DERIVANDO RELACIONES EVOLUTIVAS ENTRE LAS PROTEÍNAS DE CUBIERTA DE MEMBRANA MEDIANTE COMPARACIONES ESTRUCTURALES

Tesis entregada a la Pontificia Universidad Católica de Chile en cumplimiento parcial de los requisitos para optar al Grado de Doctor en Ciencias Biológicas con mención en Genética Molecular y Microbiología

Por
FERNANDO IGNACIO GUTIÉRREZ ESPINOSA

Director de la Tesis
Francisco Melo

Co-Director de la Tesis
Damien Devos

Septiembre 2021

A mis padres,
Por ser el principal motor de mi vida
y por su apoyo incondicional en las
decisiones que he tomado…

FACULTAD DE CIENCIAS BIOLÓGICAS
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

LA DEFENSA FINAL DE LA TESIS DOCTORAL TITULADA

"Derivando relaciones evolutivas entre las proteínas de cubierta de membrana mediante comparaciones estructurales"

Presentada por el Candidato a Doctor en Ciencias Biológicas
Mención Genética Molecular y Microbiología

SR. FERNANDO IGNACIO GUTIÉRREZ ESPINOSA

Ha sido aprobada por el Tribunal Examinador, constituido por los profesores abajo firmantes, calificándose el trabajo realizado, el manuscrito sometido y la defensa oral con nota 7,0.

DR. JUAN A. CORREA M.
Decano
Facultad de Ciencias Biológicas-UC

DR. ANDREAS P. SCHÜLLER
Coordinador Comité de Tesis
Facultad de Ciencias Biológicas-UC

DR. FRANCISCO MELO L.
Director de Tesis
Facultad de Ciencias Biológicas-UC

Damien Devos
Firmado digitalmente por Damien Devos
Fecha: 2021.10.16 08:48:38 +02'00'

DR. DAMIEN P. DEVOS
Co-Director de Tesis
Universidad Pablo de Olavide-CSIC
España

DR. LUIS LARRONDO C.
Miembro Comité de Tesis
Facultad de Ciencias Biológicas-UC

DR. CÉSAR RAMÍREZ S.
Miembro Externo Comité de Tesis
Instituto de Ingeniería Biológica y Médica-UC

Santiago de Chile, 14 de septiembre de 2021.

AGRADECIMIENTOS

Para el desarrollo de este trabajo, quiero agradecer enormemente la contribución de mi tutor el profesor Francisco Melo y mi cotutor el profesor Damien P. Devos, quienes han contribuido enormemente con sus comentarios y/o sugerencias al desarrollo de la herramienta creada en el transcurso de esta tesis doctoral. Ambos me han otorgado una valiosa experiencia profesional y familiar, además de un apoyo incondicional que me ha servido de soporte durante estos años para adentrarme de lleno en el área de la comparación estructural de proteínas y al mundo de las proteínas de cubierta de membrana.

Igualmente, quiero agradecer el compañerismo de los miembros de mi laboratorio y de aquellos que he conocido durante mis viajes al laboratorio del profesor Damien, quienes también me han mostrado un trato amable y familiar, con una mirada fresca al mundo científico. Además, agradezco el financiamiento recibido durante estos años en el desarrollo de esta carrera por medio de la beca CONICYT y también del proyecto REFRACT que me han permitido estudiar y costear mis gastos durante el transcurso del doctorado.

Por último, quiero agradecer el enorme apoyo emocional e incondicional que me han dado mis padres, y que, gracias a su crianza, he podido llegar hasta aquí sin dejar de lado mis valores, convicciones y creencias que me hacen una mejor persona. Y a las demás personas que me han ayudado de una u otra forma a recorrer este camino, les dedico este trabajo que se leerá a continuación.

Contenido

LISTA DE FIGURAS

Lista General

Artículo: "MOMA2: Improving structural similarity detection beyond the twilight zone"

LISTA DE TABLAS

Lista General

Artículo: "MOMA2: Improving structural similarity detection beyond the twilight zone"

ABREVIACIONES

ACE1	Ancestral Coatomer Element 1
AUC	Area Under the ROC Curve
CCV	Clathrin-Coated Vesicles
COPI	Coat complex protein I
COPII	Coat complex protein II
FECA	First Eukaryotic Common Ancestor
IFT	Intraflagelar transport
LECA	Last Eukaryotic Common Ancestor
MALISAM	Manual alignments for structurally analogous motifs
MC	Membrane Coat
MOMA2	Morphing & Matching 2
NPC	Nuclear pore complex
NTR	Nuclear transport receptors
PDB	Protein Data Bank
RMSD	Root-mean-square deviation
$RMSD_{pond}$	The weighted mean of the RMSD
SDM	Structural Distance Metric
SO	Structural Overlap
SPAH	Stacked Pairs of Alpha Helices
SSEs	Secondary structure elements

RESUMEN

Las proteínas de cubierta de membranas (MC) cumplen un rol esencial en el transporte intracelular, permitiendo el intercambio de materiales entre los compartimientos especializados dentro de las células eucariotas. Las proteínas MC poseen una combinación única de dominios β-propeller/α-solenoide que les otorga la capacidad de deformar membranas. Su arquitectura y las señales de secuencia señalan que estas proteínas probablemente evolucionaron a partir de un ancestro común, pero la gran divergencia estructural lleva a sugerir que estas proteínas han cambiado considerablemente durante la evolución. Esto constituye un reto sobre todo para determinar relaciones evolutivas usando las herramientas disponibles basadas en comparación de secuencias y estructuras producto de la extrema divergencia observada entre estas proteínas. Actualmente no existe un claro consenso acerca de los diferentes tipos de arquitectura que existen entre las proteínas de cubierta de membranas y, en el estado del arte, solamente se han generado árboles filogenéticos cualitativos según las relaciones parciales encontradas entre sus miembros sin considerar comparaciones estructurales a gran escala de manera exhaustiva.

Por ello, en el desarrollo de esta tesis se creó una nueva herramienta computacional para la comparación flexible de estructura de proteínas llamada MOMA2, que permite evaluar las similitudes estructurales entre proteínas muy divergentes mediante matrices de elementos de estructuras secundarias que nos permiten capturar la topología de las proteínas. Este programa ha sido diseñado para capturar las similitudes estructurales entre proteínas distantes mediante comparaciones estructurales flexibles, permitiéndonos clasificarlas según sus dominios y visualizar también el desplazamiento de cuerpo rígido de los sub-fragmentos encontrados equivalentes. Por ende, hemos utilizado esta herramienta para generar un nuevo esquema de

clasificación de las proteínas de cubierta de membrana basado en la similitud estructural de sus dominios, permitiéndonos a la vez, construir árboles filogenéticos basados en comparaciones estructurales que confirman las relaciones conocidas, además permiten extender aquellas relaciones ya existentes y finalmente descubrir nuevas relaciones entre sus miembros. Por último, esta herramienta ha permitido reformular un modelo parsimonioso considerando las posibles combinaciones de dominios que pudieron dar origen a los diferentes arreglos de dominios que están presentes en las proteínas MC.

En síntesis, el desarrollo de esta nueva herramienta nos ha permitido específicamente plantear los posibles pasos que dieron origen a las proteínas de cubierta de membrana a partir de las comparaciones de estructurales a pesar de la extrema divergencia que poseen estas proteínas. Por otro lado, desde una visión más general, esta herramienta nos abre nuevas posibilidades para explorar relaciones que aún permanecen ocultas debido a la extrema divergencia que poseen sus secuencias, pero que aún conservan sus similitudes estructurales.

ABSTRACT

Membrane coat proteins (MC) play an essential role in intracellular transport, allowing the exchange of materials between specialized compartments within eukaryotic cells. MC proteins have a unique combination of β-propeller/α-solenoid domains that gives them the ability to deform membranes. Sequence signal and architecture indicate these proteins probably evolved from a common ancestor, but major structural divergence suggest they have diverged considerably during evolution. The substantial divergence constitutes a challenge, primarily to available tools based on comparison of sequences and structures, to determine evolutionary relationships between the MC proteins. Thus, there is no clear consensus about the different classes between membrane coat proteins. Only qualitative phylogenetic trees have been generated according to their members' partial relationships without considering large-scale structural comparisons exhaustively.

For this reason, in the development of this thesis, we created a new flexible protein structure alignment tool called MOMA2 to evaluate the structural similarities between highly divergent proteins through matrices of elements of secondary structures that allow us to capture the topology of proteins. This program was designed to capture structural similarities in distantly related proteins through flexible comparisons, allowing us to classify them according to their domains and visualize the rigid-body movements of the equivalent sub-fragments. Therefore, we implemented this tool to generate a new classification scheme for membrane coat proteins based on their domains' structural similarity. It was used to build phylogenetic trees based on structural comparisons that confirm known relationships, allowing us to extend and discover new relationships between them. Lastly, MOMA2 was used to reformulate a parsimonious

model considering the possible combinations of domains that could give rise to the different arrangements of domains that are present in the MC proteins.

Overall, the development of this new program has specifically allowed us to propose the possible steps that gave rise to membrane coat proteins from structural comparisons despite the extreme divergence that these proteins possess. On the other hand, viewed from a more general perspective, this tool opens new possibilities for exploring relationships that remain hidden due to their sequences' extreme divergence but still retain structural similarities.

INTRODUCCIÓN

1.1. ¿Hasta qué punto podemos inferir la evolución divergente de las proteínas?

Todas las especies actuales se han desarrollado continuamente a partir de un número limitado de especies ancestrales (Darwin, 2004). Incluyendo sus sistemas biológicos, que han evolucionado en diversos complejos proteicos, mostrando a la vez, una enorme complejidad debido a la especialización y diferenciación de sus componentes que posiblemente derivan a partir una maquinaria más sencilla que se encontraba presente en organismos ancestrales. Posiblemente, el surgimiento de esta complejidad fue el resultado de numerosos experimentos naturales, mutaciones ocurridas durante el transcurso de varios millones de años de evolución. Estos cambios pueden haber sido pequeños en un principio como la sustitución de uno o varios residuos en su secuencia aminoacídica, o mayores como inserciones o deleciones de varios sub-fragmentos o subdominios, o la duplicación y/o fusión de distintos genes (Patthy, 2009). El conocimiento derivado de las estructuras resueltas de las proteínas nos revela que los residuos internos varían con una menor tasa de cambio que aquellos residuos que se encuentran en la superficie (Patthy, 2009). Además, estos cambios que se acumulan en la superficie de las proteínas van afectando sus funciones originales o interacciones con otras moléculas. En consecuencia, las proteínas que son lejanamente relacionadas adoptan un plegamiento similar, a pesar de que estas han divergido considerablemente a nivel de composición aminoacídica cambiando su función original.

Las familias de proteínas que generalmente reportan similitudes que se ven reflejadas en sus secuencias de aminoácidos, en la forma de sus estructuras o en sus funciones biológicas, pueden

llevar a suponer con cierto grado de certeza que descienden de un ancestro común, es decir, que son homólogas. En la Biología, el concepto de homología se refiere únicamente a que diferentes entidades provienen de ancestro evolutivo común, de allí, puede que sí o no, haya una similitud que se vea reflejada en diversos niveles, ya sea estructural, funcional, fisiológica o en su desarrollo. De hecho, en el estado del arte se considera solamente que la presencia entre un par de proteínas de una similitud mayor al 30% de identidad de secuencia, como evidencia suficiente para afirmar que son homólogas. Pero no hay que confundir que estas proteínas que poseen un alto porcentaje de identidad de secuencia también poseen un alto grado de homología o una homología significativa. La homología no tiene un alto o bajo grado de similitud, el concepto de homología se refiere a que si un grupo de proteínas comparte un ancestro común o no (Koonin and Galperin, 2013). Igualmente, la significancia estadística no puede ser alta o baja, es o no es. Son los indicadores, es decir el *p-value* o el porcentaje de identidad de secuencia que pueden tener valores altos o bajos, pero la significancia estadística y la homología son estados binarios. Por consiguiente, si un grupo de proteínas muestra solamente una fuerte similitud estructural y funcional a pesar de que comparten una muy baja similitud a nivel de secuencia es ampliamente aceptada como evidencia suficiente para afirmar que existe homología entre ellas. Sin embargo, si un grupo de proteínas comparte sólo una fuerte similitud estructural o sólo poseen en común una misma función, no se considera que la evidencia disponible sea suficiente para sugerir que estas proteínas comparten un ancestro en común. De hecho, los pares de proteínas que poseen un origen evolutivo distinto pueden adoptar un plegamiento estructural similar por convergencia posiblemente debido a que los procesos biológicos en los que participan puede favorecerlos a converger adoptando una misma forma. Por ejemplo, los dos motivos βαβββ presentes en la proteína OPPA de *S. typhimurium* y la proteína FosA de *Pseudomonas aeruginosa* son similares

tanto en arquitectura como en topología, pero el motivo βαββββ presente en OPPA tiene un origen hibrido a diferencia del motivo βαββββ de FosA, debido a que una de sus hebras β es una inserción al núcleo central conformado por las otras hebras y la hélice que están presentes en los miembros de la superfamilia de las proteínas de unión a Lisinas/Arginina/Ornitina, siendo considerado este par como proteínas análogas (Cheng *et al.*, 2008).

Actualmente, está ampliamente aceptado que la estructura es más conservada que la secuencia durante la evolución (Murzin, 1998). De hecho, hay numerosos ejemplos de proteínas que muestran una baja similitud a nivel de secuencia, pero aún adoptan una estructura similar, presentando los mismos residuos, o similares, en sus sitios activos con similares mecanismos catalíticos, por ejemplo, las lisozimas encontradas en diferentes organismos, desde bacteriófagos hasta mamíferos (Koonin and Galperin, 2013). Incluso, aquellas proteínas que han divergido más allá de los límites de detección de los métodos actuales basados en la comparación de secuencias, aún podrían retener una arquitectura similar sugiriendo que comparten un ancestro en común (es decir, que pueden conservar una similar combinación de dominios) (Murzin, 1998). Las razones para esta preservación estructural pueden depender de qué tan conservadores fueron los cambios según sus parámetros fisicoquímicos y la presión selectiva que sufrieron durante su evolución. Una teoría sugiere que una función esencial para un organismo cualquiera puede influenciar a un grupo de proteínas a retener su arquitectura ancestral a pesar de que, durante el transcurso del tiempo, los cambios aleatorios en sus secuencias se van acumulando hasta que eventualmente, las similitudes observadas entre sus secuencias aminoacídicas derivadas de un ancestro en común van desapareciendo (Murzin, 1998; Patthy, 2009).

Debido a que las estructuras son más conservadas que las secuencias, la información obtenida a partir de las estructuras de las proteínas es de vital importancia para detectar relaciones distantes entre proteínas muy divergentes a pesar de que sus secuencias han variado considerablemente. Actualmente, existe una enorme variedad de herramientas disponibles que utilizan diferentes acercamientos para estimar si la similitud reportada entre un par de estructuras es significativa o no. Por ello, a continuación, analizaremos a grandes rasgos las similitudes y diferencias que comparten estas herramientas para inferir relaciones estructurales entre las proteínas.

1.2. Programas utilizados para la búsqueda de relaciones evolutivas entre las proteínas

La búsqueda de relaciones entre las proteínas constituye un proceso clave para la Biología Celular y Molecular moderna porque nos permite comprender la historia evolutiva de varios complejos proteicos, así como entender, cómo sus funciones asociadas han divergido durante el transcurso de varios millones de años de evolución. Actualmente, para realizar estas búsquedas, existen principalmente dos enfoques: uno de ellos constituye las búsquedas realizadas a nivel de secuencias, mientras que el otro enfoque está asociado a las comparaciones estructurales.

Las búsquedas basadas en secuencias han sido exitosas en estos últimos años, debido a la enorme cantidad de secuencias disponibles en las bases de datos y a los avances implementados en el desarrollo de algoritmos de búsqueda más precisos que nos permiten realizar búsquedas sensibles y robustas contra las bases de datos de secuencias. Dentro de estas herramientas, se encuentran programas como Psi-Blast que están basados en perfiles de secuencias. Mientras que otros programas, como HMMER y HHPred, implementan modelos probabilísticos llamados modelos de Markov ocultos (HMM) para identificar secuencias de proteínas homólogas, donde el primero compara una secuencia contra un perfil de HMM y el segundo compara un par de

perfiles HMM (Altschul and Koonin, 1998; Eddy and Wheeler, 2007; Söding *et al.*, 2005). Estos programas han permitido extender la detección de familias de proteínas más a allá de la zona de penumbra (menor a 25% de identidad de secuencia) (Rost, 1999). Sin embargo, cada vez nos estamos acercando a los límites de cuánta información puede ser obtenida solamente a través de las secuencias aminoacídicas. Tal como se había mencionado anteriormente, las secuencias divergen más rápidamente que las estructuras y por ello, a una larga distancia evolutiva, podemos encontrar proteínas estructuralmente muy similares a pesar de que comparten una baja señal a nivel de secuencia. Por ejemplo, podemos mencionar casos como las comparaciones realizadas entre las proteínas ClpP proteasa y la enoyl-CoA hidratasa, o entre la hemoglobina y el citocromo b (Patthy, 2009).

Por otro lado, el conocimiento derivado a partir de las estructuras tridimensionales de las proteínas es otro tipo de información que tenemos disponible para inferir relaciones evolutivas entre las proteínas y en algunas ocasiones, es capaz de revelar conexiones inesperadas entre ellas. Por ejemplo tenemos el caso de la relación encontrada entre los componentes del citoesqueleto bacteriano y eucariota, particularmente entre las proteínas FtsZ y tubulina (van den Ent *et al.*, 2001) (Figura 1). El alineamiento de secuencia obtenido con el programa BLAST señala que este par comparte una baja similitud de secuencia y que el alineamiento resultante no es significativo, cubriendo sólo un pequeño porcentaje de la proteína *query* (el largo del alineamiento es de 103 pares de residuos alineados reportando un *e-value* 1,2). Sin embargo, cuando superponemos sus estructuras con un programa de alineamiento estructural como TOPMATCH, se observa claramente que estas proteínas son estructuralmente muy similares entre sí mostrando un alineamiento más largo que el obtenido con el programa BLAST (el largo del alineamiento es de 217 pares de residuos alineados con una desviación del error igual a 2,61

Å). Probablemente, existen muchos ejemplos como éste, donde las relaciones evolutivas entre pares de proteínas divergentes permanecen indetectables debido a las limitaciones de los programas de alineamiento de secuencias y a la falta de información estructural.

Por otra parte, la búsqueda de relaciones evolutivas a partir de comparaciones estructurales se está convirtiendo en una práctica común en este último tiempo. Cada año se observa un incremento exponencial del número de estructuras disponibles en la base de datos de la "*Protein Data Bank*" (PDB) (Figura 2). A medida que se han ido depositando más estructuras en la base de datos de la PDB, se ha hecho evidente la necesidad de desarrollar nuevos métodos de alineamiento estructural que permitan utilizar esta fuente de información para diversos propósitos, incluyendo inferir la función de las proteínas resueltas recientemente interpretando a nivel bioquímico su información estructural, así como en términos de sus relaciones evolutivas.

Pero ¿qué es un alineamiento estructural?, ¿cuáles son las ventajas y desventajas que poseen los métodos actuales?, y ¿cómo desarrollar un nuevo método para determinar relaciones estructurales en proteínas que presentan largas diferencias estructurales?, estas son las preguntas que responderemos a continuación.

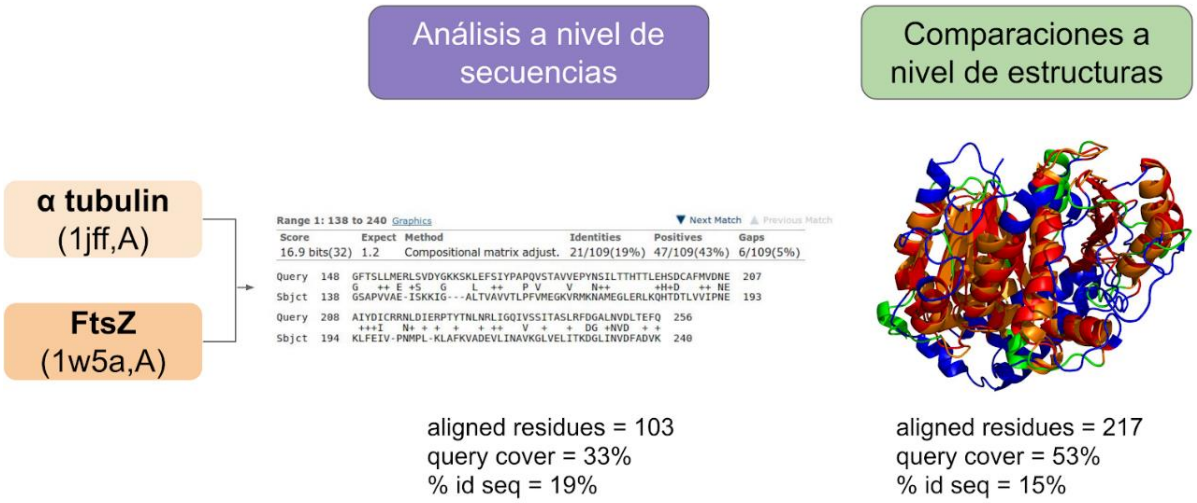


Figura 1. Ejemplo de la relación encontrada entre los componentes del citoesqueleto bacteriano y eucariota usando alineamientos de secuencia y estructura.
El alineamiento de las secuencias de las proteínas tubulina α (código PDB 1jff, cadena A) y FtsZ (código PDB 1w5a, cadena A) fue calculado con BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Adicionalmente, la superposición estructural de estas proteínas fue calculada con el programa TOPMATCH, donde los residuos en color rojo y naranja corresponden a los pares de residuos alineados entre las proteínas *query* y *target* respectivamente, mientras que los residuos no alineados fueron coloreados en azul y verde.

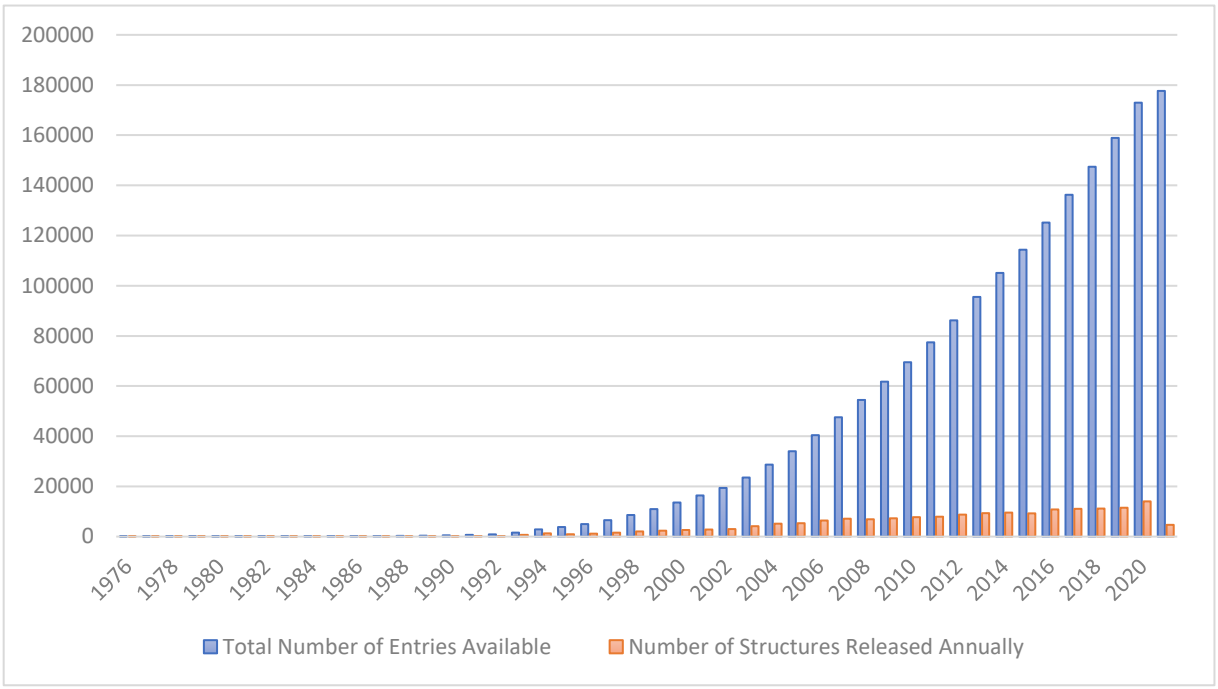


Figura 2. Crecimiento general del número de estructuras liberadas en la base de datos de la Protein Data Bank.
Estadísticas obtenidas a partir de la información publicada por la base de datos de Protein Data Bank (https://www.rcsb.org/stats/growth/growth-released-structures) hasta mayo del 2021. En total existen 182418 estructuras disponibles hasta la fecha.

1.3. ¿Qué es un alineamiento estructural?

Para establecer claramente que es un alineamiento estructural es importante resolver la confusión que existe comúnmente entre los términos comparación estructural, alineamiento estructural y superposición estructural (Bourne and Shindyalov, 2005). La comparación estructural se refiere al análisis de dos o más estructuras que han sido resueltas por diferentes métodos experimentales (como cristalografía de rayos X, resonancia magnética nuclear, o microscopia electrónica) buscando similitudes entre sus estructuras tridimensionales. Mientras que, el alineamiento estructural se refiere al establecimiento de equivalencias con cierto nivel

de detalle entre un par de estructuras. Estas equivalencias pueden ser de varios tipos (Burkowski, 2008):

- Entre átomos individuales.
- A nivel de residuos (especificado por las coordenadas de sus átomos Cα, Cβ o el centro de masa de sus cadenas laterales).
- Considerando la posición o la orientación de sus elementos de estructura secundaria (por ejemplo, entre las α-hélices o hebras β).
- O por medio de la similitud de los pliegues a nivel de estructura terciaria.

Por otra parte, la superposición estructural a diferencia del alineamiento estructural asume que ya se conoce cuáles residuos o elementos de estructura secundaria son equivalentes entre sí al sobreponer un par de estructuras. Generalmente, las posiciones de los elementos equivalentes se consideran puntos de anclaje en un espacio tridimensional que luego mediante el uso de un algoritmo de minimización (por ejemplo, el algoritmo de Kabsch) se determina la matriz de transformación que minimiza el error de la raíz media cuadrática (o RMSD por *root-mean-square deviation*) entre los puntos emparejados (Kabsch, 1976). En otras palabras, el alineamiento estructural ya se conoce y solamente se requiere ajustar las estructuras de tal forma que coincidan los elementos o residuos considerados equivalentes. Este problema algorítmico es mucho más fácil de resolver (de hecho, existe una solución exacta para hacerlo) que el problema de definir que aminoácidos o elementos de estructura secundaria que son equivalentes (Burkowski, 2008).

