Universidad de Concepción
Dirección de Postgrado
Facultad de Ciencias Biológicas
Programa de Magister en Bioquímica y Bioinformática

IDENTIFICACIÓN Y CARACTERIZACIÓN DE GRUPOS FUNCIONALES EN LA FAMILIA DE TRANSPORTADORES DE AZÚCARES

Tesis para optar al grado de Magíster en Bioquímica y Bioinformática

MARTINA ANDREA OPPLIGER MUÑOZ
CONCEPCIÓN – CHILE
2021

Profesor Guía: Alexis Salas Burgos
Departamento de Farmacología, Facultad de Ciencias Biológicas
Universidad de Concepción

Esta tesis ha sido realizada en el en el Laboratorio de Bioinformática del Departamento de Farmacología la Facultad de Ciencias Biológicas, Universidad de Concepción.

Profesor tutor

_____
Dr. Alexis Salas
Facultad de Ciencias Biológicas
Universidad de Concepción

Comisión Evaluadora

_____
Dra. Valeska Ormázabal
Facultad de Ciencias Biológicas
Universidad de Concepción

_____
Dr. Felipe Zúñiga
Facultad de Farmacia
Universidad de Concepción

_____
Dr. Cristian Hernández
Facultad de Ciencias Naturales y Oceanográficas
Universidad de Concepción

Director de Programa

_____
Dra . Amparo Uribe
Facultad de Ciencias Biológicas
Universidad de Concepción

AGRADECIMIENTOS

A mi compañero de vida, Fernando, por acompañarme y apoyarme con paciencia y amor durante todo este proceso. Fuiste, por lejos, mi pilar de apoyo más importante.

A mis amados padres, que se esforzaron por entregarme la mejor educación a nivel académico y, por sobre todo, valórico. Porque desde la niñez me enseñaron a conducirme en la vida con rectitud antes que todo, Gracias. Sin su apoyo jamás hubiera llegado a estas instancias.

Al resto de mi familia y amigos, por confiar siempre en mí y por el constante ánimo entregado.

A mis profesores de colegio que participaron de mis procesos formativos más tempranos, que me enseñaron lo maravillosa y entretenida que podía ser la ciencia y que, en definitiva, me animaron a seguir este camino.

Y, finalmente, a mis profesores tutores y a la directora del programa, por su guía, apoyo e infinita paciencia.

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

2-DG: 2-Deoxiglucosa

3-OMG: 3-Oximetilglucosa

6-DG: 6-Deoxiglucosa

a-MG: a-Metilglucósido

APC: Corrección por producto promedio

ASC: Corrección por suma media

b-NG: nonil b-D-glucopiranósido

CitoB: Citocalasina B

CON: Normalización por entropía conjunta

D-Glc: D-Glucosa

DHA: Ácido dehidroascórbico

EIG12: Epilepsia idiopática generalizada 12

GLUT1DS: Síndrome de deficiencia de GLUT1

GroPCho: Glicerofosfocolina

GroPIns: Glicerofosfoinositol

ICA: Análisis de componentes independientes

KDD: Proceso de descubrimiento de conocimiento en bases de datos

Km: Constante de Michaelis

MFS: Superfamilia de facilitadores principales

MI: Información mutua

MINCON: Normalización por entropía conjunta mínima

MSA: Alineamiento múltiple de secuencias

MTAA: Modelo de transporte de acceso alternante

NaGlcN: N-Acetilglucosamina

NIDDM: Diabetes mellitus no insulino-dependiente

RHUC2: Hipouricemia renal 2

SCA: Análisis de acoplamiento estadístico

SDP: Posición determinante de especificidad

SLC: Sistema de clasificación de portadores de solutos

SUM: Normalización por suma de entropías

RESUMEN

Los azúcares como la glucosa y polioles relacionados son compuestos esenciales en el metabolismo de todos los seres vivos, por lo que su transporte a través de las membranas biológicas es sustancial para la homeostasis y supervivencia. Su transporte es mediado por transportadores pertenecientes a la familia de transportadores de azúcares, TC 2.A.1.1, la cual abarca a diversos homólogos de los genes SLC2 que codifican para los conocidos GLUTs en humanos y animales. Se han desarrollado diversas investigaciones para la caracterización funcional, en su sentido más amplio, de diversos miembros de esta familia presentes tanto en procariontes como en eucariontes, existiendo especial énfasis en los GLUTs debido a su importancia médica y farmacológica. Si bien en los últimos años y gracias a la disponibilidad de estructuras tridimensionales representativas de algunos miembros de la familia se ha podido avanzar aceleradamente en la comprensión de las bases moleculares que sustentan la selectividad de sustratos y mecanismo de transporte en la familia, a la fecha no existe un estudio integrador que permita levantar principios generalizadores de estas bases, considerando homólogos de GLUTs en bacterias, hongos, protozoos y plantas. La información estructural y funcional de diversos miembros levantada y alojada en bases de datos permite proponer la agrupación de los transportadores de esta familia a partir de características claves para la selectividad y el transporte. En este trabajo se comparan los agrupamientos de miembros de esta familia representativos de todos los reinos de la vida obtenidos mediante aproximaciones filogenéticas y análisis cluster de regiones funcionales, y se evalúa el nivel de asociación entre los grupos identificados y las anotaciones funcionales con evidencia empírica. Así también, se identifican los residuos que más contribuyen a los agrupamientos observados por los métodos empleados.

ABSTRACT

Sugars such as glucose and related polyols are essential compounds in the metabolism of all living organisms, so their transport through biological membranes is essential for homeostasis and survival. Its transport is mediated by transporters belonging to the sugar porter (SP) family, TC 2.A.1.1, which covers various homologs of the SLC2 genes that encodes the known GLUTs in humans and animals.Various studies have been developed for the functional characterization, in its broadest sense, of various members of this family, with special emphasis on GLUTs due to their medical and pharmacological importance. Although in recent years and thanks to the availability of representative three-dimensional structures of some family members rapid progress has been made in understanding the molecular bases that support substrate selectivity and transport mechanism in the family, to date there is no integrative study that allows general principles of these bases, which requires GLUTs homologs in bacteria, fungi, protozoa and plants to be considered. The structural and functional information of various members collected over the years and stored in databases allows us to propose the clustering of sugar transporters based on key characteristics for their specificity and transport. In this work, the clustering of members of this family representative of all kingdoms of life are compared by means of phylogenetic approximations and clustering of functional sites, and the level of association between the clusters identified and functional labels with empirical evidence is evaluated. The residues that are most likely to determine the clustering observed are also identified.

1 INTRODUCCIÓN

El presente trabajo busca contribuir a una mejor comprensión de las bases moleculares determinantes de la funcionalidad de los miembros de la familia de transportadores de azúcares, proteínas evolutivamente relacionados con los genes SLC2 que codifican para los conocidos GLUTs en humanos y que se encuentran asociadas, como su nombre sugiere, al transporte de azúcares en los distintos seres vivos, siendo esenciales para su metabolismo y supervivencia.

El interés de este trabajo viene dado por la utilidad que este conocimiento representa para el desarrollo de aplicaciones en áreas como la medicina, farmacología e industria biotecnológica. Patrones de expresión anormales, así como variantes de algunos GLUTs humanos se han asociado a la manifestación de una serie de enfermedades como los síndromes de deficiencia de GLUT1, síndrome de tortuosidad arterial, hipouricemia renal, diabetes y cáncer; representando objetivos importantes para el desarrollo de medicamentos y herramientas diagnósticas (Augustin, 2010; Ferreira et al., 2019; Mike Mueckler & Thorens, 2013; Yan, 2017). De similar importancia son los homólogos presentes en patógenos como bacterias y parásitos, debido a su uso potencial como blancos terapéuticos y al papel que pueden jugar en el desarrollo de mecanismos de resistencia a fármacos (Diallinas, 2014; Majd et al., 2018). Así, este conocimiento es fundamental tanto para una mejor comprensión de los mecanismos fisiopatológicos asociados como para guiar el diseño racional de fármacos. También resulta informativo para inferir la función de transportadores poco caracterizados de la familia (Yan, 2015) así como para guiar la modificación de la selectividad de transportadores, lo cual tiene diversas aplicaciones biotecnológicas. Por ejemplo, uno de los objetivos de la industria de biocombustibles es desarrollar cepas recombinantes de *Saccharomyces cerevisiae* que también pueden fermentar, además de hexosas, xilosa y arabinosa de manera eficiente (Moysés et al., 2016; Rottmann et al., 2018; Zhao et al., 2020).

Comprender las bases moleculares de la función de proteínas involucra no sólo el estudio funcional detallado de proteínas individuales, sino también el descubrimiento de relaciones, patrones, principios generalizadores a partir de estudios comparativos que nos permitan levantar modelos lo suficientemente robustos como para poder actuar sobre estas bases. Esto implica la identificación y comprensión del rol de residuos y motivos que actúan como *determinantes funcionales*, patrones de aminoácidos que distinguen un grupo o clase funcional de otro (Fetrow & Babbitt, 2018). Sin embargo, la definición de grupos funcionales en proteínas no es una tarea fácil, pues el concepto de función es complejo y su correcta caracterización no es trivial.

En este trabajo se plantea la definición y caracterización de grupos funcionales en la familia de transportadores de azúcares, contrastando los resultados obtenidos por dos aproximaciones metodológicas empleadas para estos fines: análisis filogenético y análisis clúster de regiones funcionales, esta última surgida recientemente como una nueva aproximación al estudio de la función de proteínas y que hace uso de herramientas propias del aprendizaje automático. Las bases conceptuales, teóricas y metodológicas que sustentan el uso de esta última aproximación, se presentan en los siguientes apartados. En primera instancia se exponen los aspectos y desafíos relacionados a la caracterización funcional de transportadores, seguidos de los métodos existentes para abordar esta tarea. Se discuten las principales limitaciones de los métodos computacionales tradicionales empleadas para estos fines y las ventajas que comporta el uso de esta nueva aproximación, para finalizar contextualizando lo expuesto a la familia en estudio y presentando la propuesta de trabajo.

1.1 Caracterización funcional de proteínas de transporte.

El concepto de función biológica de una proteína es un concepto amplio, por lo que su caracterización abarca diversos niveles. Desde el nivel más global hasta el más específico podemos considerar las funciones sistémica, fisiológica, celular y, finalmente, la función bioquímica o molecular (Watson et al., 2005). Las tres primeras involucran aspectos como la determinación de los principales tejidos y niveles en los que es expresado en un organismo, las vías metabólicas en las que participa, su movilización y localización subcelular, entre otros aspectos relacionados; teniendo por último la función bioquímica o molecular, foco de investigación de este trabajo.

La caracterización de la *función bioquímica o molecula*r de una proteína involucra la comprensión de los detalles moleculares de su actividad biológica, es decir, la identificación y caracterización de sitios funcionales y mecanismos asociados a tal actividad (Fetrow & Babbitt, 2018). En transportadores, la actividad biológica corresponde a la translocación selectiva de sustratos a través de membranas celulares, por lo que la caracterización de la función de un transportador puede dividirse en dos componentes relacionados:

(1) La determinación del sustrato o rango de sustratos para los cuales es selectivo.
(2) La caracterización de los mecanismos de unión y/o translocación de sustratos y otras moléculas o interacciones reguladoras.

Una mirada general de los métodos usados y conceptos relacionados a la determinación experimental de él o los sustratos fisiológicos de un transportador (es decir, los sustratos que transporta en el contexto celular-metabólico en el que se encuentra) y otros aspectos de la actividad de transporte se presenta en el apartado 1.2. Así, si bien existen diversas técnicas que permiten la caracterización de proteínas transportadoras a nivel experimental, la complejidad de las proteínas de membrana dificulta esta tarea, la cual suele ser costosa y demandar largos

periodos de tiempo (Lundstrom, 2006), lo que contrasta con el rápido crecimiento en el número de secuencias identificadas como proteínas de membrana, las cuales representan entre un 20%-30% de los marcos de lectura abierto de los genomas de todos los organismos (Stevens & Arkin, 2000; Wallin & Heijne, 1998), siendo entre un 5-15% proteínas transportadoras (T. J. Lee et al., 2008; Majd et al., 2018).

Frente a la necesidad de inferir la función molecular de nuevas secuencias y clasificarlas según funcionalidad, desde el nacimiento de la biología computacional en los años 80 se han desarrollado diversos métodos computacionales que buscan apoyar esta tarea. Una visión general de los métodos computacionales clásicos empleados en la anotación funcional de genes y definición de grupos funcionales junto a las principales limitaciones asociadas a estos métodos se abordan en el apartado 1.3; para luego tratar metodologías basadas en aprendizaje automático.

Los métodos descritos así como el propuesto en este trabajo se basan en principios biológicos inherentes a la función molecular, y que conviene tenerlos presentes a la hora de analizar sus alcances y limitaciones. Éstos son:

- Principio de secuencia-estructura-función: Un principio básico de la biología es que la secuencia de aminoácidos de las proteínas determina su estructura tridimensional y función bioquímica, ya que tanto el plegamiento como la función se configuran a partir de patrones de interacciones físicas entre los átomos que constituyen la macromolécula. Este principio, descrito por primera vez por Anfinsen (1973), dio nacimiento a la biología computacional, abriendo la posibilidad de usar la secuencia de aminoácidos de una proteína para predecir sus propiedades funcionales y estructurales.

- Principio de Jerarquía: Este principio reconoce niveles de especificidad funcional que pueden ser descritos con distintos grados de detalle, abriendo la posibilidad de definir grupos, clases o familias funcionales de proteínas -el último concepto suele ser el más utilizado en la literatura- de manera jerárquica (Fetrow & Babbitt, 2018), en donde los grupos definidos en los niveles inferiores de la jerarquía comparten mayores propiedades funcionales, es decir, son grupos *isofuncionales* (concepto utilizado por Boari de Lima et al., 2016) como se esquematiza en la **Figura 1**. El sistema de clasificación de la Comisión de Enzimas se basa en este principio: usando criterios meramente funcionales, a cada enzima le asigna un código de 4 dígitos en donde W representa 1 de 6 tipos principales de reacciones químicas (clases) a las cuales pertenece la enzima, X indica un nivel de reacción más detallado, es decir, subclase, y los últimos dos dígitos indican sub-subclases (típicamente especificidad de reacción y especificidad del sustrato, respectivamente).

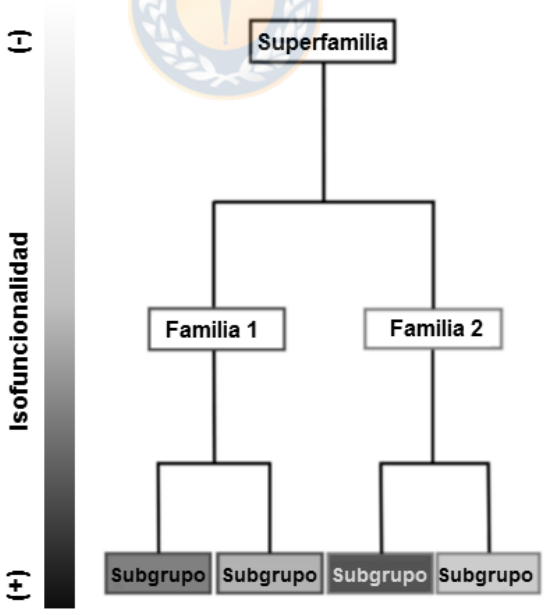


Figura 1. **Esquematización del principio de jerarquía y niveles de isofuncionalidad en familias de proteínas.**

- Principio de evolución molecular de la función: La diversidad funcional de proteínas en sus distintos niveles nace de mecanismos de evolución molecular. Es decir, se entiende que la arquitectura de las familias de proteínas es concebida a partir de la divergencia de secuencia acumulada como consecuencia de mecanismos de especiación, principalmente mutacionales como sustituciones, eventos de inserción y deleción de bases o segmentos de ADN, duplicación de genes, entre otros; sumado a la presión selectiva evolutiva ejercida sobre cada proteína (Nei & Kumar, 2000; Rausell et al., 2010). La teoría del reloj molecular establecida por Zuckerkandl y Pauling en la década de los 60 (citada por Morgan, 1998) establece que las secuencias biológicas evolucionan según tasas de cambio medibles y relativamente constantes, permitiendo el estudio de la evolución molecular a partir de análisis de secuencias de ADN y proteínas, mediante el desarrollo de modelos evolutivos.

- Principio de interacciones determinantes: Sólo una pequeña fracción de los aminoácidos de una proteína son determinantes para su función. Varios estudios demuestran que en los sitios funcionales, configurados a partir del plegamiento de una proteína, existen residuos claves que confieren la función molecular en estudio. Uno de los estudios clave que fundamentan este principio es el de Bogan & Thorn (1998), quienes estudiaron la energía libre de unión en interfaces de interacción proteína-proteína, encontrando que ésta no se distribuye uniformemente entre las interfaces; sino que existen '*hot spots*' de energía de unión formados por un pequeño subconjunto de residuos de las interfaces. Estudios en transportadores también respaldan este principio (ver sección 1.2).

1.2 Métodos experimentales empleados en la caracterización funcional de transportadores

1.2.1. Caracterización de la actividad de transporte

La actividad de transporte llevada a cabo por una proteína transportadora se asocia a distintos conceptos: selectividad, especificidad, mecanismos de transporte y de regulación. La selectividad se define como la capacidad de permitir el transporte de cierto sustrato o rango de sustratos, negando el paso de otros químicamente disímiles. La especificidad es una propiedad relacionada que corresponde a la capacidad de preferencia de transporte entre sustratos químicamente relacionados para los cuales existe selectividad (Heinz, 1978), derivando dos conceptos asociados: afinidad, referente a la capacidad de unión o reconocimiento de sustrato; y capacidad de transporte, asociado a la velocidad de translocación. Ambas propiedades determinan, en conjunto, la eficiencia del transporte. Comprender el mecanismo de transporte engloba dilucidar los mecanismos moleculares tanto de la etapa de reconocimiento de sustratos como la etapa de translocación y liberación de los mismos, junto con conocer la energética asociada. Por último, la regulación implica dilucidar mecanismos alostéricos o cooperativos.

La caracterización experimental de estas propiedades se realiza principalmente mediante estudios cinéticos para lo cual existen distintos tipos de ensayos. En general, en estos ensayos se mide la entrada o salida de un sustrato o análogo en el tiempo en condiciones específicas de temperatura, pH y en presencia o ausencia de otras moléculas (co-sustratos, inhibidores, etc) desde un sistema que exprese de manera natural o artificial el transportador deseado, estos sistemas pueden ser células, vesículas de membrana o liposomas. La variación en la concentración de sustrato se puede seguir mediante señales de radiactividad, fluorescencia, entre otros; utilizándose típicamente marcas radioactivas para sustratos que no son iónicos (Xie, 2008).

En caso de utilizarse células, la actividad de transporte observada puede deberse a la presencia de distintos sistemas de transporte selectivos para el sustrato en cuestión, por lo que se busca la expresión enriquecida del gen en células heterólogas y que no expresan transportadores con actividades relacionadas. Se suelen utilizar como sistemas de expresión células de *S.cerevisiae*, oocitos de *Xenopus* o líneas celulares como células CHO para transportadores eucariontes, y cepas de *E.coli* para procariontes. De todas maneras, en estos casos la actividad de transporte se suele confirmar mediante ensayos de pérdida y ganancia de función.

Cabe destacar que para bacterias y levaduras, la actividad de transporte de un gen putativo para un sustrato que puede ser utilizado como fuente de energía celular, se puede inferir analizando los fenotipos de crecimiento de cepas que expresan o no el transportador, al ser cultivadas en medios enriquecidos con el sustrato en cuestión como fuente de carbono. Estos estudios se suelen complementar con análisis de inducción de la expresión del gen a determinadas concentraciones de sustrato. Estos métodos se suele emplear para caracterizar sistemas de transporte de azúcares y polioles en bacterias y levaduras.

Los análisis cinéticos reportados en literatura suelen ser particulares o comparativos, y las principales métricas asociadas corresponden a tasas de entrada/salida (cantidad de sustrato transportado/ sistema de expresión /unidad de tiempo) o a parámetros cinéticos, como se detalla en la **Tabla 1**. El estudio de la cinética del transporte de sustratos permite desarrollar modelos cinéticos en coherencia, cuya validez dependerá de su capacidad de reproducir los parámetros y otros comportamientos cinéticos observados experimentalmente.

Tabla 1. Estudios cinéticos de la actividad de transporte

	Análisis	
Métricas	**Particular**	**Comparado**
Curvas de progreso y tasas de entrada o salida	Se evalúa la capacidad de transporte de un sustrato a partir de una determinada concentración inicial de éste en el compartimento interno o externo del sistema de expresión a determinadas condiciones del medio.	Se comparan las tasas de entrada/salida para distintas condiciones. Los principales análisis son: Comparación de actividad de transporte de un sustrato frente a la presencia de potenciales inhibidores o activadores a una determinada concentración. Se suelen comunicar mediante gráficas de inhibición porcentual, siendo el 100% de actividad la observada para el sustrato sin presencia de otras moléculas. Comparación de transporte de un sustrato mediado por variantes del transportador. Las variantes pueden ser, por ejemplo, homólogos o transportadores con mutaciones sitio-dirigidas.
Parámetros cinéticos	Se evalúa la cinética de transporte de un sustrato analizando la dependencia del flujo de la concentración de sustrato. Para cinéticas de saturación, se determinan parámetros Km y Vmax análogos a los determinados en enzimas. Para ello se pueden emplear distintos procedimientos.	La comparación de parámetros cinéticos determinados para distintos sustratos de un mismo transportador permite evaluar su especificidad. Se suelen comparar los valores de Km para evaluar afinidad por sustrato. Comparaciones similares se realizan entre homólogos, en el contexto de mecanismos de transporte comunes. Para moléculas reguladoras o co-sustratos, permite estudiar los mecanismos de transporte y de regulación (inhibición o activación), de manera análoga a los métodos empleados para enzimas.

Los transportadores se clasifican según criterios cinéticos en canales y carriers. Estos últimos, según la energética y mecanismo de transporte se clasifican en bombas, asociadas al transporte activo primario; cotransportadores y contratransportadores (o intercambiadores), asociados al transporte activo secundario; y facilitadores, asociados al transporte transporte facilitado o equilibrativo. Estos tres últimos se agrupan bajo la denominación de 'portadores' (Saier, 2000). Los carriers, a diferencia de los canales que median el transporte relativamente libre de sustratos a través de la membrana, exhiben una cinética de saturación que puede ser descrita por modelos análogos a los empleados para describir la cinética enzimática. La cinética de saturación de transporte de un sustrato a partir del análisis de velocidades iniciales suponiendo condición de estado estacionario y con concentración de sustrato despreciable en el compartimento o lado de la membrana hacia donde se dirijeel flujo de sustrato medido, permite la determinación de las constantes de Michaelis (Km) de entrada y salida, y las velocidades de transporte o tasas de flujo máximas (Vmáx) de entrada y salida, dependiendo si se estudia el influjo o el eflujo de sustrato. Sin embargo, mantener concentraciones despreciables de sustrato en el compartimento opuesto al flujo medido se complica, sobre todo para sistemas de transporte equilibrativo en donde, si los flujos de entrada y salida son similares y la concentración de sustrato empleada es baja, comienzan a ocurrir flujos bidireccionales. Existen otros procedimientos empleados para el estudio de facilitadores que permiten la determinación de parámetros cinéticos para un sustrato marcado isotópicamente considerando presencia de sustrato no marcado en el compartimento opuesto al origen del flujo medido. Para que estos procedimientos fueran reproducibles y no hubiera ambigüedad en su comunicación, se desarrolló la siguiente nomenclatura (Cura & Carruthers, 2012; Konings et al., 1996): Los compartimentos separados por la membrana se nominan como 1 y 2, indicando compartimento interno y externo, respectivamente. Se indica si se medirá la entrada (21) o salida (12) de sustrato. Las concentraciones de sustrato en los compartimentos pueden ser 'cero' o despreciable, denotado por *zero (o z)* o bien estar en concentración saturante o

‘infinita’, denotado por *infinite (o i)*. Se utiliza la nomenclatura *cis (o c)* para indicar el compartimento o lado de la membrana en el que se origina el flujo del sustrato medido, y *trans (o t)* para indicar el compartimento o lado de la membrana hacia donde se dirije el flujo de sustrato medido. Los procedimientos empleados para la construcción de curvas de saturación corresponden a ensayos de tipo *Zero trans entry/exit, infinite trans entry/exit e infinite cis entry/exit.* Por último, se denota como *equilibrium exchange* (o ee) la condición en la que se mide el flujo de sustrato en condiciones de equilibrio, es decir, con concentraciones iguales de sustrato en ambos compartimentos. Un esquema de los ensayos, nomenclatura asociada y parámetros equivalentes se presenta en la **Tabla 2**. La medida de la entrada y salida de sustrato considerando concentraciones despreciables de sustrato en el lado ‘trans’ de la membrana corresponderán a experimentos de tipo “*Zero trans entry*” o ‘zt 21’ y “*Zero trans exit*” o ‘zt 12’, respectivamente Las velocidades quedan determinadas por:

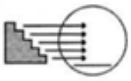
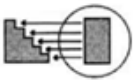
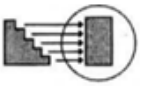
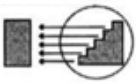
$$v_{21}^{zt} = \frac{V_{21}^{zt}\,[S]_2}{K_{21}^{zt}+[S]_2} \qquad v_{12}^{zt} = \frac{V_{12}^{zt}\,[S]_1}{K_{12}^{zt}+[S]_1}$$

Donde los valores de V y K corresponden a los valores de Vmáx y Km de entrada y salida. Este último corresponde a la concentración de sustrato requerida para que la velocidad medida en tal condición sea la mitad de la Vmax, y se se suelen interpretar como una medida de la afinidad aparente entre sustrato y transportador en sus conformaciones endo y exofaciales (Naftalin, 2018; Stein & Litman, 2014) (conformaciones asociadas al modelo de transporte de acceso alternante tratado más adelante, ver **Figura 3**), por lo que su comparación para distintos sustratos de un mismo transportador permite caracterizar tal componente de especificidad, utilizándose típicamente para categorizar al transporte de alta o baja afinidad. El análisis de Vmax, por otro lado, permite categorizar al transporte de alta o baja capacidad, evaluándose la eficiencia a partir del análisis de ambos parámetros. La misma batería de ensayos puede utilizarse para caracterizar al resto de portadores y la interpretación cinética de los parámetros cinéticos obtenidos a partir de ellos depende del modelo cinético que se asuma y que de cuenta de estos mismos.

Actualmente el modelo de transporte de acceso alternante (MTAA) es el modelo general que ha permitido describir el transporte mediado por diversos carriers con mayor éxito, y se presenta para los distintos mecanismos de transporte en la **Figura 2** (Bonting & De Pont, 1981; Sperelakis, 2012; Stein & Lieb, 1986; Stein & Litman, 2014). Este modelo nace de modelos análogos levantados por diversos autores con diferentes nombres y alcances cinéticos: modelo de carrier móvil propuesto por Widdas, (1952) para el transporte facilitado de azúcares, modelo de 'poro mediado por compuerta' (Patlak, 1957), modelo de 'confórmero alternante' (Vidaver, 1966), modelo de 'carrier simple' (Lieb & Stein, 1974), entre otros. El modelo, como su nombre indica, involucra una interpretación molecular del proceso de transporte. Si bien se levantaron interpretaciones moleculares alternativas que derivan las mismas ecuaciones cinéticas, fueron refutadas posteriormente. Esta interpretación fue primeramente introducida por Patlak y Vidaver, pero Jardetzky (1966) la popularizó al publicar en *Nature* su modelo a partir de sus estudios de bombas dependientes de ATP, enumerando tres condiciones estructurales para que una molécula polimérica (proteína) mediara el transporte de sustratos:

> (1) Debe contener una hendidura o una cavidad en el interior de la molécula, lo suficientemente grande como para admitir una molécula pequeña. (2) Debe poder asumir dos configuraciones diferentes, de modo que la cavidad molecular esté abierta hacia un lado de la membrana en una configuración y hacia el lado opuesto en la otra. (3) Debe contener un sitio de unión para las especies transportadas en su cavidad molecular, cuya afinidad para las especies transportadas sea diferente en las dos configuraciones (p. 969).

Tabla 2. Ensayos para la caracterización cinética de transportadores

Condición		Esquema	Vmax[a]	Km[a]
Zero trans	*Entry*		$V_{21}^{zt}{}_{(en)}$	$K_{21}^{zt}{}_{(en)}$
	Exit		$V_{12}^{zt}{}_{(ex)}$	$K_{12}^{zt}{}_{(ex)}$
Infinite cis	*Entry*		$V_{21}^{ic}{}_{(en)}$	$K_{21}^{ic}{}_{(ex)}$
	Exit		$V_{12}^{ic}{}_{(ex)}$	$K_{12}^{ic}{}_{(en)}$
Infinite trans[b]	*Entry (Forward exchange)*		$V_{21}^{it} = V_{ee}$	$K_{21}^{it}{}_{(en)}$
	Exit (Backward exchange)		$V_{12}^{it} = V_{ee}$	$K_{12}^{it}{}_{(ex)}$
Equilibrium exchange			V_{ee}	K_{ee}

Adaptado de Bonting & De Pont, 1981. En los esquemas, la circunferencia representa el límite del compartimento interno. En gris, los gradientes de concentración de sustrato evaluados. Las flechas indican la dirección del flujo medido. [a] Los subíndices *en* y *ex* indican parámetros de entrada y salida equivalentes entre sí. [b] Conocidos también como ensayos de contraflujo (*counterflow assays*).

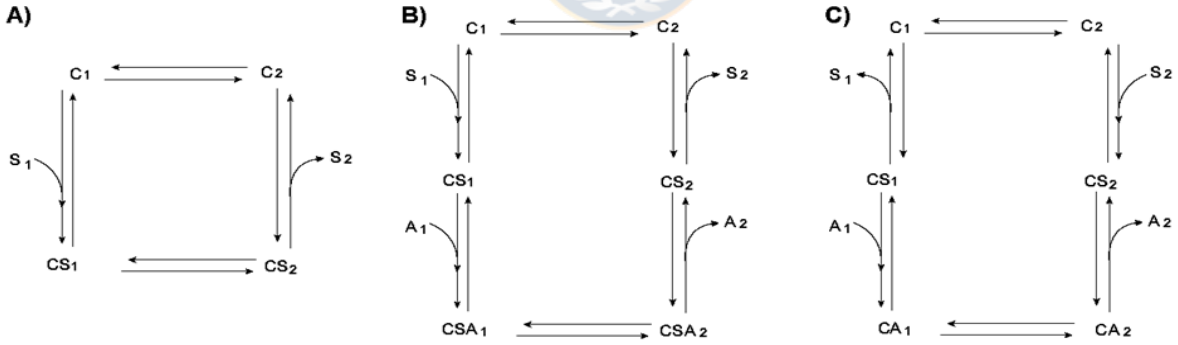


Figura 2. **Modelo de acceso alternante para distintas cinéticas de transporte.**
'C' representa a los confórmeros del Carrier, dispuesto para la unión de sustrato en uno u otro lado de la membrana (A) Modelo cinético para transporte facilitado de sustrato 'S'. (B) Modelo cinético para cotransporte de sustratos 'S' y 'A' (C) Modelo cinético para contratransporte de sustratos 'S' y 'A'. Las velocidades de cada paso se definen a partir del producto de su respectiva constante cinética y las concentraciones de las especies involucradas. Las ecuaciones cinéticas derivadas de los modelos y considerando las condiciones impuestas por cada procedimiento permite derivar ecuaciones simplificadas que toman la forma de una curva de saturación michaeliana.

El modelo involucra la existencia de un único sitio de unión central, el cual se alterna a ambos lados de la membrana, debiendo pasar por al menos un estado conformacional intermedio en el que el sitio de unión al sustrato está ocluido por ambos lados, sin haber acceso directo de un lado de la proteína al otro, como se observa en los canales. Esto implica la transición entre, al menos, 3 estados conformacionales principales: exofacial, ocluido y endofacial, como se muestra en la **Figura 3.**

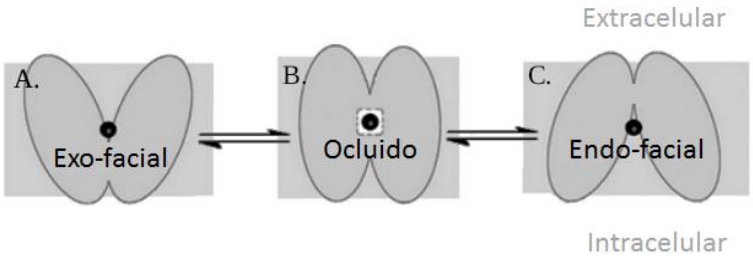


Figura 3. **Transiciones estructurales en MTAA.** Para completar un ciclo de transporte, un transportador debe someterse a cambios conformacionales. Se identifican 3 estados conformacionales principales: A. exo-facial. B. ocluido. C. Endo-facial. El sustrato se representa por una esfera negra.

La tercera condición que introduce Jardetzky implica que los flujos de transporte de entrada y salida de un sustrato mediados por Carrier son asimétricos, lo que se refleja en valores de Kzt 12 y Kzt 21 diferentes para un determinado sustrato. Si bien la asimetría es esperable para la mayoría de transportadores independiente de su mecanismo al tratarse de conformaciones distintas, existen transportadores que exhiben simetría en sus constantes cinéticas.

Cabe destacar también el fenómeno de trans-aceleración, fenómeno observado en facilitadores y referente a la estimulación del influjo o del eflujo de sustrato por la presencia de sustrato en el compartimento *trans.* Este fenómeno se traduce en Kee y Vee mayores que los valores V21 y V12. En caso de ser un facilitador asimétrico, Kee tendrá valores más próximos la constante Kzt de mayor valor, debido a que en condiciones de intercambio en equilibrio ambas conformaciones (endo y exofacial) deben estar saturadas para que la velocidad sea máxima, y la saturación de la conformación de menor afinidad requiere concentraciones mayores de sustrato (Naftalin, 2018). Se debe tener precaución al utilizar los

parámetros cinéticos para estudios comparativos, no se debe dejar de considerar su dependencia de la temperatura y otras condiciones del medio, la asimetría esperable para la mayoría de los sistemas de transporte o el fenómeno de trans-aceleración A su vez, se debe tener en cuenta que para co y contratransportadores, los valores de los parámetros cinéticos son dependientes de la concentración de co-sustrato empleada y se suelen reportar como valores de Km y Vmax aparentes a concentraciones fisiológicas o saturantes de los co-sustratos.

1.2.2. Caracterización de residuos y sitios funcionales

La caracterización de residuos y sitios funcionales de transportadores está estrechamente ligada a la caracterización de la topología y estructura tridimensional de éstos, pues el conocimiento de la disposición espacial de los aminoácidos facilita la caracterización sitios de unión y translocación.

Estudios tempranos basados en construcción de perfiles de hidrofobicidad y predicción de segmentos transmembrana a partir de la secuencia de transportadores, junto a análisis bioquímicos que permitían evaluar la topología predicha y determinar residuos críticos para la función de transporte, permitieron levantar modelos estructurales, en principio, coherentes con el MTAA introducido en el apartado anterior, mucho antes de que se resolvieran estructuras cristalográficas de transportadores por cristalografía de rayos X. Las principales técnicas empleadas para estos propósitos son el escaneo de mutagénesis de alanina y cisteína, donde esta última ha demostrado ser una gran herramienta para asignar roles funcionales o estructurales a residuos específicos en transportadores. De estos estudios, por lejos los más relevantes fueron los llevados a cabo por el grupo de investigación de Kaback sobre LacY, primer gen con una actividad de transporte específica que se clonó y secuenció (Büchel et al., 1980). En LacY, cada uno de los 417 residuos aminoacídicos fue mutado, encontrándose que sólo menos de 10 residuos ubicados en los TM 4,7,8,9 y 10 pueden abolir completamente la función de transporte (Frillingos et al., 1998), esto en concordancia con el principio de interacciones determinantes. Posterior a esto y

a partir de análisis detallados de las estructuras cristalográficas de proteínas de transporte que se han resuelto por cristalografía de rayos X desde principios de siglo -existiendo estructuras representativas de distintos estados conformacionales y co-cristalizados con sustratos o inhibidores- , complementados con estudios computacionales basados en modelamiento y dinámica molecular y otros métodos experimentales; se han levantado distintos modelos para explicar aspectos específicos de mecanismos de transporte propios de algunos sistemas representativos de superfamilias o familias de transportadores, dentro de los cuales encontramos los mecanismos denominados *Rocker-switch, Gated-pore* y *elevator* (Colas et al., 2016; Diallinas, 2014), como se muestra en la **Figura 4.**

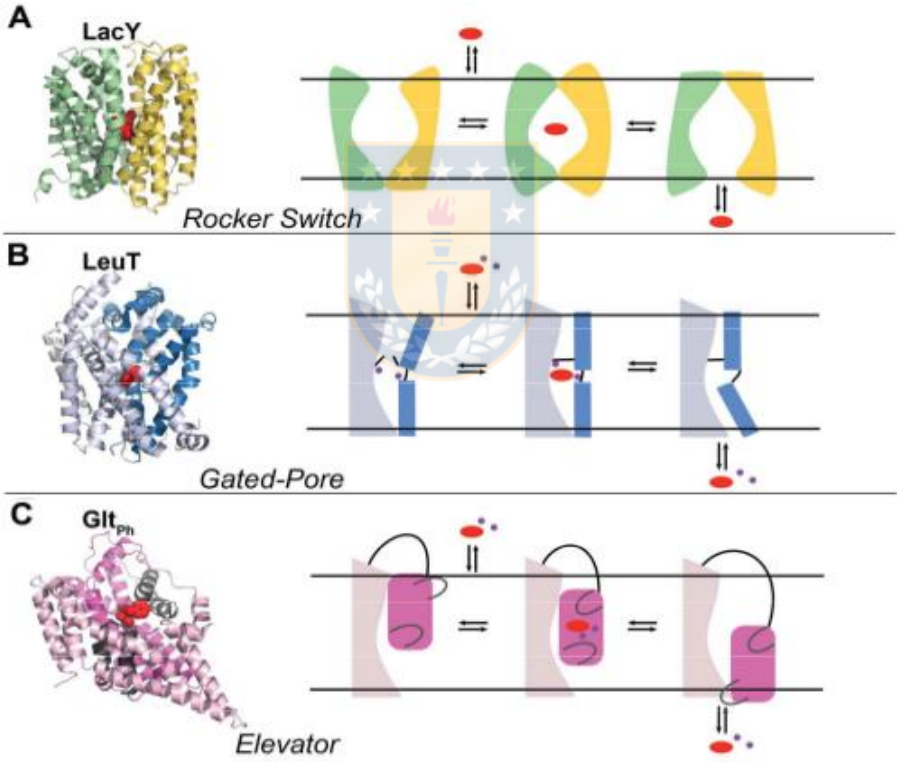


Figura 4. **Mecanismos de transporte actuales.** Adaptado de Colas et al. (2016). Sistemas representativos: (A) LacY, representativo de la MFS (TC 2.A.1). (B) LeuT, representativo de la familia NSS (TC 2.A.22). (C) GltPh, representativo de la familia DAACS (TC 2.A.23).

1.3 Métodos bioinformáticos empleados en la caracterización funcional de genes

1.3.1. Bases de datos y nivel de evidencia de las anotaciones funcionales

El aumento exponencial de información genómica levantada desde proyectos de secuenciación desembocó en la creación de bases de datos que alojan información de secuencia de los marcos de lectura abiertos identificados y de los productos génicos derivados de estos de diversas especies (secuencias de nucleótidos y de proteínas, respectivamente). Algunos, hipotéticos, otros con algún nivel de caracterización experimental. Así, además de la secuencia y el organismo de origen, existen bases de datos que complementan la información de secuencia con anotaciones funcionales con distintos niveles de evidencia.

Actualmente, la base de datos de secuencias de proteínas con anotaciones curadas más comúnmente citada es Swiss-Prot (Bairoch & Apweiler, 1997), la cual es mantenida por el Instituto Suizo de Bioinformática y el Instituto Europeo de Bioinformática (EBI). Dado que mantener la anotación de alta calidad de Swiss-Prot limitó su crecimiento, en 1997 se introdujo una base de datos complementaria llamada TrEMBL. TreMBL consta de entradas anotadas automáticamente por computadora desde la base de datos de secuencias de nucleótidos EBI; por lo que TrEMBL es significativamente más grande que Swiss-Prot. Ambas bases de datos se encuentran incorporadas en la base de datos UniProt (https://www.uniprot.org/). Esta última posee etiquetas de evidencia asociadas a la descripción de la función, de estar presente, para cada entrada. Estas etiquetas corresponden a códigos propios de la ontología ECO (*Evidence and Conclusion Ontology,* Giglio et al., 2019), conformado por dos componentes: el origen de la evidencia y, cuando corresponde, la fuente de información, que suelen ser artículos científicos representados mediante su ID de PubMed (véase https://www.uniprot.org/help/evidences).

Además del sistema ECO, destaca la base de datos de Gene Ontology (GO) (Ashburner et al., 2000). Esta ontología de genes se considera el esfuerzo más ambicioso realizado para integrar toda la información funcional de las proteínas. El objetivo del consorcio GO es proporcionar un sistema de vocabulario y etiquetas que abarque las anotaciones funcionales relacionadas con todos los niveles ya descritos, por ejemplo, anotaciones relacionadas con la localización subcelular, y (por supuesto), la función molecular. Estas anotaciones funcionales son integradas en la base de datos UniProt para cada una de sus entradas.

Gene ontology proporciona un sistema de códigos jerárquicos para clasificar el nivel de evidencia de las anotaciones funcionales, de manera similar a la ontología ECO (véase http://geneontology.org/docs/guide-go-evidence-codes/). Las tres categorías principales son códigos de evidencia experimental, los cuales abarcan evidencia levantada, en el caso de los transportadores, mediante métodos abordados en la sección 1.2; códigos basados en análisis computacionales, en donde se abarcan tanto métodos basados en similaridad como métodos de aprendizaje automático (entre otros); y códigos de anotaciones basadas en inferencia filogenética.

1.3.2. Métodos basados en similaridad

Los métodos de transferencia de anotaciones funcionales basados en similaridad entre proteínas fueron los primeros en desarrollarse y son, por lejos, los más utilizados hoy en día (Brown & Sjölander, 2006; Fetrow & Babbitt, 2018). Se han desarrollado una gran cantidad de plataformas que transfieren anotaciones a una proteína cuantificando su similaridad contra proteínas alojadas en bases de datos mediante distintas métricas y criterios. Estas aproximaciones, a partir del principio de secuencia-estructura-función, asumen que proteínas suficientemente similares en secuencia y estructura realizan funciones similares (D. Lee et al., 2007). Se pueden clasificar *grosso modo* en métodos que miden similaridad de secuencia, métodos basados en identificación de motivos o dominios, y métodos basados en similaridad estructural (Watson et al., 2005). Estos últimos suelen emplearse

cuando los dos primeros no arrojan resultados exitosos, o bien complementando y enriqueciendo los primeros análisis. En el primer caso, la predicción y análisis de la estructura de la proteína puede proporcionar pistas funcionales o confirmar asignaciones funcionales tentativas inferidas desde la secuencia, pues la estructura presenta mayor grado de conservación frente a grandes distancias evolutivas. Existen, evidentemente, métodos que integran los distintos tipos de análisis.

En breve, los primeros, se operacionalizan mediante alineamientos de secuencias, ejemplificados por los pioneros y conocidos algoritmos de alineamiento de pares de secuencia global y local Needleman-Wunsch (1970) y Smith-Waterman (1981), respectivamente, existiendo algoritmos basados en programación dinámica, que buscan alineamientos óptimos; y algoritmos heurísticos, que no son tan exactos, pero aceleran el tiempo de cómputo. Estos algoritmos emplean matrices de puntuación para ácidos nucleicos o de sustitución en el caso de proteínas - siendo las más conocidas las variantes de PAM (Dayhoff et al., 1978) y BLOSUM (Henikoff & Henikoff, 1992), pero también existiendo métricas construidas a partir de secuencias de proteínas de membrana, como las matrices PHAT y SLIM (Müller et al., 2001; Ng et al., 2000) - para modelar el ritmo al que un carácter en una secuencia cambia a otro carácter con el tiempo; además de puntajes de penalidad para el inicio y extención de inserciones (*apertura-penalty)*. Los paquetes de programa y algoritmos más conocidos para cálculos de similitud a partir de alineamientos globales y/o locales son BLAST (Altschul et al., 1997) y FASTA (Pearson, 1990). Éstos se encuentran disponibles en los sitios web https://blast.ncbi.nlm.nih.gov/Blast.cgi y https://www.ebi.ac.uk/Tools/sss/fasta/, respectivamente. Actualmente, la herramienta de búsqueda y comparación de secuencias más popular, tanto para aminoácidos como para secuencias de ácidos nucleicos es BLAST, la cual entrega porcentajes de identidad y métricas como *E-value* y *Bit-score*, que evalúan la significancia estadística del alineamiento. Ambos proporcionan información sobre la probabilidad de que un alineamiento determinado entre secuencias no tenga significancia biológica, es decir, que su similaridad sea producto del azar, permitiendo inferir homología mediante el

establecimiento de valores límite. Por ejemplo, la base de datos Pfam define un valor de *E-value* de ~10^{-2} como valor límite seguro para establecer homología (Punta et al., 2012). El sistema de clasificación de la comisión de transporte (TC), por otro lado, emplea un criterio equivalente a un valor límite de ~10^{-20} para establecer homología entre dos secuencias (Chang et al., 2004). El establecimiento de la significancia de la similaridad es necesario –si bien no siempre suficiente, como se argumenta más adelante– para inferir funcionalidades comunes a partir de ésta, por lo que estos métodos de anotación funcional también se conocen como métodos basados en homología.

Existen también métodos de detección de homólogos remotos basados en la construcción de perfiles de secuencia. Los más conocidos son PSI-BLAST (Altschul et al., 1997) y métodos basados en modelos ocultos de markov (HMMs). El funcionamiento de PSI-BLAST es el siguiente: Dada una secuencia de consulta y una base de datos de secuencias, se identifican homólogos por los métodos ya descritos, a partir de las cuales se construye una alineamiento múltiple de secuencias (MSA) -para los cuales existen una diversidad de algoritmos-, y se genera un perfil de puntaje para cada posición del alineamiento (PSSM, por sus siglas en inglés), que calcula la probabilidad de ocurrencia de cada aminoácido en cada una de las posiciones del alineamiento. Este perfil, que caracteriza al conjunto de proteínas relacionadas, se utiliza para buscar homólogos más remotos mediante un algoritmo similar a BLAST. Este proceso se itera hasta convergencia de las secuencias identificadas.

Los métodos basados en HMM como el algoritmo HMMER (Eddy, 1998), por otro lado, permiten modelar familias o motivos/dominios de proteínas homólogas mediante la construcción de perfiles estadísticos que consideran la distribución de aminoácidos y otras características de un MSA de proteínas. El concepto de HMM, en sus aspectos técnicos, es más complejo que los perfiles basados en PSSM, pero son muy similares en cuanto a su sentido biológico, permitiendo también la identificación de homólogos remotos mediante procesos iterativos. Una de sus ventajas es que permite asociar un valor de probabilidad de una secuencia de

pertenecer a una familia de proteínas o de presentar un motivo/dominio modelado por este método.

Pasando a ese último tema, los métodos de identificación de motivos/dominios de secuencia buscan la presencia de motivos o dominios, los cuales describen patrones de conservación asociados a un grupo de homólogos previamente caracterizados, en las secuencias nóveles a caracterizar. Actualmente existe una gran variedad de bases de datos de motivos/dominios con sus respectivas plataformas de búsqueda, variando en las formas en que representan y manipulan tales motivos y dominios. Las plataformas más sencillas se basan en la caracterización supervisada de motivos o de conjuntos de motivos lineales (firmas o *fingerprints),* representativos de un grupo de proteínas relacionadas evolutiva y funcionalmente a partir de un MSA de las secuencias, mediante expresiones regulares o métodos similares, como es el caso de PROSITE (Falquet et al., 2002) y PRINTS (Attwood, 2002). Métodos más refinados y/o que permiten automatización de procesos se basan en la construcción de perfiles de motivos, dominios o familias de proteínas mediante PSSMs o HMMs. Entre ellas encontramos la ya mencionada base de datos PFAM y PRODOM (Servant et al., 2002), por mencionar algunas de las principales (cabe destacar que en los últimos años PROSITE también ha integrado perfiles como forma de caracterización de motivos y dominios). De las plataformas que integran diversos métodos y bases de datos, destaca INTERPRO, cuyos sostenedores trabajan para integrar métodos basados en aprendizaje automático (Finn et al., 2017).

Todos los métodos anteriormente descritos se basan en la asunción mencionada anteriormente: proteínas suficientemente similares, a nivel local o global, realizan funciones similares. Sin embargo, esta asunción posee variadas limitaciones, nacidas principalmente del principio de interacciones determinantes, de la forma en que se evalúa lo 'suficientemente similar' y de la misma ambigüedad del concepto de función. A continuación explico esta idea con mayor detalle, contextualizada al ámbito de los transportadores:

En el caso de enzimas, si bien varios estudios han demostrado que secuencias homólogas con un porcentaje de identidad superior al 40% poseen una elevada probababilidad de desempeñar funciones comunes, empleando como criterio funcional los números del sistema de clasificación EC, existen contraejemplos donde una pequeña cantidad de residuos en proteínas muy similares están asociados con cambios dramáticos en la actividad enzimática (Pearson, 2013). Así también, si bien la mayoría de los transportadores son selectivos para un sustrato o un grupo de sustratos con estructuras químicas similares, la especificidad de sustrato o la afinidad de unión pueden variar dramáticamente, incluso entre miembros de una misma familia de transportadores que presentan una similaridad de secuencia y estructura elevada. Se ha demostrado que el cambio de un solo residuo aminoacídico en el sitio de unión de sustrato puede alterar la especificidad drásticamente (Diallinas, 2014; Majd et al., 2018). Más aún, es sabido que los sistemas de transporte suelen ser menos específicos que los sistemas enzimáticos comparables, es decir, su capacidad para seleccionar entre sustratos similares es menos pronunciada en transportadores, lo que dificulta aún más la transferencia de anotaciones funcionales por homología. Sin embargo, para la mayoría de los transportadores alojados en las bases de datos de secuencias, las anotaciones funcionales sin evidencia experimental suelen basarse sólo en homología de secuencia, lo que lleva a que las especificidades de algunos transportadores sean asignados de manera incorrecta (Majd et al., 2018). Muy a menudo, las anotaciones erróneas surgen de la transferencia de una función molecular más detallada de lo que realmente se puede inferir por homología o por la presencia de motivos representativos de una familia de homólogos. De hecho, los métodos de transferencia de anotaciones funcionales basados en homología se consideran una de las principales fuentes de error de anotación en las bases de datos debido a una aplicación excesivamente liberal del principio de similaridad (Fetrow & Babbitt, 2018; Zmasek & Eddy, 2002), lo que conduce a propagaciones de error en las bases de datos y termina por restarles valor informativo.

1.3.3. Métodos basados en inferencia filogenética

Las anotaciones funcionales basadas en análisis filogenéticos se basan en el principio de evolución molecular de la función: se derivan de un modelo evolutivo explícito que pretende explicar la ganancia y pérdida de la función génica en ramas específicas de un árbol filogenético. El procedimiento general, los alcances y limitaciones de este método son abordados por Brown & Sjölander (2006). En síntesis, la inferencia filogenética de la función proteica es un proceso de varios pasos que involucra selección de homólogos (por métodos basados en similaridad ya descritos, por lo que se puede considerar una extensión de los métodos basados en similaridad), alineamiento múltiple de las secuencias (MSA), y construcción de árboles filogenéticos. Paso siguiente, se integran las anotaciones funcionales de los homólogos utilizados para la reconstrucción filogenética en la topología del árbol; lo que permite definir grupos funcionales, en donde cada grupo o familia funcional se corresponde con clados delimitados del árbol, como se muestra en la **Figura 5.** Finalmente, se infiere la función de una proteína evaluando su posición en la filogenia: se asume que la proteína en estudio compartirá propiedades funcionales con los miembros del clado en el que quedó agrupada, las cuales se definen a partir de las anotaciones recuperadas. Considerando la diversidad funcional del clado, se suelen transferir las anotaciones del ortólogo más cercano, pues se asume que los parálogos poseen mayor divergencia funcional, pues al originarse por duplicación génica, poseen mayor libertad para fijar mutaciones y desarrollar nuevas funciones (Pearson, 2013). Si se conocen residuos críticos para la función de estos homólogos y éstos no se encuentran representados en la secuencia en estudio, se infiere la carencia de tal función (véase http://geneontology.org/docs/guide-go-evidence-codes/).

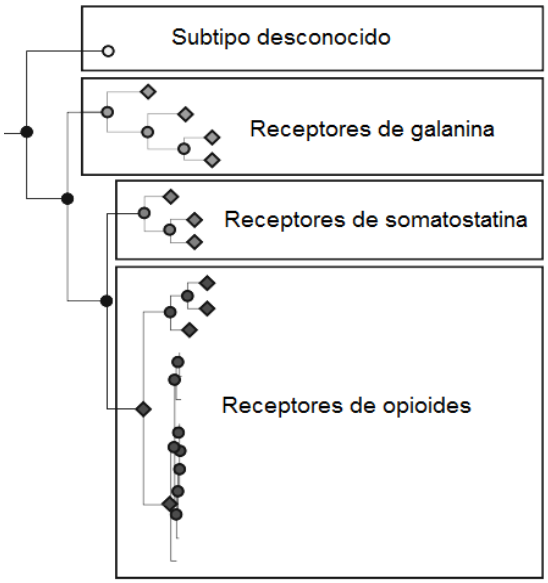


Figura 5. **Ejemplo de definición de grupos funcionales e inferencia de la función de proteínas mediante filogenética.** Adaptado de Brown & Sjölander (2006). Filogenia construida para receptores acoplados a proteína G.

Estos métodos poseen varias limitaciones técnicas como las asociadas a la evaluación de la calidad y/o exactitud del MSA o de la topología del árbol, o la propagación errónea de anotaciones funcionales al integrar anotaciones sin evidencia experimental directa. Más que abordarlas todas, quiero resaltar la siguiente, en total relación con las limitaciones de los métodos presentados en la sección anterior: si bien algunos atributos de las familias de proteínas como la estructura tridimensional se conservan a través de grandes distancias evolutivas, otros como la especificidad del sustrato pueden modificarse en función de sustituciones de aminoácidos en posiciones críticas. Así, la correspondencia entre la topología de la filogenia y los grupos o clases funcionales suele ser mayor a niveles más generales de jerarquía funcional. Esto se puede ver claramente reflejado en el sistema de clasificación de transportadores de la comisión de transporte (su formulación y nomenclatura se comenta en el **Anexo 1**). Si bien estos métodos han demostrado ser robustos para la definición de superfamilias y familias de proteínas con funcionalidades comunes, en niveles más detallados de funcionalidad estos métodos pueden presentar problemas, sin exhibirse una relación filogenética-funcional directa. A pesar de que, de acuerdo el escenario

actualmente aceptado de evolución divergente, esta relación filogenética-funcional debiera tender a verse en la mayoría de los casos; las reconstrucciones filogenéticas toman en cuenta información de las secuencias de proteínas a nivel global y, al ser muchos los factores funcionales y estructurales que conducen la evolución de una familia de proteínas; la divergencia específica asociada a la función de interés puede enmascararse dentro de la filogenia, pues sólo podemos observar una filogenia, que surge de una combinación de todas las diferentes restricciones (Landgraf et al., 2001; Pazos et al., 2006). Un ejemplo claro de esta situación se evidencia en la **Figura 6,** que muestra la clasificación filogenética de dominios SH3 encargados del reconocimiento de péptidos, caracterizados por unirse a motivos ricos en prolina en proteínas. Cesareni et al. (2002) clasificó estos dominios en diferentes clases funcionales dependiendo del tipo de péptido al que se unen. A pesar de tener un origen evolutivo común y una estructura general similar, su árbol filogenético no refleja la clasificación funcional (véanse ejemplos similares en Pazos et al., 2006). Cabe preguntarse entonces si las filogenias de las familias o subfamilias de transportadores definidos por la comisión de transporte mantienen o no esta relación funcional.

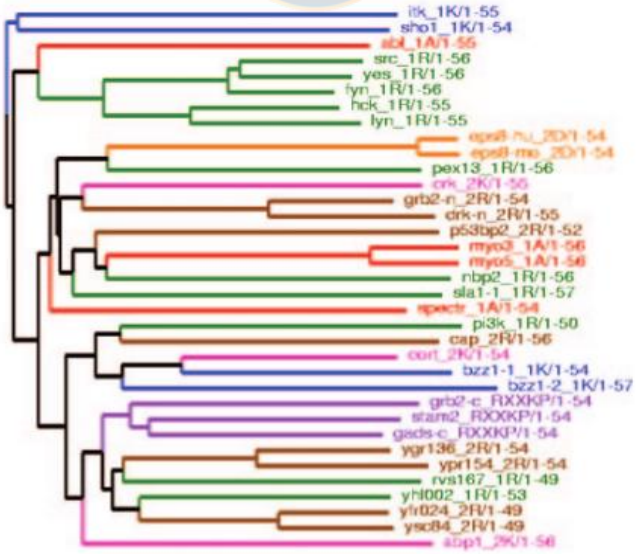


Figura 6. **Ejemplo de discrepancia entre la clasificación filogenética y funcional (Pazos et al., 2006).** Análisis filogenético de dominios SH3. Los colores reflejan las clases funcionales definidas por Cesareni et al. (2002)

1.4 Aproximaciones basadas en aprendizaje automático

Si bien la gran cantidad de información de secuencia, estructural y funcional de proteínas alojadas en bases de datos biológicos constituye una oportunidad para el levantamiento de modelos explicativos y predictivos de las propiedades funcionales de proteínas, al ser humano se le hace imposible hacerlo sin apoyo de una computadora. En este contexto, desde las ciencias de la computación y la información surge el concepto de descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Databases*, KDD); el cual hace referencia a los procesos de extracción de alto nivel de información o conocimiento desde un gran volumen de datos de bajo nivel (Brusic & Zeleznikow, 1999). Este proceso integra diversas etapas, las cuales se abordan brevemente a continuación:

1. Selección de un conjunto de datos de partida: Implica la selección del conjunto de datos y variables a partir de las que realizará el descubrimiento.

