciados por los valores de las otras CVs, aquí no representadas.

Anexo 16: Meta4Y6NUDPGlc5. Proyecciones de energía:

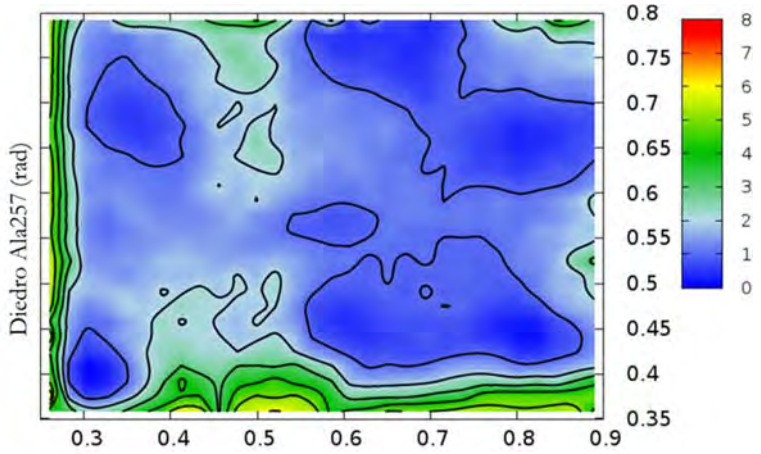
Isotermas: 1 kcal/mol ± 0,5 kcal/mol



Proyección de la energía para el diedro de la ala257 sobre la distancia.



Proyección de la energía para el diedro de la His258 sobre la distancia. Muy claramente son visibles los estados de diedro His258 -1 para la conformación LA y 1 para la conformación LI, siendo de igual energía ambos estados, cada uno en su correspondiente conformación de bucle.

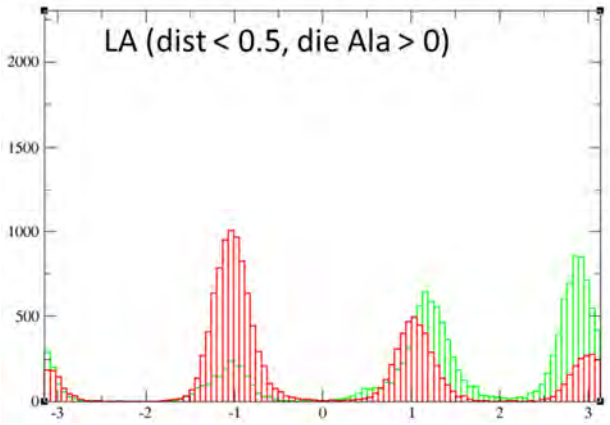


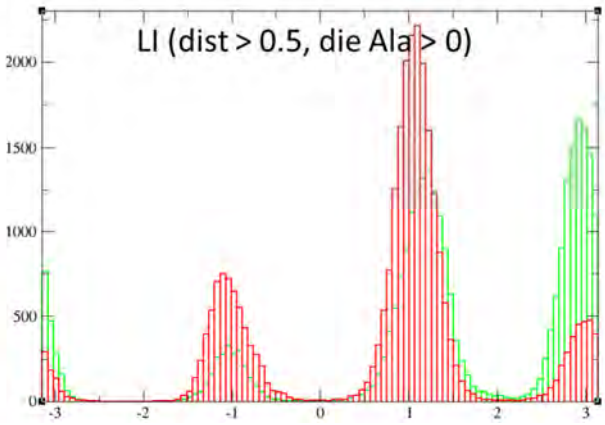
Proyección de la energía para la distancia metal-His258 sobre la distancia. Hasta 6 metaestados equiprobables. La distancia His258-metal de 0,4 nm, estable en la conformación LA según la MD del complejo ternario tiene la misma energía que distancias de 0,65 y 0,8 en conformación LI e incluso 0,7 en conformación LA.

Anexo 17. Poblaciones de diedros Arg256-His258 en las simulaciones metadinámicas.

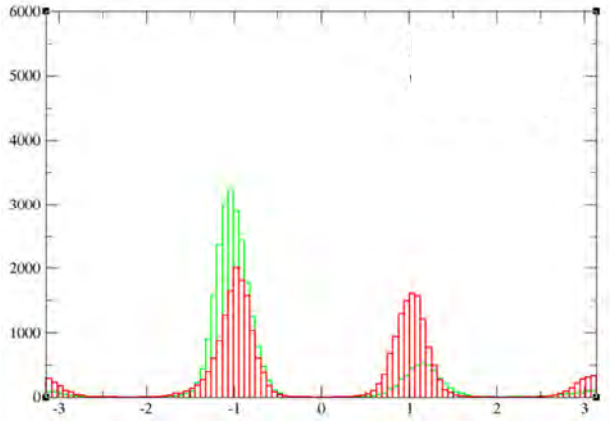
- **Izquierda**: Valores a una distancia menor de 0,5 (conformación LA).
- **Derecha**: Valores a una distancia mayor de 0,5 (conformación LI).
- **Verde:** Diedro Arg256. **Rojo:** Diedro His258.

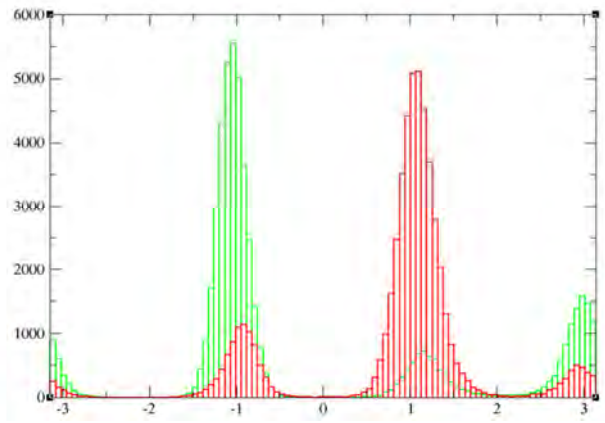
Simulación Meta4DDZ



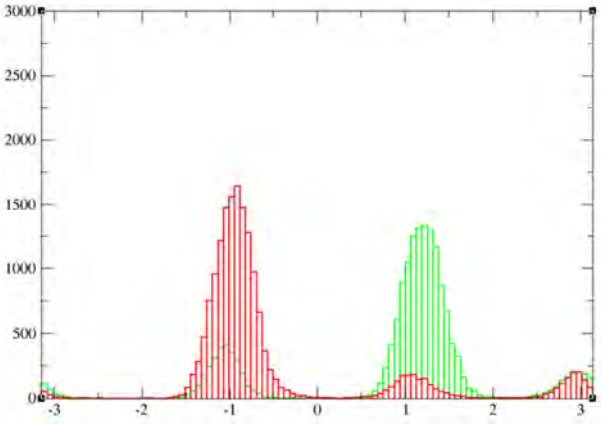


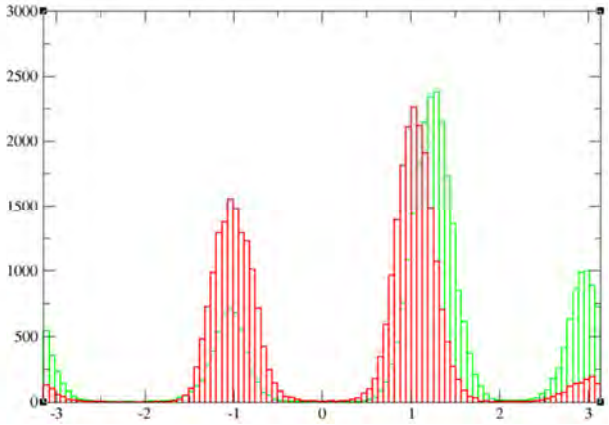
Simulación Meta4Y6NApo



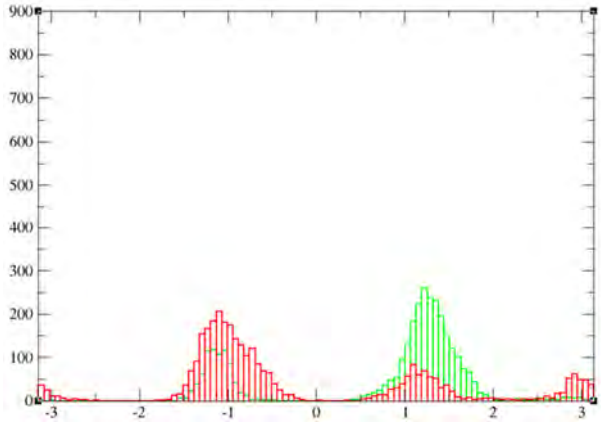


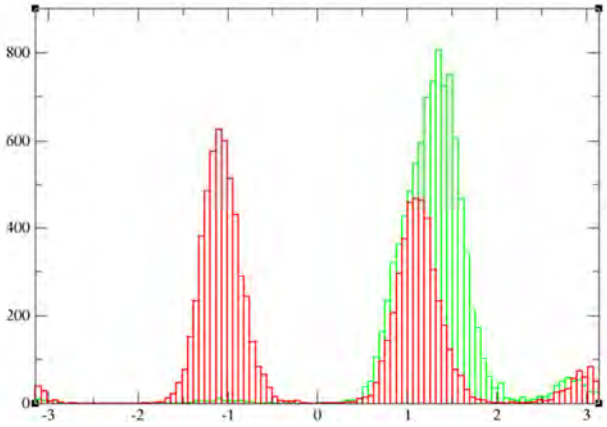
Simulación Meta4Y6N1 (Complejo ternario)



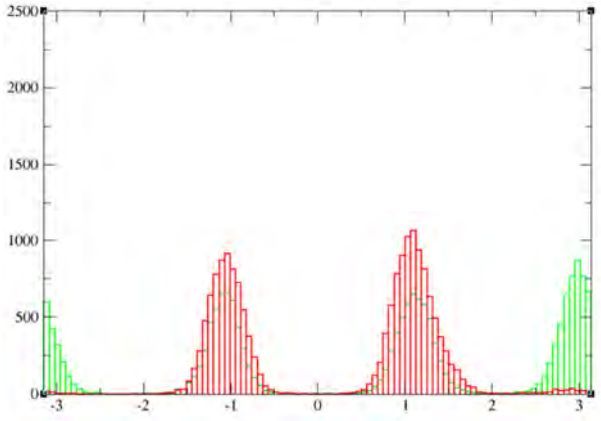


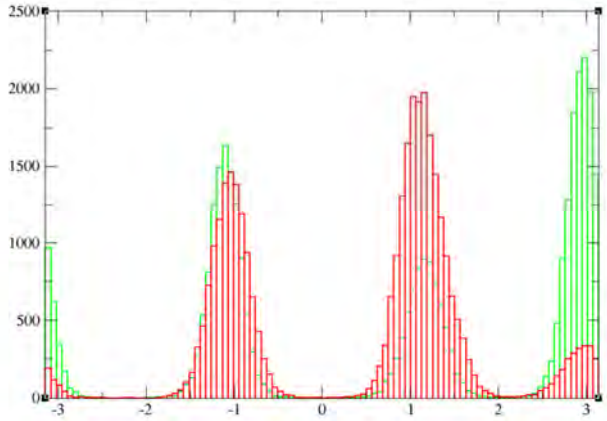
Simulación Meta4Y6N2 (diedro UDPGlc libre)



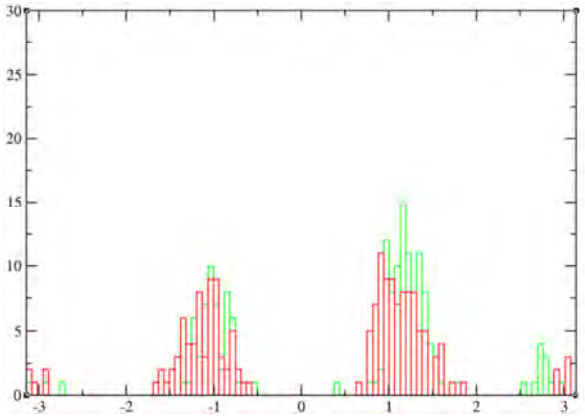


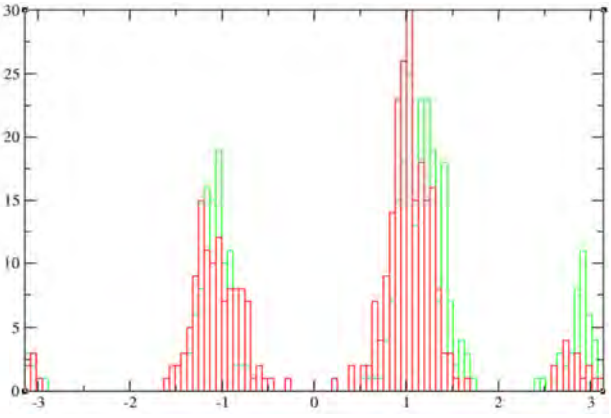
Simulación Meta4Y6N3 (diedro UDPGlc fijado)



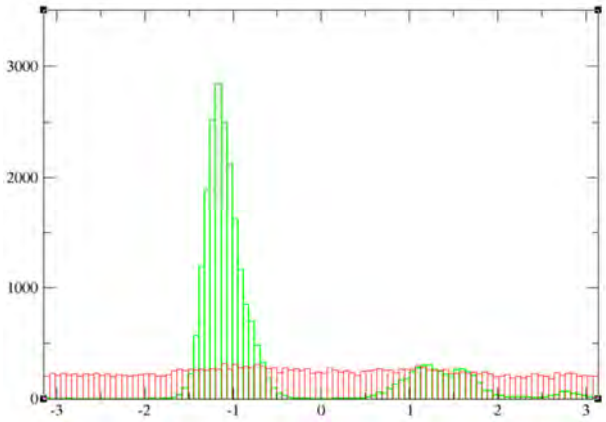


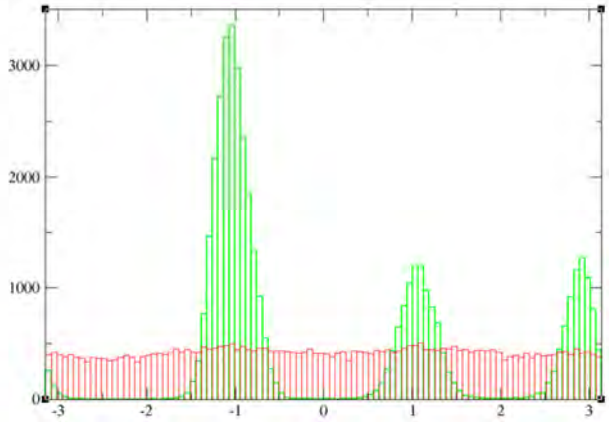
Simulación Meta4Y6N4 (PGA)



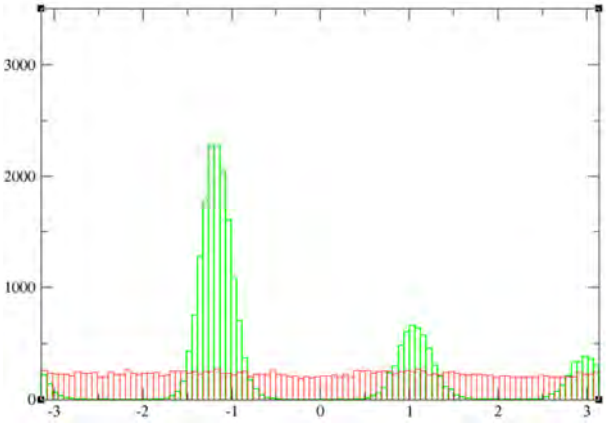


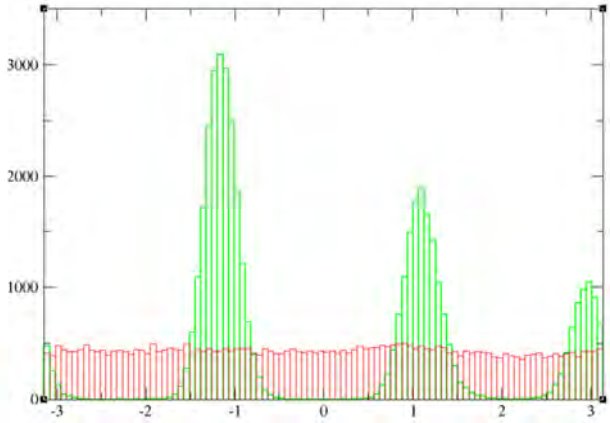
Simulación MetaD4Y6NApo4:



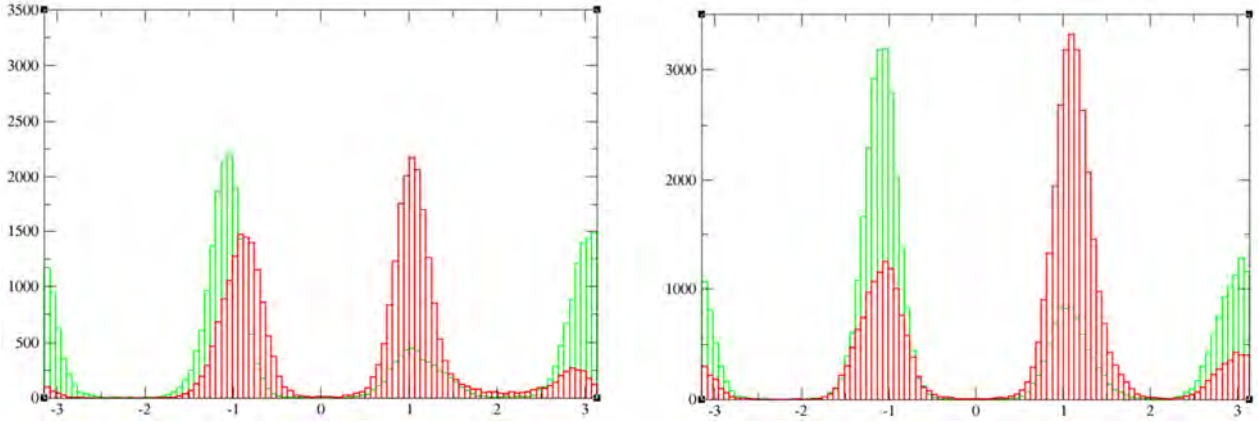


Simulación MetaD4Y6N5:

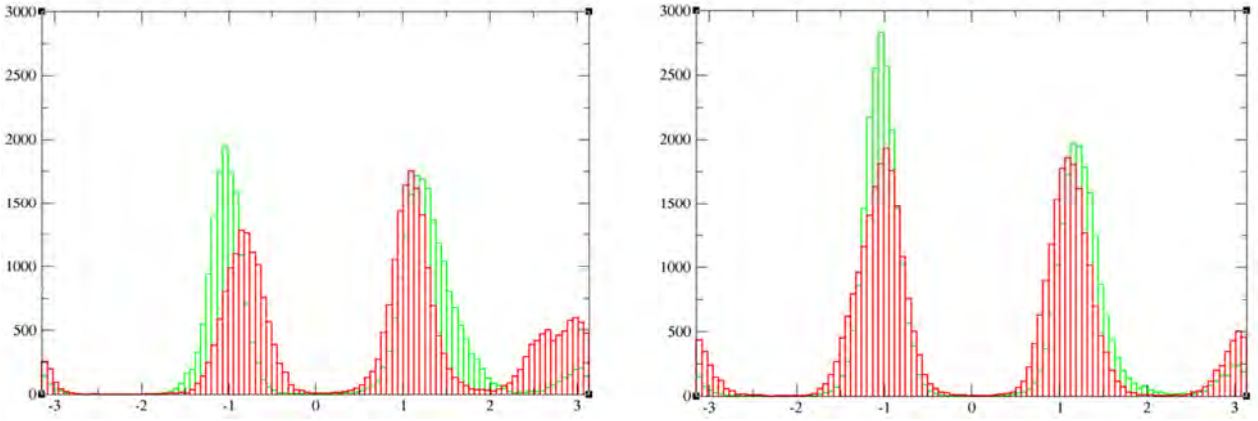




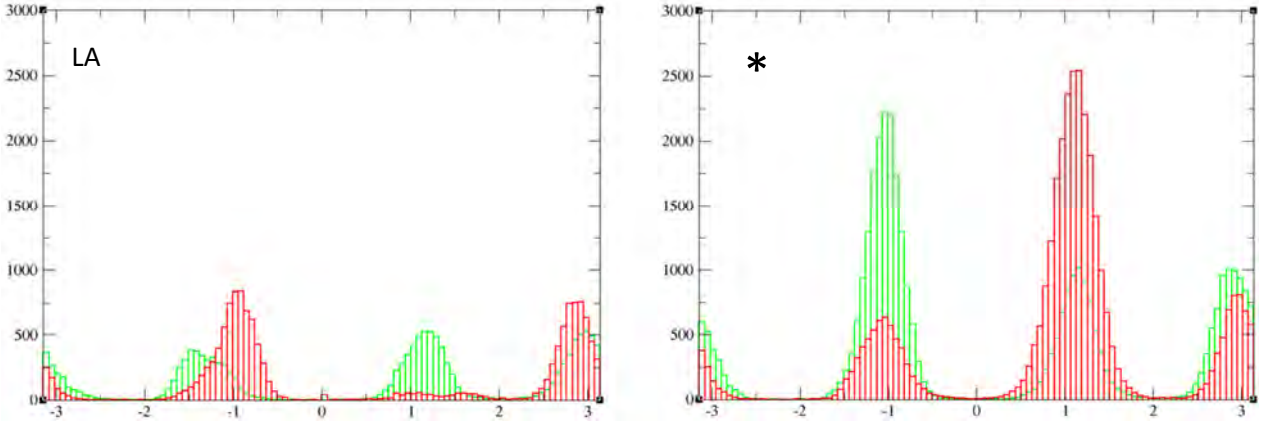
Simulación MetaD4Y6N6: Monómero A.



Simulación MetaD4Y6N6: Monómero B.



Simulación MetaD4Y6N7: CV Distancia Metal-His.



* Histograma para el metaestado LA (0,27 nm < distancia < 0,33 nm) y distancia metal-His258 0,35-0,45, considerado representativo de la forma LA en dinámica molecular clásica del complejo ternario (MD4Y6N).

Anexo 18. Estructuras GTA analizadas.