Para resolver el problema de la equivalencia de los residuos o elementos de estructura secundaria, los programas de alineamiento estructural consideran al menos tres o cuatro pasos para poder alinear un par de estructuras (Bourne and Shindyalov, 2005):

1. Primero representan las proteínas A y B (cadenas de las proteínas, dominios o varias cadenas a la vez) en un espacio independiente de coordenadas que pueden ser fácilmente comparadas.
2. Luego, comparan las proteínas A y B por medio de algún algoritmo que les permite comparar las relaciones geométricas que poseen (por ejemplo, distancias entre sus Cα, ángulos diedros entre sus elementos de estructura secundaria, área de contacto de los residuos, etc.).
3. Después, estos programas evalúan el alineamiento entre A y B calculando en el proceso una medida de similitud estructural llamada comúnmente *score*. Típicamente, se emplea una estrategia para optimizar este *score*. Por ejemplo, el algoritmo usado podría derivar el mejor alineamiento estructural o calcular algún tipo de correspondencia para los átomos de A y B para maximizar el *score*.
4. Finalmente, se mide la significancia estadística del alineamiento estructural con respecto algún conjunto al azar de comparaciones estructurales entre proteínas no relacionadas. La significancia estadística comúnmente es representada por medio de un *p-value* o comparándolo con una distribución de puntajes para obtener un Z-score (Hasegawa and Holm, 2009).

Estos programas tratan de resolver de forma general tres clases de problemas con sus resultados, los cuales se describen a continuación:

1. Establecer las equivalencias entre los residuos o los elementos de estructura secundaria.
2. Optimizar el alineamiento para reportar la mejor superposición estructural.
3. Identificar las estructuras más parecidas a la proteína *query* cuando esta se compara contra una base de datos de proteínas.

Mientras que, de forma específica, cada programa trata de resolver un problema de acuerdo con el propósito que fue diseñado. Por ejemplo, programas como FlexProt o FATCAT fueron creados para analizar cambios conformacionales de las proteínas flexibles, mientras que SHEBA o GANSTA+ fueron creados para detectar permutaciones circulares entre las proteínas mediante comparaciones estructurales (Ye and Godzik, 2003; Bliven and Prlić, 2012). A continuación, se describirán de forma breve, los diferentes tipos de programas de alineamiento estructural que existen actualmente.

1.4. Diferentes tipos de programas de alineamiento estructural

En estas últimas décadas, se ha incrementado considerablemente el número de herramientas disponibles para la comparación estructural, de hecho, existen decenas de programas que han sido desarrollados para comparar las estructuras de las proteínas (Hasegawa and Holm, 2009). Estos programas se diferencian en gran medida por el tipo de información que utilizan como entrada y el enfoque que usan para comparar las estructuras. Existen varios programas y/o servidores web como TOPMATCH, DALI, FATCAT o TM-align que son métodos de alineamiento estructural basados en residuos que utilizan las coordenadas cartesianas de los carbonos α para superponer un par de estructuras (Sippl and Wiederstein, 2008; Holm and Sander, 1995; Ye and Godzik, 2003; Zhang and Skolnick, 2005), mientras otros programas de alineamiento estructural como VAST, SSAP, GANGSTA+ o QP-tableau search (Gibrat *et al.*,

1996; Orengo and Taylor, 1996; Guerler and Knapp, 2008; Stivala *et al.*, 2009), para comparar las proteínas utilizan representaciones vectoriales de los elementos de estructura secundaria presentes en las estructuras analizadas (como α-hélices y hebras-β). Los métodos basados en el alineamiento de residuos son generalmente más precisos, pero más costosos computacionalmente que aquellos métodos basados en representaciones vectoriales de elementos de estructura secundaria, donde estos últimos poseen la ventaja de simplificar la información estructural requiriendo así menos poder de cómputo (Bourne and Shindyalov, 2005). Una excepción a esta tendencia es el programa TOPMATCH que apoyándose de las propiedades intrínsecas de la métrica que implementa, puede estimar la similitud de una comparación estructural de un par de proteínas valiéndose de otras comparaciones calculadas previamente con alguna de ellas (sabiendo la similitud estructural entre A con B y A con C puede estimar la similitud estructural entre B y C), puede realizar búsquedas eficientes y rápidas contra la Protein Data Bank (Wiederstein *et al.*, 2014). Mientras que otros programas como YAKUSA han tenido que sacrificar precisión en sus comparaciones para realizar búsquedas eficientes contra las base de datos de estructuras (Carpentier *et al.*, 2005; Zhang *et al.*, 2010).

Estos métodos también se pueden clasificar en tres tipos según el enfoque que usan para alinear las estructuras de las proteínas que pueden ser por medio de comparaciones rígidas, flexibles o elásticas (Figura 3). Aunque la mayoría de los programas de alineamiento estructural tratan las proteínas como cuerpos rígidos, es bien conocido que las proteínas son estructuras flexibles que pueden sufrir cambios conformacionales como parte de su función, por ello, para alinear las estructuras de las proteínas han aparecido los métodos de comparación flexible en estas dos últimas décadas. Entre los métodos más conocidos encontramos herramientas como FATCAT y FlexProt que consideran las proteínas como cuerpos flexibles, detectando

inicialmente torceduras (o *twists*) presentes al alinear un par de estructuras, para luego extraer los sub-fragmentos equivalentes de una estructura *target* para superponerla con respecto a una estructura *query* mediante translaciones y rotaciones independientes a lo largo de los puntos pívot (Shatsky *et al.*, 2004; Ye and Godzik, 2003; Salem *et al.*, 2010) (Figura 4). Estos programas usan algoritmos de encadenamiento para ir ensamblando los sub-fragmentos equivalentes denominados también pares alineados de fragmentos o AFP. FlexProt busca el conjunto más largo de los AFP introduciendo varios puntos de quiebre (o *hinges*)(Shatsky *et al.*, 2004). Mientras que FATCAT usa programación dinámica para el encadenamiento de los AFP, de modo que calcula el máximo puntaje de conectar una AFP con cualquiera de los alineamientos que termina en la AFP a fin de obtener el alineamiento estructural flexible que introduce el menor número de *twists* (Ye and Godzik, 2003).

Dentro de los programas de alineamiento estructural existe otro grupo que pueden inferir de forma intuitiva los pares de sub-fragmentos equivalentes sin definir *twists* entre las proteínas. Estos métodos se conocen como métodos de comparación elástica e incluyen los mapas de contacto o matrices de distancias (Figura 3). Los métodos elásticos pueden representar las estructuras de las proteínas usando matrices 2D que contienen las distancias entre todos los átomos Cα (como DALI) o entre todos los elementos de estructura secundaria (o VAST)(Holm and Sander, 1995; Gibrat *et al.*, 1996). Otros programas pueden usar solamente las diferencias angulares entre sus elementos de estructura secundaria como TableauSearch (Konagurthu *et al.*, 2008) o tomar solamente el área de contacto entre sus residuos como CAB-align (Terashi and Takeda-Shitaka, 2015). Estos programas son considerados métodos elásticos porque pueden encontrar regiones locales equivalentes entre las proteínas que se alinean sin considerar el marco de coordenadas, además no son afectados por los movimientos locales de sus dominios

(Hasegawa and Holm, 2009). Entre los métodos de comparación elástica, los programas más conocidos son DALI y VAST que colapsan las matrices de distancias obtenidas de las proteínas en regiones de *overlap* de tamaño fijo, las cuales luego se unen si hay sobreposición de los fragmentos adyacentes. Estos programas usan finalmente el algoritmo descrito por Holm y Sander para obtener la superposición óptima (Holm and Sander, 1995). Por otro lado, programas como TableauSearch o QP tableau search representan las estructuras de las proteínas usando matrices de elementos de estructura secundaria que denominan *tableaux* (Konagurthu *et al.*, 2008; Kamat and Lesk, 2007). Este tipo de representación permite capturar la esencia del patrón de plegamiento de las hebras β y las α-hélices que están en contacto, permitiendo realizar comparaciones rápidas contra conjuntos de estructuras de proteínas (Lesk, 1995). Estas herramientas tratan de encontrar el bloque equivalente (denominado *subtableaux*) que reporta la máxima similitud estructural de la matriz de diferencia, en lugar de inferir los bloques equivalentes para posteriormente combinarlos en un alineamiento flexible. Estos métodos usan dos pasos anidados de programación dinámica para alinear las matrices o emplean algoritmos para resolver problemas NP-complejos como la programación de enteros cuadráticos (QIP) o la programación lineal en enteros (ILP) para obtener el *subtableaux* no lineal que reporta la máxima similitud, para posteriormente usar adicionalmente el programa MUSTANG para obtener una superposición estructural a partir de las regiones equivalentes encontradas de las comparaciones de matrices (Stivala *et al.*, 2009; Konagurthu *et al.*, 2008). Por último, están los métodos basados en mapas de contacto de residuos como CMO (*maximum contact map*), como GR-align y CAB-align que a pesar de que pueden detectar regiones flexibles entre las proteínas que alinean a nivel de las matrices, estos programas entregan una superposición de las

estructuras que comparan. Estos programas son empleados para identificar relaciones estructurales a nivel de residuos en proteínas relacionadas (Terashi and Takeda-Shitaka, 2015).

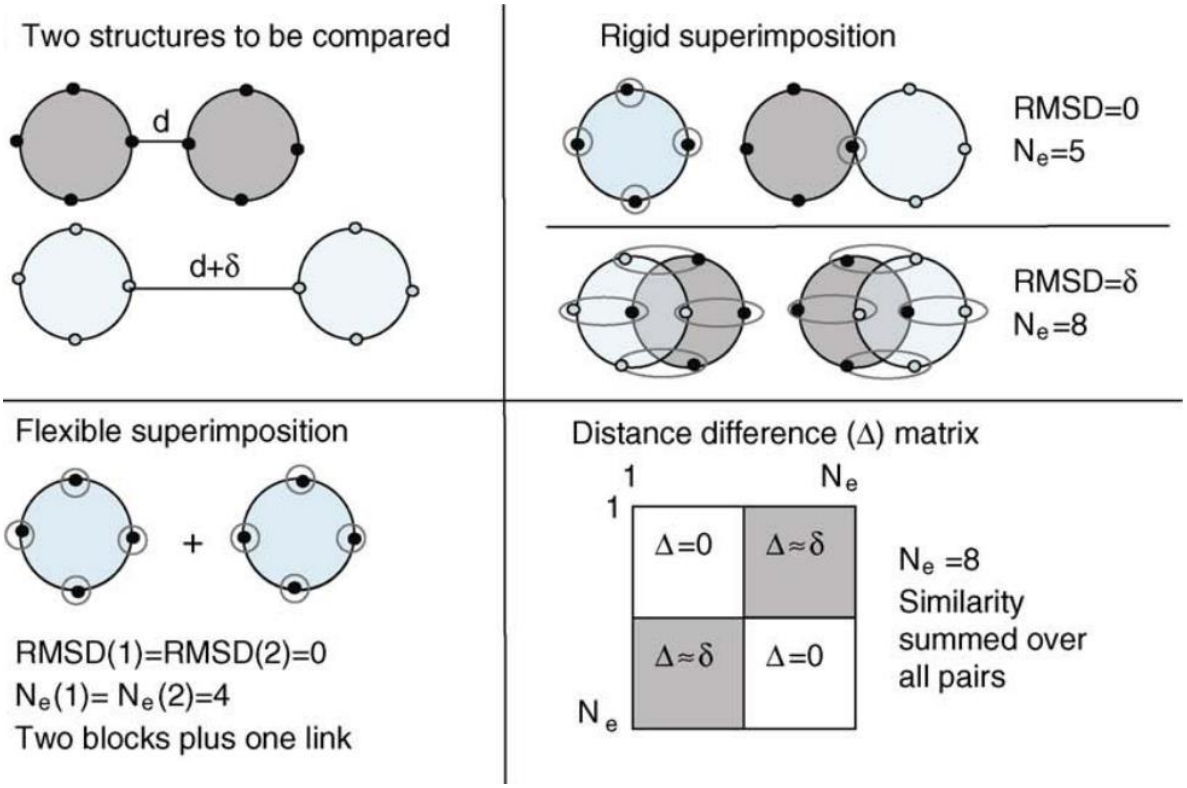


Figura 3. Ilustración esquemática de las comparaciones estructurales rígidas, flexibles y elásticas.
Arriba a la izquierda: la estructura inferior de un color más claro está relacionada a la estructura superior de color gris, presentando una translación (δ) del conjunto circular de puntos del lado derecho con respecto a los conjuntos circulares del lado izquierdo. Arriba a la derecha: se muestra que la superposición rígida debe tener un balance al reportar un alineamiento con un menor valor de RMSD o un conjunto más grande de puntos equivalentes (N_e). En este caso, se puede obtener una superposición perfecta de los cinco puntos o una superposición peor de los ocho puntos. Las elipses grises destacan aquellos pares que son considerados equivalentes. Abajo a la izquierda: la superposición flexible divide la estructura en varias subestructuras rígidas y luego aplica diferentes transformaciones de cuerpo rígido para superponer cada subestructura. Abajo a la derecha: las matrices de distancia o mapas de contacto son representaciones de las estructuras de las proteínas que son independientes del marco de coordenadas. Estas matrices de distancia pueden identificar tanto la conservación estructural como el movimiento entre las subestructuras. Esta figura ha sido modificada del artículo de Hasegawa y Holm (Hasegawa and Holm, 2009).

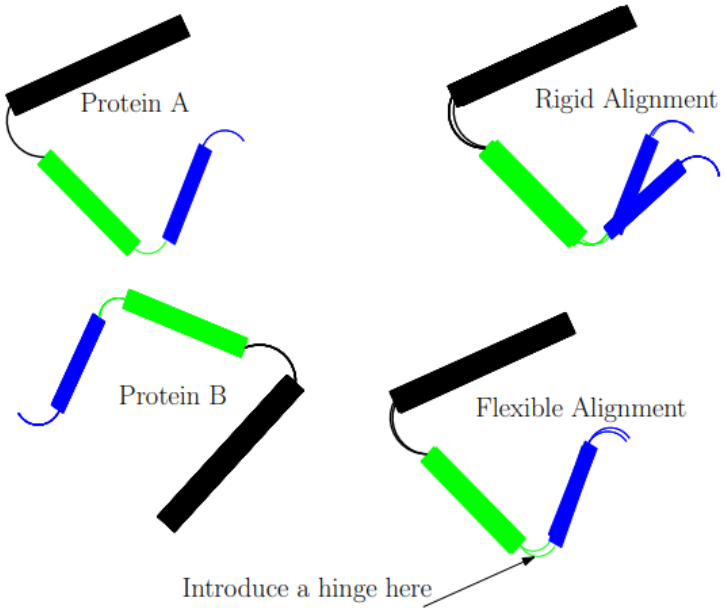


Figura 4. Alineamiento estructural flexible.
Las proteínas A y B poseen tres fragmentos estructurales similares, y son superpuestas por alineamientos rígidos y flexibles. El alineamiento rígido (arriba a la derecha) no es capaz de alinear el fragmento azul, pero el alineamiento flexible (abajo a la derecha) puede hacerlo fácilmente introduciendo un quiebre entre el bloque rígido (los fragmentos negro y verde) y el fragmento azul. Esta figura fue obtenida del artículo de Salem y colaboradores (Salem et al., 2010).

1.5. Ventajas y desventajas de los programas de alineamiento estructural para búsqueda de relaciones estructurales

Actualmente existe una enorme variedad de programas de alineamiento estructural donde cada uno de ellos cuentan con ventajas y desventajas dependiendo para el propósito que fueron diseñados inicialmente. En su gran mayoría, estos métodos usualmente tratan a las proteínas

como objetos estáticos, no siendo capaces de alinear correctamente proteínas relacionadas que presentan una gran variación estructural (Brohawn *et al.*, 2008a). Las proteínas son estructuras flexibles, ya que pueden presentar transiciones alostéricas, quiebres entre sus elementos de estructura secundaria o cambios conformacionales, siendo difícil para los programas de alineamiento rígido reportar similitudes significativas entre proteínas que poseen un alto grado de flexibilidad.

En las décadas recientes, han aparecido diversos métodos de comparación flexible para estudiar la flexibilidad estructural de las diferentes conformaciones de una proteína, o para establecer relaciones evolutivas en proteínas que presentan una gran variación estructural, siendo de gran ayuda para analizar a estos problemas a diferencia de los métodos de comparación rígida. Sin embargo, los programas de alineamiento flexible cuentan con algunas desventajas que es necesario tomar en consideración. Algunos programas de alineamiento flexible pueden requerir demasiado poder de cómputo, tomándole demasiado tiempo de cálculo en computadores de gama media para realizar consultas contra una base de datos con respecto a los programas de alineamiento rígido, dado que deben calcular adicionalmente los puntos de quiebre para establecer similitudes estructurales entre los sub-fragmentos equivalentes descritos. Segundo, estos métodos están diseñados para reportar el alineamiento estructural más largo con el menor RMSD introduciendo varios puntos de quiebre, aunque en algunos casos si se penaliza la inserción de *twists* (FATCAT al calcular su puntaje de alineamiento) igualmente estos métodos pueden sobreestimar la similitud reportada entre proteínas no relacionadas. La Figura 5 muestra, por ejemplo, el alineamiento estructural generado con FATCAT entre una hemoglobina con parte del extremo C-terminal de la nucleoporina Nup188. Aunque estas proteínas pertenecen a la clase “all α”, estas poseen plegamientos distintos y pertenecen a

diferentes familias de proteínas no presentando una conexión evolutiva entre ellas. Sin embargo, FATCAT reporta un alineamiento estructural significativo entre ambas (score = 194.67 con un *p-value* = 0.0115 < 0.05), introduciendo cinco *twists* para superponer ambas estructuras alineando pequeños pares de sub-fragmentos que están compuestos por 1 o 2 pares alineados de SSE. Este ejemplo señala claramente que la similitud reportada entre proteínas no relacionadas se puede sobreestimar al introducir varios *twists* o movimientos de cuerpo rígido, y produciendo en el proceso, superposiciones compuestas a veces por pequeños pares de sub-fragmentos que difícilmente se pueden considerar como significativos. Como resultado, estos *hits* están más arriba en el ranking que aquellas proteínas con las que sí existe una relación evolutiva, debido a que estos métodos pueden forzar el alineamiento estructural seleccionando varios pares de sub-fragmentos equivalentes donde sólo se alinearon uno o dos pares de elementos de estructura secundaria para reportar un alineamiento estructural más largo. Para enfrentar este problema, se pueden utilizar los mapas de contacto porque estos pueden determinar de forma intuitiva y eficiente los sub-fragmentos equivalentes en proteínas relacionadas según los patrones observados en las matrices de diferencias. Sin embargo, para determinar las mejores combinaciones de los sub-fragmentos equivalentes se requiere desarrollar un algoritmo adicional que permita obtener las mejores combinaciones sin tratar de explorar todas las posibles soluciones sin sobreposición. Sin embargo, como vimos anteriormente los métodos elásticos no reportan una superposición estructural la cual es necesaria para poder analizar en detalle las características de los alineamientos estructurales.

Por último, la gran mayoría de los métodos de alineamiento tanto rígidos como flexibles están restringidos a la comparación de dominios individuales en vez de realizar comparaciones de multidominios considerando que la gran mayoría de las proteínas están compuestas por más

de un dominio, con un promedio 1.7 dominios por cada proteína en genomas eucarióticos (Zmasek and Godzik, 2012).

Las desventajas observadas en los métodos actuales de comparación estructural muestran la necesidad de desarrollar una herramienta de comparación elástica que permita identificar similitudes estructurales entre proteínas que se encuentran más allá de los límites de comparación de las herramientas actuales, y a la vez, que sea rápida y precisa, permitiendo la integración de alineamientos locales de sub-fragmentos mediante alineamientos flexibles.

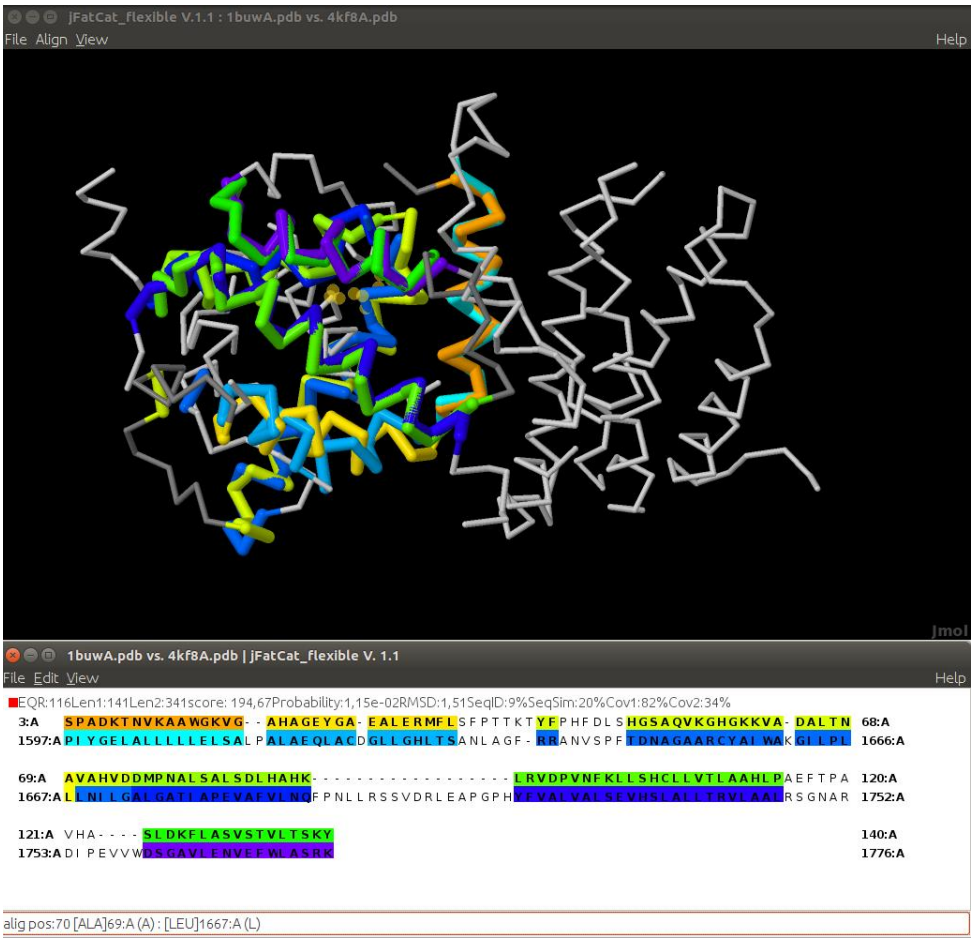


Figura 5. Ejemplo de una superposición flexible generada al comparar dos proteínas no relacionadas.
Superposición estructural calculada con la aplicación *standalone* de FATCAT (jFATCAT) entre la hemoglobina (código PDB 1buw, cadena A) y la NUP188C (código PDB 4kf8, cadena A). Los pares de sub-fragmentos equivalentes son indicados con varios pares de colores tanto en la superposición estructural como también en el alineamiento estructural de sus residuos.

1.6. ¿Cómo diseñar un programa de alineamiento estructural para estudiar las relaciones evolutivas en proteínas divergentes?

Para la búsqueda de relaciones estructurales entre proteínas divergentes es necesario desarrollar un método de alineamiento estructural que nos permita inferir las características conservadas entre las proteínas distantemente relacionadas. Este método debe ser capaz de

realizar alineamientos flexibles de las proteínas e identificar de forma intuitiva los movimientos rígidos entre sus dominios o cambios sustanciales debido a su flexibilidad. Además, necesitamos que este método nos permita identificar sub-fragmentos significativos y descartar aquellos que sean pequeños porque probablemente sus *matches* sean por azar. Por ende, requerimos de un nuevo método que pueda detectar relaciones evolutivas más allá de los límites de los métodos actuales de comparación rígida, y que lo pueda hacer de forma eficiente y precisa.

Para desarrollar este nuevo método, primero es necesario representar las estructuras de las proteínas usando una representación matricial para poder compararlas fácilmente. En 1980, Lesk y Chothia notaron que el patrón de contacto de los residuos es la característica más profundamente conservada en proteínas lejanamente relacionadas (Lesk and Chothia, 1980). Los métodos de alineamiento estructural basados en esta observación proporcionan los mejores resultados para buscar similitudes estructurales entre proteínas relacionadas que han divergido considerablemente (Konagurthu *et al.*, 2006; Terashi and Takeda-Shitaka, 2015). En 1995, Arthur Lesk fue el primero en desarrollar una representación tabular que comprime la información obtenida de las estructuras de las proteínas en una matriz 2D (Lesk, 1995). Esta representación utiliza una codificación de doble-cuadrante para codificar la orientación relativa de todos los pares de elementos de estructura secundaria (SSEs), es decir, sus ángulos diedros para representarlos posteriormente en una matriz.

Sin embargo, la comparación de las matrices de SSEs no es un problema trivial. Se han implementado distintas soluciones para comparar las matrices o mapas de contacto, incluyendo métodos basados en análisis de grafos, programación cuadrática lineal, programación dinámica iterativa, o programación dinámica para evaluar y comparar matrices de distintas dimensiones (Terashi and Takeda-Shitaka, 2015; Stivala *et al.*, 2009; Konagurthu *et al.*, 2006; Salem *et al.*,

2010). No obstante, estos métodos son lentos para realizar búsquedas contra las bases de datos o son imprecisos para encontrar *matches* locales entre las proteínas (Stivala *et al.*, 2009). Para evitar estos inconvenientes, es necesario implementar un nuevo método que implemente dos pasos secuenciales de programación dinámica, utilizando el primer paso para calcular un puntaje para cada par de SSE con la finalidad de construir una matriz de puntaje que se pueda emplear para alinear las secuencias de los SSEs de las proteínas en un segundo paso de programación dinámica. Estos dos pasos nos permitirán redimensionar las matrices SSE iniciales en matrices de tamaño idéntico para luego comparar sus ángulos o las distancias. Esto dará lugar a la creación una nueva matriz de diferencias denominada Δsubmatriz, donde los sub-fragmentos equivalentes podrán ser reconocidos visualmente al destacar las regiones con menores diferencias angulares o de distancias alrededor de la diagonal de la Δsubmatriz. En este punto, es importante mencionar que este alineamiento contiene la unión de todos los *matches* locales de elementos de estructura secundaria, el cual no puede ser interpretado directamente como una superposición estructural entre dos cuerpos rígidos. Por ello, el siguiente paso sería obtener un alineamiento estructural flexible a partir de la matriz de diferencias.