2. Limpieza y preprocesamiento de datos: Implica tareas como transformar datos a formatos con los que se pueda trabajar, normalización, filtrado y curado de datos según criterios coherentes con el problema de estudio, entre otros.

3. Ingeniaría de características: Implica la definición y selección de características útiles para representar los datos según el problema de aprendizaje, los cuales pueden clasificarse en problemas supervisados y no supervisados, como se abordará en el siguiente apartado.

4. Minería de datos: Constituye la etapa clave en donde se *aprende* de los datos. Es en esta etapa en donde se hace uso de métodos basados en aprendizaje automático, métodos que permiten que las computadoras aprendan de los datos de manera automática. Esta etapa incluye la selección de métodos que se utilizarán para el proceso. Implica decidir qué modelos y parámetros son los más apropiados para el problema en estudio.

5. Interpretación y evaluación´de resultados: Esta etapa implica tareas tales como la identificación de patrones y reglas de agrupamiento o clasificación. También implica posibles iteraciones adicionales a cualquiera de los pasos. Este paso también involucra la búsqueda de la mejor forma de representación y visualización de los patrones y modelos extraídos.

Un esquema de estos pasos se presenta en la **Figura 7.**

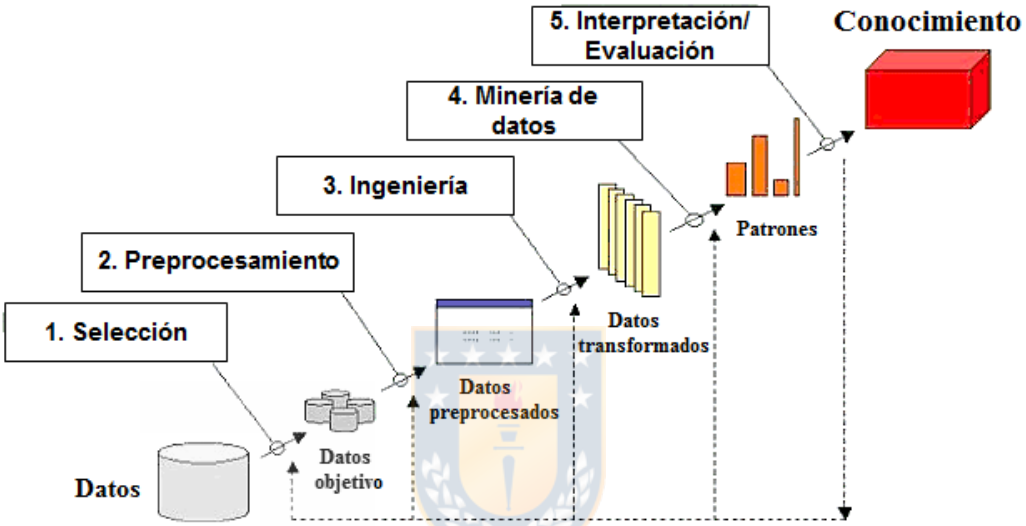


Figura 7. **Etapas del proceso de KDD.**

1.4.1. Problemas de aprendizaje y sus aplicaciones en bioinformática

Los problemas de aprendizaje propios de la minería de datos se pueden clasificar en problemas de agrupamiento, estimación y clasificación (Raza, 2012). Dentro de los problemas de estimación encontramos la estimación de parámetros y la predicción de valores futuros. Los problemas de clasificación implican la determinación de la pertenencia de un elemento a un grupo o clase dentro de un conjunto de clases predefinidas. La clasificación de una proteína transportadora dentro de un grupo o clase funcional es un problema de clasificación.

En los últimos años, se han desarrollado diversos métodos de clasificación de proteínas transportadoras en clases funcionales mediante métodos de aprendizaje automático, a partir de información de secuencia (Mishra et al., 2014). Sin embargo, estos métodos implican conocer *a priori* las clases o grupos funcionales de las proteínas a partir de las que se construye el modelo, es decir, los algoritmos de clasificación se levantan a partir de datos cuya etiqueta de clase es conocida (aprendizaje supervisado). Algunas de las limitaciones a la hora de desarrollar estos algoritmos son las siguientes:

- Dificultad en la definición de clases a partir de evidencia experimental, por su escasez. Brown & Sjölander (2006) analizaron las anotaciones funcionales de más de 300,000 proteínas alojadas en UniProt, encontrando que sólo el 3% de ellas presentaban evidencia experimental.

- Dificultad en la definición de clases que reflejen niveles detallados de funcionalidad, debido a la amplia especificidad de los transportadores y los distintos tipos y niveles de detalle de las anotaciones funcionales. Los estudios suelen dejar fuera del modelo transportadores con más de un sustrato anotado, y/o no consideran niveles de especificidad de sustrato (p.e. alta afinidad, baja afinidad), definiendo clases funcionales más generales. En efecto, los trabajos orientados a desarrollar clasificadores de proteínas transportadoras, a la fecha, suelen considerar de tres a siete clases de sustratos, incluyendo comúnmente iones metálicos, aminoácidos y azúcares; sin existir consenso en el número de clases. Cabe destacar que los azúcares siempre se consideran una gran clase, sin definirse subclases en esta categoría en ninguno de los estudios consultados.

1.4.2. Análisis cluster de regiones funcionales

En el caso de desconocerse los grupos o clases en los que se agrupa naturalmente un conjunto de datos, nos encontramos frente a un problema de agrupamiento o *clustering.* Definir y caracterizar estos grupos es el objetivo del aprendizaje no supervisado. Recientemente, se han levantado nuevas propuestas frente al problema de la función molecular de proteínas, las cuales buscan agrupar proteínas utilizando información de secuencia, estructura u otras métricas, correlacionando estos grupos con algún nivel de función molecular (Fetrow & Babbitt, 2018). El objetivo de estos enfoques no es predecir la función *per se*, sino agrupar las proteínas en grupos funcionales relevantes, contribuyendo posteriormente a dilucidar patrones que explican funciones comunes y permitiendo la anotación de genes desde el supuesto de que se puede transferir información funcional detallada entre miembros de un mismo grupo. Una vez definidos los grupos, éstos pueden utilizarse para entrenar algoritmos de clasificación y para identificar las características que mayormente contribuyen a la discriminación entre un grupo y otro, que en proteínas suelen corresponder a patrones de aminoácidos. Algunos de los métodos empleados para estos fines se abordan en el apartado 1.4.3.

Comprendiendo que la etapa de selección de características consiste en la selección de un subconjunto de los atributos de un conjunto de datos con el fin de filtrar ruido, atributos redundantes o irrelevantes para el problema de aprendizaje, cabe preguntarse qué características deben seleccionarse a la hora de agrupar proteínas según funcionalidad. Para problemas de aprendizaje supervisado, existe una gran variedad de métodos capaces de seleccionar las características que contribuyen en mayor medida al problema de clasificación (Dash & Liu, 1997). Sin embargo, no existe un '*gold standard*' para la selección de características en problemas de agrupamiento, ésta dependerá de la aplicación particular que se esté trabajando. A la hora de agrupar proteínas según funcionalidad, trabajos recientes han plantado enfoques de agrupamiento basados en los conceptos de perfilación de sitios activos o de regiones funcionales, es decir, desarrollo de

agrupamiento con foco en residuos funcionales, en donde la selección y diseño de características desde estos residuos se utilizan para optimizar la separación de grupos y posterior caracterización de sitios activos de subfamilias que comparten características funcionales (Boari de Lima et al., 2016; de Melo-Minardi et al., 2010; Fetrow & Babbitt, 2018; Landgraf et al., 2001). En este sentido, la selección de características está ligada a identificación de regiones funcionales en proteínas. Quizá la forma más intuitiva para abordar este problema corresponde a la selección de las posiciones de un MSA correspondientes a residuos que se encuentran a una cierta distancia de un sitio activo conocido, a partir del análisis tridimensional una estructura representativa de la familia de proteínas en estudio, que se encuentre co-cristalizada con su sustrato. Ésta fue la aproximación utilizada por Landgraf (2001), el cual seleccionó los residuos que se encontraban a menos de 10 A del sitio activo definido desde estructuras representativas de 35 familias de proteínas, y comparó las relaciones entre secuencias dispuestas en los filogramas globales y regionales obtenidos mediante métodos basados en distancias, como se muestra en la **Figura 8.** En este trabajo se decidió emplear esta aproximación, identificando residuos de reconocimiento de sustrato mediante revisión bibliográfica y análisis de estructuras tridimensionales representativas de la familia en estudio.

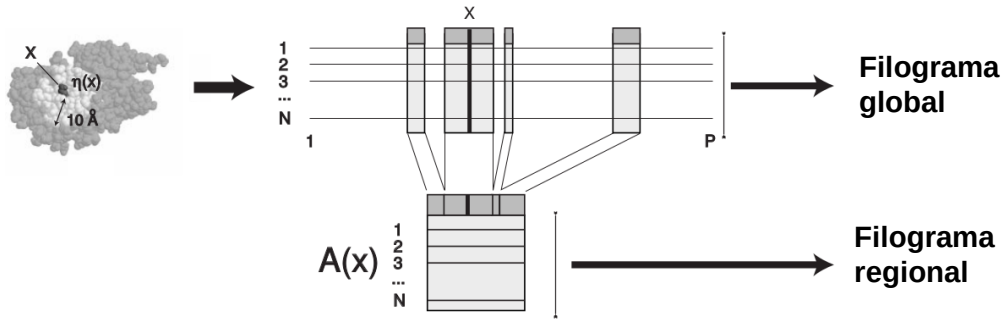


Figura 8. **Análisis cluster de regiones funcionales.** Adaptado de Landgraf et al., (2001). El esquema de trabajo consiste en la construcción de un MSA regional a partir de un MSA representativo de una familia de proteínas, seleccionando las posiciones que se encuentran a menos de 10 A del sitio activo definido por una estructura tridimensional representativa de la familia, co-cristalizada con sustrato o ligando. Posteriormente se comparan los agrupamientos obtenidos desde los filogramas globales y regionales.

1.4.3. Caracterización de familias de proteínas desde secuencias

La caracterización de una familia de proteínas, a nivel de secuencia, permite guiar la formulación de hipótesis explicativas de la diversidad funcional que presenta, así como el desarrollo de futuras herramientas de predicción de sitios funcionales (Kalinina et al., 2009; Nemoto et al., 2013). Esta tarea suele desarrollarse mediante el análisis de la información codificada en un MSA de la familia, desde el cual, operacionalmente, se pueden identificar y caracterizar tres tipos de residuos potencialmente determinantes de los distintos aspectos funcionales inherentes a la actividad biológica en estudio, como se esquematiza en la **Figura 9** (De Juan et al., 2013; Teppa et al., 2012). A partir de la identificación y análisis de estos residuos, se puede inferir residuos de interfaz con mayor importancia funcional, así como identificar redes de interacciones físicas entre residuos, información que puede guiar el análisis de las relaciones estructura-función. A continuación se abordan consideraciones teóricas y metodológicas en relación cada uno de ellos.

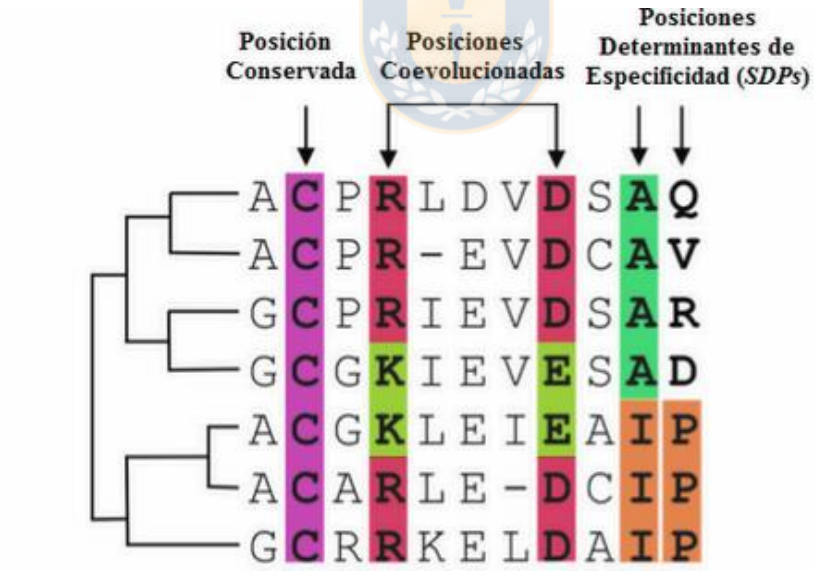


Figura 9. **Tipos de residuos con importancia funcional identificables a partir de MSA de una familia de proteínas (Teppa et al., 2012).**

Posiciones conservadas en toda la familia: Suelen representar residuos con un rol funcional o estructural conservado en toda la familia de proteínas. Las posiciones totalmente conservadas tienden a forman parte del núcleo estructural de la proteína y también se encuentran en regiones funcionales, como en sitios de interacción y sitios catalíticos, lo que se puede discriminar mediante análisis topológicos-estructurales. Dependiendo de su grado de conservación, se puede inferir su mayor o menor importancia en la determinación de la diversidad funcional de una familia. Suelen identificarse mediante análisis de entropía de Shannon (Shannon, 1948). El análisis de entropía de Shannon (H) para cada posición i de un MSA se define según la siguiente ecuación (Cover & Thomas, 1991):

$$H_i = -\sum_x p_i(x)\ log_2\ p_i(x)$$

Donde Pi(x) es la fracción de residuos del aminoácido tipo x presentes en la posición i, y x es el número de tipos de aminoácidos (20). H varía de 0 (solo un residuo presente en esa posición) a 4.322 (los 20 residuos están igualmente representados en esa posición). Se requiere un número mínimo de secuencias (~ 100) para que H describa la diversidad de una familia de proteínas. Las posiciones con H mayor a 2.0 se suelen considerar variables, mientras que aquellas con H menor a 2 se consideran conservadas. Las posiciones altamente conservadas son aquellas con H menor a 1.0 (Litwin & Jores, 1992). Cabe destacar que los valores máximos de entropía dependen del valor de la base logarítmica que se utilice para realizar el cálculo. Si bien suele utilizarse la escala logarítmica de base 2, si se determina la base igual al número de clases representadas en la distribución, por ejemplo, igual a 20 (representando las 20 clases de aminoácidos), la entropía máxima tendrá un valor igual a 1.

Se han desarrollado distintos métodos alternativos de evaluación de conservación derivados del análisis de entropía de Shannon, con el fin de incluir ciertas consideraciones o superar algunas limitaciones (véase Capra & Singh, 2007). Una de las limitantes de esta métrica es la suposición de una posición k del

alineamiento sin ninguna restricción evolutiva exhibirá una distribución de probabilidad equitativa para todos los aminoácidos. En respuesta a esta limitante se levantaron métodos que consideran la inclusión de distribuciones de fondo mediante cálculos de divergencia de Kullback-Leibler, también conocida como entropía relativa (K. Wang & Samudrala, 2006). La entropía relativa (RE) mide la diferencia en contenido de información entre ambas distribuciones de aminoácidos, según la ecuación:

$$rE_i^{A/B} = \sum_x p_i^A(x)\ log\ \frac{p_i^A(x)}{p_i^B(x)}$$

Donde pi(x) A y pi(x) B son las probabilidades observadas del aminoácido tipo x en la posición i del alineamientos A y en la distribución de fondo B, respectivamente. Cabe tener presente esta métrica, ya que constituyen la base teórica para métodos abordados posteriormente.

Posiciones correlacionadas: La presión evolutiva asociada a la mantención de una proteína estructuralmente estable y funcionalmente activa da lugar a mutaciones correlacionadas entre pares de residuos. Estas posiciones, aunque separadas en la estructura primaria, pueden encontrarse en contacto en la estructura terciaria, bien porque mantienen interacciones que dan estabilidad estructural o bien confieren funcionalidad, por ejemplo, co-participando en el reconocimiento de sustratos. Las principales metodologías para inferir residuos que coevolucionan se basan en análisis de información mutua, análisis de acoplamiento directo y estadístico (Colell et al., 2018; De Juan et al., 2013). La información mutua (MI) es una medida de reducción de la incertidumbre de una variable aleatoria, X, debido al conocimiento del valor de otra variable aleatoria Y, calculada según la siguiente ecuación (Cover & Thomas, 1991):

$$\mathrm{MI}(X,Y) = H(X) - H(X|Y)$$

Donde H(X) es la entropía de X y H(X|Y) es la entropía condicional de X, dado Y. El valor de MI es simétrico, es decir M I(X, Y) = MI(Y, X). En un MSA, la MI entre dos posiciones ij de un alineamiento se calcula según:

$$MI_{ij} = \sum_{xi,xj} p_{ij}(xi,xj)\ ln\ \frac{p_{ij}(xi,xj)}{p_i(xi)p_j(xj)}$$

Donde pij(xi,xj) corresponde a la distribución conjunta de los pares de aminoácidos xi,xj posibles; mientras que pi(xi) y pj(xj) corresponden a las distribuciones observadas del aminoácido xi en la posición i y del aminoácido xj en la posición j, respectivamente; por lo que MI mide la divergencia Kullback-Leibler de la distribución conjunta y el término factorizado pi(xi) pj(xj). Este valor refleja el grado en que el conocimiento del aminoácido en una posición nos permite predecir la identidad del aminoácido en la otra posición. Si X e Y son independientes, su MI es cero. De lo contrario, la información mutua es positiva, donde valores mayores indican una mayor interdependencia. El rango de valores que podrá tomar la MI entre dos variables está determinada por:

$$0 \leq \mathrm{MI}(X,Y) \leq \min\{H(X),H(Y)\}.$$

Donde min{H(X),H(Y)} es el valor mínimo de entropía conjunta de X e Y, y estará determinado por la base logarítmica utilizada para el cálculo de entropía. Al usar una base logarítmica=20 (representando las 20 clases de aminoácidos), los valores de entropía conjunta podrán tener valores entre 0 y 2 (Dunn et al., 2008). Si bien la MI es una medida intuitiva para identificar sitios de mutaciones correlacionadas, la correlación observada entre dos posiciones de un alineamiento i y j (Cij) surge de varias causas subyacentes separadas, que pueden expresarse mediante un modelo lineal de la siguiente forma (Ackerman et al., 2012; Atchley et al., 2000):

$$C_{ij} = C_{\text{filogenia}} + C_{\text{estructura}} + C_{\text{función}} + C_{\text{interacciones}} + C_{\text{estocástico}}$$

En breve, C estructura y C función representan la correlación que se origina de restricciones necesarias para la estabilidad estructural y actividad molecular, y corresponden a las señales más importantes que los análisis de coevolución intentan extraer del MSA. C filogenia corresponde a la correlación que se origina de relaciones filogenéticas entre secuencias homólogas cercanas. En alineamientos biológicamente significativos, no hay dos posiciones que puedan ser totalmente independientes. Incluso cuando las mutaciones en los dos sitios ocurren con total independencia, la historia evolutiva compartida contribuye a valores positivos de MI entre estos residuos. Esta fuente puede ser limitada en cierto grado al excluir secuencias muy similares de especies estrechamente relacionadas, pero no puede ser eliminado. Estos primeros tres componentes no son independientes entre sí, por lo que el componente C interacciones del modelo cuantifica la interdependencia entre ellos. Finalmente, C estocástico corresponde a las correlaciones que no pueden ser explicadas por los otros componentes del modelo, y se asocian a correlaciones nacidas del número limitado de secuencias analizadas y a factores entrópicos del alineamiento: posiciones con mayor variabilidad o entropía tienden a tener valores de MI más altos, tanto aleatorios como no aleatorios. También, la MI se sobreestima con tamaños de muestra pequeños (Dunn et al., 2008; Martin et al., 2005). A su vez, si consideramos una restricción evolutiva de una posición del alineamiento debido a que se encuentra en contacto, es decir, interaccionando físicamente con un residuo en otra posición, hablamos de una interacción o acoplamiento directo. Sin embargo, la correlación observada entre dos residuos, en este contexto, puede verse aumentada por interacciones indirectas, es decir, un residuo i interacciona con un residuo j, que a su vez interacciona con un residuo k. La correlación entre i y k se verá aumentada por el acoplamiento indirecto, mediado por j (Lunt et al., 2010). Así, es deseable poder discriminar entre la MI compuesta por C estructura y C función (MI-*sf*) y el resto de componentes, a la cual se le llama MI de fondo (MI-*b*), así como poder

discriminar entre acoplamientos directos e indirectos Para ello se han desarrollado una serie de correcciones y métodos aplicables a las matrices de MI calculadas desde alineamientos, y se encuentran disponibles para su uso en el programa ProDy (Bakan et al., 2014). Algunas de las principales se resumen en la **Tabla 3,** las cuales han demostrado mejorar la capacidad para discriminar entre MI-sf y MI-b, y mejorar la predicción de contactos funcionales entre residuos en la estructura tridimensional eliminando factores de acoplamiento indirecto (De Juan et al., 2013).

Tabla 3. Correcciones de matrices de MI

Corrección	Ecuación	Comentarios
Corrección por producto promedio (APC)	$MIc_{ij} = MI_{ij} - APC$ $APC = \frac{MI(i,\bar{x})MI(j,\bar{x})}{\overline{MI}}$	APC y ASC permiten estimar MI-b, por lo que su sustracción permite obtener MI-sf. Para su discriminación se requiere un mínimo de 125 secuencias que no sean altamente similares, para que el efecto por tamaño pequeño de muestra sea despreciable. En general, los métodos APC y ASC muestran buen rendimiento, siendo APC el que muestra mayor correspondencia con los mapas de contacto (Dunn et al., 2008).
Corrección por suma media (ASC)	$MIc_{ij} = MI_{ij} - ASC$ $ASC = MI(i,\bar{x}) + MI(j,\bar{x}) - \overline{MI}$	
Normalización por entropía conjunta mínima	$MIc_{ij} = \frac{MI_{ij}}{\min\{H(X,Y)\}}$	Estas normalizaciones permiten remover la influencia de entropía discutida. Los valores de min{H(X,Y)} poseen la mayor correlación con los valores de MI, sin embargo, las correcciones por entropía conjunta y por suma de entropía poseen mayor poder discriminante. Sin embargo, la normalización por H(X,Y) entrega valores altos para pares de columnas con mutaciones únicas. Para sitios donde varias mutaciones de aminoácidos son igualmente viables como mutaciones compensatorias, estos sitios podrían ser discriminados negativamente por esta normalización (Martin et al., 2005).
Normalización por entropía conjunta	$MIc_{ij} = \frac{MIc_{ij}}{H(X,Y)}$	
Normalización por suma de entropías	$MIc_{ij} = \frac{MIc_{ij}}{H(X) + H(Y)}$	

Tabla 3. (Continuación).

Corrección	Ecuación	Comentarios
Análisis de acoplamiento directo (DCA)	$DI_{ij} = \sum_{xi,xj} p_{ij}^{(dir)}(xi,xj)\ ln\ \frac{p_{ij}^{(dir)}(xi,xj)}{p_i(xi)p_j(xj)}$	El DCA calcula la MI para cada par de posiciones ij, selecciona residuos con mayor MI y corrije sus valores de correlación mediante un análisis estadístico iterativo basado en conocimiento empírico que permite eliminar los efectos indirectos de correlación entre ambas posiciones, hasta convergencia, obteniendo un valor de distribución conjunta corregido $p_{ij}^{(dir)}$. Finalmente, se calcula la matriz de información directa, DI ij, de manera análoga a MI (Lunt et al., 2010).

Los análisis de entropía y MI pueden ayudar a caracterizar mejor un sitio activo definido mediante criterios estructurales y a centrar el análisis en los residuos que mayormente puedan contribuir a la diversidad funcional de la familia en estudio. Sin embargo, estos análisis no son capaces de identificar redes de interacciones de residuos correlacionados.

Esta última tarea es abordada exitosamente por el análisis de acoplamiento estadístico (SCA) que, a diferencia del DCA, no busca elimiar los acoplamientos indirectos, sino identificar redes independientes de residuos acoplados entre sí. El análisis se basa, de manera similar a los análisis de MI, en la construcción de una matriz simétrica de correlación entre los distintos pares de posiciones de un alineamiento ponderada por conservación, midiendo la correlación entre dos sitios, Cij, y siguiendo la misma nomenclatura empleada con anterioridad, según la siguiente ecuación (Reynolds et al., 2013):

$$SCA_{ij} = \sum_{xi,xj} \sqrt{\emptyset_i \emptyset_j [p_{ij}(xi,xj) - p_i(xi)p_j(xj)]^2}$$

Donde $\emptyset_i$ y $\emptyset_j$ representan funciones de ponderación por grado de conservación de las posiciones i y j, respectivamente.

Posterior a su cómputo, el análisis aborda el problema de discriminar entre las correlaciones con sentido biológico y las correlaciones de fondo nacidas del número limitado de secuencias analizadas mediante la descomposición espectral de la matriz de correlación asociada en sus valores propios, los cuales representan grupos de residuos correlacionados. La creación de matrices de correlación aleatorias equivalentes y su posterior descomposición en valores propios permite identificar un valor de corte para discriminar entre valores propios nacidos del ruido estadístico y los de sentido biológico. La identificación y análisis de los residuos que más contribuyen a los vectores propios asociados permiten la identificación de redes de residuos acoplados estadísticamente.

Este método ha demostrado ser eficaz en la identificación de redes de residuos conectados físicamente a lo largo de la estructura tridimensional y asociados a funciones moleculares específicas, llamados sectores proteicos por los autores, como se muestra en la **Figura 10**. Se ha visto que sólo alrededor del 20% de los aminoácidos de una proteína conforman sectores (Halabi et al., 2009; McLaughlin Jr et al., 2012; Reynolds et al., 2011). Su implementación se encuentra disponible en el programa pySCA (Rivoire et al., 2016) (http://reynoldsk.github.io/pySCA/).

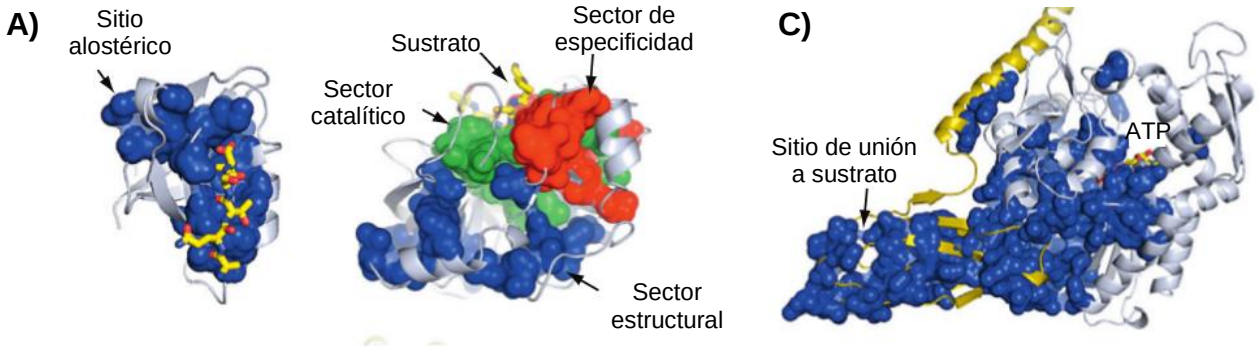


Figura 10. Sectores funcionales en tres familias de proteínas definidos por SCA (Reynolds et al., 2013). (A) Un solo sector en el dominio PDZ conecta el sitio de unión del ligando a un sitio alostérico distante. (B) Tres sectores cuasi-independientes en la familia de serina proteasas S1A asociados a funciones diferenciales. (C) Un solo sector en la familia de chaperonas Hsp70 conecta funcionalmente el sitio de unión de ATP y el sitio de unión al ligando, ambos en dominios estructurales diferentes.

Los análisis anteriores permiten caracterizar familias de proteínas y sitios funcionales a nivel general. Sin embargo, una vez se han identificado grupos funcionales dentro de una familia, es deseable identificar los residuos diferenciales entre los grupos, es decir, posiciones que se conservan sólo dentro de las subfamilias o grupos particulares. Estas posiciones, si se conoce su importancia funcional en relación a la actividad bioquímica en el contexto global de la familia, se infiere que son determinantes de las características funcionales específicas de esa subfamilia, por lo que en literatura se conocen como posiciones determinantes de especificidad (*Specificity-Determinant Positions, SDPs)* (Rausell et al., 2010), y se esquematizan en la **Figura *9.***

Para su identificación, si bien la mayoría de las metodologías requieren una definición previa de las subfamilias (aprendizaje supervisado), el análisis de comportamiento mutaciona (del Sol Mesa et al., 2003) permite una identificación de potenciales SDPs sin una definición de grupos previa. Este método se basa en la comparación entre el comportamiento mutacional de cada posición k del alineamiento y el comportamiento mutacional global del alineamiento, con base en el supuesto de que las SDPs tendrán un comportamiento mutacional similar al comportamiento mutacional de toda la familia. Para esto se construyen dos matrices, M1 y M2, como se muestra en la **Figura 11**. La primera matriz, M1, contiene la cuantificación de la similaridad entre aminoácidos de los pares de proteínas para una posición k del alineamiento, la cual se cuantifica mediante la métrica de McLachlan (aunque pueden usarse métricas definidas por el usuario). La segunda matriz, M2, corresponde a una matriz equivalente que cuantifica las similitudes funcionales entre los pares de proteínas correspondientes. La similitud funcional se establece mediante la cuantificación de la similitud de secuencia de los pares de proteínas implícitas en el alineamiento, descansando en la idea de que el escenario de evolución divergente de la función es representado por las similaridades implícitas en el MSA. Sin embargo, si se posee información empírica de la función de las proteínas representadas, se puede entregar una matriz de similitud funcional externa. Por ejemplo, se puede utilizar la similitud química entre

sustratos medidas mediante el coeficiente de Tanimoto. Si no existe información de similitud funcional cuantificada, se pueden utilizar los valores 0 para proteínas diferentes y 1 para proteínas definidas como similares.

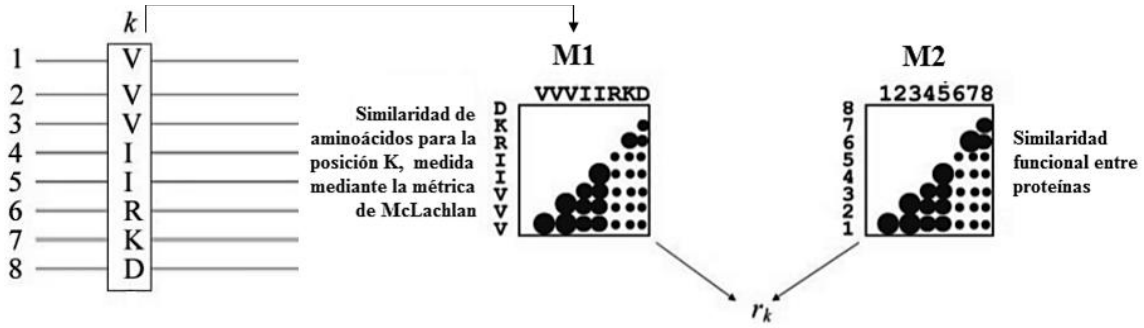


Figura 11. **Análisis de comportamiento mutacional para la detección de sitios funcionales.**

Luego, ambas matrices se comparan mediante un criterio de correlación por rangos no paramétrico, calculando el coeficiente de correlación por rangos de Spearman, rk, entre cada posición k del MSA y la similaridad funcional según:

$$r_k = \frac{\sum_{i,j}(A'_{ijk} - \bar{A}') \cdot (F'_{ij} - \bar{F}')}{\sqrt{\sum_{i,j}(A'_{ijk} - \bar{A}')^2} \cdot \sqrt{\sum_{i,j}(F'_{ij} - \bar{F}')^2}},$$

Donde A'ijk es la similitud entre los aminoácidos de las proteínas i y j en la posición k; F'ij es la similitud funcional entre estas proteínas, A' y F' son los valores de orden o posición respectivos y Ā', F̄' son los valores promedio de las matrices M1 y M2 jerarquizadas por rango. El valor rk es, por lo tanto, una medida de importancia de un residuo dado k, donde los valores más altos corresponden a posiciones k cuyo comportamiento mutacional presenta mayores coeficientes de correlación con cambios en la función.

En cuanto a los métodos supervisados para detección de SDPs asociados a cada subfamilia, destacan los métodos basados en análisis multivariable y en análisis de entropía y conservación. Ejemplo del primero y uno de los que ha demostrado ser de los más efectivos para detectar residuos que funcionalmente distinguen a una subfamilia de otra es S3Det (Rausell et al., 2010). Este método se basa en análisis de correspondencia múltiple (MCA), que es conceptualmente equivalente al análisis de componentes principales (PCA) pero es más adecuado para tratar datos categóricos, como identidades de aminoácidos. S3det representa cada aminoácido de cada proteína de manera binaria. Esta información se integra a un espacio vectorial, integrando también la información de clase mediante una matriz de clasificación funcional entregada por el usuario. Luego, reduce los datos a las dimensiones más informativas mediante MCA, para posteriormente identificar aquellos residuos asociados con las subfamilias en el espacio de posición equivalente. Si no se entrega información de clase, S3Det realiza un análisis de agrupamiento sobre el espacio vectorial reducido de secuencias, utilizando el algoritmo *k-means* para identificar potenciales subfamilias, sobre las cuales realiza el análisis.

Ejemplo del segundo caso destacan los métodos tipo *Sequence Harmony* (Pirovano et al., 2006), método basado en la cuantificación de entropía relativa entre distintas subfamilias para la selección de características. Para que un aminoácido sea totalmente discriminante entre subfamilias, debe estar presente en la subfamilia A y ausente en la subfamilia B o viceversa. Usando la ecuación de entropía relativa presentada anteriormente, la ausencia total de un aminoácido en una distribución y su presencia en la otra genera resultados matemáticos no deseados (entiéndase división por cero). Para solucionar esto, se introduce el concepto de armonía de secuencia o *Sequence Harmony (SH)*, que mide la entropía relativa entre la distribución de una subfamilia (A) y la distribución conjunta de ambas subfamilias (A + B) según la ecuación:

$$SH_i^{A/B} = \sum_x p_i^A(x)\ log\ \frac{p_i^A(x)}{p_i^A(x) + p_i^B(x)}$$

En general SH A/B /= SH B/A. Para corregir esto, se calcula el promedio de ambas entropías relativas para una posición dada i, según la ecuación:

$$SH_i = \frac{1}{2}(SH_i^{\frac{A}{B}} + SH_i^{\frac{B}{A}})$$

Distribuciones idénticas en una posición tienen armonía máxima con valor 1, por lo que los sitios diferenciales entre subfamilias poseen valores de armonía bajos. Este método implica el análisis de dos subfamilias. Su generalización para el análisis de varias subfamilias es integrado en el programa Multi-Harmony (Brandt et al., 2010) (http://ibi.vu.nl/programs/shmrwww/).

Cabe destacar que estos métodos permiten identificar los residuos que más contribuyen a la topología o agrupación observada tanto por filogenias globales como por métodos de agrupamiento basados en regiones funcionales, lo que permite analizar si existe correspondencia entre los residuos determinantes de filogenia (*tree determinants*, del Sol Mesa et al., 2003) y determinantes funcionales empíricos, constituyendo una vía para evaluar en qué medida la filogenia refleja una relación funcional.

1.4 *Caracterización funcional de la familia SP*

La familia SP se encuentra representada por 139 miembros alojados en la base de datos TCDB bajo el código TC.2.A.1.1 (http://www.tcdb.org/), existiendo 290 entradas en SwissProt y más de 90.000 entradas en TrEmbl asociadas a este mismo código. Entradas asociadas a la familia en las bases de datos Pfam, Interpro y Prosite se presentan en la **Tabla 4.**

La Familia SP se encuentra evolutivamente relacionada con la superfamilia de facilitadores principales (MFS, *TC 2.A.1*), constituida a partir de los estudios computacionales que revelaron la relación de homología existente entre transportadores de azúcares de bacterias y mamíferos, ejemplificados por el transportador de glucosa humano GLUT1 y los cotransportadores azúcar:protón de *E.coli* GalP, XylE y AraE (Henderson & Maiden, 1990; Maiden et al., 1987). Si bien los primeros miembros asociados a la MFS corresponden a transportadores de la familia SP, estudios posteriores determinaron que transportadores con mecanismos de transporte variados y de otros diversos metabolitos tales como fármacos, neurotransmisores, aminoácidos, nucleósidos, iones orgánicos e inorgánicos, entre muchos otros, también pertenecían a la MFS (Pao et al., 1998), existiendo miembros que funcionan como receptores además de o en lugar de sus funciones de transporte (Saier Jr et al., 2015). Actualmente, corresponde a la superfamilia de portadores más grande y diversa conocida, abarcando 89 familias en la base de datos TCDB, y 15-16 familias del sistema SLC (Perland & Fredriksson, 2017; Yan, 2015).

Tabla 4. Dominios y motivos en bases de datos asociados a la familia SP

Base de datos	ID de motivo/dominio	Descripción
Pfam	PF00083	Perfil HMM asociado a transporte de azúcares y otros metabolitos conformado por 85332 secuencias. Códigos TC asociados: 2.A.1, 2.A.2.
Interpro	IPR005828	Dominio MFS asociada al transporte de azúcares, polioles y ácidos orgánicos en procariontes y eucariontes. Conformada por 432 secuencias revisadas y más de 220.000 secuencias no revisadas.
	IPR003663	Dominio característico de subfamilia de la entrada anterior, asociada al transporte de azúcares e inositoles, conformada por 251 secuencias revisadas y más de 99.000 secuencias no revisadas.
	IPR005829	Motivo conservado en miembros de la familia SP, pero también presente en otros transportadores tales como miembros de la familia de transportadores de cationes y aniones orgánicos SLC22 y en bombas de eflujo de fármacos.
Prosite	PS00216	Motivo de secuencia asociada a transportadores de azúcares, identificada en 460 secuencias alojadas en SwissProt: 228 verdaderos positivos, 145 falsos positivos, 126 falsos negativos.
	PS00217	Motivo de secuencia asociada a transportadores de azúcares, identificada en 442 secuencias alojadas en SwissProt: 244 verdaderos positivos, 195 falsos positivos, 139 falsos negativos.
Prints	PR00171	Firma de secuencia conformada por 5 motivos asociada a transportadores de azúcares, presente en transportadores GLUT 1-5.
	PR00172	Firma de secuencia conformada por 2 motivos asociada a transportadores de azúcares, presente de manera más ubicua en toda la familia.

Los miembros de la MFS poseen una estructura tridimensional común altamente conservada. Estudios tempranos de predicción de estructura secundaria e hidrofobicidad de las secuencias de los miembros de la MFS a principios de los 90 ya predecían una topología común (Goswitz & Brooker, 1995), caracterizada por la presencia de, generalmente, 12 segmentos transmembrana (TM) y la cual fue confirmada posteriormente tras la aparición de las primeras estructuras cristalográficas, como se muestra en la **Figura 12.** Así, independientemente de bajas identidades de secuencia y diferencias en mecanismos de transporte y selectividad, las estructuras cristalográficas de sus miembros exhiben este plegamiento tridimensional común, conocido como plegamiento MFS (Yan, 2013). Esto queda ejemplificado mediante la comparación de las estructuras de la lactosa permeasa (LacY; Abramson et al., 2004) y el transportador de glicerol-3-fosfato (GlpT; Huang et al., 2003), cuya identidad de secuencia es solo del 21% y sus mecanismos de acción difieren (contratransporte y cotransporte, respectivamente). A pesar de esto, sus estructuras son muy similares, y ambas son similares a la estructura general de baja resolución de OxlT, obtenido mediante microscopía electrónica (EM) (Hirai et al., 2003). En efecto, el alineamiento estructural entre ambas proteínas exhibe un Cα-RMSD entre las regiones estructuralmente conservadas de solo 3,7 Å (Vardy et al., 2004).

En el plegamiento MFS, las 12 hélices transmembrana (TM 1-12) se organizan en dos dominios estructurales similares, el dominio N-terminal (TM 1-6) y el dominio C-terminal (TM7-12) Estos dominios se pueden subdividir en dos repeticiones invertidas de tres hélices codificadas como hélices A, B y C (Quistgaard et al., 2016) (**Figura 12**). El sitio de unión y translocación de las moléculas transportadas se conforma en la interfaz de los dominios N y C terminales, en donde los TM 1, 4, 7 y 10 (hélices A) se posicionan hacia el centro del transportador, conformando el canal de transporte; mientras que las TMs 2, 5, 8, y 11 (hélices B) median la interfaz entre los dos dominios (Yan, 2017). A la fecha, se han logrado cristalizar más de 70 estructuras de transportadores MFS en presencia/ausencia de distintos sustratos/ligandos (https://blanco.biomol.uci.edu/mpstruc/) y en conformaciones

representativas de los tres estados esperables para un mecanismo de acceso alternante: exofacial, ocluido y endofacial, como se muestra en la **Figura 12.** Cabe destacar, sin embargo, que de las estructuras cristalizadas, existen conformaciones endo y exo faciales parcialmente ocluídas.

Las estructuras tridimensionales resueltas por cristalografía de rayos X pertenecientes a la familia SP listadas en la **Tabla 5** también comparten este plegamiento altamente conservado, reflejado en el alineamiento estructural presentado en la **Figura 13**. Cabe destacar que la familia SP exhibe, además del plegamiento común de 12 TMs, un dominio helicoidal intracelular formado por 5 regiones helicoidales (IC1-5), el cual ha mostrado ser esencial para la activación intracelular en algunos de sus miembros (Yan, 2015). El análisis de estas estructuras, en conjunto con diversos estudios experimentales y computacionales enfocados principalmente en GLUTs y el transportador XylE de *E.coli*, ha permitido identificar residuos claves en el proceso de unión y translocación de estos transportadores, identificándose residuos claves en las hélices A, B y también en regiones intracelulares La integración de estos datos ha permitido refinar el modelo de transporte asociado a esta familia, conocido como *Rocket switch* (ver **Figura 5**) y que involucra una transición rígida de los dominios N y C terminales entre los estados conformacionales descritos, al modelo *clamp and switch,* que identifica curvamientos de las hélices A tras la unión de sustrato, acercándose entre sí por interacciones de residuos claves que conforman la 'compuerta delgada' (*thin gate*) y dando origen a las conformaciones parcialmente ocluídas. Estas interacciones orquestan el acercamiento e interacción de las hélices B adyacentes, conformando la 'compuerta gruesa' (*thick gate*) y la transición al estado conformacional opuesto, completando un ciclo de transporte (Deng et al., 2015; Nomura et al., 2015; Quistgaard et al., 2016).

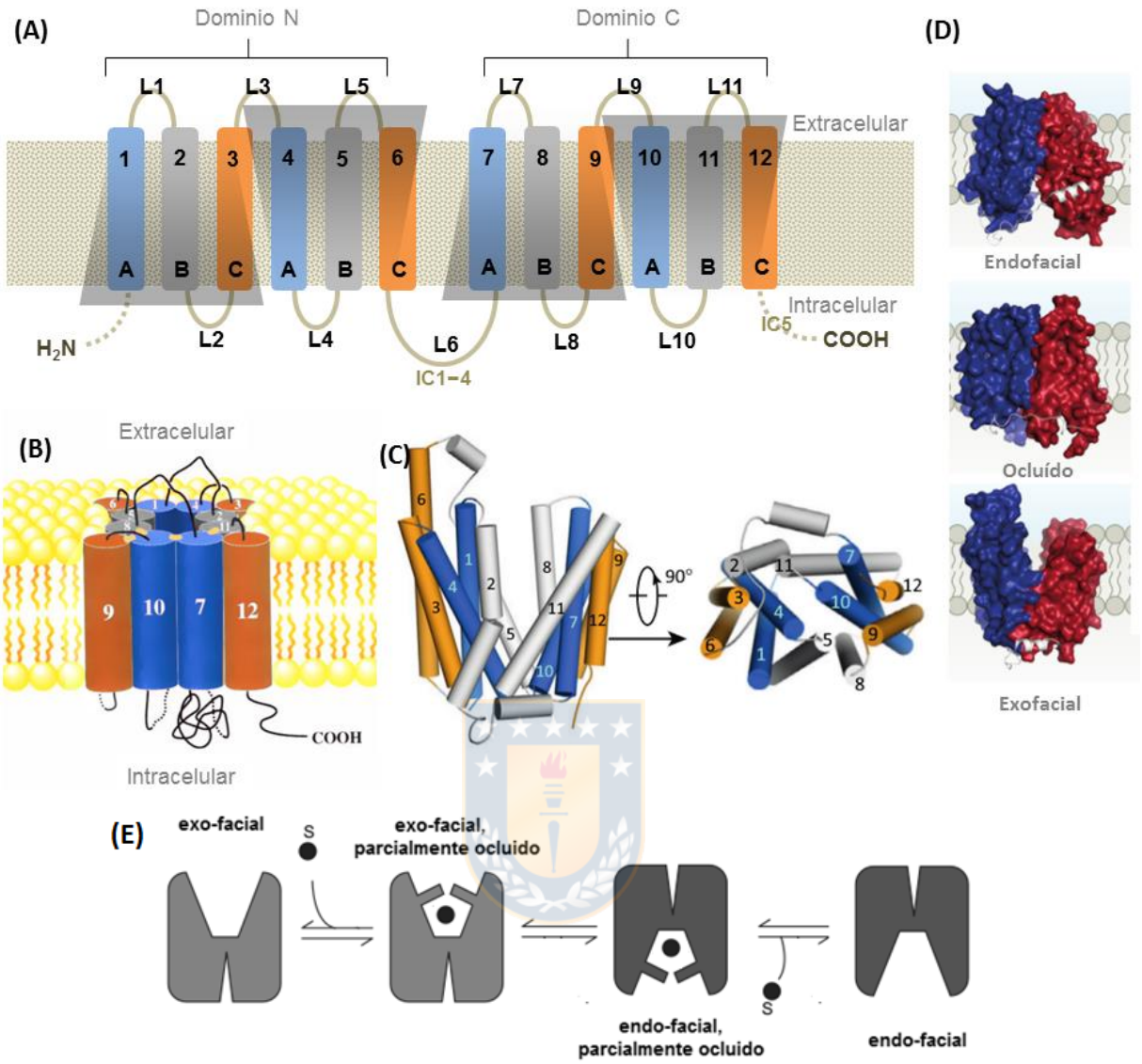


***Figura 12*. Plegamiento MFS.** (A) Topología MFS. Hélices A, B y C en azul, gris y naranja, respectivamente. (B) Topología predicha por Goswitz & Brooker (1995) a partir de análisis de secuencias. (C) Plegamiento MFS consenso a partir de estructuras resueltas por cristalografía de rayos X (Yan, 2013). (D) Estructuras de EmrD, LacY y FucP representativas de las conformaciones ocluidas, endo y exofaciales, respectivamente (Quistgaard et al., 2016). (E) Esquema mecanismo *clamp and switch.*

Tabla 5. Estructuras cristalográficas de la familia SP ǂ

Proteína	Organismo	PDBǂ	Mutante	Resolu-ción (A)	Conformación	Co-cristalizado
XylE	*E.coli*	4GBY	--	2.81	Exo-facial	D-Xilosa
		4GC0	--	2.60	Exo-facial	6-Br-6-deoxi-D-Glc
		4GBZ	--	2.89	Exo-facial	D-Glc
		4JA3	--	3.80	Endo-facial(o)*	Apo
		4JA4	--	4.20	Endo-facial	Apo
GlcP	*S.epidermidis*	4LDS	--	3.20	Endo-facial	Apo
GLUT1	*H.sapiens*	4PYP	N45T/E329Q	3.17	Endo-facial	b-NG
		5EQI	--	3.0	Endo-facial	CitoB
		5EQG	--	2.9	Endo-facial	Glut-i1
		5EQH	--	2.99	Endo-facial	Glut-i2
GLUT3	*H.sapiens*	4ZW9	N45T	1.50	Exo-facial(o)*	α,β-D-Glc.
		4ZWB	N45T	2.4	Exo-facial	Maltosa
		4ZWC	N45T	2.6	Exo-facial	Maltosa
GLUT5	*B.taurus*	4YB9	N51A	3.20	Endo-facial	Apo
	R.norvegicus	4YBQ	N50Y	3.27	Exo-facial	Fv
STP10	*A.thaliana*	6H7D	--	2.4	Exo-facial(o)*	D-Glc.

ǂ Se consideran estructuras depositadas en PDB hasta marzo, 2020.
* La letra 'o' hace referencia a conformaciones parcialmente ocluídas.

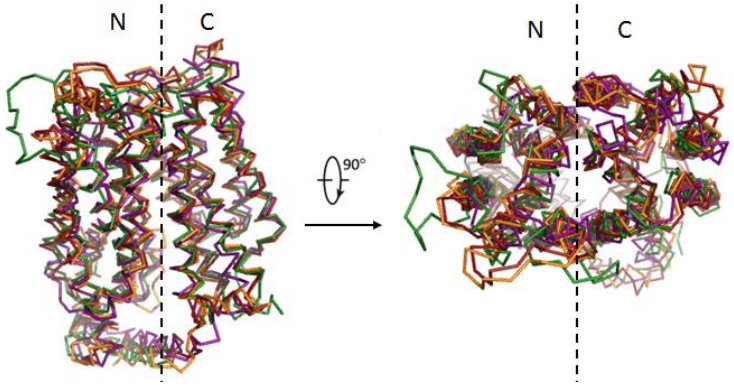
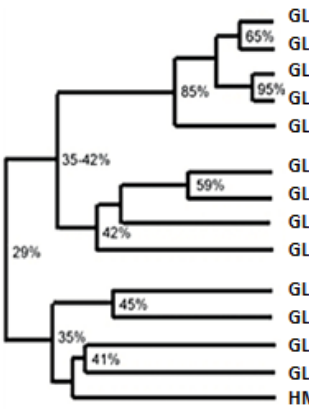


Figura 13. **Alineamiento estructural de transportadores Glut1, Glut5, GlcP y XylE *de H.sapiens, B.Taurus, S.epidermidis* y *E.coli.*** Se presentan en conformaciones endofaciales (códigos PDB 5EQI, 4YB9, 4LDS y 4JA4; colores rojo, naranja, morado y verde, respectivamente). Glut5, GlcP y XylE presentan un Cα-RMSD respectivo de 1.1, 2.4 y 1.0 A contra Glut1.

La familia es funcionalmente diversa, lo que queda evidenciado al analizar las propiedades de los transportadores GLUTs en humanos y ortólogos cercanos, representados por 14 isoformas (GLUT1-14) y clasificados en 3 clases según criterios de similaridad de secuencia. Su clasificación y principales características funcionales se resumen en la **Figura 14.** Destaca lo amplio y diverso de su selectividad: en general, estos transportadores pueden mediar el transporte de diferentes hexosas con distintas especificidades Algunos son específicos para otros compuestos como glucosamina, urato, xilosa, inositol y ácido ascórbico (Augustin, 2010; Mike Mueckler & Thorens, 2013; Reckzeh & Waldmann, 2019). Otros miembros de la familia encontrados en levaduras, plantas y parásitos protozoarios también exhiben diversos mecanismos de transporte y especificidades, encontrando transportadores asociados a transporte facilitado, a cotransporte azúcar:protón (como los transportadores de *E.coli* ya mencionados) y a contratransporte azúcar:protón, así como hexosas, pentosas, disacáridos, inositoles, polioles, quinatos y otros compuestos dentro de los sustratos identificados (Leandro et al., 2009; Palma et al., 2007)**.**



	Principal sitio expresión	Función fisiológica	Cinética*	Sustratos principales	
GLUT1	Eritrocitos	Captación basal	F, As, T+	Glucosa, galactosa	Clase I
GLUT4	Adipocitos, músculo, corazón	Respuesta a insulina	F, S, T-	Glucosamina, glucosa	
GLUT3	Neuronas	Captación neuronal	F, S, *s.i*	Glucosa, galactosa	
GLUT14	Testículos	*Desconocido*	- -	Glucosa, DHA	
GLUT2	Hígado, páncreas, intestino	Sensor de glucosa	F, *s.i, s.i*	Glucosa, galactosa	
GLU5	Testículos, intestino, músculo	Transporte de fructosa	F, *s.i, s.i*	Fructosa	Clase II
GLUT7	Intestino, testículos, próstata	*Desconocido*	F, *s.i, s.i*	Glucosa, fructosa	
GLUT9	Hígado, riñón	Homeostasis de urato	F, *s.i, s.i*	Urato	
GLUT11	Páncreas, riñón, placenta	*Desconocido*	F, *s.i, s.i*	Glucosa, fructosa	
GLUT6	Cerebro, médula, leucocitos	Transporte lisosomal	F, *s.i, s.i*	Glucosa	Clase III
GLUT8	Testículos, neuronas, adipocitos	Transporte de Trealosa	F, *s.i, s.i*	Trealosa, glucosa	
GLUT10	Hígado, páncreas	*Desconocido*	F, *s.i, s.i*	Glucosa	
GLUT12	Corazón, próstata	Respuesta a insulina	F, *s.i, s.i*	Glucosa	
HMIT	Cerebro	Transporte de inositol	Co, *s.i*, --	Inositol:H+	

Figura 14. **Clases de GLUTs y características funcionales.** Adaptado de Mueckler & Thorens, 2013; Reckzeh & Waldmann, 2019. (A) Filogenia con similaridad de secuencia entre isoformas obtenidas por ClustalW. (B) Propiedades funcionales. *Se indica mecanismo de transporte, presencia o ausencia de simetría, presencia o ausencia de trans-aceleración. F: Facilitador, Co: Cotransportador, S: Simetría, As: Asimetría. T+: Exhibe trans-aceleración. T-: No exhibe trans-aceleración. *s.i:* sin información.

1.6 *Planteamiento del problema y propuesta de trabajo*

Si bien la familia SP es una de las familias de transportadores más estudiadas y caracterizadas a lo largo de la historia, aún no se ha desarrollado un modelo integrador con capacidad explicativa y predictiva de la diversidad de especificidades presente en la familia. Lo primero, necesario para guiar el desarrollo de aplicaciones en medicina y biotecnología; lo segundo para apoyar la anotación de genes, pues muchos de los transportadores asociados a esta familia no se encuentran caracterizados a este nivel. De los 139 transportadores alojados en TCDB, un 22% no poseen sustratos anotados a partir de evidencia empírica, y a otros varios se les desconoce su preferencia por sustrato. Esto se amplía si consideramos la basta diversidad de homólogos identificados y alojados en bases de datos como UniProt y Pfam, en donde la mayoría posee anotaciones inferidas por métodos basados en homología. La amplia y variable especificidad exhibida por los GLUTs deja en evidencia que estos métodos difícilmente permiten transferir un nivel detallado de función molecular e inferir sus sustratos predominantes, limitando su utilidad. Esto también queda en evidencia al analizar los motivos y dominios asociados a la familia alojados en bases de datos como Prosite o Interpo (**Tabla 4**) y que corresponden en su mayoría a motivos con alto grado de conservación en la totalidad de la familia y, por tanto, no permiten discriminar especificidades.

Si bien existen estudios comparativos y experimentales que han logrado dilucidar el rol de residuos claves en la especificidad de sustrato de algunos miembros de la familia, ejemplificados por las investigaciones realizadas por Ferreira et al., 2019; Madej et al., 2014; A. Manolescu et al., 2005; A. R. Manolescu et al., 2007; C. y Wang et al., 2015; la mayoría se limita al estudio de GLUTs sin considerar o bien considerando un pequeño número de representantes del resto de transportadores en la familia, por lo que no existe un estudio integrador que permita ampliar la visión que se tiene sobre la problemática. Por todo lo anterior, en este trabajo se propone hacer uso de la información estructural y funcional alojada en la literatura y las bases de datos para definir grupos funcionales dentro de la familia SP,

evaluando métodos filogenéticos y de análisis clúster de regiones funcionales. Constituye, por tanto, un esfuerzo por generar conocimiento a partir de la integración y análisis de la información funcional levantada por diversas investigaciones. Un esquema del problema y propuesta de investigación se presenta en la **Figura 15.** Las preguntas que se busca responder en esta tesis son las siguientes:

- ¿Es posible identificar grupos funcionales relevantes dentro de la familia SP mediante aproximaciones filogenéticas?
- ¿En qué medida refleja la filogenia de la familia SP una relación filogenética-funcional?
- ¿Poseen las aproximaciones bioinformáticas basadas en agrupamiento de sitios funcionales un mejor rendimiento a la hora de definir grupos funcionales relevantes en la familia SP?

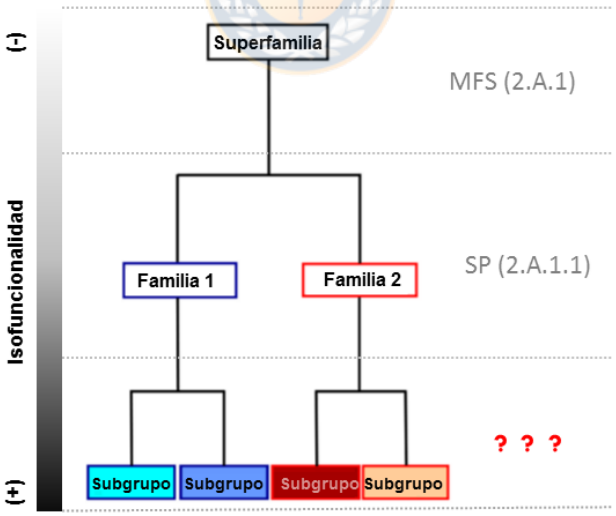


Figura 15. **Esquema del problema y propuesta de investigación.** La información estructural y funcional de diversos miembros levantada y alojada en bases de datos permite proponer la agrupación de los transportadores de la familia a partir de características claves para la selectividad y el transporte.