PDB ID	Descripción CAZY	Uniprot	Familia GT
1WEO	Celulosa sintasa A7 (Irx3;At5g17420)	Q9SWW6	GT2
4FIX	UDP-Galf; galactan b-(1,5)- / b-(1,6)-galactofuranosyltransferase (GlfT2;GlfT;Rv3808c)	O53585	GT2
4FIY	UDP-Galf; galactan b-(1,5)- / b-(1,6)-galactofuranosyltransferase (GlfT2;GlfT;Rv3808c)	O53585	GT2
2Z86	chondroitin polymerase (KfoC;K4CP)	H2KYQ5	GT2
2Z87	chondroitin polymerase (KfoC;K4CP)	H2KYQ5	GT2
3BCV	BF2801	Q5LBM4	GT2
4HG6	Celulosa sintasa subunit A (BcsA;RSP_0333)	Q3J125	GT2
1FG5	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1G8O	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1G93	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1GWV	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1GWW	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1GX0	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1GX4	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1K4V	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1O7O	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1O7Q	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1VZT	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1VZU	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1VZX	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2JCF	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2JCJ	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2JCK	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2JCL	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2JCO	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2VFZ	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2VS3	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2VS4	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2VS5	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2VXL	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2VXM	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
2WGZ	UDP-Gal: b-galactoside a-1,3-galactosyltransferase (GGTA1)	P14769	GT6
1LZ7	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1LZJ	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1R7U	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1R7X	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1R80	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1R82	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1ZIZ	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1ZJ0	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1ZJ1	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1ZJ2	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1ZJ3	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1ZJP	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2A8U	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2I7B	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2O1F	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2O1G	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RIT	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RIX	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RIY	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RIZ	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RJ0	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RJ1	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RJ4	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RJ5	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RJ6	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RJ7	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RJ8	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2RJ9	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
2Y7A	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0C	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0D	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0E	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0F	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0G	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0H	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0I	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0J	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0K	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3I0L	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3SXA	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6

3SXB	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3SXC	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3SXD	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
3U0X	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRA	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRB	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRD	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRE	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRH	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRL	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRM	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRO	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRP	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4FRQ	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4GBP	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4KC1	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4KC2	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4KC4	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
4KXO	B-specific a-1,3-galactosyltransferase (GTB)	Q9UQ63	GT6
1LZ0	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1LZI	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1R7T	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1R7V	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1R7Y	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1R81	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1WSZ	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1WT0	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1WT1	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1WT2	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1WT3	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1XZ6	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1ZHJ	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1ZI1	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1ZI3	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1ZI4	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1ZI5	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
1ZJO	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
2A8W	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3IOH	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3IOI	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3IOJ	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3SX3	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3SX5	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3SX7	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3SX8	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3SXE	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3SXG	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3U0Y	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3V0L	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3V0M	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3V0N	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3V0O	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3V0P	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3V0Q	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3ZGF	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
3ZGG	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
4C2S	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
4FQW	A-specific a-1,3-N-acetylgalactosaminyltransferase (GTA)	Q15951	GT6
4AYJ	UDP-GalNAc: 2'-fucosyl lactose a-N-acetylgalactosaminyltransferase (BoGT6a;BACOVA_01932)	A7LVT2	GT6
4AYL	UDP-GalNAc: 2'-fucosyl lactose a-N-acetylgalactosaminyltransferase (BoGT6a;BACOVA_01932)	A7LVT2	GT6
4CJ8	UDP-GalNAc: 2'-fucosyl lactose a-N-acetylgalactosaminyltransferase (BoGT6a;BACOVA_01932)	A7LVT2	GT6
4CJB	UDP-GalNAc: 2'-fucosyl lactose a-N-acetylgalactosaminyltransferase (BoGT6a;BACOVA_01932)	A7LVT2	GT6
4CJC	UDP-GalNAc: 2'-fucosyl lactose a-N-acetylgalactosaminyltransferase (BoGT6a;BACOVA_01932)	A7LVT2	GT6
1FGX	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1FR8	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1J8W	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1J8X	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1J92	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1J94	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1JN8	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1JNA	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1JNC	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1KYB	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1L7W	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1NF5	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1NHE	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1NKH	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7

1NMM	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1NQI	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1NT7	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1NWG	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1O0R	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1O23	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1OQM	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1TVY	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1TW1	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1TW5	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
1YRO	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
2FYC	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
2FYD	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
4KRV	b-1,4-galactosyltransferase T1 (B4Gal-T1;GgtB2;GalT)	P08037	GT7
2AE7	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
2AEC	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
2AES	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
2AGD	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
2AH9	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
2FY7	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
2FYA	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
3EE5	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4EE3	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4EE4	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4EE5	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4EEA	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4EEG	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4EEM	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4EEO	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4L41	UDP-Gal: b-GlcNAc b-1,4-galactosyltransferase T1 (b4GalT1;beta4GalT-I;GalT)	P15291	GT7
4IRP	UDP-Gal: xylosylprotein b-1,4-galactosyltransferase-I / VII / 7 (GalT-I;B4GALT7)	Q9UBV7	GT7
4IRQ	UDP-Gal: xylosylprotein b-1,4-galactosyltransferase-I / VII / 7 (GalT-I;B4GALT7)	Q9UBV7	GT7
3LW6	xylosylprotein b-4-galactosyltransferase I / 7 (bGalT7/CG11780)	Q9VBZ9	GT7
4LW3	xylosylprotein b-4-galactosyltransferase I / 7 (bGalT7/CG11780)	Q9VBZ9	GT7
4LW6	xylosylprotein b-4-galactosyltransferase I / 7 (bGalT7/CG11780)	Q9VBZ9	GT7
4M4K	xylosylprotein b-4-galactosyltransferase I / 7 (bGalT7/CG11780)	Q9VBZ9	GT7
3Q4S	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3QVB	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3RMV	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3RMW	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3T7M	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3T7N	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3T7O	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3U2T	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3U2U	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3U2V	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3U2W	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
3U2X	glycogenin (Gyg;Gyg1;GYG1)	P46976	GT8
4UEG	glycogenin 2 (Gyg2)	O15488	GT8
1G9R	UDP-Gal: a-1,4-galactosyltransferase (LgtC)	P96945	GT8
1GA8	UDP-Gal: a-1,4-galactosyltransferase (LgtC)	P96945	GT8
1SS9	UDP-Gal: a-1,4-galactosyltransferase (LgtC)	P96945	GT8
1LL0	glycogenin (Gyg)	P13280	GT8
1LL2	glycogenin (Gyg)	P13280	GT8
1LL3	glycogenin (Gyg)	P13280	GT8
3USQ	glycogenin (Gyg)	P13280	GT8
3USR	glycogenin (Gyg)	P13280	GT8
3V8Y	glycogenin (Gyg)	P13280	GT8
3V8Z	glycogenin (Gyg)	P13280	GT8
3V90	glycogenin (Gyg)	P13280	GT8
3V91	glycogenin (Gyg)	P13280	GT8
3TZT	Apre_0416	C7RG54	GT8
1FO8	UDP-GlcNAc: a-1,3-mannosyl-glycoprotein b-1,2-N-acetilglucosaminiltransferasa I (Mgat1;Gnt1)	P27115	GT13
1FO9	UDP-GlcNAc: a-1,3-mannosyl-glycoprotein b-1,2-N-acetilglucosaminiltransferasa I (Mgat1;Gnt1)	P27115	GT13
1FOA	UDP-GlcNAc: a-1,3-mannosyl-glycoprotein b-1,2-N-acetilglucosaminiltransferasa I (Mgat1;Gnt1)	P27115	GT13
2AM3	UDP-GlcNAc: a-1,3-mannosyl-glycoprotein b-1,2-N-acetilglucosaminiltransferasa I (Mgat1;Gnt1)	P27115	GT13
2AM4	UDP-GlcNAc: a-1,3-mannosyl-glycoprotein b-1,2-N-acetilglucosaminiltransferasa I (Mgat1;Gnt1)	P27115	GT13
2AM5	UDP-GlcNAc: a-1,3-mannosyl-glycoprotein b-1,2-N-acetilglucosaminiltransferasa I (Mgat1;Gnt1)	P27115	GT13
2APC	UDP-GlcNAc: a-1,3-mannosyl-glycoprotein b-1,2-N-acetilglucosaminiltransferasa I (Mgat1;Gnt1)	P27115	GT13
1S4N	GDP-Man: a-1,2-mannosyltransferase (Kre2;Mnt1;YDR483w;D8035.26)	P27809	GT15
1S4O	GDP-Man: a-1,2-mannosyltransferase (Kre2;Mnt1;YDR483w;D8035.26)	P27809	GT15
1S4P	GDP-Man: a-1,2-mannosyltransferase (Kre2;Mnt1;YDR483w;D8035.26)	P27809	GT15
1H7L	SpsA (BSU37910)	P39621	GT21
1H7Q	SpsA (BSU37910)	P39621	GT21
1QG8	SpsA (BSU37910)	P39621	GT21
1QGQ	SpsA (BSU37910)	P39621	GT21
1QGS	SpsA (BSU37910)	P39621	GT21

2FFU	UDP-GalNAc:polypeptide a-N-acetylgalactosaminyltransferase T2 (GalNAc-T2;ppGalNAcT2;GalNT2; pp-GalNAc-T2)	Q10471	GT27
2FFV	UDP-GalNAc:polypeptide a-N-acetylgalactosaminyltransferase T2 (GalNAc-T2;ppGalNAcT2;GalNT2; pp-GalNAc-T2)	Q10471	GT27
4D0T	UDP-GalNAc:polypeptide a-N-acetylgalactosaminyltransferase T2 (GalNAc-T2;ppGalNAcT2;GalNT2; pp-GalNAc-T2)	Q10471	GT27
4D0Z	UDP-GalNAc:polypeptide a-N-acetylgalactosaminyltransferase T2 (GalNAc-T2;ppGalNAcT2;GalNT2; pp-GalNAc-T2)	Q10471	GT27
4D11	UDP-GalNAc:polypeptide a-N-acetylgalactosaminyltransferase T2 (GalNAc-T2;ppGalNAcT2;GalNT2; pp-GalNAc-T2)	Q10471	GT27
5AJN	UDP-GalNAc:polypeptide a-N-acetylgalactosaminyltransferase T2 (GalNAc-T2;ppGalNAcT2;GalNT2; pp-GalNAc-T2)	Q10471	GT27
5AJO	UDP-GalNAc:polypeptide a-N-acetylgalactosaminyltransferase T2 (GalNAc-T2;ppGalNAcT2;GalNT2; pp-GalNAc-T2)	Q10471	GT27
5AJP	UDP-GalNAc:polypeptide a-N-acetylgalactosaminyltransferase T2 (GalNAc-T2;ppGalNAcT2;GalNT2; pp-GalNAc-T2)	Q10471	GT27
1XHB	UDP-GalNAc: polypeptide a-N-acetylgalactosaminyltransferase T1 (GalNAc-T1;ppGalNAcTase-1;Galnt1)	O08912	GT27
2D7I	UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase 10 (GalNT10;ppGalNAcT10;pp-GalNAc-T10)	Q86SR1	GT27
2D7R	UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase 10 (GalNT10;ppGalNAcT10;pp-GalNAc-T10)	Q86SR1	GT27
1FGG	UDP-GlcA: b-1,4-galactosyl-xylosylprotein b-1,3-glucuronosyltransferase 3 (GlcAT-I;B3GAT3)	O94766	GT43
1KWS	UDP-GlcA: b-1,4-galactosyl-xylosylprotein b-1,3-glucuronosyltransferase 3 (GlcAT-I;B3GAT3)	O94766	GT43
3CU0	UDP-GlcA: b-1,4-galactosyl-xylosylprotein b-1,3-glucuronosyltransferase 3 (GlcAT-I;B3GAT3)	O94766	GT43
1V82	UDP-GlcA: galactosylgalactosylxylosylprotein 3-b-glucuronyltransferase 1 (GlcAT-P;B3gat1)	Q9P2W7	GT43
1V83	UDP-GlcA: galactosylgalactosylxylosylprotein 3-b-glucuronyltransferase 1 (GlcAT-P;B3gat1)	Q9P2W7	GT43
1V84	UDP-GlcA: galactosylgalactosylxylosylprotein 3-b-glucuronyltransferase 1 (GlcAT-P;B3gat1)	Q9P2W7	GT43
2D0J	UDP-GlcA: galactosylgalactosylxylosylprotein 3-b-glucuronosyltransferase 2 (GlcAT-S;B3GAT2)	Q9NPZ5	GT43
2ZU7	a-mannosyltransferase (MPG synthase; PH0927)	O58689	GT55
2ZU8	a-mannosyltransferase (MPG synthase; PH0927)	O58689	GT55
2ZU9	a-mannosyltransferase (MPG synthase; PH0927)	O58689	GT55
2WVK	mannosyl-3-phosphoglycerate synthase (MpgS;TTC0588)	Q84B24	GT55
2WVL	mannosyl-3-phosphoglycerate synthase (MpgS;TTC0588)	Q84B24	GT55
2WVM	mannosyl-3-phosphoglycerate synthase (MpgS;TTC0588)	Q84B24	GT55
1OMX	a-N-acetylhexosaminyltransferase (Extl2)	Q8C089	GT64
1OMZ	a-N-acetylhexosaminyltransferase (Extl2)	Q8C089	GT64
1ON6	a-N-acetylhexosaminyltransferase (Extl2)	Q8C089	GT64
1ON8	a-N-acetylhexosaminyltransferase (Extl2)	Q8C089	GT64
2BO4	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2BO6	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2BO7	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2BO8	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2XW2	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2XW3	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2XW4	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2XW5	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2Y4J	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2Y4K	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2Y4L	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
2Y4M	GDP-Man:mannosylglycerate synthase (a-mannosyltransferase) (Mgs;Rmar_1220)	Q9RFR0	GT78
3E25	UDP-Glc: glucosyl-3-phosphoglycerate synthase (Rv1208;GpgS;MtGpgS;MTCI364.20)	P9WMW8	GT81
3E26	UDP-Glc: glucosyl-3-phosphoglycerate synthase (Rv1208;GpgS;MtGpgS;MTCI364.20)	P9WMW8	GT81
4DDZ	UDP-Glc: glucosyl-3-phosphoglycerate synthase (Rv1208;GpgS;MtGpgS;MTCI364.20)	P9WMW8	GT81
4DE7	UDP-Glc: glucosyl-3-phosphoglycerate synthase (Rv1208;GpgS;MtGpgS;MTCI364.20)	P9WMW8	GT81
4DEC	UDP-Glc: glucosyl-3-phosphoglycerate synthase (Rv1208;GpgS;MtGpgS;MTCI364.20)	P9WMW8	GT81
3CKJ	MAP2569c (GpgS;MaGpgS)	Q73WU1	GT81
3CKN	MAP2569c (GpgS;MaGpgS)	Q73WU1	GT81
3CKO	MAP2569c (GpgS;MaGpgS)	Q73WU1	GT81
3CKQ	MAP2569c (GpgS;MaGpgS)	Q73WU1	GT81
3F1Y	GDP-Man: mannosyl-3-phosphoglycerate synthase (MpgS)	B7SY86	GT81
3O3P	GDP-Man: mannosyl-3-phosphoglycerate synthase (MpgS)	B7SY86	GT81

Anexo 19. Metodología de identificación de las diferentes RV.

FAM	Nº EC	ACEPTOR	GRUPO	NOMBRE HMM	IDENTIFICACIÓN RV
GT2	2.4.1.-	?	Z	GT2_Z_A0Q5C5	A partir del alineamiento múltiple de la familia GT2 (con nº EC).
				GT2_Z_Q8KZ90	
GT2	2.4.1.-	?	Z	GT2_Z_UndP	
GT2	2.4.1.199	Man	Z	GT2_Z_Man	
GT2	2.4.2.53	UndP	Z	GT2_Z_UndP	
GT2	2.4.1.-	?	I	GT2_I	GTAHMM
GT2	2.4.1.-	?	A	GT2_A	GTAHMM
GT2	2.4.1.-	?	F	GT2_F_O07340	GTAHMM
				GT2_F_O87183	GTAHMM
GT2	2.4.1.305	GlcNAc	a	GT2_a_GlcNAc	GTAHMM
GT2	2.4.1.288	Gal	b	GT2_b_Gal	GTAHMM
GT2	2.4.1.-	?	G	GT2_G	GTAHMM sobre WP_ 012594069.1 y alineamiento con GT2_G
GT2	2.4.1.-	?	B	GT2_B	GTAHMM
GT2	2.4.1.-	?	S	GT2_S_P33697	A partir del alineamiento múltiple de la familia GT2 (con nº EC).
				GT2_S_P33700	
				GT2_S_Q9XBL5	
GT2	2.4.1.-	?	C	GT2_C	GTAHMM
GT2	2.4.1.-	?	L	GT2_L	GTAHMM sobre WP_039650701.1 y alineamiento con GT2_L
GT2	2.4.1.-	?	D	GT2_D	GTAHMM
GT2	2.4.1.-	?	V	GT2_V	GTAHMM sobre WP_039650701.1 y alineamiento con GT2_V
GT2	2.4.1.117	DolP	V	GT2_V_DolP	
GT2	2.4.1.83	DolP	V	GT2_V_DolP	
GT2	2.4.1.17	Ines	Y	GT2_Y_Ines	GTAHMM sobre con AEI35694.1 y alineamiento con GT2_Y
GT2	2.4.1.-	?	U	GT2_U	GTAHMM
GT2	2.4.1.-	?	E	GT2_E	**Sin RV**
GT2	2.4.1.157	DAG	X	GT2_X_DAG	GTAHMM
GT2	2.4.1.-	?	N	GT2_N	GTAHMM
GT2	2.4.1.-	?	P	GT2_P	GTAHMM (solo hasta la mitad de la RV, motivo FEYA, el resto manualmente)
GT2	2.4.1.12	Glc1	P	GT2_P_Glc1	GTAHMM
GT2	2.4.1.16	GlcNAc	P	GT2_P_GlcNAc	GTAHMM
GT2	2.4.1.212	GlcA	P	GT2_P_GlcA	GTAHMM
GT2	2.4.1.34	Glc2	P	GT2_P_Glc2	GTAHMM
GT2	2.4.1.-	?	F	GT2_F	GTAHMM
GT2	2.4.1.-	?	R	GT2_R	GTAHMM
GT2	2.4.1.289	GlcNAc	R	GT2_R_GlcNAc	GTAHMM
GT2	2.4.1.-	?	M	GT2_M_P37782	GTAHMM
				GT2_M_Q9I4K5	GTAHMM
GT2	2.4.1.-	?	O	GT2_O	GTAHMM sobre con WP_040286324.1 y alineamiento con GT2
GT2	2.4.1.-	?	Q	GT2_Q	GTAHMM
GT2	2.4.1.-	?	H	GT2_H	GTAHMM
GT2	2.4.1.-	?		GT2_A0PJ42	GTAHMM sobre con WP_038546081.1 y alineamiento con GT2 (con nº EC)
				GT2_A8E1V5	GTAHMM sobre con WP_006186106.1 y alineamiento con GT2 (con nº EC)
				GT2_B4ERA8	GTAHMM
				GT2_O01346	GTAHMM
				GT2_O07339	GTAHMM
				GT2_O31986	GTAHMM
				GT2_P33702	Alineamiento entre ellos e identificación manual
				GT2_P74820	
				GT2_Q0P9C6	GTAHMM

				GT2_Q1L2K4	GTAHMM
				GT2_Q52256	GTAHMM
				GT2_Q56914	Alineamiento entre ellos e identificación manual
				GT2_Q56915	
				GT2_Q5LFK5	GTAHMM
				GT2_Q8KB11	GTAHMM sobre con WP_039809247.1 y alineamiento con GT2_Q8KB11
				GT2_Q8R632	GTAHMM
				GT2_Q8Y949	GTAHMM sobre con ABS21693.1 y alineamiento con GT2_Q8Y949
				GT2_Q9A5M0	GTAHMM
				GT2_Q9AQI9	GTAHMM
				GT2_Q9S520	GTAHMM sobre con WP_017726152.1 y alineamiento con GT2_Q9S520
				GT2_Q9X9S1	GTAHMM sobre con EZP26224.1 y alineamiento con GT2_ Q9X9S1
				GT2_Q9XC63	GTAHMM sobre con WP_042093620.1y alineamiento con GT2_Q9XC63
GT2	2.4.1.157	DAG		GT2_DAG	GTAHMM
GT2	2.4.1.287	Rha		GT2_Rha	GTAHMM
GT2	2.4.1.303	GlcNAc		GT2_GlcNAc	GTAHMM
GT6	2.4.1.88	GalNAc	B	GT6_B_GalNAc	GTAHMM y alineamiento con GT6
GT6	2.4.1.88	GalNAc		GT6_GalNac	
GT6	2.4.1.309	Fuc	A	GT6_A_Fuc_Q5JBG6	
GT6	2.4.1.37	Fuc		GT6_Fuc	
GT6	2.4.1.40	Fuc		GT6_Fuc	
GT6	2.4.1.87	Gal		GT6_Gal	
GT7	2.4.1.174	GlcA	A	GT7_A_GlcA	GTAHMM y alineamiento con GT7
GT7	2.4.1.175	GlcA/GalNAc	A	GT7_A_GlcA/GalNAc	
GT7	2.4.1.244	GlcNAc	A	GT7_A_GlcNAc	
GT7	2.4.1.-	?		GT7_GlcNAc/Glc	
GT7	2.4.1.133	Xyl		GT7_Xyl	
GT7	2.4.1.22	Glc		GT7_GlcNAc/Glc	
GT7	2.4.1.274	Glc		GT7_GlcNAc/Glc	
GT7	2.4.1.275	GlcNAc		GT7_GlcNAc1	
GT7	2.4.1.38	GlcNAc		GT7_GlcNAc/Glc	
GT7	2.4.1.90	GlcNAc		GT7_GlcNAc/Glc	
GT8	2.4.1.-	?	H	GT8_H_Glc_lip	GTAHMM y alineamiento con GT8_H
GT8	2.4.1.-	?	H	GT8_H_Glc_lip	
GT8	2.4.1.44	Glc_lip	H	GT8_H_Glc_lip	
GT8	2.4.1.58	Sugar_lip	H	GT8_H_Glc_lip_P19817	
				GT8_H_Glc_lip_Q9ZIS5	
GT8	2.4.1.58	Sugar_lip	H	GT8_H_Glc_lip	
GT8	2.4.1.-	?	A	GT8_A	GTAHMM sobre con BAP71969.1 y alineamiento con GT8_A
GT8	2.4.1.-	?	G	GT8_G	GTAHMM
GT8	2.4.1.186	Gly	K	GT8_K	**Sin RV**
GT8	2.4.1.-	?	I	GT8_I	GTAHMM
GT8	2.4.1.123	MioIno	I	GT8_I_MioIno	GTAHMM
GT8	2.4.1.43	GalA	L	GT8_L_GalA	GTAHMM sobre con NP_189150.1 y alineamiento con GT8_L
GT8	2.4.1.-	?	D	GT8_D	GTAHMM y alineamiento con GT8 (con nº EC)
GT8	2.4.1.-	?	B	GT8_B	
GT8	2.4.1.17	Ines	J	GT8_J_Ines	
GT8	2.4.2.-	Glc-EGFlike	M	GT8_M_Glc-EFGlike	
GT12	2.4.1.92	NeuNAc		GT12_NeuNAc	GTAHMM (solo hasta la mitad de la RV, el resto manualmente)
GT13	2.4.1.-	?	A	GT13_Man	GTAHMM
GT13	2.4.1.101	Man	B	GT13_Man	
GT13	2.4.1.101	Man	C	GT13_Man	
GT13	2.4.1.101	Man		GT13_Man	
GT15	2.4.1.131	Man		GT15_Man	GTAHMM

GT21	2.4.1.-	?		GT21_SphngNAc	GTAHMM
GT21	2.4.1.80	SphngNAc		GT21_SphngNAc	GTAHMM
GT24	2.4.1.-	?		GT24	GTAHMM
GT27	2.4.1.41	Peptide		GT27_Pep	GTAHMM
GT27	2.4.1.-	GalNAc		GT27_GalNAc	GTAHMM
GT43	2.4.-.-	?		GT43_Gal	GTAHMM y alineamiento con GT43 (con nº EC)
GT43	2.4.1.135	Gal		GT43_Gal	
GT43	2.4.2.-	?		GT43_Gal	
GT55	2.4.1.217	PAG		GT55_PAG	GTAHMM
GT64	2.4.1.223	GlcA		GT64_GlcA	GTAHMM y alineamiento con GT64 (con nº EC)
GT64	2.4.1.224	GlcA		GT64_GlcA	
GT78	2.4.1.269	GA		GT78_GA	GTAHMM
GT81	2.4.1.-	?	A	GT81_PGA	GTAHMM y alineamiento con GT81 (con nº EC)
GT81	2.4.1.266	PGA	A	GT81_PGA	
GT81	2.4.1.-	?	B	GT81_PGA	
GT81	2.4.1.217	PGA	B	GT81_PGA	
GT81	2.4.1.266	PGA	B	GT81_PGA	
GT81	2.4.1.266	PGA	B	GT81_PGA	
GT81	2.4.1.268	GA	C	GT81_GA	
GT82	2.4.1.-	?		GT82	GTAHMM (solo hasta la mitad de la RV, el resto manualmente)

Anexo 20. Longitud de cada perfil, número de secuencias iniciales y tras la búsqueda contra la base de datos CAZy.