Anteriormente, hemos visto que algunos métodos basados en mapas de contacto no lo reportan o se valen de otros programas para alinear los sub-fragmentos equivalentes con el fin de reportar una superposición estructural. Mientras que algunos métodos flexibles sobreestiman la similitud estructural entre proteínas no relacionadas introduciendo varios *twists* para alinear sub-fragmentos pequeños para reportar un alineamiento estructural más largo. Considerando estos puntos es necesario crear una nueva metodología que nos permita generar un alineamiento estructural flexible a partir del alineamiento de matrices permitiendo:

1. Diferenciar los sub-fragmentos significativos al analizar la composición de los bloques presentes en la matriz de diferencias.
2. Obtener las mejores combinaciones de los bloques sin sobreposición o *subtableaux* locales sin realizar una búsqueda combinatorial.
3. Superponer localmente cada par de sub-fragmentos asociado a cada bloque destacado en la Δsubmatriz.
4. Obtener el mejor alineamiento estructural basado en la mejor combinación de *matches* locales de los sub-fragmentos que reporta el alineamiento estructural más largo.

Para el paso 1, se requiere desarrollar filtros que permitan identificar los bloques significativos, es decir, reconocer sub-fragmentos alineados que provienen de proteínas relacionadas y no relacionadas. Para el paso 2, se requiere de un algoritmo que permita inferir combinaciones de bloques de la Δsubmatrix sin sobreposición y que además pueda identificar las mejores combinaciones sin explorar todas las posibles soluciones. Para el paso 3, se puede utilizar el algoritmo de Kabsch para superponer localmente los sub-fragmentos encontrados equivalentes donde sus residuos pueden ser considerados como nubes de puntos que se pueden superponer inicialmente para luego optimizar superposición de forma iterativa para reportar la mejor superposición local (P. J. Besl and McKay, 1992; Kabsch, 1976). Por último, en el paso 4, se puede usar un programa que evalúe de forma local cada par de sub-fragmentos alineados para identificar posteriormente la mejor combinación que dará lugar a una superposición flexible, para ello podemos usar STOVCA que es un programa usado para estandarizar los resultados obtenidos de los alineamientos estructurales de diversos programas (Slater *et al.*, 2012).

Finalmente, a partir de estas comparaciones se pueden calcular diversos puntajes, por ejemplo, se puede obtener un puntaje de similitud estructural solamente de los alineamientos de las matrices o simplemente se puede obtener la similitud relativa o el porcentaje de *overlap* estructural de la superposición estructural. Por el contrario, se puede obtener un *p-value* a partir del puntaje de similitud comparando la probabilidad de obtener dicho puntaje según las comparaciones realizadas con respecto a conjunto de proteínas no relacionadas.

Esta nueva herramienta nos permitirá identificar similitudes estructurales entre proteínas que se encuentran más allá de los límites de las herramientas actuales, a la vez nos permitirá integrar de forma rápida y precisa los alineamientos locales de sub-fragmentos mediante alineamientos flexibles. En especial, si se quiere usar esta herramienta para determinar relaciones estructurales entre proteínas muy divergentes. A continuación, veremos un caso difícil para las herramientas disponibles pero relevante para comprender el origen de las células eucariontes.

1.7. Un caso difícil: las proteínas de cubierta de membrana

Un caso difícil para detectar relaciones evolutivas con las herramientas actuales de alineamiento de secuencia y de estructura son las proteínas de cubierta de membrana (o proteínas MC por sus siglas en inglés de “membrane coat”). Las proteínas MC junto con otros complejos proteicos como las proteínas de la familia de los CATCHR de los complejos “*tethering*”, las proteínas SNARE, las proteínas BAR y las Ras-GTPasas, participan activamente en conjunto para controlar el tráfico intracelular entre los diferentes organelos presentes en los eucariotas (Rout and Field, 2017). Entre las proteínas de cubierta de membrana se destacan principalmente cuatro complejos donde cada uno de ellos juega un rol clave en el desarrollo del sistema de endomembranas eucariótico, entre ellos tenemos los complejos clatrina/adaptinas (CCV), los complejos coatomer I (COPI), los complejos coatomer II (COPII) que están asociados a las

cubiertas de las vesículas permitiendo el tráfico del cargo de nutrientes en diferentes direcciones en el interior de las células, mientras que los complejos de nucleoporinas que forman parte del poro nuclear permiten el tráfico selectivo de moléculas del núcleo al citoplasma (Faini *et al.*, 2013; Debler *et al.*, 2008) (Figura 6).

Los tres primeros complejos proteicos (CCV, COPI y COPII) están implicados en diferentes rutas de tráfico en las células eucariotas. Estos complejos son importantes para recolectar y transportar el cargo entre los organelos, cubriendo y deformando las membranas de las vesículas. Los complejos COPI y COPII están involucrados principalmente en el transporte retrógrado y anterógrado de vesículas entre el retículo endoplasmático y el aparato de Golgi, mientras que los complejos CCV se encargan de empaquetar las proteínas en vesículas de la red trans-Golgi para luego ser transportadas hacia los organelos endocíticos o hacia la superficie celular siendo indispensable para los mecanismos de endocitosis y exocitosis (Dacks *et al.*, 2009).

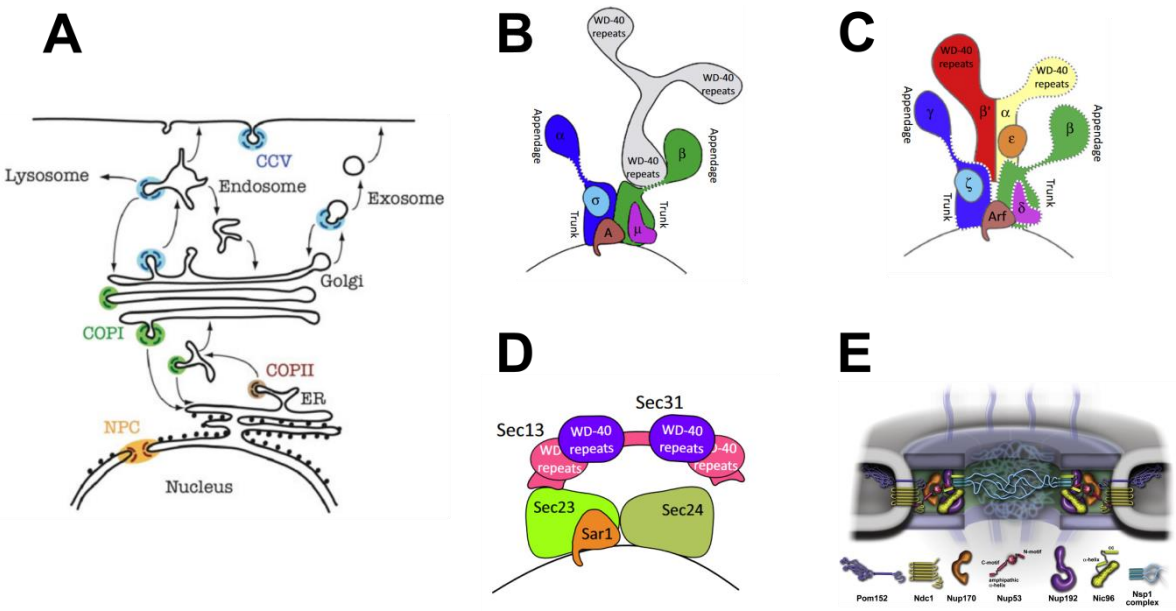


Figura 6. Localización de las proteínas de cubierta de membrana en las células eucariotas. (A) En la parte izquierda se representa la localización de los complejos de cubierta de membrana en las células eucariotas y como estos forman parte del sistema eucariótico de endomembranas. Los componentes de los complejos B) clatrina/adaptinas (CCV), C) coatomer I, D) coatomer II (COPII) y E) el poro nuclear, son representados a la derecha de la imagen. Los diagramas de los componentes que conforman los complejos fueron obtenidos de los artículos de B, C, y D) Faini et al. (2013), Trends in Cell Biology y E) Amlacher, S. et al. (2012), Cell; mientras que el esquema del sistema eucariótico de endomembranas fue obtenido del libro Fuerst (Devos, 2013).

Además, estos complejos siguen el mismo esquema básico de reclutamiento de sus componentes sobre las membranas a través de una pequeña Rab-GTPasa (las proteínas Arf o Sar que pertenecen a la misma familia de las Rab), seguido posteriormente por la selección del cargo y la deformación de las membranas dando lugar a la aparición de diferentes tipos de vesículas (Bonifacino and Glick, 2004; Jékely, 2003).

Los miembros de la familia MC también se extienden a aquellas proteínas que no forman parte del transporte de las vesículas pero que están asociadas a las membranas, por ejemplo, tenemos el cuarto complejo, el *scaffold* del poro nuclear (NPC, de Nuclear Pore Complex). El

poro nuclear a diferencia de los complejos que constituyen parte de la cubierta de las vesículas es un masivo complejo conformado por más de 500 copias de 30 diferentes proteínas llamadas nucleoporinas (NUPs). Estas proteínas se organizan en varios sub-complejos bioquímicamente estables llamados bloques de cubierta que se conectan radialmente formando anillos concéntricos: un anillo externo que está conformado por arreglos cabeza-cola del complejo Y (este complejo está compuesto por siete nucleoporinas y cuya estructura se parece a una Y), seguido por un anillo interno y el anillo de la membrana. El *scaffold* del poro nuclear está conformado por los anillos internos y externos cuya estructura central está compuesta mayormente por proteínas que presentan dominios SPAH (por *Stacked Pairs of Alpha-Helices*, también llamados α-solenoide) y β-propeller o una combinación de ambos.

Entre las características que podemos extraer de estos complejos es que cada uno de ellos se subdividen principalmente en dos subgrupos. El primer subgrupo lo constituyen los elementos estables que cubren la superficie de las vesículas (o *cage*), incluyendo también aquellas proteínas que constituyen el *scaffold* del poro nuclear. Las proteínas que forman parte del *cage* de las vesículas de CCV, COPI y COPII (como clatrina, Sec13/Sec31, COPβ' y COPα, respectivamente) y algunas Nups de los anillos internos y externos del NPC muestran una misma arquitectura a pesar de que entre ellos existe una baja o nula señal de similitud a nivel de secuencia. Por otro lado, el segundo subgrupo se refiere a aquellos elementos móviles y/o flexibles que en las vesículas establecen la conexión entre el *cage* con la membrana (conocidos como adaptadores) o aquellos elementos que permiten el transporte selectivo de moléculas entre el núcleo y el citoplasma a través del poro nuclear. Entre los elementos adaptadores se encuentran las adaptinas y los componentes adaptadores del complejo COPI, en cambio,

aquellos elementos que permiten el transporte selectivo a través del poro corresponden a los receptores de transporte nuclear (o NTR por sus siglas en inglés de *nuclear transport receptors*).

1.8. Arquitectura característica de las proteínas de cubierta de membrana

Los métodos de alineamiento de secuencia más sensibles usados para inferir relaciones evolutivas, como HHsearch (Söding *et al.*, 2005), son inútiles para inferir relaciones extensas entre las proteínas MC debido a la baja o casi nula señal que existe a nivel de secuencia entre los componentes que forman parte de los complejos *cage* en las vesículas y del *scaffold* del NPC. Las secuencias de estas proteínas son muy diferentes, tanto que no se pueden alinear con los programas convencionales de alineamiento de secuencias reportando alineamientos no muy diferentes de aquellos obtenidos al alinear un par de secuencias al azar. Sin embargo, las predicciones estructurales a nivel de los elementos de estructura secundaria han revelado similitudes, dado que las proteínas MC exhiben un arreglo de dominios que es particular sólo a las proteínas de cubierta de membrana de los eucariotas (Devos *et al.*, 2004). Las similitudes que podemos apreciar entre las proteínas MC es que en el extremo N-terminal presentan un dominio β-propeller seguido por un dominio α-solenoide en el extremo C-terminal (Devos *et al.*, 2004). Los dominios β-propeller están constituidos por seis o siete grupos de cuatro hebras beta antiparalelas llamadas *blades* que se encuentran enrolladas alrededor de un eje central formando una estructura similar a la hélice de un motor (de ahí viene el nombre de *propeller*). Mientras que los dominios α-solenoide denominados también como SPAH (Stacked Pairs of Alpha Helices) consisten en un arreglo más o menos compacto de varios pares de α-hélices formando una estructura flexible (Field *et al.*, 2011) (Figura 7). Considerando la presencia de esta arquitectura en las proteínas que constituyen estos complejos, se ha llegado a suponer que estos complejos comparten un origen evolutivo común.

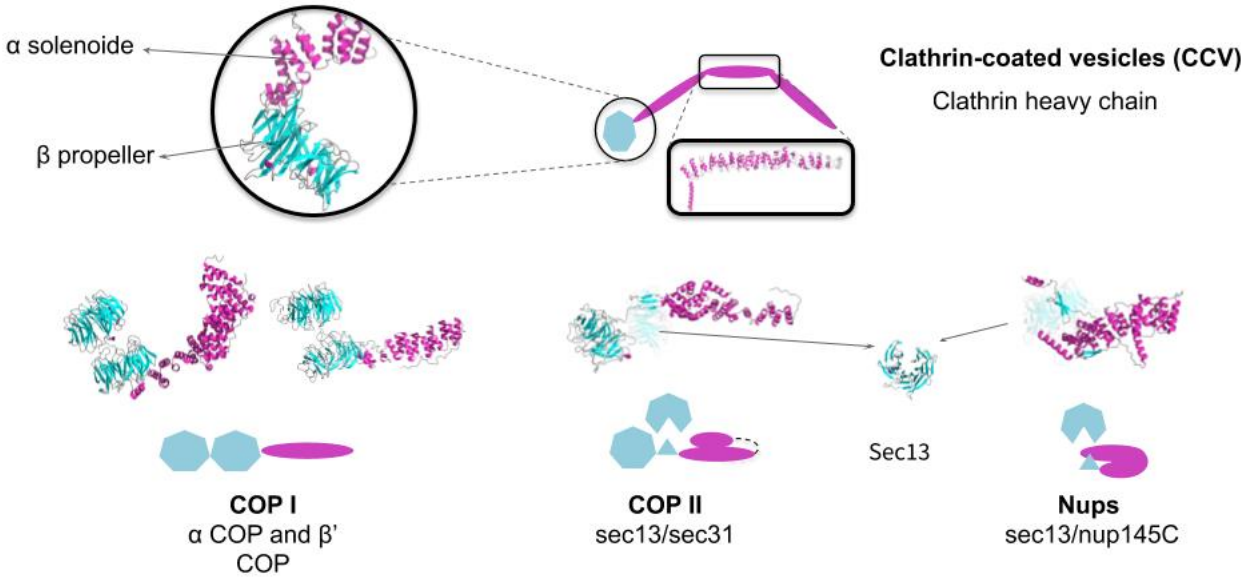


Figura 7. Arquitectura característica de las proteínas de cubierta de membrana.
Algunos de los componentes que forman parte de los complejos de cubierta de membrana presentan una característica combinación de dominios: un dominio β-propeller seguido de un dominio α-solenoide los cuales son representados con los colores celeste y rosado, respectivamente. Además, se muestra la interacción del β-propeller incompleto Sec13 con la proteína Sec31 del complejo COPII con la proteína Nup145C del complejo Y del poro nuclear.

Sin embargo, las subunidades que conforman estos complejos presentan una gran variedad de formas y tamaños. Por ejemplo, estas proteínas pueden presentar uno o dos dominios β-propeller en el extremo N-terminal (COP-β', -α y algunos componentes del Transporte Intraflagelar o complejo IFT) o un dominio β-propeller puede contener inclusiones de α-hélices entre sus *blades* (Nup 120, Nup157, Nup170 y Nup133) (Figuras 4 y 5). Otras proteínas poseen en el extremo N-terminal un único *blade* de cuatro hebras beta que complementa en *trans* a un dominio β-propeller incompleto (como Sec13 quien complementa a Nup145C) (Figuras 7 y 8). En algunas nucleoporinas y la proteína Sec31 de COPII sus dominios SPAH se pliegan sobre sí mismos en forma de J, mientras que en otras proteínas este dominio es corto y compacto (como en las COP-β', -α y -ε), o largo y extendido (por ejemplo, la cadena pesada de clatrina), mientras que en otras proteínas este dominio se tuerce en forma curvada (Nup192, Nup188, carioferinas y adaptinas) (Figuras 7 y 8).

Debido a la flexibilidad y versatilidad que poseen los dominios de las proteínas que componen estos complejos de cubierta de membrana, éstos a su vez se pueden ensamblar en bloques repetidos formando una red que se puede acomodar a diferentes curvaturas y adoptar distintos tamaños (Figura 8). Evolutivamente, este tipo de arquitectura es maleable beneficiándose de la flexibilidad que poseen sus dominios α-solenoide para contraerse o extenderse en el largo, permitiéndole adoptar diferentes tipos de ensambles y también se benefician de la rigidez que otorgan los dominios β-propeller a las estructuras que forman (Rout and Field, 2017). Esta simple combinación de motivos ha sido adaptada por las proteínas de cubierta de membrana para producir complejos proteicos que presentan distintas formas y funciones (Figura 8).

La alta plasticidad que poseen los dominios SPAH y la variedad de formas que exhiben la combinación de los dominios β-propeller/SPAH, constituye a la vez un gran reto en la detección de relaciones evolutivas entre las proteínas MC, principalmente debido a las limitaciones que poseen los actuales métodos de alineamiento de secuencias y estructuras para detectar similitudes significativas en proteínas que han divergido considerablemente (Field *et al.*, 2011). Sin embargo, la arquitectura única que poseen estas proteínas ha permitido establecer dos hipótesis que están entrelazadas y que tratan de explicar el origen de estos complejos.

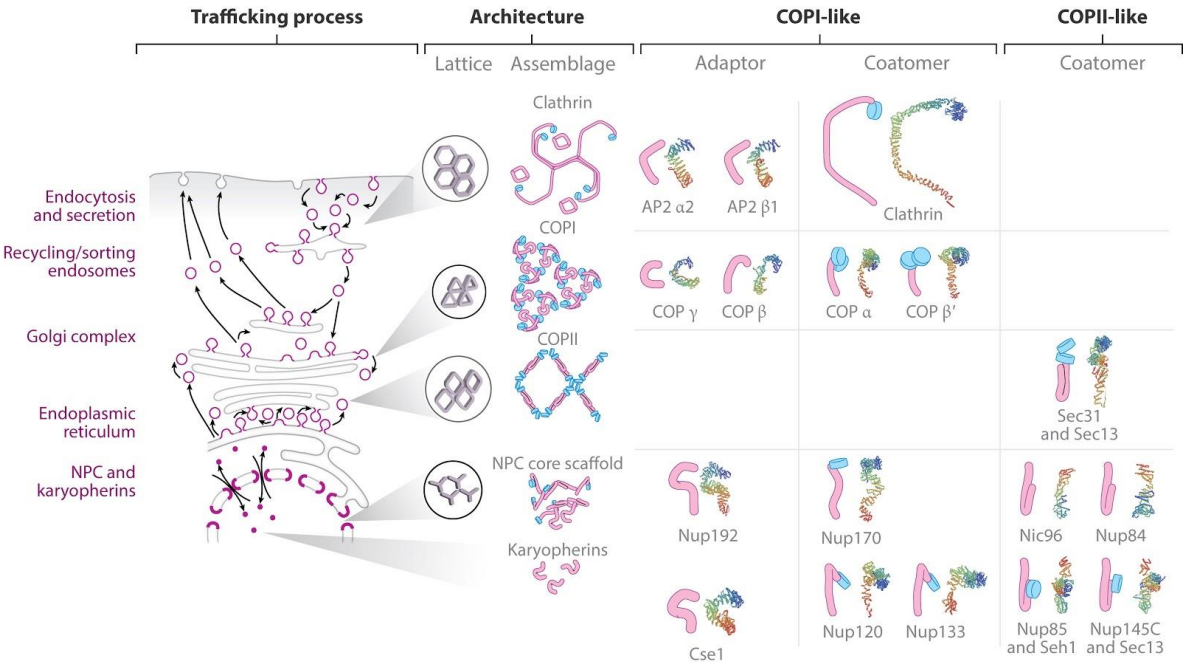


Figura 8. Anatomía molecular de los complejos de cubierta de membrana.
La parte izquierda de la figura muestra un esquema que enfatiza los eventos de transporte y los compartimientos que se encuentran asociados estos complejos, mientras la parte media del esquema muestra la arquitectura de ensamblaje de estos complejos en su asociación con la membrana. En el extremo derecho del esquema se detallan de forma esquemática la arquitectura que poseen los monómeros de estos complejos que se clasifican en dos grupos (COPI-like y COPII-like). La arquitectura de los monómeros que constituyen estos complejos es representada por la presencia de las combinaciones de los dominios β-propeller y α-solenoide los cuáles son destacados respectivamente con los colores celeste y rosado. Esta figura fue obtenida del artículo de Rout y Field, 2017. Annu. Rev. Biochem.

1.9. Hipótesis que explican el origen de las proteínas de cubierta de membrana

Una característica principal que poseen los eucariotas y que los distingue claramente de los procariotas es la presencia de un complejo sistema de endomembranas que les permite diferenciar funcionalmente su volumen celular. Por lo que entender su origen es clave para descifrar el origen de los eucariotas.

Estos compartimientos especializados se encuentran divididos en dos tipos característicos según su origen, los organelos endosimbióticos y los organelos autógenos (Jékely, 2007). Las mitocondrias y cloroplastos son organelos endosimbióticos que están presentes en organismos fotosintéticos cuyo origen se debió a la interacción endosimbiótica entre un organismo procariota que fue englobado por otro microorganismo. Mientras que los organelos autógenos se suponen que se formaron en el interior de las células de organismos ancestrales que poseían la capacidad de manipular sus membranas, entre estos organelos se encuentran el núcleo, el retículo endoplasmático, el aparato de Golgi, los lisosomas y endosomas. El conocimiento acerca del origen de los organelos autógenos es escaso comparado con la evidencia disponible para aquellos organelos que tienen un origen endosimbiótico (Jékely, 2007). Recientemente, se han propuesto dos teorías complementarias que enriquecen nuestra comprensión acerca la evolución del sistema de endomembranas y el rol de las proteínas de cubierta de membrana: la hipótesis de la paralogía de los organelos (OPH) y la hipótesis del protocoatomer (PCH) (Figuras 9 y 10).

La OPH propone que los organelos autógenos surgieron como resultado de la duplicación de genes y la neo-funcionalización de la maquinaria de tráfico preexistente en organismos ancestrales (Dacks and Field, 2007) (Figura 9). Mientras, que la PCH estipula que la adquisición

de una proteína ancestral llamada *protocoatomer* fue clave para la aparición del sistema de endomembranas permitiendo a los ancestros de los eucariotas actuales manipular sus membranas (Devos *et al.*, 2004)(Figura 10). Siguiendo con la hipótesis del OPH, la duplicación y la divergencia del *protocoatomer* permitió la aparición de diferentes variaciones de las proteínas MC observadas en la actualidad.

Por un lado, la evidencia que sustenta la OPH señala que algunas familias de genes parálogos que están en contacto con el sistema de endomembranas como las proteínas Rabs, sintaxinas, adaptinas y Arf-GAPs sufrieron duplicaciones de sus genes que ocurrieron antes de la aparición de la primera célula eucariota (Boehm and Bonifacino, 2001; Elias *et al.*, 2012; Dacks and Doolittle, 2002). Por el otro lado, la evidencia disponible que sustenta la PCH se basa en la presencia del arreglo característico de dominios β-propeller y SPAH que presentan las proteínas de cubierta de membrana o la interacción de proteínas que presentan estos dominios aislados, así como la presencia de elementos compartidos que interactúan entre los distintos complejos sugieren que estas proteínas comparten un ancestro evolutivo en común (Devos, 2013; Dokudovskaya *et al.*, 2011). Estas afirmaciones están sustentadas mediante predicciones de estructura secundaria, cristalografía de rayos X y microscopía electrónica que apoyan la PCH (Kim *et al.*, 2018; Brohawn *et al.*, 2008a; Devos *et al.*, 2004). En los últimos años, en especial, las predicciones de elementos de estructura secundaria ha permitido identificar la presencia de esta arquitectura particular en otras proteínas de complejos asociados al sistema de endomembrana, como los complejos CORVET/HOPS, complejos SEA y el complejo IFT, cuyas funciones están asociadas al anclaje de los endosomas al citoesqueleto, la autofagia en las vacuolas y el transporte intraflagelar, respectivamente (Algret *et al.*, 2014; Dam *et al.*, 2013a; Jékely and Arendt, 2006).

No obstante, las proteínas de cubierta de membrana muestran una enorme variación a nivel de secuencia que es reflejada en las diferencias observadas en su estructura, arquitectura, interacción y la formación de las cubiertas (Faini *et al.*, 2013)(Figura 8). Debido a esta extrema divergencia, las proteínas de cubierta de membrana son pobremente alineadas por los programas actuales basados en alineamiento de secuencias. Además, debido a la baja señal a nivel de secuencia que poseen estas proteínas, las posibles conexiones entre estas proteínas han sido por lejos intratables con los métodos disponibles impidiendo la directa evaluación de las hipótesis propuestas para estas proteínas usando análisis filogenéticos (González-Sánchez *et al.*, 2015). También, la evaluación directa de su similitud estructural entre los componentes de los diferentes complejos es difícil debido en parte a la carencia de estructuras que representen todas las proteínas de cubierta de membrana conocidas y las limitaciones de las herramientas actuales debido a que estas proteínas han divergido considerablemente mostrando una gran flexibilidad.

A pesar de estas limitantes, igualmente se ha podido establecer una relación entre las proteínas de cubierta de membrana y el origen del poro nuclear. La información disponible hasta la fecha ha permitido plantear una nueva teoría acerca del origen del poro nuclear y que tiene una fuerte relevancia con el origen de los eucariotas.