2 HIPÓTESIS

Considerando los siguientes antecedentes ya discutidos:

- Sólo una fracción de los aminoácidos de una proteína son determinantes para la función molecular en estudio.
- La divergencia específica asociada a tales residuos y, por lo tanto, a la función de interés, puede verse enmascarada dentro de la filogenia global de los genes en estudio.
- El plegamiento MFS, altamente conservado de la familia, presenta un poro de transporte conformado por un sitio de unión y translocación de sustrato central, cuya exposición se alterna a ambos lados de la membrana, según el modelo de transporte actualmente aceptado.

La hipótesis este trabajo es:

Los residuos involucrados en el reconocimiento de sustrato a lo largo del poro de transporte determinan la especificidad por sustrato, por lo que su agrupamiento permite definir grupos con mayor isofuncionalidad que los definidos mediante análisis filogenético global de la familia.

3 OBJETIVOS

Para responder a esta hipótesis, se plantearon los siguientes objetivos, general y específicos:

3.1 Objetivo general

Evaluar el uso de herramientas bioinformáticas para Identificar y caracterizar grupos funcionales de transporte en la familia de transportadores de azúcares.

3.2 Objetivo específicos

1. Identificar residuos con impacto en la actividad de transporte e involucrados en el reconocimiento de sustrato en la familia de transportadores de azúcares

2. Obtener y curar anotaciones funcionales en relación a la especificidad de los transportadores pertenecientes a la familia SP.

3. Comparar el nivel de isofuncionalidad observado entre los grupos definidos mediante análisis filogenético global de la familia y mediante agrupamientos desde residuos de reconocimiento de sustrato.

4. Caracterizar la familia y los grupos definidos a nivel de secuencia (patrones de conservación, comportamiento mutacional, correlación de residuos y residuos responsables de la segregación de los grupos obtenidos)

4 MATERIALES Y MÉTODOS

4.1 Datos iniciales y extensión de homólogos

A partir de las 139 secuencias representativas de los distintos miembros de la familia de transportadores de azúcares alojadas en la base de datos TCDB (http://www.tcdb.org/) bajo el código 2.A.1.1, se realizó una extensión de homólogos contra la base de datos SwissProt empleando el programa BLASTp y utilizando como valor de corte un E-value igual o menor a 10^{-30}. Proteínas alojadas en SwissProt no identificadas mediante BLASTp pero asociadas a la familia por la entrada IPR003663 de Interpro y/o etiquetadas bajo el código TC de la familia en Uniprot se incluyeron en en el conjunto de datos.

4.2 Análisis exploratorio y filtrado de secuencias

Para el análisis exploratorio y filtrado de secuencias se evaluó la identidad y similitud de cada par de proteínas en el conjunto de datos, así como la topología de cada secuencia. En el primer caso se empleó el programa Fasta36, eliminándose del conjunto de datos proteínas con una redundancia del 100%. También se computó la similitud global entre secuencias empleando la aproximación matricial implementada en el programa PySCA (Rivoire et al., 2016) para futuros análisis. La predicción de topología de las secuencias se llevó a cabo utilizando los programas Memsat3 (Jones, 2007), Memsat-svm (Nugent & Jones, 2009) y TOPCONS (Tsirigos et al., 2015). Este último entrega los resultados de los algoritmos OCTOPUS, Philius, PolyPhobius, SCAMPI y SPOCTOPUS, siendo la predicción de TOPCONS una predicción consenso de las anteriores. Se determinó el mejor método de predicción a partir de dos criterios de análisis: distribución del número de TMs predichas para las secuencias y análisis comparativo entre las predicciones y la topología conocida de secuencias con estructuras tridimensionales resueltas, la cual se encuentra anotada en la base de datos

PDBTM (Kozma et al., 2012). Las proteínas identificadas como fragmentos proteicos de parálogos ya presentes en el conjunto de datos a partir del análisis de identidad y topología fueron descartadas de los análisis posteriores.

4.3 Alineamiento múltiple de secuencias

A partir del conjunto filtrado de secuencias se obtuvo un alineamiento consenso tras la aplicación de distintos algoritmos y parámetros de alineamiento basados en consistencia, implementados en el programa Mafft v.7310 (Katoh & Standley, 2013), resumidos en la **Tabla 6**. El programa Mafft permite el uso de matrices de sustitución definidas por el usuario, lo que permitió el uso de las matrices de sustitución PHAT y SLIM (Müller et al., 2001; Ng et al., 2000), construidas a partir de proteínas transmembrana. Posteriormente se evaluaron los MSA considerando los indicadores de calidad establecidos en la **Tabla 7**, mediante el desarrollo de funciones en Python y el uso de funciones de análisis implementadas en el programa T-Coffee (Notredame et al., 2000). Tras esto, se combinaron los mejores alineamientos utilizando el programa MergeAlign (Collingridge & Kelly, 2012) (http://mergealign.appspot.com/), evaluando los mismos indicadores para confirmación de optimización.

Tabla 6. Métodos de alineamiento utilizados

Programa	Algoritmos	Parámetros*	Comentarios
Mafft	E-INS-I *Global Pair* *Local Pair*	Matriz de sustitución: BLOSUM62 Iteraciones: 1000 Otros: Por defecto Matriz de sustitución: PHAT Iteraciones: 1000 Otros: Por defecto Matriz de sustitución: SLIM Iteraciones: 1000 Otros: Por defecto	Algoritmo E-INS-I recomendado para proteínas con múltiples dominios conservados (en este caso, TMs). Los algoritmos *Global y Local Pair* corresponden, como su nombre indica, a algoritmos de alineamiento golbal y local, respectivamente.

* Parámetros empleados para cada uno de los algoritmos

Tabla 7. Indicadores empleados para evaluación de calidad de alineamientos

Programa	Indicadores	Comentario
Prody	Perfil de ocupancia	Permiten evaluar de manera exploratoria el nivel de ocupancia del alineamiento en zonas críticas (en este caso, TMs) y el correcto alineamiento de sitios conservados conocidos en la familia y en la MFS.
Código personal*	N° de sitios con 100% de ocupancia y N° de sitios conservados	
Código personal*	Porcentaje de aperturas por TMs.	Calculado para topologías de PDBs de la familia y para topologías predichas para todo el conjunto de secuencias. Criterio de mayor prioridad, ya que las TM en el contexto del plegamiento MFS deben presentar un mínimo de aperturas.
T-Coffee	TCS	El puntaje de consistencia transitiva (TCS) evalúa el nivel de robustez e incertidumbre del alineamiento obtenido para cada posición, entregando además un puntaje promedio para el alineamiento.

* Código desarrollado en Python en el transcurso de esta tesis.

Paralelamente y para complementar los análisis de caracterización de la familia a nivel de secuencia, se obtuvo el MSA asociado al perfil PF00083 de Pfam, que agrupa secuencias relacionadas al transporte de azúcares y otros metabolitos. Ambos alineamientos fueron refinados según criterios de redundancia, cobertura contra la secuencia de hGLUT1 y/u ocupancia de columnas según se requiriera, como se indica en la **Tabla 8**, empleando la función refineMSA de Prody**.**

Tabla 8. Parámetros de refinamiento de alineamientos múltiples

MSA de entrada	Parámetros de refinamiento[b]	MSA de salida[c]
PF00083	Label=GTR1_HUMAN, Rowocc=0.8, Seqid=0.98	PfamSP
Este estudio[a]	Label=P11166, Rowocc=0.7	SP
SP	Colocc = 0.6	SP-SCA

a. MSA obtenido a partir del proceso de refinamiento detallado
b. Visitar http://prody.csb.pitt.edu/_modules/prody/sequence/msa.html#refineMSA para mayor detalle de parámetros de refinamiento.
c. Nombre de MSA asignado para facilitar redación y lectura de métodos y resultados.

4.4 Identificación de residuos funcionales

Se procedió a identificar residuos de unión de sustratos y variantes con impacto en la actividad de transporte. Para el primer caso, se anotaron los residuos de unión a sustratos identificados a partir de las estructuras co-cristalizadas con estas moléculas alojadas en la base de datos PDB. Para el segundo caso, se examinaron las anotaciones funcionales integradas en la base de datos UniProt de las variantes encontradas en patologías o bien construidas mediante mutación sitio-dirigida, junto a una descripción de su impacto en la actividad de transporte. Ambas anotaciones se complementaron mediante revisión bibliográfica de artículos asociados.

4.6 Anotación funcional de sustratos

A partir del código Uniprot de cada secuencia se automatizó la descarga y tabulación de anotaciones relativas a la taxonomía de los organismos de procedencia de las secuencias (nombre de gen, especie, género, filo, dominio), junto a las anotaciones funcionales relacionadas directamente con la función molecular, desde las bases de datos Uniprot (que integra las anotaciones de las ontologías GO y ECO) y TCDB. De las anotaciones extraídas desde las bases de datos, las correspondientes a anotaciones funcionales son (1) Sustratos anotados en la ontología GO, con sus respectivos niveles de evidencia y citas de artículos científicos asociados; (2) Valores de parámetros cinéticos (Km, Vmax) alojados en Uniprot, con citas de artículos científicos asociados; (3) Descripción general de la función alojada en TCDB, con citas de artículos científicos asociados; (4) Descripción general de la función alojada en UniProt, con etiquetas de evidencia de la ontología ECO y/o citas de artículos científicos asociados. A partir de tales anotaciones, se clasificaron las proteínas en dos clases: Transportadores con y sin caracterización experimental a nivel de sustrato. Los criterios para tal clasificación a partir de la información de las anotaciones funcionales se presentan en la **Tabla 9**. Para que una proteína se considerara dentro de la primera clase (con caracterización experimental de los sustratos transportados), debía cumplir con al

menos uno de los criterios enunciados. Para las proteínas que no presentaban sustratos anotados en GO y/o no poseían valores cinéticos asociados, pero poseían una descripción funcional a nivel de sustrato en TCDB o Uniprot con un artículo científico asociado, se hizo una revisión bibliográfica de estos artículos y otros referenciados en los primeros, para complementar y curar las anotaciones funcionales. También se realizó esto para transportadores que tenían más de un sustrato anotado, pero sin información cinética, para poder jerarquizar, de existir tal información, su especificidad por estos sustratos (p.e. glucosa > fructosa). Para los transportadores que tuvieran asociados sustratos generales o ambiguos (por ejemplo “azúcares”, “hexosas”) se buscó, en la medida de lo posible, aumentar el nivel de detalle de la anotación, mediante revisión bibliográfica. Los datos recopilados fueron formateados e integrados a la tabulación inicial.

Tabla 9. Criterios empleados en la curación de anotaciones funcionales

Criterios	Transportadores con caracterización experimental a nivel de sustrato	Transportadores sin caracterización experimental a nivel de sustrato
Descripción UniProt	A nivel de sustrato, con citas incluidas.	Ninguna o Pobre. Sustratos probables, putativos, o definidos por similaridad. (ECO:0000250, 0000305)
Descripción TCDB	A nivel de sustrato, con citas incluidas.	Ninguna o Pobre. Sustratos probables o putativos.
Códigos de evidencia GO	Experimentales (IDA, IMP, IGI)	No-experimentales (IBA, IEA, ISS, ISO, IC)
Revisión Bibliográfica	Se encuentran artículos con caracterización funcional a nivel de sustratos.	No se encuentran artículos con caracterización funcional a nivel de sustratos.

4.6 Obtención de agrupamientos

Para evaluar el nivel de asociación entre las relaciones evolutivas de los genes de la familia y sus propiedades funcionales, se procedio a la recosntrucción filogenética de las secuences desde el alineamiento SP, eliminando columnas que tuvieran menos del 10% de ocupancia. Se empleó el método de máxima verosimilitud, implementado en el programa RAxML 7.2.7 (Stamatakis, 2014). Se modelaron las tasas evolutivas entre sitios según distribución gamma, permitiendo que el programa determinara el modelo evolutivo más acorde al conjunto de datos mediante criterio de inferencia bayesiana. Para cada reconstrucción se realizaron 1000 pasos de *bootstrap.*

Para analizar las relaciones existentes entre los sitios de unión de sustrato en la familia se obtuvo un sub-alineamiento de las secuencias en estudio que considera sólo las posiciones asociadas al reconocimiento de sustrato. A partir de éste se obtuvo un filograma aparente, empleando el mismo método de reconstrucción filogenética mencionado anteriormente. También se procedó a desarrollar un análisis clúster del sub-alineamiento, empleando el algoritmo k-means tras descomposición espectral del alineamiento mediante MCA, análisis implementado en el programa S3Det (Rausell et al., 2010). El análisis clúster se desarrolló para el sub-alineamiento sin filtro de redundancia y aplicando un filtro de 95% de redundancia para el sitio de unión, en ambos casos de manera iterativa hasta obtención de agrupamientos con no más de 100 secuencias por grupo.

4.7 Caracterización de la familia y de los agrupamientos a nivel de secuencia

Se caracterizó el perfil de la familia y de los residuos de reconocimiento de sustrato en términos de patrones de conservación y correlación de residuos. La Identificación de sitios conservados se desarrolló mediante análisis de entropía de Shannon y entropía relativa (RE) para los alineamientos SP y PfamSP, implementados en los programas Prody (Bakan et al., 2014) y pySCA (Rivoire et al., 2016) respectivamente. Para la identificación de pares y redes de residuos

correlacionados se llevaron a cabo análisis basados en MI y SCA, empleando las funciones disponibles para estos fines en los programas ya mencionados. También se identificaron los residuos con mayor comportamiento mutacional, empleando para ello el programa Xdet

La identificación y comparación de los grupos definidos desde filogenia y el análisis cluster del sitio de unión, así como la evaluación de su nivel de isofuncionalidad, se realizó mediante inspección visual de la topología de la filogenia, anotada con los grupos obtenidos mediante clústering y las etiquetas funcionales empíricas de las secuencias (mecanismo de transporte, sustratos transportados, sustrato predominante, parámetros cinéticos asociados) obtenidas mediante la revisión de anotaciones en base de datos y artículos científicos. Para la anotación y análisis del árbol se empleó el servidor itol (https://itol.embl.de).

La caracterización de los grupos identificados se realizó mediante la identificación de los residuos determinantes de las agrupaciones obtenidas, empleando los métodos de selección de características integrados en los programas S3Det (Rausell et al., 2010) y Multi-Harmony (Brandt et al., 2010). Para cada grupo se construyeron perfiles representativos de los sitios funcionales identificados, visualizados como LOGOs utilizando el programa WebLogo3 (https://github.com/WebLogo/weblogo), integrando la información de los residuos identificados.

5 RESULTADOS

5.1 Datos iniciales y extensión de homólogos

La búsqueda de homólogos mediante BLASTp, tras eliminar duplicados en resultados de búsqueda, arrojó un total de 186 secuencias no incluídas en el conjunto de datos inicial. Mediante el análisis de las entradas IPR003663 y de las proteínas en Uniprot anotadas bajo el código TC 2.A.1.1 se identificaron otras 33 secuencias asociadas a la familia y alojadas en SwissProt. También se incluyó el cotransportador de Fructosa:H+ de *Saccharomyces pastorianus,* código Q9HFF8, identificado mediante revisión bibliográfica como un transportador de la familia presente en levaduras bien caracterizado a nivel de especificidad y que presenta una mayor afinidad por fructosa que glucosa. Esto dio origen a un conjunto inicial conformado por 359 secuencias.

5.2 Análisis exploratorio y filtrado de secuencias

A partir de los alineamientos de pares obtenidos mediante el algoritmo Smith-waterman implementado en el programa Fasta36, se identificaron 7 pares de secuencias idénticas, descartándose del análisis posterior una de cada par, para eliminar redundancia. Los pares identificados se indican en la **Tabla 10.**

Tabla 10. Pares de secuencias redundantes

ID de pares*	Comentario	ID de pares*	Comentario
P0AE24, P0AE25	AraE de *E.coli,* cepas K12 y O157:H7, respectivamente.	Q2U2Y9, B8NIM7	Probables permeasas de quinato, qutD de *Aspergillus oryzae* y *Aspergillus flavus*, respectivamente.
P0AGF4, P0AGF5	XylE de *E.coli,* cepas K12 y O157:H7, respectivamente.	F1R0H0, A8KB28	Misma entrada Uniprot. Cambio de ID en el tiempo.
P0AEP1, P0AEP2	GalP de *E.coli,* cepas K12 y O6:H1, respectivamente.	Q3T9X0, Q5ERC7	Misma entrada Uniprot. Cambio de ID en el tiempo.
O74713, A0A1D8PCL1	HGT1 de Candida albicans, cepas distintas.		

* ID de Uniprot. Se procedió a eliminar la segunda secuencia de cada par.

Posteriormente se filtraron las secuencias según criterios topológicos. Las secuencias que fueron identificadas como secuencias con menos de 12 TMs por todos los métodos de predicción testeados (TOPCONS, Memsat3, Memsat-svm, OCTOPUS, Philius, PolyPhobius, SCAMPI y SPOCTOPUS), presentadas en la **Tabla 11**, fueron descartadas de análisis posteriores, quedando un conjunto de datos con un total de 345 secuencias con 12 TMs predichas por al menos uno de los métodos testeados. Éstas presentaron una media de identidad y de similaridad de 27.5% y 59.2%, respectivamente.

Tabla 11. Proteínas con N° de TMs predichas menor a 12

ID Uniprot	Org.	Gen	Long.	Método*								Comentario
				a	b	c	d	e	f	g	h	
Q09039	*T. brucei*	THT2C	337	4	6	7	7	7	7	7	7	Fragmento de THT2.
Q7XA64	*A. thaliana*	At3g05155	327	8	8	8	8	8	8	8	7	Producto génico tipo ERD6 con evidencia a nivel de transcrito.
P40441	*S. cerevisiae*	HXT12	457	11	11	11	11	11	11	11	11	Producto génico incierto.
P40440	*S. cerevisiae*	YIL171W	109	1	1	1	1	1	1	1	1	Producto génico incierto.
Q8TFG1	*S. pombe*	GHT7	518	11	11	11	11	11	11	11	11	Producto génico inferido por homología.
Q9XT10	*S. scrofa*	SLC2A4	174	4	4	4	4	4	4	4	3	Fragmento homólogo a genes SLC2A4
Q9XST2	*C. lupus*	SLC2A4	162	5	5	5	5	5	5	4	5	Fragmento homólogo a genes SLC2A4

* a, Topcons; b, Memsat-svm; c, Memsat3; d, Octopus; e, Phillius; f, Polyphobius; g, Scampi; h, spoctopus.

La longitud de las 345 secuencias analizadas se encuentra en el rango de 400 a 1440 residuos aminoacídicos, con una media de 543 aa y una mediana de 523 aa. Para un mejor análisis, las secuencias se clasificaron según su pertenencia a cinco categorías taxonómicas, cuatro de ellas considerando los reinos definidos por Cavalier Smith (1998) y en consistencia con los sistemas de clasificación actuales: bacteria (B), protozoa (Pr), fungi (F) y animalia o metazoa (M). La quinta categoría corresponde al grupo conocido como archaeplastida o primoplantae (Pp) en los sistemas de clasificación actuales e incluye plantas y algas verdes, rojas y glaucofitas. En la **Figura 16** se presenta la distribución en la longitud de las secuencias para cada categoría.

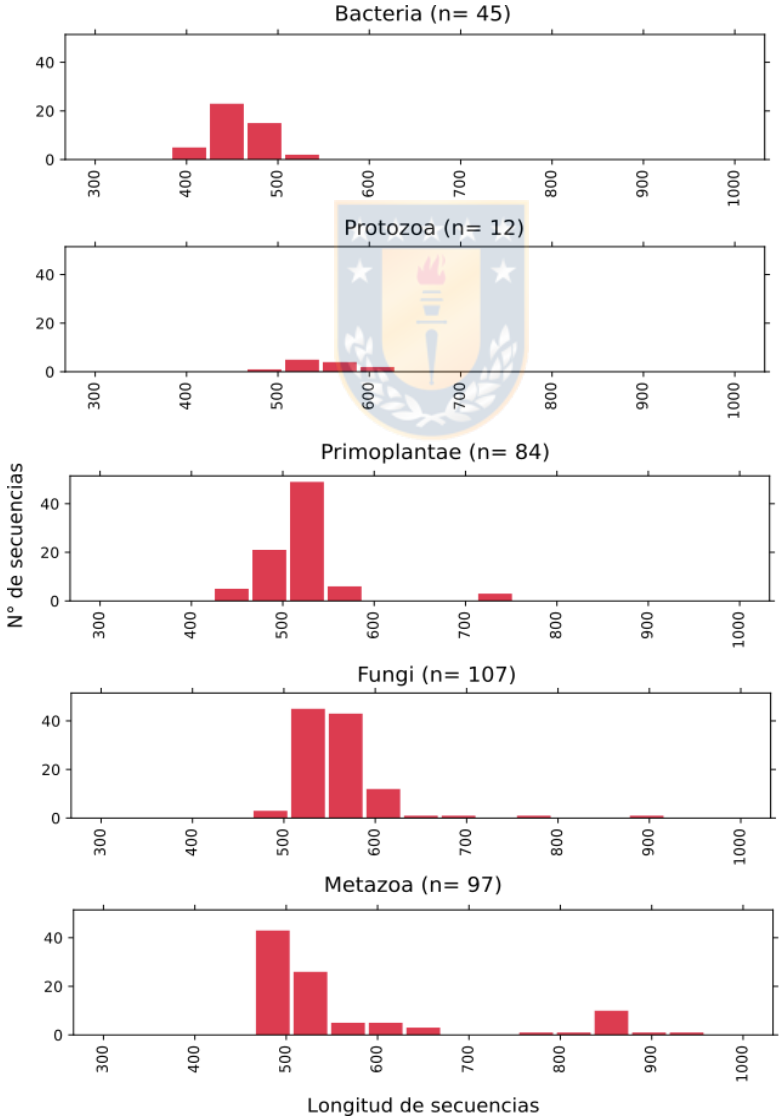


Figura 16. **Distribución de longitud de secuencias por categoría taxonómica.**

En cuanto al análisis topológico de las secuencias, en primera instancia se procedió a determinar el método con mejor capacidad predictiva. Primeramente se analizó la distribución del número de TMs predicho para las 345 secuencias por cada uno de los métodos testeados. Los resultados se presentan en la **Figura 17**. Se observa que Memsat-svm, seguido por Topcons, Polyphobius y Phillius son los que identifican una topología de 12 TMs para un mayor número de secuencias, que es lo esperable para la familia y el plegamiento MFS que la caracteriza. Por otro lado, Memsat3 es el que presenta peor capacidad de predicción siguiendo este criterio.

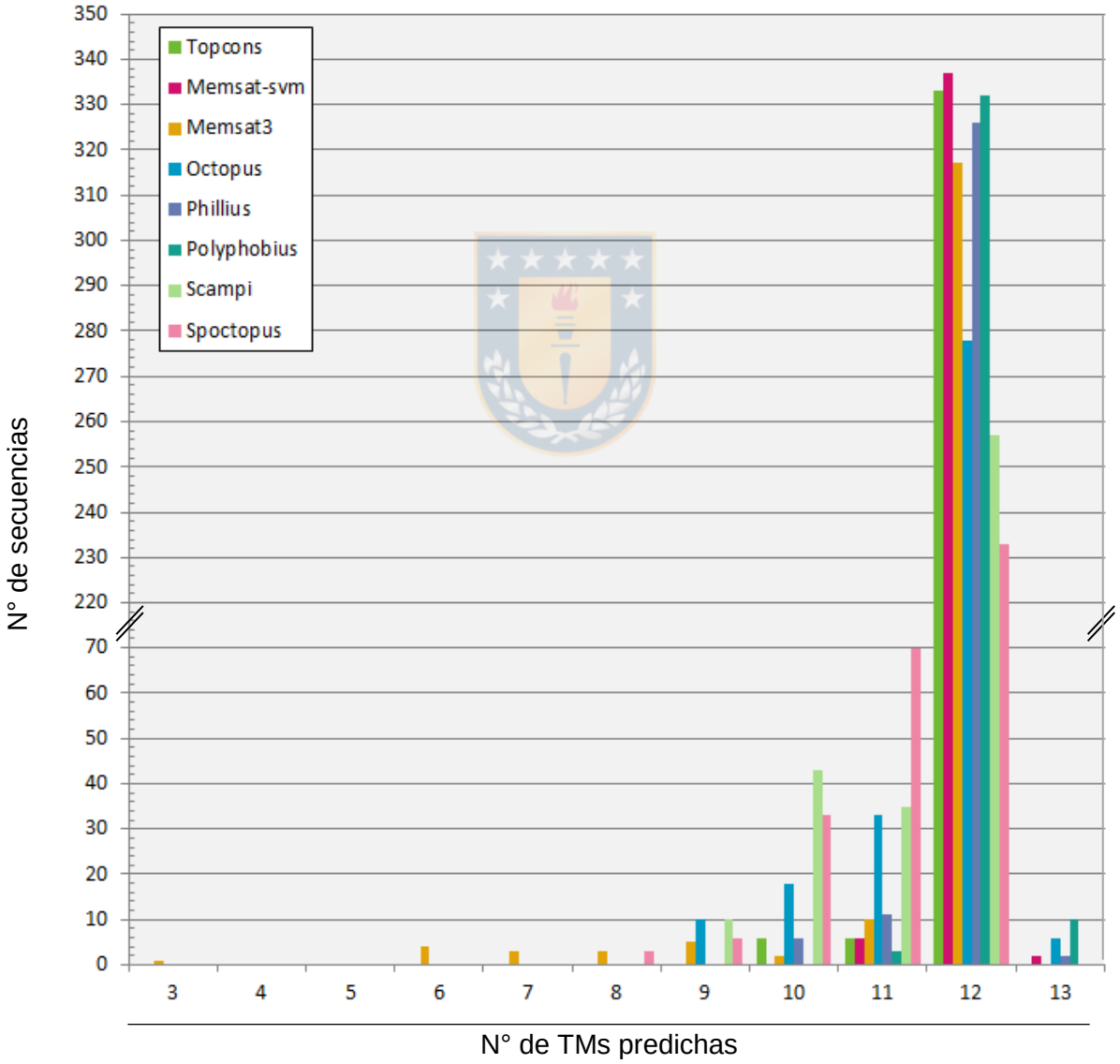


Figura 17. **Distribución del número de TMs predichas por cada método testeado.**

Como segundo criterio se analizó la concordancia entre la topología conocida de los miembros de la familia con estructuras tridimensionales resueltas a partir de cristalografía de rayos X y la predicha por cada método. Para esto se desarrolló una función en Python que entrega el número de residuos y el promedio de residuos cuya clasificación topológica discrepa de la topología estructural conocida y anotada en la base de datos PDBTM, además de una visualización comparativa entre ambas, ejemplificada por la topología de hGLUT1 predicha por Memsat-SVM en la **Figura 18.** El porcentaje de discrepancia de las predicciones obtenida por cada método para estructuras representativas de la familia se presenta en la **Tabla 12.** Así también, para cada método de predicción, se tomaron las secuencias con una topología putativa de 12 TMs y se analizó la distribución del largo para cada uno de los 25 segmentos topológicos: segmentos N y C terminales, 12 TMs y 11 tramos de conexión entre TMs –de aquí en adelante lazos– y se compararon con las longitudes de los segmentos exhibidos por estructuras cristalográficas pertenecientes a la familia SP y a la MFS, considerando su inicio y término según lo establecido en la base de datos PDBTM. La distribución para las estructuras cristalográficas analizadas y para cada uno de los métodos se presenta en la **Tabla 13** y **Figura 19**, respectivamente.

Tabla 12. Porcentajes de discrepancia entre topologías SP conocidas y predicciones obtenidas por cada método

Proteina	PDB	Topcons	Memsat-SVM	Memsat3	Octopus	Phillius	Polyphobius	Scampi	Spoctopus
$_{ec}$XylE	4GBY	14.05	**7.94**	12.02	12.02	14.05	10.39	15.27	11.61
$_{se}$GlcP	4LDS	19.31	**7.32**	10.57	23.17	11.59	10.37	20.93	25.81
$_{h}$GLUT1	4PYP	11.9	**6.85**	12.3	21.17	10.08	9.48	23.59	18.35
$_{h}$GLUT3	4ZW9	15.76	**9.92**	16.54	17.32	11.87	13.42	15.37	17.7
$_{bt}$GLUT5	4YB9	18.16	19.42	23.8	**17.75**	20.25	19.83	20.25	16.28
$_{at}$STP10	6H7D	**17.94**	21.3	23.54	18.39	20.4	**17.94**	19.28	19.73
$\bar{x}$		16.19	**12.13**	16.46	18.30	14.71	13.57	19.12	18.25

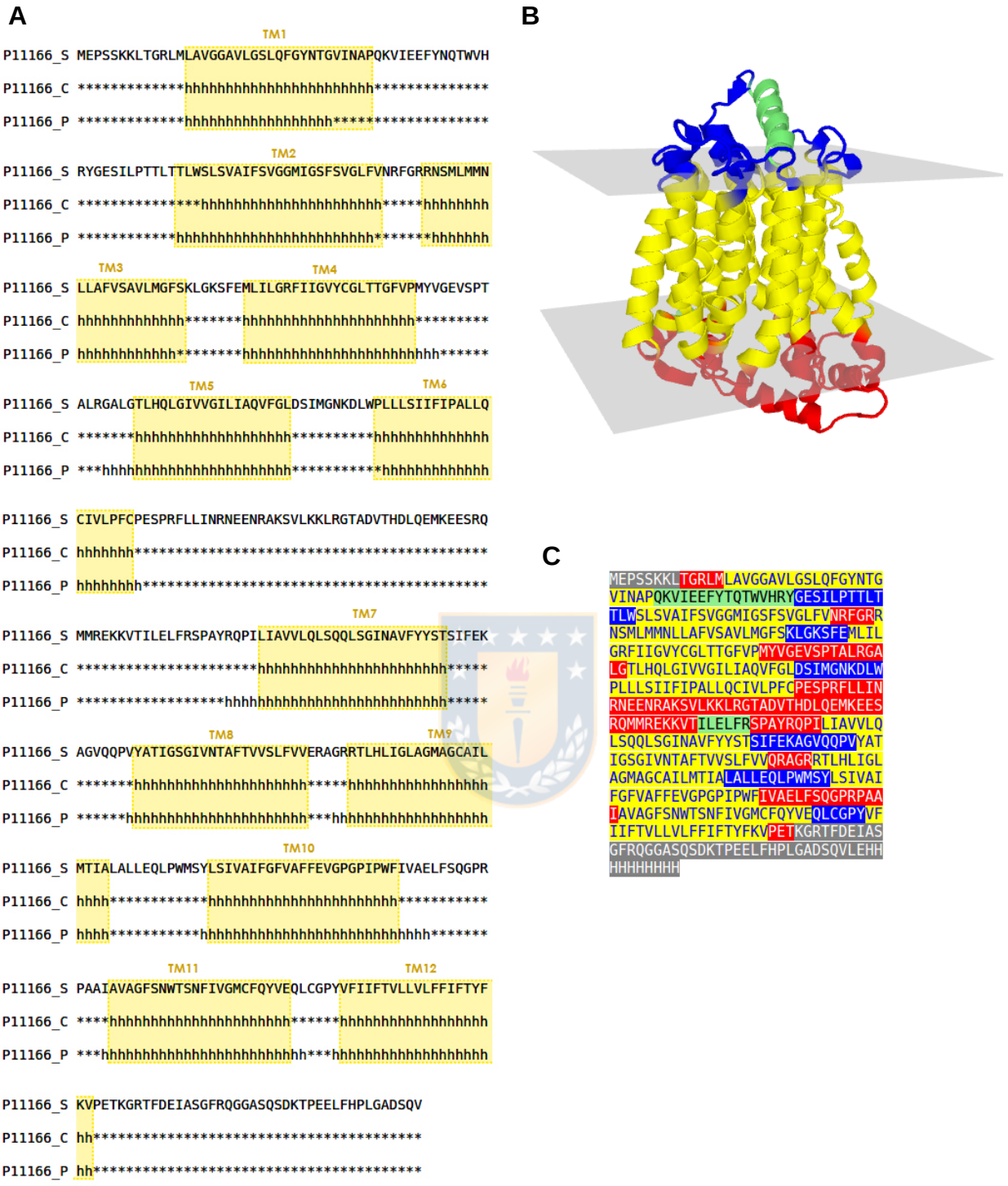


Figura 18. Comparación entre topología real y predicción por MEMSAT-SVM para hGLUT1. (A) Códigos '_S','_C' y '_P' aluden a la secuencia, topología de cristal (PDB 4PYP) alojada en PDBTM y predicción, respectivamente. (B). Visión de cristal en PDBTM. (C). Anotación topológica entregada por PDBTM.

Tabla 13. Estadísticos de longitud de TMs para estructuras cristalográficas MFS y topologías predichas por Memsat-svm.

A)	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM8	TM9	TM10	TM11	TM12
Min.	12	13	12	12	15	13	15	14	14	14	13	13
Max.	25	24	22	24	23	22	27	24	22	25	23	22
$\tilde{x}$	19.3	18.0	17.5	19.7	18.8	17.5	20.4	18.6	18.1	20.2	18.9	17.6
SD	2.7	2.6	2.2	2.6	2.1	2.0	2.6	2.2	2.1	2.6	2.5	2.0

B)	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM8	TM9	TM10	TM11	TM12
Min.	14	16	14	16	15	15	17	16	15	18	16	13
Max.	23	23	21	24	23	20	25	24	22	24	23	20
$\tilde{x}$	19	19	18	21	19	18	21	19	19	22	20	18
SD	2	2	2	2	2	1	2	2	2	2	2	2

C)	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM8	TM9	TM10	TM11	TM12
Min.	16	18	18	16	16	16	16	18	17	16	16	19
Max.	31	28	30	28	31	25	31	31	29	31	31	28
$\tilde{x}$	21.6	21.9	19.6	24.4	23.4	21.6	25.1	22.1	23.8	27.3	24.8	21.0
SD	3.4	1.2	1.0	1.6	1.8	1.6	2.5	1.6	2.2	2.5	2.2	1.4

(A) Estadísticos para 51 cristales MFS con 12 TMs alojados en base de datos PDB. (B) Estadísticos para 16 cristales SP alojados en PDB (Tabla 5) . (C) Estadísticos para topologías predichas por Memsat-svm. Los códigos PDB asociados a la tabla A se presenta en **Anexo 2.**

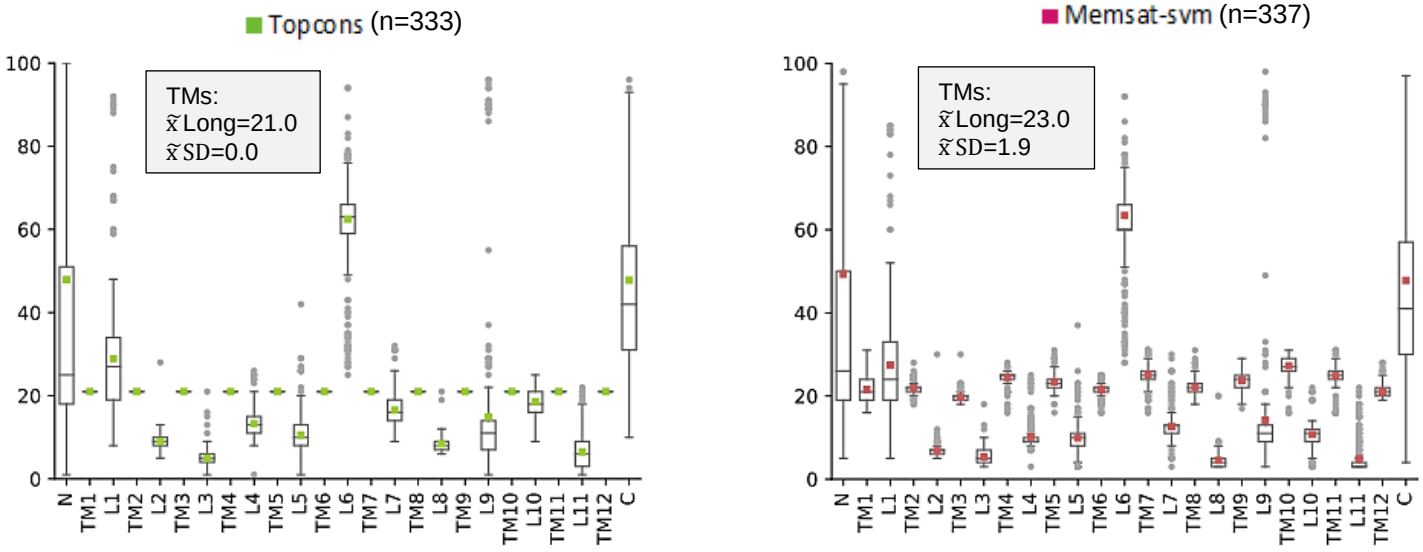


Figura 19. Distribución del largo de los segmentos topológicos predichos por los distintos métodos testeados. En rectángulo gris, se indica longitud y desviación estandar promedio para los segmentos transmembrana.

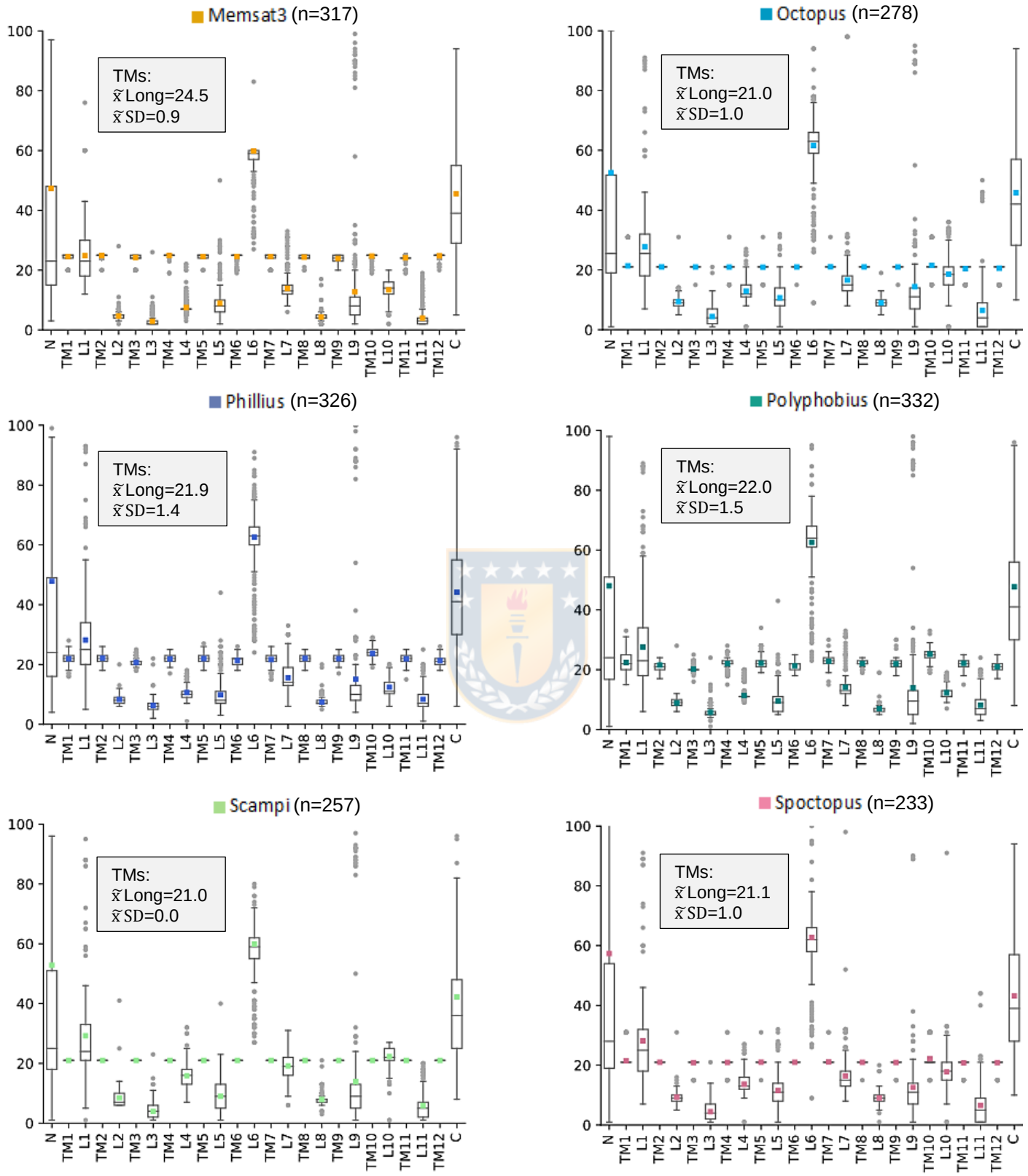


Figura 19 (Continuación).

En general y a partir del análisis de discrepancia en la clasificación de residuos, se observa que Memsat-SVM presenta la mejor precisión, seguido por Polyphobius. Para los casos de Memsat-SVM, los reiduos cuya predicción discrepa con la topología PDB corresponden a residuos flanqueantes de los TM, siendo principalmente discrepancias por sobreestimación del largo de los segmentos que se encuentran embebidos en la membrana, que es el criterio usado en PDBTM para definir el inicio y final de los TMs. Esto probablemente debido a que el programa es capaz de reconocer tramos de las hélices que se extienden más allá de la membrana. Sin embargo el núcleo de los TM predichos presenta buena concordancia con las estructuras cristalográficas. En base a lo anterior, se decidió utilizar la topología predicha por Memsat-SVM para los análisis posteriores que requirieran tal información. Para las secuencias en las que Memsat-SVM predijo un número de TMs distinto de 12, correspondientes a las entradas P42417, Q926Q9, A0A0H3NW06, Q06222, A0A075BFV8, Q3UHK1, Q32NG5 y Q10917 de Uniprot, se utilizó la predicción de Polyphobius, a excepción de la última entrada, en donde se usó la predicción de Phillius, ya que Polyphobius también predecía un número de TMs distinto de 12.

Con respecto a la longitud de los segmentos topológicos de la familia, destacan los segmentos N y C terminales, junto a los lazos 1 y 6, como los segmentos con mayor variabilidad en tamaño a lo largo de la familia, lo cual se observa independientemente del método de predicción. Se evaluó la existencia de diferencias en la distribución de los largos de los segmentos entre las categorías taxonómicas representadas. Los resutados se grafican en la **Figura 20,** las tablas de datos asociadas se presentan en el **Anexo 3.**

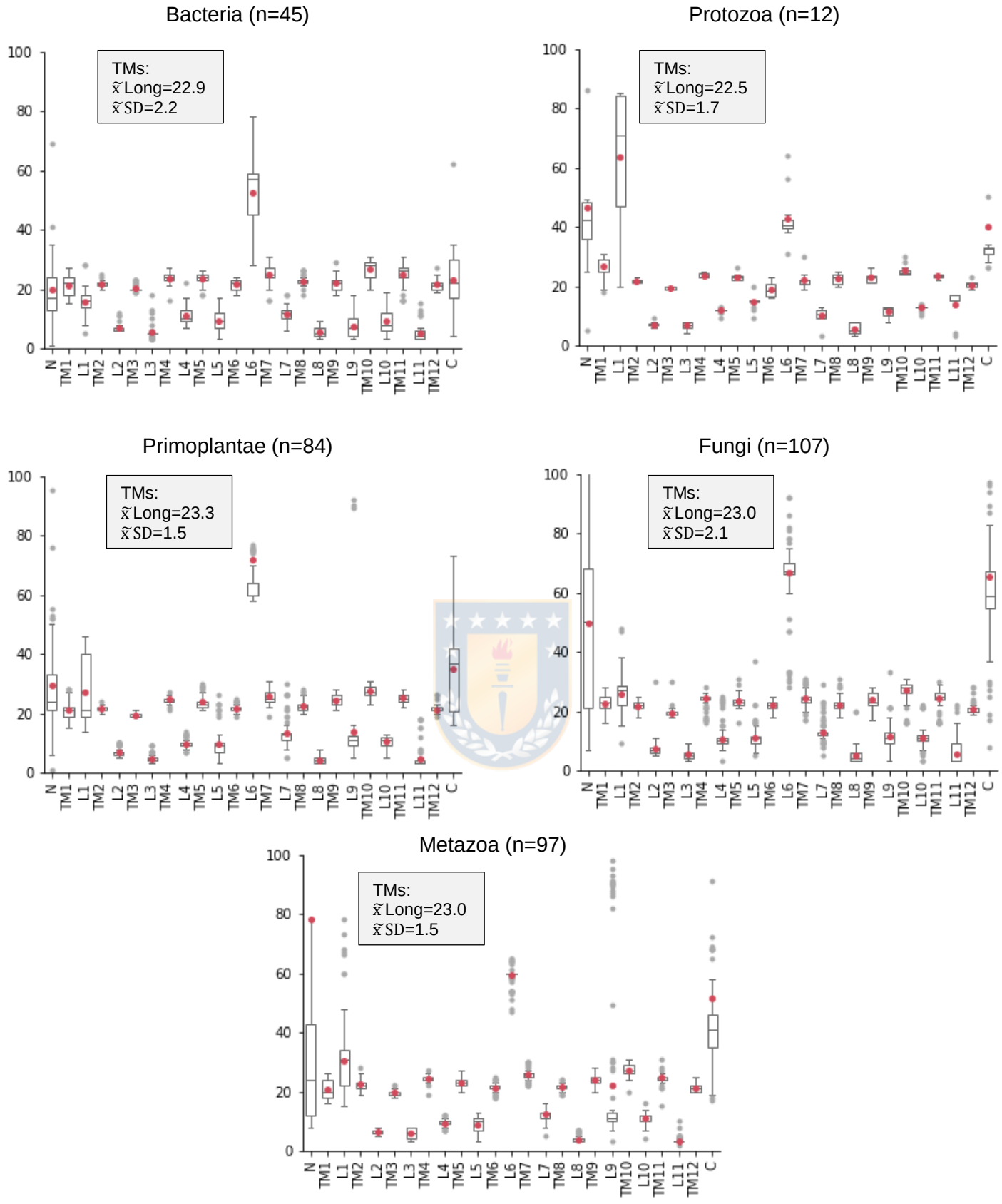


Figura 20. Distribución del largo de los segmentos topológicos por categoría taxonómica. Se usan los largos predichos por el mejor método de predicción. En rectángulo gris, se indica longitud y desviación estandar promedio para los segmentos transmembrana.

5.3 Alineamientos múltiples de secuencias

Para los alineamientos obtenidos a partir de los distintos algoritmos y matrices de sustitución empleados, se determinaron una serie de indicadores para su evaluación. Específicamente, se deteminó la longitud del alineamiento, el puntaje TCS, el número de sitios conservados en un 90 y 80%, el número de sitios con un 100% de ocupancia y el porcentaje de aperturas por TMs. Para este último se utilizaron dos aproximaciones: (1) evaluación del número de aperturas según la topología de secuencias con estructuras cristalográficas resueltas y (2) evaluación del número de aperturas para todas las secuencia, utilizando la información de topología predicha por el mejor método de predicción (ver sección 5.1). Los valores de los indicadores para cada alineamiento se presentan en la **Tabla 14.** Los alineamientos construidos a partir de las matrices de sustitución PHAT y SLIM presentarom mejor calidad independiente del algoritmo empleado, considerando el porcentaje de aperturas en TMs como el criterio de mayor importancia para su evaluación. En base a lo anterior, se obtuvieron dos alineamientos consenso mediante el empledo del programa MergeAlign, el primero a partir de la fusión de los mejores alineamientos obtenidos por cada algoritmo, y el segundo a partir de la fusión de los 6 alineamientos construidos con las matrices PHAT y SLIM (Cons1 y Cons2, respectivamente), y se determinaron los mismos indicadores para su evaluación (**Tabla 14**). Cons2 presnetó menores porcentajes de aperturas por TMs y similares valores para el resto de indicadores, por lo que fue seleccionado para desarrollar los análisis posteriores. Finalmente se inspeccionó el alineamiento de manera manual, identificándose 4 secuencias conflicto que introducían aperturas por apertura importantes en algunos TMs, y que a partir de análisis de su topología no se encontraban alineados correctamente, como se muestra en la **Figura 21(A-D)**. Estas secuencias, correspondientes a las entradas Uniprot G0R6T1, K0DZ95, B6HNK5 y P44610, fueron eliminadas del alineamiento, dando origen al alineamiento refinado final con 341 secuencias, que presentó aumento en el puntaje TCS y disminución de aperturas en TMs, como se muestra en la **Tabla 14** y en el análisis de ocupancia del alineamiento, ver **Figura 21(E).**

Tabla 14. Indicadores de calidad para los alineamientos evaluados.

Indicador	*e-ins-i*			*local pair*			*global pair*			Cons1*	Cons2*	**Ref***
	Bloss.62	Phat	Slim	Bloss.62	Phat	Slim	Bloss.62	Phat	Slim			
Long.	3084	3272	3234	3149	3109	3332	3055	3226	3414	3109	3129	**3073**
TCS	692	689	690	692	690	687	691	690	689	690	690	**695**
C90	12	11	11	12	11	11	12	12	12	11	11	**11**
C80	25	24	25	25	25	24	25	25	25	25	25	**25**
N° sitios c/100% oc.	201	233	233	207	228	250	207	236	244	228	230	**248**
$\bar{x}$(% aperturas por TMs PDBs)	10.7	5.8	3.9	8.7	5.0	5.2	10.0	5.3	5.2	5.1	4.7	**1.8**
$\bar{x}$(%aperturas por TMs predicciones)	13.6	8.9	6.9	11.6	6.4	8.3	14.4	6.9	8.7	6.5	6.3	**3.8**
$\bar{x}$	12.1	7.4	**5.4**	10.2	**5.7**	6.7	12.2	**6.1**	7.0	5.8	5.5	**2.8**

* Cons1, alneamiento obtenido a partir de la fusión de los alineamientos destacados en gris.
* Cons2, alineamiento obtenido a partir de la fusión de los alineamientos Phat y Slim.
* Ref, alineamiento refinado obtenido tras eliminación de secuencias conflicto.

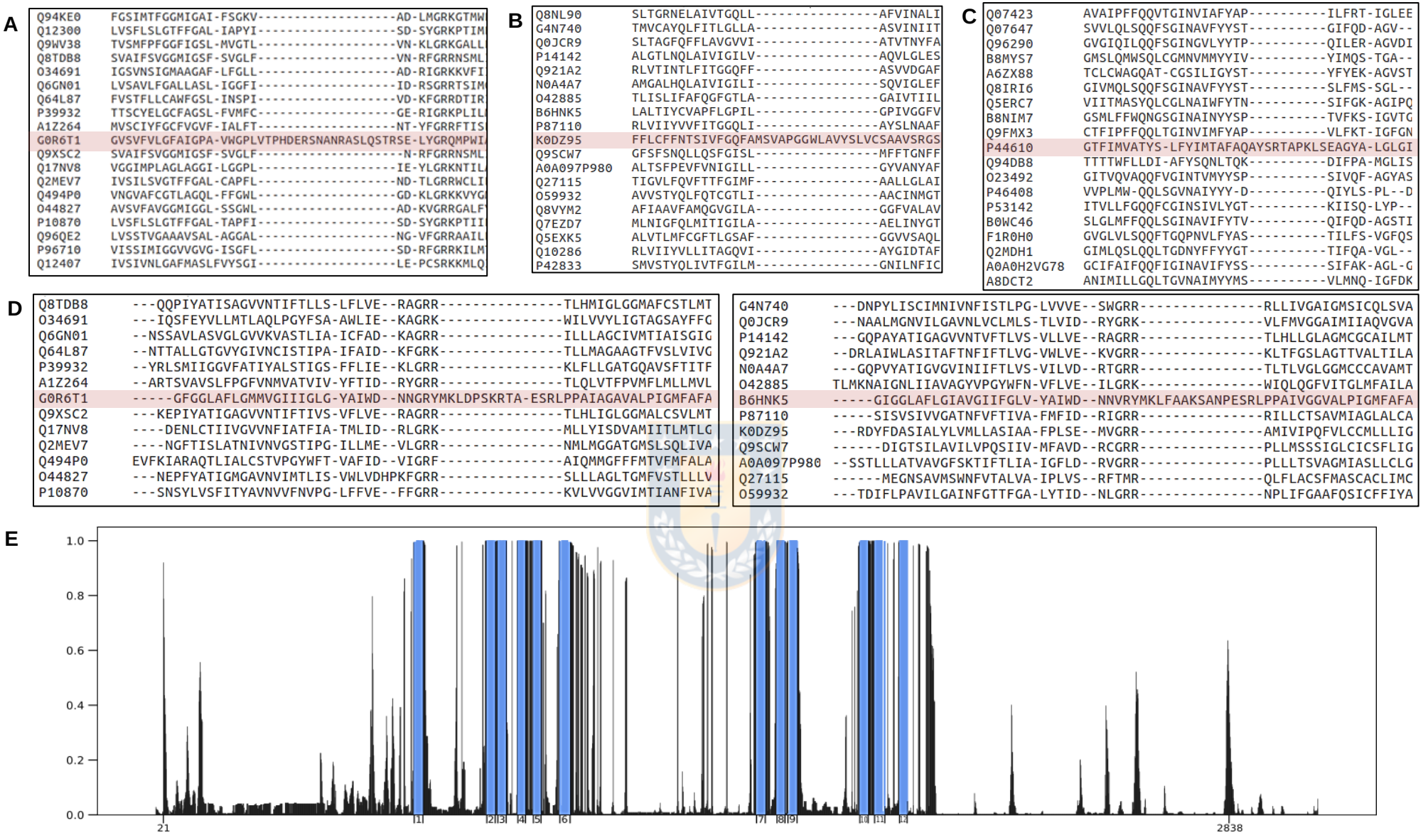


Figura 21. Obtención de alineamiento refinado. (A) Sec. G0R6T1 generando apertura de TM2. (B) Sec. G0R6T1 generando apertura de TM5. (C) Sec. P44610 generando apertura de TM7. (D) Sec. G0R6T1 y B6HNK5 generando aperturas en TM9. (E) Ocupancia del alineamiento refinado, tras eliminación de secuencias conflicto. Segmentos transmembrana destacados en azul.