HMM	LONGITUD HMM	POBLACIÓN INICIAL	POBLACIÓN FINAL
GT2_A	36	1	117
GT2_A_GLCNAC	23	1	1
GT2_A0PJ42	45	1	8
GT2_A8E1V5	22	1	4
GT2_B	56	1	23
GT2_B_GAL	37	1	226
GT2_B4ERA8	29	1	11
GT2_C	27	1	40
GT2_C	28	1	122
GT2_D	40	1	4
GT2_DAG	34	2	9
GT2_E	22	1	956
GT2_F_O07340	32	1	10
GT2_F_O87183	35	1	8
GT2_G	28	2	97
GT2_GLCNAC	25	1	8
GT2_H_B1B4J9	26	1	6
GT2_H_Q5J7C7	26	1	6
GT2_I	33	2	20
GT2_L	33	2	70
GT2_L_Q8KWR0	36	1	14
GT2_M_P37782	29	1	13
GT2_M_Q9I4K5	27	1	85
GT2_N	32	2	108
GT2_O	31	1	83
GT2_O01346	35	1	15
GT2_O07339	27	1	10
GT2_O31986	25	1	15
GT2_P	31	38	2856
GT2_P_C3U576	29	1	17
GT2_P_GLC1	38	23	1555
GT2_P_GLC1_Q6RCS2	32	1	8
GT2_P_GLC2	30	1	153
GT2_P_GLCA	25	14	124
GT2_P_GLCNAC	29	56	1024
GT2_P_GLCNAC_P78746	25	1	21
GT2_P_GLCNAC_Q8V735	30	1	25
GT2_P33702	30	1	22
GT2_P74820	24	1	4
GT2_Q	28	2	508
GT2_Q0P9C6	28	1	41
GT2_Q1L2K4	44	1	3
GT2_Q52256	27	1	1
GT2_Q56914	27	1	12
GT2_Q56915	25	1	14
GT2_Q5LFK5	18	1	6
GT2_Q8KB11	26	1	19
GT2_Q8R632	55	1	5
GT2_Q8Y949	23	1	126
GT2_Q9A5M0	26	1	5
GT2_Q9AQI9	28	1	7
GT2_Q9S520	29	1	51
GT2_Q9X9S1	28	1	1
GT2_Q9XC63	23	1	5
GT2_R_GLCNAC	52	1	263
GT2_R_P74817	40	1	3
GT2_RHA	26	1	194
GT2_S_P33697	31	1	29
GT2_S_P33700	45	1	26
GT2_S_Q9XBL5	43	1	5

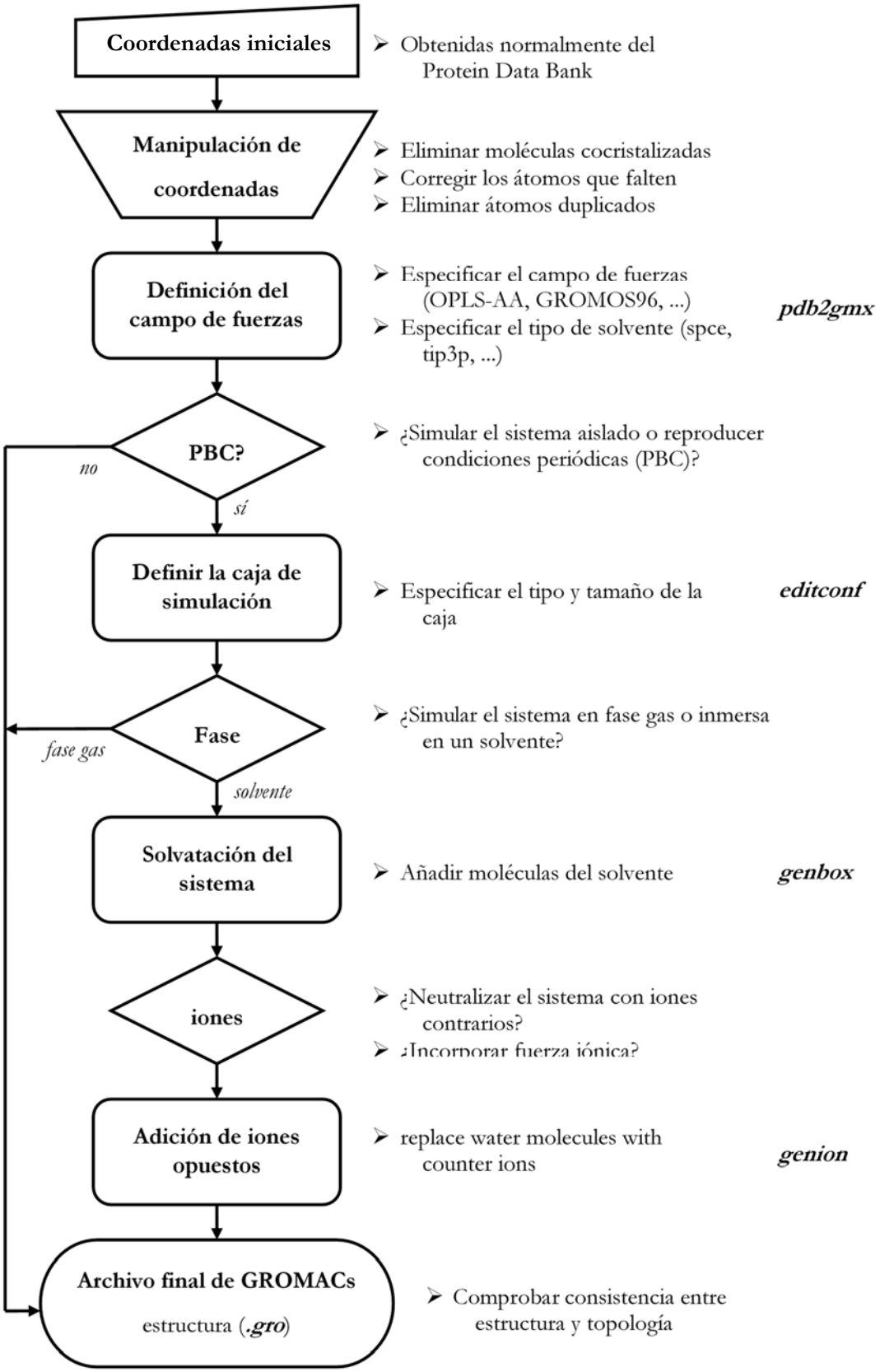
GT2_U	34	5	97
GT2_V_B5U882	32	1	64
GT2_V_DOLP	32	17	1086
GT2_X_DAG	28	2	89
GT2_Y_INES	32	1	9
GT2_Z_A0Q5C5	12	1	49
GT2_Z_MAN	14	1	73
GT2_Z_Q8KZ90	12	1	74
GT2_Z_UNDP	12	5	426
GT6_A_FUC	27	1	4
GT6_B_GALNAC	27	3	46
GT6_FUC	27	2	125
GT6_GAL	27	6	27
GT7_A_GLCA	12	2	17
GT7_A_GLCA/GALNAC	12	1	21
GT7_A_GLCNAC	9	2	8
GT7_GLCNAC/GLC	11	19	85
GT7_GLCNAC1	11	2	13
GT7_XYL	10	2	32
GT8_A	82	1	107
GT8_C	20	2	54
GT8_D	8	1	3
GT8_G	44	3	3
GT8_H_GLC_LIP	15	8	393
GT8_H_GLC_LIP_P19817	14	1	268
GT8_H_GLC_LIP_Q9ZIS5	15	1	57
GT8_H_O24967	30	1	70
GT8_H_O25962	45	1	68
GT8_H_Q48484	26	1	57
GT8_I_MIOINO	20	9	133
GT8_J_INES	8	1	2
GT8_L_GALA	20	1	106
GT8_M_GLC-EFGLIKE	20	1	23
GT12_NEUNAC	38	4	25
GT13_MAN	15	18	59
GT15_MAN	34	6	878
GT21_SPHNGNAC	28	12	103
GT24	23	6	83
GT27_PEP	49	52	216
GT27_PEP_Q9R0C5	45	1	13
GT43_GAL	31	12	69
GT55_PGA	30	5	57
GT64_GLCA	22	10	67
GT78_GA	26	1	4
GT81_GA	35	1	299
GT81_PGA	22	7	46
GT82	19	1	16

Anexo 21. Optimización de resultados tras ser sometidos al *pipeline* de CAZy,

NOMBRE HMM	EVALUE CUTOFF	RECUBRIMIENTO	RESIDUOS CUBIERTOS	HITS	FALSOS POSITIVOS	PRECISIÓN (%)
GT2_A	1,00E-16	0,6	22	244	57	77
GT2_A_GLCNAC	1,00E-07	0,6	14	1	0	100
GT2_A0PJ42	1,00E-07	0,8	36	8	0	100
GT2_A8E1V5	1,00E-16	0,6	13	3	1	67
GT2_B	1,00E-07	0,6	34	38	4	89
GT2_B_GAL	1,00E-14	0,6	22	386	66	83
GT2_B4ERA8	1,00E-08	0,6	17	9	3	67
GT2_C	1,00E-13	0,6	16	354	43	88
GT2_C	1,00E-08	0,75	21	191	12	94
GT2_D	1,00E-08	0,6	24	11	6	45
GT2_DAG	1,00E-07	0,65	22	4	0	100
GT2_E	1,00E-15	0,6	13	1293	152	88
GT2_F_O07340	1,00E-07	0,65	21	11	4	64
GT2_F_O87183	1,00E-07	0,6	21	5	0	100
GT2_G	1,00E-07	0,6	17	53	27	49
GT2_GLCNAC	1,00E-18	0,6	15	2	1	50
GT2_H_B1B4J9	1,00E-07	0,6	16	10	2	80
GT2_H_Q5J7C7	1,00E-06	0,7	18	7	3	57
GT2_I	1,00E-07	0,8	26	30	1	97
GT2_L	1,00E-07	0,6	20	35	1	97
GT2_L_Q8KWR0	1,00E-08	0,6	22	9	7	22
GT2_M_P37782	1,00E-07	0,6	17	14	3	79
GT2_M_Q9I4K5	1,00E-08	0,7	19	174	17	90
GT2_N	1,00E-06	0,8	26	211	6	97
GT2_O	1,00E-06	0,75	23	157	17	89
GT2_O01346	1,00E-09	0,7	25	119	4	97
GT2_O07339	1,00E-11	0,6	16	6	1	83
GT2_O31986	1,00E-15	0,6	15	11	0	100
GT2_P	1,00E-05	0,75	23	3981	268	93
GT2_P_C3U576	1,00E-07	0,7	20	163	16	90
GT2_P_GLC1	1,00E-07	0,75	29	3838	892	77
GT2_P_GLC1_Q6RCS2	1,00E-06	0,8	26	14	0	100
GT2_P_GLC2	1,00E-07	0,7	21	407	78	81
GT2_P_GLCA	1,00E-07	0,7	18	667	64	90
GT2_P_GLCNAC	1,00E-06	0,7	20	2834	124	96
GT2_P_GLCNAC_P78746	1,00E-09	0,6	15	190	9	95
GT2_P_GLCNAC_Q8V735	1,00E-06	0,7	21	19	0	100
GT2_P33702	1,00E-06	0,7	21	80	6	93
GT2_P74820	1,00E-15	0,6	14	3	0	100
GT2_Q	1,00E-06	0,75	21	626	52	92
GT2_Q0P9C6	1,00E-06	0,65	18	63	2	97
GT2_Q1L2K4	1,00E-16	0,6	26	3	0	100
GT2_Q52256	1,00E-05	0,75	20	1	7	0
GT2_Q56914	1,00E-07	0,6	16	13	0	100
GT2_Q56915	1,00E-07	0,8	20	15	4	73
GT2_Q5LFK5	1,00E-10	0,6	11	21	3	86
GT2_Q8KB11	1,00E-07	0,6	16	168	35	79
GT2_Q8R632	1,00E-07	0,6	33	16	1	94
GT2_Q8Y949	1,00E-12	0,6	14	49	2	96
GT2_Q9A5M0	1,00E-09	0,7	18	22	1	95
GT2_Q9AQI9	1,00E-07	0,8	22	5	1	80
GT2_Q9S520	1,00E-07	0,6	17	5	1	80
GT2_Q9X9S1	1,00E-14	0,6	17	9	2	78
GT2_Q9XC63	1,00E-10	0,6	14	6	0	100
GT2_R_GLCNAC	1,00E-06	0,8	42	543	155	71
GT2_R_P74817	1,00E-07	0,6	24	4	0	100
GT2_RHA	1,00E-07	0,6	16	431	60	86
GT2_S_P33697	1,00E-16	0,6	19	90	15	83
GT2_S_P33700	1,00E-17	0,6	27	100	3	97
GT2_S_Q9XBL5	1,00E-10	0,6	26	13	2	85
GT2_U	1,00E-07	0,6	20	227	20	91
GT2_V_B5U882	1,00E-05	0,75	24	181	68	62
GT2_V_DOLP	1,00E-05	0,75	24	4002	594	85
GT2_X_DAG	1,00E-14	0,6	17	235	32	86
GT2_Y_INES	1,00E-07	0,6	19	15	5	67
GT2_Z_A0Q5C5	1,00E-01	0,75	9	32	0	100
GT2_Z_MAN	1,00E-07	0,6	8	47	4	91

GT2_Z_Q8KZ90	1,00E-05	0,75	9	649	72	89
GT2_Z_UNDP	1,00E-05	0,75	9	748	179	76
GT6_A_FUC	1,00E-11	0,6	16	4	0	100
GT6_B_GALNAC	1,00E-08	0,6	16	372	18	95
GT6_FUC	1,00E-12	0,7	19	674	36	95
GT6_GAL	1,00E-12	0,6	16	347	51	85
GT7_A_GLCA	1,00E-04	0,75	9	499	30	94
GT7_A_GLCA/GALNAC	1,00E-05	0,7	8	450	21	95
GT7_A_GLCNAC	1,00E-04	0,8	7	413	56	86
GT7_GLCNAC/GLC	1,00E-07	0,6	7	780	56	93
GT7_GLCNAC1	1,00E-06	0,6	7	162	9	94
GT7_XYL	1,00E-01	0,7	7	300	10	97
GT8_A	1,00E-07	0,7	57	40	0	100
GT8_C	1,00E-07	0,6	12	554	33	94
GT8_D	1,00E-05	0,6	5	32	9	72
GT8_G	1,00E-20	0,8	35	4	4	0
GT8_H_GLC_LIP	1,00E-09	0,6	9	108	13	88
GT8_H_GLC_LIP_P19817	1,00E-05	0,8	11	104	14	87
GT8_H_GLC_LIP_Q9ZIS5	1,00E-06	0,75	11	119	11	91
GT8_H_O24967	1,00E-07	0,6	18	297	7	98
GT8_H_O25962	1,00E-07	0,6	27	290	5	98
GT8_H_Q48484	1,00E-07	0,6	16	140	11	92
GT8_I_MIOINO	1,00E-05	0,75	15	445	22	95
GT8_J_INES	1,00E-05	0,8	6	110	16	85
GT8_L_GALA	1,00E-01	0,8	16	1206	68	94
GT8_M_GLC-EFGLIKE	1,00E-06	0,7	14	563	28	95
GT12_NEUNAC	1,00E-07	0,65	25	366	31	92
GT13_MAN	1,00E-06	0,8	12	629	29	95
GT15_MAN	1,00E-07	0,6	20	1163	53	95
GT21_SPHNGNAC	1,00E-08	0,75	21	773	52	93
GT24	1,00E-09	0,6	14	1230	68	94
GT27_GALNAC	1,00E-09	0,65	32	340	21	94
GT27_PEP	1,00E-06	0,65	29	4801	278	94
GT43_GAL	1,00E-06	0,7	22	1163	52	96
GT55_PAG	1,00E-08	0,6	18	63	14	78
GT64_GLCA	1,00E-05	0,75	17	1282	51	96
GT78_GA	1,00E-06	0,75	20	8	0	100
GT81_GA	1,00E-06	0,8	28	53	32	40
GT81_PGA	1,00E-08	0,7	15	1013	159	84
GT82	1,00E-12	0,6	11	11	1	91

Anexo 22. Generación de archivos GROMACs desde PDB



Scripts

Script Phylip.scr

Programa que genera 4 carpetas y divide el conjunto de datos que utilizará Phylip en 4, para aprovechar los 4 CPUs del servidor de trabajo, al finalizar llama al programa Phylip1(2, 3 y 4).scr. Al llamar al *script* hay que pasarle el número del *dataset*.

```
#!/bin/bash

D=$1
DataSet=$(( D/4 ))
Seed1=3031
Seed2=4011
Seed3=173
Seed4=1943
cd Phylip1
./Phylip1.scr $DataSet $Seed1 &
cd ..
cd Phylip2
./Phylip2.scr $DataSet $Seed2 &
cd ..
cd Phylip3
./Phylip3.scr $DataSet $Seed3 &
cd ..
cd Phylip4
./Phylip4.scr $DataSet $Seed4 &
cd ..
```

Script Phylip1.scr

(Igual para Phylip2.scr, Phylip3.scr y Phylip4.scr). Programa que lanza el paquete Phylip propiamente dicho. Toma los parámetros de los archivos seqboot.comands, protdist.comands y neighbor.comands. Agrupa todos los datos y después lanza el programa Arbol.scr.

```
#!/bin/bash

source /etc/profile.modules
module load PHYLIP/3.69

DataSet=$1
Seed=$2
#grep "|" ../infile | awk '{tmp=$1; print tmp}' > codigos
#sort -d codigos | uniq -d | sed "s/|/\\\|/" > Cod_repetidos
#rm codigos
cp ../infile .
echo "r
$DataSet
2
y
$Seed" > seqboot.comands
seqboot < seqboot.comands
cp outfile outfile.seqboot
mv outfile infile
echo "m
d
$DataSet
2
y" > protdist.comands
protdist < protdist.comands
cp outfile outfile.protdist
mv outfile infile
echo "m
$DataSet
$Seed
y" > neighbor.comands
neighbor < neighbor.comands
if ! [ -e ../intree ]
then
  touch ../intree
fi
cat outtree >> ../intree
touch FIN
cd ..
./Arbol.scr $DataSet
```

Script Arbol.scr

Genera el árbol consenso.

```
#!/bin/bash

source /etc/profile.modules

module load PHYLIP/3.69

DataSet=$1

if [ -e ./Phylip1/FIN ] && [ -e ./Phylip2/FIN ] && [ -e ./Phylip3/FIN ] && [ -e ./Phylip4/FIN
]
then
  echo "2
y" > consense.comands
  consense < consense.comands
  cp intree intree.neighbor
  cp outtree outtree.neighbor
  mv outtree intree
  echo "n
n
y
w
u
q" > retree.comands
  retree < retree.comands
  mv outtree outtree.xml
fi
```

Parámetros PHYLIP

```
seqboot.comands
r
2500 #Conjunto de datos dividido en 4 (original pasado a Phylip.scr: 10000)
2
y
3031 #Semilla, proviene de Phylip1(2, 3 o 4).scr.

consense.comands
2
y

protdist.comands
m
d
2500 #Conjunto de datos dividido en 4 (original pasado a Phylip.scr: 10000)
2
y

neighbor.comands
m
2500 #Conjunto de datos dividido en 4 (original pasado a Phylip.scr: 10000)
3031 #Semilla, proviene de Phylip1(2, 3 o 4).scr.
y

retree.comands
n
n
y
w
u
q
```

Script UPDATE_SS.SH

Puesto que el software para visualización de alineamientos JalView no dispone de una herramienta para actualizar anotaciones de estructura secundaria cuando se modifica un alineamiento, diseñamos este programa para poder hacerlo. El archivo Promals.aln.fasta contiene el alineamiento, cada secuencia tiene un archivo con su estructura secundaria *.pdb.dssp.clean (H: hélice, E: hoja beta, -: sin estructura)

```
#!/bin/bash

echo "JALVIEW_ANNOTATION"
echo "# Created:" `date`
echo ""

for code in `cat codigos`
do

# busca la secuencia correspondiente al código ($code) y la pone en una sola línea

        sequence=`grep -A 200 $code Promals.aln.fasta | awk
        'NR>1{if(substr($0,1,1)==">"){exit}else{print $0}}' | tr -d "\r\n"`

# pone la secuencia ($sequence) en un array. el elemento 0 del array será la longitud de la
secuencia

        sequence_array=(`echo $sequence | awk '{long=split($0,cc,"");print
        long;for(i=1;i<=long;i++)print cc[i]}'`)
        sequence_length=${sequence_array[0]}

# busca la estructura secundaria correspondiente al código ($code) y la pone en una sola línea

        secondary=`cat $code.pdb.dssp.clean | tr -d "\r\n"`

# pone la estructura secundaria ($secondary) en un array. el elemento 0 del array será la
longitud de la estructura secundaria

        secondary_array=(`echo $secondary | awk '{long=split($0,cc,"");print
        long;for(i=1;i<=long;i++)print cc[i]}'`)
        secondary_length=${secondary_array[0]}
        echo "NO_GRAPH" $code | awk '{printf("%s\t%s\t",$1,$2)}'
        contador=0

        for i in `seq 1 $sequence_length`
        do
                if [ "${sequence_array[$i]}" == "-" ]
                then
                        echo -n "|"
                else
                        contador=$(( contador + 1 ))
                        if [ "${secondary_array[$contador]}" == "-" ]
                        then
                                        echo -n "|"
                        else
                                        echo -n "${secondary_array[$contador]}|"
                        fi
                fi
        done
        echo ""
        if [ $contador != $secondary_length ]
        then
                echo "ERROR! - code: $code - Sequence length and Secondary Structure length
                do not match!" >&2
        echo "                       The Secondary Structure adjustment may be incorrect for this
code!" >&2
        fi
done
```

Script MontajeNter.scr

Automatización del montaje del sistema y equilibrado para las simulaciones de MD de los modelos de MG517 para la región Nter.

```
# Se indican las directivas de trabajo para el ordenador Picasso:

#!/bin/bash
# @ job_name         = Modelo1MD
# @ wall_clock_limit = 86400
# @ total_tasks      = 16
# @ initialdir       = .
# @ error            = errors_%j
# @ output           = outputs_%j
# @ tasks_per_node   = 2

JOB_ID=$SLURM_JOBID
NAME=Mdelo1MD
CLASS=default
CPU_LIMIT=18000
INIT_DIR=`pwd`
JOB_TIME=`date +%Y%m%d_%H%M%S`
NPROCS=$SLURM_NPROCS

# Se carga el paquete de programas GROMACS

source /gpfs/apps/GROMACS/4.5.3/64/bin/GMXRC

# Se inicia la preparación de archivos y el montaje del sistema. El script debe estar situado
en una carpeta que contenga SOLAMENTE el archivo pdb inicial, el resto de directorios se
generarán dentro de esta carpeta.

# Guarda el nombre del fichero que se usará para el resto de procesos

name=`ls -a | awk '{tmp=$0; if (match(tmp,".pdb")>0) {split(tmp,a,".pdb"); print a[1]}}'`

# Creamos los directorios de trabajo

mkdir setup dry wet ionwet ionwet/em1 ionwet/posres ionwet/press ionwet/heat ionwet/npt
cp $name.pdb setup/$name.pdb
cd setup

# Generación del "Force Field": Convierte el PDB a un archivo GROMACS. Este paso no es necesario
en realidad, ya que tanto los .gro como los .itp y .top han sido creados previamente y modificados
incluyendo por superposición el UDPGlc en el .gro y el resto de archivos modificados a mano para
incluir el UDPGlc y depositados en la carpeta GRO_TOP_ITP que sustituirá a los generados por
pdb2gmx.

pdb2gmx -ignh -ff amber03 -f $name.pdb -o $name.gro -p $name.top -water tip3p

# Se mdifican los archivos generados para que incluyan el UDP_Glucosa
cp -f ../GRO_TOP_ITP/$name.gro $name.gro
cp -f ../GRO_TOP_ITP/$name.top $name.top
cp -f ../GRO_TOP_ITP/posre.itp posre.itp
cp -f ../GRO_TOP_ITP/posre_UDPGLC.itp posre_UDPGLC.itp
cp -f ../GRO_TOP_ITP/UDP_GLC.itp UDP_GLC.itp
cp -f ../GRO_TOP_ITP/UDP_GLC.itp ../ionwet/UDP_GLC.itp

# Define la "Caja de simulación". Definimos las "Periodic boundary conditions (PBC)"

editconf -bt cubic -f $name.gro -o $name\_dry.gro -d 0.9

# Se hace una copia de seguridad del archivo de topologías y se copian los archivos necesarios
para la solvatación

cp $name.top ../dry/$name\_dry.top
cp $name\_dry.gro ../dry/$name\_dry.gro

# Solvatamos el sistema

genbox -cp $name\_dry.gro -cs spc216.gro -o $name\_wet.gro -p $name.top
cp $name.top ../wet/$name\_wet.top
cp $name\_wet.gro ../wet/$name\_wet.gro
```

```
# Se añaden los iones (GenIon)

cp /gpfs/projects/url48/url48368/mdp/em.mdp em.mdp

grompp -f em.mdp -c $name\_wet.gro -p $name.top -o $name\_wet.tpr -po $name\_wet.mdp

genion -s $name\_wet.tpr -o $name\_ionwet.gro -pot $name\_ionwet_pot.gro -p $name.top  -pname
NA -nname CL -norandom -conc 0.150 -neutral

cp $name.top ../ionwet/$name\_ionwet.top
cp $name\_ionwet.gro ../ionwet/$name\_ionwet.gro

cd ..
echo $name > Nombre.txt
```