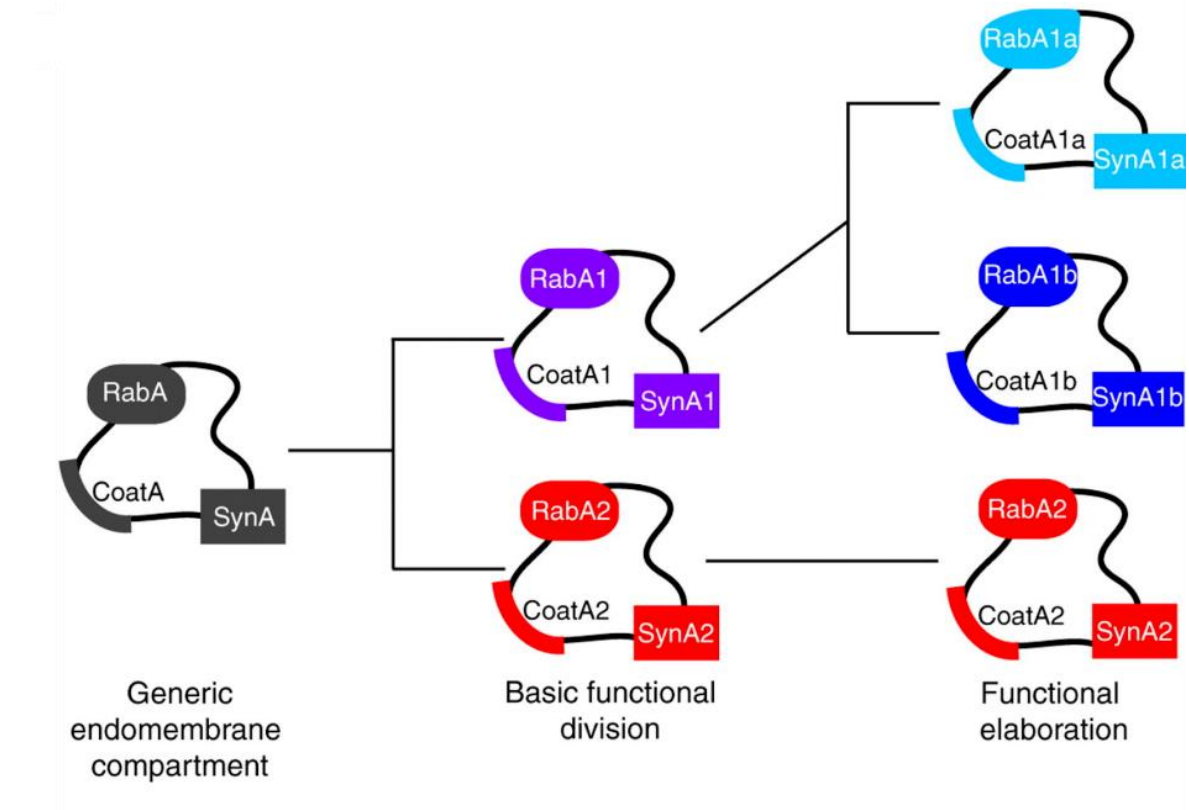


Figura 9. Hipótesis de la paralogía de los organelos (OPH).
Este modelo supone que un compartimiento progenitor A (compartimiento en color gris) contenía al menos un miembro de la familia de las Rab, sintaxinas y de las proteínas de cubierta de membrana. Posteriormente, las duplicaciones de los genes que codificaban estos factores facilitaron a la aparición de una división funcional de un compartimiento en dos compartimientos (componentes de color rojo y violeta). Este proceso continuo gradualmente, donde en algunos casos, los compartimientos resultantes se diferenciaron en un mayor grado dando origen a compartimientos con funciones especializadas a pesar de que derivan de un compartimiento en común (componentes de color celeste y azul). Mientras que, en otros casos, el grado de divergencia fue menor y la función del compartimiento resultante no cambio tanto con respecto a la función que presentaba el compartimiento ancestral del cual divergió (componentes de color rojo). Este proceso general pudo dar eventualmente origen a los distintos organelos que se observan en las células de los organismos eucariotas. Esta figura fue modificada de la imagen publicada en el artículo de Dacks y Field, 2007. Journal of Cell Science.

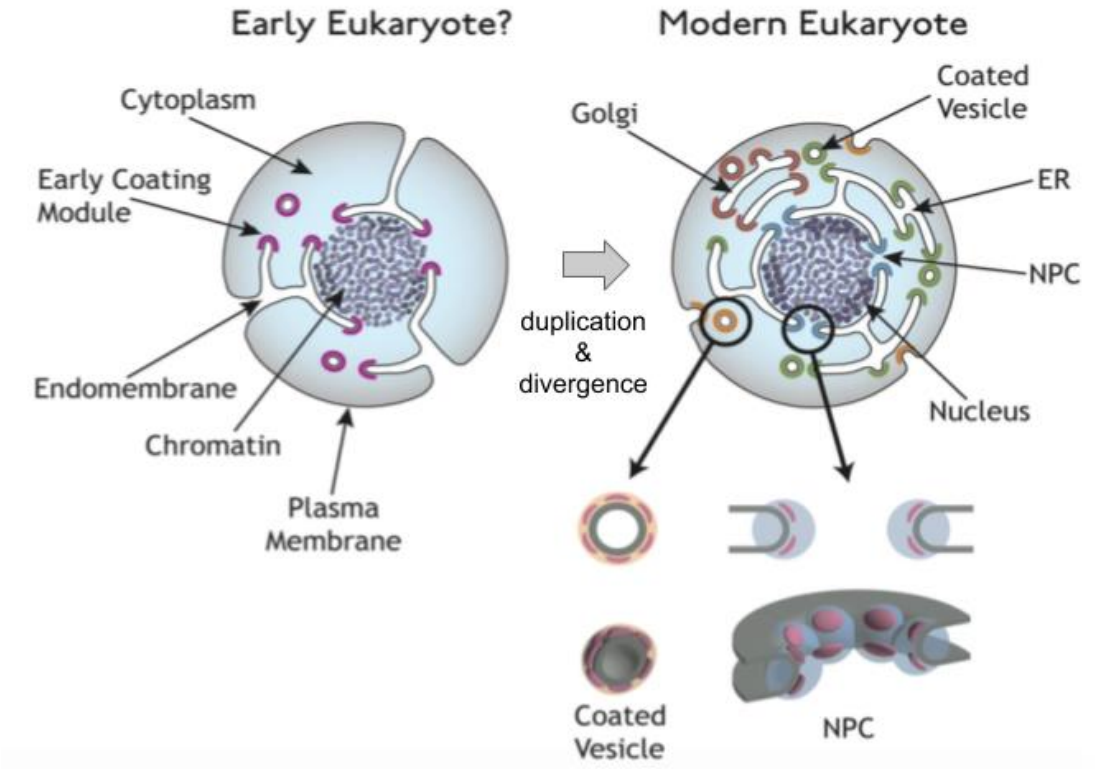


Figura 10. Hipótesis del Protocoatomer (PCH).
En la parte izquierda se muestra que una célula pre-eucariota ancestral adquirió un sistema de proteínas que le otorgo la capacidad de curvar las membranas permitiéndole además moldear su membrana plasmática en compartimientos internos. En la parte derecha, se muestra que eventos de duplicación génica y divergencia de este sistema de proteínas condujo a la diversificación de este módulo en otros módulos que se observan en los eucariotas modernos. Estas proteínas están asociadas funciones especializadas como la endocitosis (módulo de color naranjo), el transporte entre retículo endoplasmático y el aparato del Golgi (módulos de color verde y café) y a la formación del poro nuclear (módulo de color azul). Los módulos presentes en las cubiertas de las vesículas y el poro nuclear son necesarios para estabilizar las membranas curvadas. Esta figura fue obtenida de la imagen publicada en el artículo de Devos et al., 2004. PloS Biology.

1.10. Origen y evolución del poro nuclear

A parte de la existencia de un complejo sistema de endomembranas, la presencia de compartimento que encapsule la información genética de un organismo es otra de las características distintivas que diferencian a los eucariotas de los procariotas. En las últimas décadas se han planteado diversas teorías acerca del origen del núcleo, desde un origen endosimbiótico bacteriano o viral, pero la comunidad científica en su mayoría está de acuerdo que el origen del núcleo es autógeno (Jékely, 2007). Recientemente se ha propuesto una hipótesis alternativa acerca de su origen y la aparición del poro nuclear. Esta nueva hipótesis sugiere que el núcleo es un producto tardío de la evolución de lo eucariotas y que antes de la aparición del último ancestro común de los eucariotas (LECA, por "Last Eukaryotic Common Ancestor"), hace un billón y medio de años, no existía un núcleo tal como lo conocemos hoy (Field and Rout, 2019). Esta hipótesis postula que el poro nuclear surgió por la duplicación e interacción de dos familias de proteínas que poseían una arquitectura similar a los complejos que forman parte de las cubiertas de vesículas (pre-COPI y pre-COPII) que se originaron del protocoatomer ancestral. Posteriormente, los componentes de estos complejos evolucionaron después de varios eventos de duplicación, divergencia y pérdida secundaria de sus genes conduciendo a la aparición del poro nuclear (Figura 11B y 11C). Las pistas que apoyan esta hipótesis se encuentran en la forma estructural que poseen los dominios que están presentes en las proteínas que forman parte de los anillos internos y externos del poro nuclear (Field and Rout, 2019). El núcleo del *scaffold* del poro nuclear posee una estructura modular que se ha mantenido generalmente invariable en la mayoría de los eucariotas. Este enorme complejo proteico compuesto por 30 tipos distintos de proteínas está conformado principalmente por ocho

sub-complejos que pueden ser divididos verticalmente en dos columnas de subunidades repetidas de parálogos que probablemente surgieron a través de duplicaciones (Figura 11D) (Kim *et al.*, 2018; Field and Rout, 2019). Cada una de estas subunidades se han clasificado en dos tipos, llamados respectivamente en tipo I y II, considerando la forma de la arquitectura que presentan siendo similar a las que poseen algunas subunidades de las proteínas de cubierta de vesículas.

El pliegue de tipo I es parecido a la arquitectura que presentan los componentes de los complejos COPI/clatrina/adaptinas, mientras que el pliegue de tipo II es parecido a la arquitectura que poseen las subunidades del complejo COPII como Sec31 y Sec16. Las proteínas Nup120, Nup133, Nup157 y Nup170 presentan arquitecturas características al pliegue de tipo I donde un continuo dominio SPAH se proyecta lejos del dominio β-propeller, mientras que las nucleoporinas Nup192 y Nup188, que interactúan con Nup157 y Nup170, muestran una similitud estructural parecida a las adaptinas. Por otro lado, las proteínas Nic96, Nup145C, Nup84 y Nup85 poseen la distintiva y discontinua arquitectura de tipo II, donde la proteína Nup145C interactúa con la proteína Sec13 de forma similar a la forma en como interactúa Sec31 con Sec13 en el complejo COPII. Además, las nucleoporinas del tipo II poseen dominio SPAH que adopta un pliegue compuesto por tres módulos llamado ACE1 (por *ancestral coatomer element 1*). Este pliegue compuesto principalmente por α-hélices y se encuentra dividido en tres partes denominadas como corona, tronco y cola (Brohawn *et al.*, 2008a). Por otra parte, las carioferinas que están asociadas al transporte del cargo a través del poro nuclear comparten una estructura similar a las adaptinas sugiriendo posiblemente que estas proteínas se originaron y co-evolucionaron en conjunto con el poro nuclear (Stuwe *et al.*, 2014; Sampathkumar *et al.*, 2013). Aunque el orden en el cuál surgieron estas duplicaciones no puede ser completamente

resuelto, está claro que una versión temprana del poro nuclear presentaba al menos un par de estructuras de tipo I y II que posteriormente por duplicación y divergencia dieron origen a las subunidades parálogas que observamos en el *scaffold* del poro nuclear (Field and Rout, 2019).

Una hipótesis adicional sugiere que la función que posee el poro nuclear como un canal selectivo fue resultado de la evolución de los módulos de protocoatomer ancestrales que constituyen el *scaffold* del poro nuclear. Esta hipótesis llamada *connector-to-FG* sugiere que el poro nuclear ancestral estaba compuesto inicialmente por módulos de protocoatomer cuyos dominios α-solenoides interactuaban con conectores intrínsecamente desordenados. Estos módulos formaban inicialmente un canal permeable que presentaba una baja selectividad (Hayama *et al.*, 2017). Posteriormente por eventos de duplicación y divergencia, la evolución de estos módulos expuestos en el interior del canal incrementó la compartimentalización del núcleo, permitiendo a la vez que los conectores desordenados se especializaran dando origen a las nucleoporinas FG. Como resultado, el poro nuclear se convirtió en una barrera selectiva y permeable que regula el intercambio eficiente del cargo entre el núcleo y el citoplasma en los eucariotas modernos (Hayama *et al.*, 2017).

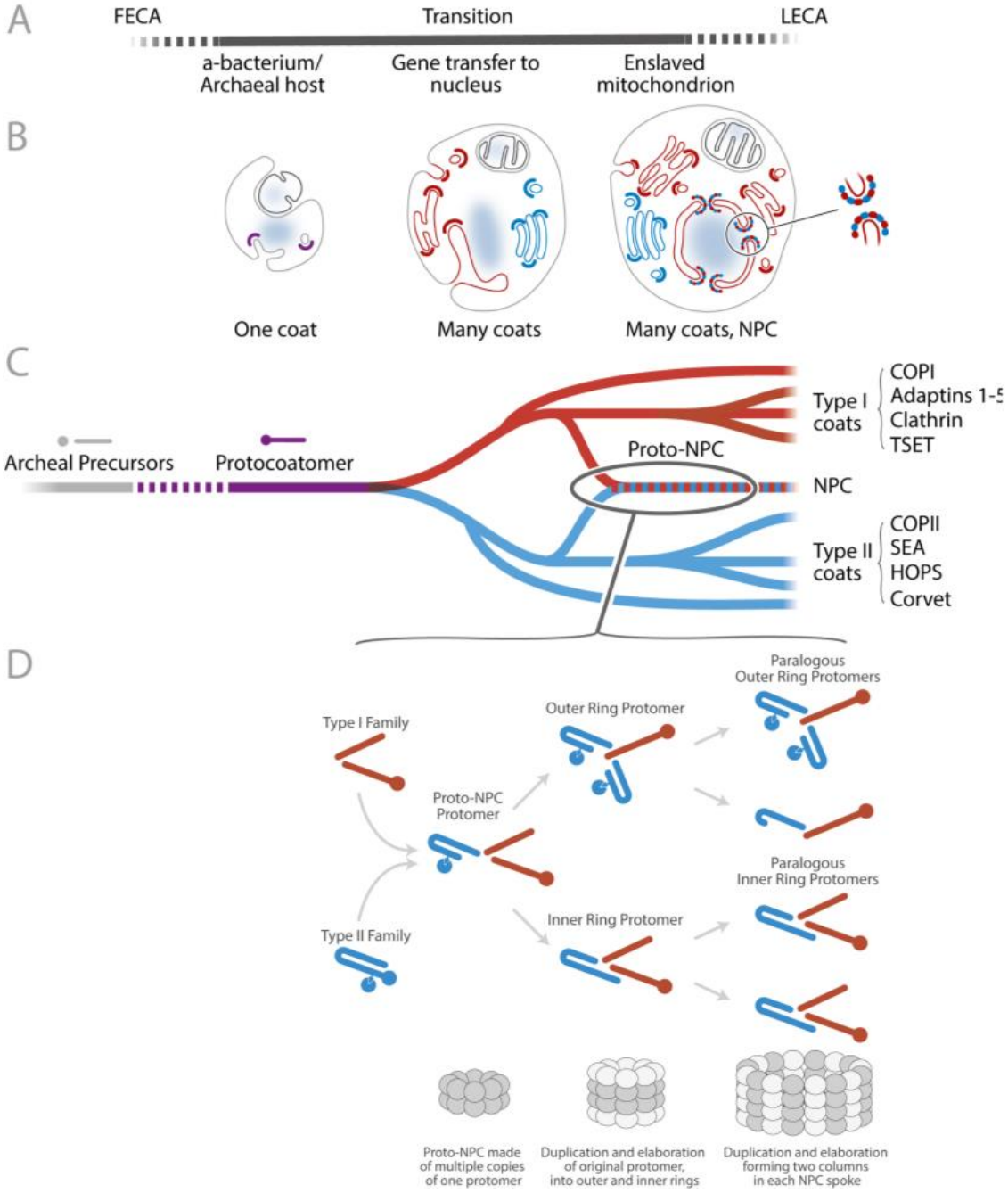

A
FECA
Transition
LECA
a-bacterium/
Archaeal host
Gene transfer to
nucleus
Enslaved
mitochondrion
B
One coat
Many coats
Many coats, NPC
C
Archeal Precursors
Protocoatomer
Proto-NPC
NPC
Type I
coats
COPI
Adaptins 1-5
Clathrin
TSET
Type II
coats
COPII
SEA
HOPS
Corvet
D
Type I Family
Type II Family
Proto-NPC
Protomer
Outer Ring Protomer
Inner Ring Protomer
Paralogous
Outer Ring Protomers
Paralogous
Inner Ring Protomers
Proto-NPC made
of multiple copies
of one protomer
Duplication and elaboration
of original protomer,
into outer and inner rings
Duplication and elaboration
forming two columns
in each NPC spoke

Figura 11. Hipótesis acerca del origen del poro nuclear a partir de la amalgama tardía de complejos de cubierta de membrana del tipo I y II.
En la parte superior del esquema se ilustra con cierto grado de incertidumbre el período de transición entre FECA (First Eukaryotic Common Ancestor) y LECA (Last Eukaryotic Common Ancestor), donde ocurrió el secuestro gradual de la mitocondria mediante la transferencia de genes desde el genoma de la bacteria original al genoma del huésped (A). En la parte de arriba de la zona central se destaca la localización de los dos tipos de protocoatomer (tipo I en color rojo y del tipo II en azul) que probablemente derivan del protocoatomer ancestral (en color violeta) y que estuvieron presentes durante esta transición, destacando que el poro nuclear está constituido por una amalgama de los tipos I y II (B). En la parte inferior de la zona central de la imagen se propone la rama evolutiva posible de los complejos de cubierta de membrana. Sugiriendo que el protocoatomer derivó de una fusión de genes de Arqueas, indicando varios ejemplos específicos de complejos de cubierta que actualmente presentan los dos tipos de arquitectura (C). La parte inferior del esquema ilustra la posible ruta que condujo a la formación de los anillos internos y externos del poro nuclear dando énfasis a los posibles eventos de duplicación y la unión de sus subunidades dando origen al scaffold del poro nuclear en los eucariotas modernos (D). Esta figura fue obtenida del artículo de Field y Rout, 2019, F1000Research.

Estas nuevas teorías acerca del origen del poro nuclear plantean una respuesta plausible relacionada con el origen del núcleo y el rol esencial de las proteínas de cubierta de membrana en dicho proceso. No obstante, actualmente existe un conocimiento limitado de las relaciones que existen entre los complejos de cubierta de membrana y el poro nuclear. A continuación, se describirá cuáles son las relaciones evolutivas conocidas entre las proteínas de cubierta de membrana y por qué es necesario contar con un nuevo método de comparación estructural para inferir nuevas relaciones entre estas proteínas.

1.11. ¿Cuáles son los pasos evolutivos entre las proteínas de cubierta de membrana?

Análisis filogenéticos realizados para algunas de las familias de proteínas involucradas en el tráfico intracelular, incluyendo también algunas proteínas de los complejos de cubierta de membrana han demostrado que los eventos de duplicación y fusión de genes que dieron origen a sus parálogos ocurrieron antes de la aparición de FECA (Field and Dacks, 2009). Estos

resultados sugieren que el sistema de endomembranas que estaba presente en FECA ya estaba desarrollado mostrando una gran complejidad. Por lo que nos lleva a preguntarnos ¿cuál es el origen de los complejos proteicos que contribuyeron al desarrollo del complejo sistema de endomembranas actual? y en especial, ¿cuál es el origen de las proteínas de cubierta de membrana eucariotas?

Análisis genómicos realizados en varias arqueas pertenecientes al clado *Asgard*, han señalado proteínas ortólogas a componentes clave del sistema de endomembranas, sugiriendo que las arqueas son posiblemente los precursores de estas familias de proteínas que tienen un rol clave en el sistema de endomembranas (Spang *et al.*, 2015; Zaremba-Niedzwiedzka *et al.*, 2017). Por ejemplo, se han descubierto en algunas arqueas proteínas que presentan el dominio *longin* que se encuentra principalmente presente en las proteínas SNARE en los eucariotas. También, se han encontrado proteínas homólogas al dominio BAR, a la familia de las GTPasas y a las proteínas de los complejos ESCRT, proponiendo que estas proteínas asociadas al sistema de endomembranas probablemente tienen un origen arquea (Dacks and Robinson, 2017).

No obstante, en el caso de las proteínas de cubierta de membrana aún no está claro si derivan de las arqueas. Zaremba-Niedzwiedzka y colaboradores en 2017, sugieren que hay proteínas de cubierta de membrana en las arqueas *Asgard* al encontrar grupos de genes cercanos que codifican los dominios WD40 (dominios β-propeller) adyacentes a dominios TPR (dominios α-solenoide). Sin embargo, sus marcos abiertos de lectura se encuentran en orientaciones opuestas posiblemente por azar, donde sus dominios no se encuentran fusionados en un solo gen como es el caso de las proteínas de cubierta de membrana que se encuentran presentes en los eucariotas. Además, las proteínas que poseen sólo un dominio β-propeller o SPAH pueden estar presentes en todos los organismos, por lo que no sería raro encontrar estos dominios aislados en

las arqueas. Estas observaciones llevan a suponer que no existe evidencia suficiente para afirmar que las proteínas de cubierta de membrana eucariotas deriven de las arqueas. Sin embargo, lo que es realmente interesante es que se han encontrado precursores de proteínas que interactúan directamente con las proteínas de cubierta de membrana. En estos años recientes se han encontrado homólogos remotos de los complejos adaptadores de COPII (proteínas homólogas a Sec23/Sec24) que están codificados en los genomas de un subconjunto de *Thorarchaeota* de las arqueas *Asgard* (Zaremba-Niedzwiedzka *et al.*, 2017). Esto no significa que los componentes que presentan la arquitectura de los protocoatomer en los complejos coatomer II deriven de las arqueas. En primer lugar, estos complejos adaptadores de COPII no poseen un pliegue similar a los dominios α-solenoide. Segundo, no poseen homología detectada a nivel de secuencia y estructura con los complejos adaptadores de Clatrina y COPI, sugiriendo que poseen un origen independiente (Faini *et al.*, 2013). Los resultados de este estudio sugieren que las arqueas fueron una fuente clave de varios precursores como los complejos Sec23/Sec24, dominios longin, RAS y ESCRT para el desarrollo de la maquinaria de tráfico intracelular en los eucariotas, pero el origen de los precursores de las proteínas de cubierta de membrana eucariotas aún sigue en discusión.

Por otra parte, aún no se conocen cuáles fueron los pasos intermedios que dieron origen a las proteínas de cubierta de membrana actuales, solamente se proponen que estas transiciones ocurrieron antes de la expansión de los eucariotas. Por ejemplo, los análisis filogenéticos y genómicos comparativos realizados a algunas proteínas de cubierta de membrana como Sec16, Sec31 y Sec13 sugieren que las transiciones que dieron su origen probablemente ocurrieron antes de la aparición de LECA (Schlacht and Dacks, 2015).

De hecho, la presencia de rasgos compartidos entre algunos componentes de los complejos de cubierta de membranas eucariotas ha llevado a los científicos a suponer modelos cualitativos acerca de las posibles relaciones evolutivas que existen entre estos complejos (Rout and Field, 2017; Schlacht and Dacks, 2015; Field and Dacks, 2009) (Figura 12). Los primeros modelos propuestos sugieren la presencia de dos grandes grupos o familias en la historia evolutiva de las proteínas de cubierta de membrana (Schlacht and Dacks, 2015). La primera familia llamada COPI incluye los complejos clatrina/adaptinas, COPI y TSET, mientras que la segunda familia llamada COPII incluye los componentes de COPII, el poro nuclear, los complejos SEA y HOPS/CORVET. La evidencia reciente acerca de los complejos IFT muestra que existe incertidumbre con respecto a qué familia pertenece, conectando ambos grupos (Schlacht and Dacks, 2015).

En cambio, otro modelo señala que existen tres familias de proteínas que derivan del protocoatomer en lugar de dos, una tercera familia llamada *adaptin-like* que incluye a las nucleoporinas Nup192, Nup188 con las carioferinas y adaptinas en un solo grupo (Sampathkumar *et al.*, 2013). Este trabajo sugería dos posibles escenarios que describen la evolución de los complejos de cubierta de membrana considerando las similitudes estructurales encontradas entre Nup192 con las familias de las proteínas carioferinas, β-cateninas y adaptinas (Figura 13). El primer escenario describe que la aparición de los complejos MC fue producto de la reintegración y la mezcla de tres tipos diferentes de complejos de cubierta ancestrales a partir de los cuales derivan todas las proteínas de cubierta de membrana, mientras que el segundo escenario más parsimonioso sugiere que estas tres arquitecturas evolucionaron juntas, y luego por eventos de duplicación y pérdida secundaria dio lugar a la aparición de los distintos tipos de proteínas de cubierta de membrana que se encuentran presentes en los eucariotas modernos.

Una hipótesis más reciente que se describió anteriormente en el capítulo 1.10, sugiere que el poro nuclear está conformado por proteínas de cubierta de membrana del tipo I y II. No obstante, las diferencias observadas en la clasificación de estos complejos nos indica que aún no existe un claro consenso acerca del origen de estos complejos o los distintos tipos de proteínas de cubierta de membrana que existen debido a que aún no se conocen claramente todas las relaciones evolutivas que hay entre ellos.

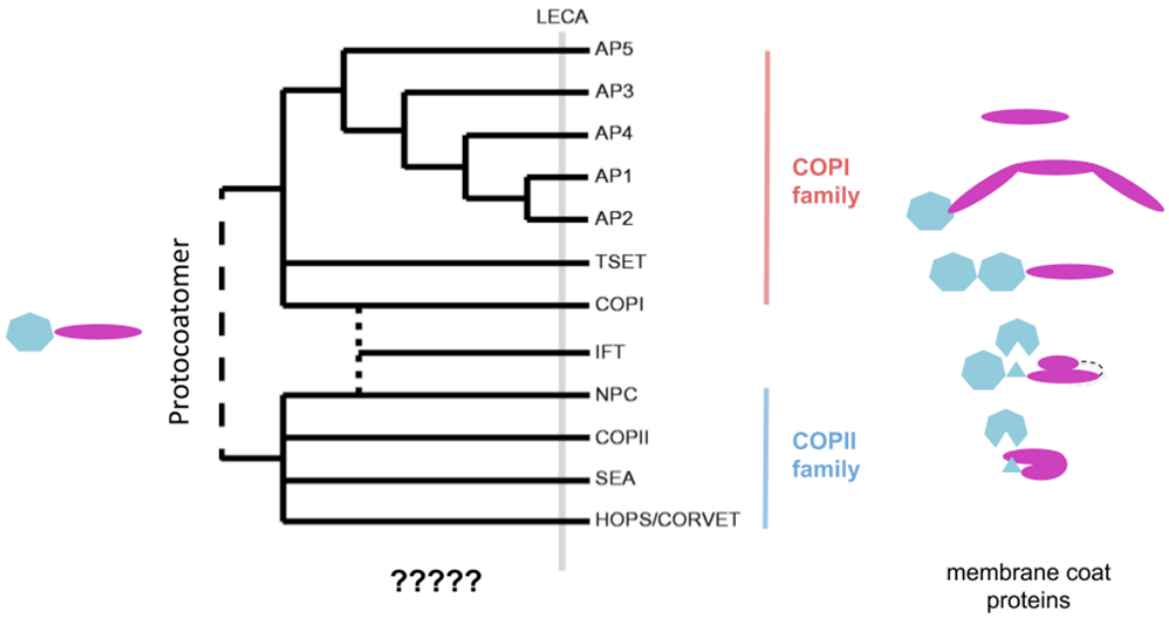


Figura 12. Modelo cualitativo de las relaciones entre los complejos de cubierta de membrana basado en sus características compartidas.
Los complejos COPII, SEA, HOPS/CORVET y el poro nuclear son agrupados en un mismo grupo (llamado familia de COPII) tomando en cuenta la presencia de proteínas Sec13 en estos complejos y la existencia de subunidades que comparten similitud estructural con la proteína Sec31. Mientras que los complejos COPI, TSET y las adaptinas constituyen la familia de la clase COPI debido a la forma tetramérica que comparten estos complejos. Las líneas punteadas en el dendrograma indican que existe cierta incertidumbre con respecto a la relación que poseen el complejo IFT con los demás complejos derivados del protocoatomer. Además, estas líneas indican que también hay incertidumbre acerca de cuál fue su raíz de origen. Esta imagen fue adaptada de la imagen publicada en el artículo de Schlacht and Dacks, 2015. Genome Biol. Evol.