5.4 Identificación de residuos funcionales

Los residuos con impacto en la actividad de transporte de los miembros de la familia SP identificado mediante estudios experimentales, ya sea que participen directamente en el reconocimiento de sustratos, se encuentren asociados a otros aspectos funcionales o bien , se presentan en la **Tabla 15.**

Tabla 15. Residuos con rol funcional en la familia SP

Taxa[a]	Proteína[b]	Sitio[c]	Ref. hGlut1[c]	Ubicación[c]	Rol funcional	Referencias
B	ecXylE (P0AGF4)	F24	F26	TM1	Unión de Xilosa. Mutagénesis de Ala reduce su transporte a un 25% o menos. Todos conservados en GLUTs 1-4, excepto Q415.	Sun *et al*. (2012)
		Q168	Q161	TM5		
		Q288	Q282	TM7		
		Q289	Q283	TM7		
		N294	N288	TM7		
		Y298	Y292	TM7		
		W392	W388	TM10		
		Q415	N411	TM11		
		W416	W412	TM11		
		I171	I164	TM5	Participan en unión de glucosa, además de los anteriores.Todos conservados en GLUTs 1-4, excepto Q175.	Sun *et al*. (2012)
		Q175	I168	TM5		
		F383	F379	TM10		
		R133	R126	TM4	R133C/H/L y R341W inducen pérdida de transporte de Xilosa.	Sun *et al*. (2012)
		R341	R333	L8		
		G83	G91	L2	Residuos propios de motivos conservados conocidos en la familia con impacto en la actividad de transporte.	Sun *et al*. (2012)
		R84	R92	L2		
		E153	E146	L4		
		R160	R153	L4		
		S223	S210	L6		
		R225	R212	L6		
		G340	G332	L8		
		R341	R333	L8		
		E397	E393	L10		
		R404	R400	L10		
	ecGalP (P0AEP1)	W371	W388	TM10	Importantes en transporte de glucosa.	Patching *et al*. (2008)
		W395	W412	TM11		

Tabla 15. (Continuación)

Ta-xa[a]	Prote-ína[b]	Sitio[c]	Ref. hGlut1[c]	Ubica-ción[c]	Rol funcional	Referencias
B	cgAraE (C4B4V9)	I178 V179	V173 F174	TM5 TM5	I178S aumenta transporte de Xilosa en un 103%. V179D aumenta transporte de xilosa en un 66%.	Wang *et al*. (2013)
Pr	ldGTR1 (Q01440)	D19 E121	N29 E146	TM1 TM4	Críticos para transporte de inositol. Mutagénesis mantiene km similar y disminuye drásticamente Vmax.	Seyfang *et al*. (1997)
	lmGT2 (O61059)	A206 A365	K114 S296	L3 L7	Confieren especificidad por ribosa. En lmGT3 corresponden a residuos de Treonina. Mutantes T>A aumentan especificidad por ribosa.	Naula *et al*. (2010)
	ciGXS1 (Q2MEV7)	F40	F26	L1	Mutantes F40V aumentan velocidad de crecimiento en Xilosa por sobre glucosa	Young *et al*. (2012)
F	scGal2 (P13181)	N376	N317	TM8	N376S y N376F aumentan afinidad por xilosa, este último en mayor medida. Además, N376F elimina transporte de glucosa.	Farwick *et al*. (2014)
		T219	V165	TM5	T219G/S/N permiten transporte de xilosa sin inhibición por glucosa.	Farwick *et al*. (2014)
		Y446 W455	F379 W388	TM10 TM10	Confieren especificidad por galactosa	Kasahara *et al.* (1997)
	scHXT7 (P39004)	F79	F26	TM1	F79S aumenta velocidad de transporte de xilosa.	Apel *et al*. (2016)
		T213 Q209 D340	V165 Q161 I287	TM5 TM5 TM7	Residuos determinantes de la alta afinidad por glucosa. D340C confiere mayor afinidad que WT.	Kasahara & Kasahara, (2010); Kasahara *et al*. (2011)
		N370	N317	TM8	N370S aumentan afinidad por xilosa y elimina transporte de glucosa.	Farwick *et al*. (2014)
	scHXT2 (P23585)	L59 L61 L201 N331 F366	G10 L12 L162 I287 V322	TM1 TM1 TM5 TM7 TM8	Residuos determinantes de la alta afinidad por glucosa. N331 es crítico para el transporte.	Kasahara *et al*. (2007)

Tabla 15. (Continuación)

Ta-xa[a]	Prote-ína[b]	Sitio[c]	Ref. hGlut1[c]	Ubica-ción[c]	Rol funcional	Referencias
F	scHXT2 (P23585)	F431 Y440	F379 W388	TM10 TM10	F431Y e Y440W confieren capacidad de transportar galactosa. Mutación análoga F462Y en SNF3 no confiere respuesta a galactosa.	Kasahara *et al.* (1997)
	scSNF3 (P10870)	I374 F462	V290 F379	TM7 TM10	I374V conlleva pérdida parcial de *sensing* de fructosa y manosa. F462Y conlleva pérdida absoluta de *sensing* de fructosa, manteninedo respuesta a glucosa. No confiere respuesta a galactosa.	Dietvorst *et al*. (2010)
P	atSTP10 (Q9LT15)	F39 L43 Q177 I184 Q295 Q296 N301 N332 F401 G406 W410 N433 T437	F26 T30 Q161 I168 Q282 Q283 N288 N317 F379 G384 W388 N411 N415	TM1 TM1 TM5 TM5 TM7 TM7 TM7 TM8 TM10 TM10 TM10 TM11 TM11	Unión de glucosa.	Paulsen *et al*. (2019)
M	hGLUT1 (P11166)	N34		TM1	Conservado en GLUTs clases I y II. Variantes en GLUT1DS1 y 2. En GLUT1DS1, N34S disminuye a un 55% el transporte de glucosa.	D. Wang *et al*.(2000); D. Wang *et al*. (2005)
		T60 S66 G75 M77		L1 TM2 TM2 TM2	Variantes en GLUT1DS1 e IGE. En GLUTDS1, G75W disminuye transporte de glucosa a un 35%. En IGE, T60M y M77T disminuyen medianamente transporte de glucosa.	D. Wang *et al*. (2000); Arsov *et al*. (2012)
		G91 R92		L2 L2	Variante G91D en GLUT1DS1 disminuye transporte de 3-OMG. R92 posee rol en cambio conformacional inducido por sustrato.	Klepper *et al*. (2001); Schürmann et al. (1997)

Tabla 15. (Continuación)

Taxa[a]	Proteína[b]	Sitio[c]	Ref. hGlut1[c]	Ubicación[c]	Rol funcional	Referencias
M	hGLUT1 (P11166)	M96		TM3	Mutagénesis disminuye actividad de transporte a menos del 10%.	Mueckler et al. (2004)
		R126		TM4	Variantes en GLUT1DS 1 y 2, R126L /H/C disminuye transporte de 3-OMG y DHA.	D. Wang *et al*.(2000); Brockmann *et al*. (2001); Suls *et al*. (2009)
		G130		TM4	En GLUT1DS1, G130S disminuye el transporte de glucosa a un 75%.	Dong Wang *et al*. (2005)
		T137		TM4	Unión de Cito-B.	Kapoor *et al*. (2016)
		Y143 E146		TM4 TM4	Mutagénesis disminuye actividad de transporte a menos del 10%. Variante E146K en paciente GLUT1DS1.	D. Wang *et al*. (2000)
		R153		L4	En GLUT1DS1, R153C disminuye la actividad de transporte de glucosa a un 44%.	Pascual *et al*. (2002); Dong Wang *et al*. (2005)
		Q161		TM5	Unión de GLUT-i1	Kapoor *et al*. (2016)
		L169		TM5	Ausente en GLUT1DS1, disminuye transporte de glucosa a un 48%.	Dong Wang *et al*. (2005)
		L204 P205		TM6 TM6	Mutagénesis de cisteína disminuye actividad de transporte a menos del 10%.	Mueckler & Makepeace, (2009)
		R218		L6	En IGE; R218S disminuye medianamente transporte de glucosa.	Arsov *et al*. (2012)
		R223		L6	Estabiliza configuración endofacial. R223Q no afecta transporte de glucosa. R223Q/P/W no afecta transporte de 2DG. R223P disminuye transporte de 3-OMG ↓Vmax, =Km. R223W perjudica fosforilación por PKC.	Deng *et al*. (2014); Arsov *et al*. (2012); E. E. Lee *et al*. (2015); Suls *et al*. (2009); Leen *et al*. (2010)
		S226		L6	Sitio fosforilación por PKC, activada por TPA. S226 aumenta transporte de 3-OMG inducido por TPA, mientras que S226A no.	E. E. Lee *et al*. (2015)

Tabla 15. (Continuación)

Taxa[a]	Proteína[b]	Sitio[c]	Ref. hGlut1[c]	Ubicación[c]	Rol funcional	Referencias
M	hGLUT1 (P11166)	R232		L6	En IGE, R232C disminuye transporte de 3-OMG.	Striano *et al*. (2012)
		E243		L6	En IGE; E243V disminuye marcadamente transporte de glucosa.	Arsov *et al*. (2012)
		A275		TM7	En GLUT1DS2, A275T disminuye transporte de glucosa,↓Vmax, =Km.	Weber *et al*. (2008)
		Q282		TM7	Inicio motivo QQLS conservado en GLUTs clase I y II. Ausente en GLUT1DS2, disminuye actividad de transporte ↓Vmax, =Km. Unión de Cito-B. Unión de b-NG junto con Q283.	Weber *et al*. (2008); Kapoor *et al*. (2016); Deng *et al*. (2014)
		G286		TM7	G286D en GLUT1DS1 induce perdida de actividad de transporte de glucosa.	Flatt *et al*. (2011)
		N288		TM7	Unión de Cito-B y de b-NG. Mutagénesis de cisteína disminuye actividad de transporte a menos del 10%	Kapoor *et al*. (2016); Deng *et al*. (2014); Mueckler & Makepeace. (2009)
		Y293		TM7	Mutagénesis fija conformación exofacial.	Mori *et al.* (1994)
		T295		L7	T295M en GLUT1DS1, disminuye el transporte de glucosa a un 75%.	Dong Wang *et al*. (2005)
		T310 G314		TM8 TM8	G314S en GLUT1DS2, disminuye el transporte de glucosa, ↓Vmax, =Km.	Wang et al. (2003) Weber *et al*. (2008)
		N317		TM8	Unión de b-NG.	Deng *et al*. (2014)
		S324		TM8	S324L en GLUT1DS2 disminuye transporte de 3-OMG ↓Vmax, =Km.	Suls *et al*. (2009)
		R333		L8	Variantes em GLUTDS1 y 2. R333W en GLUT1DS1, disminuye el transporte de glucosa a un 43%.	Schneider *et al*. (2009); Dong Wang *et al*. (2005)
		F379		TM10	Unión de Glut-i2.	Kapoor *et al*. (2016)

Tabla 15. (Continuación)

Ta-xa[a]	Prote-ína[b]	Sito[c]	Ref. hGlut1[c]	Ubica-ción[c]	Rol funcional	Referencias
M	hGLUT1 (P11166)	G382 G384 P385 I386		TM10 TM10 TM10 TM10	Mutagénesis de cisteína en posiciones 382, 385 y 386 disminuye actividad de transporte a menos del 10%. G384 participa en unión de Cyto-B.	Mueckler & Makepeace. (2009) Kapoor *et al*. (2016)
		W388		TM10	Unión de Cito-B, GLUT-i1, GLUT-i2.	Kapoor *et al*. (2016)
		N411		TM11	N411S en EIG12 disminuye marcadamente transporte de glucosa. Unión de Cito-B.	Arsov *et al*. (2012); Kapoor *et al*. (2016)
		N415		TM11	Unión de b-NG.	Deng *et al*. (2014)
		I435		TM12	Ausente en GLUT1DS2. Pérdida de actividad de transporte de glucosa.	Flatt *et al*. (2011)
		R458		C	R458W en EIG12 disminuye marcadamente transporte de glucosa.	Arsov *et al*. (2012)
	hGLUT2 (P11168)	V197	V165	TM5	V197I en NIDDM abole la actividad de transporte en transportadores expresados en oocitos de Xenopus.	M. Mueckler *et al*. (1994)
	hGLUT3 (P11169)	Q159 Q280 Q281 N286 N315 E378 W386	Q161 Q282 Q283 N288 N317 E380 W388	TM5 TM7 TM7 TM7 TM8 TM10 TM10	Unión de glucosa mediante puentes de hidrógeno con grupos hidroxilo.	Deng *et al*. (2015)
		F24 I162 I166 I285 F289 F377	F26 I164 I168 I287 F291 F379	TM1 TM5 TM5 TM7 TM7 TM10	Residuos hidrofóbicos y aromáticos en entorno de unión de glucosa	Deng *et al*. (2015)

Tabla 15. (Continuación)

Taxa[a]	Proteína[b]	Sitio[c]	Ref. hGlut1[c]	Ubicación[c]	Rol funcional	Referencias
	rGLUT5 (P43427)	Y31 Q166 Q287 H386 S391 H418	F26 Q161 Q282 F379 G384 N411	TM1 TM5 TM7 TM10 TM10 TM11	Unión de fructosa inferida por disminución de fluorescencia de triptófano. Q166 y S391 contribuyen en menor medida a la unión.	Nomura *et al*. (2015)
	hGLUT7 (Q6PXP3)	I302	V290	TM7	Motivo NAI/V conservado en GLUTs clase I y II. I en GLUT2,5,7,9. I302V conlleva pérdida de transporte de fructosa, mientras que transporte de glucosa se ve inalterado. Potencial interacción de I302 con W77 en TM2 (W65 en hGLUT1).	Manolescu *et al*. (2005)
	hGLUT9 (Q9NRM0)	L75 T125 R171 R198 R380	T30 S80 R126 R153 R333	TM1 TM2 TM4 L4 TM9	Variantes presentes en RHUC2. L75R, T125M, R171C, R198C y R380W reducen marcadamente transporte de urato.	Dinour *et al*. (2010); Dinour *et al*. (2012)
	mGLUT10 (Q8VHD6)	G128	P149	L4	G128E induce perdida de transporte de DHA en mitocondria.	Lee *et al*. (2010)
	hGLUT11 (Q9BYW1)	V292	V290	TM7	Mutante DSV>DSI conlleva pérdida de transporte de glucosa y fructosa. Mutante DSV>NAI/V mantiene transporte tanto de fructosa como glucosa.	Manolescu *et al*. (2007)

[a] Categoría taxonímica. B, bacteria; Pr, protozoa; F, fungi; M, metazoa; Pp, primoplantae.
[b] Nombre proteína con prefijo de especie en minúscula y código Uniprot. Prefijos: ec, *E.coli;* cg, *C. glutamicum*; ld, *L.donovani;* lm, *L.mexicana*; ci, *C.intermedia;* sc, *S. cerevisiae;* at*, A.thaliana*; r, *R.norvegicus*; m, *M.musculus;* h, *H.sapiens.*
[c] Numeración según secuencia alojada en Uniprot. Ubicación topológica según topología establecida para PDB de hGLUT1 4PYP en PDBTM.

Para determinar en qué medida los ressiduos corresponden a sitios homólogos, a partir del alineamiento SP obtenido en pasos anteriores se identificó, para cada caso, el residuo de la secuencia de hGLUT1 con el que se encontraba alineado **(Tabla 15).** Los residuos identificados se reducen a 71 sitios en el contexto del alineamiento, lo cual corresponde a un 14,4% de los aminoácidos que conforman

la secuencia de hGLUT1. De éstos, se identificaron los residuos que participan directamente en el reconocimiento de sustrato a lo largo del poro de transporte y residuos contigüos con conocido impacto en actividad de transporte, a partir de los cuales se obtuvo un sub-alineamiento desde el que se desarrolló la reconstrución filogenética local y análisis clúster. Los residuos identificados y seleccionados se presentan en el contexto de la topología y estructura de hGLUT1, en la **Figura 22**. En la sección 5.7.1 se integra a esta información los análisis de conservación de residuos (**Figura 27**).

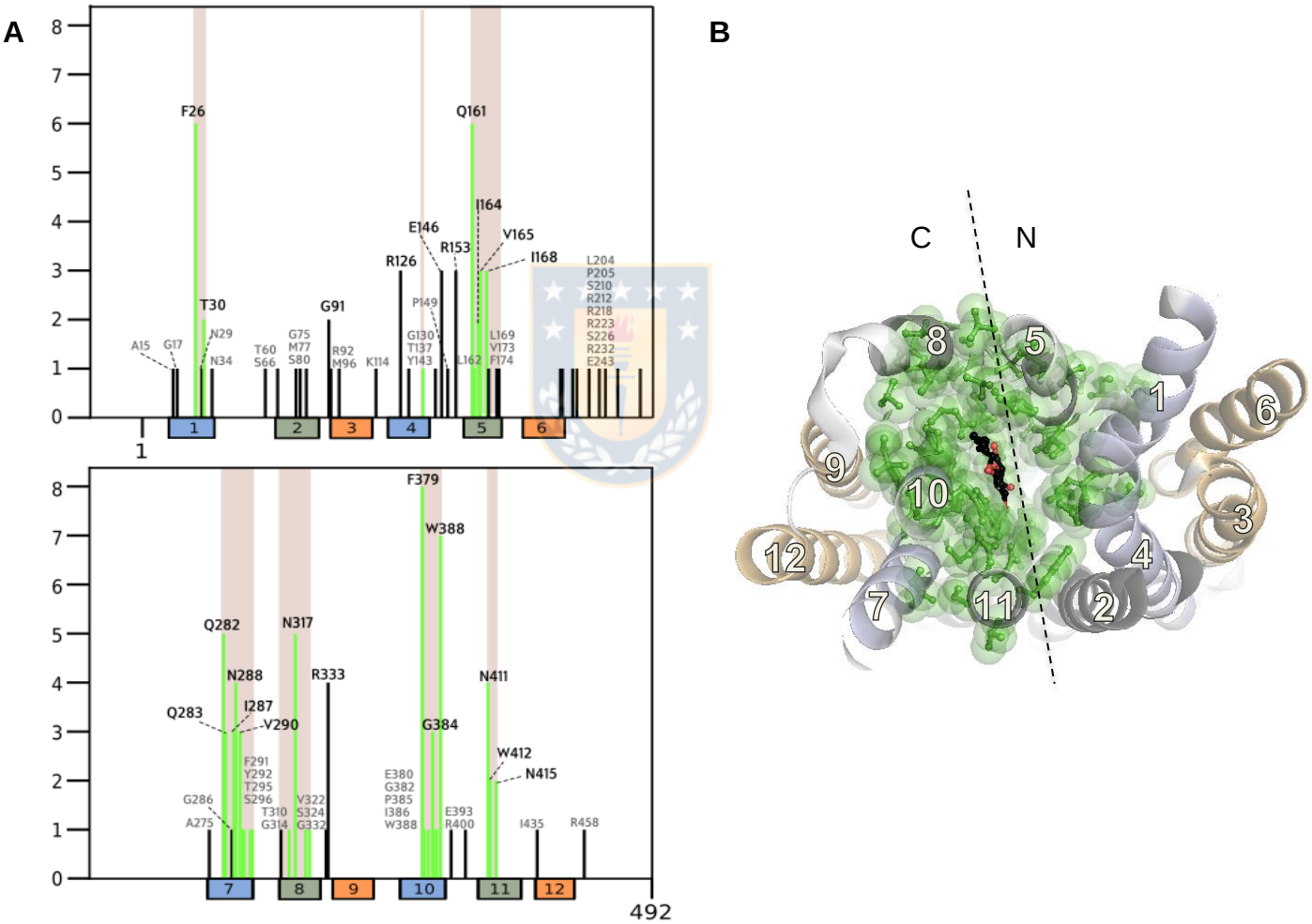


Figura 22. Residuos con rol funcional en la familia SP. (A) Se grafica, para los 71 residuos identificados, el número de transportadores de la familia a los que se le conoce rol funcional con evidencia empírica, según la revisión bibliográfica realizada (Tabla 15), proyectados en topología de hGLUT1. En verde, residuos a menos de 10 A de β-NG en PDB 4PYP. En fondo gris, residuos con conocida participación en unión de sustrato o impacto en actividad de transporte (cavidad central), seleccionados para sub-alineamiento. (B) Residuos seleccionados resaltados en verde en PDB 4PYP, visión endofacial.

5.5 Análisis de anotaciones funcionales

A partir de la obtención y curación de anotaciones funcionales para los 341 transportadores analizados se encontró que, considerando los criterios definidos en la **Tabla 9**, 187 de ellos poseían anotaciones funcionales relativas a la selectividad de sustrato en donde al menos una presenta evidencia experimental, constituyendo el 55% del total. Esto, sin embargo, considera la curación y ampliación de las anotaciones desde la revisión bibliográfica realizada. En la **Tabla 16,** para cada categoría taxonómica, de los 341 transportadores analizados, se indican la cantidad de transportadores con y sin caracterización experimental a nivel de sustrato según los criterios enunciados en la sección métodos, junto a la gama de molćulas transportadas o con afinidad de unión, anotados en las bases de datos.

En la **Figura 23** se presenta la distribución de etiquetas de evidencia de la ontología GO asociadas a las anotaciones funcionales relativas a la selectividad para ambos grupos de proteínas (con y sin caracterización experimental a nivel de sustrato). A su vez, para una mejor idea del nivel de anotación alcanzado por los métodos no experimentales, en la **Tabla 17** se presenta un recuento de las anotaciones asociadas a cada tipo de etiqueta.

Tabla 16. Recuento de transportadores SP con y sin anotaciones funcionales experimentales a nivel de sustrato.

Categoría Taxonómica	N° Sec	Nivel de evidencia de anotación de sustratos*	Nivel de evidencia de existencia de proteína†	Sustratos anotados
Bacteria	44	26 [6→15(+9)]	12 14	Glucosa, fructosa,galactosa, fucosa, manosa, xilosa, arabinosa, arabinobiosa, inositol, hidroxibenzoatos, a-ketoglutarate, arabinatol, ribotol
		18	7 11	
Protozoa	12	10 [0→8(+8)]	5 5	Glucosa, fructosa, galactosa, manosa, ribosa, glucosamina, NaGlcN, inositol
		2	2 0	
Fungi	104	67 [5→34(+29)]	40 27	Glucosa, fructosa, galactosa, galacturonato, gluconato, xilosa, celobiosa, lactosa, manosa, maltotriosa, inositol, lactato, glicerol, fosfato, manitol, xilitol, sorbitol, turanosa, quinato, trehalosa, GroPCho, GroPIns, otros
		37	14 23	
Primoplantae	84	47 [12→26(+14)]	42 5	Glucosa, fructosa, galactosa, manosa, fucosa, glucuronato, xilosa, ribosa, ribulosa, ramnosa, sucrosa, arabinosa, eritrosa, eritriol, xilitol, sorbitol, inositol, glicerol, manitol, fosfato, arsenito
		38	33 5	
Metazoa	97	37 [13→25(+12)]	37 0	DHA, arabinosa, fructosa, fucosa, galactosa,glucosamina, glucosa, inositol, manosa, xilosa, trehalosa, urato
		60	42 18	

***** Con y sin evidencia experimental en gris y blanco, respectivamente. En corchetes, número de proteínas con parámetros cinéticos reportados para al menos un sustrato en Uniprot y total aumentado tras revisión bibliográfica.

† Número superior, cantidad de proteínas con evidencia experimental (a nivel de transcrito o de proteína). Número inferior, cantidad de proteínas sin evidencia experimental (predichas o inferidas por homología).

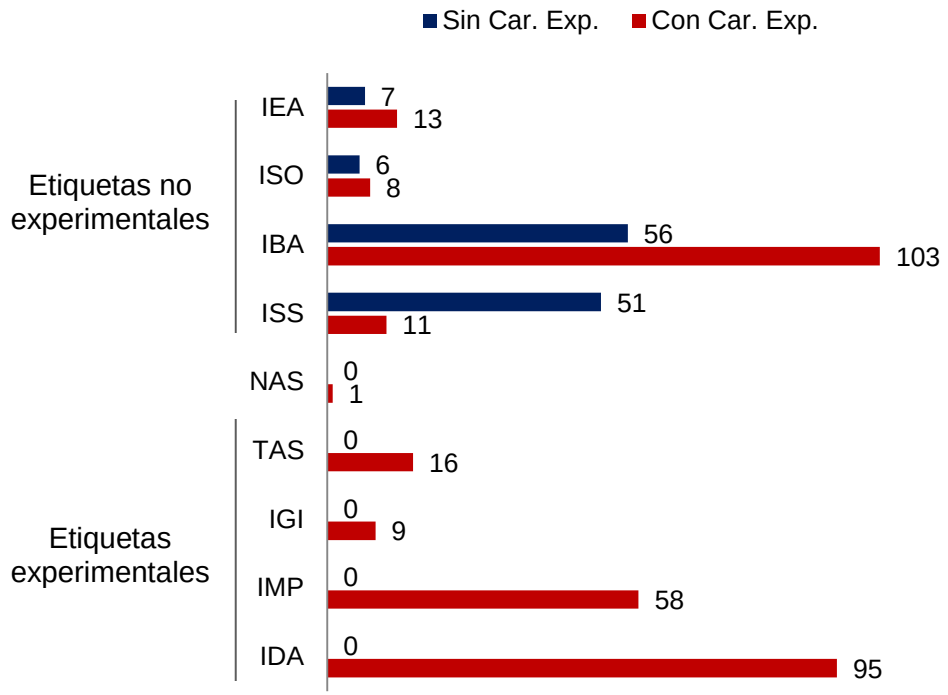


Figura 23. Distribución de etiquetas de evidencia para las anotaciones funcionales alojadas en GO. IEA, inferido por anotación electrónica. ISO, inferido por ortología de secuencia. IBA, inferido por aspecto biológico de ancestro. ISS, inferido por similaridad de secuencia. NAS, sentencia de autor no trazable. TAS, sentencia de autor trazable. IGI, inferido por interacción génica. IMP, inferido por fenotipo mutado. IDA, inferido por ensayo directo.

Tabla 17. Recuento de anotaciones funcionales asociadas a etiquetas de evidencia no experimental

	Anotaciones asociadas a proteínas	
Etiqueta	**Con caracterización experimental a nivel de sustrato**	**Sin caracterización experimental a nivel de sustrato**
IBA	Carbohidrato (73), Hexosa (22), Inositol (4), Ácido carboxílico (2), Glucosa (1), Fructosa (1).	Carbohidrato (49), Inositol (3), Hexosa (2), Ácido carboxílico (2).
ISS	Glucosa (4), Fructosa (4), DHA (1), α-glucosido (1), Maltosa (1).	Glucosa (19), Trehalosa (13), Fructosa (10), DHA (5), Fosfato (2), Azucar (1), Urato (1).
ISO	Glucosa (5), Galactosa (1), Monosacarido (1), xenobiotico (1).	Fosfato (4), Inositol (1), Glicerofosfodiester (1).
IEA	Glucosa (4), Fosfato (3), DHA (1), Fructosa (1), Manosa (1), Sialicilato (1), Glicerol-3-fosfato (1), Xenobiotico (1).	Glucosa (3), DHA (2), Fosfato (2).

Para los sistemas de transporte revisados con caracterización experimental a nivel de sustrato, en **Anexo 4** se indican los sustratos que se sabe pueden ser transportados o bien reconocidos sin existir caracterización experimental directa del transporte. Caen dentro del primer caso, las moléculas cuyo transporte se confirma por ensayos cinéticos o análisis de fenotipos de crecimiento, en medios enriquecidos con sustrato, de cepas con ausencia-presencia de expresión del transportador y estudios relacionados. En el segundo caso se encuentran las moléculas cuya unión se deduce desde ensayos de competición y cinéticas de inhibición. Las moléculas se anotan separadas por coma, de no conocerse su preferencia de unión. Para los casos en que su preferencia de unión se puede conocer o inferir a partir del análisis de parámetros de afinidad (km, ki) o de ensayos de competición de un mismo estudio, también se indica, mediante el empleo del símbolo '>'. A su vez, las moléculas que se sabe no son reconocidas o muy débilmente reconocidas se escriben anteponiendo la palabra *'no*'. Ej: *no-galactosa.* En este caso se incluyen moléculas que poseen valores de Km o Ki > 250 mM, que no muestran inhibición o bien mantienen un 90% o más de la actividad de transporte en ensayos de competición, o moléculas que en análisis de fenotipos de crecimiento no se integran y no permiten el desarrollo celular. No se explicita en la tabla información de capacidades o eficiencias de transporte, no obstante tal información puede encontrarse en las referencias respectivas, según corresponda.

5.6 Análisis filogenético y *clustering*

Una visión global de los árboles sin enraizar obtenidos a partir de la reconstrucción filogenética de las secuencias en estudio desde el alineamiento SP y el sub-alineamiento que considera residuos del sitio de unión se muestran en la **Figura 24(A) y 24(B),** respectivamente. Los distintos clados y hojas se colorean según la categoría taxonómica a la que pertenecen las especies asociadas. A partir del análisis de anotaciones taxonómicas y funcionales se identificaron ramas asociadas al transporte de un tipo de sustrato claramente definido o bien a familias de genes caracterizadas en ciertas especies. La posición relativa de los GLUTs en cada árbol se indica en rojo. Para enraizar el árbol se empleó la rama que agrupa tanto secuencias asociadas al código TC 2.A.1.1 en TCDB y SwissProt como a secuencias que se encuentran clasificadas bajo otros códigos en la base de datos TDCB. Esta rama agrupa a secuencias asociadas al transporte de glicerofosfodiésteres y fosfato pertenecientes a la familia de cotransportadores fosfato:H+ código TC.2.A.1.9 (PHS *family*), secuencias asociadas a la familia de cotransportadores de ácidos aromáticos:H+ alojadas en TCDB bajo el código 2.A.1.15 (AAHS *family*) y secuencias asociadas a la familia de transportadores de polioles lineales representadas por las proteínas transportadoras de D-arabinatol y ribitol de *Klebsiella pneumoniae* y el transportador de csbX de *Bacillus subtilis*, alojadas bajo el código 2.A.1.18 de TCDB (PP *family*). Se decidió usar esta rama para enraizar el árbol, debido a su longitud y su buen nivel de soporte (*bootstrap* de 72) y la presencia de estos grupos de divergencia funcional. El filograma enraizado y anotado se presenta en la **Figura 25**, junto con la matriz de similaridad global de las secuencias obtenidas desde el alineamiento. Se adjunta además código QR para acceder a filograma ampliado, con secuencias enumeradas según su posición en el árbol (1 a 341), y con los sustratos con evidencia experimental conocidos anotados para cada proteína. En el **Anexo 5** se presentan las matrices de identidad y similaridad para los distintos clados de interés, calculados mediante el programa Fasta36.

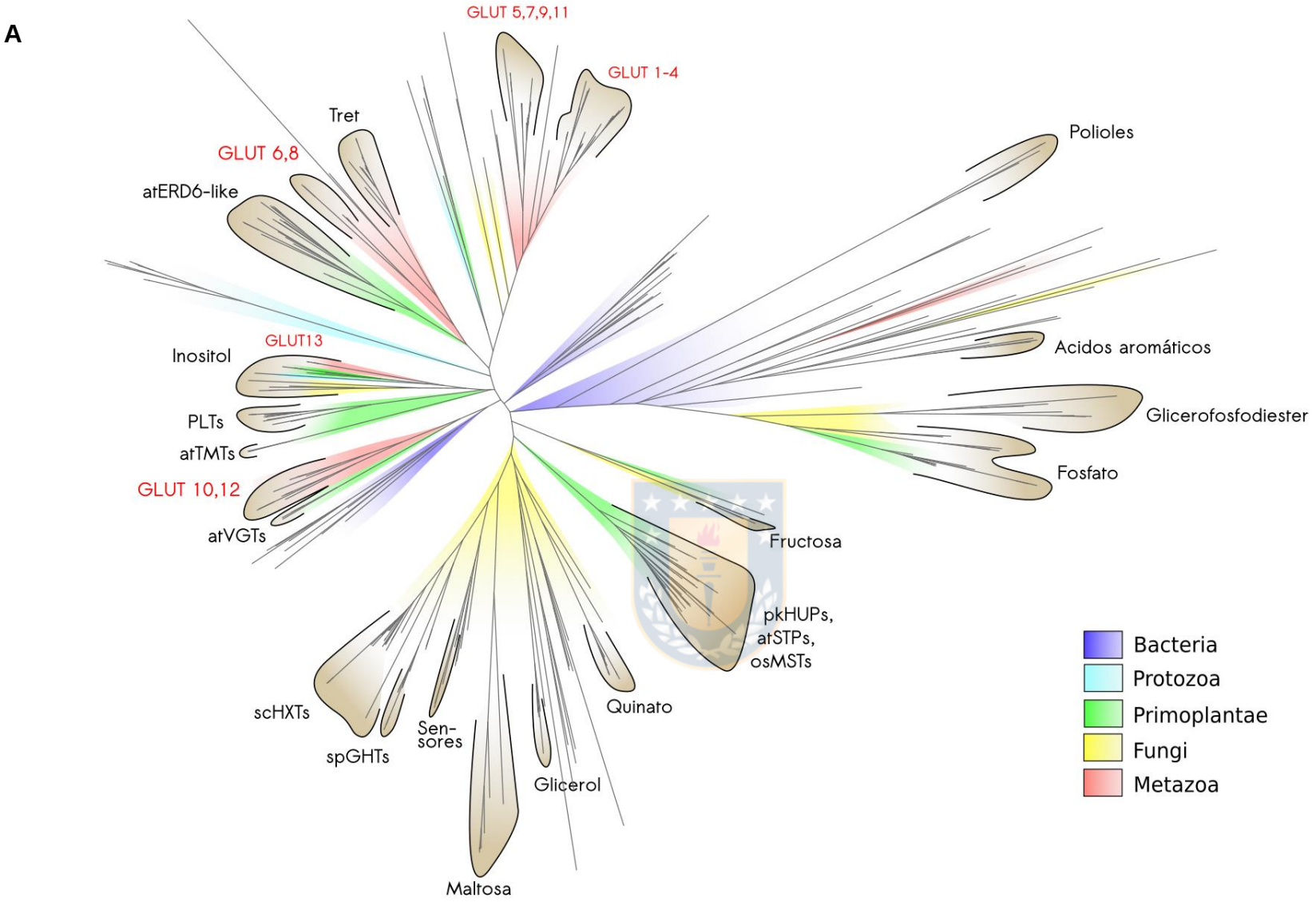


Figura 24. Filogramas de la familia SP y secuencias relacionadas. Reconstruciones filogenéticas para el alineamiento global (A) y el sub-alineamiento integrado por residuos de reconocimiento de sustrato (B) de las 341 secuencias en estudio, mediante método de máxima verosimilitud implementado en programa RAxML 7.2.7. Se anotan ramas asociadas al transporte de un tipo de sustrato claramente definido o bien a familias de genes bien caracterizadas en ciertas especies. La posición relativa de los GLUTs se resalta en rojo.

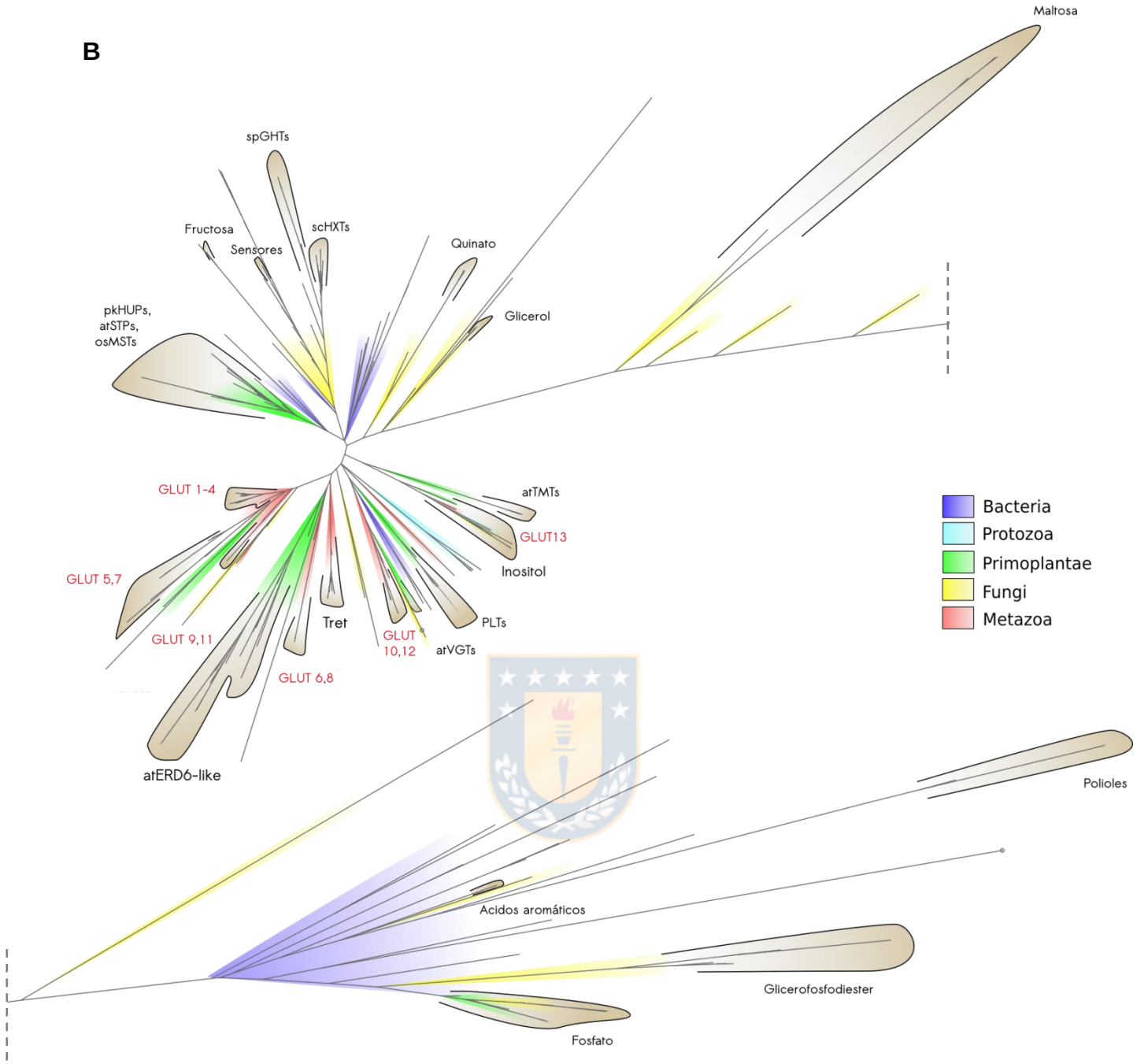


Figura 24 (Continuación).

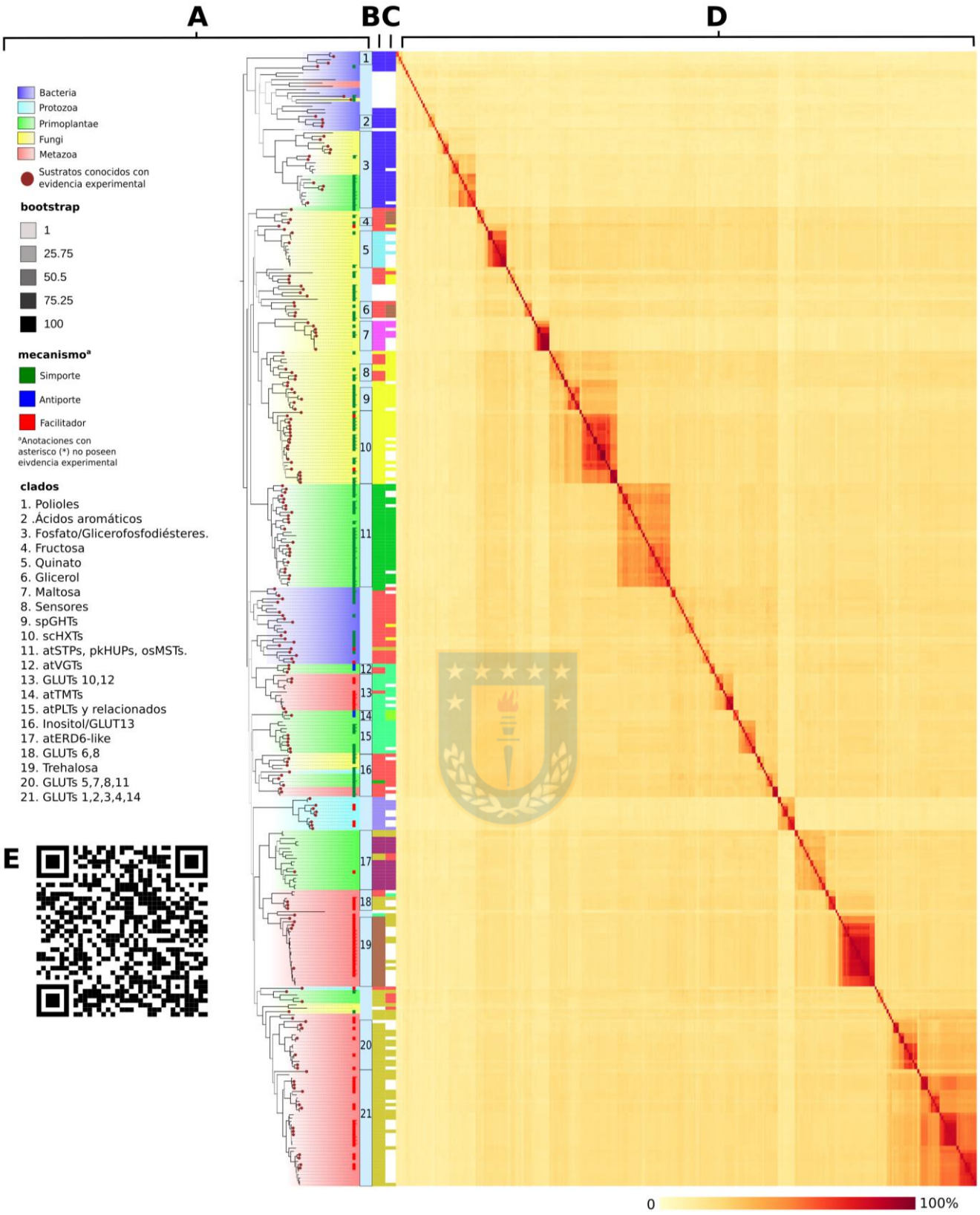


Figura 25. Filograma y análisis cluster de la familia SP. (A) Filograma enraizado y anotado de las 341 secuencias en estudio. Anotaciones y simbologías explicadas en imagen. (B, C) Clústers identificados mediante MCA y k-means del sub-alineamiento de residuos de reconocimiento de sustrato, sin aplicar y tras previa aplicación de filtro de redundancia del 95%, respectivamente. (D) Matriz de similaridad global de las secuencias. (E) Filograma ampliado con sustratos con evidencia experimental conocidos anotados.

5.7 Caracterización de la familia y de los agrupamientos a nivel de secuencia

5.7.1. Análisis de conservación

A partir del análisis de entropía de Shannon para los alineamientos SP y PfamSP se identificaron sitios conservados y altamente conservados, definidos como sitios con una entropía de Shannon (SE) menor o igual a 1.0 y 0.5, respectivamente. Los sitios conservados para ambos casos se presentan en la **Tabla 18.** Se observa correspondencia en la gran mayoría de posiciones, con tendencia a presentar valores similares, a excepción de la posición 479. El resto de posiciones presentan una diferencia de entropía máxima de 0.36, para las posiciones 27 y 134. Se identifican también, en ambos alineamientos, los 6 motivos conservados en la familia descritos en literatura y bases de datos. Posteriormente, se procedió a realizar un análisis de entropía relativa (RE) para el alineamiento SP. En este caso, las posiciones con valores más elevados corresponden a sitios más conservados, siendo los sitios cuya distribución de frecuencia de aminoácidos más se desvía de la distribución en la base de datos no redundante de Blast. Ambos perfiles de entropía se contrastan en la **Figura 26**. Se puede observar que ambos perfiles son similares, sin embargo el perfil de RE identifica mayor cantidad de sitios conservados considerando un valor de corte de 1.5 (64), incluyendo a la gran mayoría de los sitios identificados por entropía de Shannon.

Debido a la presencia en el filograma de una rama que agrupa secuencias asociadas en la base de datos TCDB a las familias de transportadores de fosfato (TC 2.A.1.9), ácidos aromáticos (TC 2.A.1.15) y polioles (TC 2.A.1.18), además de secuencias identificadas como propias de la familia SP, se analizó el nivel de conservación de estos motivos en tales secuencias. En la **Figura 27(A)** se muestra la rama aludida con las posiciones del alineamiento correspondientes a estos motivos anotados para las secuencias. En la **Figura 27(B)** se muestra el mismo análisis generalizado para todo el filograma.

Tabla 18. Residuos conservados en la familia SP

Ref. hGlut1		H SP	H PfamSP	Ref. hGlut1		H SP	H PfamSP
G27	TM1	0.36	0.72	P211	L6	0.66	0.49
T47	L1	0.93	2.28	A224	L6	0.90	0.79
G75	TM2	0.75	0.89	L228	L6	0.92	1.34
G79	TM2	0.78	0.65	D240	L6	0.92	2.44
G91	L2	0.21	0.2	G286	TM7	0.73	0.95
R92	L2	0.52	0.4	Y293	TM7	0.70	1.18
R126	TM4	0.41	0.61	G302	L7	0.60	1.17
G130	TM4	0.08	0.26	G332	L8	0.53	0.55
G134	TM4	0.37	0.73	R333	L8	0.50	0.22
E146	L4	0.19	0.1	R334	TM9	0.94	0.85
R153	L4	0.29	0.21	E393	L10	0.40	0.23
G154	TM5	0.61	0.87	R400	L10	0.52	0.38
G167	TM5	0.16	0.34	P453	C	0.68	1.05
W186	L5	0.51	0.62	E454	C	0.34	0.33
P187	TM6	1.24	0.96	T455	C	0.58	0.82
P208	L6	0.83	1.76	G457	C	0.95	1.32
E209	L6	0.74	0.67	P479	C	0.22	n.d
S210	L6	0.69	0.52	D489	C	0.98	n.d

*Se resaltan los 6 motivos de secuencia conocidos para la familia.

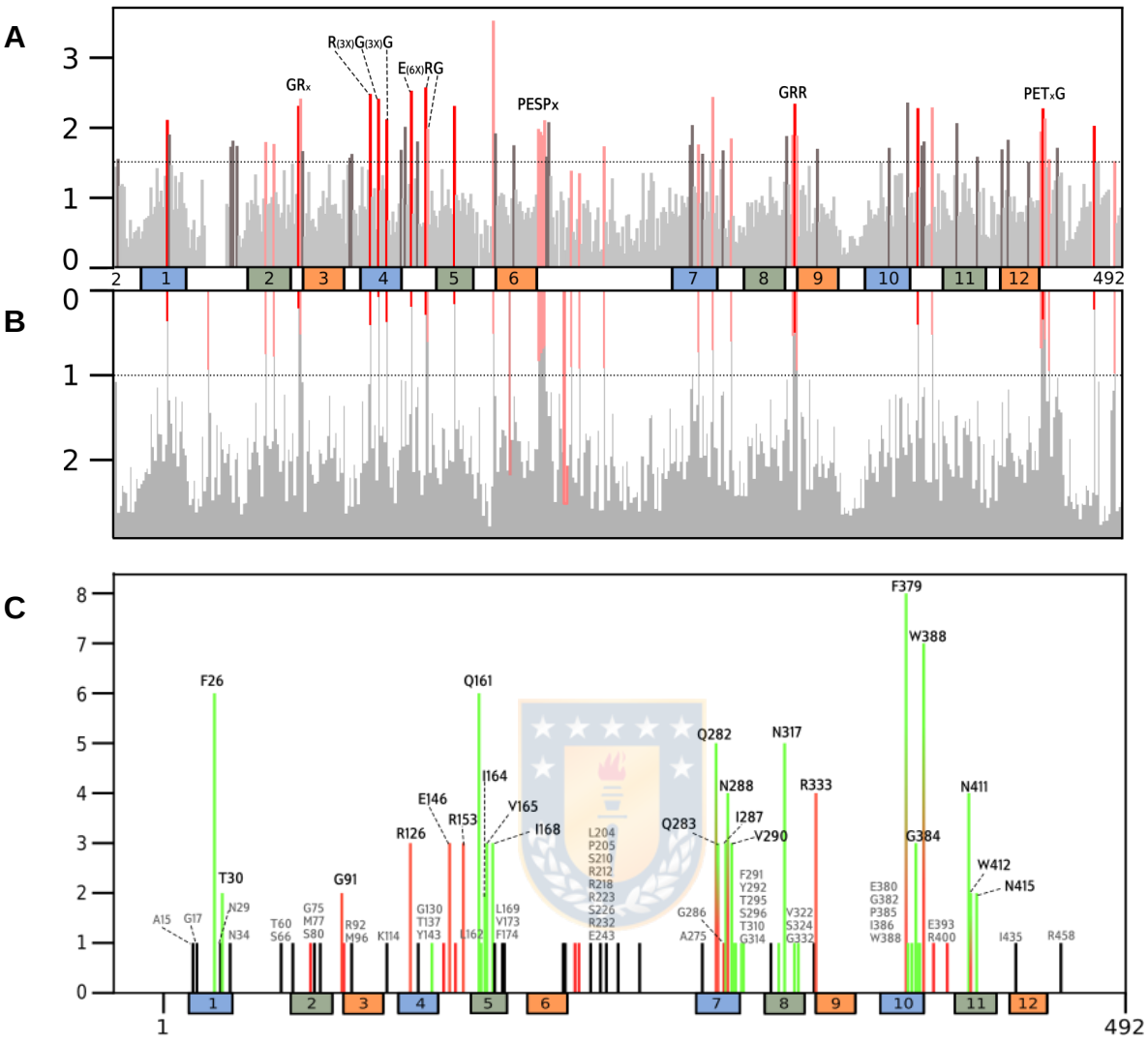


Figura 26. Perfiles de entropía de Shannon y entropía relativa de la familia SP. Se usa como referencia la secuencia de hGLUT1**.** (A) Perfil de entropía relativa. Línea horizontal indica valor de corte de 1.5. (B) Perfil de entropía de Shannon. Para ambos perfiles se destecan en rojo y rosado los sitios conservados identificados por análisis de entropía de Shannon, con valores de entropía menores o iguales a 0.5 y a 1, respectivamente. En el perfil de entropía relativa se destacan en gris oscuro los sitios con un valor de entropía relativa mayor o igual a 1.5, y se indican los motivos conservados conocidos en la familia. (C) Anotación de residuos conservados identificados tanto por análisis de entropía de Shannon en perfil de residuos funcionales de la familia (Figura 22). Los residuos conservados se destacan en rojo.

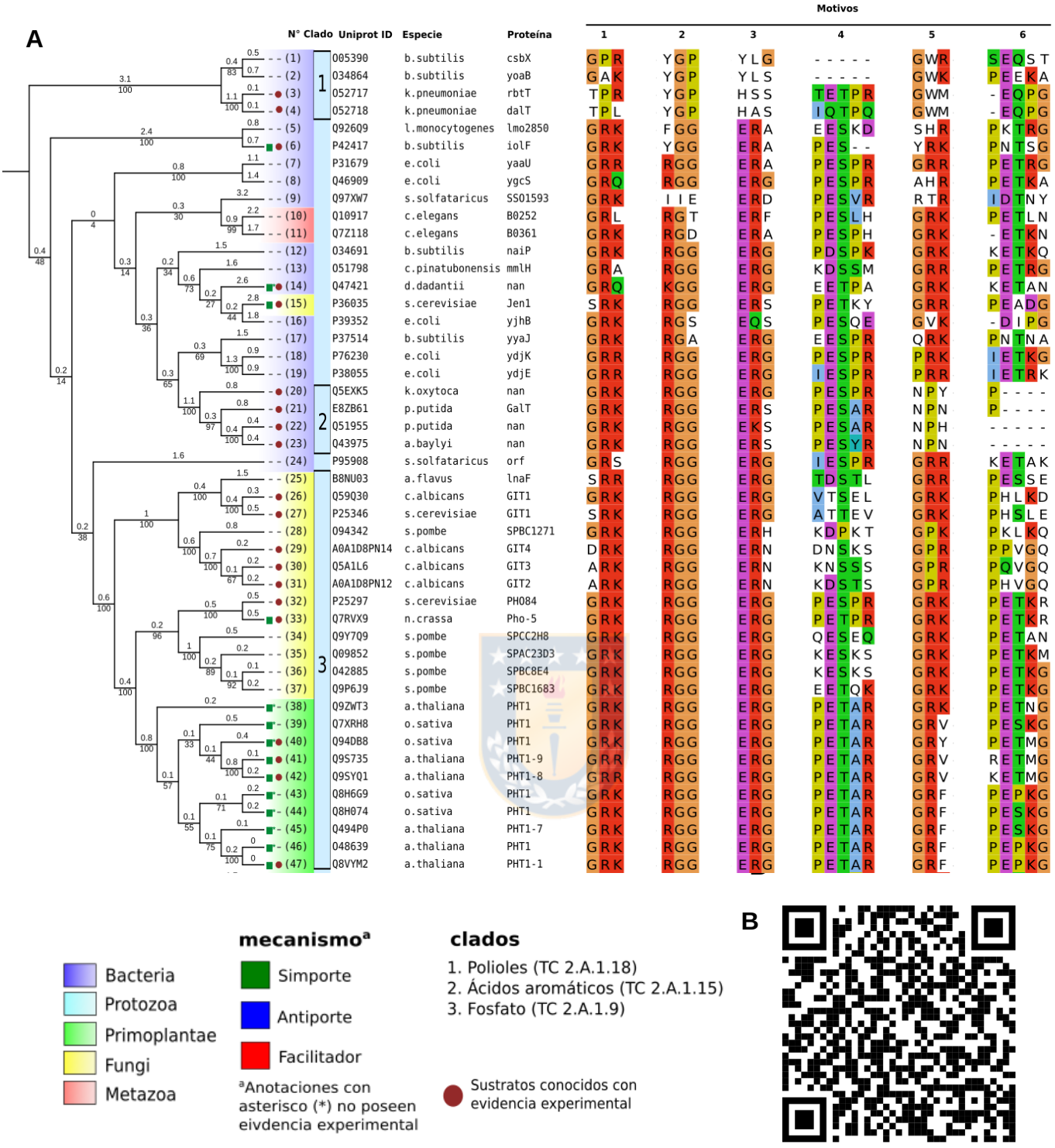


Figura 27. Anotación de motivos conservados en filograma de la familia. (A) Análisis de presencia de motivos conservados para rama que agrupa a secuencias pertenecientes a las familias TC 2.A.1.9, 15 y 18. Motivos 1 a 6 corresponden a residuos 91-93; 126, 130, 134; 146,153, 154; 208-212; 332-334 y 453-457 en hGLUT1, respectivamente. (B) Análisis de motivos generalizado para todo el filograma.

5.7.2. Análisis de correlación de residuos mediante métodos basados en Información Mutua

Para identificar pares de posiciones con un perfil mutacional correlacionado se computaron, para los alineamientos SP y PfamSP, las matrices MI, ver **Figura 28(A)**, aplicando luego las correciones por APC, ASC, las normalizaciones por entropía conjunta, entropía conjunta mínima, suma de entropías y análisis de información directa. Para cada caso, tras aplicar normalización por z-score, se ordenaron los pares de resíduos según su z-score de mayor a menor.

Para evaluar qué método presentó el mejor nivel de discriminación entre MI-sf y MI-b para ambos conjuntos de datos, se determinó el porcentaje de pares de residuos con valores de z-score mayores o iguales a determinados valores de corte que corresponden a contactos efectivos (verdaderos positivos, VP) en la estructura tridimensional de hGLUT1, 4PYP, contactos definidos como pares de residuos cuyos átomos se encuentran a distancias iguales o menores a 10 A. Se testearon valores de corte de 2SD en adelante y hasta el número entero máximo compartido por todos los métodos para el mismo conjunto de datos, en intervalos de 1 SD. Los resultados de los métodos testeados para ambos alineamientos se presentan en la **Figura 29.**

Se observa que, para el alineamiento PfamSP, el método DI es el que presenta mejor precisión, observándose que los pares de residuos correlacionados identificados por sobre las 2SD alcanzan una tasa de VP cercano al 85% según el criterio mencionado, logrando una reproducción cercana del mapa de contacto del plegamiento MFS, como se muestra en la **Figura 28(B, C)**; y superando el 95% de VP los pares que se encuentran por sobre los 8SD. En seguido lugar se encuentran los métodos MI-APC y MI-ASC, donde los pares de residuos identificados por sobre las 4SD alcanzan una tasa de VP cercana al 90%, lo cual concuerda con los valores de corte reportados en literatura (Dunn et al., 2008).

Como se muestra en la **Figura 28(D),** la matriz de correlación obtenida mediante MI-APC también presenta una buena correspondencia con los mapas de contacto experimentales. En cambio, para el alineamiento SP, MI-APC y MI-ASC presentan un relativo mejor rendimiento que DI, observándose nuevamente la mejor precisión por sobre los 4SD, alcanzando cerca de un 55% de VP. Destaca, sin embargo, la baja en la tasa de VP por sobre las 5SD, lo que implica los pares de posiciones mejor rankeados tienden a encontrarse a más de 10A de la estructura tridimensional evaluada, lo cual puede surgir de acoplamientos indirectos. En este caso, no se logra observar una clara correspondencia entre el perfil de correlación obtenido mediante MI-APC y los mapas de contacto experimentales, ver **Figura 28(D)**. El relativo pobre desempeño del método DI ha de deberse al limitado número de secuencias del alineamiento, puesto que este método ha demostrado requerir un número de secuencias efectivas del orden de 1000 o superior para la predicción efectiva de mapas de contacto (Morcos et al., 2011), lo cual sí cumple el alineamiento PfamSP.

A partir de lo anterior se decidió trabajar con las matrices DI y MI-APC del alineamiento PfamSP para la identificación de redes de correlación asociadas a los residuos funcionalmente importantes identificados mediante revisión bibliográfica, las cuales se muestran en la **Figura 30**.

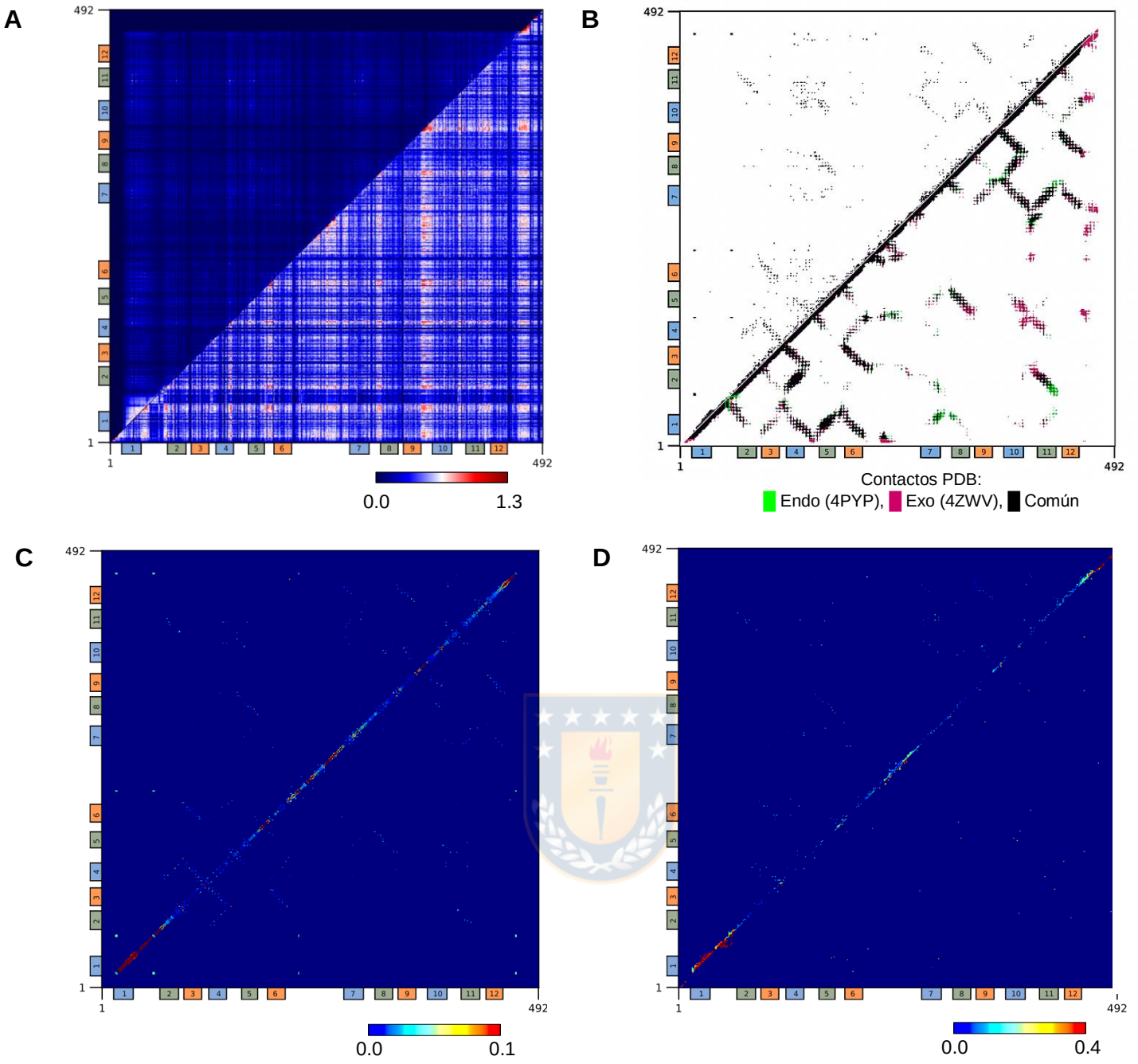


Figura 28. Análisis de correlación de residuos mediante métodos basados en Información Mutua. (A) Información mutua de los alineamientos PfamSP y SP en matriz triangular superior e inferior, respectivamente. Se observa evidente ruido estadístico. (B) Comparación entre mapa de contacto predicho por DI desde alineamiento PfamSP (matriz triangular superior) y mapas de contacto de GLUT1 y GLUT3 en conformación endo y exofacial (matriz triangular inferior, PDBs 4PYP y 4ZWV respectivamente). Se grafican pares de residuos con z-score >2.Para PDBs, contactos definidos como pares de residuos con una distancia menor o igual a 10 A, considerando todos los átomos. (C) Perfil de correlación obtenido mediante DI para alineamiento PfamSP. Se grafican pares de residuos con z-score>4. (D) Perfil de correlación obtenido mediante método MI-APC para los alineamientos PfamSP y SP en matriz triangular superior e inferior, respectivamente.