Script MontajeHelice.scr

```
#!/bin/bash
# @ job_name         = helice_1
# @ wall_clock_limit = 86400
# @ total_tasks      = 16
# @ initialdir       = .
# @ error            = errors_%j
# @ output           = outputs_%j
## @ tasks_per_node   = 2

JOB_ID=$SLURM_JOBID
NAME=MD_Helix_1
CLASS=default
CPU_LIMIT=18000
INIT_DIR=`pwd`
JOB_TIME=`date +%Y%m%d_%H%M%S`
NPROCS=$SLURM_NPROCS

source /gpfs/apps/GROMACS/4.5.3/64/bin/GMXRC

# El script debe estar situado en una carpeta que contenga el archivo pdb inicial Y SOLO ESTE
ARCHIVO, el resto de directorios se generarán dentro de esta carpeta

# Guarda el nombre del fichero que se usará para el resto de procesos

name=`ls -a | awk '{tmp=$0; if (match(tmp,".pdb")>0) {split(tmp,a,".pdb"); print a[1]}}'`

# Creamos los directorios de trabajo
mkdir setup
mkdir dry
mkdir wet
mkdir ionwet
mkdir ionwet/em1
mkdir ionwet/posres
mkdir ionwet/press
mkdir ionwet/heat
mkdir ionwet/npt

cp $name.pdb setup/$name.pdb
cd setup

# Generación del "Force Field": Convierte el PDB a un archivo GROMACS
pdb2gmx -ignh -ff amber03 -f $name.pdb -o $name.pdb -p $name.top -water tip3p

# Define la "Caja de simulación". Definimos las "Periodic boundary conditions (PBC)"
editconf -bt triclinic -f $name.pdb -o $name\_dry.pdb -d 0.9

# Se hace una copia de seguridad del archivo de topologías y se copian los archivos necesarios
para la solvatación
cp $name.top ../dry/$name\_dry.top
cp $name\_dry.pdb ../dry/$name\_dry.pdb

# Solvatamos el sistema

genbox -cp $name\_dry.pdb -cs spc216.gro -o $name\_wet.pdb -p $name.top

cp $name.top ../wet/$name\_wet.top
cp $name\_wet.pdb ../wet/$name\_wet.pdb

# Se añaden los iones (GenIon)

cp /gpfs/projects/url48/url48368/mdp/em.mdp em.mdp
grompp -f em.mdp -c $name\_wet.pdb -p $name.top -o $name\_wet.tpr -po $name\_wet.mdp
genion -s $name\_wet.tpr -o $name\_ionwet.pdb -pot $name\_ionwet_pot.pdb -p $name.top -pname
NA -nname CL -norandom -conc 0.150 -neutral

cp $name.top ../ionwet/$name\_ionwet.top
cp $name\_ionwet.pdb ../ionwet/$name\_ionwet.pdb

cd ..
echo $name > Nombre.txt
```

Script Producción.scr

Automatización del equilibrado para las simulaciones de MD de los modelos de MG517 para la región Nter y las hélices en solvente acuoso de la región Cter, en el superordenador Picasso.

```
# Se indican las directivas de trabajo para el ordenador Picasso:

#!/bin/bash
# @ job_name         = helice_1
# @ wall_clock_limit = 86400
# @ total_tasks      = 16
# @ initialdir       = .
# @ error            = errors_%j
# @ output           = outputs_%j
# @ tasks_per_node   = 2

JOB_ID=$SLURM_JOBID
NAME=MD_Helix_1
CLASS=default
CPU_LIMIT=18000
INIT_DIR=`pwd`
NPROCS=$SLURM_NPROCS

source /gpfs/apps/GROMACS/4.5.3/64/bin/GMXRC

name=`cat Nombre.txt`

# Archivo de parámetros de simulación (.mdp). Primero se copia el archivo con las condiciones
ya escritas y después se genera el archivo y se carga mdrun.

cd ionwet/em1
cp /gpfs/projects/url48/url48368/mdp/em.mdp em2.mdp

# Pone el nombre del archivo como título
sed "s/TEMPLATE/$name/" em2.mdp > em1.mdp
rm -f em2.mdp

grompp -f em1.mdp -c ../../setup/$name\_ionwet.gro -p ../$name\_ionwet.top -o $name\_em1.tpr -
po $name\_em1.mdp

srun mdrun -s $name\_em1.tpr -np $NPROCS -deffnm $name

cd ..
mkdir em2
cd em2

cp ../em1/em1.mdp em2.mdp

grompp -f em2.mdp -c ../em1/$name.gro -p ../$name\_ionwet.top -o $name\_em2.tpr -po
$name\_em2.mdp

srun mdrun -s $name\_em2.tpr -np $NPROCS -deffnm $name

cd ..
mkdir em3
cd em3

cp ../em2/em2.mdp em3.mdp

grompp -f em3.mdp -c ../em2/$name.gro -p ../$name\_ionwet.top -o $name\_em3.tpr -po
$name\_em3.mdp

grompp -f em3.mdp -c ../em2/$name.gro -p ../$name\_ionwet.top -o $name\_em3.tpr -po
$name\_em3.mdp

srun mdrun -s $name\_em3.tpr -np $NPROCS  -deffnm $name

cd ..
mkdir em4
cd em4

cp ../em3/em3.mdp em4.mdp
```

```
grompp -f em4.mdp -c ../em3/$name.gro -p ../$name\_ionwet.top -o $name\_em4.tpr -po
$name\_em4.mdp

srun mdrun -s $name\_em4.tpr -np $NPROCS -deffnm $name

cd ..
mkdir em5
cd em5

cp ../em4/em4.mdp em5.mdp

grompp -f em5.mdp -c ../em4/$name.gro -p ../$name\_ionwet.top -o $name\_em5.tpr -po
$name\_em5.mdp

srun mdrun -s $name\_em5.tpr -np $NPROCS4 -deffnm $name

# Equilibración del solvente: Mantenemos fija la molécula para que se rellene con aguas.
Primero se copia el archivo con las condiciones ya escritas y después se genera el archivo y
se carga mdrun.

cd ../posres
cp /gpfs/projects/url48/url48368/mdp/posres.mdp posres2.mdp
cp ../../setup/posre.itp ../posre.itp
cp ../../setup/posre_UDPGLC.itp ../posre_UDPGLC.itp
# Pone el nombre del archivo como tÃ-tulo
sed "s/TEMPLATE/$name/" posres2.mdp > posres.mdp
rm -f posres2.mdp

grompp -f posres.mdp -c ../em5/$name.gro -p ../$name\_ionwet.top -o $name\_posres.tpr -po
$name\_posres.mdp

srun mdrun -s $name\_posres.tpr -np $NPROCS -deffnm $name

# Equilibrando la temperatura.
# Primero se copia el archivo con las condiciones ya escritas y después se genera el archivo y
se carga mdrun

cd ../heat
cp /gpfs/projects/url48/url48368/mdp/heat.mdp heat2.mdp

# Pone el nombre del archivo como tÃ-tulo
sed "s/TEMPLATE/$name/" heat2.mdp > heat.mdp
rm -f heat2.mdp

grompp -f heat.mdp -c ../posres/$name.gro -t ../posres/$name.trr -p ../$name\_ionwet.top -o
$name\_heat.tpr -po $name\_heat.mdp

srun mdrun -s $name\_heat.tpr -np $NPROCS -deffnm $name

# Equilibrando la presión. Primero se copia el archivo con las condiciones ya escritas y
después se genera el archivo y se carga mdrun

cd ../press
cp /gpfs/projects/url48/url48368/mdp/press.mdp press2.mdp

# Pone el nombre del archivo como título
sed "s/TEMPLATE/$name/" press2.mdp > press.mdp
rm -f press2.mdp

grompp -f press.mdp -c ../heat/$name.gro -t ../heat/$name.trr -p ../$name\_ionwet.top -o
$name\_press.tpr -po $name\_press.mdp

srun mdrun -s $name\_press.tpr -np $NPROCS -deffnm $name

# Mantenemos la temperatura y presión constantes

cd ../npt
cp /gpfs/projects/url48/url48368/mdp/npt.mdp npt2.mdp

# Pone el nombre del archivo como título
sed "s/TEMPLATE/$name/" npt2.mdp > npt.mdp
rm -f npt2.mdp

grompp -f npt.mdp -c ../press/$name.gro -t ../press/$name.trr -p ../$name\_ionwet.top -o
$name\_npt.tpr -po $name\_npt.mdp

srun mdrun -s $name\_npt.tpr -np $NPROCS -npme 16 -ddorder interleave -deffnm $name
```

Script MolecularDynamics.scr MD Modelos MG517

Programa con el que se automatiza el lanzamiento de simulaciones. Cuando finaliza una se prepara una nueva carpeta y se lanza su continuación.

```
#!/bin/bash

of=` cat num.txt`

tpbconv -s ../npt$of/Modelo1_npt.tpr -extend 6000 -o Modelo1_npt.tpr

srun mdrun -s Modelo1_npt.tpr -np $NPROCS  -npme 16 -ddorder interleave  -deffnm Modelo1 -cpi
../npt$of/Modelo1.cpt -noaddpart

fin=` tail -1 Modelo1.log | grep "Finished mdrun" | wc -l`

if [ $fin -gt 0 ]

then

  of=` echo $of | awk '{printf "%d",$1}'`
  af=$(( of + 1 ))
  nf=$af
  af=` echo $af | awk '{printf "%04d",$1}'`
  nof=$af
  nxf=$(( of + 2 ))
  nxf=` echo $nxf | awk '{printf "%04d",$1}'`

  mkdir ../npt$nxf
  cp JOB.scr ../npt$nxf/
  cp MolecularDynamics.scr ../npt$nxf/
  cd ../npt$nxf/
  echo $af > num.txt
  mnsubmit JOB.scr

  cd ../npt$nof
  echo 0 | trjconv -f Modelo1.xtc -o Modelo1_dt10_whole$nf.xtc -dt 10 -s Modelo1_npt.tpr -pbc
whole

else

  touch ERROR

fi
```

Script Metadinámica enlazado de ejecuciones. Hélice anfipática inserta en membrana.

```
#!/bin/bash

of=`cat num.txt`
nombre=System6_Meta9VC_DC

cp ../META_INP.dat .

tpbconv_d -s ../npt$of/$nombre.tpr -extend 7500 -o $nombre.tpr

srun mdrun_d -s $nombre.tpr -np $NPROCS -deffnm $nombre -cpi ../npt$of/$nombre.cpt -plumed
META_INP.dat

if [ `grep "Finished" $nombre.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi

of=` echo $of | awk '{printf "%d",$1}'`
af=$(( of + 1 ))
nf=$af
af=` echo $af | awk '{printf "%03d",$1}'`
nof=$af
nxf=$(( of + 2 ))
nxf=` echo $nxf | awk '{printf "%03d",$1}'`

mkdir ../npt$nxf
cp JOB_VC_DC9.scr ../npt$nxf/
cp MolecularDynamics.scr ../npt$nxf/
cp HILLS ../npt$nxf/
cd ../npt$nxf/
echo $af > num.txt
jobsubmit JOB_VC_DC9.scr

cd ../npt$nof
echo 0 | trjconv_d -f $nombre.xtc -o $nombre\_wh$nf.xtc -s $nombre.tpr -pbc whole

if [ -e $nombre\_wh$nf.xtc ]
then
        rm -f $nombre.xtc
fi
```

Script Equilibrado.scr. MD Cristal 4DDZ

```
#!/bin/bash

mkdir em1
mkdir posres
mkdir press
mkdir heat
mkdir npt

name=`GpgSAPO_LAdis`

cd em1
cp ../../mdp/em.mdp em1.mdp

grompp_d -f em1.mdp -c ../$name\_ionwet.gro -p ../$name\_ionwet.top -o $name\_em1.tpr -po
$name\_em1.mdp -maxwarn 2

srun mdrun_d -s $name\_em1.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi

cd ..
mkdir em2
cd em2

cp ../em1/em1.mdp em2.mdp

grompp_d -f em2.mdp -c ../em1/$name.gro -p ../$name\_ionwet.top -o $name\_em2.tpr -po
$name\_em2.mdp -maxwarn 2

srun mdrun_d -s $name\_em2.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi

cd ..
mkdir em3
cd em3

cp ../em2/em2.mdp em3.mdp

grompp_d -f em3.mdp -c ../em2/$name.gro -p ../$name\_ionwet.top -o $name\_em3.tpr -po
$name\_em3.mdp -maxwarn 2

srun mdrun_d -s $name\_em3.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi

cd ..
mkdir em4
cd em4

cp ../em3/em3.mdp em4.mdp

grompp_d -f em4.mdp -c ../em3/$name.gro -p ../$name\_ionwet.top -o $name\_em4.tpr -po
$name\_em4.mdp -maxwarn 2

srun mdrun_d -s $name\_em4.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi
```

```
cd ..
mkdir em5
cd em5

cp ../em4/em4.mdp em5.mdp

grompp_d -f em5.mdp -c ../em4/$name.gro -p ../$name\_ionwet.top -o $name\_em5.tpr -po
$name\_em5.mdp -maxwarn 2

srun mdrun_d -s $name\_em5.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi

cd ..
cd posres
cp ../../mdp/posres.mdp posres.mdp

grompp_d -f posres.mdp -c ../em5/$name.gro -p ../$name\_ionwet.top -o $name\_posres.tpr -po
$name\_posres.mdp -maxwarn 2

srun mdrun_d -s $name\_posres.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi

cd ..
cd heat
cp ../../mdp/heat.mdp heat.mdp

grompp_d -f heat.mdp -c ../posres/$name.gro -t ../posres/$name.trr -p ../$name\_ionwet.top -o
$name\_heat.tpr -po $name\_heat.mdp -maxwarn 2

srun mdrun_d -s $name\_heat.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi

cd ..
cd press
cp ../../mdp/press.mdp press.mdp

grompp_d -f press.mdp -c ../heat/$name.gro -t ../heat/$name.trr -p ../$name\_ionwet.top -o
$name\_press.tpr -po $name\_press.mdp -maxwarn 2

srun mdrun_d -s $name\_press.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi

cd ..
cd npt
cp ../../mdp/npt.mdp npt.mdp

grompp_d -f npt.mdp -c ../press/$name.gro -t ../press/$name.trr -p ../$name\_ionwet.top -o
$name\_npt.tpr -po $name\_npt.mdp -maxwarn 2

srun mdrun_d -s $name\_npt.tpr -np $NPROCS -deffnm $name

if [ `grep "Finished" $name.log | wc -l` = 0 ]
then
        echo "ERROR"
        exit
fi
```

Script Equilibrado.scr. MD modelo 4Y6N

```
#!/bin/bash

# 11 pasos alternados de steepest descend y conjugated gradients.

grompp_d -f ../mdp/em_steep.mdp -c ../GRO_posre_itp_top/GpgS_ternario_ionwet.gro -p
../GRO_posre_itp_top/GpgS_ternario.top -o em1_descent.tpr

srun mdrun_d -s em1_descent.tpr -np $NPROCS -deffnm em1_descent -plumed ../META_INP.dat

rm -f mdout.mdp

for i in `seq 1 10`
do
        grompp_d -f ../mdp/em_cg.mdp -c em$i\_descent.gro -p
../GRO_posre_itp_top/GpgS_ternario.top -o em$i\_gradients.tpr

        srun mdrun_d -s em$i\_gradients.tpr -np $PROCS -deffnm em$i\_gradients -plumed
../META_INP.dat

        rm -f mdout.mdp

        e=$((i + 1))

        grompp_d -f ../mdp/em_steep.mdp -c em$i\_gradients.gro -p
../GRO_posre_itp_top/GpgS_ternario.top -o em$e\_descent.tpr

        srun mdrun_d -s em$e\_descent.tpr -np $NPROCS -deffnm em$e\_descent -plumed
../META_INP.dat

        rm -f mdout.mdp
done

grompp_d -f ../mdp/em_cg.mdp -c em$e\_descent.gro -p ../GRO_posre_itp_top/GpgS_ternario.top -o
em$e\_gradients.tpr

srun mdrun_d -s em$e\_gradients.tpr -np $NPROCS -deffnm em$e\_gradients -plumed
../META_INP.dat

rm -f mdout.mdp

cd posre

grompp -f ../mdp/posres.mdp -c ../em/em11_gradients.gro -p
../GRO_posre_itp_top/GpgS_ternario.top -o GpgS_posres -n ../mdp/index.ndx

mdrun -s GpgS_posres.tpr -nt 12 -deffnm GpgS_posres

cd ..
cd free

grompp -f ../mdp/free.mdp -c ../posre/GpgS_posres.gro -p
../GRO_posre_itp_top/GpgS_ternario.top -o GpgS_free -n ../mdp/index.ndx -t
../posre/GpgS_posres.trr

mdrun -s GpgS_free.tpr -nt 12 -deffnm GpgS_free

cd ..
cd heat

grompp -f ../mdp/heat.mdp -c ../free/GpgS_free.gro -p ../GRO_posre_itp_top/GpgS_ternario.top -
o GpgS_heat -n ../mdp/index.ndx -t ../free/GpgS_free.trr

mdrun -s GpgS_heat.tpr -nt 12 -deffnm GpgS_heat

cd ..
cd press

grompp -f ../mdp/press.mdp -c ../heat/GpgS_heat.gro -p ../GRO_posre_itp_top/GpgS_ternario.top
-o GpgS_press -n ../mdp/index.ndx -t ../heat/GpgS_heat.trr

mdrun -s GpgS_press.tpr -nt 12 -deffnm GpgS_press
```

Script Extracción de códigos UniProt de CAZy GTAs

```
#El listado “GTAs.txt” contiene los números de las familias GTAs (2, 6, 7, 8, 12, 13, 21, 24,
27, 43, 55, 64, 78, 81, 82, 84)

#!/bin/bash

arc=archaea
bac=bacteria
euk=eukaryota
vir=virus

for i in `cat GTAs.txt`
do
   mkdir GT$i\_seqs
   echo Bajando http://www.cazy.org/GT$i\_$arc.html
   curl http://www.cazy.org/GT$i\_$arc.html > GT$i\_seqs/$arc.html
   pages=`grep "rel='nofollow'>" GT$i\_seqs/$arc.html | grep -v "lien_pagination" | awk
'{split($0,var1,"<"); split(var1[1],var2,">"); print var2[2]}'`
   if [ -n "$pages" ]
   then
      pages=$((pages - 1))
      for e in `seq 1 $pages`
      do
         curl http://www.cazy.org/GT$i\_$arc.html?debut_PRINC=$e\000 >> GT$i\_seqs/$arc.html
      done
   else
      grep "class='lien_pagination' rel='nofollow'" GT$i\_seqs/$arc.html | awk
'{split($0,var3,"><a href="); split(var3[2],var4,"class"); print var4[1]}' | tr -d "'" >
GT$i\_seqs/paginas.txt
      for a in `cat GT$i\_seqs/paginas.txt`
      do
         curl http://www.cazy.org/$a >> GT$i\_seqs/$arc.html
      done
   rm -f GT$i\_seqs/paginas.txt
   fi
   echo Bajando http://www.cazy.org/GT$i\_$bac.html
   curl http://www.cazy.org/GT$i\_$bac.html > GT$i\_seqs/$bac.html
   pages=`grep "rel='nofollow'>" GT$i\_seqs/$bac.html | grep -v "lien_pagination" | awk
'{split($0,var1,"<"); split(var1[1],var2,">"); print var2[2]}'`
   if [ -n "$pages" ]
   then
      pages=$((pages - 1))
      for e in `seq 1 $pages`
      do
         curl http://www.cazy.org/GT$i\_$bac.html?debut_PRINC=$e\000 >> GT$i\_seqs/$bac.html
      done
   else
      grep "class='lien_pagination' rel='nofollow'" GT$i\_seqs/$bac.html | awk
'{split($0,var3,"><a href="); split(var3[2],var4,"class"); print var4[1]}' | tr -d "'" >
GT$i\_seqs/paginas.txt
      for a in `cat GT$i\_seqs/paginas.txt`
      do
         curl http://www.cazy.org/$a >> GT$i\_seqs/$bac.html
      done
   rm -f GT$i\_seqs/paginas.txt
   fi
   echo Bajando http://www.cazy.org/GT$i\_$euk.html
   curl http://www.cazy.org/GT$i\_$euk.html > GT$i\_seqs/$euk.html
   pages=`grep "rel='nofollow'>" GT$i\_seqs/$euk.html | grep -v "lien_pagination" | awk
'{split($0,var1,"<"); split(var1[1],var2,">"); print var2[2]}'`
   if [ -n "$pages" ]
   then
      pages=$((pages - 1))
      for e in `seq 1 $pages`
      do
         curl http://www.cazy.org/GT$i\_$euk.html?debut_PRINC=$e\000 >> GT$i\_seqs/$euk.html
      done
   else
      grep "class='lien_pagination' rel='nofollow'" GT$i\_seqs/$euk.html | awk
'{split($0,var3,"><a href="); split(var3[2],var4,"class"); print var4[1]}' | tr -d "'" >
GT$i\_seqs/paginas.txt
      for a in `cat GT$i\_seqs/paginas.txt`
      do
         curl http://www.cazy.org/$a >> GT$i\_seqs/$euk.html
      done
```

```
   rm -f GT$i\_seqs/paginas.txt
   fi
   echo Bajando http://www.cazy.org/GT$i\_$vir.html
   curl http://www.cazy.org/GT$i\_$vir.html > GT$i\_seqs/$vir.html
   pages=`grep "rel='nofollow'>" GT$i\_seqs/$vir.html | grep -v "lien_pagination" | awk
'{split($0,var1,"<"); split(var1[1],var2,">"); print var2[2]}'`
   if [ -n "$pages" ]
   then
      pages=$((pages - 1))
      for e in `seq 1 $pages`
      do
         curl http://www.cazy.org/GT$i\_$vir.html?debut_PRINC=$e\000 >> GT$i\_seqs/$vir.html
      done
   else
      grep "class='lien_pagination' rel='nofollow'" GT$i\_seqs/$vir.html | awk
'{split($0,var3,"><a href="); split(var3[2],var4,"class"); print var4[1]}' | tr -d "'" >
GT$i\_seqs/paginas.txt
      for a in `cat GT$i\_seqs/paginas.txt`
      do
         curl http://www.cazy.org/$a >> GT$i\_seqs/$vir.html
      done
   rm -f GT$i\_seqs/paginas.txt
   fi
done
```