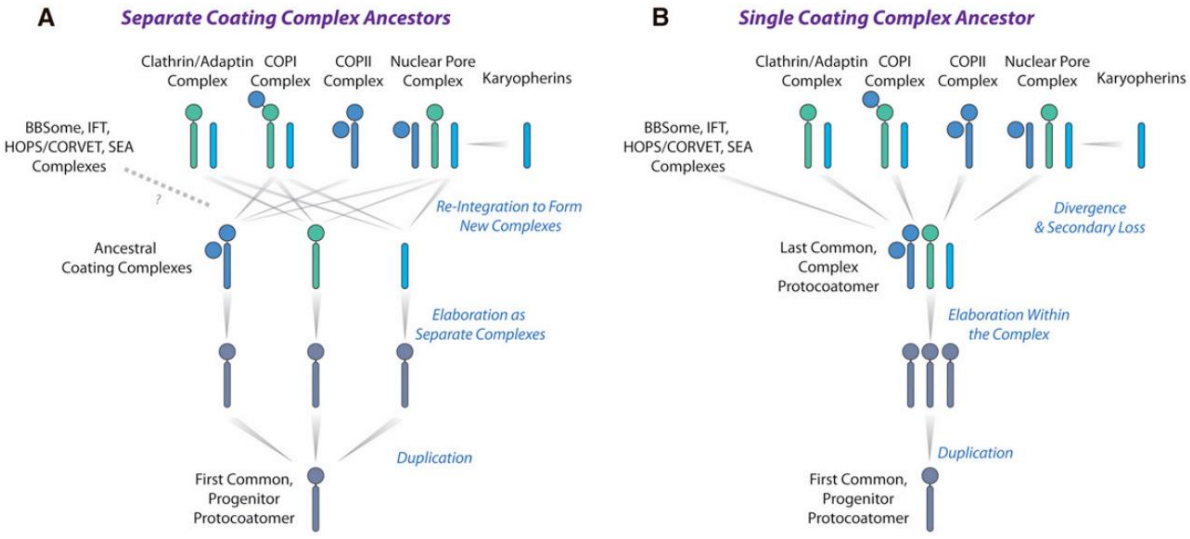


Figura 13. Relaciones evolutivas entre los complejos de cubierta de membrana.
Se describen dos posibles escenarios que son mostrados en A y B según las relaciones estructurales encontradas entre Nup192 con las familias de proteínas de las carioferinas, β-cateninas y adaptinas. Los dominios β-propeller y α-solenoide presentes en estas proteínas son representados con círculos y varillas, respectivamente. Esta figura fue obtenida Sampathkumar et al., 2013. Cell Press.

Los recientes descubrimientos acerca de los complejos de cubierta de membrana nos dan la oportunidad de considerar algunas de las relaciones evolutivas que existen entre los distintos complejos, así como sugerir los posibles pasos tempranos en la evolución del sistema de endomembranas.

La presencia de elementos comunes que forman parte de la composición de distintos complejos como el poro nuclear, COPII y SEA es un claro indicio de que comparten un ancestro en común. Por ejemplo, los β-propeller incompletos Sec13 y Seh1 interactúan con algunas nucleoporinas de complejo Y del poro nuclear (Sec13-Nup145C y Seh1-Nup85), con la proteína Sec31 del complejo COPII (Sec13-Sec31) y con las proteínas asociadas a los complejos SEA (Debler *et al.*, 2008; Algret *et al.*, 2014; Dokudovskaya *et al.*, 2011). Esto nos lleva a especular

sobre la cercana relación que poseen los complejos COPII, SEA y el poro nuclear debido a la presencia de componentes comunes que interactúan con su maquinaria como Sec13 y Seh1. También, la presencia de algunas Nups del poro nuclear, la proteína Vps39 del complejo HOPS y algunos componentes de SEA como SEA4 que poseen una arquitectura y estructura similar a Sec31 del complejo COPII, sugiriendo una relación evolutiva cercana entre estos complejos (Schlacht and Dacks, 2015). Recientemente, se ha descubierto evidencia acerca de la presencia de componentes compartidos de otros complejos asociados al sistema de endomembranas que interactúan con la envoltura nuclear, el retículo endoplasmático y la vacuola, por ejemplo, las proteínas Rab32, Sintaxina 17 y PACS-2 cuyas funciones se encuentran asociadas a estos organelos (Schlacht and Dacks, 2015). Todas estas características sugieren una tentativa conexión entre los complejos HOPS y SEA con el poro nuclear y con el complejo COPII.

También, la similitud a nivel de secuencia entre algunos de los componentes de los complejos COPI, clatrina/adaptinas y TSET nos da cuenta de una relación cercana entre estos complejos. La estructura tetramérica que poseen los complejos COPI, clatrina/adaptinas y TSET y la presencia de homología basada en secuencia entre los componentes del complejo adaptador de COPI (F-COP) con las adaptinas sugieren una conexión entre los complejos clasificados en la familia COPI (Field and Dacks, 2009). La conexión entre ambas familias está dada por la presencia de algunas proteínas de la misma familia de las GTPasas (como las proteínas Arf y Sar1) que son necesarias para el ensamblaje de los complejos COPI y COPII en las cubiertas de las vesículas (Field and Dacks, 2009). De la misma forma, la presencia de componentes del *scaffold* del poro nuclear que presentan una arquitectura similar a las proteínas de las familias COPI y COPII sugieren una conexión entre ambas familias (Field and Rout, 2019).

Sin embargo, la conexión que existe entre algunos complejos o grupos de proteínas con las familias del tipo COPI o COPII (o entre el tipo I y II) es incierta, considerando por ejemplo el complejo Intraflagelar y las carioferinas. Algunos estudios basados en análisis filogenéticos y estudios comparativos muestran conexiones entre algunos componentes del complejo IFT con algunas proteínas del complejo COPI (Dam *et al.*, 2013a). Sin embargo, otros estudios sugieren que existe una conexión funcional y evolutiva entre el complejo IFT con el poro nuclear, porque varios parásitos unicelulares poseen una estructura combinada entre parte del núcleo y los cilios llamada "*karyomastigont*", mientras que en las células de algunos mamíferos, las nucleoporinas forman un complejo en la base del cilio (Schlacht and Dacks, 2015; Kee *et al.*, 2012; Walker *et al.*, 2011).

En el caso de las carioferinas, en algunos modelos son clasificados en tipo I o COPI-like, mientras que en otros es agrupada dentro de un tercer tipo (del tipo adaptina o *adaptin-like*) (Sampathkumar *et al.*, 2013; Field and Rout, 2019). En un estudio se ha mostrado que las Nup192 y Nup188 comparten propiedades funcionales y estructurales con estas proteínas sugiriendo que existe una relación evolutiva entre las Nups estacionarias con la maquinaria de transporte nuclear soluble (Andersen *et al.*, 2013). Además, estas Nups y las carioferinas comparten una estructura similar a las adaptinas sugiriendo que estas proteínas también se originaron y evolucionaron conjuntamente para formar el poro nuclear (Field and Rout, 2019).

A pesar de toda la información disponible acerca de las características comunes presentes entre las proteínas de cubierta de membrana, aún es muy difícil determinar la raíz de su árbol evolutivo y los pasos intermedios que dieron origen mediante un enfoque cuantitativo. Esto se debe en parte a la divergencia observada tanto a nivel de secuencia como de estructura que existen entre estos complejos y, por otro lado, a la carencia de estructuras cristalográficas

para la mayoría de sus componentes. Además, las herramientas computacionales con las que contamos actualmente no nos permiten detectar similitudes significativas entre proteínas que han divergido considerablemente durante la evolución, como es el caso de las proteínas de cubierta de membrana. Debido a ello, ha permanecido elusivo inferir cuales son los posibles pasos que dieron origen a las proteínas de cubierta de membrana actuales y cuyo descubrimiento tiene el potencial de clarificar el orden en el cual evolucionó el sistema de endomembranas eucariótico. ¿Cuántas otras relaciones entre proteínas divergentes han permanecido indetectables debido a las limitaciones de las herramientas actuales?

Para comprender mejor las relaciones evolutivas existentes entre proteínas muy divergentes, hemos desarrollado una nueva herramienta bioinformática que permite una comparación flexible y precisa de proteínas (Gutiérrez *et al.*, 2016). Durante el transcurso del proyecto, esta herramienta ha sufrido de varias mejoras para comparar proteínas divergentes y determinar sus posibles relaciones evolutivas usando la información derivada principalmente de sus estructuras. Además, para llevar a cabo dichas comparaciones se tuvieron que afinar sus parámetros iniciales y se definieron nuevas estadísticas para evaluar la significancia de las similitudes estructurales reportadas. A partir de las similitudes encontradas, hemos propuesto un nuevo modelo cuantitativo de las relaciones encontradas entre los complejos MC con la finalidad de esclarecer los posibles pasos evolutivos que existen entre estos complejos, cuyos resultados se discutirán en detalle en este manuscrito. Finalmente, esperamos que la aplicación de esta nueva versión desarrollada de la herramienta de comparación estructural nos ayudará a comprender de mejor forma cómo las proteínas de cubierta de membrana contribuyeron en particular al surgimiento de la complejidad en el sistema de endomembranas actual en las células eucariotas y a la Eucariogénesis en general.

HIPÓTESIS Y OBJETIVOS

2.1. Hipótesis

En vista de los antecedentes señalados anteriormente, se propuso la siguiente hipótesis:

"El desarrollo de una nueva y eficiente herramienta para la comparación flexible de estructuras de proteínas nos permitirá obtener evidencia cuantitativa que apoye la teoría de que las proteínas de cubierta de membrana evolucionaron a partir de escenario parsimonioso del Protocoatomer ancestral".

2.2. Objetivo General

Desarrollar una herramienta bioinformática que nos permita evaluar las relaciones estructurales entre proteínas muy divergentes y utilizarla para validar e inferir nuevas relaciones entre las proteínas de cubierta de membrana eucariotas.

2.3. Objetivos Específicos

Este proyecto de tesis cuenta principalmente con 6 objetivos específicos los cuales se describen a continuación:

1. Desarrollar e implementar una herramienta bioinformática para la comparación eficiente y flexible de estructuras de proteínas.
2. Validar la utilidad de la nueva herramienta desarrollada para inferir relaciones evolutivas o de similitud estructural en ejemplos conocidos de proteínas relacionadas y no relacionadas.
   2.1. Identificar los apareamientos locales significativos a partir de la superposición estructural de un par de proteínas, mediante una serie de reglas establecidas a través del aprendizaje automático de un conjunto de datos de entrenamiento provenientes de proteínas homólogas y no homólogas.

2.2. Determinar la lista de combinaciones de apareamientos locales sin solapamiento.

2.3. Reportar la mejor combinación de apareamientos estructurales locales en una superposición global.

2.4. Establecer una métrica de similitud y metodología ad hoc para evaluar la significancia estadística de los alineamientos estructurales obtenidos con la herramienta.

3. Utilizar la nueva herramienta desarrollada para crear un nuevo esquema de clasificación para las proteínas de cubierta de membrana conocidas.
4. Desarrollar un modelo cuantitativo que explique las relaciones encontradas entre las proteínas de cubierta de membrana según las similitudes observadas.
5. Proponer los posibles pasos evolutivos entre las proteínas MC según las relaciones encontradas.
6. Publicar y liberar abiertamente la herramienta desarrollada para que los usuarios puedan detectar relaciones evolutivas entre proteínas divergentes.

METODOLOGÍA Y RESULTADOS

La metodología implementada y los resultados obtenidos de la ejecución de este proyecto son descritos a continuación en dos artículos científicos, uno publicado y otro en proceso de redacción, donde ambos se centran en la metodología desarrollada para explorar las relaciones estructurales entre proteínas relacionadas muy divergentes. Estos dos artículos son descritos en los dos primeros capítulos de esta tesis, abarcando así el cumplimiento de los objetivos 1 y 2 del proyecto. Los resultados de las comparaciones realizadas entre las proteínas de cubierta de membrana utilizando la herramienta desarrollada son descritos en un tercer y último capítulo de esta tesis, abarcando el cumplimiento de los objetivos 3 y 4 de este proyecto. Finalmente, en el último capítulo de esta sección se ejemplificará la funcionalidad y modo de uso de la nueva herramienta desarrollada para realizar comparaciones estructurales flexibles entre las proteínas, cumpliendo con el objetivo 5 del proyecto.

A continuación, cada capítulo de esta sección incluye un breve resumen introduciendo el contenido de cada artículo o trabajo en curso realizado durante el desarrollo del proyecto.

3.1. DESARROLLAR UNA HERRAMIENTA PARA LA COMPARACIÓN ESTRUCTURAL FLEXIBLE DE PROTEÍNAS DIVERGENTES

3.1.1. Resumen

Existe actualmente un gran número de herramientas disponibles para comparar las estructuras de las proteínas, donde cada una de ellas presentan ventajas y desventajas con respecto a las demás dependiendo para que propósito fueron diseñadas originalmente. En este capítulo se incluye un artículo publicado que muestra el desarrollo de un programa computacional para la comparación flexible y rápida de estructuras de proteínas que puede ser usado para derivar relaciones estructurales entre proteínas muy divergentes como las proteínas de cubierta de membrana. Este programa codifica la orientación espacial de los elementos de estructura secundaria que componen las estructuras de las proteínas en matrices 2D. Luego, esta aplicación mediante el alineamiento de las matrices puede detectar pares de sub-fragmentos comunes que pueden ser alineados mediante apareamientos locales para reportar una superposición flexible. Esto ofrece una enorme ventaja debido a que las comparaciones a nivel de matrices son rápidas permitiendo realizar búsquedas exhaustivas en un menor tiempo contra una base de datos de estructuras, y también permiten capturar el plegamiento de las proteínas pudiendo ser empleadas para clasificar proteínas dentro de una misma familia. La implementación de esta herramienta sentó las bases para el posterior desarrollo de una nueva metodología mucho más compleja que nos permitió derivar relaciones estructurales significativas. A continuación, se describe con mayor detalle las características principales de este método y su potencial para búsqueda de relaciones estructurales entre proteínas divergentes.

3.1.2. PAPER: "Efficient and automated large-scale detection of structural relationships in proteins with a flexible aligner"

Gutiérrez *et al. BMC Bioinformatics* (2016) 17:20
DOI 10.1186/s12859-015-0866-8

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Efficient and automated large-scale detection of structural relationships in proteins with a flexible aligner

Fernando I. Gutiérrez[1,2], Felipe Rodriguez-Valenzuela[1], Ignacio L. Ibarra[1,3], Damien P. Devos[2,3*] and Francisco Melo[1*]

Abstract

Background: The total number of known three-dimensional protein structures is rapidly increasing. Consequently, the need for fast structural search against complete databases without a significant loss of accuracy is increasingly demanding. Recently, TopSearch, an ultra-fast method for finding rigid structural relationships between a query structure and the complete Protein Data Bank (PDB), at the multi-chain level, has been released. However, comparable accurate flexible structural aligners to perform efficient whole database searches of multi-domain proteins are not yet available. The availability of such a tool is critical for a sustainable boosting of biological discovery.

Results: Here we report on the development of a new method for the fast and flexible comparison of protein structure chains. The method relies on the calculation of 2D matrices containing a description of the three-dimensional arrangement of secondary structure elements (angles and distances). The comparison involves the matching of an ensemble of substructures through a nested-two-steps dynamic programming algorithm. The unique features of this new approach are the integration and trade-off balancing of the following: 1) speed, 2) accuracy and 3) global and semiglobal flexible structure alignment by integration of local substructure matching. The comparison, and matching with competitive accuracy, of one medium sized (250-aa) query structure against the complete PDB database (216,322 protein chains) takes about 8 min using an average desktop computer. The method is at least 2–3 orders of magnitude faster than other tested tools with similar accuracy. We validate the performance of the method for fold and superfamily assignment in a large benchmark set of protein structures. We finally provide a series of examples to illustrate the usefulness of this method and its application in biological discovery.

Conclusions: The method is able to detect partial structure matching, rigid body shifts, conformational changes and tolerates substantial structural variation arising from insertions, deletions and sequence divergence, as well as structural convergence of unrelated proteins.

Keywords: Protein structure comparison, Protein structure search, Flexible structural alignment

* Correspondence: damienpdevos@gmail.com; fmelo@bio.puc.cl
[2]Centre for Organismal Studies (COS), Heidelberg University, Heidelberg, Germany
[1]Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Alameda 340, Santiago, Chile
Full list of author information is available at the end of the article

Background

Structural comparison between proteins is a fundamental and common practice in structural biology with many applications, such as the identification of new domains, the classification into structural families and the detection of evolutionary relationships between protein structures that cannot be found by sequence comparisons. For example, the homology between prokaryotic and eukaryotic cytoskeletal filaments (FtsZ/Tubulin and MreB/Actin) or the paralogy between proteins such as hemoglobin and myoglobin where only revealed once the 3D structures of these proteins were solved and compared [1, 2]. Since the determination of the first structures in the 1970s to the present day, the number of solved protein structures in the Protein Data Bank (PDB) has continued to grow at an exponential rate, with more than one hundred thousand structures available today. To facilitate the organization and analysis of this large amount of information, different structure comparison methods and tools have been developed [3]. However, the rise in number of known structures makes the comparison of query structures against the database increasingly costly (both for time and computational requirements) using existing tools.

Depending on the representation of proteins, current structural alignment methods use two main approaches: methods based at the level of residues or Cα atoms (DALI, Structal, TopMatch, MAMMOTH, CE, MUSTANG, FATCAT, TM-align) [4–11] or based on secondary structure representations (VAST, SSAP, GANGSTA+, QP tableau search) [12–15]. One of the major advantages of methods based on secondary structure representations is that they are generally faster, as there is typically at least one order of magnitude fewer secondary structure elements than residues within a protein. However, residue-based methods are generally more accurate [16].

Structure comparison methods are increasingly successful at detecting more divergent relationships [3]. Significant improvements have also been achieved in terms of speed when searching against large databases [17]. Despite this success, current structural comparison tools have a few major drawbacks that limit their utility for detecting cases of remote homology where protein structures might have diverged considerably. First, they treat proteins as rigid bodies and cannot accommodate the large structural variations observed over long evolutionary divergence, for example, the relationship between the nucleoporins and vesicle coats [18]. Additional structural variations that might be due to protein flexibility or allosteric transitions are difficult to detect with the current methods. Finally, they are usually restricted to the comparison of individual domains and do not consider multi-domain proteins. How many distant structural relationships remain undetected because the tools are not sensitive enough? Our goal was to detect protein structure similarities that are beyond the reach of current tools based on rigid body superposition and, at the same time, to be able to do it efficiently and with competitive accuracy.

To that end, we have developed an efficient flexible aligner tool to compare protein structures based on matrices that contain a simple description of the geometrical arrangement of secondary structure elements. Arthur Lesk was the first to describe a tabular representation, which comprises the information about the relative orientation of the elements of secondary structure (interaxial angle) using a coarse-grained and discrete double quadrant codification [19]. The concept is that the sequential order of secondary structure elements and the geometry of interacting pairs capture the essence of the protein fold. The secondary structure elements and their respective angles and distances can be encoded in a matrix. The secondary structure elements are recorded in order of appearance along the main diagonal of the matrix. Each off-diagonal position contains the angles and distances between the pairs of secondary structure elements. The comparison of these matrices allows a faster structural matching than when using a protein representation at the residue/atomic level. However, secondary structure geometry matrices comparison is an NP-hard problem. Various implementations to solve this problem have been presented, including quadratic and linear integer programming [15, 20, 21]. Those methods are very precise at extracting maximally similar submatrices, but this is at the expense of speed when comparing against a large number of matrices such as the complete PDB database. In 2008, Konagurthu proposed the TableauSearch method to detect similarities between matrices using two steps of dynamic programming [20]. TableauSearch is faster than previous methods, but this comes at the expense of accuracy and of lacking the ability to find local matches as compared to global ones [15]. This method is not limited to element pairs that are in contact and uses the scheme previously proposed by Lesk described above [19].

We present and release here a new computer application called MOMA (from MOrphing & MAtching). This tool relies on a new algorithm that incorporates several innovations, which are: 1) it considers the continuous value of the angles instead of the discrete and coarse-grained quadrant codification proposed by Lesk and implemented in TableauSearch; 2) the incorporation of a user-defined maximum distance cutoff to consider contacts between secondary structure elements, 3) a modified two-step dynamic programming algorithm that allows for the maximization of the rigid union of several local and compatible structural

matches and 4) a new procedure to solve the integration of several rigid and globally incompatible local matches into a flexible and global solution. This new algorithm, as implemented in MOMA computer application, results in a fully automated and highly efficient global flexible structural aligner, which is able to find structural similarity between distantly related proteins with high accuracy.

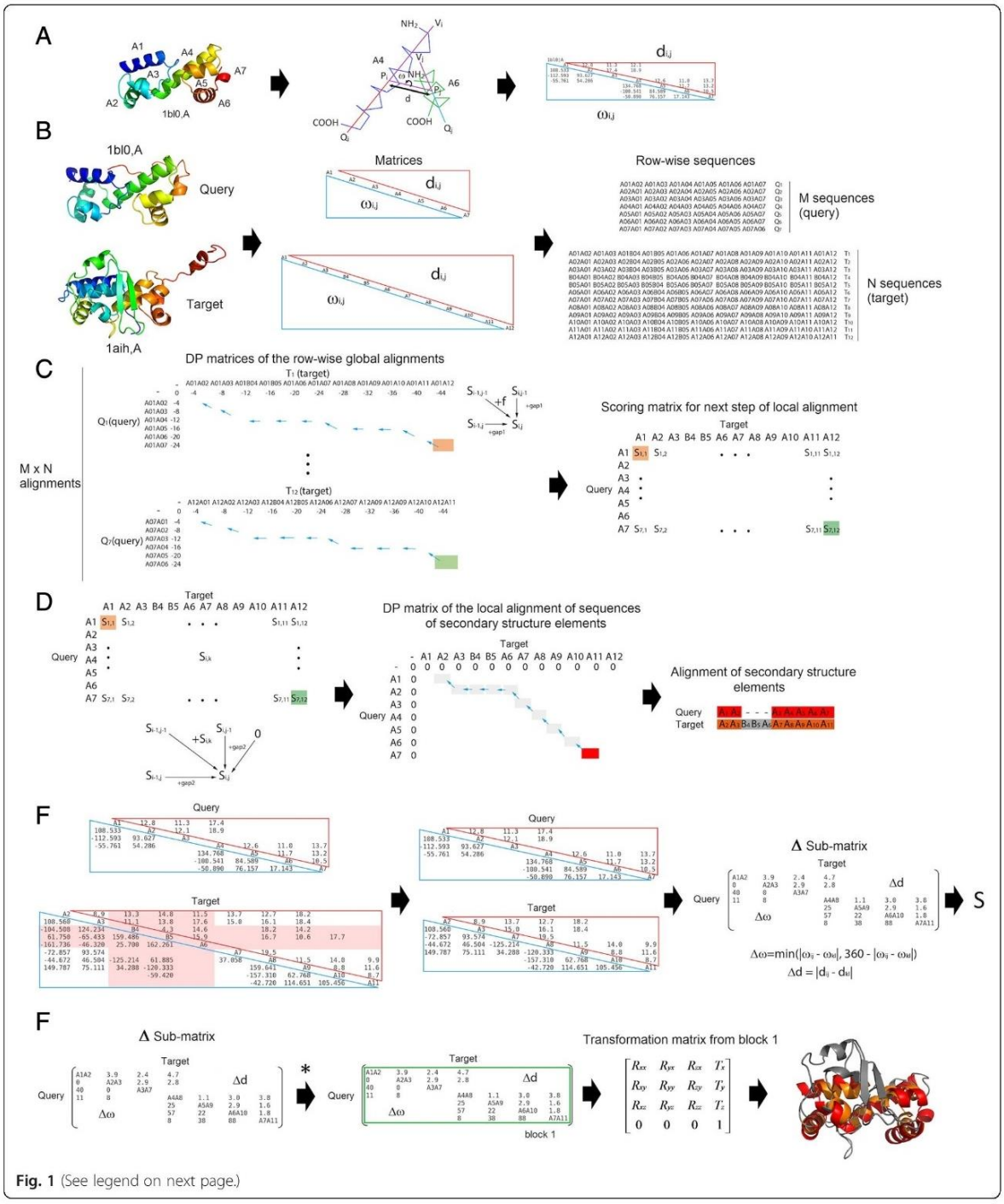


Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Flowchart of the method as implemented in MOMA. **a** Example of structure of MarA (PDB code 1bl0) and the matrix representation of its folding pattern. The relative orientation of any two secondary structural elements (for example, A4 and A6 helices) is specified by the angle (w) between the vectors along their axes (*left bottom of the matrix*). This is recorded only for those SSE pairs found in close proximity (d < D), as measured by the distance (d) between midpoints of the vectors (*upper right of the matrix*). **b** These matrices are built for the query (1BL0 chain A) and the target (1AIH chain A) structures. After that, row-wise matrices containing all possible SSE pairs in each structure are also built. Query and target proteins render matrices of [M, M-1] and [N, N-1] pairs, where M and N correspond to the total number of SSEs found in the query and target structures, respectively. **c** A first step of global or semi-global dynamic programming (DP) algorithm is executed to build DP matrices for each query row against each target row, thus generating a total of *MxN* DP matrices. In this step, scoring rules and restraints based on angular and distance information of all SSE pairs in each structure are used (see Methods for details). From each DP matrix, only the maximum score value is selected and recorded into a new scoring matrix that is going to be used in a second and final step of a dynamic programming algorithm. In the case of a global alignment, this value is obtained from the bottom right cell of the DP matrix. In the case of the semi-global alignment, this value is obtained from the most right column or the most bottom row of the DP matrix. **d** A local dynamic programming algorithm and the previously built scoring matrix are now used to align the secondary structure elements of the query and the target structures. **e** Unaligned SSE elements from the query and target structures are removed from the initial 2D matrices, thus rendering two matrices of identical dimensions, which can now be compared directly. A delta sub-matrix is built and from it a global matching score calculated (see Methods for details). **f** Finally, a new algorithm (*) is used to infer the list of all incompatible rigid local matches (*blocks*), which are independently superposed with the Kabsch algorithm. In this particular and simple example only one local match or block is found. Details of the algorithm for finding local matching blocks are provided as Supplemental Material (Additional file 1: Figure S1). The resulting superposition is represented with aligned elements in red (*query*) and orange (*target*). Residues not aligned are displayed in grey color

Results and discussion

Overview of the new method

This article describes a fully automated and highly efficient method for the flexible comparison of two protein structure chains. The method relies on the matching of secondary structure elements between the protein chains, based on a two-step dynamic programming algorithm that combines local and global matching procedures. The results obtained when applying this method consist on a single and global structural alignment that integrates all rigid local matches found between the two input protein structures. A general overview of the method is provided in Fig. 1. A detailed description of each step of the method is provided in Methods section.