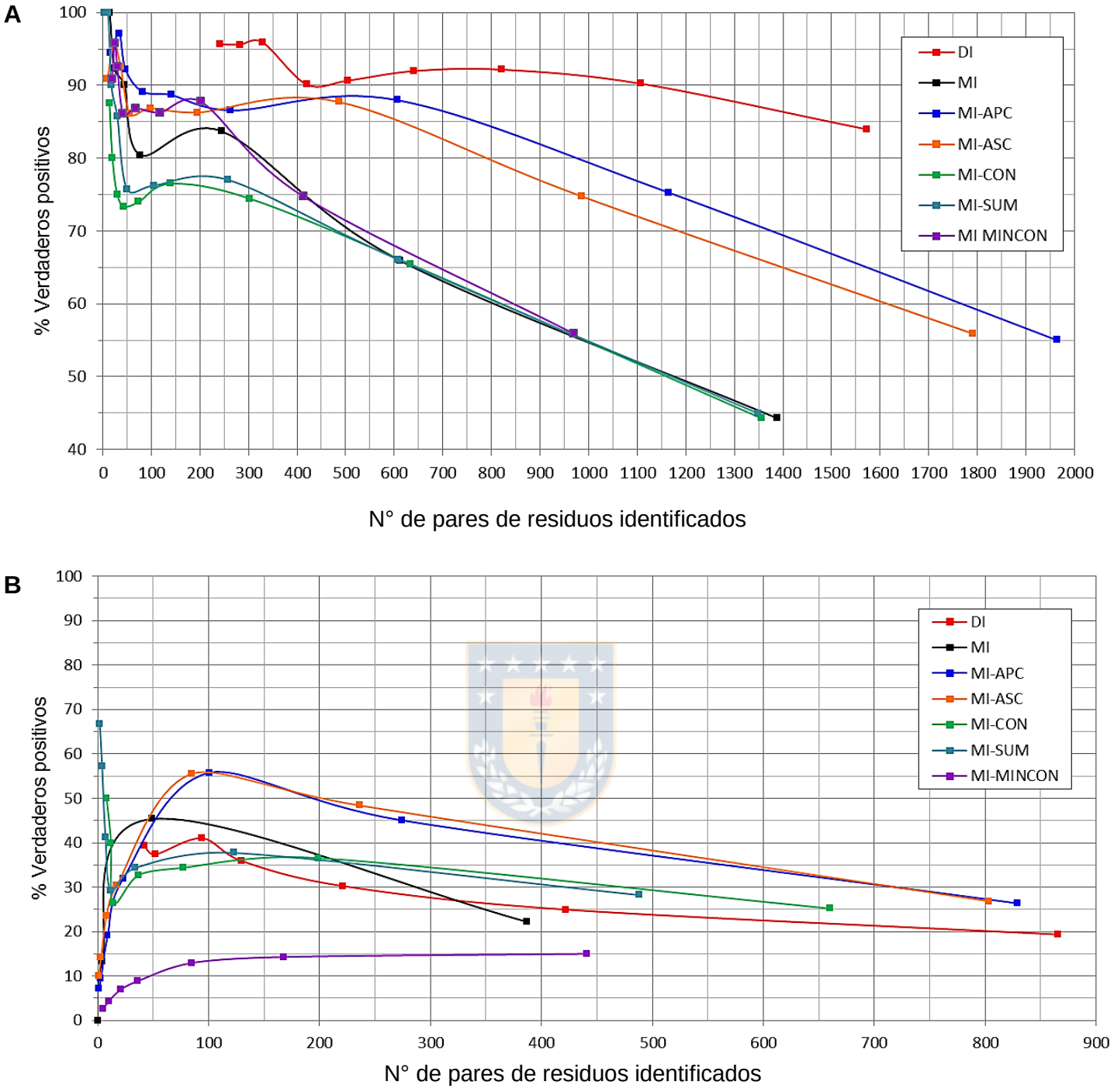


Figura 29. Evaluación de métodos basados en MI. (A) Resultados para alineamiento PfamSP. (B) Resultados para alineamiento SP. Para cada método se grafica en la abscisa, de derecha a izquierda, el número de pares de residuos cuyo valor de MI normalizado por z-score se encuentra por sobre los 2SD en adelante, en intervalos de 1SD, y en la ordenada, el porcentaje de pares de residuos que corresponden a contactos efectivos en la estructura de hGLUT1, PDB 4PYP, definidos como pares de residuos cuyos átomos se encuentran a distancias iguales o menores a 10 A.

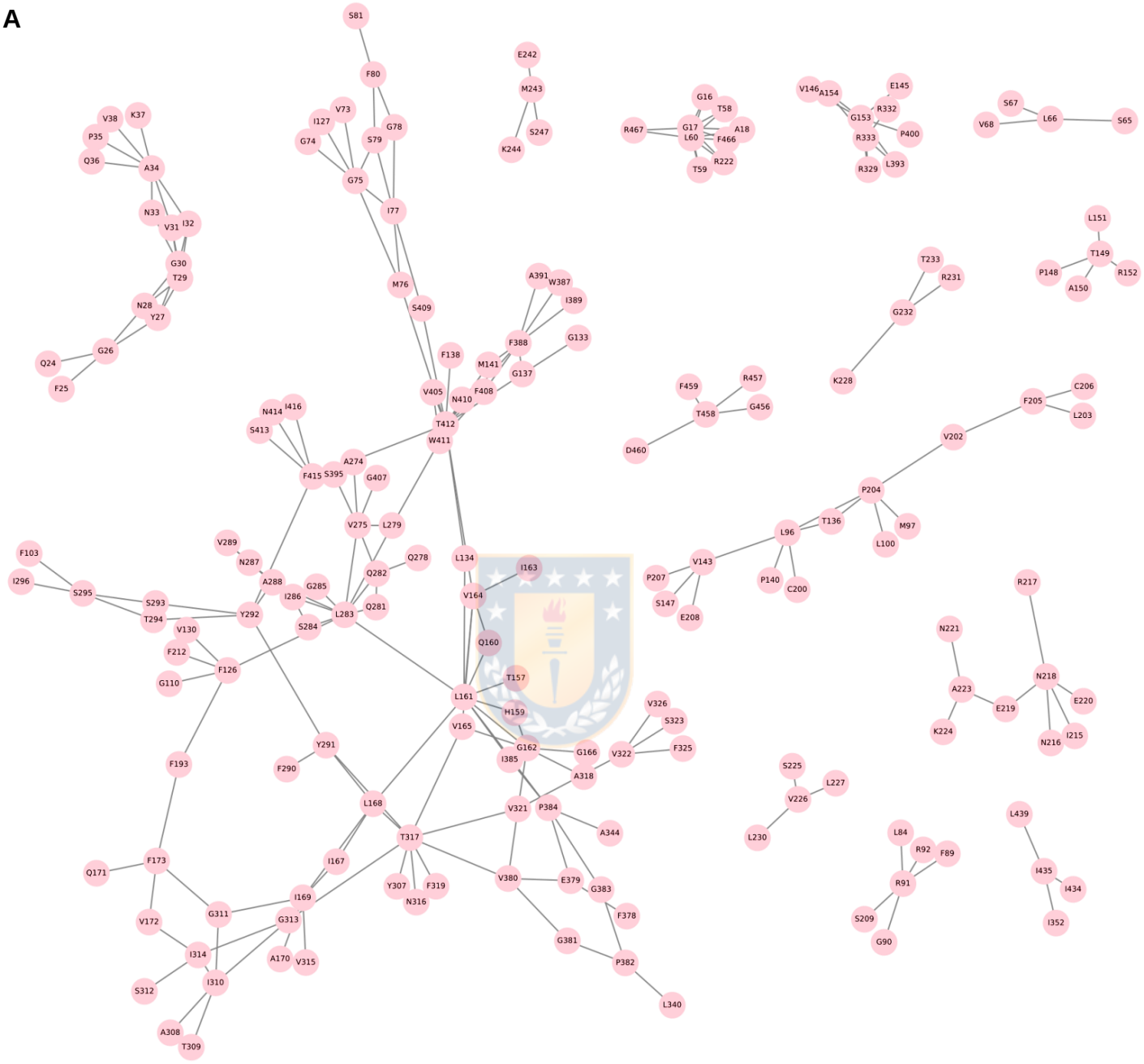


Figura 30. Red de correlación de residuos funcionalmente relevantes en la familia SP. Redes de correlación obtenidas mediante (A) información directa, (B) MI-APC para el alineamiento PfamSP. Se grafican pares de residuos con valores de z-score ≥4.

B

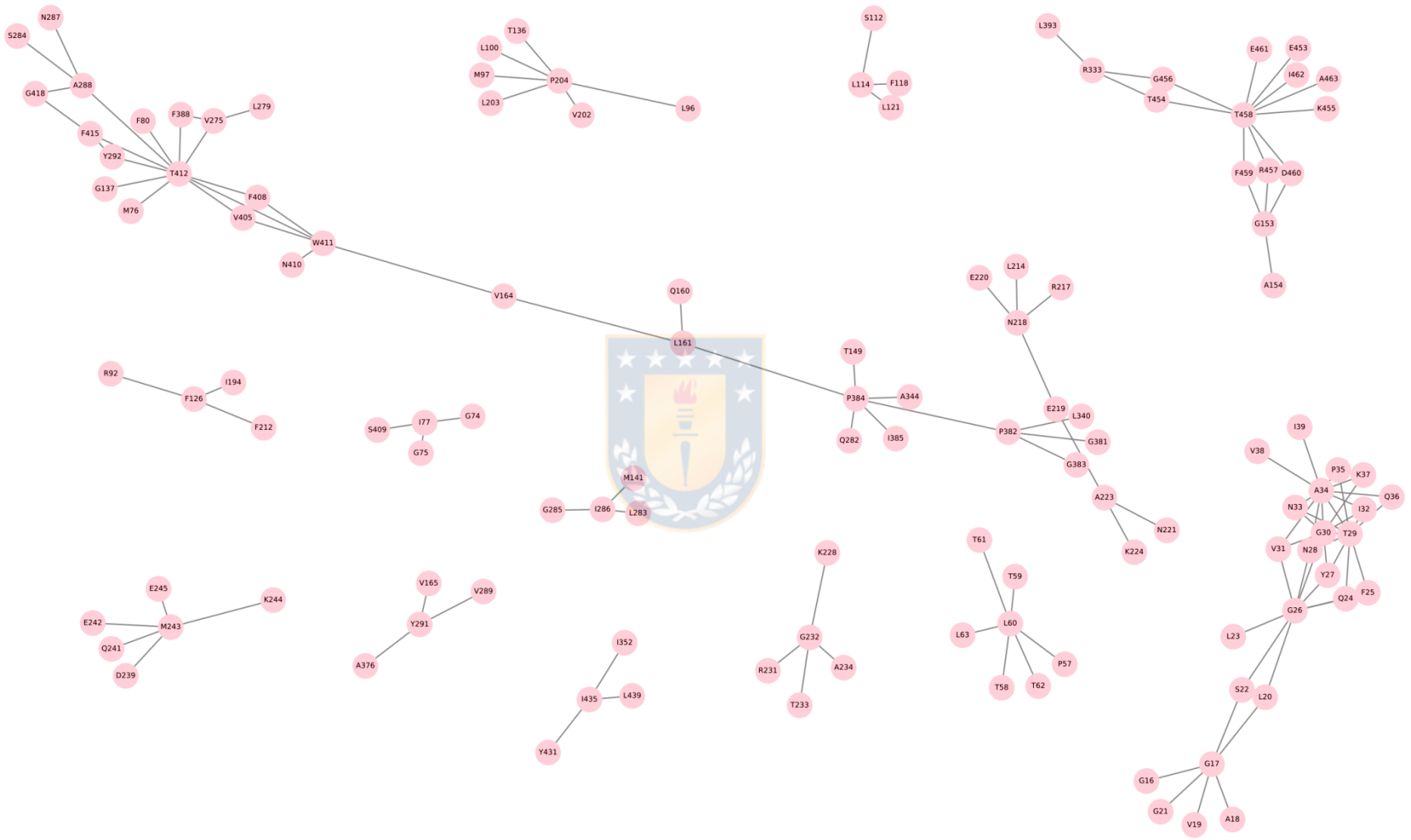


Figura 30. (Continuación).

5.7.3. Análisis de correlación de residuos mediante SCA

La descomposición espectral de la matriz de correlación del alineamiento SP obtenida mediante SCA reveló 11 autovalores (λ), de un total de 341, que emergen sobre la línea espectral de los autovalores obtenidos a partir de 100 aleatorizaciones de la matriz, como se muestra en la **Figura 31**. Estos componentes se asocian a grupos de residuos correlacionados cuyos acoplamientos se consideran estadísticamente y, en potencialidad, biológicamente significativos, puesto que no son reproducidos por el proceso de aleatorización. Para definir las posiciones del alineamiento con contribuciones significativas a cada uno de los componentes, se ejecuta un análisis de componentes independientes (ICA, véase Rivoire et al, 2016), el cual transforma los k primeros autovectores de la matriz de correlación en k componentes independientes (IC). ICA produce una representación en que la mayoría de los grupos de residuos que están débilmente correlacionadas se agrupan cerca del origen del espacio ICA. Para identificar los grupos de residuos cuya correlación es significativa, se ajusta el espacio ICA a la distribución t y se seleccionan los residuos asociados a los IC encontrados por sobre el valor de corte de p = 0,95; ver **Figura 32.**

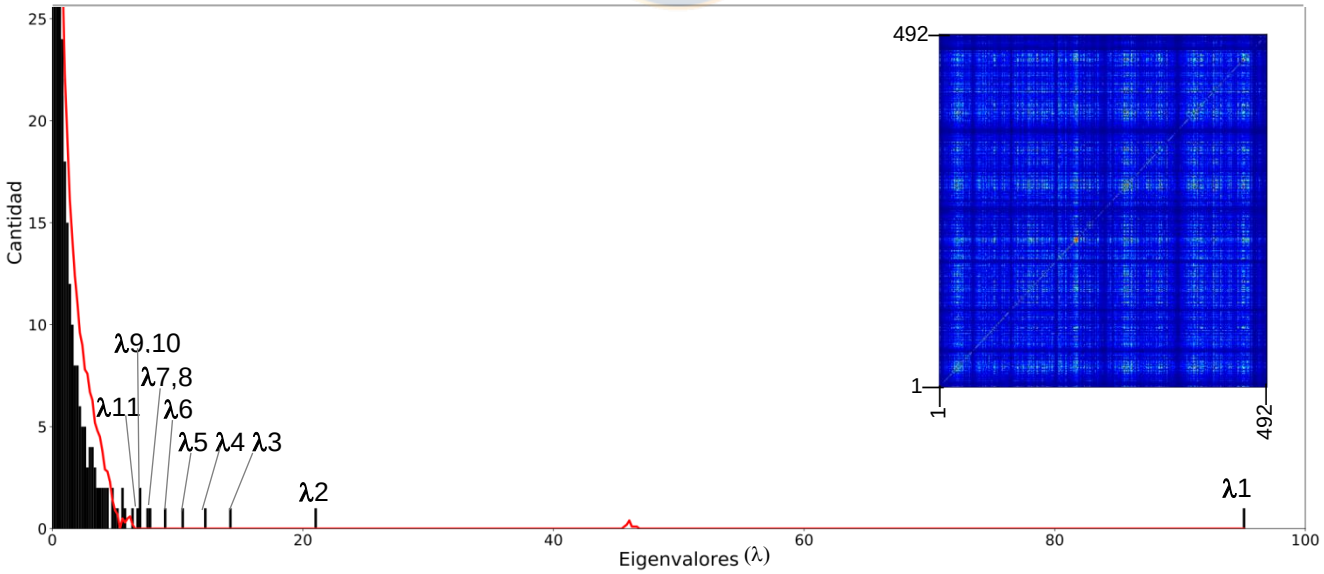


Figura 31. Autovalores de la matriz de SCA para el alineamiento SP: Visión general de la matriz SCA en inserto de esquina superior derecha. En negro, histograma de los autovalores de la matriz SCA. En rojo, línea espectral de los autovalores determinados a partir de 100 aleatorizaciones de la matriz del alineamiento original.

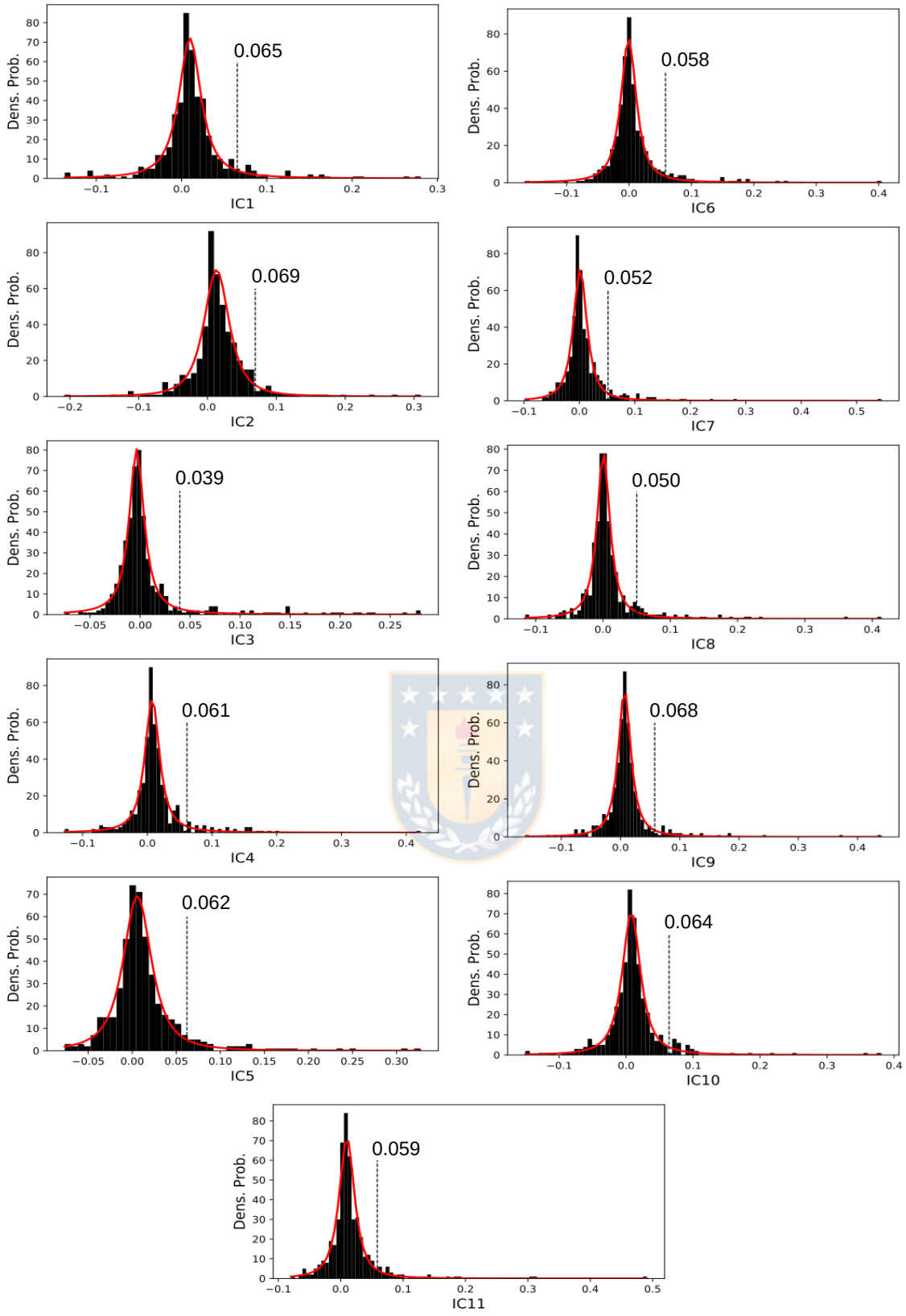


Figura 32. Definición de grupos de residuos correlacionados en la familia SP. Por cada espacio IC se ajusta una distribución t (línea roja), seleccionando los residuos significativos según un valor de corte de 0.95 (línea punteada).

Luego se examinó la estructura de la matriz SCA considerando sólo las posiciones identificadas que contribuyen a los autovectores superiores (λ1-11), **Figura 33(A)**. Esto permite examinar qué grupos de residuos son estadísticamente independientes y cuáles representan desgloses jerárquicos de un grupo mayor. El grupo 1, muestra un claro acoplamiento con el grupo 6, y este último con el grupo 8. Estos dos últimos muestran, a su vez, algún grado de asociación con el grupo 3. El grupo 2 no muestra correlación con otros grupos y se considera independiente. El sector 7 también parece surgir de la fragmentación del sector 6- El resto de componentes muestran mayor independencia. Para una mejor visualización de lo descrito, los grupos se muestran reordenados en la **Figura 33(B).** En la **Tabla 19** se presentan los residuos asociados a cada grupo, usando como referencia la numeración de residuos de hGLUT1.

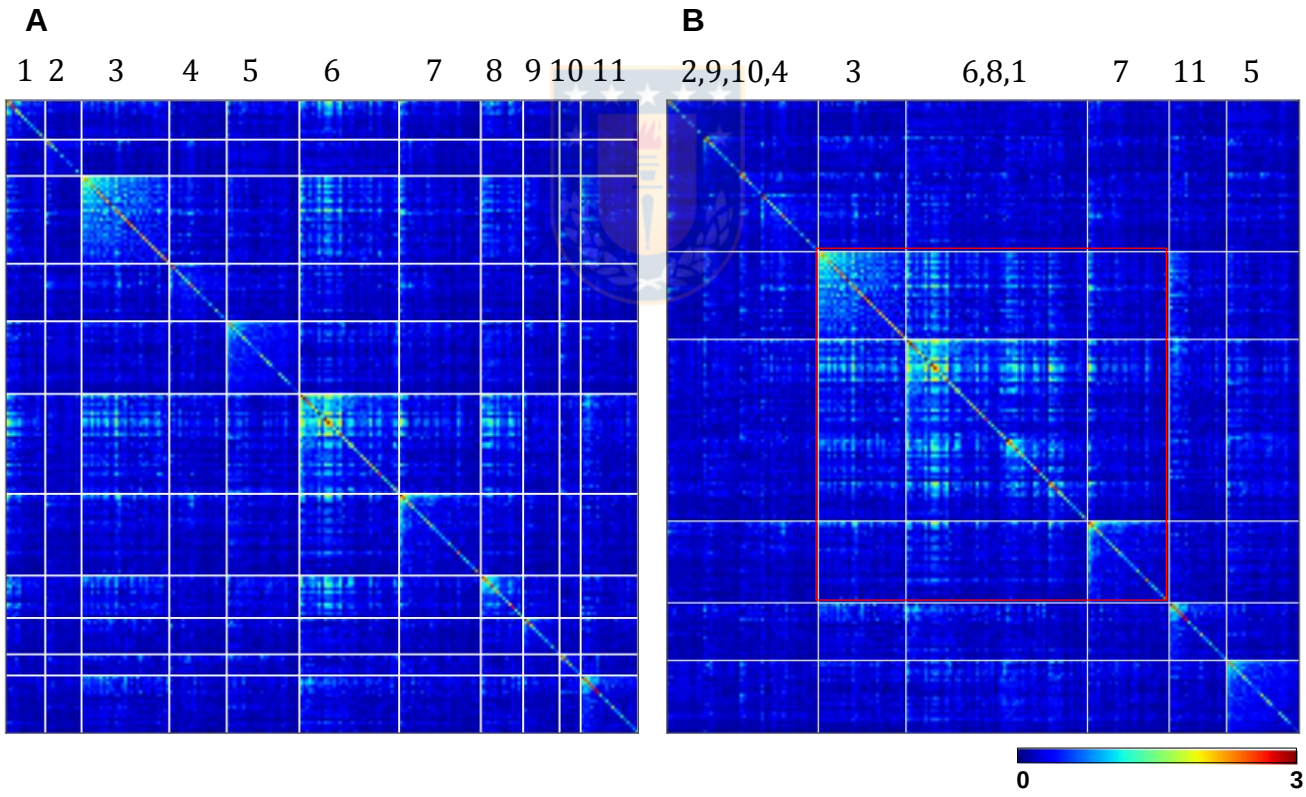


Figura 33. Matriz SCA de los grupos de residuos acoplados identificados. (A) Submuestra de la matriz SCA original, considerando sólo posiciones que contribuyen a los grupos de residuos identificados. Conrresponden a un total de 209 residuos, agrupados según su pertenencia a cada grupo. Éstos se muestran delimitados por líneas blancas y enumerados en zona superior (B) Rearreglo de los grupos para mejor visualización de interdependencia. Grupos con mayor correlación resaltados en recuadro rojo.

Tabla 19. Grupos de residuos correlacionados identificados mediante SCA

Grupo	N° residuos	Residuos*
1	13	L67, S148, **Q161, V165**, F206, G233, V237, L284, **N288**, **N317**, R334, **E380**, F416
2	12	T9, S68, G79, N88, M121, V381, V425, Q427, I436, F444, T459, K477
3	29	M1, A19, G22, **G91**, **R92**, **R126**, **G130**, G134, **E146**, **R153**, G154, G167, P208, E209, P211, **R212**, L214, A224, D240, G302, **G332, R333**, **E393**, E454, T455, K456, G457, D461, E462
4	19	Q25, P36, Y44, R93, T158, Q172, L189, **S210**, F213, E220, **R223**, M244, V277, **S324**, H337, G343, A348, F460, A464
5	24	V16, S23, **N29**, V39, E42, V83, G84, F127, Y132, G145, G175, L176, S178, S191, V203, I259, S285, A371, F373, **G382**, P383, Q397, V418, F467
6	33	**F26**, G27, G31, A35, A70, A103, S106, S113, G125, C133, V140, P141, **Y143, P149**, **I164**, N219, L228, **R232**, D236, R269, **Q282, Q283, G286, I287, V290**, Y293, L336, **G384, W388**, A392, L394, F434, F447
7	27	**T30**, I33, V69, **G75, T137, L204**, L231, **E243**, M251, F263, **A275**, L278, Q279, L280, **Y292, T295,** F375, P387, V391, A403, **W412, N415**, Q423, G430, V452, I463, P479
8	14	**I168**, S281, F298, **G314**, V316, T321, V328, M344, V376, **F379, P385, N411**, F437, F450
9	12	L24, Y28, **S80**, A107, L109, I123, **L169**, A171, A197, G340, A377, **R400**
10	7	**N34,** L122, M142, P187, T352, S396, Y432
11	19	V32, I40, M110, G111, L156, **L162**, W186, L190, P196, L262, P271, I272, I297, **T310**, V323, E329, F395, F422, P453
Total =	209	

*En rojo, residuos conservados según análisis por entropía de Shannon. Resaltados en negro, residuos con algún rol funcional descrito en literatura (Tabla 15).

El grupo de residuos n°1 está conformado por 13 residuos, de los cuales 5 corresponden a residuos localizados en la cavidad central y que poseen un rol directo en la unión de sustrato (resaltados en rojo, **Figura 34**), cuatro de ellos mediante interacciones de coordinación de grupos hidroxilo. Otros 6, no polares, se ubican inemediatamente contigüos a residuos con algún impacto o importancia funcional en la actividad de transporte según lo descrito en bibliografía (resaltados en púrpura, **Figura 34**)**.** Destaca que los residuos identificados que no se ubican en la zona media del transportador, lo hacen hacia el intracelular, dos de ellos, F206 y R331, en la zona terminal de las TMs 6 y 9, respectivamente, y otros tres en hélices que conforman parte del dominio helicoidal intracelular. Estos residuos, distanciados físicamente entre sí, quedan conectados físicamente al considerar los grupos de residuos que presentan interdependencia al analizar la matriz de acoplamiento estadístico (grupos 3,6,8 y 7). Al identificar estos grupos una gran mayoría de los residuos con un rol funcional conocido en la actividad de transporte (ver **Tabla 19**), se centra el análisis en ellos. La proyección de estos grupos y la de los grupos restantes en la estructura de hGLUT1 se presenta en la **Figura 35(A)** y **(B),** respectivamente.

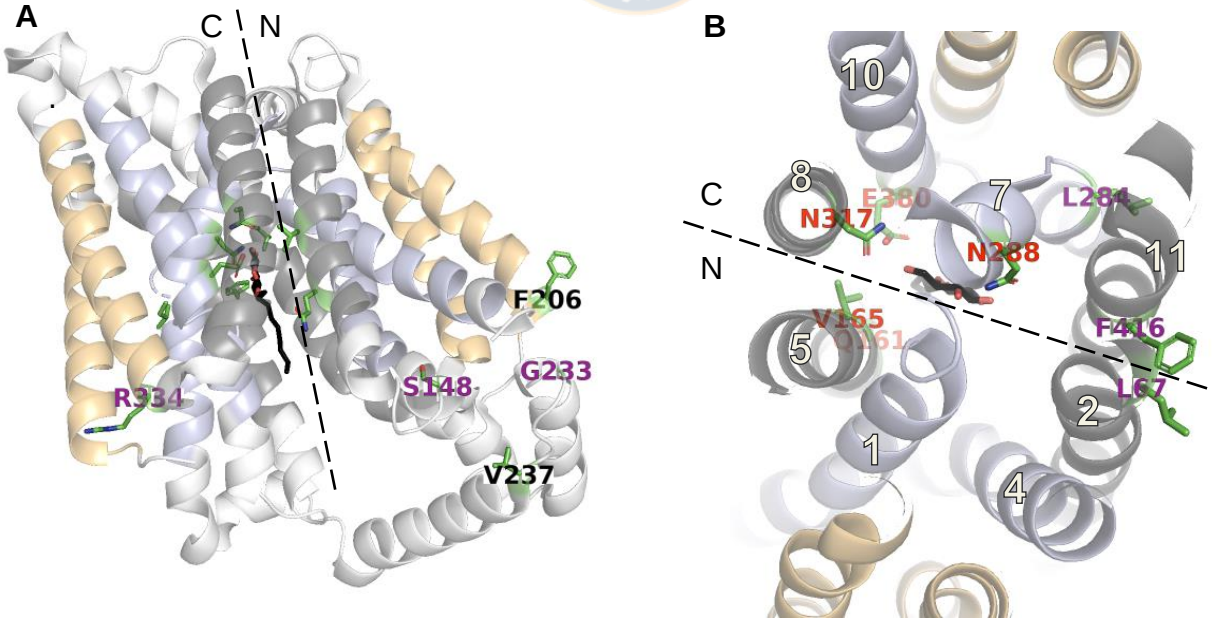


Figura 34. Disposición de grupo n°1 de residuos acoplados en estructura de hGLUT1. (A) Visión lateral. (B) Visión exofacial.

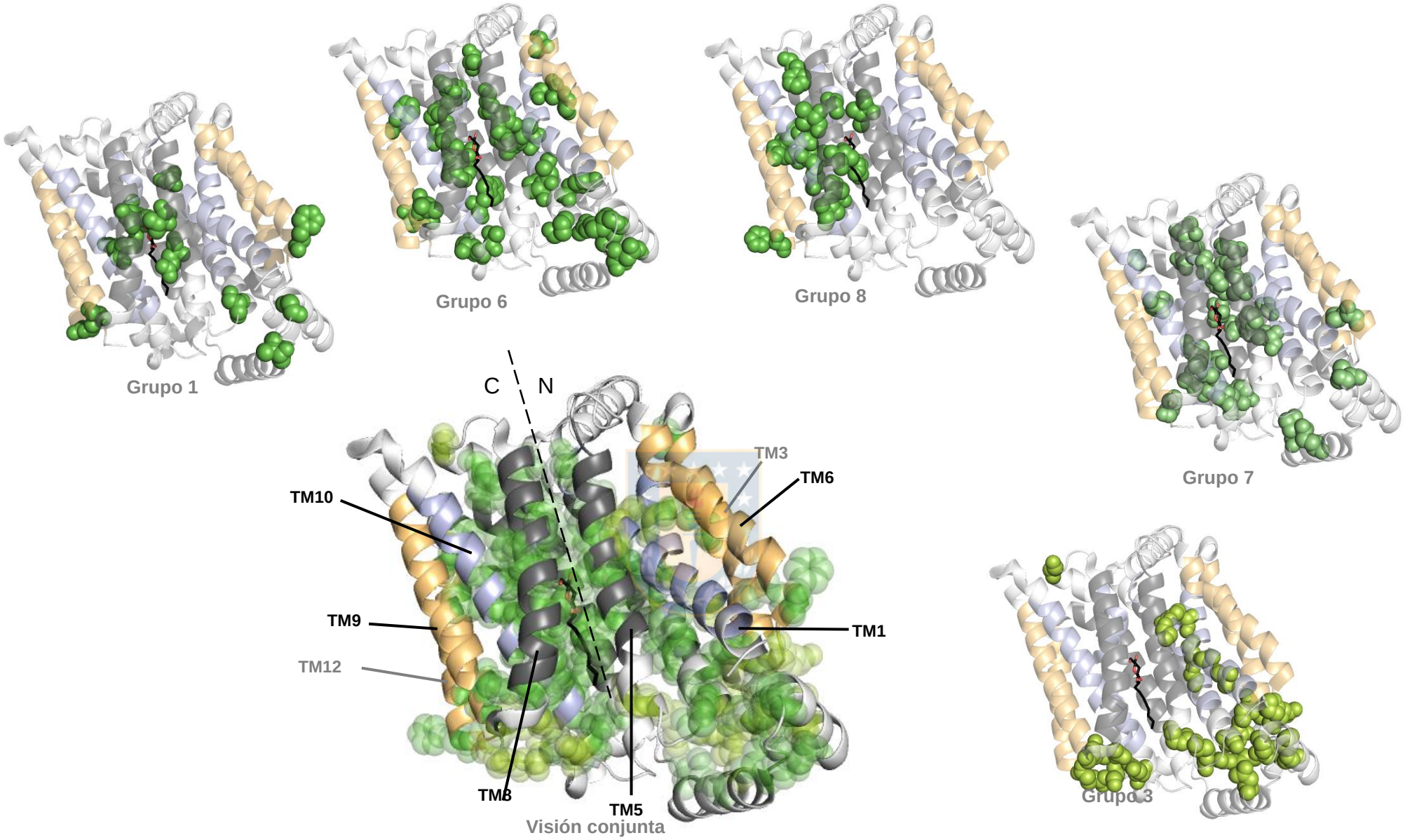
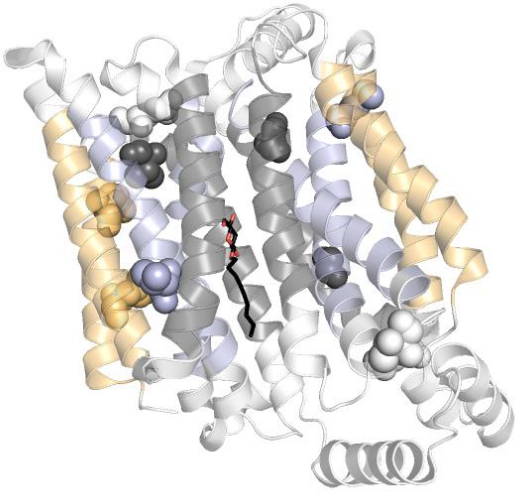
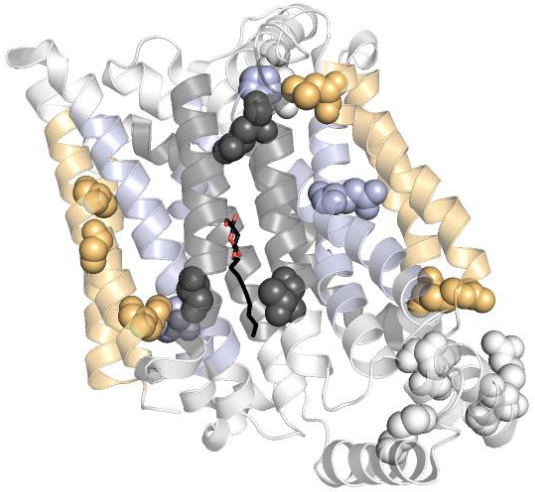


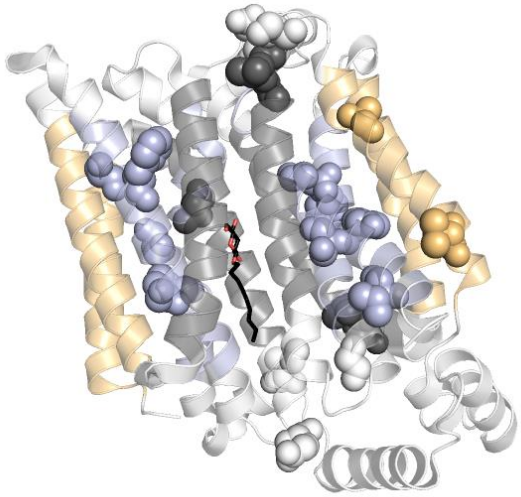
Figura 35. Proyección de grupos de residuos acoplados en estructura de hGLUT1. Los grupos de residuos acoplados n°1,3,6,7 y 8 forman un sector estructural en plegamiento MFS. Los grupos de residuos se proyectan en la estructura de hGLUT1, 4PYP, co-cristalizada con b-NG.



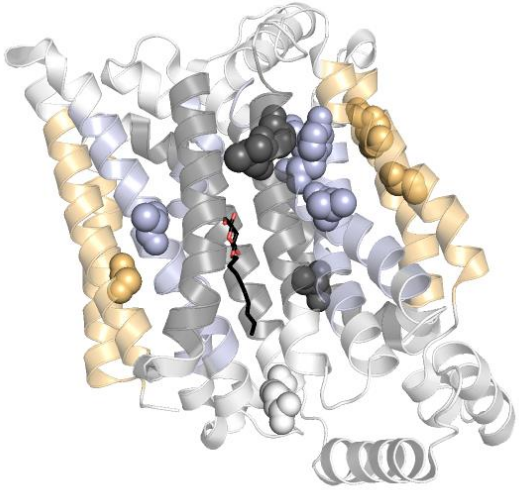
Grupo 2



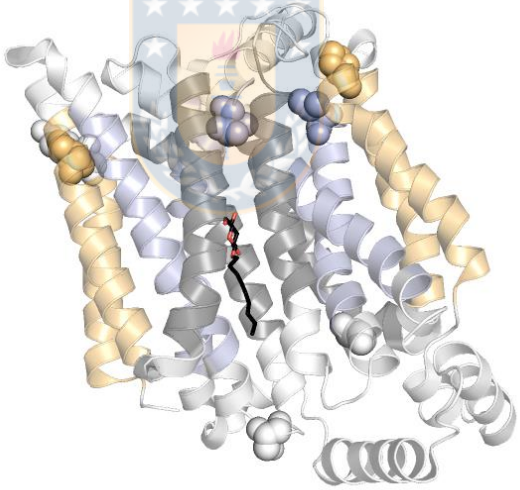
Grupo 4



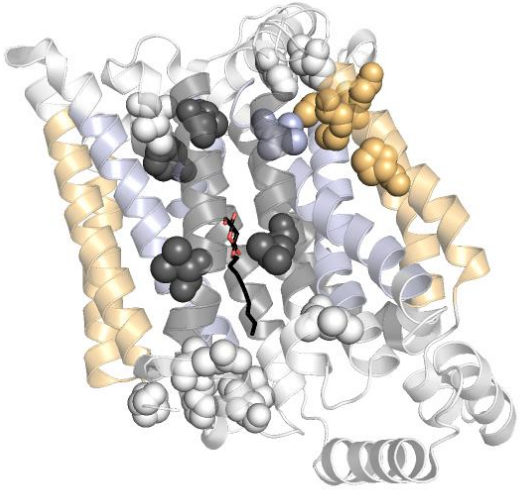
Grupo 5



Grupo 9



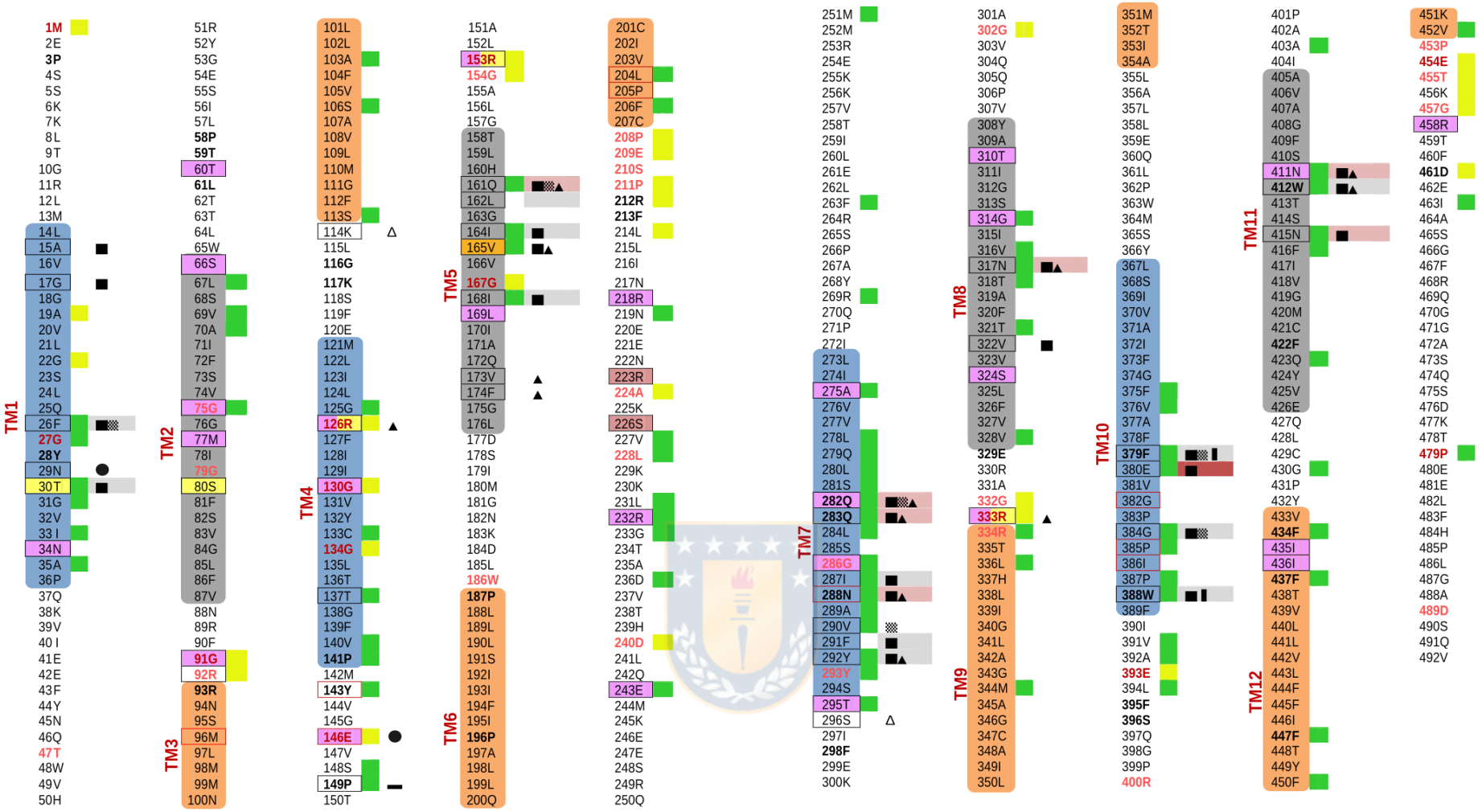
Grupo 10



Grupo 11

Figura 35 (Continuación).

Se observa que el grupo 3 comprende a una gran mayoría de residuos con un elevado grado de conservación en la familia y que se disponen principalmente en las hélices que conforman el dominio IC. La red de residuos acoplados abarca los conocidos motivos conservados en la familia R(3X)G(3X)G y E(6X)RG presentes a lo largo del TM4 y lazo *4*, los motivos GRR ubicados al inicio de las hélices 3 y 9, y los motivos PESPR y PETKG ubicados final de las hélices 6 y 12, respectivamente (Quistgaard et al., 2013). Los grupos 6 y 7 abarcan residuos principalmente ubicados en hélicas A de la cavidad central y residuos dispuestos hacia el intracelular que conectan físicamente con el grupo 3. El grupo 8, a diferencia de los anteriores, involucra sólo residuos que se encuentran en el dominio C-terminal, a excepción de L168, en TM5. Estos grupos de residuos acoplados estadísticamente forman un sector conectado físicamente a lo largo de la estructura, cuya conexión no es evidente a nivel de estructura secundaria, como se muestra en la **Figura 36,** la cual integra las anotaciones funcionales obtenidas mediante revisión bibliográfica y de conservación. Se indican en amarillo los residuos pertenecientes al grupo 3 y en verde los pertenecientes al resto de grupos, denominados subsector conservado y subsector variable de aquí en adelante y respectivamente, para mayor claridad. Destaca que la mayoría de los residuos identificados exterminetalmente como determinantes en la actividad de transporte y en la unión de sustrato se encuentran estadísticamente acoplados. Además, estos residuos se encuentran estadísticamente acoplados a residuos cuyas variantes se han asociado a la manifestación de distintas enfermedades, o bien a residuos que se encuentran inmediatamente contigüos a éstos. Esta red de interacciones permite epxlicar el efecto de estas variantes, a pesar de no encontrarse directamente involucrados en la unión de sustrato.



Anotaciones funcionales:

: Residuos cuya mutagénesis de cisteína en GLUT1 reduce el transporte a menos del 10%.

: Residuos involucrados en fosforilación por PKC(GLUT1).

: Residuos con variantes contrasentido identificadas en GLUT1DS1 y2 y en EIG12 (GLUT1).

: Residuos con variantes identificadas en RHUC2 (GLUT9).

: Residuo con variante identificada en NIDDM (GLUT2).

Residuos con impacto en transporte de:

Glucosa (■), fructosa (▒), galactosa (▌), xilosa (▲), ribosa(Δ), inositol (●), DHA (▬).

Residuos involucrados en la unión de sustrato:

No polares (■), polares (■), cargados (■)

Conservación:

SE ≤ 0.5

0.5 <SE ≤ 1.0

RE ≥ 1.5

Figura 36. Proyección de anotaciones funcionales, análisis de conservación y correlación de residuos en topología de hGLUT1. Hélices A, B y C resaltadas en azul, gris y naranja, respectivamente. Sus posiciones de inicio y fin se presentan según lo establecido en base de datos PDBTM. Residuos conservados destacados en rojo, rosado o negro según criterios de conservación descritos en imagen. Residuos con relevancia funcional conocida enmarcados en rectángulos, códigos de colores y símbolos explicados en imagen. Residuos asociados a los subsectores conservados y variables en amarillo y verde, respectivamente.

5.7.4. Análisis de comportamiento mutacional

Los primeros 30 residuos identificados con mayor coeficiente de correlación por rangos de Spearman (rk), con respecto a la matriz de similaridad del alineamiento SP, calculado mediante el programa Xdet tras aplicar filtro de redundancia del 95% se listan en la **Tabla 20.** Para evaluar la significancia estadística del valor de correlación asociado a cada posición, se calcula su *z-score* y *p-value* con respecto a una distribución de valores de fondo para cada posición, generadas a partir de 1000 aleatorizaciones del alineamiento de entrada. Los residuos con algún rol funcional descrito en literatura (**Tabla 15**) se resaltan en negro. Se destacan también residuos conservados, determinados mediante análisis de entropía de Shannon (H) y entropía relativa (RE). Los primeros 30 residuos identificados muestran valores de rk superiores a 0.45 y valores de z-score elevados, que confirman sus significancia estadística. De éstos, el 50% corresponden a residuos funcionales con directa participación en el reconocimiento de sustrato a nivel de la cavidad central descritos en literatura, y un porcentaje similar posee un significativo nivel de conservación en la familia. Destaca que la gram mayoría del resto de residuos igualmente se ubican en las hélices A y B que participan en la constitución del canal central, adyacentes a los residuos con roles funcionales descritos, como se muestra en la **Figura 37.**

Tabla 20. Residuos con mayor coeficiente de correlación en análisis de comportamiento mutacional.

Residuo	rk	z-score	p-value	Residuo	rk	z-score	p-value
G286	**0.704**	**17.60**	**0.00E+00**	**F379**	**0.521**	**12.73**	**0.00E+00**
Q283	**0.663**	**16.78**	**0.00E+00**	F375	0.517	14.81	0.00E+00
V418	0.644	18.81	0.00E+00	T321	0.512	14.84	0.00E+00
Q282	**0.620**	**14.59**	**0.00E+00**	**Q161**	**0.505**	**13.30**	**0.00E+00**
W388	**0.580**	**13.71**	**0.00E+00**	**G314**	**0.501**	**15.05**	**0.00E+00**
F26	**0.576**	**15.96**	**0.00E+00**	**G384**	**0.497**	**12.78**	**0.00E+00**
S23	0.573	22.63	0.00E+00	V376	0.492	14.03	0.00E+00
P211	0.566	13.09	0.00E+00	G31	0.478	12.78	0.00E+00
P141	0.564	14.11	0.00E+00	**W412**	**0.477**	**12.05**	**0.00E+00**
Y293	0.560	15.32	0.00E+00	F447	0.469	10.96	0.00E+00
Y143	**0.552**	**15.13**	**0.00E+00**	G27	0.466	10.42	0.00E+00
N288	**0.546**	**14.49**	**0.00E+00**	**N411**	**0.465**	**12.84**	**0.00E+00**
Y292	**0.542**	**13.94**	**0.00E+00**	**T137**	**0.464**	**12.44**	**0.00E+00**
G154	0.538	11.52	0.00E+00	L278	0.461	12.34	0.00E+00
A35	0.526	17.65	0.00E+00	G125	0.458	13.20	0.00E+00

: H<0.5. : H<1.0. RE>1.5

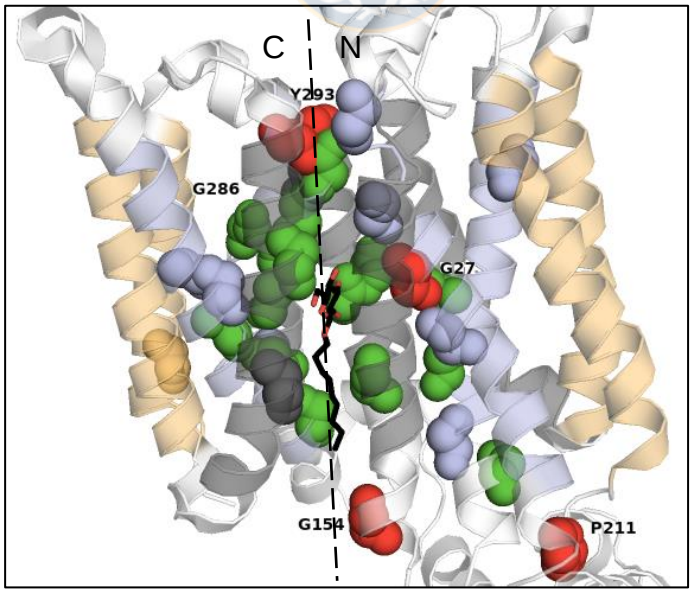


Figura 37. Residuos con mayor coeficiente de correlación en análisis de comportamiento mutacional proyectados en estructura de hGLUT1. Código PDB, 4pyp. Primeros 30 residuos con mayor coeficiente de correlación visualizados como esferas. Residuos involucrados en reconocimiento de sustrato y residuos con entropía de Shannon menor a 1 en verde y rojo, respectivamente.

5.7.5. Caracterización de agrupamientos

En la **Figura 38** se muestra, para los distintos grupos identificados mediante análisis cluster de la región de reconocimiento de sustrato presentados en la **Figura 25 (B),** los logos de secuencia asociados, indicándosse los residuos responsables de la segregación de los grupos identificados por el programa S3det. También se integran anotaciones de los residuos con mayor comportamiento mutacional identificados mediante el programa Xdet y los residuos con un elevado grado de conservación. Son pocos los residuos identificados como responsables de la segregación de los grupos por el programa S3Det y no se corresponden con los identificados mediante comportamiento mutacional. Tampoco se observan grupos de residuos que presenten patrones de conservación diferenciales que se correspondan con los grupos identificados de manera específica, sin embargo, se observan algunas tendencias, además de residuos conservados en la mayoría de los grupos que presentan mayor variabilidad en algunos grupos. Por ejemplo, la posición equivalente a N288 en hGLUT1 se encuentra muy conservada en la mayoría de los grupos, exceptuando el grupo correspondiente al transporte de maltosa y los grupos que integran a los transportadores GLUT10,12, PLTs, TMTs y ERLD6-like de *arabidopsis thaliana.* Para poder indagar de manera más específica en posiciones potencialmente determinantes de las diferencias de especificidad por los sustratos fructosa, inositol y DHA, se seleccionaron y agruparon secuencias con especificidades conocidas para estos tres sustratos, los cuales se presentan en la **Tabla 21**, y se analizó su armonía de secuencia (SH) para los distintos aminoácidos de reconocimiento de sustrato. Los resultados se presentan en la **Tabla 22,** destacándose las posiciones con armonías de secuencia menores a 0.5. Se observan posiciones en donde algunos grupos aceptan una diversidad de aminoácidos, mientras otros una elevada conservación. También se observan algunas posiciones con patrones de conservación que a diferencian un grupo de diferenciales entre los tres grupos, como la posición 168, lo que sugiere su relevancia a la hora de discirminar entre la especificidad por estos sustratos.

Tabla 21. Grupos de transportadores con especificidad por inositol, fructosa y DHA

N° filograma*	UNIPROT ID	Organismo	Proteína	Grupo
171	Q8NL90	*C.glutamicum*	iolT2	Inositol
172	Q8NTX0	*C.glutamicum*	iolT1	Inositol
174	O34718	*B.subtilis*	iolT	Inositol
177	A0A0H3NKX8	*S.typhimurium*	iolT2	Inositol
178	A0A0H3NW06	*S.typhimurium*	iolT1	Inositol
214	P87110	*S.pombe*	ITR2	Inositol
215	P30605	*S.cerevisiae*	ITR1	Inositol
216	P30606	*S.cerevisiae*	ITR2	Inositol
217	Q01440	*L.donovani*	GTR1	Inositol
218	Q8VZR6	*A.thaliana*	INT1	Inositol
219	Q9C757	*A.thaliana*	INT2	Inositol
220	O23492	*A.thaliana*	INT4	Inositol
222	Q96QE2	*H.sapiens*	GLUT13	Inositol
224	Q921A2	*R.norvegicus*	GLUT13	Inositol
49	Q8NJ22	*K.lactis*	frt1	Fructosa
50	Q9HFF8	*S.pastorianus*	FSY1	Fructosa
292	Q9BYW1	*H.sapiens*	GLUT11	Fructosa
295	Q9NRM0	*H.sapiens*	GLUT9	Fructosa
298	Q6PXP3	*H.sapiens*	GLUT7	Fructosa
303	P22732	*H.sapiens*	GLUT5	Fructosa
305	P43427	*R.norvegicus*	GLUT5	Fructosa
306	Q9WV38	*M.musculus*	GLUT5	Fructosa
255	Q9JIF3	*M.musculus*	GLUT8	Glucosa,DHA
256	Q9JJZ1	*R.norvegicus*	GLUT8	Glucosa,DHA
310	P14672	*H.sapiens*	GLUT4	Glucosa,DHA
317	P11168	*H.sapiens*	GLUT2	Glucosa,DHA
318	P14246	*M.musculus*	GLUT2	Glucosa,DHA
324	P11166	*H.sapiens*	GLUT1	Glucosa,DHA
325	P17809	*M.musculus*	GLUT1	Glucosa,DHA
326	P11167	*R.norvegicus*	GLUT1	Glucosa,DHA
332	P32037	*M.musculus*	GLUT3	Glucosa,DHA
333	Q07647	*R.norvegicus*	GLUT3	Glucosa,DHA
335	Q8TDB8	*H.sapiens*	GLUT14	Glucosa,DHA
337	P11169	*H.sapiens*	GLUT3	Glucosa,DHA

*Se indica posción en el filograma de la familia, Figura 25(E).