Script Extracción de secuencias fasta UniProt de CAZy GTAs

```
#El listado “GTAs.txt” contiene los números de las familias GTAs (2, 6, 7, 8, 12, 13, 21, 24,
27, 43, 55, 64, 78, 81, 82, 84)

#!/bin/bash

arc=archaea
bac=bacteria
euk=eukaryota
vir=virus

for i in `cat GTAs.txt`
do
   echo bajando secuencias GT$i\_$arc
   mkdir GT$i\_seqs/$arc
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$arc.html | awk
'{num=split($0,target,"br"); for (i=1; i<=num; i++) {split(target[i],uni,"/");
split(uni[5],cod,"\""); print cod[1]}}' > tmp
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$arc.html | awk
'{num=split($0,target,"br"); for (i=1; i<=num; i++) {split(target[i],uni,"/");
split(uni[5],cod,"\""); print cod[1]}}' >> tmp
   sort tmp | uniq > GT$i\_seqs/$arc\_codigosUniprot
   for e in `cat GT$i\_seqs/$arc\_codigosUniprot`; do echo GT_$i\_$arc\_$e >>
All_UP_in_CAZY.txt; echo $e >> All_UP_in_CAZY_just_codes.txt; done

   for u in `cat GT$i\_seqs/$arc\_codigosUniprot`
   do
      curl http://www.uniprot.org/uniprot/$u.fasta > tmp
      cod=`grep ">" tmp | cut -c 5-10`
      if [ -z $cod ]
      then
         echo GT_$i\_$arc\_$u >> UP_obsoletes_in_CAZY.txt
         echo $u >> UP_obsoletes_in_CAZY_just_codes.txt
      else
         mv -f tmp GT$i\_seqs/$arc/GT$i\_$arc\_$u.fasta
      fi
      if [ "$u" != "$cod" ]
      then
         echo GT_$i\_$arc\_$u changed to GT_$i\_$arc\_$cod >> UP_moved_in_CAZY.txt
         echo $u >> UP_moved_in_CAZY_just_codes.txt
      fi
   done
   echo bajando secuencias GT$i\_$bac
   mkdir GT$i\_seqs/$bac
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$bac.html | awk
'{num=split($0,target,"br"); for (i=1; i<=num; i++) {split(target[i],uni,"/");
split(uni[5],cod,"\""); print cod[1]}}' > tmp
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$bac.html | awk
'{num=split($0,target,"br"); for (i=1; i<=num; i++) {split(target[i],uni,"/");
split(uni[5],cod,"\""); print cod[1]}}' >> tmp
   sort tmp | uniq > GT$i\_seqs/$bac\_codigosUniprot
   for e in `cat GT$i\_seqs/$bac\_codigosUniprot`; do echo GT_$i\_$bac\_$e >>
All_UP_in_CAZY.txt; echo $e >> All_UP_in_CAZY_just_codes.txt; done

   for u in `cat GT$i\_seqs/$bac\_codigosUniprot`
   do
      curl http://www.uniprot.org/uniprot/$u.fasta > tmp
      cod=`grep ">" tmp | cut -c 5-10`
      if [ -z $cod ]
      then
         echo GT_$i\_$bac\_$u >> UP_obsoletes_in_CAZY.txt
         echo $u >> UP_obsoletes_in_CAZY_just_codes.txt
      else
         mv -f tmp GT$i\_seqs/$bac/GT$i\_$bac\_$u.fasta
      fi
      if [ "$u" != "$cod" ]
      then
        echo GT_$i\_$bac\_$u changed to GT_$i\_$bac\_$cod >> UP_moved_in_CAZY.txt
        echo $u >> UP_moved_in_CAZY_just_codes.txt
      fi
   done
   echo bajando secuencias GT$i\_$euk
   mkdir GT$i\_seqs/$euk
```

```
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$euk.html | awk
'{num=split($0,target,"br"); for (i=1; i<=num; i++) {split(target[i],uni,"/");
split(uni[5],cod,"\""); print cod[1]}}' > tmp
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$euk.html | awk
'{num=split($0,target,"br"); for (i=1; i<=num; i++) {split(target[i],uni,"/");
split(uni[5],cod,"\""); print cod[1]}}' >> tmp
   sort tmp | uniq > GT$i\_seqs/$euk\_codigosUniprot
   for e in `cat GT$i\_seqs/$euk\_codigosUniprot`; do echo GT_$i\_$euk\_$e >>
All_UP_in_CAZY.txt; echo $e >> All_UP_in_CAZY_just_codes.txt; done

   for u in `cat GT$i\_seqs/$euk\_codigosUniprot`
   do
      curl http://www.uniprot.org/uniprot/$u.fasta > tmp
      cod=`grep ">" tmp | cut -c 5-10`
      if [ -z $cod ]
      then
         echo GT_$i\_$euk\_$u >> UP_obsoletes_in_CAZY.txt
         echo $u >> UP_obsoletes_in_CAZY_just_codes.txt
      else
         mv -f tmp GT$i\_seqs/$euk/GT$i\_$euk\_$u.fasta
      fi
      if [ "$u" != "$cod" ]
      then
         echo GT_$i\_$euk\_$u changed to GT_$i\_$euk\_$cod >> UP_moved_in_CAZY.txt
         echo $u >> UP_moved_in_CAZY_just_codes.txt
      fi
   done
   echo bajando secuencias GT$i\_$vir
   mkdir GT$i\_seqs/$vir
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$vir.html | awk
'{num=split($0,target,"br"); for (i=1; i<=num; i++) {split(target[i],uni,"/");
split(uni[5],cod,"\""); print cod[1]}}' > tmp
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$vir.html | awk
'{num=split($0,target,"br"); for (i=1; i<=num; i++) {split(target[i],uni,"/");
split(uni[5],cod,"\""); print cod[1]}}' >> tmp
   sort tmp | uniq > GT$i\_seqs/$vir\_codigosUniprot
   for e in `cat GT$i\_seqs/$vir\_codigosUniprot`; do echo GT_$i\_$vir\_$e >>
All_UP_in_CAZY.txt; echo $e >> All_UP_in_CAZY_just_codes.txt; done

   for u in `cat GT$i\_seqs/$vir\_codigosUniprot`
   do
      curl http://www.uniprot.org/uniprot/$u.fasta > tmp
      cod=`grep ">" tmp | cut -c 5-10`
      if [ -z $cod ]
      then
         echo GT_$i\_$vir\_$u >> UP_obsoletes_in_CAZY.txt
         echo $u >> UP_obsoletes_in_CAZY_just_codes.txt
      else
         mv -f tmp GT$i\_seqs/$vir/GT$i\_$vir\_$u.fasta
      fi
      if [ "$u" != "$cod" ]
      then
        echo GT_$i\_$vir\_$u changed to GT_$i\_$vir\_$cod >> UP_moved_in_CAZY.txt
        echo $u >> UP_moved_in_CAZY_just_codes.txt
      fi
   done
done
```

Script Extracción de la información de texto que contiene cada entrada UniProt de CAZy GTAs

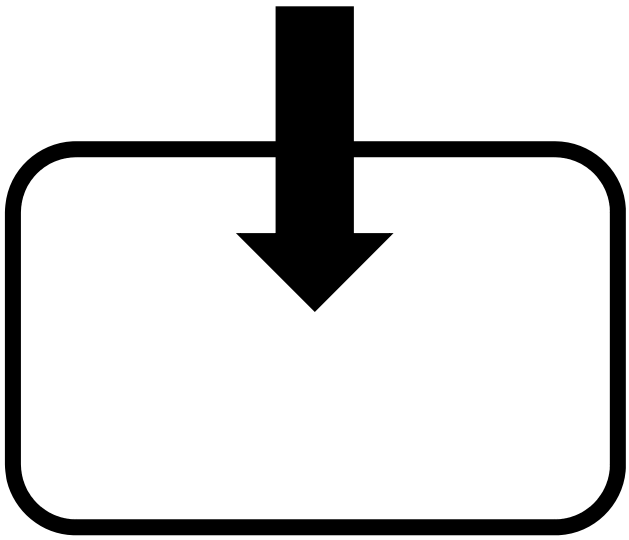
```
#!/bin/bash

arc=archaea
bac=bacteria
euk=eukaryota
vir=virus

for i in `cat GTAs.txt`
do
   echo bajando secuencias GT$i\_$arc
   mkdir GT$i\_seqs/$arc
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$arc.html | awk '{split($3,uni,"/");
split(uni[5],cod,"\""); print cod[1]}' > GT$i\_seqs/$arc\_codigosUniprot
   ls GT$i\_seqs/$arc/GT$i\_$arc\_*.fasta | cut -d "_" -f 4 | cut -d "." -f 1 > tmp
   for u in `cat tmp`
   do
      if [ ! -f GT$i\_seqs/$arc/GT$i\_$arc\_$u.txt ]
      then
        echo "Escribiendo GT"$i"_seqs/"$arc"_"$u".txt"
        curl http://www.uniprot.org/uniprot/$u.txt > GT$i\_seqs/$arc/GT$i\_$arc\_$u.txt
      fi
   done
   echo bajando secuencias GT$i\_$bac
   mkdir GT$i\_seqs/$bac
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$bac.html | awk '{split($3,uni,"/");
split(uni[5],cod,"\""); print cod[1]}' > GT$i\_seqs/$bac\_codigosUniprot
   ls GT$i\_seqs/$bac/GT$i\_$bac\_*.fasta | cut -d "_" -f 4 | cut -d "." -f 1 > tmp
   for u in `cat tmp`
   do
      if [ ! -f GT$i\_seqs/$bac/GT$i\_$bac\_$u.txt ]
      then
        echo "Escribiendo GT"$i"_seqs/"$bac"_"$u".txt"
        curl http://www.uniprot.org/uniprot/$u.txt > GT$i\_seqs/$bac/GT$i\_$bac\_$u.txt
      fi
   done
   echo bajando secuencias GT$i\_$euk
   mkdir GT$i\_seqs/$euk
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$euk.html | awk '{split($3,uni,"/");
split(uni[5],cod,"\""); print cod[1]}' > GT$i\_seqs/$euk\_codigosUniprot
   ls GT$i\_seqs/$euk/GT$i\_$euk\_*.fasta | cut -d "_" -f 4 | cut -d "." -f 1 > tmp
   for u in `cat tmp`
   do
      if [ ! -f GT$i\_seqs/$euk/GT$i\_$euk\_$u.txt ]
      then
        echo "Escribiendo GT"$i"_seqs/"$euk"_"$u".txt"
        curl http://www.uniprot.org/uniprot/$u.txt > GT$i\_seqs/$euk/GT$i\_$euk\_$u.txt
      fi
   done
   echo bajando secuencias GT$i\_$vir
   mkdir GT$i\_seqs/$vir
   grep "http://www.uniprot.org/uniprot/" GT$i\_seqs/$vir.html | awk '{split($3,uni,"/");
split(uni[5],cod,"\""); print cod[1]}' > GT$i\_seqs/$vir\_codigosUniprot
   ls GT$i\_seqs/$vir/GT$i\_$vir\_*.fasta | cut -d "_" -f 4 | cut -d "." -f 1 > tmp
   for u in `cat tmp`
   do
      if [ ! -f GT$i\_seqs/$vir/GT$i\_$vir\_$u.txt ]
        then
        echo "Escribiendo GT"$i"_seqs/"$vir"_"$u".txt"
        curl http://www.uniprot.org/uniprot/$u.txt > GT$i\_seqs/$vir/GT$i\_$vir\_$u.txt
      fi
   done
done

rm -f tmp*
```

Script Extracción del nº EC y otra información de cada entrada UniProt de CAZy GTAs

```
#!/bin/bash

rm -f GTAs_EC.txt

echo
"Family;Kingdom;Organism;Protein_name;EC_clean;Donnor;Acceptor;Size;UniProt_cod;CAZY_cod;Revie
wed;EC_dirty" > GTAs_EC.txt

uni=`ls GT*_seqs/*/*.fasta`

for i in $uni
do
  route=`echo $i | awk '{tmp=$0; split(tmp,var1,"."); print var1[1]}'`
  if [ `grep "Fragment" $i | wc -l` -eq 0 ]; then
    famil=`echo $i | awk '{tmp=$0; split(tmp,var1,"/"); split(var1[3],var2,".");
split(var2[1],var3,"_"); print var3[1]}'`
    kingd=`echo $i | awk '{tmp=$0; split(tmp,var1,"/"); split(var1[3],var2,".");
split(var2[1],var3,"_"); print var3[2]}'`
    CodUni2=`echo $i | awk '{tmp=$0; split(tmp,var1,"/"); split(var1[3],var2,".");
split(var2[1],var3,"_"); print var3[3]}'`
    name=`grep "^OS" $route.txt | awk '{split($0,fields3," "); print fields3[2] " "
fields3[3]}'`
    revie=`grep "^ID" $route.txt | awk '{split($0,fields4," "); print fields4[3]}'`
    size=`grep "^ID" $route.txt | awk '{split($0,fields5," "); print fields5[4]}'`
    CodUni=`grep "^AC" $route.txt | awk '{split($0,fields6," ");split(fields6[2],name,";");
print name[1]}'`
    rname=`grep "^DE" $route.txt | awk '{split($0,fields7,"Full=");split(fields7[2],name,";");
print name[1]}'`
    grep "^DE" $route.txt | awk '{split($0,fields8,"EC=");split(fields8[2],nameEC,";"); print
nameEC[1]}' > tmp
    grep "^CC" $route.txt | awk '{if (match($0,"CC   -!-") == 1  && match($0,"CC   -!-
CATALYTIC ACTIVITY:") == 0) {flag=0}; if (match ($0,"CC   -!- CATALYTIC ACTIVITY:")) {flag=1;
split($0,lig,":"); print lig[2]}; if (flag==1){split($0,ligcont,"CC       "); print
ligcont[2]}}' > tmp2
    cat tmp2 | tr -d "\n" > tmp3
    Donnor=`cat tmp3 | awk '{num=split($0,d,".");for (i=1; i<=num; i++) {split
(d[i],part1,"=");split (part1[1],part2,"+"); print part2[1]}}'`
    Accep=`cat tmp3 | awk '{num=split($0,d,".");for (i=1; i<=num; i++) {split
(d[i],part1,"=");split (part1[1],part2,"+"); print part2[2]}}'`
    EC=`cat tmp | awk '{split($0,EC," "); print EC[1]" "EC[3]" "EC[5]" "EC[7]}'`
    ECdirty=`grep "^DE" $route.txt | awk
'{split($0,fields8,"EC=");split(fields8[2],nameEC,";"); print nameEC[1]}'`
    echo
$famil";"$kingd";"$name";"$rname";"$EC";"$Donnor";"$Accep";"$size";"$CodUni";"$CodUni2";"$revi
e";"$ECdirty >> GTAs_EC.txt
  fi

done
rm -f tmp*
```

Scrip para realizar la "prueba de solidez" de los perfiles HMM

```
#!/bin/bash

for e in `ls * | grep ".sto" | awk '{split($0,cod,".sto"); print cod[1]}'`; do hmmbuild --
enone --amino $e.hmm $e.sto; done

for e in `ls * | grep ".sto" | awk '{split($0,cod,".sto"); print cod[1]}'`
do
        mkdir $e
        mkdir $e/sto
        mkdir $e/Not_$e\_hmm

        for i in `ls *hmm | grep -v "$e"`; do cp -f $i $e/Not_$e\_hmm; done

        for i in `cat $e.sto | awk '{split($0,cod,":"); print cod[4]}'`
        do
                mkdir $e/$i\_hmmpress
                grep -v $i $e.sto > $e/sto/$e\_but_$i.sto

                hmmbuild --enone --amino $e/sto/$e\_but_$i.hmm $e/sto/$e\_but_$i.sto

                cod=`grep $i $e.sto`
                echo "# STOCKHOLM 1.0" > $e/$i\_hmmpress/$i.sto
                echo $cod | tr -d "\r" >> $e/$i\_hmmpress/$i.sto
                echo "//" >> $e/$i\_hmmpress/$i.sto
        done

        for i in `cat $e.sto | awk '{split($0,cod,":"); print cod[4]}'`
        do
                cat $e/Not_$e\_hmm/*hmm $e/sto/$e\_but_$i.hmm >>
$e/$i\_hmmpress/$e\_but_$i\_DB.hmm

                hmmpress $e/$i\_hmmpress/$e\_but_$i\_DB.hmm

                hmmscan --max --tblout $e/$i\_hmmpress/$i.out
$e/$i\_hmmpress/$e\_but_$i\_DB.hmm $e/$i\_hmmpress/$i.sto

        done
done
```



Inputs

Modelado MG517.

Input para MODELLER. Ejemplo del modelo híbrido 2Z86_2 (zona conservada), 3CU0 (RV) para la secuencia Q9ZB73.

```
from modeller import *
from modeller.automodel import *

log.verbose()
env = environ(rand_seed=-8127)

class MyModel(loopmodel):

    def special_restraints(self, aln):              # Eliminación de restricciones para K34
        rsr = self.restraints
        at = self.atoms
       rsr.unpick(self.residue_range('34:', '34:'))
       rsr.condense()

a = MyModel(env, alnfile='Alineamiento.pir',
              knowns=('2Z86_2','3CU0'), sequence='Q9ZB73',
              assess_methods=(assess.DOPE, assess.GA341, assess.DOPEHR,
assess.normalized_dope),
              loop_assess_methods=(assess.DOPE, assess.DOPEHR, assess.GA341,
assess.normalized_dope))

a.starting_model = 1
a.ending_model = 10                  # 10 modelos diferentes por cada híbrido
a.library_schedule = autosched.slow
a.max_var_iterations = 300
a.md_level = refine.very_slow
a.final_malign3d = True                     # Alineamiento de las estructuras a la plantilla
a.loop.starting_model = 1
a.loop.ending_model   = 10                  # 10 versiones de loop por cada modelo
a.loop.library_schedule = autosched.slow
a.loop.max_var_iterations = 300
a.loop.md_level = refine.very_slow          # Modo de refinamiento: muy lento
a.loop.final_malign3d = True         # Alineamiento de los loops a la plantilla
a.make()
```

Modelado hélices

Input para MODELLER.

```
# Example for model.build_sequence(), secondary_structure.alpha()

from modeller import *
from modeller.optimizers import conjugate_gradients
from modeller.optimizers import molecular_dynamics

# Set up environment
e = environ()
e.libs.topology.read('${LIB}/top_heav.lib')
e.libs.parameters.read('${LIB}/par.lib')

# Build an extended chain model from primary sequence, and write it out
m = model(e)
m.build_sequence('Incluir la secuencia a modelar')
m.write(file='extended-chain.pdb')

# Make stereochemical restraints on all atoms
allatoms = selection(m)
m.restraints.make(allatoms, restraint_type='STEREO', spline_on_site=False)

# Constrain all residues to be alpha-helical
# (Could also use m.residue_range() rather than m.residues here.)
m.restraints.add(secondary_structure.alpha(m.residues))

# Get an optimized structure with CG, and write it out
cg = conjugate_gradients()
md = molecular_dynamics()

cg.optimize(allatoms, max_iterations=100)
cg.optimize(allatoms, max_iterations=100)

m.write(file='alpha-helix1.pdb')

md.optimize(allatoms, temperature=100, max_iterations=1000000)
md.optimize(allatoms, temperature=300, max_iterations=1000000)
md.optimize(allatoms, temperature=500, max_iterations=1000000)
md.optimize(allatoms, temperature=750, max_iterations=1000000)
md.optimize(allatoms, temperature=1000, max_iterations=1000000)
md.optimize(allatoms, temperature=1300, max_iterations=1000000)
md.optimize(allatoms, temperature=1600, max_iterations=1000000)

m.write(file='alpha-helix2.pdb')

md.optimize(allatoms, temperature=1600, max_iterations=1000000)
md.optimize(allatoms, temperature=1300, max_iterations=1000000)
md.optimize(allatoms, temperature=1000, max_iterations=1000000)
md.optimize(allatoms, temperature=750, max_iterations=1000000)
md.optimize(allatoms, temperature=500, max_iterations=1000000)
md.optimize(allatoms, temperature=300, max_iterations=1000000)
md.optimize(allatoms, temperature=100, max_iterations=1000000)

m.write(file='alpha-helix3.pdb')

cg.optimize(allatoms, max_iterations=1000)
cg.optimize(allatoms, max_iterations=1000)
cg.optimize(allatoms, max_iterations=1000)
cg.optimize(allatoms, max_iterations=1000)
cg.optimize(allatoms, max_iterations=1000)

m.write(file='alpha-helix4.pdb')
```

Modelado GpgS. Cristal 4DDZ, generación del .ini

```
from modeller import *
from modeller.automodel import *

log.verbose()
env = environ(rand_seed=-8153)

class MyModel(loopmodel):

    def select_atoms(self):
        return selection(self.residue_range('144:A', '159:A'), self.residue_range('272:A',
'278:A'), self.residue_range('443:B', '458:B'), self.residue_range('571:B', '577:B'))

    def special_patches(self, aln):
        self.patch(residue_type='DISU', residues=(self.residues['156:A'],
self.residues['455:B']))


a = MyModel (env, alnfile='Alineamiento.pir',
              knowns=('GpgSApo'), sequence='P9WMW9',
              assess_methods=(assess.DOPE, assess.GA341,
assess.DOPEHR,assess.normalized_dope),
              loop_assess_methods=(assess.DOPE, assess.DOPEHR, assess.GA341,
assess.normalized_dope))

a.starting_model = 1
a.ending_model = 10
a.library_schedule = autosched.slow
a.max_var_iterations = 300
a.md_level = refine.very_slow
a.final_malign3d = True                 # Align new structures to template, and write them to
_fit.pdb

a.loop.starting_model = 1               # First loop model
a.loop.ending_model   = 10              # Last loop model
a.loop.library_schedule = autosched.slow
a.loop.max_var_iterations = 300
a.loop.md_level = refine.very_slow # Loop model refinement level
a.loop.final_malign3d = True
a.make()
```

Modelado GpgS. Cristal 4DDZ

```
from modeller import *
from modeller.automodel import *

log.verbose()
env = environ(rand_seed=-8153)

class MyModel(loopmodel):

    def select_atoms(self):
        return selection(self.residue_range('144:A', '159:A'), self.residue_range('272:A',
'278:A'), self.residue_range('443:B', '458:B'), self.residue_range('571:B', '577:B'))

    def special_patches(self, aln):
        self.patch(residue_type='DISU', residues=(self.residues['156:A'],
self.residues['455:B']))


a = MyModel (env, alnfile='Alineamiento.pir',
              knowns=('GpgSApo'), sequence='P9WMW9', inifile='INI_GpgSApoLA.pdb',
              assess_methods=(assess.DOPE, assess.GA341,
assess.DOPEHR,assess.normalized_dope),
              loop_assess_methods=(assess.DOPE, assess.DOPEHR, assess.GA341,
assess.normalized_dope))

a.starting_model = 1
a.ending_model = 10
a.library_schedule = autosched.slow
a.max_var_iterations = 300
a.md_level = refine.very_slow
a.final_malign3d = True                 # Align new structures to template, and write them to
_fit.pdb

a.loop.starting_model = 1               # First loop model
a.loop.ending_model   = 10              # Last loop model
a.loop.library_schedule = autosched.slow
a.loop.max_var_iterations = 300
a.loop.md_level = refine.very_slow # Loop model refinement level
a.loop.final_malign3d = True
a.make()
```

Modelado GpgS: Cristal 4DDZ. Alineamiento de entrada

```
sequence:Q7U0E1:1:A:598:B:GT81:Mycobacterium bovis:0.00:0.00
TTWHRPGWTIGELEAAKAGRTISVVLPALNEEATIESVIDSISPLVDGLVDELIVLDSGSTDDTEIRAIASG
ARVVSREQALPEVPVRPGKGEALWRSLAATSGDIVVFIDSDLINPHPLFVPWLVGPLLTGEGIQLVKSFYRR
PLQVSDVTSGVCATGGGRVTELVARPLLAALRPELGCVLQPLSGEYAASRELLTSLPFAPGYGVEIGLLIDT
FDRLGLDAIAQVNLGVRAHRNRPLDELGAMSRQVIATLLSRCGIPDSGVGLTQFLPGGPDDSDYTRHTWPVS
LVDRPPMKVMR
/
TTWHRPGWTIGELEAAKAGRTISVVLPALNEEATIESVIDSISPLVDGLVDELIVLDSGSTDDTEIRAIASG
ARVVSREQALPEVPVRPGKGEALWRSLAATSGDIVVFIDSDLINPHPLFVPWLVGPLLTGEGIQLVKSFYRR
PLQVSDVTSGVCATGGGRVTELVARPLLAALRPELGCVLQPLSGEYAASRELLTSLPFAPGYGVEIGLLIDT
FDRLGLDAIAQVNLGVRAHRNRPLDELGAMSRQVIATLLSRCGIPDSGVGLTQFLPGGPDDSDYTRHTWPVS
LVDRPPMKVMR*

>P1;GpgSApoBC
structureX:GpgSApoBC:1:A:598:B:GT81:Mycobacterium tuberculosis:0.00:0.00
TTWHRPGWTIGELEAAKAGRTISVVLPALNEEATIESVIDSISPLVDGLVDELIVLDSGSTDDTEIRAIASG
ARVVSREQALPEVPVRPGKGEALWRSLAATSGDIVVFIDSDLINPHPLFVPWLVGPLLTGEGIQLVKSFYR-
---------------GGRVTELVARPLLAALRPELGCVLQPLSGEYAASRELLTSLPFAPGYGVEIGLLIDT
FDRLGLDAIAQVNLGVRAHRNRPLDELGAMSRQVIATLLSRCGIPDSGVGLTQFL-------DYTRHTWPVS
LVDRPPMKVMR
/
TTWHRPGWTIGELEAAKAGRTISVVLPALNEEATIESVIDSISPLVDGLVDELIVLDSGSTDDTEIRAIASG
ARVVSREQALPEVPVRPGKGEALWRSLAATSGDIVVFIDSDLINPHPLFVPWLVGPLLTGEGIQLVKSFYR-
---------------GGRVTELVARPLLAALRPELGCVLQPLSGEYAASRELLTSLPFAPGYGVEIGLLIDT
FDRLGLDAIAQVNLGVRAHRNRPLDELGAMSRQVIATLLSRCGIPDSGVGLTQFL-------DYTRHTWPVS
LVDRPPMKVMR*
```