Calibration of parameter values

The results of our method, as implemented in MOMA, strongly depend on the value of three parameters, which are the constant that limits the score calculated from the angular difference (C) and the gap-opening penalties for the two steps of dynamic programming (g1 and g2). By optimization of the different combinations of these parameters, we found that the best results were obtained with a C constant value of 45 and a gap-opening penalty of −4 for both steps of dynamic programming (Additional file 1: Table S1). With these parameter values, only 2 out of 100 alignments from HOMSTRAD database have a QS index smaller or equal to 0.5 and the average QS index was 0.9436 (Additional file 1: Table S2). The failure of MOMA to correctly align the corresponding SSE pairs in these two cases is due to an inaccurate assignment of secondary structure elements by DSSP computer program. In some cases, DSSP does not assign the exact start and end points of SSEs. In other cases, long helices and strands with some bending are split into two or more non-contiguous SSEs [21].

Another relevant parameter in the matrix comparison step of our method is the distance cutoff (D) used to define SSE pairs in contact [15]. We tested different values of distance thresholds in the HOMSTRAD set to define the best performing one (Additional file 1: Figure S1 and Table S1). If the distance cutoff value was smaller or equal than 12 Å, several matrices could not be aligned because too few SSE pairs were considered (ie. few contacts are found near the main diagonal of the matrix). Most of the information required to identify a folding pattern is contained in adjacent positions near the main diagonal in the matrices [22].

On the other hand, if the distance cutoff was set to values greater than 20 Å, the average QS index decreased (Additional file 1: Table S1). Therefore, a value of 20 Å was finally used as the maximum distance cutoff to define a contact between two SSEs.

After fixing the previous parameter values, and to evaluate if the raw score reported by MOMA was better than the relative similarity score, we then carried out searches using the seven most common folds as a query against a subset of 19,602 domains from ASTRAL 2.03 (95% sequence-identity cutoff; for details see Methods). The ROC curve analysis of these two scores showed that the relative similarity is slightly better than the raw score (Additional file 1: Figure S2 and Table S3). Thus, we defined relative similarity as the measure to be used for fold assignment by default in our method, as implemented in MOMA.

Testing the new method

As a first test of our method with the fixed parameter values described above, we used as a query the seven

most common folds and searched against the 19,602 domains in ASTRAL 95 % sequence identity dataset. ROC analysis of structure similarity matching results shows that, irrespectively of the query, the method has an excellent performance in terms of accuracy at the fold, family and superfamily levels (Additional file 1: Table S3). Execution time increases exponentially with the total number of SSE elements assigned in the structures (Additional file 1: Figure S4).

Benchmarking with other methods

The representative set of 100 protein queries was compared against the ASTRAL 2.03 40 % sequence identity dataset (which contains a total of 11,121 domains; for details see Methods) with SHEBA, YAKUSA, QP tableau search, GANGSTA+, Structal, TopMatch and MOMA computer programs. The performance of these methods was assessed by ROC curve analysis based on the normalized scores reported by each of them and adopting the SCOP classification as the gold standard [23]. We also measured the execution time required by these computer programs to perform a search against the full ASTRAL dataset of 11,121 domains with the 100 query structures.

In terms of AUC and maximum accuracy values, both at the fold and superfamily levels, Structal, TopMatch and MOMA are the best performing methods, followed by GANGSTA+, QP tableau search, SHEBA and Yakusa (Table 1; Additional file 1: Figure S3). In terms of accuracy, at the fold and superfamily levels, MOMA has the best performance among methods that use a geometric secondary structure representation of 3D protein structure such as QP tableau search and GANGSTA+, or when compared to currently the fastest methods for 3D structure matching such as YAKUSA and SHEBA. MOMA requires a variable amount of time to complete the search, which depends on the number of SSEs present in the matrix (Additional file 1: Figure S4), but in this large benchmark set MOMA is much faster than all tested methods (at least by one or two-three orders of magnitude faster than most of the tested methods) (Table 1).

A detailed analysis of ROC curves reveals that SHEBA is a more specific classifier than MOMA, GANGSTA+ and QP tableau search, exhibiting a very low rate of false positives at the fold and superfamily levels. However, these methods have a higher sensitivity when compared to SHEBA. GANGSTA+ has an excellent performance and is better than QP tableau to search for proteins with the same fold, but QP tableau search is better than GANSGTA+ at a rate of false positives >0.6 for the superfamily level.

At the fold level, Yakusa is always worst than SHEBA, QP tableau search, GANGSTA+ and MOMA. However, Yakusa has a slight advantage than SHEBA at a rate of false positives >0.5 for the superfamily level.

The statistical analysis of the AUC curves reveals that the difference observed in the performance of MOMA with other computer programs is statistical significant at the 95 % confidence level (Additional file 1: Table S4).

As for the running time of each method, MOMA is the fastest of the methods tested. For example, it takes only 8 min and 28 s to search the 100 queries against the whole ASTRAL 40 %, while all other methods take more than 45 min, hours or even days of execution time (Table 1). We note that Structal, GANGSTA+, QP tableau search, and SHEBA are infeasible to run queries on very large datasets, such as the PDB database, which was one of the goals that motivated us to develop this method. Although QP tableau search can calculate the exact solution of the comparison of two matrices and GANGSTA+ can generate non-sequential protein structure alignments based in SSEs, MOMA has a better performance and is much faster than these two methods.

Table 1 Performance benchmark analysis of MOMA with different methods

Methods	AUC		ACC		*fp		*tp		time
	Fold	Superfamily	Fold	Superfamily	Fold	Superfamily	Fold	Superfamily	
Structal	0.956	0.969	0.902	0.919	0.076	0.060	0.880	0.898	10d 21h (1,842x)
TopMatch	0.955	0.974	0.883	0.911	0.121	0.069	0.887	0.891	2d (339x)
MOMA	0.940	0.956	0.872	0.889	0.139	0.113	0.884	0.891	8m 28s (1x)
GANGSTA+	0.916	0.911	0.845	0.851	0.101	0.058	0.791	0.761	5d 6h 49m (895x)
QP tableau search	0.877	0.918	0.791	0.831	0.224	0.188	0.805	0.850	2d 7h 27m (391x)
SHEBA	0.870	0.889	0.841	0.875	0.052	0.042	0.734	0.793	6h 51m (48x)
FATCAT flexible	0.837	0.911	0.743	0.825	0.220	0.211	0.706	0.862	27d 2h 38m (4,614x)
YAKUSA	0.790	0.858	0.727	0.794	0.155	0.088	0.609	0.677	48m (5.7x)

Area under ROC curve (AUC), maximal accuracy (ACC), false positive (fp) and true positive (tp) rates for each method are reported (*these values are calculated at the same threshold that gives the maximum accuracy reported as ACC). The execution time needed to compare the 100 queries against the 11,121 domains in the ASTRAL SCOP 40% sequence identity dataset is shown in the last column of the table. Execution times are reported in seconds (s), minutes (m), hours (h) and days (d) (in parenthesis, the speed gain factor of MOMA when compared to other methods is displayed, where "x" means number of times faster)

Biological applications

Rigid body shift caused by a rearrangement of domains

A well-known case that illustrates an example of rigid body movement between two structural domains is provided by the comparison of structures of calmodulin with and without Ca^{2+} ion (PDB codes and 2bbm and 1cfc, respectively). Both structures have 4 EF-hands, which consist of a helix-loop-helix motif that interact with $Ca^{2+}$and are organized into two distinct globular domains (N-terminal and C-terminal domains) [24]. These two domains are connected by a linker that is unstructured. This specific case is difficult to align due to the flexibility of the 6 loops and of the central linker. In the calmodulin-Ca^{2+}structure, the two calcium-binding domains are wrapped around a binding peptide in a "close" conformation while in the Ca^{2+}free structure, a rotation around the axis of the linker leaves the two domains in an "open" conformation. Other flexible aligners such as Flexprot [25] and FATCAT [10], required the introduction of four or more rigid-body movements (twists) around pivot points (hinges) to obtain a good superposition of these two structures. In a single step, MOMA is able to automatically detect the conserved N-terminal and C-terminal domains, as shown in the matrix alignment, despite the different relative orientation of the two domains (Fig. 2).

Simple but significant structural rearrangement

The case of two functionally unrelated proteins illustrates the capacity of MOMA to obtain the global alignment of two structurally similar domains whose relative orientation is not conserved. The putative oxidoreductase from *Pseudomonas putida* (PDB code 3l6d) and the human Cytokine-Like Nuclear Factor N-Pac (PDB code 2uyy) share two almost identical structural domains, which are separated by a connecting linker (Fig. 3). This linker is composed by two or three helices in bacterial and human proteins, respectively. The differential number of helices present in the linkers orients the two domains differently in the bacteria and human proteins. This simple structural rearrangement is a challenging problem for structural similarity detection methods, because the orientation of the two domains is different in both proteins. Rigid structure comparison tools can only identify the matching of these domains as two separate solutions, in the rare cases where more than one solution is reported (ie. TopMatch).

The power of MOMA resides in the fact that the structural similarity between both structural domain pairs is automatically detected and reported in a single step. In addition, the source of the conformational difference is also readily detected and highlighted in the alignment matrix (ie. helix 20 of 2uyyA cannot be aligned to a missing corresponding helix in 3l6dA).

Complex structural rearrangement

A more impressive example of structural rearrangement detection occurs in the case of Sec31 subunit from the COPII coated vesicle complex and the nucleoporin Nic96. Despite a lack of detectable sequence similarity [26], it is now generally accepted that coated vesicles proteins and nucleoporins have a common origin [18, 26].

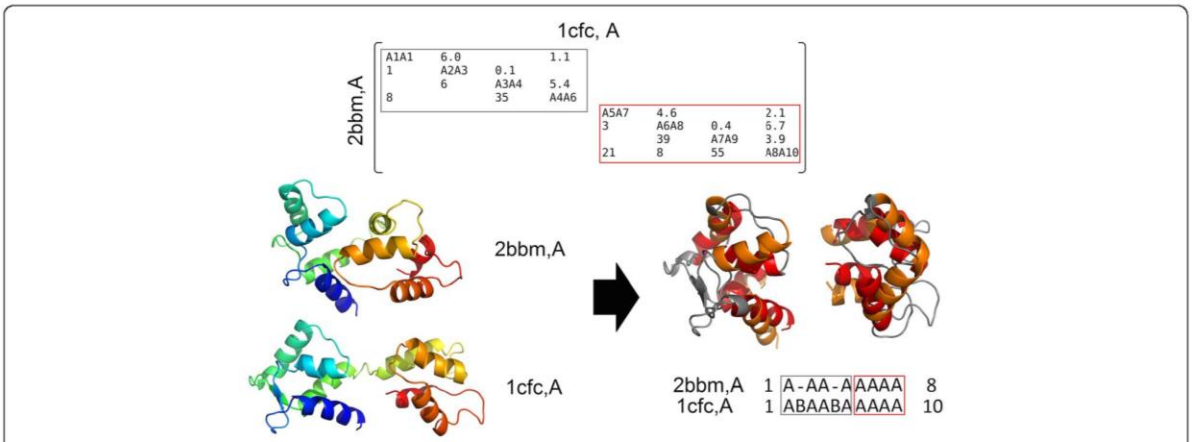


Fig. 2 Example of a rigid body shift caused by the rearrangement of two structural domains upon ligand binding. Structure superposition generated with MOMA of the Calmodulin-target peptide complex (the query; PDB code 2BBM, chain A) and the calcium-free Calmodulin (the target; PDB code 1CFC, chain A). Top Panel: The conserved domains are shown in the alignment of SSEs and the respective sub-matrices surrounded by grey and red blocks. Bottom Panel: The two structures and the superposition of their aligned domain pairs are shown respectively in rainbow color representation (*Left*) and with the SSE pairs structurally aligned in red and orange colours (*Right*). Non-aligned residues are shown in grey. The alignment of SSE elements is also represented with aligned blocks highlighted (*Bottom Right*). The structural superposition of these two domain pairs required different rotation matrices and translation vectors. In this example MOMA was executed with the following parameters: g1 = −4, g2 = −4 and C = 90

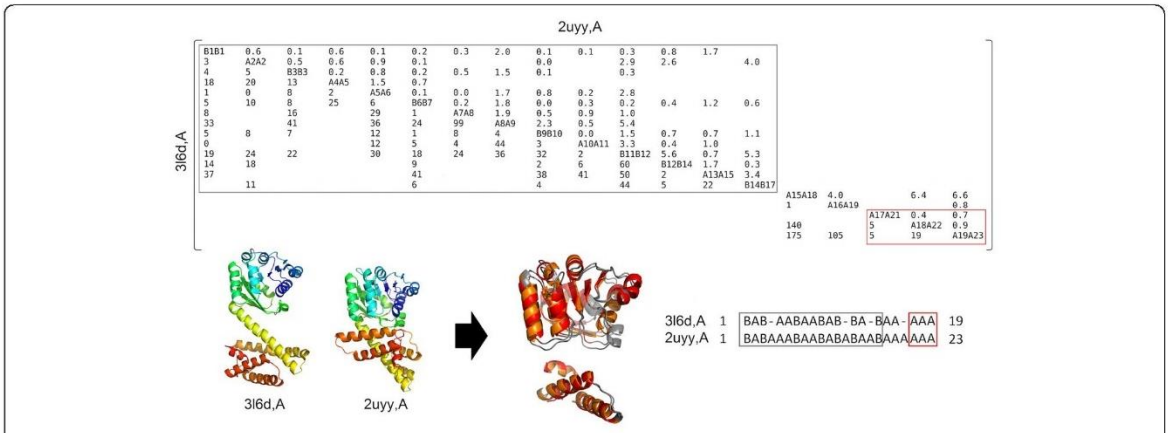


Fig. 3 Example of a simple but significant structural re-arrangement. Structure superposition of the putative oxidoreductase from Pseudomonas putida (the query; PDB code 3L6D, chain A) and the human Cytokine-like Nuclear Factor N-Pac (the target; PDB code 2UYY, chain A) generated with MOMA. The conserved domains are shown in the alignment and the sub-matrices remarked by grey and red blocks (*Top Panel*). The two structures are rainbow colored (*Bottom Left*) and the resulting SSE pairs aligned are shown in red and orange colours (*Bottom Middle*). The alignment of SSE elements is also represented with aligned blocks highlighted (*Bottom Right*). Each superposition was carried out with different rotation matrices and translation vectors. In this example MOMA was executed with the following parameters: g1 = −4, g2 = −4 and C = 45

However, considerable divergence has occurred since the event of gene duplication, up to a point that sequence similarity cannot be detected any longer, even by the most recent and powerful methods [27]. This sequence divergence has had important consequences on the structural conformation, interactions and cages formed in these two proteins [27]. This is the type of structural divergence that we aimed to detect efficiently and automatically, and thus the main motivation behind the development of the new method reported here. Nic96 (PDB code 2qx5) and Sec31 (PDB code 2pm7) are mainly composed of pairs of α-helices that are stacked on each other, hence termed SPAH domain (for Stacked Pairs of Alpha-Helices; also referred to as α-solenoid domain). Both proteins adopt a roughly linear shape that can be divided into three sections of conserved local structure (Fig. 4). However, those three conserved sections are preceded, followed and separated by other sections that exhibit considerable structural

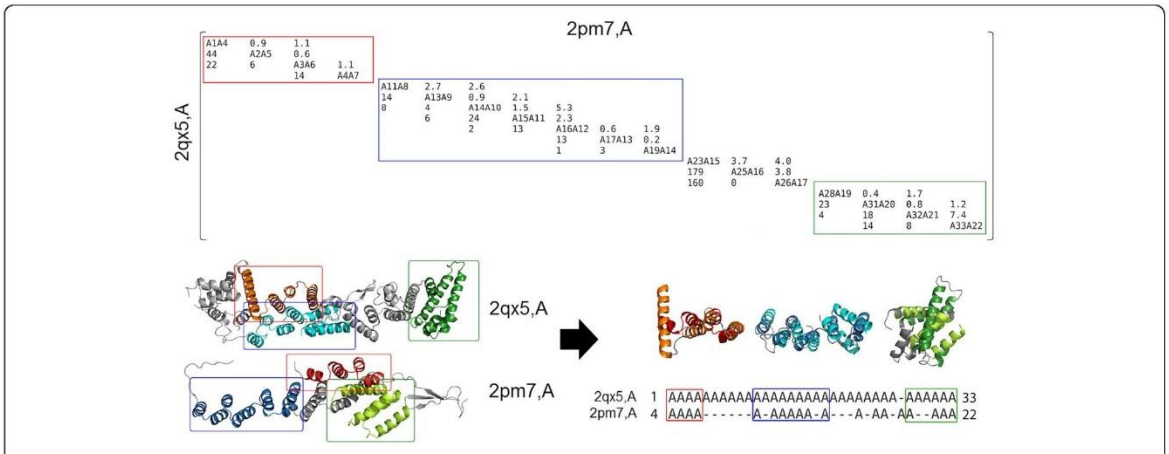


Fig. 4 Example of a complex structural re-arrangement. Structural superposition with MOMA of proteins Sec31 of the COPII complex of coat vesicle (the query; PDB code 2QX5, chain A) and the nucleoporin Nic96 (the target; PDB code 2PM7, chain A). Top Panel: The structure matching domains are shown in the alignment of SSEs and the respective sub-matrices surrounded by red, blue and green colored boxes. Bottom Panel: The three independent protein structure sectors that match between the two proteins, identified by MOMA, are highlighted by green, orange and blue colored boxes (*Bottom Left*). The alignment of SSE pairs and the independent structural superposition of the three matching block pairs are shown in red-orange, cyan-blue and light–dark green colours (*Bottom Right*). In this example MOMA was executed with the following parameters: g1 = −5, g2 = −5 and C = 45

deviation. Sections 1 and 2 are separated by a compact globular U-turn in Nic96, while this linker is unstructured in Sec31. The linker between sections 2 and 3 is composed by 9 α-helices in Nic96, but only by 3 α-helices in Sec31. These substantial structural modifications imply that section 1 is interacting only with section 2 in Nic96, while in Sec31 section 1 interacts almost exclusively with section 3. The relative orientation between the three blocks is also very different in both proteins. Despite these considerable global structural differences, the local structural similarity of the three blocks is clear and represents a legacy of their common ancestry [26]. To the best of our knowledge, MOMA is the only existing tool that is able to readily detect this intricate structural conservation in an automated fashion, which was the initial motivation of this work. The result obtained for this example case with MOMA clearly illustrates the power and potential for biological discovery of the new method reported here.

Strengths and weaknesses of the method

The speed, accuracy and flexible alignment capability of the method described here are their distinctive strengths. The method, as implemented in MOMA computer tool, is able to detect distant structural relationships in proteins in an automated fashion and efficiently, which makes it suitable to search the complete PDB for biological discovery. Among the weaknesses is the fact that MOMA is a single chain and topology-dependent protein structure alignment tool (ie. it depends on the connectivity order of SSEs). Few other tools, such as TopMatch and Structal have the capability of aligning protein structures in a topology-independent manner, but this comes at the cost of a longer execution time (these computer programs are 2–3 orders of magnitude slower than MOMA). TopMatch is the only tool currently available that is capable of aligning multiple protein chains, but the alignments are rigid and not flexible, which is a drawback in order to find domain movements or significant structural re-arrangements as exemplified here.

Structal was the most accurate tool in our benchmark (Table 1; Additional file 1: Table S4). A detailed analysis of the benchmark differences observed between Structal and MOMA shows that out of the 4,340 and 3,882 positive cases reported by Structal and MOMA, respectively, a total of 3,618 positive cases are common to both methods. There are 722 and 264 positive cases reported only by Structal and MOMA, respectively. Out of the 722 positive cases that Structal reports and MOMA fails to detect, 36.1 % is because of topological re-arrangements and 16.7 % is because there are too short or very few SSEs in the structures. In 11.1% of the cases, MOMA fails to detect the positive cases because of large differences in secondary structure definitions between the target and the query structures. It is noteworthy to mention that the use of STRIDE [28] or DSSP to assign SSEs produced, in a general basis, no significant difference on the performance of MOMA in our benchmark test (Additional file 1: Table S5). However, the accuracy of our method does depend directly on the assignment of SSEs, as well as on its use to represent protein structures and on its intrinsic topology-dependency. On the other hand, this simplified representation translates into a significant gain of speed without an important loss in accuracy (MOMA was 1,842 times faster than Structal in our benchmark test, but only 3 % less accurate). Finally, it is important to mention that the method described here produces structural alignments of secondary structure elements and not structural alignments at the residue-level. Therefore, if required, MOMA could be used in a first stage for fast database search on the task of fold or superfamily assignment and then, afterwards and only for positive matches, a more sophisticated software tool able to incorporate topology re-arrangements and to provide residue-level structure alignment, could be executed in a nested and sequential manner.

It is noteworthy to mention that this new method is not only restricted to protein structure comparison and could be implemented for many other applications that require the maximization of global shape matching between two three-dimensional objects with significant conformational variation, provided that those objects can be represented with vectors of different types which are relevant to describe the shape of the object, but with the limitation that vector order is a constraint of the method (ie. the method is topology-dependent).

Conclusions

We have developed a new structural comparison algorithm based on the spatial arrangement of secondary structure elements and shown that it allows the efficient retrieval of similar folding patterns in database searches. MOMA exhibits a high sensitivity to detect distant structural similarities without compromising its performance at identifying proteins that share a common fold.

In this regard, the development of a new combined global/semi-global and local structural alignment method that relies on a two-level nested dynamic programming algorithm and involves a new scoring scheme based on the continuous angular difference of SSE pairs close in 3D space instead of the previously used discrete quadrant codification, significantly improved the accuracy to find global similarities based on local matches in protein structures.

Methods

Protein structure and benchmark datasets

We used different protein structure datasets to first optimize the value of some parameters and then to evaluate the implementation of our method. First, to calibrate internal parameter values of the program, we used a subset of 100 pairwise structural alignments obtained from HOMSTRAD database [29] as previously described [30]. We kept only those alignments with a percentage sequence identity equal or less than 25 % and an average sequence length equal or greater than 150 residues (Additional file 1: Table S6). In this calibration process, a measure of similarity (the QS index) was maximized (see below). Second, to define the similarity score used and reported by our method, we used a small set of seven protein structures that represent the most common folds according to TOPS database [15, 20]. These seven proteins were used as a query to search against the ASTRAL SCOPe 2.03 95 % sequence identity protein domain database that contains 19,602 entries [31] (released October 2013). Receiver operating characteristic (ROC) curve analysis was performed and the area under the curve (AUC) measure was used to define the best performing score for classifying at the fold, family and superfamily level the query structures (see below).

Finally, to evaluate the performance of MOMA and other methods at classifying protein structures at the fold and superfamily levels, we used a representative set of 100 proteins extracted from the ASTRAL SCOPe 2.03 95 % sequence identity protein domain database described above (19,602 entries). These 100 proteins were used as a query to search for common structural matches against a non-redundant subset obtained from ASTRAL SCOPe 2.03 protein domain database [31] (released October 2013) with a 40 % sequence identity cutoff, which contains a total of 11,121 entries, none of them being any of the 100 query proteins. In this benchmark, we also carried out ROC curve analysis to assess and compare the performance of the methods (see below). All datasets described in this paper are available as supplementary data at: http://melolab.org/supdat/moma.

Computer software and methods

We used the DSSP program [32] to assign the secondary structure of proteins and the Numpy Python library to calculate the vectors and interaxial angles between the secondary structure elements. Moreover, we evaluated and compared MOMA against six methods based on their performance at classifying protein structures with similar folds or belonging to the same superfamily. The tested software implementing different methods were TopMatch [6], SHEBA [33], Yakusa [34], QP tableau search [15], Structal [5, 35], FATCAT [10] and GANGSTA+ [14]. These computer programs were used with their default parameter values. All calculations were carried out using an Intel Core i7 2.64 GHz processor with 12 GB RAM memory and Ubuntu 13.04 Linux operating system.

Method description

To construct a 2D matrix from the 3D structure of a protein, the secondary structural elements (SSE) are assigned with the DSSP program, version 2.0.4 [32]. Only α-helices and β-strands with more than four and three residues, respectively, are considered in the analysis. Different types of α-helices (π, 3_{10} and α) are treated equivalently and always assigned as a common α-helix type. Next, each secondary structure element is represented as a vector from its amino to carboxyl terminus by linear square fitting of an axis through the Cα coordinates with the singular value decomposition method [36].

After that, the interaxial angle between each pair of SSE vectors and the Euclidean distance between the midpoints of the axes is computed (Fig. 1a). The interaxial angle (ω) is the shortest rotation (clockwise or anticlockwise) required for the reorientation of the nearest vector that eclipses the farther vector, its value is restricted between −180° and 180° and was calculated as previously described [21]. Finally, the angle and distance between each pair of SSEs are recorded in the two halves of a 2D matrix: 1) the angle half-matrix and 2) the distance half-matrix. Two SSEs are only considered to be in contact if the distance between the midpoints of their linear axes is below a user-defined cutoff (see below). The diagonal positions are labeled by the elements of secondary structure, numbered by order of appearance in the amino acid sequence, from NH2 to COOH terminus (where 'A' stands for α-helix and 'B' for β-strand). All off-diagonal positions in the matrix are either blank, if the SSE pairs are not in contact, or they contain the observed angle or distance value of the corresponding SSE pair (Fig. 1a).

To compare 2D matrices of different size, we implemented a different method than that of TableauSearch [20] for submatrix matching. Our method aligns the two matrices with a nested dynamic programming algorithm. The first step of the method is aimed at discovering putatively equivalent SSE pairs by comparing each row in the query matrix with each row in the target matrix, with a global or semi-global alignment and a constant gap opening penalty value model (denominated g1). The rows are treated as linear sequences of SSE pairs (Fig. 1b). Therefore, each element in a row represents a pair of different SSEs in a protein. If the query and target structures contain M and N elements of secondary structure, then a total of M and N rows are generated

from the query and target structures, respectively. Consequently, in this step of the method, a total of *MxN* global or semi-global alignments are calculated (Fig. 1c).