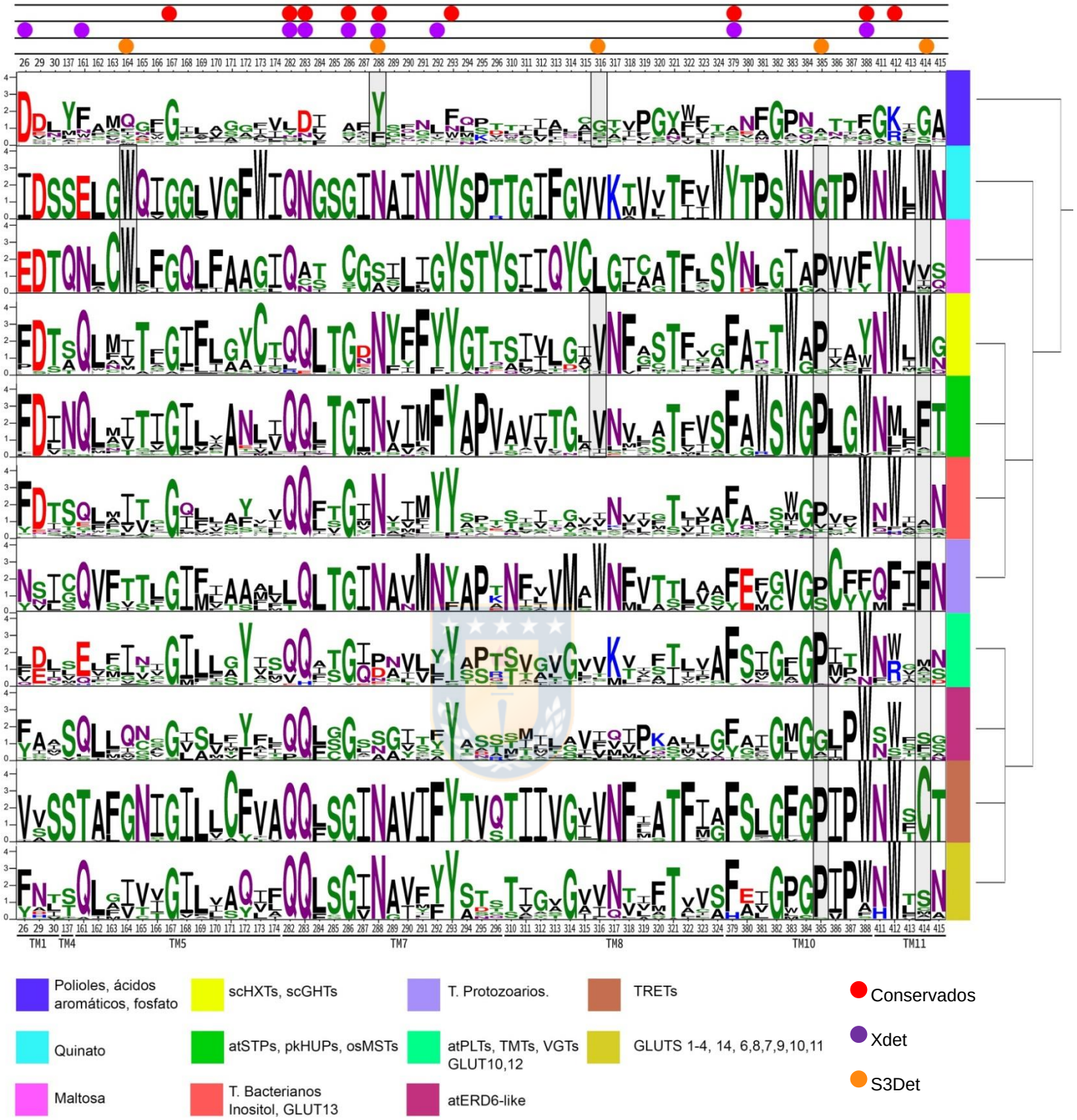


Figura 38. Caracterización de residuos de reconocimiento de sustrato para los distintos grupos identificados. Se muestran los logos de secuencia de los residuos de reconocimiento de sustrato para los grupos definidos mediante análisis cluster. Se enumeran según secuencia de referencia, hGLUT1 y se emplea mismo código de colores que en Figura 25(B). En rojo, residuos conservados identificados por entrpía de Shannon y RE. En morado, residuos con elevado comportamiento mutacional. En naranja, residuos responsables de segregación de grupos. Se resalta en rectángulo los residuos identificados como discriminantes para uno o varios grupos específicos.

Tabla 22. Armonía de secuencia para residuos de reconocimiento de sustrato

Ubicación	Residuo	SH	z-score[a]	Inositol[b]	Fructosa[b]	Glucosa, DHA[b]
TM1	26	0.56	-9.31	F	Yfs	F
TM1	29	0.63	-7.57	Ds	Nd	Nad
TM1	**30**	**0.42**	**-8.75**	**T**	**QLV**	**Tli**
TM4	137	0.56	-5.16	Sag	GSlmy	ST
TM5	**161**	**0.49**	**-5.66**	**GNSTVEq**	**Qea**	**Q**
TM5	162	0.82	-1.94	Livf	FILv	L
TM5	**163**	**0.34**	**-5.77**	**ALWMFs**	**Fn**	**Gma**
TM5	164	0.91	0.14	Ivl	Itv	Ilv
TM5	**165**	**0.44**	**-11.75**	**VT**	**ATci**	**V**
TM5	**166**	**0.41**	**-6**	**Gftvs**	**LVi**	**TVi**
TM5	167	1	nan	G	G	G
TM5	**168**	**0.29**	**-14.22**	**Q**	**EVl**	**I**
TM5	169	0.82	-1.95	FLiv	Lvf	L
TM5	170	0.69	-1.78	FILav	Vflmt	Vll
TM5	171	0.73	-3.81	Aivs	AG	As
TM5	172	0.6	-5.46	FYas	Qy	Qy
TM5	173	0.73	-1.84	GLVisy	VIL	IV
TM5	174	0.72	-1.58	IVlcf	VFil	Fila
TM7	282	0.8	-4.26	Q	GQmy	Q
TM7	283	0.96	-1.68	Q	Qe	Q
TM7	284	0.77	-4.49	LFiv	L	Lf
TM7	285	0.68	-5.55	STv	CTS	S
TM7	286	1	nan	G	G	G
TM7	287	0.8	-0.91	Itwv	IVln	Iv
TM7	288	0.96	-1.68	N	Nd	N
TM7	289	0.59	-4.71	STav	As	Ag
TM7	290	0.79	-2.17	Ilmv	Iv	IV
TM7	**291**	**0.37**	**-9.26**	**Mnq**	**MYnw**	**Fm**
TM7	292	0.92	-0.52	Y	Yaf	Yf
TM7	293	0.92	-1.58	Yf	Y	Y
TM7	294	0.55	-6.62	Sag	Atm	Sa
TM7	**295**	**0.24**	**-8.56**	**APqsgt**	**SDn**	**Tn**
TM7	**296**	**0.48**	**-4.7**	**Telrsiv**	**SQT**	**GSt**
TM8	310	0.72	-3.11	Satn	Tis	Ts
TM8	311	0.6	-3.77	Ivsl	MAilv	Iv
TM8	**312**	**0.43**	**-6.21**	**AVIfpw**	**Gsv**	**Gst**
TM8	**313**	**0.37**	**-5.11**	**ITVflnp**	**Tsg**	**SVA**
TM8	314	0.72	-4.76	SGA	G	G
TM8	315	0.61	-3.33	AFGV	Agsv	Iva

Tabla 21 (Continuación).

Ubicación	Residuo	SH	z-score[a]	Inositol[b]	Fructosa[b]	Glucosa, DHA[b]
TM8	316	0.57	-3.58	ITcvlm	AVci	Vi
TM8	317	0.71	-2.96	Nqsa	ELN	Nq
TM8	**318**	**0.43**	**-5.47**	**AVFl**	**LVit**	**Tmv**
TM8	319	0.87	0.28	LVIaf	FILV	Allv
TM8	**320**	**0.45**	**-8.98**	**FGa**	**GMat**	**F**
TM8	321	0.81	-2.59	STag	Ta	T
TM8	322	0.62	-3.23	FILav	IVfm	AVl
TM8	323	0.56	-6.06	IV	CFLTPV	Vl
TM8	324	0.67	-4.95	AGs	TAS	Sa
TM10	379	0.51	-6.22	FYqm	FHa	F
TM10	**380**	**0.4**	**-8.16**	**Asq**	**FAcgs**	**Ea**
TM10	**381**	**0.5**	**-3.6**	**LPgsvt**	**SVGIL**	**Vl**
TM10	382	0.84	-0.69	Gafs	Gst	G
TM10	**383**	**0.4**	**-6.34**	**IMlw**	**Py**	**Pw**
TM10	384	0.6	-6.22	Gns	ASg	G
TM10	385	0.6	-4.61	PTVcn	GPat	P
TM10	386	0.57	-5.52	Vagim	LVI	I
TM10	387	0.83	-1.67	Pvat	TP	P
TM10	388	0.66	-7.14	W	WAfgs	W
TM11	411	0.6	-5.23	Nglm	LHmn	N
TM11	412	0.91	-2.48	W	Wy	W
TM11	**413**	**0.13**	**-9.81**	**Iagmv**	**Li**	**Tvf**
TM11	414	0.54	-2.33	FGSAiv	Sfmtw	Scm
TM11	415	0.85	-1.33	Ns	Nla	Na

a. El Z-score de SH se calcula permutando las etiquetas de grupo y ejecutando el análisis para 100 aleatorizaciones. Se calculan la desviación media y estándar de estos conjuntos permutados. Una puntuación Z de -1 significa que la puntuación SH está 1 desviación estándar por debajo de la "media aleatoria". A puntajes más negativos mayor significancia estadística.

b. Presenta tipos de residuos presentes en cada uno de los grupos. Los tipos de residuos se clasifican en orden decreciente de frecuencia. Los tipos de residuos que ocurren con una frecuencia menor a la mitad de la más alta se escriben en minúsculas.

6 DISCUSIÓN

Comprender las relaciones entre la secuencia, estructura y función de las proteínas es el objetivo que persigue la biología estructural desde sus inicios. A pesar de los grandes avances en la materia tras el explosivo aumento en la cantidad de secuencias y estructuras proteicas disponibles para su estudio, conocer qué información de las secuencias proteicas es necesaria y suficiente para determinar sus distintas propiedades funcionales sigue siendo un desafío actual para la biología estructural y computacional; desafío que ha ido abordándose de manera particular para las distintas familias de proteínas conocidas. Para el caso de la familia de transportadores de azúcares, a pesar de que ha sido ampliamente estudiada, la caracterización y comprensión de las bases moleculares de la especificidad por sustrato en la literatura revisada se limita a análisis individuales de sus miembros o bien análisis comparativos centrados en algunos residuos funcionales que, en definitiva, abarcan pocas secuencias, generalmente las más representativas de la familia, sin considerar de manera integrativa la gran diversidad de homólogos presentes en los distintos reinos de la vida. Nuestro trabajo se orientó a la identificación y caracterización de grupos de secuencias con funcionalidades comunes, y se basó en la necesidad de avanzar en el desarrollo de herramientas bioinformáticas predictivas y explicativas de la especificidad de transporte, entendiendo y en concordancia con lo ya expuesto, que las actuales son limitadas. Las interrogantes asociadas a la etapa de identificación a las que se buscó dar respuesta en este trabajo son las relativas a la existencia de correlatividad entre las relaciones evolutivas de las secuencias y su especificidad por sustrato, la existencia de relaciones divergentes respecto de las primeras para los residuos involucrados en el reconocimiento de sustrato y si éstas poseen una mejor correlatividad con la función molecular. A continuación se discuten los principales hallazgos, limitaciones y proyecciones de este trabajo en torno a estas preguntas.

6.1 Sobre las relaciones filogenéticas de la familia

El primer paso para estudiar las relaciones evolutivas inherentes a una familia de proteínas es determinar las secuencias que serán consideradas como miembros de ésta, lo cual no es una tarea trivial. Si bien existen trabajos que estudian las relaciones evolutivas de transportadores relacionados por similaridad y funcionalidad con los miembros más representativos y conocidos de la familia SP, suelen ser análisis filogenéticos de homólogos presentes en una misma especie y que integran, en algunos casos, especies cercanas; como los análisis filogenéticos de GLUTs en humanos (Augustin, 2010), de la familia de transportadores de hexosas en *Saccharomyces cerevisiae* (Boles & Hollenberg, 1997; Wieczorke et al., 1999), de polioles en *Arabidopsis thaliana* (Klepek et al., 2005) y de trehalosa en *Drosoóphila melanogaster* (Kanamori et al., 2010), por dar algunos ejemplos. Existen pocos estudios que analicen relaciones evolutivas más divergentes dentro de la familia. El principal que responde a este desafío es el pionero análisis realizado por Pao et al (1998) en el contexto de los trabajos desarrollados en el laboratiorio de Saier y colaboradores para levantar el sistema de clasificación de transportadores, estudio que permitió clasificar las distintas secuencias presentes en diversos filos y organismos identificadas en ese momento como miembros de la MFS en 17 familias nombradas según propiedades funcionales comunes de sus miembros mejor caracterizados, empleando como criterio clasificador la similaridad de secuencias y análisis filogenéticos basado en distancias y siendo la primera familia bautizada como '*Sugar Porter Family'.* Si bien el estudio de Pao *et al* (1998) consideró secuencias presentes en los distintos reinos de la vida, el objetivo principal era analizar las relaciones interfamiliares y el análisis filogenético de la familia SP en citado trabajo se realizó sólo con 20 secuencias representativas, por lo que las conclusiones relativas a la divergencia evolutiva y funcional derivables de éste fueron limitadas. Tras este trabajo, destaca el reciente realzado por Jia *et al* (2019), en el cual aprovechando el incremento en el número de secuencias disponibles en bases de datos, realizó un análisis filogenético con 65 secuencias, integrando las 14 isoformas de GLUTs humanos y las isoformas homólogas presentes en especies

modelos del reino metazoa: ratón, pez cebra, mosca de la fruta y el nemátodo *C.elegans.* Este estudio, si bien avanza en el esclarecimiento de las relaciones evolutivas de la familia, se limita al reino metazoa y, al igual que el estudio de Pao et al (1998), no integra mayores anotaciones ni análisis relativos a la divergencia en la función molecular de las secuencias estudiadas.

En este estudio se avanzó en el esclarecimiento de las relaciones evolutivas de esta familia de multigenes, considerando una mayor cantidad de secuencias y especies. Se usaron como base las 139 secuencias asignadas a la familia en la base de datos TCDB para la posterior búsqueda de homólogos mediante BLASTp contra la base de datos de secuencias curadas SwissProt, debido a que la primera incluye secuencias presentes desde bacterias hasta animales. Para la búsqueda de homólogos se empleó como valor de corte un valor de expectación de $1x10^{-30}$, considerado altamente restrictivo en comparación al empleado por TCDB para establecer homología entre miembros de una misma familia, equivalente a un valor de expectación de $1x10^{-19}$ (Chang et al., 2004). Empleando este criterio se integraron otras 168 secuencias al análisis. Se integraron además secuencias anotadas bajo el código 2.A.1.1 Swissprot y secuencias Swissprot alojadas en InterPro bajo el código IPR003663, que agrupa a transportadores de azúcares e inositoles (ver **Tabla 4**).

Para el análisis filogenético se requiere un MSA de entrada de 'buena calidad', esto es, capaz de alinear residuos análogos entre las distintas secuencias, representando así correctamente su divergencia. Anterior a la construcción del MSA se filtraron las secuencias según criterios topológicos y de redundancia. Los aspectos técnicos relativos a estas etapas fueron previamente discutidos en las secciones 5.2 y 5.3, por lo que sólo se remarcarán y comentarán algunas ideas relevantes. Relativo al primer criterio de filtrado, se mantuvieron sólo las secuencias que poseen 12 TMs conocidas o predichas por el mejor método de predicción, que en este caso resultó ser MEMSAT-SVM, hallazgo en concordancia con lo reportado por Salas-Burgos (2011). Esto permitió eliminar fragmentos y así

poder evaluar la calidad del alineamiento considerando todo el plegamiento MFS. Con respecto al segundo criterio, no se usaron filtros de redundancia restrictivos, eliminando sólo secuencias idénticas, esto para evitar pérdida de información funcional de secuencias de interés, bajo la premisa de que eventualmente se podrían encontrar secuencias con porcentajes elevados de identidad, vale decir superiores al 95% y con propiedades de especificidad distintivas, aspecto que se aborda en el siguiente apartado. Relativo al MSA obtenido, se constata que los alineamientos construidos a partir de las matrices de sustitución PHAT y SLIM (Müller et al., 2001; Ng et al., 2000) presentaron mejor calidad que los construdios mediante la matriz BLOSSUM62 e independientemente del algoritmo empleado, considerando el porcentaje de aperturas en TMs como el criterio de mayor importancia para su evaluación. Esto porque la topología de proteínas de membrana posee mayores restricciones estructurales y funcionales a nivel de las TM, y se espera que se encuentren mayormente conservadas. Esto se encuentra en concordancia con el análisis topológico de las secuencias, donde si bien el rango de longitud observado va de 400 a 1440 residuos aminoacídicos, las mayores variaciones en longitud se encuentran a nivel de los segmentos N y C terminales y del segmento que conecta los dominos N y C **(Figuras 19 y 20).** Este resultado, si bien esperable al considerar que estas matrices fueron construidas a partir de proteínas transmembrana, por lo que se consideran adecuadas para el problema de alineamiento en cuestión, entrega mayores precedentes para sugerir su uso a la hora de realizar esta tarea con proteínas transportadoras. Además, el resto de indicadores evaluados presentan estadísticas muy similares con BLOSSUM62 (**Tabla 14**), por lo que la preferencia de su uso puede no resultar evidente si no se consideran aspectos topológicos. El uso del programa MergeAlign (Collingridge & Kelly, 2012) a partir de los alineamientos mejor evaluados mostroó ser útil para obtener un alineamiento con mejores estadísticos, aumentando el número de posiciones con 100% de ocupancia, el puntaje TCS y disminuyendo el porcentaje de aperturas por TMs en las secuencias, como se observa en la **Tabla 14**. La calidad del alineamiento resultante se logra constatar también con la elevada ocupancia de los TM en el perfil de ocupancia (**Figura 21**),

el reconocimiento de los motivos conservados propios de la familia (**Figura 27(B)**) y la correspondencia entre sitios funcionales (**Tabla 15**), para los cuales se evaluó consistencia entre el alineamiento de a pares obtenido mediante Fasta36 y el observado en el alineamiento SP, obtenido a partir de la selección de los sitios que mostraran cobertura con la secuencia de hGLUT1 (**Tabla 8**) .

A partir del análisis de la topología del filograma sin enraizar obtenido desde el alineamiento SP destaca, como se describe en la sección 5.6, la presencia de una rama larga que agrupa tanto secuencias bacterianas reonocidas como miembros de la familia SP en TCDB (secuencias n° 5,6,7,8,9,17,18,19 y 24 en **Figura 25**) como a secuencias que se encuentran clasificadas bajo los códigos TC.2.A.1.9, 2.A.1.15, y 2.A.1.18 de TCDB, corresondientes a las familias de cotransportadores de fosfato:H+ (PHS), cotransportadores de ácidos aromáticos:H+ (AAHS) y polioles (PP), rama utilizada para enraizar el árbol por razones ya comentadas. Aparecen además dos secuencias, numeradas en el filograma como 14 y 15, asociadas a los códigos TC 2.A.1.6 y 2.A.1.12, correspondientes a las familias de cotransportadores de metabolitos:H+ (MHS) y de sialato:H+ (SHS). Interesantemente en el estudio de Pao et al (1998) tanto la familia PHS, SHS y MHS nacen de la misma rama que la familia SP, lo que sustenta el uso de esta rama como grupo externo. La familia AAHS y PP no aparece en integrada en tal análisis. Estos resultados suguieren que los transportadores asignados a la familia SP presentes en esta rama se encuentran más asociados a estas familias, por lo que su clasificación debe ser revisada. Es interesante también destacar que la gran mayoría de las secuencias que conforman esta rama se integraron al análisis desde la búsqueda de secuencias anotadas bajo el código ‘2.A.1.1’ en la sección ‘Familia y dominios’ de la base de datos SwissProt o desde las secuencias pertenecientes a la entrada de Interpro analizada. Esto muestra una clara discrepancia en la clasificación asignada para estas secuencias entre estas bases de datos y TCDB, lo que invita a investigar más profundamente para esclarecer las relaciones de divergencia entre estas familias.

El enraizamiento del árbol muestra la bifurcación de la familia en dos ramas principales, la primera (secuencias 48 a 161, **Figura 25**) agrupando sólo a secuencias presentes en hongos y plantas. La segunda (secuencias 162 a 341, **Figura 25**), más diversa, agrupa secuencias de bacterias, hongos, plantas, protozoos y animales. La ramificación de esta segunda rama principal en 4 sub-ramas con secuencias de plantas, hongos y animales sugiere que eventos de duplicación tempranos previos a la divergencia de estos grupos filogenéticos dieron origen a la diversidad funcional de la familia, con algunos clados con especificades mas diversas y amplias, y otras más restringidas**.** Las relaciones entre la filogenia y la función se discuten en el siguiente apartado.

El hallazgo más importante asociado a este análisis es el relativo a las relaciones entre los genes que codifican para los GLUTs en las distintas especies analizadas. Destaca que los transportadores GLUTs 1-4,14,5,7,9 y 11, GLUTs 6,8, GLUTs 10 y 12 y GLUT13 no se encuentran ubicados en una única de estas 4 sub-ramas, sino que en sub-ramas separadas, lo que sugiere que su divergencia se originó por eventos de duplicación previos a la divergencia de estos grupos taxonómicos y a la específica del reino metazoa. Así, este estudio apoya parcialmente lo planteado por Jia *et al* (2019), en la que se postula una nueva clasificación para los GLUTs: una nueva clase I, que abacra a los GLUTs 1-4,14,5,7,9 y 11, una clase II que abarca a los GLUTs 10, 12 y 13 y una clase III que abarca los GLUTs 6,8, junto a los transportadores de trehalosa. En este estudio existe concordancia con las clases I y III, sin embargo la topología del árbol sugiere una mayor divergencia entre los GLUTs 10, 12 con respecto a GLUT13, que se posiciona en un clado particular que integra transportadores de inositol de protozos, hongos y plantas, todos ellos cotransportadores inositiol:H+ (clado 16, secuencias 212 a 224, **Figura 25**).

6.2 Sobre la diversidad funcional de la familia y las relaciones estructura-función asociadas

La revisión de los sustratos anotados con evidencia experimental para las secuencias en estudio refleja su diversidad en todas las categorías taxonómicas representadas **(Tabla 16),** incluyéndose hexosas, pentosas, disacáridos, polioles lineales y ácidos orgánicos. Con respecto a la existencia de correlatividad entre las relaciones evolutivas de las secuencias y su especificidad por sustrato, en el filograma global de la familia se observan ramas que agrupan secuencias asociadas al transporte de un tipo de sustrato claramente definido, como los clados asociados al transporte de quinato, maltosa, glicerol e inositol (clados 5,6,7 y 16, ver **Figura 25**). Estos grupos de secuencias presentan porcentajes de identidad y similaridad mínimos de 61,7 y 84,4; 43,5 y 74,7; 35,3 y 66,6 y 26,2 y 60,2; respectivamente (ver **Anexo 4**), cumpliendo en general con la premisa de que homólogos con identidades superiores al 40% poseen una elevada probababilidad de desempeñar funciones comunes (Pearson, 2013). Sin embargo, también se observan otras ramas que agrupan secuencias con selectividades más amplias en donde éstas, si bien se asocian al transporte de una gama de sustratos estructuralmente relacionados, se observan algunas con patrones similares de especificidad (al menos en términos de afinidad, sugerido por los valores de km), mientras que otras presentan cambios en las preferencias de sustratos o bien pérdidas de capacidad de transporte de un sustrato determinado en comparación al resto de secuencias con las que se agrupa. Esto, a pesar de los elevados valores de identidad y similaridad compartidos. Casos ejemplo de lo anterior son el transportador GLUT10 de *H.sapiens* y el de *M.músculus (*secuencias 191 y 192 en filograma*),* que a pesar de que comparten una identidad del 77,4% y una similaridad de 88,9%, el primero ha mostrado no tener capacidad de transporte de DHA (Rumsey et al., 1997), a diferencia del segundo (Lee et al., 2010). Un caso similar se observa en el clado 9, que agrupa a la familia de transportadores GHT de *S.pombe.* Estos transportadores poseen un porcentaje de identidad y similaridad mínimo de 54,6 y 80,6 respectivamente, sin embargo GHT6 muestra

una ligera mayor afinidad por fructosa que glucosa, a diferencia de los GHT2, 1 y 5. Por lo demás, GHT3 evidenció ser un transportador de gluconato (Heiland et al., 2000) . Destaca también el clado 10 del filograma, que agrupa a la familia de transportadores de hexosas de *S. cerevisiae* y otros transportadores relacionados, con una selectividad amplia, permitiendo el transporte de diversos azúcares. A pesar que el porcentaje de identidad y similaridad mínimo encontrado entre las secuencias que conforman el clado es de 51,5 y 78,5 respectivamente, los distintos transportadores muestran variaciones en sus especificidades relativas al transporte de glucosa, manosa, fructosa, galactosa, xilosa y polioles lineales. Una situación similar se observa en el clado 11 que agrupa a los transportadores STP de *A.thaliana* y otros relacionados, con porcentajes de identidad y similaridad mínimos de 40.1 y 73.7 entre las secuencias, respectivamente.

Con respecto a los agrupamientos obtenidos a partir de residuos funcionales, al comparar los filogramas sin enraizar, se observa que el filograma específico de residuos de reconocimiento de sustrato reproduce, en general, los mismos grupos de secuencias y sus relaciones, variando las distancias relativas entre grupos. Esto en concordancia con los resultados encontrados para la mayoría de las familias analizadas por Landgraf *et al* (2001). Destaca que la rama que agrupa a los transportadores de polioles, fosfatos y ácidos aromáticos se observa mucho más distante del resto de transportadores, lo cual sugiere una mayor divergencia respecto del resto de la familia en el sitio de unión.

Los grupos identificados mediante *clustering* de residuos de reconocimiento de sustrato empleando el programa S3Det también muestran una elevada correspondencia con el filograma global. En definitiva, no se encontró que estos residuos mostraran relaciones divergentes respecto del filograma global, lo que sugiere que los residuos del poro de transporte guían en gran medida las relaciones filogenéticas observadas. Esto, probablemente debido a que estos residuos concentran y explican la mayor proporción de la varianza de las secuencias, conclusión soportada por los resultados obtenidos mediante S3det, el

análisis de acoplamiento estadístico (SCA) y el análisis de comportamiento mutacional. La matriz SCA de los grupos de residuos acoplados identificados y su disposición en la estructura de hGLUT1 (**Figuras 33 y 35**, respectivamente) muestran que la arquitectura de la familia se compondría por la descomposición jerárquica de un único sector principal, donde el grupo de residuos asociado al primer autovector y que explica el mayor porcentaje de varianza de los datos integra a residuos con un rol directo en la unión de sustrato o residuos inemediatamente contigüos a éstos y con algún impacto o importancia funcional en la actividad de transporte, según lo descrito en bibliografía (**Figura 34**). Los residuos con mayor comportamiento mutacional también se posicionan hacia el poro de transporte, estando varios de ellos directamente involucrados en el reconocimiento de sustrato (**Figura 37**). Al ser estos residuos los que poseen mayores coeficientes de correlación con cambios drásticos en las relaciones de similaridad derivadas del alineamiento, permiten inferir la misma conclusión que la desprendida del análisis SCA. Por úlimo, el análisis ejecutado mediante S3Det aplica el algoritmo de *clustering k-means* sobre el espacio vectorial reducido de secuencias obtenido tras la previa aplicación de un MCA, análogo al análisis de componentes principales (PCA). Como el algoritmo *k-means* corresponde a las soluciones discretas del agrupamiento obtenido mediante PCA para un conjunto de datos (Ding & He, 2004), y las soluciones obtenidas mediante este análisis suelen mostrar una elevada correspondencia con la historia evolutiva de las familias de proteínas (De Juan, 2013), la elevada correspondencia entre los grupos observados mediante filogenia y el *clústering* de residuos funcionales también soporta esta idea.

Con relación a la identificación de posiciones que permitan discriminar entre los distintos grupos, destaca que existen ramas del árbol filogenético que poseen variaciones distintivas en los motivos conservados de la familia, ver **Figura 27(B).** La implicancia biológica de estas variaciones en la definición de propiedades funcionales específicas asociadas a estas secuencias queda planteada como motivo de estudio posterior. Por otro lado, el algoritmo de identificación de residuos responsables de la segregación de grupos implementado en el programa

S3Det no resultó fructífero. En contraste, el análisis de armonía de secuencia permitió identificar de manera exitosa residuos que pueden explicar la diferencia de especificidad entre fructosa, glucosa e inositol, las cuales son de las más estudiadas en literatura. Destaca la posición equivalente a I168 en hGLUT1, que presenta patrones de conservación diferenciales entre los gruos asociados al transporte de estos sustratos (**Tabla 21),** estando la glutamina totalmente conservada en los transportadoress de inositol, y totalmente ausente en los otros transportadores. También la posición equivalente a Q161 en hGLUT1 junto a otros varios residuos contigüos presentes en la TM5 presentan patrones de conservación diferenciales. Ambos residuos fueron postulados por Ferreira et al., (2019) como determinantes para el transporte de inositol mediante análisis basados en modelamiento y docking molecular. Su identificación mediante análisis de armonía de secuencia demuestra la utilidad de su uso para identificar potenciales posiciones determinantes de especificidad. Los patrones diferenciales de conservación identificados proporcionan puntos de partida para estudiar las relaciones entre estos sitios y elaborar hipótesis explicativas de la selectividad que pueden ser futuramente evaluadas y confirmadas por mutagénesis y otros estudios funcionales.

6.3 Limitaciones y proyecciones del trabajo

A pesar de los grupos identificados por *clustering* de residuos de reconocimiento de sustrato tienden a agrupar las secuencias por funcionalidades comunes, lo hacen de manera equivalente a la filogenia. Este último método sigue siendo el más empleado por los algoritmos de transferencia de anotaciones automáticas junto con los basados en similaridad de secuencia, como revela la **Figura 23**. Sin embargo, sólo permiten transferir anotaciones funcionales más generales, como se muestra en la **Tabla 17,** presentando limitaciones a la hora de transferir anotaciones más detalladas.

La presencia de especificidades mayores por un mismo tipo de sustrato en ramas divergentes del árbol, ya sea en términos de afinidad y/o de capacidad, sugieren interacciones comunes debidas probablemente a fenómenos de convergencia evolutiva. Este fenómeno, ejemplificado por las secuencias FRT1 y FSY1 de *K.lactis* y *S.pastorianus*, GHT6 de *S. Pombe* y los GLUTs 7,9 y 11 de *H.sapiens* ubicadas en los clados 4,9 y 11, respectivamente (**Figura 25**), y que presentan una mayor afinidad por fructosa que por glucosa; por los transportadores de inositol que no se agrupan con el resto ubicado en el clado 16, así como por el transporte de DHA presente en GLUT8, en los GLUTs1-4 y ausente en los GLUTs 5,7,911 más estrechamente relacionados con estos últimos; mantiente abierto el desafío: ¿es posible identificar patrones comunes que epxliquen y permitan predecir tales interacciones, abriendo la posibilidad de desarrollar herramientas bioinformáticas que no dependan de las relaciones filogenéticas y que permitan tranferir anotaciones funcionales más detalladas?.

Los esfuerzos actuales se centran en entrenar algoritmos de aprendizaje supervisado empleando información codificada en la secuencia. Éstos lamentablemente se encuentran muy limitados, debido a que su aplicación depende de la disponibilidad de secuencias agrupadas bajo una misma etiqueta funcional que además posea soporte experimental. Los escasos estudios que buscan desarrollar algoritmos de predicción de la selectividad en transportadores,

ejemplificados por los trabajos de Hu et al., (2015); Mishra et al., (2014) y Schaadt et al., (2010), emplean conjuntos de entrenamiento con clases funcionales más generales representadas por pocas secuencias (entre 10 a 40 secuencias por clase) que, como se anticipó en la sección 1.4.1, no involucran niveles detallados de especificidad. Esta asignación de clase de sustrato se realiza normalmente a través de un proceso de selección manual que suele no explicarse en los estudios, lo que plantea el desafío de levantar criterios unificados y reproducibles (Alballa & Butler, 2019). Lamentablemente, y como se constató en este estudio, actualmente las bases de datos como SwissProt no facilitan esta tarea, ya que no tienen actualizada toda la información relativa a la especificidad de sustrato disponible en literatura, como los parámetros cinéticos. Además, cuando se reportan valores como parámetros de km, no siempre se repotan las capacidades de transporte asociadas, ni se especifica en qué condiciones se determinó tal parámetro. Esto último es relevante sobre todo en el caso de proteínas transportadoras, donde pueden reportarse más de un tipo de km para un mismo sistema, ver **Tabla 2**. Tampoco existe una nomenclatura unificada que permita definir jerarquías de especificidad basadas afinidad, capacidades de transporte o eficiencias.

A pesar de lo anterior, se logró levantar información curada relativa a la especificidad de transporte del 55% de las secuencias analizadas, lo que abre la posibilidad de levantar estudios de aprendizaje supervisado que integren más información que el análisis de armonía de secuencia realizado. La principal limitante de algoritmos supervisados que se basan sólo en secuencias, es que cada aminoácido se considera como una categoría, sin considerar propiedades físicas o químicas de manera explícita y que, en definitiva, le otorgan funcionalidad. Así, existe un abanico de características que pueden extraerse a partir de la información de secuencia. Entre ellas, encontramos atributos basados en composición aminoacídica y atributos basados en propiedades fisicoquímicas específicas de aminoácidos, tales como la hidrofobicidad, la polaridad y el volumen Van der Waals (véase Chen et al., 2019), a partir de las cuales se pueden diseñar nuevos atributos contextualizados al problema de estudio y entrenar algoritmos de aprendizaje automático.

El plegamiento altamente conservado propio de la MFS también abre la posibilidad de extraer y utilizar información estructural. El reciente estudio de Narunsky et al., (2020) analizó los patrones de interacción proteína-adenina a partir de estructuras tridimensionales co-cristalizadas con sustratos que presentan el esqueleto de la base nitrogenada. La superimposición estructural de las proteínas con base en el fragmento de adenina reveló distintos patrones de interacción asociados a segmentos cortos de aminoácidos denominados 'temas' en el estudio, y cuya detección en proteínas nóveles podría revelar su función. El esclarecimiento de patrones estructurales de interacción entre azúcares y transportadores pertenecientes a la familia SP, así como del resto de familias de la MFS involucradas en el transporte de azúcares, alojadas bajo los códigos TCDB 2.A.1.5,7,12,18,20 y 2.A.2 (Saier Jr, 2000); podría permitir el desarrollo de herramientas bioinformáticas de predicción más robustas. También, podría sentar las bases para el diseño artifical de transportadores con selectividades específicas, desafío que ya posee precedentes en los recientes trabajos orientados al diseño artificial de acuaporinas (Chowdhury et al., 2018; Song & Kumar, 2019) .

7 CONCLUSIÓN

En este trabajo se pudo avanzar en el esclarecimiento de las relaciones evolutivas inherentes a las secuencias que conforman la familia de transportadores de azúcares e indagar más profundamente en las relaciones estructura-función asociadas. A partir del estudio de las relaciones exhibidas por los homólogos de GLUTs presentes en las especes estudiadas y en relación al resto de transportadores de la familia, se sugiere una nueva clasificación para estas secuencias, soportada también por estudios recientes y conformada por los siguientes 4 grupos: GLUTs1-4,5,7,9 y 11, GLUTs 6 y 8, GLUTs 10,12 y GLUT13. También se avanzó en la caracterización de la familia en términos de coevolución y patrones de conservación, identificándose residuos que pueden ser claves en la discriminación de los sustratos glucosa, fructosa, inositol y DHA.

8 BIBLIOGRAFÍA

Abramson, J., Iwata, S., & Kaback, H. R. (2004). Lactose permease as a paradigm for membrane transport proteins. Molecular Membrane Biology, 21(4), 227–236.

Ackerman, S. H., Tillier, E. R., & Gatti, D. L. (2012). Accurate simulation and detection of coevolution signals in multiple sequence alignments. PloS One, 7(10).

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Aperturaped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 25(17), 3389–3402.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science, 181(4096), 223–230.

Apel, A. R., Ouellet, M., Szmidt-Middleton, H., Keasling, J. D., & Mukhopadhyay, A. (2016). Evolved hexose transporter enhances xylose uptake and glucose/xylose co-utilization in Saccharomyces cerevisiae. Scientific Reports, 6(1), 1–10.

Arsov, T., Mullen, S. A., Rogers, S., Phillips, A. M., Lawrence, K. M., Damiano, J. A., Goldberg-Stern, H., Afawi, Z., Kivity, S., Trager, C., Petrou, S., Berkovic, S. F., & Scheffer, I. E. (2012). Glucose transporter 1 deficiency in the idiopathic generalized epilepsies. Annals of Neurology, 72(5), 807–815. https://doi.org/10.1002/ana.23702

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., & Eppig, J. T. (2000). Gene ontology: Tool for the unification of biology. Nature Genetics, 25(1), 25–29.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., & Dress, A. W. (2000). Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. Molecular Biology and Evolution, 17(1), 164–178.

Attwood, T. K. (2002). The PRINTS database: A resource for identification of protein families. Briefings in Bioinformatics, 3(3), 252–263.

Augustin, R. (2010). The protein family of glucose transport facilitators: It's not only about glucose after all. IUBMB Life, 62(5), 315–333. https://doi.org/10.1002/iub.315

Bairoch, A., & Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Research, 25(1), 31–36.

Bakan, A., Dutta, A., Mao, W., Liu, Y., Chennubhotla, C., Lezon, T. R., & Bahar, I. (2014). Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics, 30(18), 2681–2683.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., & Sonnhammer, E. L. (2000). The Pfam protein families database. Nucleic Acids Research, 28(1), 263–266.

Bhadola, P., & Deo, N. (2016). Targeting functional motifs of a protein family. Physical Review E, 94(4), 042409.

Boari de Lima, E., Meira, W., & Melo-Minardi, R. C. de. (2016). Isofunctional Protein Subfamily Detection Using Data Integration and Spectral Clustering. PLOS Computational Biology, 12(6), e1005001. https://doi.org/10.1371/journal.pcbi.1005001

Bogan, A. A., & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. Journal of Molecular Biology, 280(1), 1–9.

Boles, E., & Hollenberg, C. P. (1997). The molecular genetics of hexose transport in yeasts. FEMS Microbiology Reviews, 21(1), 85-111. https://doi.org/10.1111/j.1574-6976.1997.tb00346.x

Bonting, S. L., & De Pont, J. (1981). Membrane transport. Elsevier.

Brandt, B. W., Feenstra, K. A., & Heringa, J. (2010). Multi-Harmony: Detecting functional specificity from sequence alignment. Nucleic acids research, 38(suppl_2), W35-W40.

Brockmann, K., Wang, D., Korenke, C. G., von Moers, A., Ho, Y. Y., Pascual, J. M., Kuang, K., Yang, H., Ma, L., Kranz-Eble, P., Fischbarg, J., Hanefeld, F., & De Vivo, D. C. (2001). Autosomal dominant glut-1 deficiency syndrome and familial epilepsy. Annals of Neurology, 50(4), 476–485. https://doi.org/10.1002/ana.1222

Brown, D., & Sjölander, K. (2006). Functional classification using phylogenomic inference. PLoS Computational Biology, 2(6), e77. https://doi.org/10.1371/journal.pcbi.0020077

Brusic, V., & Zeleznikow, J. (1999). Knowledge discovery and data mining in biological databases. The Knowledge Engineering Review, 14(3), 257–277.

Büchel, D. E., Gronenborn, B., & Müller-Hill, B. (1980). Sequence of the lactose permease gene. Nature, 283(5747), 541–545.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 160–172.

Capra, J. A., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics, 23(15), 1875–1882.

Cesareni, G., Panni, S., Nardelli, G., & Castagnoli, L. (2002). Can we infer peptide recognition specificity mediated by SH3 domains? FEBS Letters, 513(1), 38–44.

Chakerian, J., & Holmes, S. (2012). Computational tools for evaluating phylogenetic and hierarchical clustering trees. Journal of Computational and Graphical Statistics, 21(3), 581–599.

Chang, A. B., Lin, R., Keith Studley, W., Tran, C. V., & Saier, M. H. (2004). Phylogeny as a guide to structure and function of membrane transport proteins. Molecular Membrane Biology, 21(3), 171-181. https://doi.org/10.1080/09687680410001720830

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., Zhu, Y., Powell, D. R., Akutsu, T., Webb, G. I., Chou, K.-C., Smith, A. I., Daly, R. J., Li, J., & Song, J. (2019). iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbz041

Chowdhury, R., Ren, T., Shankla, M., Decker, K., Grisewood, M., Prabhakar, J., Baker, C., Golbeck, J. H., Aksimentiev, A., Kumar, M., & Maranas, C. D. (2018). PoreDesigner for tuning solute selectivity in a robust and highly permeable outer membrane pore. Nature Communications, 9(1), 3661. https://doi.org/10.1038/s41467-018-06097-1

Colas, C., Ung, P. M.-U., & Schlessinger, A. (2016). SLC transporters: Structure, function, and drug discovery. Medchemcomm, 7(6), 1069–1081.

Collingridge, P. W., & Kelly, S. (2012). MergeAlign: Improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. BMC Bioinformatics, 13(1), 117.

Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. Elements of Information Theory, 2, 1–55.

Cura, A. J., & Carruthers, A. (2012). Role of monosaccharide transport proteins in carbohydrate assimilation, distribution, metabolism, and homeostasis. Comprehensive Physiology, 2(2), 863–914. https://doi.org/10.1002/cphy.c110024

Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). 22 a model of evolutionary change in proteins. In Atlas of protein sequence and structure (Vol. 5, pp. 345–352). National Biomedical Research Foundation Silver Spring MD.

De Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. Nature Reviews. Genetics, 14(4), 249–261. https://doi.org/10.1038/nrg3414

de Melo-Minardi, R. C., Bastard, K., & Artiguenave, F. (2010). Identification of subfamily-specific sites based on active sites modeling and clustering. Bioinformatics, 26(24), 3075–3082.

del Sol Mesa, A., Pazos, F., & Valencia, A. (2003). Automatic methods for predicting functionally important residues. Journal of Molecular Biology, 326(4), 1289–1302.

Deng, D., Sun, P., Yan, C., Ke, M., Jiang, X., Xiong, L., Ren, W., Hirata, K., Yamamoto, M., Fan, S., & Yan, N. (2015). Molecular basis of ligand recognition and transport by glucose transporters. Nature, 526(7573), 391–396. https://doi.org/10.1038/nature14655

Deng, D., Xu, C., Sun, P., Wu, J., Yan, C., Hu, M., & Yan, N. (2014). Crystal structure of the human glucose transporter GLUT1. Nature, 510(7503), 121–125. https://doi.org/10.1038/nature13306

Diallinas, G. (2014). Understanding transporter specificity and the discrete appearance of channel-like gating domains in transporters. Frontiers in Pharmacology, 5, 207.

Dietvorst, J., Karhumaa, K., Kielland-Brandt, M. C., & Brandt, A. (2010). Amino acid residues involved in ligand preference of the Snf3 transporter-like sensor in Saccharomyces cerevisiae. Yeast, 27(3), 131–138.

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. Proceedings of the twenty-first international conference on Machine learning, 29.

Dinour, D., Gray, N. K., Campbell, S., Shu, X., Sawyer, L., Richardson, W., Rechavi, G., Amariglio, N., Ganon, L., Sela, B.-A., Bahat, H., Goldman, M., Weissgarten, J., Millar, M. R., Wright, A. F., & Holtzman, E. J. (2010). Homozygous SLC2A9 mutations cause severe renal hypouricemia. Journal of the American Society of Nephrology: JASN, 21(1), 64–72. https://doi.org/10.1681/ASN.2009040406

Dinour, D., Gray, N. K., Ganon, L., Knox, A. J. S., Shalev, H., Sela, B.-A., Campbell, S., Sawyer, L., Shu, X., Valsamidou, E., Landau, D., Wright, A. F., & Holtzman, E. J. (2012). Two novel homozygous SLC2A9 mutations cause renal hypouricemia type 2. Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association, 27(3), 1035–1041. https://doi.org/10.1093/ndt/gfr419

Dunn, S. D., Wahl, L. M., & Gloor, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics, 24(3), 333–340.

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics (Oxford, England), 14(9), 755–763.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., & Bairoch, A. (2002). The PROSITE database, its status in 2002. Nucleic Acids Research, 30(1), 235–238.

Farwick, A., Bruder, S., Schadeweg, V., Oreb, M., & Boles, E. (2014). Engineering of yeast hexose transporters to transport D-xylose without inhibition by D-glucose. Proceedings of the National Academy of Sciences, 111(14), 5159–5164.

Ferreira, R. S., Pons, J.-L., & Labesse, G. (2019). Insights into Substrate and Inhibitor Selectivity among Human GLUT Transporters through Comparative Modeling and Molecular Docking. ACS Omega, 4(3), 4748–4760.

Fetrow, J. S., & Babbitt, P. C. (2018). New computational approaches to understanding molecular protein function. PLoS Computational Biology, 14(4). https://doi.org/10.1371/journal.pcbi.1005756

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., … Mitchell, A. L. (2017). InterPro in 2017—Beyond protein family and domain annotations. Nucleic Acids Research, 45(D1), D190–D199. https://doi.org/10.1093/nar/gkw1107

Flatt, J. F., Guizouarn, H., Burton, N. M., Borgese, F., Tomlinson, R. J., Forsyth, R. J., Baldwin, S. A., Levinson, B. E., Quittet, P., Aguilar-Martinez, P., Delaunay, J., Stewart,

G. W., & Bruce, L. J. (2011). Stomatin-deficient cryohydrocytosis results from mutations in SLC2A1: A novel form of GLUT1 deficiency syndrome. Blood, 118(19), 5267–5277. https://doi.org/10.1182/blood-2010-12-326645

Frillingos, S., Sahin-tóth, M., Wu, J., & Kaback, H. R. (1998). Cys-scanning mutagenesis: A novel approach to structure–function relationships in polytopic membrane proteins. The FASEB Journal, 12(13), 1281–1299.

Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., Mitraka, E., Schriml, L. M., Gaudet, P., & Hobbs, E. T. (2019). ECO, the Evidence & Conclusion Ontology: Community standard for evidence information. Nucleic Acids Research, 47(D1), D1186–D1194.

Goswitz, V. C., & Brooker, R. J. (1995). Structural features of the uniporter/symporter/antiporter superfamily. Protein Science, 4(3), 534–537.

Halabi, N., Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Protein sectors: Evolutionary units of three-dimensional structure. Cell, 138(4), 774–786.

Heiland, S., Radovanovic, N., Höfer, M., Winderickx, J., & Lichtenberg, H. (2000). Multiple hexose transporters of Schizosaccharomyces pombe. Journal of Bacteriology, 182(8), 2153-2162.

Heinz, E. (1978). Mechanics and energetics of biological transport. Molecular Biology, Biochemistry, and Biophysics, 29, 1.

Henderson, P. J. F., & Maiden, M. C. J. (1990). Homologous sugar transport proteins in Escherichia coli and their relatives in both prokaryotes and eukaryotes. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 326(1236), 391–410.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences, 89(22), 10915–10919.

Hirai, T., Heymann, J. A. W., Maloney, P. C., & Subramaniam, S. (2003). Structural model for 12-helix transporters belonging to the major facilitator superfamily. Journal of Bacteriology, 185(5), 1712–1718. https://doi.org/10.1128/jb.185.5.1712-1718.2003

Hu, Y., Guo, Y., Shi, Y., Li, M., & Pu, X. (2015). A consensus subunit-specific model for annotation of substrate specificity for ABC transporters. RSC Advances, 5(52), 42009-42019.

Huang, Y., Lemieux, M. J., Song, J., Auer, M., & Wang, D.-N. (2003). Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli. Science, 301(5633), 616–620.

Jardetzky, O. (1966). Simple allosteric model for membrane pumps. Nature, 211(5052), 969–970. https://doi.org/10.1038/211969a0

Jones, D. T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics (Oxford, England), 23(5), 538-544. https://doi.org/10.1093/bioinformatics/btl677

Kalinina, O. V., Gelfand, M. S., & Russell, R. B. (2009). Combining specificity determining and conserved residues improves functional site prediction. BMC Bioinformatics, 10(1), 174.

Kanamori, Y., Saito, A., Hagiwara-Komoda, Y., Tanaka, D., Mitsumasu, K., Kikuta, S., Watanabe, M., Cornette, R., Kikawada, T., & Okuda, T. (2010). The trehalose transporter 1 gene sequence is conserved in insects and encodes proteins with different kinetic properties involved in trehalose import into peripheral tissues. Insect Biochemistry and Molecular Biology, 40(1), 30-37. https://doi.org/10.1016/j.ibmb.2009.12.006

Kapoor, K., Finer-Moore, J. S., Pedersen, B. P., Caboni, L., Waight, A., Hillig, R. C., Bringmann, P., Heisler, I., Müller, T., & Siebeneicher, H. (2016). Mechanism of inhibition of human glucose transporter GLUT1 is conserved between cytochalasin B and phenylalanine amides. Proceedings of the National Academy of Sciences, 113(17), 4711–4716.

Kasahara, M., Shimoda, E., & Maeda, M. (1997). Amino acid residues responsible for galactose recognition in yeast Gal2 transporter. Journal of Biological Chemistry, 272(27), 16721–16724.

Kasahara, T., & Kasahara, M. (2010). Identification of a key residue determining substrate affinity in the yeast glucose transporter Hxt7: A two-dimensional comprehensive study. The Journal of Biological Chemistry, 285(34), 26263–26268. https://doi.org/10.1074/jbc.M110.149716

Kasahara, T., Maeda, M., Ishiguro, M., & Kasahara, M. (2007). Identification by comprehensive chimeric analysis of a key residue responsible for high affinity glucose transport by yeast HXT2. The Journal of Biological Chemistry, 282(18), 13146–13150. https://doi.org/10.1074/jbc.C700041200

Kasahara, T., Shimogawara, K., & Kasahara, M. (2011). Crucial effects of amino acid side chain length in transmembrane segment 5 on substrate affinity in yeast glucose transporter Hxt7. Biochemistry, 50(40), 8674–8681. https://doi.org/10.1021/bi200958s

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment programa version 7: Improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780.

Klepek, Y.-S., Geiger, D., Stadler, R., Klebl, F., Landouar-Arsivaud, L., Lemoine, R., Hedrich, R., & Sauer, N. (2005). Arabidopsis POLYOL TRANSPORTER5, a new member of the monosaccharide transporter-like superfamily, mediates H+-Symport of numerous substrates, including myo-inositol, glycerol, and ribose. The Plant Cell, 17(1), 204-218. https://doi.org/10.1105/tpc.104.026641

Klepper, J., Willemsen, M., Verrips, A., Guertsen, E., Herrmann, R., Kutzick, C., Flörcken, A., & Voit, T. (2001). Autosomal dominant transmission of GLUT1 deficiency. Human Molecular Genetics, 10(1), 63–68. https://doi.org/10.1093/hmg/10.1.63

Konings, W. N., Kaback, H. R., & Lolkema, J. S. (1996). Transport processes in eukaryotic and prokaryotic organisms (Vol. 2). Elsevier.

Kozma, D., Simon, I., & Tusnady, G. E. (2012). PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic acids research, 41(D1), D524-D529.

Landgraf, R., Xenarios, I., & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. Journal of Molecular Biology, 307(5), 1487–1502.

Leandro, M. J., Fonseca, C., & Gonçalves, P. (2009). Hexose and pentose transport in ascomycetous yeasts: An overview. FEMS Yeast Research, 9(4), 511–525.

Lee, D., Redfern, O., & Orengo, C. (2007). Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology, 8(12), 995–1005.

Lee, E. E., Ma, J., Sacharidou, A., Mi, W., Salato, V. K., Nguyen, N., Jiang, Y., Pascual, J. M., North, P. E., & Shaul, P. W. (2015). A protein kinase C phosphorylation motif in GLUT1 affects glucose transport and is mutated in GLUT1 deficiency syndrome. Molecular Cell, 58(5), 845–853.

Lee, T. J., Paulsen, I., & Karp, P. (2008). Annotation-based inference of transporter function. Bioinformatics (Oxford, England), 24(13), i259-267. https://doi.org/10.1093/bioinformatics/btn180

Lee, Y.-C., Huang, H.-Y., Chang, C.-J., Cheng, C.-H., & Chen, Y.-T. (2010). Mitochondrial GLUT10 facilitates dehydroascorbic acid import and protects cells against oxidative stress: Mechanistic insight into arterial tortuosity syndrome. Human Molecular Genetics, 19(19), 3721-3733. https://doi.org/10.1093/hmg/ddq286

Leen, W. G., Klepper, J., Verbeek, M. M., Leferink, M., Hofste, T., van Engelen, B. G., Wevers, R. A., Arthur, T., Bahi-Buisson, N., Ballhausen, D., Bekhof, J., van Bogaert, P., Carrilho, I., Chabrol, B., Champion, M. P., Coldwell, J., Clayton, P., Donner, E., Evangeliou, A., … Willemsen, M. A. (2010). Glucose transporter-1 deficiency syndrome: The expanding clinical and genetic spectrum of a treatable disorder. Brain: A Journal of Neurology, 133(Pt 3), 655–670. https://doi.org/10.1093/brain/awp336

Lieb, W. R., & Stein, W. D. (1974). Testing and characterizing the simple carrier. Biochimica et Biophysica Acta (BBA)-Biomembranes, 373(2), 178–196.

Litwin, S., & Jores, R. (1992). Shannon information as a measure of amino acid diversity. In Theoretical and experimental insights into immunology (pp. 279–287). Springer.

Lundstrom, K. H. (2006). Structural genomics on membrane proteins. crc Press.

Lunt, B., Szurmant, H., Procaccini, A., Hoch, J. A., Hwa, T., & Weigt, M. (2010). Inference of direct residue contacts in two-component signaling. In Methods in enzymology (Vol. 471, pp. 17–41). Elsevier.

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579–2605.

Madej, M. G., Sun, L., Yan, N., & Kaback, H. R. (2014). Functional architecture of MFS D-glucose transporters. Proceedings of the National Academy of Sciences, 111(7), E719–E727.

Maiden, M. C., Davis, E. O., Baldwin, S. A., Moore, D. C., & Henderson, P. J. (1987). Mammalian and bacterial sugar transport proteins are homologous. Nature, 325(6105), 641–643. https://doi.org/10.1038/325641a0

Majd, H., King, M. S., Palmer, S. M., Smith, A. C., Elbourne, L. D., Paulsen, I. T., Sharples, D., Henderson, P. J., & Kunji, E. R. (2018). Screening of candidate substrates and coupling ions of transporters by thermostability shift assays. ELife, 7, e38821.

Manolescu, A. R., Augustin, R., Moley, K., & Cheeseman, C. (2007). A highly conserved hydrophobic motif in the exofacial vestibule of fructose transporting SLC2A proteins acts as a critical determinant of their substrate selectivity. Molecular Membrane Biology, 24(5–6), 455–463. https://doi.org/10.1080/09687680701298143

Manolescu, A., Salas-Burgos, A. M., Fischbarg, J., & Cheeseman, C. I. (2005). Identification of a hydrophobic residue as a key determinant of fructose transport by the facilitative hexose transporter SLC2A7 (GLUT7). The Journal of Biological Chemistry, 280(52), 42978–42983. https://doi.org/10.1074/jbc.M508678200

Martin, L. C., Gloor, G. B., Dunn, S. D., & Wahl, L. M. (2005). Using information theory to search for co-evolving residues in proteins. Bioinformatics (Oxford, England), 21(22), 4116–4124. https://doi.org/10.1093/bioinformatics/bti671

McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., & Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature, 491(7422), 138–142.

Mishra, N. K., Chang, J., & Zhao, P. X. (2014). Prediction of membrane transport proteins and their substrate specificities using primary sequence information. PLoS One, 9(6), e100278.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences, 108(49), E1293-E1301.

Morgan, G. J. (1998). Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959-1965. Journal of the History of Biology, 155–178.

Mori, H., Hashiramoto, M., Clark, A. E., Yang, J., Muraoka, A., Tamori, Y., Kasuga, M., & Holman, G. D. (1994). Substitution of tyrosine 293 of GLUT1 locks the transporter into an outward facing conformation. Journal of Biological Chemistry, 269(15), 11578-11583.

Moysés, D. N., Reis, V. C. B., Almeida, J. R. M. de, Moraes, L. M. P. de, & Torres, F. A. G. (2016). Xylose fermentation by Saccharomyces cerevisiae: Challenges and prospects. International Journal of Molecular Sciences, 17(3), 207.

Mueckler, M., & Makepeace, C. (2009). Model of the exofacial substrate-binding site and helical folding of the human Glut1 glucose transporter based on scanning mutagenesis. Biochemistry, 48(25), 5934-5942.

Mueckler, M., Kruse, M., Strube, M., Riggs, A. C., Chiu, K. C., & Permutt, M. A. (1994). A mutation in the Glut2 glucose transporter gene of a diabetic patient abolishes transport activity. The Journal of Biological Chemistry, 269(27), 17765–17767.

Mueckler, M., Roach, W., & Makepeace, C. (2004). Transmembrane segment 3 of the Glut1 glucose transporter is an outer helix. Journal of Biological Chemistry, 279(45), 46876-46881.

Mueckler, Mike, & Thorens, B. (2013). The SLC2 (GLUT) family of membrane transporters. Molecular Aspects of Medicine, 34(2–3), 121–138. https://doi.org/10.1016/j.mam.2012.07.001

Müller, T., Rahmann, S., & Rehmsmeier, M. (2001). Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics, 17(suppl_1), S182–S189.

Naftalin, R. J. (2018). A critique of the alternating access transporter model of uniport glucose transport. Biophysics Reports, 4(6), 287–299.

Narunsky, A., Kessel, A., Solan, R., Alva, V., Kolodny, R., & Ben-Tal, N. (2020). On the evolution of protein–adenine binding. Proceedings of the National Academy of Sciences, 117(9), 4701-4709. https://doi.org/10.1073/pnas.1911349117

Naula, C. M., Logan, F. J., Logan, F. M., Wong, P. E., Barrett, M. P., & Burchmore, R. J. (2010). A glucose transporter can mediate ribose uptake: Definition of residues that confer substrate specificity in a sugar transporter. The Journal of Biological Chemistry, 285(39), 29721–29728. https://doi.org/10.1074/jbc.M110.106815

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443–453.

Nei, M., & Kumar, S. (2000). Molecular evolution and phylogenetics. Oxford university press.

Nemoto, W., Saito, A., & Oikawa, H. (2013). Recent advances in functional region prediction by using structural and evolutionary information–Remaining problems and future extensions. Computational and Structural Biotechnology Journal, 8(11), e201308007.

Ng, P. C., Henikoff, J. G., & Henikoff, S. (2000). PHAT: A transmembrane-specific substitution matrix. Bioinformatics, 16(9), 760–766.

Nomura, N., Verdon, G., Kang, H. J., Shimamura, T., Nomura, Y., Sonoda, Y., Hussien, S. A., Qureshi, A. A., Coincon, M., & Sato, Y. (2015). Structure and mechanism of the mammalian fructose transporter GLUT5. Nature, 526(7573), 397–401.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology, 302(1), 205-217.