Modelado GpgS. Cristal 4Y6N

```
from modeller import *
from modeller.automodel import *

log.verbose()
env = environ(rand_seed=-8153)

class MyModel(loopmodel):

    def select_atoms(self):
        return selection(self.residue_range('149:A', '162:A'), self.residue_range('277:A',
'284:A'), self.residue_range('456:B', '469:B'), self.residue_range('584:B', '591:B'))

    def special_patches(self, aln):
        self.patch(residue_type='DISU', residues=(self.residues['161:A'],
self.residues['468:B']))


a = MyModel (env, alnfile='Alineamiento.pir',
              knowns=('GpgSUDPgPGA'), sequence='P9WMW9', inifile=' INI_GpgSApoLA.pdb ',
              assess_methods=(assess.DOPE, assess.GA341,
assess.DOPEHR,assess.normalized_dope),
              loop_assess_methods=(assess.DOPE, assess.DOPEHR, assess.GA341,
assess.normalized_dope))

a.starting_model = 1
a.ending_model = 10
a.library_schedule = autosched.slow
a.max_var_iterations = 300
a.md_level = refine.very_slow
a.final_malign3d = True                 # Align new structures to template, and write them to
_fit.pdb

a.loop.starting_model = 1               # First loop model
a.loop.ending_model   = 10              # Last loop model
a.loop.library_schedule = autosched.slow
a.loop.max_var_iterations = 300
a.loop.md_level = refine.very_slow # Loop model refinement level
a.loop.final_malign3d = True
a.make()
```

Modelado del cristal 4Y6N. Alineamiento de entrada

```
>P1;P9WMW9
sequence:P9WMW9:1:A:322::GT81:MYCTU:0.00:0.00
ALPLDTTWHRPGWTIGELEAAKAGRTISVVLPALNEEATIESVIDSISPLVDGLVDELIVLDSGSTDDTEIR
AIASGARVVSREQALPEVPVRPGKGEALWRSLAATSGDIVVFIDSDLINPHPLFVPWLVGPLLTGEGIQLVK
SFYRRPLQVSDVTSGVCATGGGRVTELVARPLLAALRPELGCVLQPLSGEYAASRELLTSLPFAPGYGVEIG
LLIDTFDRLGLDAIAQVNLGVRAHRNRPLDELGAMSRQVIATLLSRCGIPDSGVGLTQFLPGGPDDSDYTRH
TWPVSLVDRPPMKVMR...
/
ALPLDTTWHRPGWTIGELEAAKAGRTISVVLPALNEEATIESVIDSISPLVDGLVDELIVLDSGSTDDTEIR
AIASGARVVSREQALPEVPVRPGKGEALWRSLAATSGDIVVFIDSDLINPHPLFVPWLVGPLLTGEGIQLVK
SFYRRPLQVSDVTSGVCATGGGRVTELVARPLLAALRPELGCVLQPLSGEYAASRELLTSLPFAPGYGVEIG
LLIDTFDRLGLDAIAQVNLGVRAHRNRPLDELGAMSRQVIATLLSRCGIPDSGVGLTQFLPGGPDDSDYTRH
TWPVSLVDRPPMKVMR...*
>P1;GpgSUDPgPGA
structureX:GpgSUDPgPGA:19:A:325:B:GT81:Mycobacterium tuberculosis:0.00:0.00
ALPLDTTWHRPGWTIGELEAAKAGRTISVVLPALNEEATIESVIDSISPLVDGLVDELIVLDSGSTDDTEIR
AIASGARVVSREQALPEVPVRPGKGEALWRSLAATSGDIVVFIDSDLINPHPLFVPWLVGPLLTGEGIQLVK
SFYR--------------TGGGRVTELVARPLLAALRPELGCVLQPLSGEYAASRELLTSLPFAPGYGVEIG
LLIDTFDRLGLDAIAQVNLGVRAHRNRPLDELGAMSRQVIATLLSRCGIPDSGVGLTQFL--------YTRH
TWPVSLVDRPPMKVMR...
/
ALPLDTTWHRPGWTIGELEAAKAGRTISVVLPALNEEATIESVIDSISPLVDGLVDELIVLDSGSTDDTEIR
AIASGARVVSREQALPEVPVRPGKGEALWRSLAATSGDIVVFIDSDLINPHPLFVPWLVGPLLTGEGIQLVK
SFYR--------------TGGGRVTELVARPLLAALRPELGCVLQPLSGEYAASRELLTSLPFAPGYGVEIG
LLIDTFDRLGLDAIAQVNLGVRAHRNRPLDELGAMSRQVIATLLSRCGIPDSGVGLTQFL--------YTRH
TWPVSLVDRPPMKVMR...*
```

Energía de minimización: Nter MG517 y hélices en solvente acuoso

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp       ; Preprocessor

; The following line passes specific messages to the program
define          = -DFLEXIBLE     ; This defines the solvent molecules as flexible

; Parameters describing what to do, when to stop and what to save
integrator      = cg       ; Algorithm (steep = steepest descent minimization; cg = conjugate
gradients; md = newtonian molecular dynamics)
init_step       = 0        ; The starting step
nsteps          = 10000    ; Maximum number of steps to perform
nstlog          = 1        ; Frequency to write energies to log file
nstxout         = 0        ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 0        ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 100      ; Frequency to write energies to output energy file (.edr)
energygrps      = System   ; Which energy group(s) to write to output energy file (.edr)
nstxtcout       = 1  ; Frequency to write compressed coordinates to output trajectory file
(.xtc)
xtc_grps        = System   ; Which compressed coordinate group(s) to write to output
trajectory file (.xtc)

; Energy minimizing specific parameters
emtol           = 1.0                  ; Stop minimization when the maximum force < 1.0
kJ/mol/nm
emstep          = 0.01                 ; initial step-size
nstcgsteep      = 1000                 ; frequency of performing 1 steepest descent step while
doing conjugate gradient energy minimization

; Molecular dynamics specific parameters
dt              = 0.001                ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear               ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                   ; frequency for center of mass motion removal
tcoupl          = no                   ; Thermostat type (berendsen, nose-hoover)
pcoupl          = no                   ; Barostat type (berendsen, Parrinello-Rahman)

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                   ; generate velocities at startup

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                  ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                 ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10               ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                  ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                  ; long range electrostatic cut-off
coulombtype     = PME                  ; Treatment of long range electrostatic interactions
pme_order       = 4                    ; Interpolation order for PME
fourierspacing  = 0.12                 ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                  ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                  ; long range Van der Waals cut-off
vdwtype         = cut-off              ; Treatment of long range van der waals interactions

; System constraints
constraints     = none                 ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
```

Equilibrado de las aguas con position restrains: Nter MG517 y hélices en solvente acuoso

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES -DPOSRESUGC ; This defines the macromolecule as rigid

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 100000                ; Maximum number of steps to perform
nstlog          = 500                   ; Frequency to write energies to log file
nstxout         = 5000     ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 5000     ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 500                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein non-protein  ; Which energy group(s) to write to output energy file
(.edr)
nstxtcout       = 500; Frequency to write compressed coordinates to output trajectory file
(.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.001                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = no                    ; Thermostat type (berendsen, nose-hoover)
pcoupl          = no                    ; Barostat type (berendsen, Parrinello-Rahman)
4.5e-5 bar-1

; Molecylar dynamics specific parameters (startup)
gen_vel         = yes                   ; generate velocities at startup
gen_temp        = 300.                  ; Maxwell distribution temperature for initial
velocities
gen_seed        = 173529                ; used to initialize random generator for random
velocitie
annealing       = no                    ; Type of annealing for each temperature group (single,
periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = none                  ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
```

Equilibrado de la temperatura: Nter MG517 y hélices en solvente acuoso

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          =                       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 750000                ; Maximum number of steps to perform
nstlog          = 500                   ; Frequency to write energies to log file
nstxout         = 5000     ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 5000     ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 500                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein non-protein  ; Which energy group(s) to write to output energy file
(.edr)
nstxtcout       = 500                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein non-protein  ; groups to couple separately to temperature bath
tau_t           = 2. 2.                 ; time constant for temperature coupling
ref_t           = 300. 300.  ; reference temperature for coupling (one for each group in
tc_grps)
pcoupl          = no                    ; Barostat type (berendsen, Parrinello-Rahman)

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = single none           ; Type of annealing for each temperature group (single,
periodic)
annealing_npoints = 6 0                 ; A list with the number of annealing reference/control
points used for each temperature group
annealing_time  = 0 20 40 60 80 100         ; List of times at the annealing reference/control
points for each group
annealing_temp  = 175 200 200 250 250 300   ; List of temperatures at the annealing
reference/control points for each group

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid              ; Method to determine neighbor list (simple, grid: faster)
nstlist         = 10             ; Frequency to update the neighbor list and long range forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds         ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Equilibrado de la presión: Nter MG517 y hélices en solvente acuoso

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          =                       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 50000                 ; Maximum number of steps to perform
nstlog          = 100                   ; Frequency to write energies to log file
nstxout         = 5000     ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 5000     ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 100                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein non-protein   ; Which energy group(s) to write to output energy file
(.edr)
nstxtcout       = 100                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein non-protein   ; groups to couple separately to temperature bath
tau_t           = 2. 2.                 ; time constant for temperature coupling
ref_t           = 300. 300.  ; reference temperature for coupling (one for each group in
tc_grps)
pcoupl          = Parrinello-Rahman     ; Barostat type (berendsen, Parrinello-Rahman)
tau_p           = 2.                    ; time constant for pressure coupling
ref_p           = 1.                    ; reference pressure for coupling
compressibility = 4.5e-5     ; For water at 1 atm and 300 K the compressibility is 4.5e-5 bar-
1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = none       ; Type of annealing for each temperature group (single, periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10         ; Frequency to update the neighbor list and long range forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Producción: Nter MG517 y hélices en solvente acuoso

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          =                       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 250000                ; Maximum number of steps to perform
nstlog          = 100                   ; Frequency to write energies to log file
nstxout         = 50000; Frequency to write full coordinates to output trajectory file (.trr)
nstvout         = 50000; Frequency to write full velocities to output trajectory file (.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 500                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein non-protein   ; Which energy group(s) to write to output energy file
(.edr)
nstxtcout       = 500                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Energy minimizing specific parameters
emtol           = 1.0                   ; Stop minimization when the maximum force < 1.0
kJ/mol/nm
emstep          = 0.01                  ; initial step-size
nstcgsteep      = 1000                  ; frequency of performing 1 steepest descent step while
doing conjugate gradient energy minimization

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein non-protein   ; groups to couple separately to temperature bath
tau_t           = 2. 2.                 ; time constant for temperature coupling
ref_t           = 300. 300. ; reference temperature for coupling (one for each group in
tc_grps)
pcoupl          = Parrinello-Rahman     ; Barostat type (berendsen, Parrinello-Rahman)
tau_p           = 2.                    ; time constant for pressure coupling
ref_p           = 1.                    ; reference pressure for coupling
compressibility = 4.5e-5                ; For water at 1 atm and 300 K the compressibility is
4.5e-5 bar-1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = none     ; Type of annealing for each temperature group (single, periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid             ; Method to determine neighbor list (simple, grid: faster)
nstlist         = 10        ; Frequency to update the neighbor list and long range forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                ; Calculate the optimal FFT plan for the grid at startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Equilibrado de la temperatura: Hélice inserta en membrana

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES_helix_lipid ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 50000                 ; Maximum number of steps to perform
nstlog          = 500                   ; Frequency to write energies to log file
nstxout         = 0         ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 0         ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0         ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 500                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = Protein_DPC_DPM NA_CL_SOL; Which energy group(s) to write to output energy
file (.edr)
nstxtcout       = 500                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = Protein DPC_DPM NA_CL_SOL; groups to couple separately to temperature bath
tau_t           = 0.1  0.1  0.1     ; time constant for temperature coupling
ref_t           = 323  323  323; reference temperature for coupling (one for each group in
tc_grps)
pcoupl          = no                    ; Barostat type (berendsen, Parrinello-Rahman)

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = none none none        ; Type of annealing for each temperature group (single,
periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds         ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Equilibrado de la presión1: Hélice inserta en membrana

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES_helix_lipid ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 250000                ; Maximum number of steps to perform
nstlog          = 100                   ; Frequency to write energies to log file
nstxout         = 0         ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 0         ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0         ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 100                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = Protein_DPC_DPM NA_CL_SOL  ; Which energy group(s) to write to output energy
file (.edr)
nstxtcout       = 100                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein non-protein   ; groups to couple separately to temperature bath
tau_t           = 0.5  0.5  0.5         ; time constant for temperature coupling
ref_t           = 323. 323. 323.        ; reference temperature for coupling (one for each
group in tc_grps)
pcoupl          = Parrinello-Rahman     ; Barostat type (berendsen, Parrinello-Rahman)
tau_p           = 5.                    ; time constant for pressure coupling
ref_p           = 1.0  1.0                  ; reference pressure for coupling
compressibility = 4.5e-5     ; For water at 1 atm and 300 K the compressibility is 4.5e-5 bar-
1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = none       ; Type of annealing for each temperature group (single, periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10         ; Frequency to update the neighbor list and long range forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Equilibrado de la presión2: Hélice inserta en membrana

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 250000                ; Maximum number of steps to perform
nstlog          = 100                   ; Frequency to write energies to log file
nstxout         = 0        ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 0        ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 100                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = Protein_DPC_DPM NA_CL_SOL  ; Which energy group(s) to write to output energy
file (.edr)
nstxtcout       = 100                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein non-protein   ; groups to couple separately to temperature bath
tau_t           = 0.5  0.5  0.5         ; time constant for temperature coupling
ref_t           = 323. 323. 323.        ; reference temperature for coupling (one for each
group in tc_grps)
pcoupl          = Parrinello-Rahman     ; Barostat type (berendsen, Parrinello-Rahman)
tau_p           = 5.                    ; time constant for pressure coupling
ref_p           = 1.0  1.0                   ; reference pressure for coupling
compressibility = 4.5e-5     ; For water at 1 atm and 300 K the compressibility is 4.5e-5 bar-
1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = none       ; Type of annealing for each temperature group (single, periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10         ; Frequency to update the neighbor list and long range forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Producción: Hélice inserta en membrana

```
; The following line passes specific messages to the program
define          =                       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 500000                ; Maximum number of steps to perform
nstlog          = 1000                  ; Frequency to write energies to log file
nstxout         = 0     ; Frequency to write full coordinates to output trajectory file (.trr)
nstvout         = 0     ; Frequency to write full velocities to output trajectory file (.trr)
nstfout         = 0       ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 1000                  ; Frequency to write energies to output energy file
(.edr)
energygrps      = Protein_DPC_DPM NA_CL_SOL  ; Which energy group(s) to write to output energy
file (.edr)
nstxtcout       = 1000                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Energy minimizing specific parameters
emtol           = 1.0                   ; Stop minimization when the maximum force < 1.0
kJ/mol/nm
emstep          = 0.01                  ; initial step-size
nstcgsteep      = 1000                  ; frequency of performing 1 steepest descent step while
doing conjugate gradient energy minimization

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein non-protein   ; groups to couple separately to temperature bath
tau_t           = 0.5  0.5  0.5         ; time constant for temperature coupling
ref_t           = 323. 323. 323.        ; reference temperature for coupling (one for each
group in tc_grps)
pcoupl          = Parrinello-Rahman     ; Barostat type (berendsen, Parrinello-Rahman)
tau_p           = 2.                    ; time constant for pressure coupling
ref_p           = 1. 1.                 ; reference pressure for coupling
compressibility = 4.5e-5                ; For water at 1 atm and 300 K the compressibility is
4.5e-5 bar-1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = none     ; Type of annealing for each temperature group (single, periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10        ; Frequency to update the neighbor list and long range forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Energía de minimización. Modelos 4DDZ

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DFLEXIBLE            ; This defines the solvent molecules as flexible

; Parameters describing what to do, when to stop and what to save
integrator      = cg                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 10000                 ; Maximum number of steps to perform
nstlog          = 1                     ; Frequency to write energies to log file
nstxout         = 0       ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 0       ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0       ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 100                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = System  ; Which energy group(s) to write to output energy file (.edr)
nstxtcout       = 1                     ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Energy minimizing specific parameters
emtol           = 1.0                   ; Stop minimization when the maximum force < 1.0
kJ/mol/nm
emstep          = 0.01                  ; initial step-size
nstcgsteep      = 1000                  ; frequency of performing 1 steepest descent step while
doing conjugate gradient energy minimization


; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions
```

Equilibrado de las aguas. Modelo 4DDZ

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES -DPOSRESUGC ; This defines the macromolecule as rigid

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 100000                ; Maximum number of steps to perform
nstlog          = 500                   ; Frequency to write energies to log file
nstxout         = 5000     ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 5000     ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 500                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein non-protein  ; Which energy group(s) to write to output energy file
(.edr)
nstxtcout       = 500                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Energy minimizing specific parameters
emtol           = 1.0                   ; Stop minimization when the maximum force < 1.0
kJ/mol/nm
emstep          = 0.01                  ; initial step-size
nstcgsteep      = 1000                  ; frequency of performing 1 steepest descent step while
doing conjugate gradient energy minimization

; Molecular dynamics specific parameters
dt              = 0.001                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = no                    ; Thermostat type (berendsen, nose-hoover)
pcoupl          = no                    ; Barostat type (berendsen, Parrinello-Rahman)

; Molecular dynamics specific parameters (startup)
gen_vel         = yes                   ; generate velocities at startup
gen_temp        = 300.                  ; Maxwell distribution temperature for initial
velocities
gen_seed        = 173529                ; used to initialize random generator for random
velocitie
annealing       = no                    ; Type of annealing for each temperature group (single,
periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = none                  ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
```

Equilibrado de la temperatura. Modelo 4DDZ

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          =                       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 750000                ; Maximum number of steps to perform
nstlog          = 500                   ; Frequency to write energies to log file
nstxout         = 5000 ; Frequency to write full coordinates to output trajectory file (.trr)
nstvout         = 5000 ; Frequency to write full velocities to output trajectory file (.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 500                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein  non-protein  ; Which energy group(s) to write to output energy file
(.edr)
nstxtcout       = 500      ; Frequency to write compressed coordinates to output trajectory
file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein non-protein   ; groups to couple separately to temperature bath
tau_t           = 2. 2.                 ; time constant for temperature coupling
ref_t           = 300. 300.             ; reference temperature for coupling (one for each
group in tc_grps)
pcoupl          = no                    ; Barostat type (berendsen, Parrinello-Rahman)
4.5e-5 bar-1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = single none   ; Type of annealing for each temperature group (single,
periodic)
annealing_npoints = 10 0                ; A list with the number of annealing reference/control
points used for each temperature group
annealing_time  = 0 20 40 60 80 100 120 140 160 180      ; List of times at the annealing
reference/control points for each group
annealing_temp  = 175 200 200 225 225 250 250 275 275 300   ; List of temperatures at the
annealing reference/control points for each group

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10             ; Frequency to update the neighbor list and long range forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Equilibrado de la presión. Modelo 4DDZ

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          =                       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 50000                 ; Maximum number of steps to perform
nstlog          = 100                   ; Frequency to write energies to log file
nstxout         = 5000     ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 5000     ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 100                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein non-protein             ; Which energy group(s) to write to output
energy file (.edr)
nstxtcout       = 100                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein non-protein   ; groups to couple separately to temperature bath
tau_t           = 2. 2.                 ; time constant for temperature coupling
ref_t           = 300. 300.             ; reference temperature for coupling (one for each
group in tc_grps)
pcoupl          = Parrinello-Rahman     ; Barostat type (berendsen, Parrinello-Rahman)
tau_p           = 2. 2.                 ; time constant for pressure coupling
ref_p           = 1. 1.                 ; reference pressure for coupling
compressibility = 4.5e-5                ; For water at 1 atm and 300 K the compressibility is
4.5e-5 bar-1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Energía de minimización. MD modelo 4Y6N. *Conjugated gradients*

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DFLEXIBLE            ; This defines the solvent molecules as flexible

; Parameters describing what to do, when to stop and what to save
integrator      = cg                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 5000                  ; Maximum number of steps to perform
nstlog          = 1                     ; Frequency to write energies to log file
nstxout         = 1        ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 0        ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 1                     ; Frequency to write energies to output energy file
(.edr)
energygrps      = System   ; Which energy group(s) to write to output energy file (.edr)
nstxtcout       = 1  ; Frequency to write compressed coordinates to output trajectory file
(.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Energy minimizing specific parameters
emtol           = 1.0                   ; Stop minimization when the maximum force < 1.0
kJ/mol/nm
emstep          = 0.01                  ; initial step-size
nstcgsteep      = 1000                  ; frequency of performing 1 steepest descent step while
doing conjugate gradient energy minimization

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions
```

Energía de minimización. MD modelo 4Y6N. *Steepest descent*

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DFLEXIBLE            ; This defines the solvent molecules as flexible

; Parameters describing what to do, when to stop and what to save
integrator      = steep                 ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 5000                  ; Maximum number of steps to perform
nstlog          = 1                     ; Frequency to write energies to log file
nstxout         = 1        ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 0        ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 1                     ; Frequency to write energies to output energy file
(.edr)
energygrps      = System   ; Which energy group(s) to write to output energy file (.edr)
nstxtcout       = 1  ; Frequency to write compressed coordinates to output trajectory file
(.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Energy minimizing specific parameters
emtol           = 1.0                   ; Stop minimization when the maximum force < 1.0
kJ/mol/nm
emstep          = 0.01                  ; initial step-size
nstcgsteep      = 1000                  ; frequency of performing 1 steepest descent step while
doing conjugate gradient energy minimization