Semi-global alignment is similar to global alignment, in the sense that it attempts to align the two sequences entirely. The difference between both methods lies in the way the alignments are scored. Semi-global alignment assigns no cost to opening end gaps in the alignment [37]. This alignment type selection depends on the difference in the number of SSEs identified in the query and target structures (ie. the size difference of the matrices). If the maximum ratio of the number of SSEs from the two structures is greater than two, a semi-global alignment is calculated; otherwise, a global alignment is built. We defined a scoring function that takes into account the value of interaxial angle (in degrees) calculated for each pair of SSEs, implicitly incorporating the distance between the two vectors. This function was defined as follows:

$$f(\omega_{ij},,\omega_{kl},,d_{ij},,d_{kl},,E_i,,E_j,,E_k,,E_l) = \begin{cases} 0,\ d_{ij} > D \text{ or } d_{kl} > D \\ -C,\ E_iE_j \neq E_kE_l \\ -C,\ \Delta\omega > 2C \\ C-\Delta\omega,\ \text{otherwise} \end{cases} \quad (1)$$

$$\Delta\omega = \min\left(\left|\omega_{ij}-\omega_{kl}\right|, 360-\left|\omega_{ij}-\omega_{kl}\right|\right) \quad (2)$$

where E_x stands for an element of secondary structure in relative position x from NH2 to COOH terminus in the protein chain, which can adopt two possible labels or values: A for alpha helix and B for beta strand; E_iE_j and E_kE_l are SSE pairs in the query and target structure, respectively; ω_{ij} and ω_{kl} are the interaxial angles between the E_iE_j pair in the query structure and between the E_kE_l pair in the target structure, respectively; d_{ij} and d_{kl} are the distances between the E_iE_j pair in the query structure and between the E_kE_l pair in the target structure, respectively; $\Delta\omega$ is the minimal angular difference between ω_{ij} and ω_{kl}, and C is an angular constant (in degree units). D is the maximum distance allowed to define that two SSEs are in contact (in Angstroms). This function is subjected to several constraints. The first constraint, $d_{ij} < D$ and $d_{kl} < D$, is introduced in order to avoid false positives when pairs of SSEs in two proteins have a similar interaxial angle, but are found at very different distances in the two structures [15] or found at very large distances in both the query and target structures. It is expected than in these cases there is no direct association between the SSE pairs in the two structures that should be used to infer fold similarity. This restriction is applied if at least one of the pairs is not in contact, as defined by the maximal distance cutoff D (a user-defined parameter). The second constraint, $E_iE_j = E_kE_l$, ensures that two SSE pairs of different types should not be matched (for example, helix-helix with strand-helix or with strand-strand) and the third constraint, $\Delta\omega < 2C$, ensures that the function takes values between C and -C. Finally, the adopted constant gap opening penalty values for the two levels of the dynamic programming algorithm were those resulting from an optimization process using one of the benchmark datasets (see section 2.6 below and Additional file 1).

The optimal score value obtained from each query and target row alignment (Fig. 1c) is taken to generate the scoring matrix that is used in the second alignment step (Fig. 1d), but this time with the local Smith-Waterman dynamic programming algorithm [38]. Here, a different constant gap opening penalty value can be adopted (denominated g2), which is another user-defined parameter required by our method. The alignment of SSE elements between the query and target structures is generated by the usual backtracking procedure (Fig. 1d).

At this point, it is important to mention that this alignment contains the union of all local structurally matching SSEs between the query and target structures, concordant to optimized, but not yet integrated global information of structurally matching SSE pairs. Therefore, the current alignment cannot be directly interpreted as a global structure alignment of two rigid bodies. In the case of highly related proteins this alignment will be accurate, but in the case of proteins with domain movements, rigid body shifts or partial structure matching, the identification of the structural regions to be matched as rigid body shifts by unique geometrical transformations is still needed.

The next step of the method consists on removing all rows and columns corresponding to non-aligned SSEs from both 2D initial matrices, the query and the target, thus rendering two matrices of identical size and shape that can be now compared directly and efficiently, in a one-to-one cell-to-cell manner (Fig. 1e). A unique 2D difference sub-matrix is now produced (called ΔSM or delta sub-matrix), which contains in the diagonal the labels for only those matching SSE pairs between the query and target structure, along with their differences in angle (upper middle triangle) and distance (lower triangle). Only the difference values for SSE pairs below a maximum parameter value, named ΔD, are reported in this difference matrix.

Structural matching score and similarity measures

A score of overall and integrated structural similarity for the query and target structures is calculated from the 2D difference sub-matrix (Fig. 1e). This score represents an estimation of the global integration of local matches. We calculate a measure of integrated structural similarity based on a Gaussian function that considers the angular

difference observed in the matrix. This raw score can be defined as:

$$S = \sum_{i}^{N} e^{-r_i^2/\sigma^2},\ r_i^2 = \Delta\omega^2 \quad (3)$$

where r_i^2 is the squared angular difference observed between two SSE pairs below distance threshold D, N is the total number of the SSE pairs aligned and σ is the scale parameter that determines the reduction rate of the score as a function of increasing angular difference. If the target structure is structurally equivalent with the query structure (ie. similar matrices), the score is equal to the total number of SSE pairs aligned. With increasing spatial deviation of the angular difference of SSE pairs aligned, the score approaches to 0.

In addition to score S, for comparing proteins of different size, we implemented two normalization functions. One of these functions is the relative similarity, S_r [39], which constitutes a global similarity measure between two proteins, and is defined by:

$$S_r = 100 \times \frac{2S}{n_q + n_t} \quad (4)$$

where n_q and n_t are the number of SSE pairs that are in contact in the query and target matrices, respectively, and S is the raw score described above. Another normalization function is the relative cover C_r [30] which represents the cover of the structural match in the smallest protein with respect to the largest protein [39], and it was implemented in the following function:

$$C_r = \frac{100 \times S}{\min(n_q, , n_t)} \quad (5)$$

The integration of the information from all these score similarity measures allows the detailed assessment of structure similarity between two protein chains, from a local and global perspective, at once.

Inference of compatible local structural matching

To obtain a flexible and global superposition of two structures, a complete list of rigid local sub-matches between the two structures must be generated (Fig. 1f). Each rigid local sub-match follows a specific geometric transformation (ie. a specific rotation matrix and translation vector pair). To that end, we have implemented an algorithm that infers all local and rigid matches from the 2D difference sub-matrix. The only constraint imposed by this algorithm is that a minimum local match must contain at least three pairs of SSE elements. Briefly, the algorithm follows the diagonal below and adjacent to the main diagonal, checking for the observed Δω values. To initiate a new local matching block, a non-null Δω value equal or smaller than 90° is needed. If the next value is equal or smaller than 90°, the algorithm extends the matching block. If the observed Δω value is absent (ie. null), then the block is trimmed. Matching blocks smaller than 3 × 3 are not considered. If the Δω value is larger than 90°, then the adjacent left-row and bottom-column cell values are checked for non-null values equal or smaller than 90°. If this is not fulfilled, the matching block is trimmed. The detailed pseudocode of this algorithm is provided as supplementary material (Additional file 1: Figure S5).

Integrated visualization of structural matches

Finally, the local matching blocks are superposed in 3D following independent geometrical transformations. To achieve this, the coordinates of the SSE vectors belonging to each local matching block are first extracted. Then, both sets of coordinates are superposed using a particular implementation of the Kabsch algorithm [40], which is based on Lagrange multipliers to solve the optimal superposition problem. This algorithm implementation was proposed by Kearsley and provides an analytical solution based on quaternions to generate the three-dimensional superposition with minimal root mean square deviation [41]. The end result is the flexible global superposition of two structures (Fig. 1f).

Parameterization of the method

The gap-opening penalties defined in the steps of dynamic programming, C constant and maximum distance cutoff are the most important parameters to compare the SSE matrices. To calibrate these parameters in our method, we aligned 100 homologous protein pairs from HOMSTRAD dataset, carrying out several tests with different combinations of parameter values.

We used the Sorensen-Dice similarity index (QS) [42] to compare the precision of the method to detect equivalent pairs of SSEs in matrix alignments, using as gold standard the HOMSTRAD superpositions. The QS index was defined as:

$$QS = \frac{2 \times M}{A + B} \quad (6)$$

where A and B are the number of SSE pairs aligned that were reported by MOMA and HOMSTRAD, respectively, and M is the number of SSE pairs aligned in common. QS index lies between 0 (all SSE pairs aligned by MOMA are different from those reported by HOMSTRAD superposition) and 1 (SSE pairs aligned by MOMA are equal to those reported by HOMSTRAD). In each test, we calculated the average QS index to

determine the best combination of parameter values (Additional file 1: Table S2).

Performance assessment

We performed standard receiver operating characteristic (ROC) curve analysis and adopted the area under the ROC curve (AUC) as the accuracy measure for each method [43]. In these tests, SCOP classification (same fold, superfamily or family) was used as the gold standard to define true positive and true negative instances. Given a protein query and considering the list of hits above a score threshold returned by a search against the datasets, we counted a hit as a true positive (TP) if the structure target had the same SCOP classification level as the protein query. Otherwise, it was classified as a false positive (FP). The statistical significance of the observed differences in classifier performance was calculated with StAR web server (http://melolab.org/star) as previously described [44].

Additional file

Additional file 1: Suplementary data, including: Table S1. Average QS values for best combinations of MOMA parameter values; **Table S2.** Comparison of structural alignments generated by MOMA with those defined in HOMSTRAD; **Table S3.** Benchmark test to assess the performance of MOMA; **Table S4.** Statistical analysis for the benchmark of MOMA with other methods; **Table S5.** Statistical analysis for the benchmark of MOMA with different methods to assign secondary structure; **Table S6.** Set of 100 distant homologous protein pairs obtained from HOMSTRAD database; **Figure S1.** Calibration of distance cutoff using the HOMSTRAD set with the best combination of parameter values; **Figure S2.** ROC curves for the small set of seven most common folds according to TOPS database; **Figure S3.** ROC curves of classification at the SCOP fold and superfamily level; **Figure S4.** Execution time of MOMA; **Figure S5.** Algorithm used for extracting the rigid local matches. (PDF 2793 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FIG programmed and implemented the method, executed all computer calculations reported in this work and contributed to the writing of this manuscript and preparation of Tables and Figures. FR-V implemented the Kabsch algorithm for optimal three-dimensional superposition of matching structures and contributed to the preparation of some Figures. ILI contributed with the testing and improvement of the algorithm for integrating local structure matches into a single global solution. DPD and FM conceived and supervised this research, structured and wrote this manuscript with the help of FIG. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by FONDECYT Chile research [grant 1141172, to F.I.G., F.R-V., I.L.I. and F.M.], and by Heidelberg University Frontier [grant 28577, project #D.801000/12.074, to D.P.D and F.I.G., respectively]. I.L.I was funded by the Marie Curie Initial Training Network PERFUME (PERoxisome Formation, Function, Metabolism) grant (grant agreement number 316723)."

Author details

[1]Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Alameda 340, Santiago, Chile. [2]Centre for Organismal Studies (COS), Heidelberg University, Heidelberg, Germany. [3]Centro Andaluz de Biología del Desarrollo (CABD), Universidad Pablo de Olavide, Sevilla, Spain.

Received: 19 August 2015 Accepted: 21 December 2015
Published online: 05 January 2016

References

1. Erickson HP. Atomic structures of tubulin and FtsZ. Trends Cell Biol. 1998; 8(4):133–7.
2. van den Ent F, Amos LA, LoĚwe J. Prokaryotic origin of the actin cytoskeleton. Nature. 2001;413(6851):39–44.
3. Hasegawa H, Holm L. Advances and pitfalls of protein structural alignment. Curr Opin Struct Biol. 2009;19(3):341–8.
4. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci. 1995;20(11):478–80.
5. Gerstein M, Levitt M. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. Proc Int Conf Intell Syst Mol Biol. 1996;4:59–67.
6. Sippl MJ, Wiederstein M. Detection of spatial correlations in protein structures and molecular complexes. Structure (London, England : 1993). 2012;20(4):718–28.
7. Ortiz AR, Strauss CEM, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci. 2002;11(11):2606–21.
8. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 1998;11(9):739–47.
9. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: a multiple structural alignment algorithm. Proteins: Struct, Funct, Bioinf. 2006;64(3):559–74.
10. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics. 2003;19 suppl 2:ii246–55.
11. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302–9.
12. Gibrat J-F, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol. 1996;6(3):377–85.
13. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. *Computer methods for macromolecular sequence analysis*. 1996.
14. Guerler A, Knapp EW. Novel protein folds and their nonsequential structural analogs. Protein Sci. 2008;17(8):1374–82.
15. Stivala A, Wirth A, Stuckey PJ. Tableau-based protein substructure search using quadratic programming. BMC bioinformatics. 2009;10:153.
16. Schwede T, Peitsch MC. Computational structural biology: Methods and applications. 1st ed. Singapore: World Scientific; 2008.
17. Wiederstein M, Gruber M, Frank K, Melo F, Sippl MJ. Structure-based characterization of multiprotein complexes. Structure. 2014;22(7):1063–70.
18. Brohawn SG, Leksa NC, Spear ED, Rajashankar KR, Schwartz TU. Structural evidence for common ancestry of the nuclear pore complex and vesicle coats. Science. 2008;322(5906):1369–73.
19. Lesk AM. Systematic representation of protein folding patterns. J Mol Graph. 1995;13(3):159–64.
20. Konagurthu AS, Stuckey PJ, Lesk AM. Structural search and retrieval using a tableau representation of protein folding patterns. Bioinformatics (Oxford, England). 2008;24(5):645–51.
21. Konagurthu AS, Lesk AM. Structure description and identification using the tableau representation of protein folding patterns. Methods in molecular biology (Clifton, NJ). 2013;932:51–9.
22. Kamat AP, Lesk AM. Contact patterns between helices and strands of sheet define protein folding patterns. Proteins. 2007;66(4):869–76.
23. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247(4):536–40.
24. Chen K, Ruan J, Kurgan L. Prediction of three dimensional structure of calmodulin. Protein J. 2006;25(1):57–70.
25. Shatsky M, Nussinov R, Wolfson HJ. Flexible protein alignment and hinge detection. Proteins: Struct, Funct, Bioinf. 2002;48(2):242–56.
26. Devos D, Dokudovskaya S, Alber F, Williams R, Chait BT, Sali A, et al. Components of coated vesicles and nuclear pore complexes share a common molecular architecture. PLoS Biol. 2004;2(12):e380.

27. Field MC, Sali A, Rout MP. Evolution: On a bender–BARs, ESCRTs, COPs, and finally getting your coat. J Cell Biol. 2011;193(6):963–72.
28. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins: Struct, Funct, Bioinf. 1995;23(4):566–79.
29. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci. 1998;7(11):2469–71.
30. Slater AW, Castellanos JI, Sippl MJ, Melo F. Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. Bioinformatics (Oxford, England). 2013;29(1):47–53.
31. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014;42(Database issue):D304–309.
32. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.
33. Jung J, Lee B. Protein structure alignment using environmental profiles. Protein Eng. 2000;13(8):535–43.
34. Carpentier M, Brouillet S, Pothier J. YAKUSA: a fast structural database scanning method. Proteins. 2005;61(1):137–51.
35. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol. 2005;346(4):1173–88.
36. Wall ME, Rechtsteiner A, Rocha LM. Singular value decomposition and principal component analysis. In: *A practical approach to microarray data analysis*. Springer. 2003: 91–109.
37. Sung W-K. Algorithms in bioinformatics: A practical introduction: CRC Press; 2009. Broken Sound Parkway, NW Suite 300, Boca Raton, FL, 33487. USA.
38. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.
39. Sippl MJ. On distance and similarity in fold space. Bioinformatics (Oxford, England). 2008;24(6):872–3.
40. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr A. 1978;34(5):827–8.
41. Kearsley SK. On the orthogonal transformation used for structural comparisons. Acta Crystallogr A. 1989;45(2):208–10.
42. Wolda H. Similarity indices, sample size and diversity. Oecologia. 1981;50(3):296–302.
43. Fawcett T. ROC graphs: Notes and practical considerations for researchers. Mach Learn. 2004;31:1–38.
44. Vergara IA, Norambuena T, Ferrada E, Slater AW, Melo F. StAR: a simple tool for the statistical comparison of ROC curves. BMC bioinformatics. 2008;9:265.

3.2. VALIDAR LA HERRAMIENTA PARA INFERIR RELACIONES ESTRUCTURALES

3.2.1. Resumen

La publicación de MOMA dio lugar al desarrollo de una nueva metodología para la comparación flexible y eficiente de estructuras de proteínas, en especial, los alineamientos a nivel de matrices ofrecen una enorme ventaja en el reconocimiento visual de regiones locales conservadas al comparar un par de estructuras. Sin embargo, el método publicado inicialmente contaba con varias falencias que debían ser corregidas para inferir relaciones estructurales entre proteínas muy divergentes. Además, este programa requería de la aplicación de una metodología robusta que le permitiera identificar pares equivalentes de sub-fragmentos a partir de las matrices de diferencias.

Al corregir estas carencias, se dio lugar al desarrollo de una nueva herramienta para la comparación de estructuras de proteínas llamada MOMA2 que cuenta con varias ventajas con respecto a su versión anterior, como una mejor definición de los pares equivalentes de elementos de estructura secundaria y un nuevo esquema de puntaje que le permiten destacar, de forma precisa, las similitudes estructurales entre proteínas relacionadas. Además, MOMA2 cuenta con una nueva metodología que le permite reconocer y evaluar varias combinaciones de pares de sub-fragmentos alineados obtenidos a partir de los alineamientos de matrices, reportando subsecuentemente superposiciones compuestas de las cuales se identifica el alineamiento estructural más largo. MOMA2 incluso puede reportar las superposiciones locales de las estructuras comparadas según el alineamiento individual de cada uno de los pares de sub-fragmentos que fueron encontrados equivalentes, permitiéndole explorar la flexibilidad de las proteínas analizadas incluyendo el desplazamiento de cuerpo rígido de sus dominios. Esta nueva

característica señala que MOMA2 puede ser usado para explorar y estudiar cuantitativamente el desplazamiento espacial de los dominios en proteínas flexibles.

A continuación, el siguiente capítulo presenta las nuevas mejoras incluidas en MOMA2 para inferir relaciones estructurales entre proteínas mediante comparaciones flexibles. Por último, cabe señalar que este capítulo fue adaptado de un artículo en estado de edición que describe el desarrollo de MOMA2 con propósito de incluirlo en este manuscrito, y se espera más adelante publicar esta herramienta.

3.2.2. PAPER EN DESARROLLO: "MOMA2: Improving structural similarity detection beyond the twilight zone"

3.2.2.1. Introduction

The detection of structural similarities between pairs of proteins to infer evolutionary relationships is an important issue today. The number of protein structures available in the Protein Data Bank is increasing exponentially these years, and this pace is foreseen to continue and even rise in the future. To explore the vast quantity of information available will require the development of novel, fast, and precise approaches for the comparison of protein structures (wwPDB consortium, 2018).

Among the hundreds of applications used to compare protein structures, several of these methods are known as rigid-body aligners because they treated proteins as rigid bodies to find the largest common substructure shared between a pair of proteins. However, proteins are intrinsically flexible molecules, and their function is often associated to flexibility, and the rigid-body alignment methods are limited when it comes to finding correct superpositions for proteins that undergo structural changes, especially when reporting significant structural similarities in flexible regions.

In the recent years, few flexible alignment methods have been developed to overcome this problem such as FlexProt, FATCAT, RAPIDO, MATT, GR-align, and CAB-align (Shatsky *et al.*, 2004; Ye and Godzik, 2003; Terashi and Takeda-Shitaka, 2015; Menke *et al.*, 2008; Malod-Dognin and Pržulj, 2014; Mosca *et al.*, 2008). These programs implement different approaches to determine equivalent sub-fragments between pairs of proteins. Some flexible methods, such as FATCAT and FlexProt, apply independent and local rigid-body transformations to align each pair of equivalent sub-

fragments to report the optimal alignment that minimize the number of twists and maximize the alignment length (Hasegawa and Holm, 2009). On the other hand, flexible aligners such as CAB-align and GR-align use residue-residue contact maps rather than 3D coordinates to superpose protein structures (Terashi and Takeda-Shitaka, 2015; Malod-Dognin and Pržulj, 2014).

However, the flexible methods face some challenges that need to be handled to explore structural relationships among remotely related proteins. Flexible aligners could overvalue the structural similarity found between unrelated pairs of proteins, inserting several twists in the input structures to superpose many small pairs of protein sub-fragments to report subsequently a large alignment (*i.e.,* alignment length is usually the variable that is optimized). These aligned sub-fragments are often too small to be significant, specially from an evolutionary point of view. Meanwhile, superpositions generated by flexible methods among remotely related proteins can include several twists to report a large alignment. Because these methods are optimized to report large structural alignments, they are not able to discriminate between related and unrelated proteins based on the structural similarity reported. Another disadvantage of these methods is that they are computationally expensive and often time consuming to determine protein flexibility. For example, the required execution time to compare protein structures of flexible methods is much larger than rigid methods such as TOPMATCH or Structal (Gutiérrez *et al.*, 2016). Also, flexible methods based on contact maps cannot report structural superpositions despite they can intuitively determine equivalent domains. For example, programs such as CAB-align and GR-align

do not have a methodology to report and evaluate the optimal superposition obtained according to local sub-fragments pairs derived from the alignment of the matrices.

To address these needs, we previously published a flexible, fast, and accurate tool called MOMA that recognizes structural similarities between pairs of protein structures based on the geometrical orientation, angles, and distances, of their secondary structure elements (SSEs)(Gutiérrez *et al.*, 2016). Unlike the contact map methods, MOMA represents protein structures into matrices retaining the structural information of their folds at the level of secondary structure elements (Gutiérrez *et al.*, 2016).

Our goal here is to present the most recent version of our final tool called MOMA2 to cope the disadvantages that possessed the current flexible alignment tools. This program includes a novel methodology to significantly identify combinations of pairs of equivalent sub-fragments to infer quantitative and statistically significant structural relationships between any pair of divergent proteins. MOMA2 improves the local superposition of the selected sub-fragments to generate the largest common substructure match and reports the local context of the equivalent sub-fragments. Additionally, MOMA2 can evaluate the conservation of the domains and characterize their movements using the information derived from matrix alignments. MOMA2 also performs fast searches against a curated database of matrices, and it has a better performance to recognize structural relationships in remotely related multi-domain proteins. The MOMA2 repository is freely available at https://hub.docker.com/r/fggutierrez2018/moma2 that includes all required Python modules, databases, and programs required to calculate the structural alignments.

3.2.2.2. Algorithm and implementation

3.2.2.2.1. Description of the algorithm

The MOMA2 algorithm is divided in two consecutive parts. The first part consists in calculating the structural similarity found between a pair of proteins using 2D matrices of secondary structure elements (SSE matrices containing angles and distances between pairs of secondary structure elements), while the second part calculates the best superposition according to the equivalent sub-fragments extracted from a 2D matrix alignment (which contains differences of angles and distances of SSE from the two proteins being compared).

In the first part of the algorithm, MOMA2 implements two secondary structure assignment programs, DSSP and KAKSI (Kabsch and Sander, 1983; Martin *et al.*, 2005), to find the precise start and endpoint of each secondary structure element in the protein structures to build the SSE matrices (Appendix 6.1.1, Supplementary Figure 1). The rows of the matrices are then aligned using dynamic programming and a precise scoring function to detect equivalent pairs of secondary structure elements (Appendix 6.1.1, Supplementary Equation 1). MOMA2 defines an equivalent SSEs based on three aspects, the angular differences of their dihedral and internal angles, and the distance differences of their centroids (Appendix 6.1.1, Supplementary Figure 2, and Equation 1). MOMA2 then performs local, semi-global, and global alignments to align the rows derived from each matrix (Appendix 6.1.1, Supplementary Figure 3). As a result, MOMA2 creates three scoring matrices to calculate three local alignments using the SSE sequences of the input proteins. Original SSE matrices are resized into an identical size

according to the aligned SSEs, where their dihedral angles and distances are subtracted to create a new matrix called Δsubmatrix. MOMA2 calculates from this matrix a similarity score called B_{score} which integrates the structural similarity reported by equivalent sub-fragment pairs (Equation 1). This score is defined as:

$$B_{score} = \frac{B_\omega + B_d}{2} \quad (1)$$

$$B_d = \sum_i^N e^{-r_i^2/\sigma_i^2}, r_i^2 = \Delta d_i^2 \ (2)$$

$$B_\omega = \sum_i^N e^{-r_i^2/\sigma_i^2}, r_i^2 = \Delta \omega_i^2 \ (3)$$

In contrast to the S_{score} reported by MOMA, the B_{score} combines two scores derived from Gaussian functions that are used to calculate similarity scores based on the angular and distance differences, Δω and Δd respectively, observed in N positions that are covered by significant sub-blocks selected in the Δsubmatrix (Equations 2 and 3). The rate of reduction of the B_ω and B_d scores is determined in function of the σ_ω and σ_d scaling parameters.

The range of the B_{scores} that a Δsubmatrix can report is [0, N(N-1)/2], where N is the number of aligned pairs of SSEs. Intuitively, a high B_{score} close to N(N-1)/2 indicates that the pair of proteins superposed are identical or structurally highly similar, while a low B_{score}, close to 0, suggests that the matrix alignment generated by comparing a pair of proteins with MOMA2 is not significant (Appendix 6.1.1, Supplementary Figure 4). The statistical significance of the detected similarity is evaluated through a p-value that

measures the probability of a Δsubmatrix of dimension *NxN* to obtain a B_{score} larger than or equal to the observed B_{scores} on Δsubmatrices calculated from random comparisons among unrelated proteins (whose dimensions are lower than or equal to *NxN*) (Appendix 6.1.1, Supplementary Figure 5). The p-values are calculated by fitting the B_{scores} with empirical distributions derived from ~2,575,000 random comparisons of 199 non-homologous proteins obtained from the TM-align dataset (Zhang and Skolnick, 2005). The p-value is calculated for each Δsubmatrix based on their B_{score}; if this value is lower than $\alpha = 0.05$, there is stronger evidence in favor of the corresponding protein structure pair is related.