Nugent, T., & Jones, D. T. (2009). Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics, 10, 159. https://doi.org/10.1186/1471-2105-10-159

Palma, M., Goffeau, A., Spencer-Martins, I., & Baret, P. V. (2007). A phylogenetic analysis of the sugar porters in hemiascomycetous yeasts. Journal of Molecular Microbiology and Biotechnology, 12(3–4), 241–248. https://doi.org/10.1159/000099645

Pao, S. S., Paulsen, I. T., & Saier, M. H. (1998). Major Facilitator Superfamily. Microbiology and Molecular Biology Reviews, 62(1), 1–34.

Pascual, J. M., Van Heertum, R. L., Wang, D., Engelstad, K., & De Vivo, D. C. (2002). Imaging the metabolic footprint of Glut1 deficiency on the brain. Annals of Neurology, 52(4), 458–464. https://doi.org/10.1002/ana.10311

Patching, S. G., Henderson, P. J., Herbert, R. B., & Middleton, D. A. (2008). Solid-state NMR spectroscopy detects interactions between tryptophan residues of the E. coli sugar transporter GalP and the α-anomer of the d-glucose substrate. Journal of the American Chemical Society, 130(4), 1236–1244.

Patlak, C. S. (1957). Contributions to the theory of active transport: II. The gate type non-carrier mechanism and generalizations concerning tracer flow, efficiency, and measurement of energy expenditure. The Bulletin of Mathematical Biophysics, 19(3), 209–235.

Paulsen, P. A., Custódio, T. F., & Pedersen, B. P. (2019). Crystal structure of the plant symporter STP10 illuminates sugar uptake mechanism in monosaccharide transporter superfamily. Nature Communications, 10(1), 1–8.

Pazos, F., Rausell, A., & Valencia, A. (2006). Phylogeny-independent detection of functional residues. Bioinformatics (Oxford, England), 22(12), 1440–1448. https://doi.org/10.1093/bioinformatics/btl104

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. Current Protocols in Bioinformatics, 42(1), 3–1.

Pérez López, C. (2004). Técnicas de análisis multivariante de datos. Aplicaciones Con SPSS, Madrid, Universidad Complutense de Madrid, 121–154.

Perland, E., & Fredriksson, R. (2017). Classification Systems of Secondary Active Transporters. Trends in Pharmacological Sciences, 38(3), 305–315. https://doi.org/10.1016/j.tips.2016.11.008

Pirovano, W., Feenstra, K. A., & Heringa, J. (2006). Sequence comparison by sequence harmony identifies subtype-specific functional sites. Nucleic Acids Research, 34(22), 6540–6548.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., & Clements, J. (2012). The Pfam protein families database. Nucleic acids research, 40(D1), D290-D301.

Quistgaard, E. M., Löw, C., Guettou, F., & Nordlund, P. (2016). Understanding transport by the major facilitator superfamily (MFS): Structures pave the way. Nature Reviews. Molecular Cell Biology, 17(2), 123–132. https://doi.org/10.1038/nrm.2015.25

Quistgaard, E. M., Löw, C., Moberg, P., Trésaugues, L., & Nordlund, P. (2013). Structural basis for substrate transport in the GLUT-homology family of monosaccharide transporters. Nature Structural & Molecular Biology, 20(6), 766-768. https://doi.org/10.1038/nsmb.2569

Rausell, A., Juan, D., Pazos, F., & Valencia, A. (2010). Protein interactions and ligand binding: From protein subfamilies to functional specificity. Proceedings of the National Academy of Sciences, 107(5), 1995-2000.

Raza, K. (2012). Application of data mining in bioinformatics. ArXiv Preprint ArXiv:1205.1125.

Reckzeh, E. S., & Waldmann, H. (2019). Development of Glucose Transporter (GLUT) Inhibitors. European Journal of Organic Chemistry.

Reynolds, K. A., McLaughlin, R. N., & Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. Cell, 147(7), 1564–1575.

Reynolds, K. A., Russ, W. P., Socolich, M., & Ranganathan, R. (2013). Evolution-based design of proteins. En Methods in enzymology (Vol. 523, pp. 213-235). Elsevier.

Rivoire, O., Reynolds, K. A., & Ranganathan, R. (2016). Evolution-Based Functional Decomposition of Proteins. PLOS Computational Biology, 12(6), e1004817. https://doi.org/10.1371/journal.pcbi.1004817

Rizzo, J., & Rouchka, E. C. (2007). Review of phylogenetic tree construction. University of Louisville Bioinformatics Laboratory Technical Report Series, 2–7.

Rottmann, T., Klebl, F., Schneider, S., Kischka, D., Rüscher, D., Sauer, N., & Stadler, R. (2018). Sugar Transporter STP7 Specificity for l-Arabinose and d-Xylose Contrasts with the Typical Hexose Transporters STP8 and STP12. Plant Physiology, 176(3), 2330–2350. https://doi.org/10.1104/pp.17.01493

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Rumsey, S. C., Kwon, O., Xu, G. W., Burant, C. F., Simpson, I., & Levine, M. (1997). Glucose transporter isoforms GLUT1 and GLUT3 transport dehydroascorbic acid. The Journal of Biological Chemistry, 272(30), 18982-18989.

Saier Jr, M. H. (2000). Families of transmembrane sugar transport proteins: MicroReview. Molecular microbiology, 35(4), 699-710.

Saier Jr, M. H., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C., & Moreno-Hagelsieb, G. (2015). The transporter classification database (TCDB): Recent advances. Nucleic Acids Research, 44(D1), D372–D379.

Salas-Burgos (2011). Desarrollo de una metodología robusta para la construcción de modelos 3D de proteinas de membrana y su aplicación en la asignación de un modelo de plegamiento espacial del transportador de ácido ascórbico SVCT2 [PhD Thesis]. Universidad de Concepción.

Saraçli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. Journal of Inequalities and Applications, 2013(1), 203.

Schaadt, N. S., Christoph, J., & Helms, V. (2010). Classifying substrate specificities of membrane transporters from Arabidopsis thaliana. Journal of chemical information and modeling, 50(10), 1899-1905.

Schneider, S. A., Paisan-Ruiz, C., Garcia-Gorostiaga, I., Quinn, N. P., Weber, Y. G., Lerche, H., Hardy, J., & Bhatia, K. P. (2009). GLUT1 gene mutations cause sporadic paroxysmal exercise-induced dyskinesias. Movement Disorders: Official Journal of the Movement Disorder Society, 24(11), 1684–1688. https://doi.org/10.1002/mds.22507

Schürmann, A., Doege, H., Ohnimus, H., Monser, V., Buchs, A., & Joost, H.-G. (1997). Role of conserved arginine and glutamate residues on the cytosolic surface of glucose transporters for transporter function. Biochemistry, 36(42), 12897-12902.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., & Kahn, D. (2002). ProDom: Automated clustering of homologous domains. Briefings in Bioinformatics, 3(3), 246–251.

Seyfang, A., Kavanaugh, M. P., & Landfear, S. M. (1997). Aspartate 19 and glutamate 121 are critical for transport function of the myo-inositol/H+ symporter from Leishmania donovani. The Journal of Biological Chemistry, 272(39), 24210–24215. https://doi.org/10.1074/jbc.272.39.24210

Smirnova, I., Kasho, V., & Kaback, H. R. (2011). Lactose permease and the alternating access mechanism. Biochemistry, 50(45), 9684–9693.

Smith, T. F., & Waterman, M. S. (1981). Comparison of biosequences. Advances in Applied Mathematics, 2(4), 482–489.

Song, W., & Kumar, M. (2019). Artificial water channels: Toward and beyond desalination. Current Opinion in Chemical Engineering, 25, 9-17.

Sperelakis, N. (2012). Cell physiology source book: Essentials of membrane biophysics. Elsevier.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Stein, W. D., & Lieb, W. R. (1986). Transport and diffusion across cell membranes.

Stein, W. D., & Litman, T. (2014). Channels, carriers, and pumps: An introduction to membrane transport. Elsevier.

Striano, P., Weber, Y. G., Toliat, M. R., Schubert, J., Leu, C., Chaimana, R., Baulac, S., Guerrero, R., LeGuern, E., Lehesjoki, A.-E., Polvi, A., Robbiano, A., Serratosa, J. M., Guerrini, R., Nürnberg, P., Sander, T., Zara, F., Lerche, H., Marini, C., & EPICURE Consortium. (2012). GLUT1 mutations are a rare cause of familial idiopathic generalized epilepsy. Neurology, 78(8), 557–562. https://doi.org/10.1212/WNL.0b013e318247ff54

Suls, A., Mullen, S. A., Weber, Y. G., Verhaert, K., Ceulemans, B., Guerrini, R., Wuttke, T. V., Salvo-Vargas, A., Deprez, L., Claes, L. R. F., Jordanova, A., Berkovic, S. F., Lerche, H., De Jonghe, P., & Scheffer, I. E. (2009). Early-onconjunto absence epilepsy caused by mutations in the glucose transporter GLUT1. Annals of Neurology, 66(3), 415–419. https://doi.org/10.1002/ana.21724

Sun, L., Zeng, X., Yan, C., Sun, X., Gong, X., Rao, Y., & Yan, N. (2012). Crystal structure of a bacterial homologue of glucose transporters GLUT1-4. Nature, 490(7420), 361–366. https://doi.org/10.1038/nature11524

Teppa, E., Wilkins, A. D., Nielsen, M., & Buslje, C. M. (2012). Disentangling evolutionary signals: Conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. BMC Bioinformatics, 13(1), 235.

Tsirigos, K. D., Peters, C., Shu, N., Käll, L., & Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Research, 43(W1), W401-407. https://doi.org/10.1093/nar/gkv485

Vardy, E., Arkin, I. T., Gottschalk, K. E., Kaback, H. R., & Schuldiner, S. (2004). Structural conservation in the major facilitator superfamily as revealed by comparative modeling. Protein Science : A Publication of the Protein Society, 13(7), 1832–1840. https://doi.org/10.1110/ps.04657704

Vidaver, G. A. (1966). Inhibition of parallel flux and augmentation of counter flux shown by transport models not involving a mobile carrier. Journal of Theoretical Biology, 10(2), 301–306. https://doi.org/10.1016/0022-5193(66)90128-7

Wang, C., Bao, X., Li, Y., Jiao, C., Hou, J., Zhang, Q., Zhang, W., Liu, W., & Shen, Y. (2015). Cloning and characterization of heterologous transporters in Saccharomyces cerevisiae and identification of important amino acids for xylose utilization. Metabolic Engineering, 30, 79–88.

Wang, C., Shen, Y., Hou, J., Suo, F., & Bao, X. (2013). An assay for functional xylose transporters in Saccharomyces cerevisiae. Analytical Biochemistry, 442(2), 241–248. https://doi.org/10.1016/j.ab.2013.07.041

Wang, D., Kranz-Eble, P., & De Vivo, D. C. (2000). Mutational analysis of GLUT1 (SLC2A1) in Glut-1 deficiency syndrome. Human Mutation, 16(3), 224–231. https://doi.org/10.1002/1098-1004(200009)16:3<224::AID-HUMU5>3.0.CO;2-P

Wang, D., Pascual, J. M., Iserovich, P., Yang, H., Ma, L., Kuang, K., Zuniga, F. A., Sun, R. P., Swaroop, K. M., Fischbarg, J., & De Vivo, D. C. (2003). Functional studies of threonine 310 mutations in Glut1: T310I is pathogenic, causing Glut1 deficiency. The Journal of Biological Chemistry, 278(49), 49015-49021. https://doi.org/10.1074/jbc.M308765200

Wang, Dong, Pascual, J. M., Yang, H., Engelstad, K., Jhung, S., Sun, R. P., & De Vivo, D. C. (2005). Glut-1 deficiency syndrome: Clinical, genetic, and therapeutic aspects. Annals of Neurology, 57(1), 111–118. https://doi.org/10.1002/ana.20331

Wang, K., & Samudrala, R. (2006). Incorporating background frequency improves entropy-based residue conservation measures. BMC Bioinformatics, 7, 385. https://doi.org/10.1186/1471-2105-7-385

Watson, J. D., Laskowski, R. A., & Thornton, J. M. (2005). Predicting protein function from sequence and structural data. Current Opinion in Structural Biology, 15(3), 275–284.

Weber, Y. G., Storch, A., Wuttke, T. V., Brockmann, K., Kempfle, J., Maljevic, S., Margari, L., Kamm, C., Schneider, S. A., Huber, S. M., Pekrun, A., Roebling, R., Seebohm, G., Koka, S., Lang, C., Kraft, E., Blazevic, D., Salvo-Vargas, A., Fauler, M., … Lerche, H. (2008). GLUT1 mutations are a cause of paroxysmal exertion-induced dyskinesias and induce hemolytic anemia by a cation leak. The Journal of Clinical Investigation, 118(6), 2157–2168. https://doi.org/10.1172/JCI34438

Widdas, W. F. (1952). Inability of diffusion to account for placental glucose transfer in the sheep and consideration of the kinetics of a possible carrier transfer. The Journal of Physiology, 118(1), 23–39. https://doi.org/10.1113/jphysiol.1952.sp004770

Wieczorke, R., Krampe, S., Weierstall, T., Freidel, K., Hollenberg, C. P., & Boles, E. (1999). Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in Saccharomyces cerevisiae. FEBS Letters, 464(3), 123-128.

Yan, N. (2013). Structural advances for the major facilitator superfamily (MFS) transporters. Trends in Biochemical Sciences, 38(3), 151–159. https://doi.org/10.1016/j.tibs.2013.01.003

Yan, N. (2015). Structural Biology of the Major Facilitator Superfamily Transporters. Annual Review of Biophysics, 44(1), 257–283. https://doi.org/10.1146/annurev-biophys-060414-033901

Yan, N. (2017). A Glimpse of Membrane Transport through Structures-Advances in the Structural Biology of the GLUT Glucose Transporters. Journal of Molecular Biology, 429(17), 2710–2725. https://doi.org/10.1016/j.jmb.2017.07.009

Young, E. M., Comer, A. D., Huang, H., & Alper, H. S. (2012). A molecular transporter engineering approach to improving xylose catabolism in Saccharomyces cerevisiae. Metabolic Engineering, 14(4), 401–411.

Zhao, Z., Xian, M., Liu, M., & Zhao, G. (2020). Biochemical routes for uptake and conversion of xylose by microorganisms. Biotechnology for Biofuels, 13(1), 21.

9 ANEXOS

Anexo 1. Sistema de clasificación de la comisión de transporte

Saier y col. (2000) desarrollaron el sistema de clasificación de transporte (TC) en forma análoga a la clasificación de la Comisión de Enzimas (EC). Este sistema es reconocido y recomendado por la Unión Internacional de Bioquímica y Biología Molecular (IUBMB). Su principal diferencia con el sistema de la Comisión de Enzimas es que basa la clasificación nó solo en características funcionales, sino que además integra análisis filogenéticos. Cada transportador se identifica con un código alfanumérico de 5 componentes. El primer componente, un número, indica la clase de transporte asociado, los principales a saber: 1, Canales y poros; 2, transportadores impulsados por gradiente electroquímico (facilitadores, co y contratransportadores); 3, transportadores asociados al transporte activo primario y 4, translocadores de grupo (existen otras tres clases, donde la última agrupa a transportadores sin caracterizar, véase Saier Jr *et al*., 2015). El segundo componente, una letra, indica la subclase de transporte, definida mediante criterios estructurales o funcionales, como se muestra en la tabla dispuesta al final del anexo.

El tercer y cuarto componente, ambos números, se asocian a criterios de clasificación filogenéticos, siguiendo una jerarquía de superfamilia> familia> subfamilia. Finalmente, el último componente numérico indica el sustrato o gama de sustratos transportados. y la polaridad del transporte asociada (entrada o salida). Cabe destacar que este sistema, frente a la existencia dos o más transportadores en la misma subfamilia que transportan el mismo sustrato o rango de sustratos usando el mismo mecanismo de transporte, les asigna un mismo código TC, independientemente de si son parálogos u ortólogos. A los homólogos secuenciados de función desconocida normalmente no se les asigna un código TC, a menos que representen una familia o subfamilia única, o bien sean de un reino de organismos no representados.

La base de datos de la comisión de transporte, TCDB (http://www.tcdb.org/) contiene información no-redundante de más de 10.000 transportadores de eucariontes y procariontes, los cuales se agrupan en más de 1000 familias, según los criterios enunciados anteriormente.

Tabla. Clases y subclases de transportadores en el sistema TC (adaptada de Saier, 2000)

1. Canales y poros
    1.A Canales tipo α
    1.B Porinas tipo barril-β
    1.C Toxinas formadoras de poros (proteínas y péptidos)
    1.D Canales sintetizados no ribosómicamente
2. Transportadores impulsados por gradiente electroquímico
    2.A Portadores (uniportadores, simportadores, antiportadores)
    2.B Portadores sintetizados no ribosómicamente
    2.C Energizadores guiados por gradiente iónico
3. Transportadores activos primarios
    3.A Transportadores impulsados por hidrólisis de enlace difosfato
    3.B Transportadores impulsados por decarboxilación
    3.C Transportadores impulsados por transferencia de metilo
    3.D Transportadores impulsados por oxidoreducción
    3.E Transportadores impulsados por absorción de luz
4. Translocadores de grupo
    4.A. Translocadores impulsados por transferencia de fosfato

Anexo 2. Códigos PDB empleados para análisis de estadísticos de longitud de TMs de familia MFS (Tabla 13)

3O7Q, 4ZP2, 4ZOW, 4M64, 4ZP0, 3O7P, 4J05, 4IU8, 4IU9, 3WDO, 4GBY, 4GBZ, 4GC0, 1PW4, 2CFQ, 2CFP, 4U4T, 4U4V, 4YB9, 4ZYR, 6S7V, 4LDS, 2Y5Y, 5GXB, 4JR9, 4JRE, 6C9W, 4YBQ, 4ZWB, 5EQI, 5AYN, 5AYM, 5AYO, 4ZWC, 4ZW9, 5EQH, 5EQG, 4PYP, 6H7D, 6GV1, 6HCL, 6G9X, 6EUQ, 6OOQ, 6OOP, 6E9N, 6OOM, 5C65, 6E9C, 6E8J, 6BTX.
Total = 51 estructuras.

Anexo 3. Estadísticos de largos de segmentos topológicos por categoría taxonómica

		N	TM1	L1	TM2	L2	TM3	L3	TM4	L4	TM5	L5	TM6	L6	TM7	L7	TM8	L8	TM9	L9	TM10	L10	TM11	L11	TM12	C
B	min	1	15	5	20	6	19	3	16	7	18	3	18	28	16	6	18	3	18	3	20	3	16	3	19	4
	max	69	27	28	25	12	23	18	27	22	26	17	24	78	31	18	26	9	29	18	31	19	31	15	27	62
	$\tilde{x}$	19.8	21.1	15.8	21.6	6.7	20.3	5.4	23.6	11.2	23.7	9.3	21.4	52.6	25	11.6	22.7	5.3	22	7.4	26.9	9.3	25	5.1	21.5	22.9
	sd	11.2	3.2	4.3	1	1.4	1	2.8	1.9	3.1	1.8	3.7	1.6	10.6	3.2	2.5	1.6	2.1	2.5	3.8	2.9	4.4	3.8	3	1.7	10.2
Pr	min	5	18	20	21	6	19	4	23	9	22	9	16	31	19	3	20	3	21	8	24	10	22	3	19	26
	max	105	31	85	23	9	20	8	25	13	26	20	23	64	30	13	25	8	26	13	30	14	24	17	23	120
	$\tilde{x}$	46.4	26.6	63.4	21.6	7.1	19.3	6.8	23.7	11.8	23.1	14.5	18.9	42.9	22.1	10.1	22.4	5.6	22.9	11.5	25.3	12.7	23.4	13.8	20.4	39.9
	sd	26.1	4.4	23.6	0.7	0.9	0.5	1.4	0.9	1.1	1.2	2.5	2.5	8.7	3	2.7	1.8	2.1	1.9	1.8	1.9	1.2	0.7	4.9	1.2	26
F	min	7	16	9	18	5	18	3	16	3	16	5	18	28	17	5	18	3	17	3	16	3	16	3	19	8
	max	142	28	48	25	30	30	9	28	25	31	37	25	130	31	29	31	20	28	33	31	22	30	22	28	344
	$\tilde{x}$	49.9	22.5	25.7	21.8	7.3	19.5	5.3	24.2	10.7	23.3	10.9	21.9	66.7	24.4	13	22.1	4.9	23.7	11.4	27.1	11.2	24.4	5.7	20.6	65.1
	sd	29.7	3	6.1	1.3	2.5	1.3	1.6	2	3.7	1.9	3.8	1.5	13.3	2.5	3.2	1.9	2.6	2.4	4.2	3.2	3.1	2.7	4.4	1.5	37.5
Pp	min	1	15	14	20	5	19	3	21	7	21	3	19	58	19	5	20	3	21	5	23	5	22	3	19	16
	max	104	28	46	24	10	21	9	27	14	30	26	25	327	31	30	28	8	28	92	31	13	28	18	26	73
	$\tilde{x}$	29.5	20.9	27.2	21.7	6.8	19.5	4.5	24.9	9.7	23.7	9.7	21.8	71.7	25.6	13.4	22.3	4.3	24.4	13.7	27.7	10.4	25.2	4.5	21.4	34.8
	sd	16.4	3.2	10.3	0.6	1.1	0.5	1.3	0.9	1.4	2	4.8	1.2	49.2	2.1	4.4	1.5	1.3	2	15	1.8	2.2	1.2	3.8	1.2	13.1
M	min	8	16	15	19	5	18	3	19	7	20	3	18	47	22	5	19	3	20	3	20	4	15	2	20	17
	max	463	26	78	28	8	22	8	27	12	27	13	25	65	30	16	24	7	28	98	31	16	31	10	25	985
	$\tilde{x}$	78	20.7	30.1	22.3	6.3	19.7	5.9	24.5	9.4	23	8.6	21.3	59.4	25.6	12.4	21.7	3.9	24	22.3	27.2	11.1	24.7	3.3	21.1	51.6
	sd	132.5	3.4	12.6	1.6	0.7	0.7	1.7	1.1	1	1.5	3	1.3	2.7	1.7	1.7	1	1.1	1.4	27.5	1.8	1.9	1.5	1	1.1	96.5
		48.8	21.6	27.3	21.9	6.8	19.7	5.3	24.4	10.2	23.4	9.9	21.6	63.2	25.1	12.7	22.1	4.5	23.7	14.5	27.2	10.8	24.7	5	21	47.6

Anexo 4. Sustratos con evidencia experimental

Cat.[a]	Uniprot ID	Organism	Name[b]	Specificities and affinities[c]	Mec.[d]	Refs.[e]
B	Q8G3X1	*Bifidobacterium longum*	glcP	glucose, 2DG > mannose > galactose > xylose > fructose > arabinose	S	16452407
	Q8NTX0	*Corynebacterium glutamicum*	iolT1	inositol (0,2) > glucose		21452034, 16997948*
	Q8NL90	*Corynebacterium glutamicum*	iolT2	inositol (0,45) , not-glucose		21452034, 16997948*
	C4B4V9	*Corynebacterium glutamicum*	araE	Arabinose, xylose, glucose	S	23928049
	A0QZX3	*Mycobacterium smegmatis*	n.d	glucose (0.019) , 2DG, not-xylose, not-arabinose, not-fructose, not-trehalose	F	18311074*
	Q0SE66	*Rhodococcus jostii*	n.d	glucose		21464575
	Q7BEC4	*Streptomyces lividans*	glcP	glucose (0,041)		15659175
	P15729	*Synechocystis sp*	gtr	glucose, fructose	S	2507869
	P96710	*Bacillus subtilis*	araE	arabinose, xylose, galactose, arabinobiose	S	9401028, 20693325
	P42417	*Bacillus subtilis*	iolF	inositol (0,327)	S	11807058*
	O34718	*Bacillus subtilis*	iolT	inositol (0,707)	S	11807058*
	O52733	*Lactobacillus brevis*	xylT	xylose (0,215) , 6-DG > glucose > arabinose > galactose, not-ribose, not-fucose	S	9835554*
	P0AE24	*Escherichia coli*	araE	arabinose (0.15-0.17) , fucose	S	11739756, 6282256*
	P0AEP1	*Escherichia coli*	galP	galactose(0.07-0.17), 2-deoxy-galactose, fucose, glucose, 2DG, arabinose, a-MG, xylose	S	23115, 15558, 6282256*, 11693918
	P0AGF4	*Escherichia coli*	xylE	xylose, glucose	S	23075985
	A0A0H3NW06	*Salmonella*	iolT1	inositol (0,5-0.8)	S	19833776*

		typhimurium				
	A0A0H3NKX8	*Salmonella typhimurium*	iolT2	inositol	S	19833776
	P21906	*Zymomonas mobilis*	glf	glucose (4.1) > fructose (39)	F	7768841
	Q43975	*Acinetobacter baylyi*		4-hydroxybenzoate		24907408
	Q51955	*Pseudomonas putida*		4-hydroxybenzoate, protocatechuate		7961399
	E8ZB61	*Pseudomonas putida*	GalT	Gallate		21219457
	Q5EXK5	*Klebsiella oxytoca*		3-hydroxybenzoate(0.006)		22729544*
	O52718	*Klebsiella pneumonae*	DalT	arabinatol	S	9324246
	O52717	*Klebsiella pneumonae*	RbtT	ribotol	S	9324246
	Q47421	*Dickeya dadantii*		Betaine(0.05), proline, ectoine, pipecolic acid		8550465, 16000740*
	A0A0H2VG78	*Staphylococcus epidermidis*	glcP	Glucose(0.029)		24127585*
Pr	Q01440	*Leishmania donovani*	GTR1	inositol (0,64)	S	9305873*
	B1PLM1	*Leishmania mexicana*	LmGT4	glucose (0,44) > fructose(7.3)		19017272*
	O76486	*Leishmania mexicana*	LmGT1	glucose sensor		25300620
	O61059	*Leishmania mexicana*	LmGT2	glucose(0,11) > ribose(0.98)	F	20601430*
	O61060	*Leishmania mexicana*	LmGT3	glucose (0.21) > ribose (5.75)	F	20601430*
	O97467	*Plasmodium falciparum*	HT1	glucose(0.48) > mannose > galactose > fructose	F	10066789*
	Q06221, Q09037	*Trypanosoma brucei*	THT1B/E	2DG(0.5), glucose(1) > mannose(0.7) > fructose(2.56) > NAGlcN(11) > glucosamine(21), not-galactose, not-xylose, not-mannitol , not-glycerol	F	8423781*, 2790048*

	Q06222	*Trypanosoma brucei*	THT2A	glucose(0.05) > 2DG(0.06) > fructose (2.54)	F	8554506*
	Q27115	*Trypanosoma vivax*	HT1	glucose, 2DG > mannose > glucosamine > fructose, not-galactose	F	8620878
F	A2R3H2	*Aspergillus niger*	gatA	galacturonate > xylose		25177540
	Q8J0V1	*Aspergillus niger*	mstA	glucose(0.03) > mannose(0.06) > xylose(0.3) > fructose(4.5)	S	14717659*
	Q8J0U9	*Aspergillus niger*	mstC	glucose	S	17526853
	O74713	*Candida albicans*	HGT1	glucose		10612724
	Q5A8J5	*Candida albicans*	HGT10	glycerol	S	19383674
	Q59Q30	*Candida albicans*	GIT1	GroPIns(0.028), not- GroPCho		21984707*
	A0A1D8PN12	*Candida albicans*	GIT2	not- GroPIns l		24114876)*
	Q5A1L6	*Candida albicans*	GIT3	GroPCho (0.045), not-GroPIns		24114876*
	A0A1D8PN14	*Candida albicans*	GIT4	GroPCho (0.016), not-GroPIns		24114876*
	Q2MDH1	*Candida intermedia*	GXF1	glucose, xylose	F	16402921
	Q2MEV7	*Candida intermedia*	GXS1	glucose, xylose	S	16402921
	Q64L87	*Debaryomyces fabryi*	Xylhp	Xylose	F	10427054
	Q400D8	*Emericella nidulans*	mstE	glucose		16418173
	P15325	*Emericella nidulans*	qutD	quinate		2976880
	A0ZXK6	*Geosiphon pyriformis*	mst1	glucose(1.2) > mannose > galactose, not-fructose, not-xylose	S	17167486
	A8DCT2	*Gibberella moniliformis*	FST1	inositol		27195938

P49374	*Kluyveromyces lactis*	HGT1	glucose(1), galactose		17156020, 12677461*
P07921	*Kluyveromyces lactis*	LAC12	lactose, galactose	S	3053697, 17156020
P18631	*Kluyveromyces lactis*	RAG1	glucose(20-50), not-galactose		2402460, 12677461*
Q8NJ22	*Kluyveromyces lactis*	frt1	fructose(0.16)	S	12677461*
C4QVV9	*Komagataella phaffii*	gt1	glycerol		27189360
Q7SCU1	*Neurospora crassa*	cdt1	cellobiose(4)		20829451*
Q7SD12	*Neurospora crassa*	cdt2	cellobiose(3.2)		20829451*
P11636	*Neurospora crassa*	qa	quinate	S	2525625
Q7RVX9	*Neurospora crassa*	Pho-5	Phosphate(0.04)	S	7732001
B1PM37	*Pichia angusta*	HXS1	hexose sensor		18310355
Q32SL4	*Pichia angusta*	MAL2	maltose		16103021
A0A097NV01	*Pneumocystis carinii*	ITR1	inositol (1), not-glucose	S	27965450*
P13181	*Saccharomyces cerevisiae*	GAL2	galactose, glucose(1.6), xylose	F	3082856, 23928049, 12702270*
P32465	*Saccharomyces cerevisiae*	HXT1	glucose(100)> not-fructose(300) > not-xylose (880); mannose, not-galactose		12702270*, 9151960,* 17180689*, 299703
P23585	*Saccharomyces cerevisiae*	HXT2	glucose(2-10)> fructose(6-20) > not-xylose (260)		12702270*, 9151960,* 17180689*

P32466	*Saccharomyces cerevisiae*	HXT3	glucose(30-60) > fructose (125)	12702270*, 9151960*
P32467	*Saccharomyces cerevisiae*	HXT4	glucose(6-9) > fructose(52) > xylose (170)	12702270*, 9151960,* 17180689*
P38695	*Saccharomyces cerevisiae*	HXT5	mannose, glucose, fructose, not-galactose	10618490
P39003	*Saccharomyces cerevisiae*	HXT6	glucose (1.2)	12702270*
P39004	*Saccharomyces cerevisiae*	HXT7	glucose(1.1-2.1) > fructose(2-5) > xylose(130); mannose, galactose	12702270*, 9151960,* 17180689*, 10618490
P40886	*Saccharomyces cerevisiae*	HXT8	mannose, glucose, fructose, not-galactose	10618490
P40885	*Saccharomyces cerevisiae*	HXT9	mannose, glucose, fructose, galactose	10618490
P43581	*Saccharomyces cerevisiae*	HXT10	mannose, fructose, galactose, glucose	10618490
P54862	*Saccharomyces cerevisiae*	HXT11	glucose, fructose, mannose, galactose, xylitol(159.1)	10618490, 26996892*
P39924	*Saccharomyces cerevisiae*	HXT13	mannitol (16.7) > sorbitol (20.4), not-hexoses	10618490, 26996892*
P42833	*Saccharomyces cerevisiae*	HXT14	galactose, not-glucose, not-fructose, not-mannose	10618490
P54854	*Saccharomyces cerevisiae*	HXT15	mannitol(11.4) > sorbitol(38.9) > xylitol(143.3); mannose, fructose, glucose, not-galactose	10618490, 26996892*
P47185	*Saccharomyces cerevisiae*	HXT16	sorbitol(152.6), not-mannitol(527.6); fructose, glucose, not-galactose	10618490, 26996892*
P53631	*Saccharomyces*	HXT17	mannitol(18.6), sorbitol	10618490,

	cerevisiae		(155.7); mannose, fructose, glucose, not-galactose		26996892*
P30605	*Saccharomyces cerevisiae*	ITR1	inositol(0.1)	S	2040626*
P30606	*Saccharomyces cerevisiae*	ITR2	inositol(0.14)		2040626*
P53048	*Saccharomyces cerevisiae*	MAL11/ AGT1	maltose, maltotriose, a-MG, trehalose, turanose, palatinose, melezitose, glucose	S	12210897, 10618490
P38156	*Saccharomyces cerevisiae*	MAL31	maltose, maltotriose, turanose, glucose, not-a-MG, not-melezitose, not-trehalose		12210897
P15685	*Saccharomyces cerevisiae*	MAL61	maltose, turanose	S	12210897, 16088872
P0CD99, P0CE00	*Saccharomyces cerevisiae*	MPH2, MPH3	maltose (4.4) > maltrotriose (7.2), a-MG, glucose, turanose, not-melezitose, not-trehalose		10618490, 12210897*
Q12300	*Saccharomyces cerevisiae*	RGT2	glucose sensor		9564039
P10870	*Saccharomyces cerevisiae*	SNF3	glucose sensor		9564039
P39932	*Saccharomyces cerevisiae*	STL1	glicerol (1.7)	S	15703210, 9398075*
P25346	*Saccharomyces cerevisiae*	GIT1	Glycerophosphoinositol, glycerophosphocholine		9691030,12912892,16141200
P25297	*Saccharomyces cerevisiae*	PHO84	Phosphate		7851439
P36035	*Saccharomyces cerevisiae*	Jen1	lactate	S	10198029
Q9HFF8	*Saccharomyces*	FSY1	Fructose(0.16)	S	10986274

		pastorianus				
	Q9P3U6	*Schizosaccharomyces pombe*	ght1	glucose(4-6) >fructose(15)		9090050*, 10735857*
	O74969	*Schizosaccharomyces pombe*	ght2	glucose(2)>fructose(10)	S	10735857*
	Q92339	*Schizosaccharomyces pombe*	ght3	gluconate	S	10735857
	P78831	*Schizosaccharomyces pombe*	ght5	glucose(0.6 mM) > fructose(50)		10735857*
	O74849	*Schizosaccharomyces pombe*	ght6	fructose(5) > glucose(8)	S	10735857*
	P87110	*Schizosaccharomyces pombe*	ITR2	inositol		9560432
	G4TIR7	*Serendipita indica*	HXT5	glucose(2.56) > fructose > xylose > mannose > galactose	S	27499747*
	A0A0K1NY69	*Wickerhamomyces anomalus*	STL1	glycerol	S	24225317
M	Q7PIR5	*Anopheles gambiae*	Tret1	trehalose(45.75)		20035867
	A9ZSY2	*Apis mellifera*	Tret1	trehalose(9.42)		20035867
	A9ZSY3	*Bombyx mori*	Tret1	trehalose(71.58)		20035867
	A1Z8N1	*Drosophila melanogaster*	Tret1	trehalose(10.94), not-maltose, not-sucrose, not-lactose	F	20035867
	A5LGM7	*Polypedilum vanderplanki*	Tret1	trehalose(114.5), not-maltose, not-sucrose, not-lactose	F	20035867
	P11166	*Homo sapiens*	GLUT1	DHA(1.1) > glucose(1.6-3) > 3OMG(3.6) > 2DG(11.6) > galactose(25)	F	2217557*,12135767*, 22365203*, 9228080*

P11168	*Homo sapiens*	GLUT2	glucosamine(0.8) > DHA(2.33) > 2DG(11-25) > glucose(17) > fructose (76-108), galactose (92) > mannose (125)	F	8457197*, 23506862*, 9477959*, 23396969*
P11169	*Homo sapiens*	GLUT3	2DG(1.4) > DHA(1.7) > 3OMG(3.5) > galactose(7); glucose, mannose, DHA, fucose, xylose, galactose, glucosamine, arabinose, not-fructose, not-inositol, not-ribose, not-NAGlcN	F	8457197*, 9477959*, 1445263*, 1420159*, 8457197*, 9228080*, 26176916
P14672	*Homo sapiens*	GLUT4	DHA (0.98) > glucosamine (3.9) > 2DG(5.2) > glucose (5.5-6.6)	F	2036379*, 12135767*, 10862609*,
P22732	*Homo sapiens*	GLUT5	glucose(0.36) > fructose (6)	F	17710649*, 1634504*
Q6PXP3	*Homo sapiens*	GLUT7	fructose (0.13) > glucose (0.3) ,not-DHA, not-galactose, not-xylose	F	16186102*, 15033637*, 23396969
Q9NY64	*Homo sapiens*	GLUT8	trehalose, fructose	F	27922102, 24519932
Q9NRM0	*Homo sapiens*	GLUT9	fructose(0.42) > glucose (0.61) > urate (0.89), not-DHA	F	17710649*, 18327257*, 23396969
O95528	*Homo sapiens*	GLUT10	2DG(0.3) ,glucose > galactose, not-DHA	F	11592815*, 23396969
Q9BYW1	*Homo sapiens*	GLUT11	fructose(0.06) > glucose(0.16), not-DHA	F	17710649*, 23396969
Q8TD20	*Homo sapiens*	GLUT12	glucose, not-DHA	F(S)	18039784, 20487568, 23396969
Q96QE2	*Homo sapiens*	GLUT13	inositol(0.33)	S	20230529*
Q8TDB8	*Homo sapiens*	GLUT14	2DG, DHA, not-fructose	F	28971850

N0A4A7	*Megalobrama amblycephala*	GLUT2	glucose	F	29560575
P17809	*Mus musculus*	GLUT1	glucose, DHA	F	18668520
P14246	*Mus musculus*	GLUT2	glucose, DHA	F	18668520
P32037	*Mus musculus*	GLUT3	glucose, DHA	F	18668520
P14142	*Mus musculus*	GLUT4	glucose	F	26629404
Q9WV38	*Mus musculus*	GLUT5	fructose(13)		12031501*
Q9JIF3	*Mus musculus*	GLUT8	DHA(3.23) >2DG(10.3), glucose > fructose (96)		23396969*
Q3T9X0	*Mus musculus*	GLUT9	glucose, urate(0.649)		14657010, 19587147*
Q8VHD6	*Mus musculus*	GLUT10	DHA	F	20639396
P11167	*Rattus norvegicus*	GLUT1	glucose, DHA(1.1)	F	12006627, 8473295*
P12336	*Rattus norvegicus*	GLUT2	fructose(16) > glucose(30)	F	8405848*
Q07647	*Rattus norvegicus*	GLUT3	glucose, 3OMG(2.87), DHA	F	8645164*
P19357	*Rattus norvegicus*	GLUT4	glucose	F	1733237
P43427	*Rattus norvegicus*	GLUT5	fructose > 2DG, glucose	F	8333543
Q9JJZ1	*Rattus norvegicus*	GLUT8	glucose,2DG(2.4)> fructose> galactose, DHA	F	10671487*, 23396969
Q921A2	*Rattus norvegicus*	GLUT13	inositol	S	20230529
O44827	*Caenorhabditis elegans*	fgt	2DG(2.8), glucose, mannose > fructose > galactose	F	23826391
A5Y0C3	*Nilaparvata lugens*	HT1	glucose(3)	F	17916500
Q26579	*Schistosoma mansoni*	GTP1	3OMG(1.3), glucose > mannose > galactose > fructose		8307988

Pp	Q2V4B9	*Arabidopsis thaliana*	At1g79820	2DG		16855027
	Q8L6Z8	*Arabidopsis thaliana*	At3g03090 / Vgt1	glucose(3.7)> fructose > galactose > xylose	A	17284600*, 18629494
	Q6AWX0	*Arabidopsis thaliana*	At5g17010/Vgt2	glucose, xylose		18629494
	Q56ZZ7	*Arabidopsis thaliana*	At5g16150	glucose(20)		10810150*
	Q0WWW9	*Arabidopsis thaliana*	At5g59250	glucose, sucrose, xylose	A	18629494, 30482788
	Q94KE0	*Arabidopsis thaliana*	ESL1	glucose, 2DG > galactose > mannose > fructose > xylose, not-inositol, not-sorbitol, not-mannitol	F	19901034
	Q8VZR6	*Arabidopsis thaliana*	INT1	inositol	S	18441213
	Q9C757	*Arabidopsis thaliana*	INT2	inositol(1.16)	S	20230529
	O23492	*Arabidopsis thaliana*	INT4	inositol(0.24)	S	16603666
	Q96290	*Arabidopsis thaliana*	MSSP1	glucose, sucrose	A	21668536, 17158605
	Q8LPQ8	*Arabidopsis thaliana*	MSSP2	glucose, sucrose	A	21668536, 17158605
	Q9XIH7	*Arabidopsis thaliana*	PLT1	Xylitol	S	19969532
	Q9XIH6	*Arabidopsis thaliana*	PLT2	xylitol	S	19969532
	Q8VZ80	*Arabidopsis thaliana*	PLT5	ribose, fucose, xylose, xilitol, rhammose, glucose (1.5), arabinose, erythrose, erythritol, sorbitol (0.5), a-	S	15598803, 15525644

			MG, inositol (3.5), mannose, fructose, galactose, glucuronate, glycerol (23.4)		
P23586	*Arabidopsis thaliana*	STP1	glucose(0.02),3OMG(0.1), 2DG, mannose, xylose, galactose > fucose > fructose > arabinose	S	2209537*, 8051137
Q9LNV3	*Arabidopsis thaliana*	STP2	galactose(0.02) > 3OMG(0.05), xylose, mannose, glucose , fructose	S	10074716*
Q8L7R8	*Arabidopsis thaliana*	STP3	glucose(2) > xylose > 3OMG > mannose > galactose > fructose	S	10.1046/j.1365-3040.2000.00538.x,
Q39228	*Arabidopsis thaliana*	STP4	glucose(0.015),3OMG(0.1),galactose, xylose, mannose, not-fructose	S	8989877*
Q9SFG0	*Arabidopsis thaliana*	STP6	glucose(0.02), mannose, fructose, galactose, xylose, ribulose	S	12529516*
O04249	*Arabidopsis thaliana*	STP7	arabinose(0.29), xylose, weak recognition of other sugars		29311272*
Q9SBA7	*Arabidopsis thaliana*	STP8	glucose(0.25) > galactose > mannose > xylose > fructose, not-ribose, not-sucrose.		29311272*
Q9SX48	*Arabidopsis thaliana*	STP9	glucose(0.084), galactose, mannose, xylose		12970485*
Q9LT15	*Arabidopsis thaliana*	STP10	glucose(0.01) >galactose >mannose, not-fructose, not-xylose, not-ribose	S	26893494*
Q9FMX3	*Arabidopsis thaliana*	STP11	glucose(0.025), mannose, galactose, xylose, not-fructose, not-ribose		15565288*
O65413	*Arabidopsis*	STP12	glucose(0.17) > galactose >	S	29311272*

	thaliana		mannose > fructose, not-xylose, not-ribose, not-sucrose		
Q94AZ2	*Arabidopsis thaliana*	STP13	glucose(0.074) > galactose > fructose > mannose > ribose, not-xylose, not-sucrose		16616142*
Q8GW61	*Arabidopsis thaliana*	STP14	galactose(0.529), not-mannose, not-glucose,3OMG, not-fructose, not-ribose		20627950*
A0A2C9WLH2	*Manihot esculenta*	STP2	mannose, fructose, glucose, galactose xylose		29587418
A0A2C9W8R5	*Manihot esculenta*	STP7	mannose, galactose, glucose, fructose, not-xylose		29587418
A0A2C9UCP7	*Manihot esculenta*	STP16	galactose		29587418
A1Z264	*Galdieria sulphuraria*	SPT1	fucose,mannitol, glucose,fructose	S	17497961
Q1XF07	*Lotus japonicus*	PLT4	xylitol(0.34)>fructose>xylose		21219252*
Q851G4	*Oryza sativa*	MST2	glucose	S	11038054
Q7EZD7	*Oryza sativa*	MST3	glucose, xylose, 3OMG	S	11038054
Q10PW9	*Oryza sativa*	MST4	galactose(0.095) > mannose(0.102) > xylose(0.135), ribose(0.142) > glucose(0.209)> fructose(0.271)		18506478*
Q6ZKF0	*Oryza sativa*	MST5	glucose, xylose, 3OMG	S	12723603
Q6Z401	*Oryza sativa*	MST6	galactose(0.049)> mannose(0.108) > ribose(0.110) > xylose (0.151)> fructose(0.257) > glucose(0.266)		18506478

Q94DB8	*Oryza sativa*	PHT1	Phosphate		12271140
P15686	*Parachlorella kessleri*	HUP1	glucose, fructose, mannose, xylose, 3OMG	S	10965467
Q39524	*Parachlorella kessleri*	HUP2	galactose, xylose, 3OMG, mannose, not-fructose		10965467
Q39525	*Parachlorella kessleri*	HUP3	glucose, fructose, mannose, xylose, 3OMG	S	10965467
A0A097P980	*Prunus salicina*	SOT1	sorbitol(0.64)		12692316*
A0A075BFV8	*Pyrus pyrifolia*	SOT2	sorbitol(0.81)		12692316*
Q8VYM2	*Arabidopsis thaliana*	PHT1-1	Phosphate(0.003), arsenate	S	9192698*
Q494P0	*Arabidopsis thaliana*	PHT1-7	arsenate	S	23108027
Q9SYQ1	*Arabidopsis thaliana*	PHT1-8	Phosphate	S	25428623
Q9S735	*Arabidopsis thaliana*	PHT1-9	Phosphate	S	25428623

a Categoría taxonómica: B,bacteria; Pr, protozoa; F, fungi; Pp, primoplantae; M, metazoa.
b Nombre común del transportador: n.d, no definido.
c Moléculas capaces de ser transportadas (subrayadas) o reconocidas (sin existir caracterización experimental directa del transporte, fuente normal) por el transportador. Km aparente de entrada conocida para un sustrato se indica en paréntesis, en mM. Si se conoce preferencia de unión, se indica con el símbolo '>'. Para moléculas no reconocidas o escasamente reconocidas por el transportador, se antepone la palabra 'not-'. Los criterios empleados para la anotación se entregan con detalle en el texto.
d Mecanismo de transporte: F, facilitador; S, cotransportador sustrato:H+. Casilla se deja en blanco de no existir caracterización o alusión explícita del mecanismo en literatura consultada.
e Códigos PMID/PMC de referencias consultadas. Se indican con * los artículos que reportan parámetros cinéticos.

Anexo 5. Matrices de similaridad e identidad para clados de interés. Matrices de identidad y de similaridad calculadas por clado correspondientes a las triangulares superiores e inferiores, respectivamente.La enumeración de los clados se corresponde con la empleada en la Figura 25. Cada matriz se acompaña de una tabla que indica, para cada secuencia, la categoría taxonómica a la que pertenece mediante el mismo código de colores ya empleado (Figura 25), su número en la matriz, en el filograma, código Uniprot, organismo y nombre del gen o producto proteico; respectivamente.

CLADO 1.

	0	1	2	3
0	100	54.1	36.3	36.5
1	83.5	100	34.6	33.7
2	75.4	71.9	100	88.3
3	73.9	71.7	98.5	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	1	O05390	B.subtilis	csbX
1	2	O34864	B.subtilis	yoaB
2	3	O52717	K.pneumoniae	rbtT
3	4	O52718	K.pneumoniae	dalT

CLADO 2.

	0	1	2	3
0	100	42.9	42.7	41.7
1	75.5	100	49.9	47.9
2	77.4	78.9	100	64.8
3	75.2	77.6	88.4	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	20	Q5EXK5	K.oxytoca	mhbT
1	21	E8ZB61	P.putida	galT
2	22	Q51955	P.putida	pcaK
3	23	Q43975	A.baylyi	pcaK

*Clado 3 en siguiente página

CLADO 4.

	0	1	2	3	4
0	100	29.6	29.7	27.6	29.2
1	65.6	100	70.6	53.8	54.1
2	65.9	88.1	100	48.8	53.1
3	63.6	80.4	78.2	100	68.3
4	62.2	80.5	80.2	87.6	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	48	A1Z264	G.sulphuraria	SPT1
1	49	Q8NJ22	K.lactis	frt1
2	50	Q9HFF8	S.pastorianus	FSY1
3	51	Q0ULF7	P.nodorum	SNOG_07407
4	52	A8DCT2	G.moniliformis	FST1

CLADO 3.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
0	100	34.5	32.8	26.6	27.1	26.3	27.3	20.7	21.5	24.1	22.7	24.1	23	22.9	22.2	21.8	23.1	23.6	20.5	23.5	22.3	22	21.9
1	72.3	100	56.5	35.8	35.8	35.4	33.5	26.8	26.3	23.9	24.4	25.2	24.8	27.5	28.6	29.2	27.3	28	28	26.6	26.2	28	27.8
2	71.3	82.5	100	30	31.4	32	29.8	24.1	23.7	20	22.9	23.4	22.6	22.8	24.1	24.4	22.6	22.4	23.7	23.6	23.2	23.3	22.6
3	63.6	67	66.5	100	44.4	44	45.1	25	24.8	27.8	22.9	25.2	25.9	22.8	25.5	26	23.4	24.7	24.9	23.1	24.9	25.5	23.8
4	66.8	65	65.8	73.5	100	73.3	70.1	21.7	25.4	29.6	24.7	23.3	21.8	24.6	24.4	25.6	21.8	24.2	25.7	23.5	25.3	25.4	25.8
5	63.9	65.5	65.1	73.3	90.2	100	73.9	22.9	23.8	26.3	23.3	22.7	21.6	25.2	26.3	25.5	24.2	24.7	24	24.4	24.1	23.6	25.8
6	64.2	64.8	65.3	72.5	88.1	92.8	100	21.9	22.8	26	23.8	24.8	23.3	24.3	25.8	23.4	24.5	25	25	24	24.1	24.8	25.3
7	52.2	53	51.5	52.9	53.5	51.5	51.6	100	55.1	34.4	36.7	37.1	35.9	36.5	37.3	34.9	34.5	34.7	36.2	35.6	35.5	35.4	34.8
8	53.4	53.6	54	53.7	52.4	53.4	51.5	78.8	100	34.4	35.4	35.1	35.6	40.2	35.5	35.6	33.9	33.5	38.3	39.1	36.3	38.2	39.3
9	52.8	53.8	51.7	56	60.1	60.2	59.2	62.9	63.5	100	55.9	55.3	55.1	33.5	31.7	33	29.5	30	32.8	31.4	31.8	32.1	31.7
10	51.8	54.4	52.9	53.2	55.7	52.2	52.2	65.8	64.6	79.6	100	72.7	70	36	33.4	32.5	31.6	33.5	36.5	34.6	35.3	35.2	35.2
11	51.2	52.7	51.2	53.7	53.1	52.2	51.9	63.7	64	80.3	87.6	100	84	35.5	33.9	34.5	30.7	29.9	36	35.9	36	35	35
12	52.1	54.4	51.9	52.4	54.8	52.7	51.3	65	65.2	78.8	86.6	94.8	100	34.7	33.5	33.2	30.2	29.9	34.9	35.6	35.3	34.9	34.5
13	55.1	56.5	56	54.7	54.1	54	53.8	66.2	69.2	60	63.4	63.6	63	100	56.9	57.6	48.6	49.6	66.9	67.2	68.5	67.4	67.8
14	52.9	56.2	55.6	57.4	54.8	55	57.4	65.1	66.5	59.1	62.9	63.8	63.4	81	100	52.7	48.4	48.1	57	58.3	58.1	58.2	58.3
15	55.2	55.4	57	57.1	55.4	55	54.4	64.9	66.7	60.1	63.5	61.2	62.5	82.1	78.4	100	46.5	46.2	58.1	57.5	58.4	60.8	61.5
16	51.4	52.7	53.9	53	54.1	53.3	52.8	64.7	66.1	57.5	59.4	60.1	58.3	76.8	73.8	75.2	100	79.2	48.4	48.1	48.6	48.3	49.4
17	51.8	54.7	54.9	53.2	55	54	53.8	62.9	65.9	58.6	60.8	59	58.7	73.5	73.7	72.7	91.1	100	50.3	48.4	48	51.2	52
18	53.7	53.9	52.8	54.8	53.8	52.2	54.5	66.5	68.5	59.9	61.6	62.3	61.6	85.4	78.1	80.9	76.7	75.8	100	73.4	75.3	72.2	73.1
19	53.2	56.5	55.6	53	54	54.1	53.3	64	67.9	58	60.2	62.8	62.5	88.1	80.1	80.5	76.6	76.1	90.9	100	72.5	73.7	75
20	54.1	53.2	55.2	54.6	53.4	53.6	53.8	65	68.6	57.1	60.7	60.6	60.8	88.3	81.5	82.3	77.6	77.1	90.8	90.6	100	79.2	78.2
21	52.1	54.7	54.2	55.8	53.4	52.8	52.4	66.3	67.6	59.6	61.8	61.2	62.1	86.6	80.7	81.9	77.7	76.9	89.1	91.5	92.1	100	94.2
22	52	54.5	54.7	56.2	54.5	53.5	53.5	65.6	68.8	60.4	60.8	61.4	62.5	87.4	80.9	81.4	77.3	76.5	89.7	90.5	91.4	97.9	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	25	B8NU03	A.flavus	lnaF
1	26	Q59Q30	C.albicans	GIT1
2	27	P25346	S.cerevisiae	GIT1
3	28	O94342	S.pombe	SPBC1271
4	29	A0A1D8PN14	C.albicans	GIT4
5	30	Q5A1L6	C.albicans	GIT3
6	31	A0A1D8PN12	C.albicans	GIT2
7	32	P25297	S.cerevisiae	PHO84
8	33	Q7RVX9	N.crassa	pho
9	34	Q9Y7Q9	S.pombe	SPCC2H8
10	35	Q09852	S.pombe	SPAC23D3
11	36	O42885	S.pombe	SPBC8E4
12	37	Q9P6J9	S.pombe	SPBC1683

Mat.	Filo.	UNIPROTID	Organismo	Proteína
13	38	Q9ZWT3	A.thaliana	PHT1
14	39	Q7XRH8	O.sativa	PHT1
15	40	Q94DB8	O.sativa	PHT1
16	41	Q9S735	A.thaliana	PHT1
17	42	Q9SYQ1	A.thaliana	PHT1
18	43	Q8H6G9	O.sativa	PHT1
19	44	Q8H074	O.sativa	PHT1
20	45	Q494P0	A.thaliana	PHT1
21	46	O48639	A.thaliana	PHT1
22	47	Q8VYM2	A.thaliana	PHT1

CLADO 5.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	100	31.5	33.7	34.4	34.2	33.6	33	33.7	33	33.9	32.9	33.3	33.3
1	64.1	100	34.7	35.1	34.5	38.3	36.7	39.6	37.7	40.2	40.2	37.9	37.9
2	63.6	68.8	100	93.1	95.7	63.4	62.5	62.2	62	63.2	63.2	63	62.8
3	64.2	68.9	98.3	100	95.5	62.9	62.1	61.8	61.7	63.3	61.9	62.3	62.1
4	63.8	68.9	98.9	98.7	100	63	62.5	62.9	62.1	63.2	62.4	62.6	62.5
5	67	72.6	87	85.6	86.7	100	76.9	79.3	78.3	77.3	75.5	75.8	75.8
6	65.8	71.7	85.5	84.1	85.7	93.5	100	83.5	83.8	78.5	79.7	80.5	80.3
7	65.9	73.2	86.9	85.1	86.4	93.5	96.5	100	84.9	81.5	81.8	82.4	82.5
8	65.4	71.4	85.7	85.1	86.2	93.4	96.3	96.5	100	81.1	82.4	82.4	82.2
9	66.1	72.6	87.9	87.3	88.5	92.9	95	93.7	93.8	100	89.7	90.7	90.7
10	65.9	72.9	88.3	86.9	88.2	93.4	96.2	94.9	95.3	97.4	100	95.9	96.1
11	65.8	70.6	87.9	86.7	88.1	93.1	96.2	94.7	95.1	97.2	99.8	100	99.8
12	65.9	70.6	87.9	86.7	88.1	93.3	96.2	94.7	95.1	97.2	99.8	100	100

	Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	53	Q64L87	D.fabryi	nan
1	54	A2R3H2	A.niger	An14g04280
2	55	P11636	N.crassa	qa
3	56	Q4U3U6	N.africana	qa
4	57	Q4U3U4	N.terricola	qa
5	58	P15325	E.nidulans	qutD
6	59	A2QQV6	A.niger	qutD
7	60	Q0D135	A.terreus	qutD
8	61	Q2U2Y9	A.oryzae	qutD
9	62	A1CPX0	A.clavatus	qutD
10	63	A1D2R3	N.fischeri	qutD
11	64	Q6MYX6	N.fumigata	qutD
12	65	B0XQS8	N.fumigata	qutD

CLADO 6.

	0	1	2	3	4
0	100	36.8	35.3	38.8	38.1
1	69.5	100	52.9	60.2	57.4
2	66.6	81.2	100	57.3	58.6
3	68.5	86.4	82.3	100	69.5
4	68	85.6	83.1	88.7	100

	Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	76	C4QVV9	K.phaffii	PAS_chr1
1	77	P39932	S.cerevisiae	STL1
2	78	Q4WC50	N.fumigata	mfsA
3	79	Q5A8J5	C.albicans	HGT10
4	80	A0A0K1NY69	W.anomalus	STL1

CLADO 7.

	0	1	2	3	4	5	6	7	8
0	100	43.9	43.5	43.5	45.1	44.2	44	44.2	44.2
1	76.3	100	57.8	58	55	55	54.9	55	55
2	75.1	82.5	100	98.7	75.5	75.4	75.2	75.4	75.4
3	74.7	82.4	99.7	100	75.7	75.5	75.4	75.5	75.5
4	75.4	81.4	90.8	90.9	100	99.8	98.3	98.6	98.5
5	74.9	81.4	91.2	91.4	99.8	100	99.5	99.5	99.7
6	74.7	81.3	91.1	91.2	98.8	99.5	100	99.7	99.8
7	74.7	81.4	91.4	91.6	99	99.7	99.8	100	99.8
8	74.9	81.4	91.2	91.4	99	99.7	99.8	100	100

	Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	82	Q32SL4	P.angusta	MAL2
1	83	P53048	S.cerevisiae	MAL11
2	84	P15685	S.cerevisiae	MAL61
3	85	P38156	S.cerevisiae	MAL31
4	86	P0CD99	S.cerevisiae	MPH2
5	87	A6ZX88	S.cerevisiae	MPH3
6	88	C8Z6M6	S.cerevisiae	MPH3
7	89	P0CE00	S.cerevisiae	MPH3
8	90	B5VF36	S.cerevisiae	MPH3

CLADO 8.