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions
```

Equilibrado de las aguas. MD modelo 4Y6N. Proteína y ligandos fijos

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES              ; This defines the macromolecule as rigid

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 100000                ; Maximum number of steps to perform
nstlog          = 1000                  ; Frequency to write energies to log file
nstxout         = 50000    ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 50000    ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 1000                  ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein Ligands_chain_A Ligands_chain_B Water_and_ions  ; Which energy
group(s) to write to output energy file (.edr)
nstxtcout       = 1000                  ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output
trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.001                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = no                    ; Thermostat type (berendsen, nose-hoover)
pcoupl          = no                    ; Barostat type (berendsen, Parrinello-Rahman)

; Molecular dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = no                    ; Type of annealing for each temperature group (single,
periodic)

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = none                  ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Equilibrado de las aguas y proteína. MD modelo 4Y6N. Proteína libre, ligandos fijados

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES_LIGAND       ; This defines the macromolecule as rigid

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 100000                ; Maximum number of steps to perform
nstlog          = 1000                  ; Frequency to write energies to log file
nstxout         = 50000    ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 50000    ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 1000                  ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein Ligands_chain_A Ligands_chain_B Water_and_ions  ; Which energy
group(s) to write to output energy file (.edr)
nstxtcout       = 1000                  ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.001                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = no                    ; Thermostat type (berendsen, nose-hoover)
pcoupl          = no                    ; Barostat type (berendsen, Parrinello-Rahman)

; Molecular dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = none                  ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Equilibrado de la temperatura. MD modelo 4Y6N.

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES_LIGAND       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 1000000               ; Maximum number of steps to perform
nstlog          = 500                   ; Frequency to write energies to log file
nstxout         = 50000    ; Frequency to write full coordinates to output trajectory file
(.trr)
nstvout         = 50000    ; Frequency to write full velocities to output trajectory file
(.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 500                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein Ligands_chain_A Ligands_chain_B Water_and_ions ; Which energy
group(s) to write to output energy file (.edr)
nstxtcout       = 500                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein Ligands_chain_A Ligands_chain_B Water_and_ions  ; groups to couple
separately to temperature bath
tau_t           = 2. 2. 2. 2.                   ; time constant for temperature coupling
ref_t           = 300. 300. 300. 300.           ; reference temperature for coupling (one for
each group in tc_grps)
pcoupl          = no                    ; Barostat type (berendsen, Parrinello-Rahman)

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
gen_temp        = 300.                  ; Maxwell distribution temperature for initial
velocities
gen_seed        = 173529                ; used to initialize random generator for random
velocitie
annealing       = single single single single          ; Type of annealing for each
temperature group (single, periodic)
annealing_npoints = 42 42 42 42               ; A list with the number of annealing
reference/control points used for each temperature group
annealing_time  = 0 25 75 100 150 175 225 250 300 325 375 400 450 475 525 550 600 625 675 700
750 775 825 850 900 925 975 1000 1050 1075 1125 1150 1200 1225 1275 1300 1350 1375 1425 1450
1500 1525 0 25 75 100 150 175 225 250 300 325 375 400 450 475 525 550 600 625 675 700 750 775
825 850 900 925 975 1000 1050 1075 1125 1150 1200 1225 1275 1300 1350 1375 1425 1450 1500 1525
0 25 75 100 150 175 225 250 300 325 375 400 450 475 525 550 600 625 675 700 750 775 825 850
900 925 975 1000 1050 1075 1125 1150 1200 1225 1275 1300 1350 1375 1425 1450 1500 1525 0 25 75
100 150 175 225 250 300 325 375 400 450 475 525 550 600 625 675 700 750 775 825 850 900 925
975 1000 1050 1075 1125 1150 1200 1225 1275 1300 1350 1375 1425 1450 1500 1525  ; List of
times at the annealing reference/control points for each group
annealing_temp  = 6.9 6.9 15 15 30 30 45 45 60 60 75 75 90 90 105 105 120 120 135 135 150 150
165 165 180 180 195 195 210 210 225 225 240 240 255 255 270 270 285 285 300 300 6.7 6.7 15 15
30 30 45 45 60 60 75 75 90 90 105 105 120 120 135 135 150 150 165 165 180 180 195 195 210 210
225 225 240 240 255 255 270 270 285 285 300 300 6.9 6.9 15 15 30 30 45 45 60 60 75 75 90 90
105 105 120 120 135 135 150 150 165 165 180 180 195 195 210 210 225 225 240 240 255 255 270
270 285 285 300 300 10.5 10.5 15 15 30 30 45 45 60 60 75 75 90 90 105 105 120 120 135 135 150
150 165 165 180 180 195 195 210 210 225 225 240 240 255 255 270 270 285 285 300 300  ; List of
temperatures at the annealing reference/control points for each group

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
```

```
ns_type         = grid                ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                  ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                 ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                 ; long range electrostatic cut-off
coulombtype     = PME                 ; Treatment of long range electrostatic interactions
pme_order       = 4                   ; Interpolation order for PME
fourierspacing  = 0.12                ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                 ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                 ; long range Van der Waals cut-off
vdwtype         = cut-off             ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds           ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs          ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                   ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Equilibrado de la presión. MD modelo 4Y6N

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES_LIGAND       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 1000000               ; Maximum number of steps to perform
nstlog          = 500                   ; Frequency to write energies to log file
nstxout         = 50000 ; Frequency to write full coordinates to output trajectory file (.trr)
nstvout         = 50000 ; Frequency to write full velocities to output trajectory file (.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 500                   ; Frequency to write energies to output energy file
(.edr)
energygrps      = protein Ligands_chain_A Ligands_chain_B Water_and_ions ; Which energy
group(s) to write to output energy file (.edr)
nstxtcout       = 500                   ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)

; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = protein Ligands_chain_A Ligands_chain_B Water_and_ions  ; groups to couple
separately to temperature bath
tau_t           = 2. 2. 2. 2.                   ; time constant for temperature coupling
ref_t           = 300. 300. 300. 300.           ; reference temperature for coupling (one for
each group in tc_grps)
pcoupl          = Parrinello-Rahman             ; Barostat type (berendsen, Parrinello-Rahman)
tau_p           = 2.                    ; time constant for pressure coupling
ref_p           = 1.                    ; reference pressure for coupling
compressibility = 4.5e-5                ; For water at 1 atm and 300 K the compressibility is
4.5e-5 bar-1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = none none none none   ; Type of annealing for each temperature group (single,
periodic)
annealing_npoints = 0 0 0 0             ; A list with the number of annealing reference/control
points used for each temperature group

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Producción. MD Modelo 4Y6N

```
; The following lines tell the program the standard locations where to find certain files
cpp             = /lib/cpp              ; Preprocessor

; The following line passes specific messages to the program
define          = -DPOSRES_LIGAND       ;

; Parameters describing what to do, when to stop and what to save
integrator      = md                    ; Algorithm (steep = steepest descent minimization; cg
= conjugate gradients; md = newtonian molecular dynamics)
init_step       = 0                     ; The starting step
nsteps          = 10000000              ; Maximum number of steps to perform
nstlog          = 5000                  ; Frequency to write energies to log file
nstxout         = 50000 ; Frequency to write full coordinates to output trajectory file (.trr)
nstvout         = 50000 ; Frequency to write full velocities to output trajectory file (.trr)
nstfout         = 0        ; Frequency to write forces to output trajectory file (.trr)
nstenergy       = 5000                  ; Frequency to write energies to output energy file
(.edr)
energygrps      = Protein Ligands_chain_A Ligands_chain_B Water_and_ions ; Which energy
group(s) to write to output energy file (.edr)
nstxtcout       = 5000                  ; Frequency to write compressed coordinates to output
trajectory file (.xtc)
xtc_grps        = System                ; Which compressed coordinate group(s) to write to
output trajectory file (.xtc)


; Molecular dynamics specific parameters
dt              = 0.002                 ; integration time step of the simulation. (0.001 ps or
0.002 ps if bonds are constrained)
comm_mode       = linear                ; remove center of mass motion (linear = translation;
angular = translation and rotation)
nstcomm         = 10                    ; frequency for center of mass motion removal
tcoupl          = nose-hoover           ; Thermostat type (berendsen, nose-hoover)
tc_grps         = Protein Ligands_chain_A Ligands_chain_B Water_and_ions  ; groups to couple
separately to temperature bath
tau_t           = 2. 2. 2. 2.                   ; time constant for temperature coupling
ref_t           = 300. 300. 300. 300.           ; reference temperature for coupling (one for
each group in tc_grps)
pcoupl          = Parrinello-Rahman             ; Barostat type (berendsen, Parrinello-Rahman)
tau_p           = 2.          ; time constant for pressure coupling
ref_p           = 1.          ; reference pressure for coupling
compressibility = 4.5e-5                ; For water at 1 atm and 300 K the compressibility is
4.5e-5 bar-1

; Molecylar dynamics specific parameters (startup)
gen_vel         = no                    ; generate velocities at startup
annealing       = none none none none   ; Type of annealing for each temperature group (single,
periodic)
annealing_npoints = 0 0 0 0             ; A list with the number of annealing reference/control

; Parameters describing how to find the non-bonding neighbors of each atom and how to
calculate the interactions
pbc             = xyz                   ; periodic boundary conditions (xyz = in all
directions; no = ignore the box: must change ns_type to simple)
ns_type         = grid                  ; Method to determine neighbor list (simple, grid:
faster)
nstlist         = 10                    ; Frequency to update the neighbor list and long range
forces
rlist           = 0.8                   ; Cut-off for making neighbor list (short range forces)
rcoulomb        = 0.8                   ; long range electrostatic cut-off
coulombtype     = PME                   ; Treatment of long range electrostatic interactions
pme_order       = 4                     ; Interpolation order for PME
fourierspacing  = 0.12                  ; maximum grid spacing for the FFT grid when using PME
optimize_fft    = yes                   ; Calculate the optimal FFT plan for the grid at
startup
rvdw            = 0.8                   ; long range Van der Waals cut-off
vdwtype         = cut-off               ; Treatment of long range van der waals interactions

; System constraints
constraints     = all-bonds             ; constrained bonds (hbonds, all-bonds, h-angles, all-
angles)
constraint_algorithm = lincs            ; algorithm used to constraint bonds (lincs, shake)
lincs_iter      = 1                     ; Number of iterations to correct for rotational
lengthening in LINCS (2 conserves better the energy)
```

Hélice anfipática inserta en membrana

Input para el *plugin* Plumed, en la simulación metadinámica de la hélice en membrana.

```
HILLS RESTART HEIGHT 4.0  W_STRIDE 20000

DISTANCE LIST <g1> <g2>    DIR Z SIGMA 0.025 NOPBC
ALIGN_ATOMS LIST <g1>

g1->
5 18 29 41 46 55 73 91 100 121 129 141 158 167 180 198 215 228 242 250 259 269 277
g1<-

g2->
296 310 338 346 360 388 396 410 438 446 460 488 496 510 538 546 560 588 596
610 638 646 660 688 696 710 738 746 760 788 796 810 838 846 860 888 896 910
938 946 960 988 996 1010 1038 1046 1060 1088 1096 1110 1138 1146 1160 1188
1196 1210 1238 1246 1260 1288 1296 1310 1338 1346 1360 1388 1396 1410 1438
1446 1460 1488 1496 1510 1538 1546 1560 1588 1596 1610 1638 1646 1660 1688
1696 1710 1738 1746 1760 1788 1796 1810 1838 1846 1860 1888 1896 1910 1938
1946 1960 1988 1996 2010 2038 2046 2060 2088 2096 2110 2138 2146 2160 2188
2196 2210 2238 2246 2260 2288 2296 2310 2338 2346 2360 2388 2396 2410 2438
2446 2460 2488 2496 2510 2538 2546 2560 2588 2596 2610 2638 2646 2660 2688
2696 2710 2738 2746 2760 2788 2796 2810 2838 2846 2860 2888 2896 2910 2938
2946 2960 2988 2996 3010 3038 3046 3060 3088 3096 3110 3138 3146 3160 3188
3196 3210 3238 3246 3260 3288 3296 3310 3338 3346 3360 3388 3396 3410 3438
3446 3460 3488 3491 3505 3533 3536 3550 3578 3581 3595 3623 3626 3640 3668
3671 3685 3713 3716 3730 3758 3761 3775 3803 3806 3820 3848 3851 3865 3893
3896 3910 3938 3941 3955 3983 3986 4000 4028 4031 4045 4073 4076 4090 4118
4121 4135 4163 4166 4180 4208 4211 4225 4253 4256 4270 4298 4301 4315 4343
4346 4360 4388 4391 4405 4433 4436 4450 4478 4481 4495 4523 4526 4540 4568
4571 4585 4613 4616 4630 4658 4661 4675 4703 4706 4720 4748 4751 4765 4793
4796 4810 4838 4841 4855 4883 4886 4900 4928 4931 4945 4973 4976 4990 5018
5021 5035 5063 5066 5080 5108 5111 5125 5153 5156 5170 5198 5201 5215 5243
5246 5260 5288 5291 5305 5333 5336 5350 5378 5381 5395 5423 5426 5440 5468
5471 5485 5513 5516 5530 5558 5561 5575 5603 5606 5620 5648 5651 5665 5693
5696 5710 5738 5741 5755 5783 5786 5800 5828 5831 5845 5873 5876 5890 5918
5921 5935 5963 5966 5980 6008 6011 6025 6053 6056 6070 6098 6101 6115 6143
6146 6160 6188 6191 6205 6233 6236 6250 6278
g2<-

COORD LIST <g3> <g4>  NN 5 MM 6 D_0 0.4 R_0 0.04 SIGMA 4

g3->
1 10 145 148 151 172 220 282
g3<-

g4->
3490 3492 3493 3494 3497 3499 3516 3518 3535 3537 3538 3539 3542 3544 3561 3563
3580 3582 3583 3584 3587 3589 3606 3608 3625 3627 3628 3629 3632 3634 3651 3653
3670 3672 3673 3674 3677 3679 3696 3698 3715 3717 3718 3719 3722 3724 3741 3743
3760 3762 3763 3764 3767 3769 3786 3788 3805 3807 3808 3809 3812 3814 3831 3833
3850 3852 3853 3854 3857 3859 3876 3878 3895 3897 3898 3899 3902 3904 3921 3923
3940 3942 3943 3944 3947 3949 3966 3968 3985 3987 3988 3989 3992 3994 4011 4013
4030 4032 4033 4034 4037 4039 4056 4058 4075 4077 4078 4079 4082 4084 4101 4103
4120 4122 4123 4124 4127 4129 4146 4148 4165 4167 4168 4169 4172 4174 4191 4193
4210 4212 4213 4214 4217 4219 4236 4238 4255 4257 4258 4259 4262 4264 4281 4283
4300 4302 4303 4304 4307 4309 4326 4328 4345 4347 4348 4349 4352 4354 4371 4373
4390 4392 4393 4394 4397 4399 4416 4418 4435 4437 4438 4439 4442 4444 4461 4463
4480 4482 4483 4484 4487 4489 4506 4508 4525 4527 4528 4529 4532 4534 4551 4553
4570 4572 4573 4574 4577 4579 4596 4598 4615 4617 4618 4619 4622 4624 4641 4643
4660 4662 4663 4664 4667 4669 4686 4688 4705 4707 4708 4709 4712 4714 4731 4733
4750 4752 4753 4754 4757 4759 4776 4778 4795 4797 4798 4799 4802 4804 4821 4823
4840 4842 4843 4844 4847 4849 4866 4868 4885 4887 4888 4889 4892 4894 4911 4913
4930 4932 4933 4934 4937 4939 4956 4958 4975 4977 4978 4979 4982 4984 5001 5003
5020 5022 5023 5024 5027 5029 5046 5048 5065 5067 5068 5069 5072 5074 5091 5093
5110 5112 5113 5114 5117 5119 5136 5138 5155 5157 5158 5159 5162 5164 5181 5183
5200 5202 5203 5204 5207 5209 5226 5228 5245 5247 5248 5249 5252 5254 5271 5273
5290 5292 5293 5294 5297 5299 5316 5318 5335 5337 5338 5339 5342 5344 5361 5363
5380 5382 5383 5384 5387 5389 5406 5408 5425 5427 5428 5429 5432 5434 5451 5453
5470 5472 5473 5474 5477 5479 5496 5498 5515 5517 5518 5519 5522 5524 5541 5543
5560 5562 5563 5564 5567 5569 5586 5588 5605 5607 5608 5609 5612 5614 5631 5633
5650 5652 5653 5654 5657 5659 5676 5678 5695 5697 5698 5699 5702 5704 5721 5723
5740 5742 5743 5744 5747 5749 5766 5768 5785 5787 5788 5789 5792 5794 5811 5813
5830 5832 5833 5834 5837 5839 5856 5858 5875 5877 5878 5879 5882 5884 5901 5903
```

```
5920 5922 5923 5924 5927 5929 5946 5948 5965 5967 5968 5969 5972 5974 5991 5993
6010 6012 6013 6014 6017 6019 6036 6038 6055 6057 6058 6059 6062 6064 6081 6083
6100 6102 6103 6104 6107 6109 6126 6128 6145 6147 6148 6149 6152 6154 6171 6173
6190 6192 6193 6194 6197 6199 6216 6218 6235 6237 6238 6239 6242 6244 6261 6263
g4<-

ALPHABETA NDIH 38
37 39 41 42 -1.1
42 44 46 51 -1.1
51 53 55 69 -1.1
69 71 73 87 -1.1
87 89 91 96 -1.1
96 98 100 117 -1.1
117 119 121 125 -1.1
125 127 129 137 -1.1
137 139 141 154 -1.1
154 156 158 163 -1.1
163 165 167 176 -1.1
176 178 180 194 -1.1
194 196 198 211 -1.1
211 213 215 224 -1.1
224 226 228 238 -1.1
238 240 242 246 -1.1
246 248 250 255 -1.1
255 257 259 265 -1.1
265 267 269 273 -1.1
27 29 37 39 -0.7
39 41 42 44 -0.7
44 46 51 53 -0.7
53 55 69 71 -0.7
71 73 87 89 -0.7
89 91 96 98 -0.7
98 100 117 119 -0.7
119 121 125 127 -0.7
127 129 137 139 -0.7
139 141 154 156 -0.7
156 158 163 165 -0.7
165 167 176 178 -0.7
178 180 194 196 -0.7
196 198 211 213 -0.7
213 215 224 226 -0.7
226 228 238 240 -0.7
240 242 246 248 -0.7
248 250 255 257 -0.7
257 259 265 267 -0.7

#INTERVAL CV 1 LOWER_LIMIT 1.5 UPPER_LIMIT 6.0
LWALL CV 1 LIMIT 1.5 KAPPA 1000000.0
UWALL CV 1 LIMIT 6.0 KAPPA 1000000.0

INTERVAL CV 2 LOWER_LIMIT 30 UPPER_LIMIT 190
UWALL CV 2 LIMIT 190 KAPPA 1000000.0
LWALL CV 2 LIMIT 30 KAPPA 1000000.0

LWALL CV 3 LIMIT 36 KAPPA 1000.0

NOHILLS CV 3

PRINT W_STRIDE 200

ENDMETA
```

Hélice anfipática no-apolar inserta en membrana

```
HILLS RESTART HEIGHT 4.0  W_STRIDE 10000

DISTANCE LIST <g1> <g2>    DIR Z SIGMA 0.05 NOPBC
ALIGN_ATOMS LIST <g1>

g1->
4 17 28 40 45 54 62 74 80 88 96 108 125 134 147 165 182 195 209 217 226 236 244
g1<-

g2->
264 314 364 414 464 514 564 614 664 714 764 814 864 914 964 1014 1064 1114
1164 1214 1264 1314 1364 1414 1464 1514 1564 1614 1664 1714 1764 1814 1864
1914 1964 2014 2064 2114 2164 2214 2264 2314 2364 2414 2464 2514 2564 2614
2664 2714 2764 2814 2864 2914 2964 3014 3064 3114 3164 3214 3264 3314 3364
3414 3459 3504 3549 3594 3639 3684 3729 3774 3819 3864 3909 3954 3999 4044
4089 4134 4179 4224 4269 4314 4359 4404 4449 4494 4539 4584 4629 4674 4719
4764 4809 4854 4899 4944 4989 5034 5079 5124 5169 5214 5259 5304 5349 5394
5439 5484 5529 5574 5619 5664 5709 5754 5799 5844 5889 5934 5979 6024 6069
6114 6159 6204
g2<-

COORD LIST <g3> <g4>  NN 5 MM 6 D_0 0.4 R_0 0.04 SIGMA 2

g3->
1 9 112 115 118 139 187 249
g3<-

g4->
3458 3460 3461 3462 3465 3467 3484 3486 3503 3505 3506 3507 3510 3512 3529
3531 3548 3550 3551 3552 3555 3557 3574 3576 3593 3595 3596 3597 3600 3602
3619 3621 3638 3640 3641 3642 3645 3647 3664 3666 3683 3685 3686 3687 3690
3692 3709 3711 3728 3730 3731 3732 3735 3737 3754 3756 3773 3775 3776 3777
3780 3782 3799 3801 3818 3820 3821 3822 3825 3827 3844 3846 3863 3865 3866
3867 3870 3872 3889 3891 3908 3910 3911 3912 3915 3917 3934 3936 3953 3955
3956 3957 3960 3962 3979 3981 3998 4000 4001 4002 4005 4007 4024 4026 4043
4045 4046 4047 4050 4052 4069 4071 4088 4090 4091 4092 4095 4097 4114 4116
4133 4135 4136 4137 4140 4142 4159 4161 4178 4180 4181 4182 4185 4187 4204
4206 4223 4225 4226 4227 4230 4232 4249 4251 4268 4270 4271 4272 4275 4277
4294 4296 4313 4315 4316 4317 4320 4322 4339 4341 4358 4360 4361 4362 4365
4367 4384 4386 4403 4405 4406 4407 4410 4412 4429 4431 4448 4450 4451 4452
4455 4457 4474 4476 4493 4495 4496 4497 4500 4502 4519 4521 4538 4540 4541
4542 4545 4547 4564 4566 4583 4585 4586 4587 4590 4592 4609 4611 4628 4630
4631 4632 4635 4637 4654 4656 4673 4675 4676 4677 4680 4682 4699 4701 4718
4720 4721 4722 4725 4727 4744 4746 4763 4765 4766 4767 4770 4772 4789 4791
4808 4810 4811 4812 4815 4817 4834 4836 4853 4855 4856 4857 4860 4862 4879
4881 4898 4900 4901 4902 4905 4907 4924 4926 4943 4945 4946 4947 4950 4952
4969 4971 4988 4990 4991 4992 4995 4997 5014 5016 5033 5035 5036 5037 5040
5042 5059 5061 5078 5080 5081 5082 5085 5087 5104 5106 5123 5125 5126 5127
5130 5132 5149 5151 5168 5170 5171 5172 5175 5177 5194 5196 5213 5215 5216
5217 5220 5222 5239 5241 5258 5260 5261 5262 5265 5267 5284 5286 5303 5305
5306 5307 5310 5312 5329 5331 5348 5350 5351 5352 5355 5357 5374 5376 5393
5395 5396 5397 5400 5402 5419 5421 5438 5440 5441 5442 5445 5447 5464 5466
5483 5485 5486 5487 5490 5492 5509 5511 5528 5530 5531 5532 5535 5537 5554
5556 5573 5575 5576 5577 5580 5582 5599 5601 5618 5620 5621 5622 5625 5627
5644 5646 5663 5665 5666 5667 5670 5672 5689 5691 5708 5710 5711 5712 5715
5717 5734 5736 5753 5755 5756 5757 5760 5762 5779 5781 5798 5800 5801 5802
5805 5807 5824 5826 5843 5845 5846 5847 5850 5852 5869 5871 5888 5890 5891
5892 5895 5897 5914 5916 5933 5935 5936 5937 5940 5942 5959 5961 5978 5980
5981 5982 5985 5987 6004 6006 6023 6025 6026 6027 6030 6032 6049 6051 6068
6070 6071 6072 6075 6077 6094 6096 6113 6115 6116 6117 6120 6122 6139 6141
6158 6160 6161 6162 6165 6167 6184 6186 6203 6205 6206 6207 6210 6212 6229
6231
g4<-

ALPHABETA NDIH 38
24 26 28 36 -1.1
36 38 40 41 -1.1
41 43 45 50 -1.1
50 52 54 58 -1.1
58 60 62 70 -1.1
70 72 74 76 -1.1
76 78 80 84 -1.1
```

```
84 86 88 92 -1.1
92 94 96 104 -1.1
104 106 108 121 -1.1
121 123 125 130 -1.1
130 132 134 143 -1.1
143 145 147 161 -1.1
161 163 165 178 -1.1
178 180 182 191 -1.1
191 193 195 205 -1.1
205 207 209 213 -1.1
213 215 217 222 -1.1
222 224 226 232 -1.1
26 28 36 38 -0.7
38 40 41 43 -0.7
43 45 50 52 -0.7
52 54 58 60 -0.7
60 62 70 72 -0.7
72 74 76 78 -0.7
78 80 84 86 -0.7
86 88 92 94 -0.7
94 96 104 106 -0.7
106 108 121 123 -0.7
123 125 130 132 -0.7
132 134 143 145 -0.7
145 147 161 163 -0.7
163 165 178 180 -0.7
180 182 191 193 -0.7
193 195 205 207 -0.7
207 209 213 215 -0.7
215 217 222 224 -0.7
224 226 232 234 -0.7

#INTERVAL CV 1 LOWER_LIMIT 1.5 UPPER_LIMIT 6.0
LWALL CV 1 LIMIT 1.5 KAPPA 1000000.0
UWALL CV 1 LIMIT 6.0 KAPPA 1000000.0

INTERVAL CV 2 LOWER_LIMIT 30 UPPER_LIMIT 190
UWALL CV 2 LIMIT 190 KAPPA 1000000.0
LWALL CV 2 LIMIT 30 KAPPA 1000000.0

LWALL CV 3 LIMIT 36 KAPPA 1000.0

NOHILLS CV 3

PRINT W_STRIDE 100

ENDMETA
```

Constrains a los átomos que interaccionan con el Mg en el modelo 4Y6N (GpgS ternario)

Input para el *plugin* Plumed, en la simulación MD del complejo ternario GpgS, monómero B. MD4Y6N.

```
HILLS HEIGHT 0.0 W_STRIDE 10000000

#CVS MonB
#Mg-H2O
DISTANCE LIST 9227 48747 SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9227 48750 SIGMA 0.02
#Mg-Op3
DISTANCE LIST 9227 9315  SIGMA 0.02
#Mg-Op6
DISTANCE LIST 9227 9311  SIGMA 0.02
#Mg-D425_OD2
DISTANCE LIST 9227 6369  SIGMA 0.02
#Mg-H547_ND1
DISTANCE LIST 9227 8226  SIGMA 0.02

#UGC_C1-O3_3PG
DISTANCE LIST 9304 9370  SIGMA 0.02
#Mg-D425_OD1
DISTANCE LIST 9227 6367  SIGMA 0.02
#DistN-O
DISTANCE LIST 6371 8235 SIGMA 0.03
#PSIA257
TORSION LIST 8208 8210 8216 8218 SIGMA 0.1

#UGC_C1-O3_3PG
DISTANCE LIST 9249 9355  SIGMA 0.02
#Mg-D118_OD1
DISTANCE LIST 9228 1755  SIGMA 0.02

NOHILLS CV 1
NOHILLS CV 2
NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8
NOHILLS CV 9
NOHILLS CV 10
NOHILLS CV 11
NOHILLS CV 12

UWALL CV 1 LIMIT 0.198 KAPPA 45000000.0
LWALL CV 1 LIMIT 0.198 KAPPA 45000000.0
UWALL CV 2 LIMIT 0.267 KAPPA 45000000.0
LWALL CV 2 LIMIT 0.267 KAPPA 45000000.0
UWALL CV 3 LIMIT 0.251 KAPPA 45000000.0
LWALL CV 3 LIMIT 0.251 KAPPA 45000000.0
UWALL CV 4 LIMIT 0.213 KAPPA 45000000.0
LWALL CV 4 LIMIT 0.213 KAPPA 45000000.0
UWALL CV 5 LIMIT 0.226 KAPPA 45000000.0
LWALL CV 5 LIMIT 0.226 KAPPA 45000000.0

PRINT W_STRIDE 100

ENDMETA
```

MetaD4DDZApoLA

Input para el *plugin* Plumed, en la simulación metadinámica de MetaD4DDZLA.

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000

#Dist Asp412(O)-Arg535(N) MonB
DISTANCE LIST 6223 8087 SIGMA 0.03

#Diedro PSI Ala 533 MonB
TORSION LIST 8060 8062 8068 8070 SIGMA 0.2

#Diedro His 534
TORSION LIST 8077 8074 8072 8085

#Puentes de hidrogeno del bucle MonB
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 6244 8088 H<-
O-> 6223 8069 O<-

#Dist Asp113(O)-Arg236(N) MonA
DISTANCE LIST 1684 3548 SIGMA 0.03

#Diedro PSI Ala 234 MonA
TORSION LIST 3521 3523 3529 3531 SIGMA 0.2

#Diedro His 235 MonA
TORSION LIST 3538 3535 3533 3546

#Puentes de hidrogeno del bucle MonA
HBONDS LIST <Ha> <Oa> TYPE 0 SIGMA 1
Ha-> 3549 1705 Ha<-
Oa-> 1684 3530 Oa<-

NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8

UWALL CV 1 LIMIT 1 KAPPA 100000000.0

PRINT W_STRIDE 100

ENDMETA
```

MetaD4Y6NApo

Input para el *plugin* Plumed, en la simulación metadinámica de MetaD4Y6NApo1 y MetaD4Y6NApo2.

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000

#Dist Asp118(O)-Arg241(N) MonA
DISTANCE LIST 1758 3622 SIGMA 0.03

#Diedro PSI Ala 239 MonA
TORSION LIST 3595 3597 3603 3605 SIGMA 0.2

#Diedro His 240 MonA
TORSION LIST 3612 3609 3607 3620

#Puentes de hidrogeno del bucle MonA
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 3623 1778 H<-
O-> 1758 3604 O<-

#Dist Asp425(O)-Arg548(N) MonB
DISTANCE LIST 6371 8235 SIGMA 0.03

#Diedro PSI Ala 546 MonB
TORSION LIST 8208 8210 8216 8218 SIGMA 0.2

#Diedro His 534
TORSION LIST 8225 8222 8220 8233

#Puentes de hidrogeno del bucle MonB
HBONDS LIST <Hb> <Ob> TYPE 0 SIGMA 1
Hb-> 6392 8236 Hb<-
Ob-> 6371 8217 Ob<-

NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8

UWALL CV 1 LIMIT 1 KAPPA 100000000.0

PRINT W_STRIDE 100

ENDMETA
```

HILLS MetaD4Y6N FES restringido

Inicio del archivo HILLS para restringir el espacio conformacional del diedro Ala257.

#tiempo	Distancia	DiedroAla257	Altura gauss	Anchura gauss	Energía(kjul)	
0.000	0.000000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.100000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.200000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.300000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.400000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.500000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.600000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.700000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.800000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.900000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.000000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.100000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.200000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.300000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.400000000	0.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.000000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.100000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.200000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.300000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.400000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.500000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.600000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.700000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.800000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	0.900000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.000000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.100000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.200000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.300000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000
0.000	1.400000000	-1.700000000	1.000000000	0.100000000	20.000000000	0.000

Meta4Y6NApo3

Input para el *plugin* Plumed, en la simulación metadinámica por el método de BIAS-Exchange Meta4Y6NApo3. Se utilizan 4 *inpunts,* el presente es para la simulación que no tiene activada ninguna CV, cada simulación activa su CV correspondiente comentando la línea "NOHILLS" de esa CV.

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000
BIASXMD

#CV1 (CG-CB-CB-C)H240 diedro
TORSION LIST 3612 3609 3607 3620 SIGMA 0.1
#CV2 PSI A239 diedro
TORSION LIST 3595 3597 3603 3605 SIGMA 0.2
#CV3 (CG-CB-CB-C)A238 diedro
TORSION LIST 3578 3575 3573 3593 SIGMA 0.1

#CV4 Dist D118(O)-R241(N)
DISTANCE LIST 1758 3622 SIGMA 0.03

#CV5 Mg-H240(Nd1) Dist
DISTANCE LIST 9228 3613 SIGMA 0.02

#Puentes de hidrogeno del bucle
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 1779 3623 H<-
O-> 1758 3604 O<-

#Posibles interacciones de la His
#LA
#H240-R238
DISTANCE LIST 3613 3590 SIGMA 0.02
#LI
#H240-R167
DISTANCE LIST 3613 2502 SIGMA 0.02
#H240-R149
DISTANCE LIST 3613 2270 SIGMA 0.02

#Fijacion de los diedros
#P1, P2, Glc
ALPHABETA NDIH 3
9259 9256 9255 9252 1.75
9256 9255 9252 9251 -2.5
9255 9252 9251 9249 -2.84

NOHILLS CV 1
NOHILLS CV 2
NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8
NOHILLS CV 9
NOHILLS CV 10

UWALL CV 4 LIMIT 1 KAPPA 100000000.0
LWALL CV 10 LIMIT 2.9 KAPPA 100000.0

PRINT W_STRIDE 100

ENDMETA
```

Meta4Y6NApo4

Input para el *plugin* Plumed, en la simulación metadinámica de.

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000

DISTANCE LIST 1758 3622 SIGMA 0.03

TORSION LIST 3612 3609 3607 3620 SIGMA 0.1

#Puentes de hidrogeno del bucle
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 1779 3623 H<-
O-> 1758 3604 O<-

#PSI_A239
TORSION LIST 3595 3597 3603 3605 SIGMA 0.2
#PHI_A239
TORSION LIST 3593 3595 3597 3603 SIGMA 0.1

#Posibles interacciones de la His
#LA
#H240-R238
DISTANCE LIST 3613 3590 SIGMA 0.02
#LI
#H240-R167
DISTANCE LIST 3613 2502 SIGMA 0.02
#H240-R149
DISTANCE LIST 3613 2270 SIGMA 0.02


NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8

UWALL CV 1 LIMIT 0.85 KAPPA 100000000.0

PRINT W_STRIDE 100

ENDMETA
```

MetaD4Y6NUDPGlc·PGA.

Input para el *plugin* Plumed, en la simulación metadinámica de MetaD4Y6NUDPGlc·PGA. (Complejo ternario).

```
DISTANCE LIST 1758 3622 SIGMA 0.03

TORSION LIST 3595 3597 3603 3605 SIGMA 0.2

#H258 diedro
TORSION LIST 3607 3609 3612 3620

#Puentes de hidrogeno del bucle
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 1779 3623 H<-
O-> 1758 3604 O<-

#Var MonA Pocket
#Mg-H240_ND1
DISTANCE LIST 9228 3613  SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9228 48711 SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9228 48714 SIGMA 0.02
#Mg-Op3
DISTANCE LIST 9228 9257  SIGMA 0.02
#Mg-Op6
DISTANCE LIST 9228 9253  SIGMA 0.02
#Mg-D118_OD1
DISTANCE LIST 9228 1755  SIGMA 0.02
#Mg-D118_OD2
DISTANCE LIST 9228 1756  SIGMA 0.02

#Posibles interacciones de la His
#LA
#H240-R238
DISTANCE LIST 3613 3590 SIGMA 0.02
#LI
#H240-R167
DISTANCE LIST 3613 2502 SIGMA 0.02
#H240-R149
DISTANCE LIST 3613 2270 SIGMA 0.02

#Fijacion de los diedros
#P1, P2, Glc
ALPHABETA NDIH 3
9259 9256 9255 9252 1.75
9256 9255 9252 9251 -2.5
9255 9252 9251 9249 -2.84

NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8
NOHILLS CV 9
NOHILLS CV 10
NOHILLS CV 11
NOHILLS CV 12
NOHILLS CV 13
NOHILLS CV 14
NOHILLS CV 15

LWALL CV 15 LIMIT 2.9 KAPPA 100000.0
UWALL CV 1 LIMIT 1 KAPPA 100000000.0

PRINT W_STRIDE 100

ENDMETA
```

MetaD4Y6NUDPGlcFree.

Input para el *plugin* Plumed, en la simulación metadinámica de Meta4Y6N2 (UDPGlc+metal), diedro libre.

HILLS RESTART HEIGHT 1.0 W_STRIDE 10000

```
TORSION LIST 8206 8208 8210 8216 SIGMA 0.2

DISTANCE LIST 6371 8235 SIGMA 0.015

TORSION LIST 8208 8210 8216 8218 SIGMA 0.05

HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 6392 8236 H<-
O-> 6371 8217 O<-

DISTANCE LIST 6374 8220 SIGMA 0.06

#CVS MonB
#Mg-H2O
DISTANCE LIST 9227 48717 SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9227 48720 SIGMA 0.02
#Mg-Op3
DISTANCE LIST 9227 9315  SIGMA 0.02
#Mg-Op6
DISTANCE LIST 9227 9311  SIGMA 0.02
#Mg-D425_OD2
DISTANCE LIST 9227 6369  SIGMA 0.02

NOHILLS CV 1
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8
NOHILLS CV 9
NOHILLS CV 10

UWALL CV 2 LIMIT 1.10 KAPPA 100000000.0
UWALL CV 6 LIMIT 0.198 KAPPA 45000000.0
LWALL CV 6 LIMIT 0.198 KAPPA 45000000.0
UWALL CV 7 LIMIT 0.267 KAPPA 45000000.0
LWALL CV 7 LIMIT 0.267 KAPPA 45000000.0
UWALL CV 8 LIMIT 0.251 KAPPA 45000000.0
LWALL CV 8 LIMIT 0.251 KAPPA 45000000.0
UWALL CV 9 LIMIT 0.213 KAPPA 45000000.0
LWALL CV 9 LIMIT 0.213 KAPPA 45000000.0
UWALL CV 10 LIMIT 0.226 KAPPA 45000000.0
LWALL CV 10 LIMIT 0.226 KAPPA 45000000.0

PRINT W_STRIDE 100

ENDMETA
```

MetaD4Y6NUDPGlcFix.

Input para el *plugin* Plumed, en la simulación metadinámica de MetaD4Y6N3 (UDPGlc+metal), diedro fijo.

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000

DISTANCE LIST 1758 3622 SIGMA 0.03

TORSION LIST 3595 3597 3603 3605 SIGMA 0.2

#H258 diedro
TORSION LIST 3607 3609 3612 3620

#Puentes de hidrogeno del bucle
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 1779 3623 H<-
O-> 1758 3604 O<-

#Var MonA Pocket
#Mg-H240_ND1
DISTANCE LIST 9228 3613  SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9228 48711 SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9228 48714 SIGMA 0.02
#Mg-Op3
DISTANCE LIST 9228 9257  SIGMA 0.02
#Mg-Op6
DISTANCE LIST 9228 9253  SIGMA 0.02
#Mg-D118_OD1
DISTANCE LIST 9228 1755  SIGMA 0.02
#Mg-D118_OD2
DISTANCE LIST 9228 1756  SIGMA 0.02

#Posibles interacciones de la His
#LA
#H240-R238
DISTANCE LIST 3613 3590 SIGMA 0.02
#LI
#H240-R167
DISTANCE LIST 3613 2502 SIGMA 0.02
#H240-R149
DISTANCE LIST 3613 2270 SIGMA 0.02

#Fijacion de los diedros
#P1, P2, Glc
ALPHABETA NDIH 3
9259 9256 9255 9252 1.75
9256 9255 9252 9251 -2.5
9255 9252 9251 9249 -2.84

NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8
NOHILLS CV 9
NOHILLS CV 10
NOHILLS CV 11
NOHILLS CV 12
NOHILLS CV 13
NOHILLS CV 14
NOHILLS CV 15

ENDMETA
```

MetaD4Y6NPGA.

Input para el *plugin* Plumed, en la simulación metadinámica de MetaD4Y6N4 (PGA).

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000

DISTANCE LIST 1758 3622 SIGMA 0.03

TORSION LIST 3595 3597 3603 3605 SIGMA 0.2

#H258 diedro
TORSION LIST 3607 3609 3612 3620

#Puentes de hidrogeno del bucle
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 1779 3623 H<-
O-> 1758 3604 O<-

#Posibles interacciones del PGA
#
#H240(NE2)-PGA(P)
DISTANCE LIST 3616 9346 SIGMA 0.02
#N242(ND2)-PGA(P)
DISTANCE LIST 3655 9346 SIGMA 0.02
#T169(N)-PGA(O1)
DISTANCE LIST 2527 9241 SIGMA 0.02
#T169(N)-PGA(O2)
DISTANCE LIST 2527 9240 SIGMA 0.02

#Fijacion de los diedros
#P-OP1-C3-C2
#O1P-C3-C2-C1
ALPHABETA NDIH 2
9228 9231 9232 9235 -3
9231 9232 9235 9239 1.1

NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8
NOHILLS CV 9

LWALL CV 9 LIMIT 1.85 KAPPA 10000.0
UWALL CV 1 LIMIT 1 KAPPA 100000000.0

PRINT W_STRIDE 100

ENDMETA
```

MetaD4Y6NUDPGlc3.

Input para el *plugin* Plumed, en la simulación metadinámica de MetaD4Y6NUDPGlc3 (UDPGlc+metal).

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000

DISTANCE LIST 1758 3622 SIGMA 0.03

TORSION LIST 3612 3609 3607 3620 SIGMA 0.1

#Puentes de hidrogeno del bucle
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 1779 3623 H<-
O-> 1758 3604 O<-

#PSI_A239
TORSION LIST 3595 3597 3603 3605 SIGMA 0.2
#PHI_A239
TORSION LIST 3593 3595 3597 3603 SIGMA 0.1

#Var MonA Pocket
#Mg-H240_ND1
DISTANCE LIST 9228 3613  SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9228 48711 SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9228 48714 SIGMA 0.02
#Mg-Op3
DISTANCE LIST 9228 9257  SIGMA 0.02
#Mg-Op6
DISTANCE LIST 9228 9253  SIGMA 0.02
#Mg-D118_OD1
DISTANCE LIST 9228 1755  SIGMA 0.02
#Mg-D118_OD2
DISTANCE LIST 9228 1756  SIGMA 0.02

#Posibles interacciones de la His
#LA
#H240-R238
DISTANCE LIST 3613 3590 SIGMA 0.02
#LI
#H240-R167
DISTANCE LIST 3613 2502 SIGMA 0.02
#H240-R149
DISTANCE LIST 3613 2270 SIGMA 0.02


NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8
NOHILLS CV 9
NOHILLS CV 10
NOHILLS CV 11
NOHILLS CV 12
NOHILLS CV 13
NOHILLS CV 14
NOHILLS CV 15

UWALL CV 1 LIMIT 0.85 KAPPA 100000000.0

PRINT W_STRIDE 100

ENDMETA
```

MetaD4Y6NUDPGlc4

Input para el *plugin* Plumed, en la simulación metadinámica MetaD4Y6N8 (UDPGlc+metal). BIAS-Exchange 3 variables

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000
BIASXMD

#CV1 (CG-CB-CB-C)H240 diedro
TORSION LIST 3612 3609 3607 3620 SIGMA 0.1
#CV2 PSI A239 diedro
TORSION LIST 3595 3597 3603 3605 SIGMA 0.2
#CV3 (CG-CB-CB-C)A238 diedro
TORSION LIST 3578 3575 3573 3593 SIGMA 0.1

#CV4 Dist D118(O)-R241(N)
DISTANCE LIST 1758 3622 SIGMA 0.03

#CV5 Mg-H240(Nd1) Dist
DISTANCE LIST 9228 3613 SIGMA 0.02

#Puentes de hidrogeno del bucle
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 1779 3623 H<-
O-> 1758 3604 O<-

#Posibles interacciones de la His
#LA
#H240-R238
DISTANCE LIST 3613 3590 SIGMA 0.02
#LI
#H240-R167
DISTANCE LIST 3613 2502 SIGMA 0.02
#H240-R149
DISTANCE LIST 3613 2270 SIGMA 0.02

NOHILLS CV 1
NOHILLS CV 2
NOHILLS CV 3
NOHILLS CV 4
NOHILLS CV 5
NOHILLS CV 6
NOHILLS CV 7
NOHILLS CV 8
NOHILLS CV 9

UWALL CV 4 LIMIT 1 KAPPA 100000000.0

PRINT W_STRIDE 100

ENDMETA
```

MetaD4Y6NUDPGlc5

Input para el *plugin* Plumed, en la simulación metadinámica MetaD4Y6NUDPGlc4 (UDPGlc+metal). BIAS-Exchange 5 variables.

```
HILLS RESTART HEIGHT 1.0 W_STRIDE 10000
BIASXMD

#CV1 Dist D118(O)-R241(N)
DISTANCE LIST 1758 3622 SIGMA 0.03
#CV2 PSI A239 diedro
TORSION LIST 3595 3597 3603 3605 SIGMA 0.2
#CV3 (CG-CB-CB-C)H240 diedro
TORSION LIST 3612 3609 3607 3620 SIGMA 0.1
#CV4 Mg-H240(Nd1) Dist
DISTANCE LIST 9228 3613 SIGMA 0.02
#CV5 Glc diedro
TORSION LIST 9255 9252 9251 9249 SIGMA 0.2

#Puentes de hidrogeno del bucle
HBONDS LIST <H> <O> TYPE 0 SIGMA 1
H-> 1779 3623 H<-
O-> 1758 3604 O<-

#Var MonA Pocket
#Mg-H2O
DISTANCE LIST 9228 48711 SIGMA 0.02
#Mg-H2O
DISTANCE LIST 9228 48714 SIGMA 0.02
#Mg-Op3
DISTANCE LIST 9228 9257  SIGMA 0.02
#Mg-Op6
DISTANCE LIST 9228 9253  SIGMA 0.02
#Mg-D116_OD1
DISTANCE LIST 9228 1730  SIGMA 0.02
#Mg-D116_OD2
DISTANCE LIST 9228 1731  SIGMA 0.02
#Mg-D118_OD1
DISTANCE LIST 9228 1755  SIGMA 0.02
#Mg-D118_OD2
DISTANCE LIST 9228 1756  SIGMA 0.02

#Posibles interacciones de la His
#LA
#H240-R238
DISTANCE LIST 3613 3590 SIGMA 0.02
#LI
#H240-R167
DISTANCE LIST 3613 2502 SIGMA 0.02
#H240-R149
DISTANCE LIST 3613 2270 SIGMA 0.02

#Fijacion de los diedros
#Ura-penta,penta-P1,penta-P1,penta-P1,P1,P2
ALPHABETA NDIH 6
9276 9274 9273 9263 -2.6
9273 9263 9260 9259 -1.3
9263 9260 9259 9256 -2.4
9260 9259 9256 9255 -1.3
9259 9256 9255 9252 1.9
9256 9255 9252 9251 -2.5

NOHILLS CV 1
NOHILLS CV 2
NOHILLS CV 3
NOHILLS CV 4

ENDMETA
```

Docking DAG sobre modelos de MG517

Parámetros del *grid*. (**** es el nombre de la estructura)

```
npts 126 126 126                    # num.grid points in xyz
gridfld ****.maps.fld               # grid_data_file
spacing 0.375                       # spacing(A)
receptor_types A C HD Mg N NA OA P SA # receptor atom types
ligand_types A C HD N NA OA SA      # ligand atom types
receptor ****.pdbqt                 # macromolecule
gridcenter 33.224 52.216 3.316      # xyz-coordinates or auto
smooth 0.5                          # store minimum energy w/in rad(A)
map ****.A.map                      # atom-specific affinity map
map ****.C.map                      # atom-specific affinity map
map ****.HD.map                     # atom-specific affinity map
map ****.N.map                      # atom-specific affinity map
map ****.NA.map                     # atom-specific affinity map
map ****.OA.map                     # atom-specific affinity map
map ****.SA.map                     # atom-specific affinity map
elecmap ****.e.map                  # electrostatic potential map
dsolvmap ****.d.map              # desolvation potential map
dielectric -0.1465                  # <0, AD4 distance-dep.diel;>0, constant
```

Docking DAG sobre modelos de MG517

Parámetros del *docking*. (**** es el nombre de la estructura).

```
autodock_parameter_version 4.2       # used by autodock to validate parameter set
outlev 1                             # diagnostic output level
intelec                              # calculate internal electrostatics
seed pid time                        # seeds for random generator
ligand_types C HD OA                 # atoms types in ligand
fld ****.maps.fld                    # grid_data_file
map ****.C.map                       # atom-specific affinity map
map ****.HD.map                      # atom-specific affinity map
map ****.OA.map                      # atom-specific affinity map
elecmap ****.e.map                   # electrostatics map
desolvmap ****.d.map                 # desolvation map
move ligand.pdbqt                    # small molecule
about 6.3168 -30.6241 -8.2519        # small molecule center
tran0 random                         # initial coordinates/A or random
axisangle0 random                    # initial orientation
dihe0 random                         # initial dihedrals (relative) or random
tstep 2.0                            # translation step/A
qstep 50.0                           # quaternion step/deg
dstep 50.0                           # torsion step/deg
torsdof 11                           # torsional degrees of freedom
rmstol 2.0                           # cluster_tolerance/A
extnrg 1000.0                        # external grid energy
e0max 0.0 10000                      # max initial energy; max number of retries
ga_pop_size 300                      # number of individuals in population
ga_num_evals 25000000                # maximum number of energy evaluations
ga_num_generations 27000             # maximum number of generations
ga_elitism 1                  # number of top individuals to survive to next
generation
ga_mutation_rate 0.02                # rate of gene mutation
ga_crossover_rate 0.8                # rate of crossover
ga_window_size 10                    #
ga_cauchy_alpha 0.0                  # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0                   # Beta parameter Cauchy distribution
set_ga                               # set the above parameters for GA or LGA
sw_max_its 300                       # iterations of Solis & Wets local search
sw_max_succ 4                        # consecutive successes before changing rho
sw_max_fail 4                        # consecutive failures before changing rho
sw_rho 1.0                           # size of local search space to sample
sw_lb_rho 0.01                       # lower bound on rho
ls_search_freq 0.06           # probability of performing local search on
individual
set_psw1                             # set the above pseudo-Solis & Wets parameters
unbound_model bound                  # state of unbound ligand
ga_run 100                           # do this many hybrid GA-LS runs
analysis                             # perform a ranked cluster analysis
```