	0	1	2	3	4
0	100	66.1	48.7	46.3	45.6
1	88.3	100	50.6	49.5	49.6
2	78.6	77.8	100	53.2	54.6
3	78.5	79.6	80.5	100	59.8
4	78.7	80.5	83.9	81.4	100

	Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	95	Q2MEV7	C.intermedia	GXS1
1	96	A3M0N3	S.stipitis	RGT2
2	97	B1PM37	P.angusta	HXS1
3	98	P10870	S.cerevisiae	SNF3
4	99	Q12300	S.cerevisiae	RGT2

CLADO 9.

	0	1	2	3	4	5	6
0	100	68.4	54.6	56.7	69.7	69.5	71.4
1	88.6	100	58.2	58.9	72.6	72	72.1
2	81.5	85.7	100	85.1	57.6	58.8	56.7
3	82.2	85.3	94.7	100	59.3	61.5	58.6
4	87.5	90.7	82.7	80.6	100	78.3	78.1
5	86.7	89.7	83.5	83.9	93.4	100	91.1
6	88.4	89.8	82.9	82.9	91.4	97.4	100

	Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	102	O74849	S.pombe	ght6
1	103	O74969	S.pombe	ght2
2	104	Q92339	S.pombe	ght3
3	105	O59932	S.pombe	ght4
4	106	Q9P3U6	S.pombe	ght1
5	107	P78831	S.pombe	ght5
6	108	Q9P3U7	S.pombe	ght8

CLADO 10.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	100	38.8	38.8	40.1	38.6	35.9	37.9	35.2	38	38.1	38.8	38.8	37.6	38.2	37.8	39.2	38	37.3	35	35.5	35.3	35.3
1	71.9	100	60.6	59.7	61.3	63.2	57.7	57.1	62.6	60	61.1	62.7	61.2	61.2	61	58.1	62.7	56.7	53	53.2	51.5	51.7
2	73.2	84.7	100	81.8	70.2	69.7	66.3	69.2	64.8	65.2	69	69	68.5	68.8	69	65.3	73.3	67.5	54.6	54.8	55.8	55.6
3	74	84.8	93.7	100	67.1	66.9	69.1	63.7	65.2	65.8	68.7	68.7	66.4	66.7	66.9	65.1	69.1	64.5	57	57.2	57.7	58.4
4	73.8	86	91.3	87.8	100	73.5	73.4	73	64.9	65.2	76	76	69.2	69.4	69.2	67.2	68.4	69.8	54.2	54.4	54.5	54.5
5	70.6	86.6	87.4	87.3	91.6	100	72.6	72.5	65.7	66.7	70.4	70.4	68.7	69.1	69.1	67	75.1	67.5	57.6	57.8	53	53.2
6	72.5	84.3	86.4	89.5	88.9	91.2	100	86.1	66.9	65.7	75.8	75.8	75.3	75.6	75.4	66.7	70.8	67.6	55.1	55.3	54.3	54.6
7	71.8	83	88.8	85.1	86.8	90.5	95.8	100	66	66	73.8	73.8	72.8	73.2	73	65.1	69.1	66.2	54.5	54.7	53.4	56.2
8	69.7	86.7	87.9	86.1	86.6	88.1	86.9	87.5	100	97.9	71.5	71.5	69.7	70.3	70.1	64.2	67.4	65.5	52.7	52.9	53	52.9
9	70.9	84.6	88.1	85.9	86.7	88.1	86.9	87.4	99.6	100	71.3	71.3	69.7	70.3	70.1	65	67.8	65.5	53.9	54	54.5	54.7
10	69.8	85.5	88.7	87.6	91.7	90	89.4	89.7	90.7	90.9	100	99.8	82.3	82.6	82.5	70.5	75.1	73.3	58	58.2	61.7	61.9
11	69.5	86.7	88.7	87.6	91.7	90	89.4	89.7	90.7	90.9	100	100	82.3	82.6	82.5	70.5	75.1	73.3	57.6	57.8	61.7	61.9
12	70.3	84.5	87.3	86.1	87.4	90.1	89.8	87.6	90.1	89.9	93.8	93.8	100	99.3	99.1	68.6	73.7	72.5	54.7	54.9	53.6	54.7
13	69.8	84.5	87.3	86.1	87.6	90.1	89.6	87.5	90.1	89.9	93.6	93.6	99.8	100	99.8	68.9	74.1	72.8	55.1	55.3	53.9	55.1
14	70.3	84.5	87.3	86.1	87.6	89.9	89.6	87.5	90.1	89.9	93.6	93.6	99.8	100	100	68.7	73.9	72.7	55.3	55.5	54.1	55.3
15	70.7	83.6	87.1	86.7	87.5	89	87.9	88.1	86.7	87.7	90.4	90.4	88.5	88.5	88.5	100	70.4	67.6	54.9	55.1	54.2	55.5
16	71.5	86.1	88.3	88.4	88.4	90	90	88.8	88	88.5	91.5	91.5	90.1	90.1	90.1	88.9	100	74.1	59.3	59.5	62.5	60.5
17	68.6	84.8	86.8	85	89.5	88.2	88.6	87.2	86.9	86.3	90.4	90.4	91.4	91.4	91.4	88.2	91.1	100	57.4	57.8	59	60.1
18	70.3	83.3	80.8	83.6	83.9	84.1	81.9	81.7	81.2	80.9	81.7	81.1	80.8	80.6	80.6	79.9	81.5	80.4	100	99.6	90.9	90.8
19	70.5	83.3	80.8	83.6	83.9	84.1	81.9	81.7	81.2	80.9	81.7	81.1	80.8	80.6	80.6	79.9	81.5	80.6	99.8	100	91.2	91.2
20	69.4	82.8	81.1	83.4	82.9	79.6	81.1	81	80.6	80.9	83.9	83.9	78.7	78.5	78.5	80.3	83.3	81.3	96.4	96.6	100	98
21	69.4	82.8	80.6	83.9	83.1	79.6	81.2	81.5	80.7	81.1	84.1	84.1	80.4	80.2	80.2	81.3	81.2	81.8	96.6	96.8	98.9	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	109	P42833	S.cerevisiae	HXT14
1	110	Q2MDH1	C.intermedia	gxf1
2	111	P23585	S.cerevisiae	HXT2
3	112	P43581	S.cerevisiae	HXT10
4	113	P38695	S.cerevisiae	HXT5
5	114	P18631	K.lactis	RAG1
6	115	P32466	S.cerevisiae	HXT3
7	116	P32465	S.cerevisiae	HXT1
8	117	P40885	S.cerevisiae	HXT9
9	118	P54862	S.cerevisiae	HXT11
10	119	P39003	S.cerevisiae	HXT6

Mat.	Filo.	UNIPROTID	Organismo	Proteína
11	120	P39004	S.cerevisiae	HXT7
12	121	C7GWV6	S.cerevisiae	HXT4
13	122	P32467	S.cerevisiae	HXT4
14	123	A6ZT02	S.cerevisiae	HXT4
15	124	P40886	S.cerevisiae	HXT8
16	125	P53387	K.lactis	KHT2
17	126	P13181	S.cerevisiae	GAL2
18	127	P47185	S.cerevisiae	HXT16
19	128	P54854	S.cerevisiae	HXT15
20	129	P53631	S.cerevisiae	HXT17
21	130	P39924	S.cerevisiae	HXT13

CLADO 11.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0	100	56.8	56.8	48.8	55.1	52.6	46.2	49.7	51	49.7	51.3	48.9	46.5	48	43	48.1	41.5	42.6	47.3	50.1	50.7	50.7	52.1	52	50.4	51.6	49.5	49	51.6	50	50.1
1	84.5	100	82.7	48.5	48.4	49.6	45.9	44.6	47.5	45.2	51.3	48.5	46.5	47.5	41.4	46.8	43.1	41.8	45.1	48.1	48.9	50.2	48	46.6	48.2	50	47.1	46.4	47.6	46.3	45
2	83.8	96.6	100	50.3	48.2	48.5	45.7	44.7	48	46.4	47.6	46.8	45.1	46.3	41.9	46.4	42.1	42.9	44.8	47.1	48.5	49.9	47.6	45.9	47.4	48.3	46	46.3	45.7	45.7	45.8
3	80.1	78.9	79	100	82.9	81.6	57.9	54.4	60.2	61.8	60.2	56.8	56.9	55.2	50.6	54.2	47.7	50.7	53.6	55.8	55	56.4	57.7	56.5	57.8	60.7	59.3	55.4	56.2	58.6	56.1
4	82.5	77.9	78.6	95	100	82.5	57	52	59.3	59.2	61	58.8	56.1	56.5	50.7	55.5	46.5	49.7	54.8	57.3	55.8	55.7	58.8	58.1	58.3	59.9	59.8	55.3	56.5	59.5	58.4
5	83.3	79.8	78.6	94.3	94.3	100	58.9	54.4	62.1	62.5	60.3	57.1	56.4	54.9	49.6	55.4	46.7	48.6	53	56	55.4	55.4	58.1	55.6	57.6	58.2	57.4	54	54.1	57.1	57
6	78.6	79.7	79.4	85.3	83.6	84.3	100	61.9	61.4	61	53	51.9	45.8	48.1	46.3	48	41.9	45.2	50.8	49.9	53.1	51.8	50.9	51.7	51.8	52.7	51.8	49.2	50.5	53.2	52.8
7	78.5	75.9	75.1	82.5	81.7	80.8	85.6	100	57.1	57.3	50.6	50.9	46.5	46.9	47.6	51.2	40.1	43.8	48.1	53.5	51.5	51.9	50.6	51.3	49.9	50.7	52.2	47.2	50.6	52.4	50.9
8	79.7	79.6	79.7	87.1	85.8	85.2	85.9	83.9	100	87.2	52.7	51.7	48.7	48.8	45.3	50.5	43.2	47	52	51.8	53.1	52.5	55.2	52.4	52.6	54.5	53.2	49.6	53.1	53	52.9
9	80.6	79.7	79.4	87.9	85.9	85.2	86.1	84.1	97.8	100	52.6	51.4	48.8	49.9	45.2	48.5	43.8	44.6	50.8	51.3	53.5	52.7	55	51.4	53.3	53.7	52.7	49.6	52.8	52.6	53.8
10	81.5	79.6	77.8	86.3	86.1	85.5	82.2	78.8	83.9	84.7	100	80	60	59.7	50.6	57.1	49	50.2	54.8	55.2	55.1	56	55.8	56.3	56.2	58.6	57.6	54.6	57.8	56.4	55.5
11	78.5	78.2	77.5	84.9	83.7	82.8	79.2	77	81.9	82.2	93.2	100	57.8	59.4	49.8	56.1	46.8	50.1	55.3	54.4	54.4	52.9	55.6	56.9	55.7	56.8	55.8	54.7	56.8	56.8	56.7
12	79.4	79.8	78.2	84.6	84	84	78	78.6	81.6	82.1	86.1	86.2	100	76.8	46.3	51.6	41.5	44.5	49.5	53.6	51.7	54.2	53	52.7	52.2	50.9	51.8	45	53.4	52.2	51.5
13	78.3	78.5	77.4	85.6	83.9	83.5	78.2	78.9	80.6	81.2	87.9	86.8	94.8	100	48.4	55.3	42.1	47.1	51.2	52.9	54	53.3	53.4	54.7	54	55.2	53.9	47.1	52.8	53.7	53.7
14	78.1	76.5	78.2	81.2	81.4	81.1	79.1	77.4	79.8	78.8	82.4	82.2	79.3	80.8	100	65.2	46.8	52.4	53.2	49.2	49	51.4	53.8	53	55	56.8	56.7	52.6	52.3	54.1	53.7
15	80.3	76.8	77.8	85	84.5	83.9	80.2	78.5	81	80.5	85.6	84.6	82.6	83.7	88.9	100	48.2	53.1	58.4	58	57.1	55.8	58.2	56.3	58.6	62.2	64.1	56.5	59.6	58.7	57.4
16	75.5	75.5	74.8	80.6	80.2	77.7	75.4	73.7	78.1	78.9	79.5	77	75.7	73.9	80.4	80.3	100	53.3	47.9	46.6	48.3	49	49.4	48.7	49.6	50.3	51.1	47.2	48.4	47.9	49
17	75.3	76.4	75.7	80.1	80.2	79.4	76.8	74.9	77	78.1	79.4	78.5	76	76.1	81.6	82	83.8	100	51.1	49.9	49.5	51	52.8	52.5	53.1	54.1	53.1	51	51.7	54.1	55.1
18	79.6	77.8	77.6	84.6	85.6	82.5	80.2	80.4	82.3	83.2	83.3	82.8	83	81.9	82	85.1	78.7	80.4	100	64.7	65.7	64.7	61.7	63.6	61.9	65.3	64.9	58.1	61.2	59.3	59.5
19	80	77.5	77.3	83.1	82.9	81.9	80.3	80	81.5	80.5	82.6	82.3	82.2	80.3	82.3	84.8	77.7	79.2	88.6	100	73.2	73.1	62.8	62.4	60.4	64.4	64.5	58.2	61.9	60.8	61.2
20	80.7	80.4	80	85.7	85.5	84.3	82	80	83.8	83	85.3	84.1	83	84.1	81.4	83.9	79	81.7	89.2	91.3	100	85.2	63	63.1	61	65.7	63.9	58.8	61.3	60.4	61.1
21	81.3	80.4	80.3	84.4	84.3	84.2	80.4	80.4	83.3	83.9	83.9	82.3	82.2	83	82	83.2	79.1	80.8	87.1	91.5	98.2	100	61.3	62.5	61.3	67.5	63.7	59.3	62.7	60	60.4
22	82.9	79.1	79.6	84.5	84.5	84.5	82.2	79.8	83.7	84.5	83.3	81.2	82.8	80.2	81.9	83.4	78.1	79.4	84.9	84.8	85.4	84.3	100	80	81.9	69	69.2	61.5	64.7	66.9	67.1
23	80.5	80.5	80	86.6	83.9	82.9	81.4	81.1	81.7	82.3	83.7	82.6	84	82.5	83.1	83.8	80.4	80.4	86	84.6	86.1	86.8	94.2	100	80.2	67	67.3	59.2	67.6	65.1	64.6
24	79.9	80.2	78.3	85.3	84.6	84.5	82.8	80.6	83.4	84.2	83.4	81.9	83.7	81.3	82.3	84	78.8	79.8	85.4	86	86	86.1	94.4	94	100	70.5	68.5	59.6	66.5	65.6	66.1
25	79.6	78.9	77.9	87.4	86.6	85.3	82.1	79.6	82.6	82.8	85.7	85.5	82.8	83.2	84.6	87.3	78.5	81	86.8	84.6	86.6	86.3	86.4	84.5	86.2	100	82.8	67.3	72.3	69.5	68
26	78.2	78	78.4	84.7	86	83.1	81.6	79.4	81.9	82	83.4	84.3	82.7	82.7	85	86.5	78.6	79.2	86.7	85.4	86.5	85	86.4	84.8	85	94.3	100	68.5	71.1	70	70.1
27	79.1	77.4	78.2	83.5	82.9	82.4	79.4	77.5	79.5	79.6	82.9	82.9	78.7	78.5	81.5	83.1	78.4	80.2	84.5	83.6	82.3	82.4	84.9	82.1	83.7	88.5	89.6	100	64.7	64	62.5
28	81.2	80	80	84.8	86	84.8	80.2	78.4	82.8	84.1	85.1	82.9	82.8	81.1	81.5	85.2	76.4	80.5	86.4	85.1	84.4	84.3	85.1	85.3	85	88.6	88.8	86	100	69.7	69.1
29	80.3	79.6	79.4	86.4	85.2	84.1	83.6	81.1	82.5	83.1	84.2	84.1	83.7	83.6	81.3	85.3	78.1	80.3	85	84.3	86.2	84	86.6	87.2	86.4	89	88.1	84.6	88.8	100	89.3
30	80.5	79.3	79.2	86	86	85.1	82.6	81.1	82.4	83	83.5	84.6	83.9	83.6	81.3	85.4	78.4	81.2	86.2	85.5	87	85.4	86.7	86.7	87.1	88	87.8	83.7	89.5	96.8	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	131	Q39524	P.kessleri	HUP2
1	132	Q39525	P.kessleri	HUP3
2	133	P15686	P.kessleri	HUP1
3	134	Q10PW9	O.sativa	MST4
4	135	Q94AZ2	A.thaliana	STP13
5	136	A0A2C9W8R5	M.esculenta	MANES_03G180400
6	137	A0A2C9WLH2	M.esculenta	MANES_01G164600
7	138	Q9LNV3	A.thaliana	STP2
8	139	Q9SBA7	A.thaliana	STP8
9	140	Q9SFG0	A.thaliana	STP6
10	141	O04249	A.thaliana	STP7
11	142	Q10710	R.communis	STA
12	143	Q8GW61	A.thaliana	STP14
13	144	A0A2C9UCP7	M.esculenta	MANES_15G030700
14	145	Q8L7R8	A.thaliana	STP3
15	146	Q07423	R.communis	HEX6

Mat.	Filo.	UNIPROTID	Organismo	Proteína
16	147	Q0JCR9	O.sativa	MST1
17	148	Q93Y91	A.thaliana	STP5
18	149	Q39228	A.thaliana	STP4
19	150	Q9FMX3	A.thaliana	STP11
20	151	Q9LT15	A.thaliana	STP10
21	152	Q9SX48	A.thaliana	STP9
22	153	Q41144	R.communis	STC
23	154	O65413	A.thaliana	STP12
24	155	P23586	A.thaliana	STP1
25	156	Q7EZD7	O.sativa	MST3
26	157	Q6Z401	O.sativa	MST6
27	158	Q851G4	O.sativa	MST2
28	159	Q6ZKF0	O.sativa	MST5
29	160	Q94EC4	O.sativa	MST8
30	161	Q94EC3	O.sativa	MST7

CLADOS 12 Y 13.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	100	61.4	61.3	30.5	27.1	28.8	29.9	31.5	29.7	31.1	29.1	29	29.6	29.8
1	86.1	100	80	28.7	27.7	28.1	27.1	29.7	27.8	30.4	29.5	29.5	30.7	31
2	85.1	94.4	100	27.9	26.8	28.5	25.5	28	26.8	27.2	27.2	27.7	28.4	28.4
3	58.1	58.4	59.2	100	48	48.6	44.7	44.8	33.7	37.5	39.5	38.9	39.5	39.7
4	54.4	55.8	56.8	75.3	100	85.5	55.7	53	35.2	41.3	43.7	42.8	44.4	44.8
5	56	57.1	56.6	75.1	95.3	100	55.3	54.3	36.4	40.2	43.5	43	43.2	44.5
6	54.1	56.8	54.2	70.8	79.9	78.5	100	77.4	38.7	40.1	39.6	40.6	41.1	41.1
7	55	56.2	55.4	72.7	79.8	78.9	88.9	100	37.2	39.5	41.4	40.1	41.2	41.9
8	55.1	55.5	55.3	65.4	70.7	68.6	66.5	68.2	100	45.6	46.4	43.8	45.8	46.2
9	58.2	58.4	57.5	69.4	73.8	72.4	68.1	67	72.9	100	60.9	59.8	61.1	61.2
10	56.5	59	58	69.8	72.7	72	66.5	69.5	73.2	84.2	100	82.4	88.1	88.2
11	56.6	59.9	58.7	68.7	72.5	71.4	68.8	70.9	73	82.2	92.6	100	83.6	83.9
12	56.7	59.4	59.4	70.8	72.8	70.2	68.9	71.3	71.8	84.1	95	93.4	100	97.6
13	56.9	59.3	59.4	71	72.7	72.8	68.8	71.2	72	83.5	95.5	93.6	99.2	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	185	Q0WWW9	A.thaliana	At5g59250
1	186	Q6AWX0	A.thaliana	At5g17010
2	187	Q8L6Z8	A.thaliana	At3g03090
3	188	F1R0H0	D.rerio	slc2a10
4	189	Q0P4G6	X.tropicalis	slc2a10
5	190	Q6GN01	X.laevis	slc2a10
6	191	O95528	H.sapiens	SLC2A10
7	192	Q8VHD6	M.musculus	Slc2a10

Mat.	Filo.	UNIPROTID	Organismo	Proteína
8	193	Q6NWF1	D.rerio	slc2a12
9	194	Q32NG5	X.laevis	slc2a12
10	195	Q5J316	B.taurus	SLC2A12
11	196	Q8BFW9	M.musculus	Slc2a12
12	197	Q8TD20	H.sapiens	SLC2A12
13	198	Q9BE72	M.fascicularis	SLC2A12

CLADOS 14,15 Y 16.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	100	61.1	64.8	36.3	34.8	34.4	38.5	33.9	34.4	32.2	33.1	37.4	35.4	32.5	32.4	29.6	25.8	31.4	33.8	39.6	33.7	36.6	34.2	36.2	36.9	36.9
1	82.2	100	54.5	36.2	35.4	33.5	32	31.1	34.9	30.2	29.9	33.1	36.4	31.2	31.3	29.5	28.2	29.5	31.4	40.5	39.3	35.9	38.2	34.9	34.1	34.1
2	84.9	78.6	100	36.7	34.4	34.8	36.9	32	35	32.9	33.3	37.3	38.1	30.8	29	27.6	29.2	31	32.8	35.5	38.7	34.8	36.1	33.8	35	35
3	74.5	73.9	70	100	66.2	47.9	46	47.2	46.6	48.6	47.9	48.2	49.3	29	27.4	29	30.7	30.1	32.3	35.7	31.6	28.7	28.2	30.3	29.9	30.9
4	74.4	71.7	70.3	88.8	100	50.2	48.9	50.1	48.9	48.1	48.3	50.1	50.2	26.4	29.4	30.9	30.6	28.9	28.7	34.8	32.4	30.9	28.9	29.9	29.9	30.1
5	71.7	72.6	70.5	80.8	83	100	70	56.3	58.8	58.1	58.1	59	57.5	25.9	29.4	31.4	30.1	30.9	30.3	32.2	34.5	29.4	28.9	30.2	29.3	29.5
6	72.6	71.2	70.4	78.8	80.2	90	100	55.9	58	58.4	58.2	59.2	57.2	28.1	32.2	30.6	28.7	29	32.3	31.9	31	30.4	27.6	32.4	30.9	31.4
7	70.2	70.9	69.3	78.5	79.9	84.2	85.4	100	78	62.4	63.5	71	73.4	25.8	30	30.3	29.5	28.6	28.9	33.9	29.5	27.2	28	31	29.1	29.3
8	72.9	72.3	70	79.2	80	84.1	87.3	93	100	63.5	64.2	70.1	73.1	26	29.4	30.3	30.9	30.7	29.5	32.9	29.4	27.9	28.7	30.1	29.6	29.4
9	71.1	71.8	69.2	78	77.7	85.4	85.4	87.1	87.3	100	93.5	68.6	69.1	26.5	29.1	30.5	28.7	28.7	28.8	33.1	27.8	27.5	27.9	29.8	28.1	28.3
10	70.2	71.3	70	76.1	78.5	85.4	85	87.1	87	98.8	100	67.4	69.5	27	28.9	28.9	28.5	28.1	29.2	32.5	27.4	28.6	28	29.6	28.1	28.3
11	75.7	69	72.7	79.3	80.5	85.8	87	88.7	89	89.2	88.1	100	77	25.6	29.5	30.4	28.8	29.2	28.9	34.2	29.3	28.3	26.5	30.7	28.9	29
12	73.4	75.1	72.9	79.7	81.4	84.3	86.8	90.5	91.4	90.6	90	91.6	100	25.7	30.2	31	29.4	28.6	29.6	33.5	27.8	27.5	27.5	30	29.9	29.7
13	69.9	70.2	67	58.3	58.9	59.5	58.6	58.5	59.2	59.3	60.1	58	58.3	100	32	31.3	32.3	32.3	26.7	30.5	27.8	28.8	29.1	28.2	27.1	27.1
14	69.3	66.5	69.2	63	65.4	61.1	63.3	63.8	63.8	64.2	64	64.1	65.3	62.3	100	52.2	47.7	46.8	32.6	35.5	31.6	33.8	32.4	31.3	31.9	31.2
15	70.8	72.2	72	64.4	66.5	64.7	64.1	66.6	65.3	66.9	66.4	65.9	65.7	63.6	83.3	100	40.5	41.9	35	34.6	31.3	34.3	31.5	30.5	29.4	29.2
16	57.5	68.5	61	65.4	64	61.8	65.4	65.1	65.4	64.5	64.5	64.2	66.5	62.6	78.3	73.8	100	80.4	33.3	34.4	30.2	30.9	30.2	27.8	26.8	26.2
17	72.5	72	71.9	64.9	65.1	62.2	66.6	65.7	64.6	63.6	64	65.5	64.1	60.6	79.4	75.1	92.6	100	33	33.4	31.1	30.7	31.6	29.3	29.9	29.4
18	71.1	71	71.6	64.6	64.7	64.5	64	63.5	63.6	61.4	62.7	63.3	61.6	59.7	63.9	66.5	65.2	67	100	36.8	33.2	31.9	30.7	32.9	32.8	32.4
19	76.2	74.6	72.4	69.9	70.8	64.8	67.7	67.2	65	67.1	66.9	68.5	67.3	60.5	69.4	69.8	68.6	69.8	72.7	100	50	46.8	46.1	36.2	36.1	36.1
20	65.3	76.7	73.4	61.8	62	60.6	59.7	59.2	59	59.9	59.5	59.8	59.3	59.9	59.5	60.2	61.4	62	62.6	72.8	100	60.6	59.2	39.1	38.6	38.6
21	74.9	73.8	72.5	59.3	59.9	57.6	58.5	58.8	57.6	59.2	59.2	59.9	59.3	59	59.6	62	60.2	61	60.9	70.1	82.5	100	79.7	38.5	38.4	38.3
22	66.1	73.5	70.8	58.4	59.3	56.9	57.8	60	59	58.8	58.3	58.4	58.2	59.7	61	61.8	61.6	62	62.5	70.6	84	93.1	100	38.1	38.2	37.9
23	74.2	74	74.3	60.7	60.8	60.1	60	60.1	58.9	59.2	59.2	59.9	58.2	54.5	59.8	64.2	61.5	62.7	60.7	64.8	69.3	67.8	67.9	100	88.4	89.5
24	71.4	72.4	71.5	59.3	60.9	59.7	59.2	58.8	58.5	58.7	58.5	59.2	58.4	53.5	59.6	63.1	62	62.5	61.6	65.7	68.7	67.7	67.4	96	100	97.8
25	71.4	72.4	71.5	59.5	60.4	60.1	59.6	58.6	58.9	58.9	58.7	59.4	58.7	53.2	58.7	63.3	60.8	62.4	61.6	65.2	68.7	67.1	67.5	96.6	99.4	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	199	Q8LPQ8	A.thaliana	MSSP2
1	200	Q9SD00	A.thaliana	MSSP3
2	201	Q96290	A.thaliana	MSSP1
3	202	Q0WUU6	A.thaliana	PLT4
4	203	I1M280	G.max	100815073
5	204	Q9ZNS0	A.thaliana	PLT3
6	205	Q8GXR2	A.thaliana	PLT6
7	206	A0A097P980	P.salicina	SOT1
8	207	A0A075BFV8	P.pyrifolia	SOT2
9	208	Q9XIH6	A.thaliana	PLT2
10	209	Q9XIH7	A.thaliana	PLT1
11	210	Q8VZ80	A.thaliana	PLT5
12	211	Q1XF07	L.japonicus	plt4

Mat.	Filo.	UNIPROTID	Organismo	Proteína
13	212	Q04162	S.cerevisiae	YDR387C
14	213	Q10286	S.pombe	itr1
15	214	P87110	S.pombe	itr2
16	215	P30605	S.cerevisiae	ITR1
17	216	P30606	S.cerevisiae	ITR2
18	217	Q01440	L.donovani	nan
19	218	Q8VZR6	A.thaliana	INT1
20	219	Q9C757	A.thaliana	INT2
21	220	O23492	A.thaliana	INT4
22	221	Q9ZQP6	A.thaliana	INT3
23	222	Q96QE2	H.sapiens	SLC2A13
24	223	Q3UHK1	M.musculus	Slc2a13
25	224	Q921A2	R.norvegicus	Slc2a13

CLADO 17.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	100	91.6	39	41.6	40.7	47.5	45.8	42.6	45.2	41.6	41	42	42.2	43.8	40.1	40.5	42.5	41.7
1	98.2	100	40.1	42.7	41.8	47.3	46.5	44.1	43.2	41.7	40.7	40.6	39.6	44.5	41	41.8	43.9	42.9
2	74.5	74.3	100	43.8	44.2	47.8	45.7	41.8	45.1	42.4	42.2	46.1	47.1	42.7	47.5	46.8	42.8	44.5
3	70.8	71.8	78.1	100	81.4	45.6	47.5	41.8	43.7	40.5	40.2	43.8	41.4	41.2	43.1	43.1	41.7	40.8
4	70.3	71.9	78.4	93	100	45.9	44.9	42	42.8	40.2	40.4	44.2	42.1	42.1	44.6	45	45.4	45
5	78.5	78.7	81.6	76.9	77.5	100	53.2	52.3	55.3	48.8	50.1	48.6	48.5	48.5	47.9	47.9	49.2	50.1
6	77.3	78.3	83.4	78.6	78.7	84.1	100	55.6	60.3	45.9	45.2	49.4	46.4	47.2	48.1	49.3	46.3	47.8
7	72.8	73.3	76.7	72.6	72.7	80.3	84.2	100	59.5	41.2	42.7	43	42.7	41.5	44.9	43.8	45.7	45.3
8	77.1	75.4	77.7	76	73.7	82.7	86.5	83.3	100	43.8	43.1	47.7	46.6	46.2	44.3	43.8	43.6	46.4
9	72.5	72.8	77.6	73.4	73.7	80.4	77.9	73.4	74.7	100	81.5	52.5	51.7	50.4	48.7	47.8	50.2	50.3
10	71.9	73	77	72.7	73.7	79.3	77.4	73.5	75.4	94.6	100	52.5	50.8	48.3	48.9	48.8	50.2	50.8
11	75.3	73.8	82.3	74.3	76.9	82	81.1	78.4	80.6	82.8	82	100	65.3	53.6	53.2	53.9	54.4	53.5
12	77.7	74.1	80	75	74.8	80.8	78.4	75.1	77.4	83.8	82.3	89.2	100	51.4	51.7	52.1	53.8	53.3
13	72.6	73	79.1	72.9	73.7	78.6	76.8	72.6	76.1	81.3	82.2	80.3	82.5	100	52	51.4	52.5	53.7
14	74.2	73.9	82.7	76.6	77.2	80.8	80.8	77.5	77.5	82.3	80	83.4	84.3	82.8	100	90.4	68	68
15	73.3	73.1	82.9	76	76.8	81.8	81.7	77.4	78.4	82.3	81	84.3	86.2	82.2	97.6	100	66.2	66.8
16	74.6	75.2	79.7	75.3	76.1	80.4	80.6	76.3	76.5	82.3	80.9	84.1	85.3	82.7	91.7	92	100	82.9
17	73.8	73.8	80.8	74.9	76.2	78.8	80	75	78.1	80.9	79.2	84.3	85	84.1	90.6	90.7	95.6	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	235	Q9FRL3	A.thaliana	At1g75220
1	236	Q93YP9	A.thaliana	At1g19450
2	237	Q9LTP6	A.thaliana	At3g20460
3	238	Q9M0Z9	A.thaliana	At4g04760
4	239	Q8GXK5	A.thaliana	At4g04750
5	240	Q3ECP7	A.thaliana	At1g54730
6	241	P93051	A.thaliana	At2g48020
7	242	Q0WQ63	A.thaliana	At3g05150
8	243	Q8LBI9	A.thaliana	At5g18840
9	244	Q4F7G0	A.thaliana	SUGTL3
10	245	Q9SCW7	A.thaliana	SUGTL4
11	246	Q94KE0	A.thaliana	ESL1
12	247	O04036	A.thaliana	ERD6
13	248	Q8VZT3	A.thaliana	SUGTL5
14	249	Q93Z80	A.thaliana	At3g05160
15	250	Q94AF9	A.thaliana	At3g05165
16	251	Q94CI6	A.thaliana	SFP2
17	252	Q94CI7	A.thaliana	SFP1

CLADOS 18 Y 19.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	100	82	42.9	44.3	43.6	42.7	32.2	27.2	31.6	32.9	33.3	32.7	32.5	32.8	33.8	33.5	33.5	32.8	33.6	33.6	33.8	33.3	33.5	33.5	33.4	33	32.6	32.7	32.9
1	92.3	100	43	44.3	44.9	43.6	31.6	26.1	30.8	33.2	33.3	32.3	32.8	32.5	32.6	32.6	32.6	31.5	31.7	31.7	32.4	31.7	33	32.8	32.6	31.7	33.1	32.3	32.3
2	70.8	71.3	100	94.4	86.2	83.1	34.2	25.3	35.5	36.2	35.1	36.4	35.7	35.7	35.3	36.7	36.5	35.5	36.1	36.1	36.9	35.7	35.9	36.5	36.3	36.3	37.1	37.1	36.7
3	70.5	71.7	98.3	100	85.4	83.7	34.1	26.1	35.7	36.9	36.8	37.7	37.1	36.6	36.8	38.1	37.7	36.6	37	37	38.3	37.7	37.2	38.2	38	37.9	38.3	38.1	37.7
4	70.1	71.6	97.1	97.1	100	88.9	34.8	24.8	35.9	37.2	35.7	37	36.8	36.8	37.4	37.7	37.7	36.6	36.8	36.8	37.9	37.7	37.7	37.4	37.2	37	37.4	36.8	36.7
5	70	73.1	95.8	96	97.1	100	34.5	24.6	37	36.8	36.1	37.9	36.2	37	36.6	37.6	37	36.3	36.4	36.6	37.6	37.4	37.2	37.4	36.9	37.2	38.5	38.3	38.3
6	66.3	66.9	67.5	67.2	67	66	100	28.2	37.4	37.1	37.1	37.5	37.2	38.4	35.7	36.3	36.3	36.7	37	37	37.6	37.6	37.2	36.9	36.9	37.6	39.5	38.7	39.1
7	55.2	56.7	55.9	56.2	55.9	55	61.3	100	27.2	28.3	27.3	26.5	26.3	27.1	26.5	26.7	27	26.4	26.4	26.4	26.6	26.1	26.9	27	27.3	27.3	27	27.6	27.2
8	66.7	65.6	69.4	69	67.6	67.7	69.6	61.7	100	54.9	61	63.1	64.4	61.7	61	60.1	60.3	60.9	61.4	61.4	60.6	61.2	61.6	61.6	61.4	61.4	62.1	62.1	61.9
9	66.5	68.1	69.3	68.8	69.4	68.4	70.1	62.4	84.3	100	59.2	60	61.6	57.8	58.6	58.2	58.4	58.1	58.8	58.8	59.4	57.4	58.2	57.6	57.3	57.6	58.3	59.1	59.6
10	67.6	68.3	67.2	68.8	69.7	68.6	69.2	60.1	88.3	86.3	100	80.4	81.3	78.5	75.2	78	77.6	76.7	77.4	77.4	77.8	74.8	76.4	75.6	76	75.8	72.5	71.7	71.6
11	68.4	68.2	68.3	68.1	67.5	67.3	72.2	59.3	88.7	87.3	96.4	100	85.7	89.7	65.7	69.5	64.9	63.7	63.4	63.6	64.3	63.4	64.6	64.3	64.3	64.7	74.2	73.4	73.1
12	68.6	68.3	67.6	68.5	69.9	67.9	71.4	59.5	88.7	87.1	95.8	92.9	100	92.6	67.6	69.4	64.2	63.4	64.1	64.2	64.7	63.4	64.9	64.1	64.1	64.2	75.5	74.5	74.2
13	68.7	68.8	69	68.6	69.3	68.8	71.9	59.9	87.4	84.9	94.9	97.5	98.1	100	78.7	79.5	80.2	78.1	80	80	80.4	78.7	80.2	79.5	79.5	79.5	75.7	75.5	75.7
14	67.9	68.7	66.8	67.7	68.3	66.7	68.8	58.4	86.4	85.1	92.5	82.9	83.9	92.8	100	86.5	85	82.6	81.6	81.5	82.4	80.7	81.5	80.7	80.9	80.2	80.7	79.8	79.9
15	68	68	67.8	69.5	69	67.4	70.5	59.1	87.2	85.9	93.5	87.1	85.3	93.2	93	100	91.9	86.4	86.4	86.3	85.8	82.8	84	83.1	83.3	82.2	81.3	80	79
16	68.5	68.9	67.8	69.5	69	66.3	70.1	58.6	87.4	86.7	93.8	81.8	80.1	94.1	90.5	93.9	100	89	88.6	88.7	89.3	87.1	88.7	87.1	87.6	86.6	81.3	80	79.2
17	68	68	67.2	68	69.3	67.6	70.8	58.8	87.7	86	93.2	81.5	80.1	93.9	90.1	91.8	94.9	100	88.8	89	89.1	86.7	88.2	87	87.1	86.2	81.8	80.7	79.5
18	67.7	67.8	68.2	69.7	69.3	67.5	69.3	59.4	87.4	85.9	93.5	80.6	79.3	94.1	88.6	91.3	94.6	94.4	100	99.9	91.7	87.6	89.4	88.8	88.8	88.2	82.8	81.5	80.9
19	67.7	67.8	68.2	69.7	69.3	67.8	69.3	59.4	87.4	85.9	93.5	80.7	79.4	94.1	88.5	91.2	94.7	94.5	99.9	100	91.8	87.7	89.6	88.9	88.9	88.3	82.8	81.5	80.9
20	68.3	68.9	67.8	69.5	69.5	68	70.3	59	87.6	86.3	93.7	82.3	80.3	94.3	89.2	90.9	95.6	94.9	96.1	96.2	100	92.5	93.8	92.4	92.7	92.1	84.5	83	82.4
21	67.8	68.5	66.2	67.7	67.5	65.9	70.5	59	86.6	85.5	93.5	81.7	80	93.7	88.7	90.2	95.5	94	95.1	95.2	97.2	100	96.1	95.3	95.7	95.1	86.3	84.3	84.3
22	68	68.7	66.8	68.4	68.4	66.7	71	59	87.6	85.7	93.8	81.9	80.2	94.3	88.7	89.4	95.3	94	95.1	95.2	97.5	99.5	100	97.1	97.4	96.3	89.1	87	86.7
23	67.4	67.8	66.6	68.3	68	66.7	71.3	59.1	87.2	85.5	93.3	81.6	79.9	93.7	88.3	89.6	94.9	93.8	95	95.1	97.1	98.9	99.3	100	99.1	97.5	89.3	88.1	87.9
24	68.3	67.8	66.4	68.1	67.7	66.2	71.1	59.2	87	85.1	93.3	81.5	79.8	93.7	88.2	89.4	94.9	93.8	94.8	94.9	97.2	98.9	99.3	99.5	100	97.7	88.6	87.4	87.3
25	67.2	67.6	66.4	68	67.7	66.3	71.1	59.8	86.4	85.3	92.7	82.2	79.8	93.2	88.1	89.2	94.7	93.5	94.6	94.7	97	98.6	98.9	99.1	99.2	100	90.8	90	90.5
26	66.3	66.7	67.9	67.7	68	67.5	70.3	60	86.8	85.8	91.6	92.7	92.3	92.4	93.8	94.4	95.3	95.3	95.5	95.5	96.1	96.1	97.2	97	97	97.4	100	95.7	95.7
27	64.6	67	67.9	67.7	68.2	66.9	69.3	60.9	86.4	86.9	91.3	92.1	91.7	92	93	93.8	94.7	94.7	94.7	94.7	95.5	95.5	96.6	96.4	96.4	97.2	98.8	100	98.2
28	64.6	67	67.1	66.7	67.8	66.5	70	60.5	86.1	86.5	90.7	91.5	91.3	92	93	93.4	94.3	94.3	94.3	94.3	95.1	94.9	96.2	96	96	96.8	98.6	99.2	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	253	Q9UGQ3	H.sapiens	SLC2A6
1	254	Q3UDF0	M.musculus	Slc2a6
2	255	Q9JIF3	M.musculus	Slc2a8
3	256	Q9JJZ1	R.norvegicus	Slc2a8
4	257	Q9NY64	H.sapiens	SLC2A8
5	258	P58354	B.taurus	SLC2A8
6	259	A5Y0C3	N.lugens	HT1
7	260	P53403	D.melanogaster	Glut3
8	261	A9ZSY2	A.mellifera	Tret1
9	262	A9ZSY3	B.mori	Tret1
10	263	A5LGM7	P.vanderplanki	Tret1
11	264	Q7PIR5	A.gambiae	Tret1
12	265	Q17NV8	A.aegypti	Tret1
13	266	B0WC46	C.quinquefasciatus	Tret1
14	267	B4J913	D.grimshawi	Tret1

Mat.	Filo.	UNIPROTID	Organismo	Proteína
15	268	B4LPX5	D.virilis	Tret1
16	269	B4KR05	D.mojavensis	Tret1
17	270	B4MYA4	D.willistoni	Tret1
18	271	B4GAP7	D.persimilis	Tret1
19	272	Q291H8	D.pseudoobscura	Tret1
20	273	B3MG58	D.ananassae	Tret1
21	274	B4P624	D.yakuba	Tret1
22	275	B3NSE1	D.erecta	Tret1
23	276	B4HNS0	D.sechellia	Tret1
24	277	A1Z8N1	D.melanogaster	Tret1
25	278	B4QBN2	D.simulans	Tret1
26	279	Q8MKK4	D.melanogaster	Tret1
27	280	B4HNS1	D.sechellia	Tret1
28	281	B4QBN3	D.simulans	Tret1

CLADO 20.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	100	43.2	42.9	42.9	41.4	41.8	41.5	42.3	42.3	41.7	42.3	43	40.2	40.6	41.3
1	75	100	82.2	82.9	45.1	44.8	44.8	45.9	45.7	46.6	44.9	45.1	40.5	45.5	46.7
2	76.7	92.2	100	97	43.9	44.4	44.4	45.9	45.7	45.7	45.5	45.9	39.3	45.2	46.4
3	76.3	92.5	99.4	100	45.9	44	44	46.1	45.9	44.8	45.5	45.9	39.3	45.4	45.8
4	73.2	74.1	74.5	76.3	100	75.7	75.9	59	59	56.9	58	58.4	52	57.8	57.4
5	72.7	75.9	76.4	76.7	92.8	100	99.4	59	58.4	58.7	60.2	60.4	51	57.8	58.4
6	72.9	75.9	76.4	76.7	93	99.8	100	59	58.4	58.7	60.2	60.4	51.4	57.8	58.4
7	74	77.5	77.9	78.1	84.5	83.9	84.1	100	98.2	82.3	79.8	80.2	68.1	76.7	78
8	73.6	77	77.5	77.7	84.3	83.5	83.7	99.6	100	81.9	79.4	79.8	67.5	76.7	78.5
9	73	78.7	78.7	78.7	85.2	85.3	85.5	94.8	94.2	100	81.2	82.4	68.5	78.1	77.2
10	74.7	79.1	79.4	78.9	84.1	84.3	84.5	94.6	94.2	94	100	98.8	72.7	81.3	82.3
11	74.8	78.9	79.4	78.9	83.9	84.3	84.6	94.4	94	94.3	99.6	100	72.7	81.5	82.7
12	66.9	71.7	70.8	70.6	76.8	75.8	76.2	84.6	84.6	83.3	85.3	84.9	100	67.4	67.3
13	73.2	78.2	77.1	78.3	82.9	85.1	85.3	92.8	92.4	92.4	93.8	93.2	83.9	100	89.6
14	73.3	78.3	78.1	77.8	82.6	84.8	85	92.7	92.7	91.6	93.4	92.8	83.1	96.8	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	292	Q9BYW1	H.sapiens	SLC2A11
1	293	Q3T9X0	M.musculus	Slc2a9
2	294	Q5RB09	P.abelii	SLC2A9
3	295	Q9NRM0	H.sapiens	SLC2A9
4	296	P0C6A1	M.musculus	Slc2a7
5	297	A4ZYQ5	R.norvegicus	Slc2a7
6	298	Q6PXP3	H.sapiens	SLC2A7
7	299	P58353	B.taurus	SLC2A5
8	300	Q8WMN1	O.aries	SLC2A5
9	301	Q863Y9	E.caballus	SLC2A5
10	302	Q5RET7	P.abelii	SLC2A5
11	303	P22732	H.sapiens	SLC2A5
12	304	P46408	O.cuniculus	SLC2A5
13	305	P43427	R.norvegicus	Slc2a5
14	306	Q9WV38	M.musculus	Slc2a5

CLADO 21.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
0	100	35.7	36.7	38.3	36.9	37.2	32.3	31.9	32.6	32	31.2	32.3	32.5	34.4	33.3	36	35.8	36.2	35.8	35.8	36.4	36.2	36	36.8	36.8	36.5	36.1	36.2	33.5	34.7	35.1	35.1	36.1	36.7	36.1
1	72.5	100	51.9	51.5	51.4	52.1	43.5	46.7	41.9	42.3	43.3	42.8	42.5	49.9	47.9	51.3	51.1	51.5	51.5	51.5	51.5	51.5	51.5	48.2	53.1	49.4	50.3	49.7	48.6	49.7	49.7	50.6	51.3	52.2	50.7
2	69.3	80.7	100	94.5	93.7	93.7	55.9	56.2	53.2	53.1	53.5	52.4	52.9	64.4	60.6	66.2	65.5	65.9	66.1	66.1	66.1	66.1	65.9	59.9	59.3	58.9	59.8	58.6	59.1	60.3	60.7	58.1	56.9	57.6	58.2
3	70.4	80.3	99	100	95.3	95.3	56.3	57.1	53.1	52.7	52.9	52.1	52.5	64	59.8	65.6	64.9	65.5	65.9	65.9	65.9	65.9	65.7	60.1	59.3	59.6	60.2	58.8	59.6	60.7	60.9	58.3	57.2	57.6	58.2
4	69.7	80.1	98.6	98.6	100	98	55.7	56	50.9	52.5	52.5	51.9	52.2	63.8	59.5	64.8	65	65.3	66	65.7	65.3	65.8	65.6	59.1	58.9	58.5	59.4	58	59.1	60.3	60.5	58.3	56.8	57.2	57.8
5	70	80.3	98.6	99	99.6	100	56.4	56.3	52.9	52.9	52.9	52.1	52.4	64	59.7	65	64.9	65.5	65.9	66.1	65.5	66.2	65.5	59.9	59.1	59.4	60.3	58.4	58.7	60.7	60.9	58.5	57.2	57.6	58.2
6	63.7	72.4	75.9	75.7	75.7	75.7	100	63.3	61	64.1	64.5	62	62.6	53.6	54	55.2	56.3	56.1	56.3	56.3	56.9	56.5	56.7	53.2	51.8	49.1	49.3	48.5	48.2	49.3	49.3	47.9	49.3	49.9	49.7
7	66	77.7	80.4	81.1	80.3	80.7	83	100	56.6	58.6	57.2	57.5	58.4	54.4	52	54.5	54.7	54.9	55.5	54.9	55.2	55.2	55.2	53.4	52.4	50.8	51.3	50.7	51.2	50	50.2	49.8	49.5	50.6	50.4
8	64.7	74.1	78.5	78.1	77.5	78.8	83.8	82.5	100	87.3	81.4	78.3	77.7	51.8	49.3	52.7	53.7	53.7	54.4	53.5	53.7	54.2	54.2	49.6	49.2	47.7	48.6	48.8	47	47.9	47.9	47.2	48.8	49.8	50
9	65	75.2	80.2	79.7	79.5	79.8	87.3	82.7	95.1	100	87.8	80.2	79.6	52.7	49.6	53.1	53.2	53.4	53.2	53	53.2	53.2	53.2	48.8	50.1	49.3	49.5	49.4	48.7	49.5	49.5	50.3	49.3	48.9	49
10	64.8	74.4	78.7	79.1	78.4	78.7	86.5	81.7	92.8	96.4	100	81.9	81.5	53.1	50.5	53.3	53.4	53.4	53.7	53.5	53.2	52.9	52.9	50.4	48.8	49.1	49.9	50.3	49.5	50.6	50.6	50.5	49.7	50	50.2
11	65.3	74.8	78	78	77.6	78.1	85.6	80.7	91.8	93.3	94.3	100	94.8	52.4	49.3	49.8	51.7	52.5	52.3	52.3	52.1	52.6	52.6	49.2	47.6	48	48	49.5	48.4	48.7	49.1	48.5	48.6	49.2	49.2
12	65.9	75.2	78	77.9	77.1	77.6	86.1	80.2	91.6	93.1	93.1	99	100	51.8	48.5	50.7	51.2	51.2	52	52.2	51.4	51.9	51.9	49.2	48.5	48.1	47.5	48.4	47.7	48.4	48.8	48	48.7	49.1	49.1
13	70.8	82.6	84	85	85	85	78.1	80.7	76.3	78.7	77.7	77.1	77.1	100	77.8	83.3	81.4	82	81.9	81.2	81.4	81.2	81.4	68.3	72.1	67.4	65.3	66.4	67.3	68.5	68.7	68.6	69.2	70	67.3
14	69.6	82	85.1	85.9	85.1	85.3	78.7	80.3	77.6	78.6	78.4	76.5	76.5	92.9	100	78.3	77.2	77.6	77.7	77.2	77.6	77.4	77.4	68.4	65.2	64.2	62.4	62	62.3	62	62.4	62.9	62	63.3	63.1
15	69.8	82.3	86.3	86.5	85.9	85.9	78.3	80.4	78.1	79.2	78.5	78.2	78.3	93.9	94.9	100	87.6	88.2	87.9	87.6	88.2	87.8	88	65.9	70.5	66.9	66	66.5	65.2	66.7	66.7	68.4	67.2	68.1	68.9
16	68.9	84.6	86.4	86.6	85.6	86	78.9	81.6	79.5	80	78.9	78.7	77.7	92.5	94.1	95.5	100	97.2	96.5	97.4	96.5	95.9	96.1	67.4	70.7	66.7	67.4	66	63.8	65.8	65.8	67.4	67	66.9	66.7
17	68.7	84.5	86.6	86.8	86.4	86.4	79.1	81	79.9	80	79.1	79.3	77.7	92.7	94.1	95.9	99.6	100	97.2	97.8	97.6	97.2	97.2	67.6	70.7	66.9	67.4	66.2	63.6	65.8	65.8	67.6	67.4	67.2	66.9
18	67.5	84.3	86.4	86.6	86	86.4	79.1	80.9	79.5	80	78.9	78.3	77.9	92.5	94.1	95.7	98.8	99	100	99	97	97.4	97.2	67.6	70.3	67.6	67.9	66.6	63.8	66	66	67.4	67.6	67.4	67.4
19	67.9	84.8	86.4	86.6	86.6	86.6	79.1	81	79.3	79.8	79	78.7	77.7	92.3	94.1	95.1	98.8	99.2	99.8	100	97.6	97.2	97.4	67.6	70.3	67.2	67.9	66.4	63.6	66.3	66.3	67.2	67.8	67.4	67.2
20	68.2	84.3	86.8	86.8	86.6	86.6	79.3	81.3	80.3	80.4	79.2	78.9	77.5	92.5	94.5	95.7	99.2	99.6	98.8	99	100	98	98.2	68.4	71.1	67.6	67.9	66.8	64.2	66.5	66.5	67.6	67.5	66.9	66.7
21	68.2	84.3	87.2	87.2	86.4	86.4	79.7	81.5	79.6	80.2	78.6	78.8	77.8	92.3	94.1	94.9	99	99.2	99	98.8	99.4	100	99.6	68.2	70.7	67	67.7	66.2	63.8	66	66	67.4	66.9	66.5	66.6
22	68.4	84.5	87.2	87.2	86.4	86.8	79.7	81.5	79.6	80.2	78.6	78.6	77.6	92.5	94.3	94.9	99	99.2	99	98.8	99.4	100	100	68.4	70.7	67	67.7	66.2	63.8	66	66	67.4	66.9	66.5	66.6
23	69.1	76.5	83.1	84.2	83.3	83.8	77.7	79.1	78.4	78.2	77.5	76.7	76.1	88.8	90.3	86.9	88.6	88.6	88.1	88.1	89.4	89.4	89.6	100	67.9	66.8	65	68.3	63.6	67.2	67.5	65.2	65.2	67.6	64.7
24	69.1	82.3	81.9	82.3	81.5	81.7	78.7	80.2	78.4	78.9	78.2	77.3	78.1	89.9	88.1	88.9	91.1	91.3	90.6	90.9	90.9	90.4	90.4	87.5	100	71.3	70.5	72.7	71.4	72.8	73	73.9	73.1	74.3	74.3
25	68.9	80.8	83.4	83.2	82.4	82.4	75.4	78.8	77.6	78.3	77.3	76.8	77.1	87	85.9	86.9	87.6	87.6	87.4	87.8	87.6	87	87.2	85	88.2	100	94.3	83.2	81.5	82.8	83	84.4	81.1	82.2	82.4
26	68.4	80.6	83.2	83	82.8	82.8	75	78.1	77.6	77.9	76.4	76.6	76.2	85.6	85.7	86.2	87.3	87.3	87.3	87.6	87.3	87.1	87.1	84.8	88	98.8	100	81.7	81.1	82.6	82.6	84.1	80.7	81.7	81.9
27	69.6	79.4	82.9	82.9	82.7	82.7	75.8	78.1	77.4	76.9	76.2	76.4	75.9	86	86.5	87.9	87.7	88.2	87.9	87.7	87.9	87.5	87.5	88.1	89.5	95.1	94.5	100	84.4	84.9	85.5	84.4	84	84.8	85
28	68.5	79.2	81.5	81.7	81.5	80.6	75.6	79.1	76.3	77	76.9	76.6	76.8	86.1	86.2	85.4	85.4	85.9	85.4	85.4	85.9	86.1	85.9	84.6	88.6	92.9	92.9	94.5	100	94.3	94.9	85.8	83.4	83.7	83.9
29	69.2	80	83.1	83.5	83	83	77.2	79.1	77.6	78.1	78.1	77.4	77.8	87.9	87	86.7	87.2	87.6	87.2	87.9	87.6	87.6	87.6	88.6	89.8	93.9	93.5	95.6	98.6	100	99.4	88.1	84.9	85.6	85.8
30	69.2	79.8	83.3	83.5	83	83	77.2	79.1	77.6	78.1	78.1	77.4	77.8	87.9	87	86.7	87.2	87.6	87.2	87.9	87.6	87.6	87.6	88.6	89.8	93.9	93.5	95.6	98.6	100	100	88.3	85.5	85.8	86
31	70.2	81.6	82	82.2	81.8	81.8	75.1	78.9	76.7	77.9	78.3	77.2	77.6	87.6	87	87.2	88.3	88.5	88.1	88.3	88.3	87.9	88.1	85.7	89.7	94.9	94.5	96.2	95.9	96.6	96.6	100	89.2	89.8	89.6
32	69.4	82.3	81.8	82.1	81.9	81.9	75.9	78.9	76.7	75.8	76	76.1	76.2	87.6	86.7	87.4	87.5	87.9	87.9	88.3	87.4	87.7	87.9	84.9	89.4	93.3	92.9	95.1	94.7	95.5	95.5	96.5	100	91.6	91
33	68.8	82.5	82.2	82.3	81.9	81.9	76.6	77.6	77.3	76	75.6	77.1	77.5	88.1	87.6	88.1	87.7	88.1	87.9	88.7	87.7	87.2	87.4	87.7	90	93.9	93.7	95.7	94.9	95.5	95.5	97.6	98.4	100	98.8
34	68.1	80.9	82	82.1	81.7	81.7	76.4	77.5	77.7	76.5	76.3	77.1	77.7	86.3	87.8	88.3	87.4	87.8	88.3	88.3	87.4	87.7	87.9	85.4	89.8	93.7	93.5	95.5	94.5	95.1	95.1	97.2	98	99.6	100

Mat.	Filo.	UNIPROTID	Organismo	Proteína
0	307	Q26579	S.mansoni	GTP1
1	308	Q8IRI6	D.melanogaster	Glut1
2	309	Q27994	B.taurus	SLC2A4
3	310	P14672	H.sapiens	SLC2A4
4	311	P19357	R.norvegicus	Slc2a4
5	312	P14142	M.musculus	Slc2a4
6	313	Q90592	G.gallus	SLC2A2
7	314	N0A4A7	M.amblycephala	nan
8	315	P58351	B.taurus	SLC2A2
9	316	O62786	S.scrofa	SLC2A2
10	317	P11168	H.sapiens	SLC2A2
11	318	P14246	M.musculus	Slc2a2
12	319	P12336	R.norvegicus	Slc2a2
13	320	Q63ZP5	X.laevis	slc2a1
14	321	H9BPB6	S.aurata	SLC2A1
15	322	P46896	G.gallus	SLC2A1
16	323	P13355	O.cuniculus	SLC2A1
17	324	P11166	H.sapiens	SLC2A1

Mat.	Filo.	UNIPROTID	Organismo	Proteína
18	325	P17809	M.musculus	Slc2a1
19	326	P11167	R.norvegicus	Slc2a1
20	327	P20303	S.scrofa	SLC2A1
21	328	P27674	B.taurus	SLC2A1
22	329	P79365	O.aries	SLC2A1
23	330	Q6DFZ5	D.rerio	slc2a3a
24	331	P28568	G.gallus	SLC2A3
25	332	P32037	M.musculus	Slc2a3
26	333	Q07647	R.norvegicus	Slc2a3
27	334	Q9XSC2	O.cuniculus	SLC2A3
28	335	Q8TDB8	H.sapiens	SLC2A14
29	336	Q5R608	P.abelii	SLC2A3
30	337	P11169	H.sapiens	SLC2A3
31	338	P47842	C.lupus	SLC2A3
32	339	O62787	S.scrofa	SLC2A3
33	340	P47843	O.aries	SLC2A3
34	341	P58352	B.taurus	SLC2A3