The second part of the algorithm consists in calculating the best structural superposition from a Δsubmatrix (Figure 1). MOMA2 starts by identifying a list of aligned sub-fragment pairs, which are represented in a Δsubmatrix as blocks. These blocks are selected using two types of filters. The first group of filters consists of a list of cut-offs that was trained with the Perceptron algorithm, implemented in Scikit-learn Python module (Freund and Schapire, 1999; Pedregosa *et al.*, 2011), to select significant blocks based on their structural differences, angles and distances between their pairs of SSEs (Appendix 6.1.1, Supplementary Figure 6, and Supplementary Table 1). These filters were trained using a dataset composed of ~341,500 blocks calculated from comparisons between pairs of related and unrelated domains available in the ASTRAL SCOPe 2.06 dataset (with a sequence identity below 40%) (Fox *et al.*, 2013). Firstly, MOMA2 selects those blocks whose sizes are between 3 to 10 pairs of SSEs aligned, and their average angular and distance differences are lower than the established cut-offs. Subsequently, the evaluation can tolerate, to a certain extent, the local variation of

one or two pairs of aligned SSEs at the level of angles or distances without decreasing their structural similarity. Those blocks that have a considerable increment in the local variation between their SSE pairs are filtered out with the second group of filters, which consider the number of positions in the Δsubmatrix where the angular difference is larger than 45° or the measured distance is more than 5 Å. If a previously selected block overpasses the number of positions tolerated according to its dimension, the block is discarded (Appendix 6.1.1, Supplementary Table 2). The remaining blocks are highlighted with red squares in the Δsubmatrix (Figure 1).

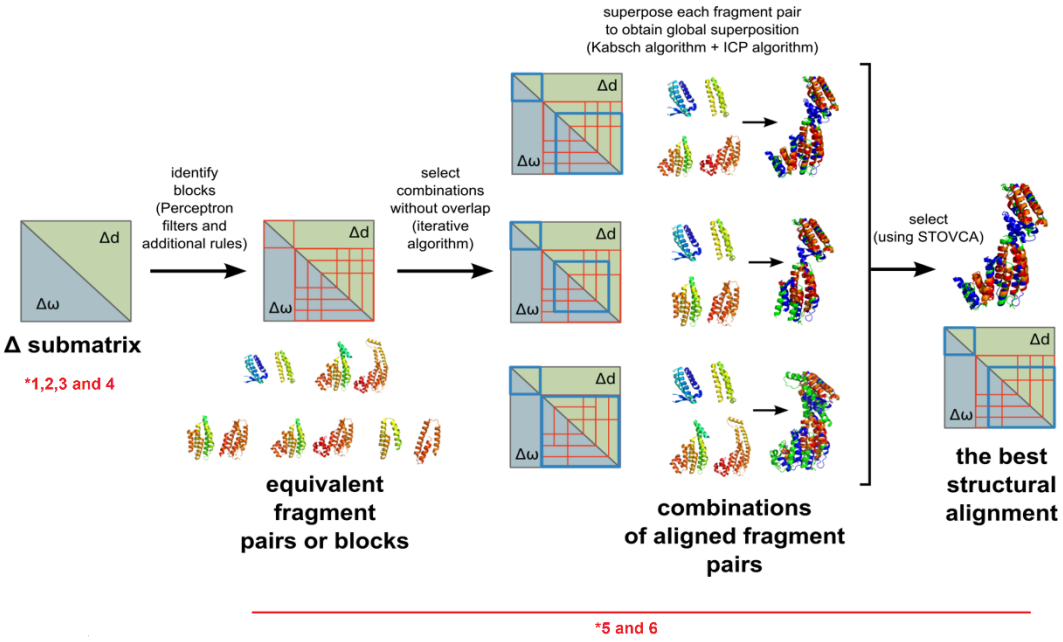


Figure 1. Flowchart implemented to calculate a structural superposition from a matrix alignment.
Firstly, MOMA2 uses optimized threshold filters and additional rules to recognize equivalent sub-fragments based on the blocks with lower angular and distance differences observed in the Δsubmatrix (Δω and Δd). The equivalent sub-fragments are highlighted in the Δsubmatrix, respectively (red color in the matrices). MOMA2 then calculates a limited number of non-overlapping combinations from these blocks. Individually, MOMA2 refines the structural alignment of each pair of aligned sub-fragments and combines them to reconstruct global solutions. Based on the results reported with the STOVCA software (Slater et al., 2012), MOMA2 chooses the solution that reports the largest structural alignment and creates a PyMOL session to visualize the aligned sub-fragments, whose blocks are represented in the Δsubmatrix (blue in the matrices). *Modifications implemented are described in detail in the Appendix 6.1.1 (red font labels and line below the matrices).

MOMA2 then implements an iterative algorithm to select a representative non-overlapping set of combinations of the blocks without exploring all possible solutions, and at the same time, reducing the search time (by default, MOMA2 defines a maximum

number of 25 combinations) (Appendix 6.1.5.1). For each combination, the local sub-fragments deducted from selected blocks are superposed individually using SSE vectors, which are aligned through geometrical transformations with the Kabsch algorithm (Kabsch, 1976). Subsequently, these initial superpositions are improved using a modified version of the Iterative Closest Point algorithm (ICP) (Paul J. Besl and McKay, 1992) (Appendix 6.1.1, Supplementary Figure 7). The ICP algorithm refines the superpositions calculated with the Kabsch method, searching in each iteration, the superposition that reports the highest number of Cα aligned, and the lowest RMSD value in comparison with superpositions calculated in the previous iterations (Appendix 6.1.5.2). Eventually, MOMA2 reports a list of global solutions by composing the partial superpositions reported for each combination. Each partial superposition is evaluated using the STOVCA software to measure the structural similarity of each block selected (Slater *et al.*, 2012) (Appendix 6.1.1, Supplementary Figure 8). All combinations are ranked using the total number of equivalent aligned residues, then MOMA2 chooses the largest structural alignment whose blocks are highlighted in blue in the Δsubmatrix (Figure 1, red color). Finally, MOMA2 creates a script session to visualize in PyMOL the largest common substructure match between a pair of proteins, and several local superpositions based on their pairs of equivalent sub-fragments to analyze the rigid displacements of their domains (Schrödinger, LLC, 2015).

3.2.2.2.2. Implementation

The MOMA2 algorithm has been implemented in a repository whose scripts were written in Python and ANSI C for visualization and statistical analysis of the SSE alignments. MOMA2 computer software has two main implementations: MOMA2-pw

and MOMA2-db. The first computer program is used to compare pairs of protein structures to identify the similarities between them, and the second one is used to perform query searches against a curated database of multiple proteins.

MOMA2-pw requires as input two structure files in PDB format to report a structural alignment for a pair of proteins, rotating and translating the rigid blocks of the target protein to those blocks of the query protein according to the best combination found. This program provides as output a list of structural measurements from this combination and other combinations tested, including the total number of equivalent pairs of aligned residues (eq), the percentage of relative similarity (Sr), the percentage of structural overlap (SO), the percentage of sequence identity (%id seq.) the overall RMSD ($RMSD_{pond}$), and the B_{score} of the matrix alignment (Supplementary data, Figure S8).

On the other hand, MOMA2-db requires as input a PDB file to search similar protein structures in a database. The MOMA2 repository contains a non-redundant dataset composed by 27,500 chain matrices (protein structures solved by X-ray crystallography or NMR spectroscopy with a sequence identity between them below 90% and a maximum resolution of 3.0 Å) obtained from a curated list of protein structures defined by PISCES public server (Wang and Dunbrack, 2005). MOMA2-db outputs a list of hits that are filtered by a specific p-value and ranked by their B_{scores}. These ranking files show some details associated with the target protein, including the protein name, the scientific name of the species, and the life domain to which it belongs. Moreover, MOMA2-db software also allows users to vary/explore different

combinations of parameters and supports a multi-threaded implementation to accelerate the process.

3.2.2.3. Results

3.2.2.3.1. Comparisons of analogous domains

A comparative analysis with other flexible and rigid-body aligner tools was performed to determine the behavior of MOMA2 to align unrelated domains and reporting their structural similarity. A set of 92 domain pairs collected from the MALISAM (Manual ALIgnments of Structurally Analogous Motifs) database were aligned with MOMA2, flexible FATCAT, FlexProt, MATT, TM, FAST, DALI, and TOPMATCH to determine if the similarity reported by these methods is overestimated using as reference the manual superpositions obtained from the MALISAM database (Appendix 6.1.2, Supplementary Table 3) (Cheng *et al.*, 2008). If the alignment length of a pair of domains aligned is higher than the number of aligned residues estimated by the manual superposition of MALISAM, the superposition obtained is considered overestimated. Subsequently, these superpositions were evaluated with STOVCA software (Slater *et al.*, 2012) to establish a standardized point of comparison in the evaluation of the length of the structural alignments reported by different methods (distance cut-off was set to 3.5 Å). Subsequently, the one-tail t-test for a paired sample was applied to compare the distributions of the length of aligned residues reported by different aligners to establish whether the similarities calculated by a particular computer program are overestimated or not overestimated in comparison with MALISAM superpositions (Table 1). The null hypothesis implemented in the t-test assumes that the mean difference between the mean of the equivalent residues reported by MALISAM

and any aligner is greater or equal to 0 (similarity is not overestimated), while the alternative hypothesis assumes that this mean difference is lower than 0 (similarity overestimated). Additionally, we performed the Bonferroni Correction (Post-hoc test) following a significant one-way ANOVA result to determine where the method differences lie. The proportion of overestimated pairs of unrelated pairs and their significant differences with respect to MALISAM by seven known flexible and rigid-body tools, and MOMA2 are shown in Table 1.

A high percentage of the analogous domains was overvalued by flexible aligners such as FlexProt, accounting for 98,91%, followed by MATT at just over 67% and FATCAT at nearly 52%, whereas MOMA2 reported the lowest percentage of overestimated pairs (at just over 13%) (Table 1). Paired comparisons performed between the samples of MALISAM, and the flexible aligners such as FlexProt, MATT and FATCAT showed sufficient evidence to support alternative hypothesis that these programs significantly overestimated the similarity reported in the comparisons of the analogous domains (p-values < 0.01), calculating in most of the cases longer structural alignments than the superpositions of MALISAM (Table 1; Appendix 6.1.2, Supplementary Table 3). On the other hand, methods such as MOMA2, TM, FAST, and TOPMATCH reported t-values greater than the critical value, suggesting that it is not sufficient evidence to reject the null hypothesis, and therefore these methods do not overestimate the similarity reported in the domains superposed. On the contrary, these programs showed shorter or similar structural alignments than MALISAM, except for DALI that overestimated the similarity in 40 of the 92 analogous pairs (Table 1; Appendix 6.1.2 Supplementary Table 3). MOMA2 only overestimated the similarity

found in 12 out of the 92 pairs analyzed, where the number of aligned residue pairs was longer than the aligned residue pairs present in the MALISAM superpositions (Appendix 6.1.2, Supplementary Table 3). After performing the one-way ANOVA test analysis, there was a significant difference in the average of the aligned residue pairs reported between the 8 methods tested and MALISAM ($F(8;819) = 35,3579$; $p<0.0001$). Post-hoc analyses revealed that the length of aligned residues reported by FATCAT (eq = 44,72±11,52), FlexProt (eq = 58,05±11,85), and MATT (eq = 47,32±12,79) were significantly larger compared with those in MALISAM (eq = 42,15±10,24). Also, post-hoc tests revealed that rigid alignment methods, such as TM (eq = 39,86±10,98), FAST (eq = 39,51±10,61), DALI (eq = 42,65±10,77) and TOPMATCH (eq = 38,54±12,32), plus MOMA2(eq = 31,23±14,56), did not significantly overestimated the similarities reported for analogous pairs compared with those in the database MALISAM (eq = 42,15±10,24), considering the length of structural alignments.

Table 1. Statistical analysis of the similarity reported in analogous pairs by flexible and rigid-body aligners.

	flexible aligners				rigid-body aligners			
	MOMA2	**FATCAT**	**FLEXPROT**	**MATT**	**TM**	**FAST**	**DALI**	**TOPMATCH**
overestimated pairs (%)	13,04	51,09	98,91	67,39	31,52	27,17	43,48	28,26
t-value	7,5543	-3,1951	-17,3265	-5,4352	3,3036	3,4388	-0,9166	3,6018
critical t	-1,6618	-1,6618	-1,6618	-1,6618	-1,6618	-1,6618	-1,6618	-1,6618
p-value (lower-tailed)	**1**	**0,001**	**7,00E-31**	**2,00E-07**	**0,9993**	**0,9996**	**0,1809**	**0,9997**

A total of 92 pairs of analogous domains were superposed with MOMA2, flexible FATCAT, MATT, FlexProt, TM, FAST, DALI, and TOPMATCH whose superpositions were compared with the manual structural alignments reported by MALISAM (Appendix 6.1.2, Supplementary Table 3). If the alignment length from a superposition generated by a computer program is higher than the alignment length reported by MALISAM, this pair is counted as overestimated. The differences of the structural alignments reported between MALISAM, and any aligner were evaluated using the paired sample t-test (1-tail). According with the Bonferroni Correction, if the p-value is less than or equal to 0,00625 (red), significantly a program reports longer structural alignments in comparison with the alignments calculated manually by MALISAM. If p-value > 0,00625 (green), then the program does not overestimate the similarity reported by MALISAM.

In the flexible comparison of pairs of unrelated proteins that have a single domain and belong to a different type of fold, it is common that some methods added many twists to force the structural alignment. A clear example is observed in the structural alignment of the domains d1st6a2 and d1jqna_ present in the MALISAM dataset (Figure 2). This example corresponds to an analogous pair composed of an interface motif and a core motif, where the domain d1st6a2 is a 4-helix bundle classified in the alpha-catenin/vinculin superfamily, while the domain d1jqna_ is a TIM beta/alpha-barrel fold. Superposition obtained from MALISAM reports 69 aligned residues with an RSMD of 2.3 Å (8.7 % of sequence identity).

Some flexible aligners such as FlexProt and MATT maximize the number of aligned residues introducing several twists to superpose the 4-helix bundle of the d1st6a2 domain, reporting 105 and 84 aligned residues, respectively. FATCAT alignment does not introduce a twist, and it is very similar to the MALISAM superposition, reporting an alignment of 79 pairs of equivalent residues with an RMSD of 1.8 Å. Otherwise, TOPMATCH reported short structural alignments compared with the superposition estimated by MALISAM, where only two equivalent pairs of α-helices were superposed, aligning a total of 42 pairs of residues with an RMSD of 2.3 Å. In this example, MOMA2 showed the shortest alignment length superposing in total 39 pairs of equivalent residues with an RMSD of 1.7 Å and reporting the greatest percentage of sequence identity of the tested methods equal to 10.26%. These results show that some flexible aligners can force the superposition of analogous domains introducing small sub-fragments pairs to superpose different protein folds, and MOMA2 can identify domains with similar folds and differentiating them from those domains that are analogous.

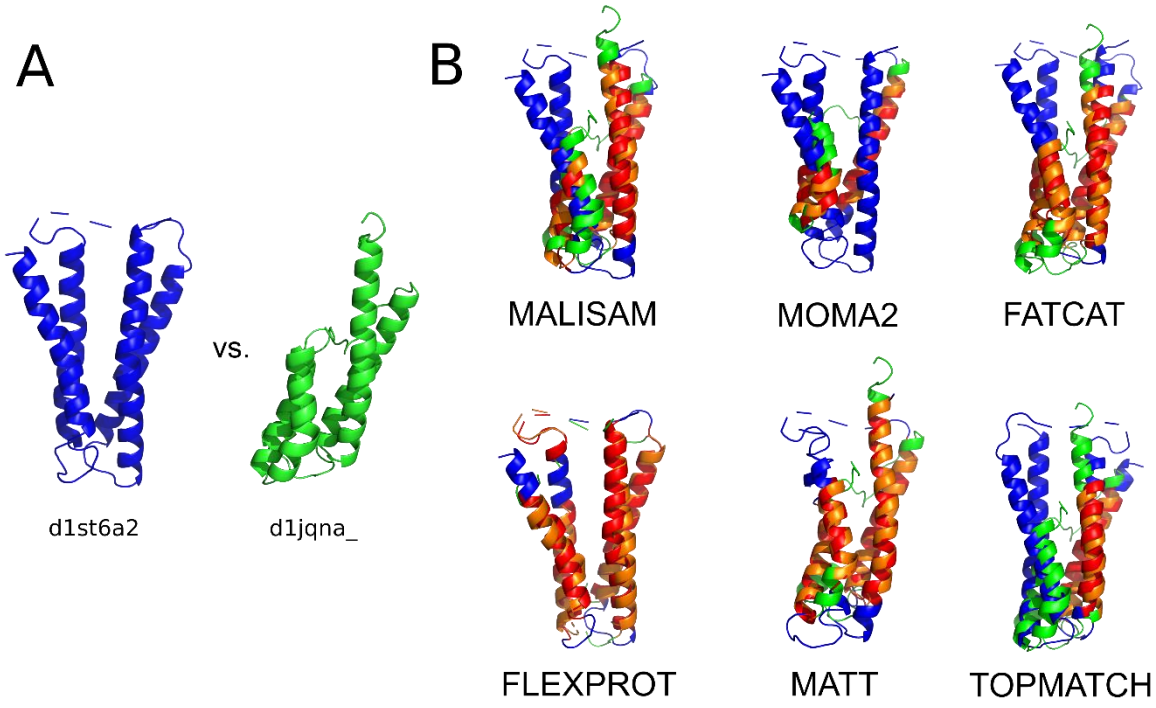


Figure 2. Flexible aligners can overestimate the similarity found between unrelated proteins that have the same composition of SSE.

This example shows a pair of unrelated domains d1st6a2 (PDB code 1st6, vinculin), and d1jqna_ (PDB code 1jqn, phosphoenolpyruvate carboxylase) (A). The vinculin structure is colored in blue, and the carboxylase structure is colored in green. As a standard, it is included the manual alignment of the MALISAM. Additionally, four flexible alignment programs such as Matt, FlexProt, flexible FATCAT, and MOMA2 were used to superpose these structures, including also a rigid-body aligner as TOPMATCH (B). The structurally aligned pairs of residues were determined using the STOVCA program where these residues are colored in red and orange, while the unaligned residues are colored in blue and green, respectively.

3.2.2.3.2. Classifying multi-domain proteins

To assess the performance of MOMA2 in the classification of multi-domain proteins, we constructed a benchmark dataset containing 2,995 multi-domain protein chain pairs obtained from the PDB database. This dataset is composed of 997 pairs of protein chains defined as related pairs, and 1,998 protein chain pairs considered as unrelated pairs. According to the SCOP classification, a pair of domains grouped into the same superfamily is inferred that they share a common ancestor based on their functional and structural features, and pair of proteins where none of their domains share the same topological arrangements probably arose from a distinct evolutionary origin. In the agreement with SCOP definitions, we classified a pair of proteins as unrelated if none of their domains have the same superfamily classification (negative case). Otherwise, we classified a pair of proteins as related if at least a pair of their domains share the same superfamily (positive case). We evaluated the performance of the flexible aligners using ROC and Precision-Recall analyses to classify the multi-domain proteins between related and unrelated pairs, considering the results obtained from the comparison of each pair of proteins, similarity scores reported by each method or according to their superpositions. Additionally, we measured the statistical differences of the areas under the ROC curves (AUC) using the StAR server (Vergara *et al.*, 2008).

Firstly, considering only the scores reported for each method we observed that MOMA2 shows the highest AUC and AUPRC values of the four methods tested in the classification of multi-domain proteins (Table 2), indicating that MOMA2 detect pairs of multi-domain proteins using only alignments of matrices with high specificity in comparison with other flexible aligners such as FATCAT or CAB-align. Also, MOMA2

reports a higher AUC value than MOMA, supporting the fact that the modifications implemented has a positive impact in the recognition of structural relationships using only matrices (Appendix 6.1.3, Supplementary Figure 9A). The Precision-Recall analysis shows that MOMA2 has an excellent early-retrieval of the true positive cases, showing a high precision at a recall rate less than or equal to 0.97 in comparison with other methods such as FATCAT, MOMA, and CAB-align (Appendix 6.1.3, Supplementary Figure 9C). The low figures observed for CAB-align are probably because this program was not able to align the contact maps of some protein pairs affecting its performance in the ROC analysis. In addition, according to the statistical analysis performed with the StAR server, all AUC differences were statistically significant suggesting that these methods behave differently when classifying the pairs of proteins as related and unrelated (Appendix 6.1.3, Supplementary Table 4).

Table 2. Benchmark of the classification of related and unrelated multidomain proteins.

Methods	AUC	AUPRC	ACC	OT	fp	tp	N	P
MOMA2	**0.9799**	**0.9759**	**0.9603**	25.5[*]	**0.009**	**0.898**	1998	997
flexible FATCAT	0.9731	0.9609	0.9295	286.6[+]	0.038	0.864	1998	997
MOMA1	0.9531	0.9397	0.9225	32[#]	0.033	0.832	1998	997
CABalign	0.9078	0.9274	0.9362	6142.3[&]	0.014	0.836	1998	997

AUC Area Under the ROC Curve
AUPRC Area Under the Precision-Recall curves
ACC Maximal accuracy
OT Optimal threshold at ACC ([#] S_{score}, * B_{score}, [+] FATCAT chaining score, [&] Normalized Similarity score)
fp false positive rate at OT
tp true positive rate at OT
N Number of negative instances
P Number of positive instances
Note that the results shown are based on alignments returned by four different methods

Secondly, the classification of multi-domain pairs according to the superpositions reported by the tested methods indicates that MATT and MOMA2 have an excellent performance in terms of AUC to classify pairs of related and unrelated proteins (Table 3). In addition, MOMA2 shows the best accuracy in the classification of homologue proteins, which is followed by FATCAT, MATT, FlexProt, and MOMA (Table 3). MOMA2 is more specific to classify pairs of multi-domain proteins than MATT at a false positive rate of less than 0.2, but MATT is more sensitive than MOMA2 at a false positive rate higher than 0.2 (Appendix 6.1.3, Supplementary Figure 9B). In AUPRC terms, MOMA2 has the best performance of the methods tested, which is followed by flexible FATCAT, FlexProt, and MOMA. This analysis suggests that MOMA2 has an excellent early retrieval of the true positive at a recall rate lower and equal to 0.98 (Appendix 6.1.3, Supplementary Figure 9D). Otherwise, MOMA has the worst AUC and AUPRC values in the benchmark test (Table 3), indicating that the modifications implemented for the identification and extraction of significant combinations of aligned sub-fragments had a positive impact in the obtention of precise superpositions from tested protein pairs. The analysis performed in the StAR server reports that there are not significant AUC differences between MOMA2 and MATT, or between MOMA and FlexProt. These results suggest that these methods have a similar performance to classify the protein pairs of the benchmark dataset (Appendix 6.1.3, Supplementary Table 5).

Table 3. Benchmark of the classification of related and unrelated proteins according to the number of aligned residues from superpositions.

Methods	AUC	AUPRC	ACC	OT	fp	tp	N	P
MATT	**0.9874**	0.9814	0.9603	131	0.025	0.930	1998	997
MOMA2	0.9861	**0.9844**	**0.9700**	77	**0.016**	**0.941**	1998	997
flexible FATCAT	0.9771	0.9761	0.9616	124	0.022	0.928	1998	997
FlexProt	0.9594	0.9480	0.9219	119	0.018	0.801	1998	997
MOMA1	0.9548	0.9268	0.9316	43	0.048	0.890	1998	997

AUC Area Under the ROC Curve
AUPRC Area Under the Precision-Recall curves
ACC Maximal accuracy
OT Optimal threshold at ACC (Number of aligned residues returned by STOVCA)
fp false positive rate at OT
tp true positive rate at OT
N Number of negative instances
P Number of positive instances
Note that the results shown are based on superpositions returned by five different methods

3.2.2.3.3. Computational time

MOMA2 is a practical option to search protein structures against a large dataset of protein in comparison with other methods such as FATCAT, FlexProt, MATT, and CAB-align. At the level of matrices, MOMA2 is faster than other flexible aligners to perform structural comparisons (Table 4). The speed of MOMA2 in the matrix comparisons depends on the time required by DSSP or KAKSI to assign the secondary structure elements to build the matrices. However, structural superpositions generated with MOMA2 are slower than the structural comparisons performed with FATCAT, FlexProt, and MATT. The bottleneck of the MOMA2 is the optimization of the superpositions obtained from the aligned sub-fragments detected in the Δsubmatrices. As described early, MOMA2 performs by default 2,500 steps of the ICP algorithm to refine their superpositions of the sub-fragments, trying at least 25 solutions to identify the best combination of blocks. The longest time required by MOMA2 to calculate these superpositions is attributable to these complex processes. Users can modify the parameters established by default to reduce the average running time used to calculate the superpositions. For example, selecting only one combination of blocks and 500 iterations of the ICP algorithm, the computation time of the pairs of the benchmark dataset is reduced to 9.53 sec/pair, 1.26x times faster than the configuration established by default, without sacrificing considerably the precision of the method for aligning the equivalent fragments identified.

Table 4. The average computational time of MOMA2 and other flexible aligners to perform comparisons of the benchmark set.

We used a general Linux computing system (Intel Core i7 8th Generation CPU at 4.1 GHz and 8GB memory) to compare 2,995 pairs of the multi-domain benchmark dataset with MOMA2, flexible FATCAT, FlexProt, Matt and CAB-align. The time of calculation considers the time to align a pair of structures with or without calculating a superposition.

Methods	Assignment of SSE	Superposition	Seconds/pair	Number of times slower	#
MOMA2	DSSP	-	0.36	1x	2995
flexible FATCAT	NA	-	3.68	10x	2995
flexible FATCAT	NA	+	3.69	12x	2995
FlexProt	NA	+	6.25	17x	2995
Matt	NA	+	8.53	24x	2995
MOMA2	DSSP	+	11.99	33x	2995
CAB-align	NA	-	21.14	58x	2995

[1] The average computational time for an alignment pair.
Number total of protein pairs aligned.
+ report a superposition
- without superposition
NA Not applies

3.2.2.3.4. Biological applications

A biological application derived from MOMA2 is the identification and characterization of domain movements in the comparison of related proteins. Different types of protein flexibility such as semi-rigid domain movements o large movements with substantial internal flexibility can be easily visualized in the Δsubmatrix, where internal rigid domains represented as blocks show greater differences in distances and angles between them indicating the presence of conformational changes between the proteins being compared. The Figure 3 illustrates the displacement of a pair of domains in two amylolytic and related enzymes composed by several domains and characterizing their local structural matches. We show the structural alignment calculated between the ***Klebsiella pneumoniae*** pullulanase and the ***Thermoactinomyces vulgaris*** α-amylase I with MOMA2. Both proteins have one or two carbohydrate-binding domains, which correspond to the N2 and N3 domains present in the pullulanase and the N domain present in the α-amylase protein (Mikami *et al.*, 2006; Abe *et al.*, 2004). These proteins also have a similar core composed by the A and C domains (Figure 3). It is interesting to note that a one rigid subdomain move of the N domains with respect to the common core in both enzymes is captured in the local structural matches of the rigid domains represented with red boxes in the Δsubmatrix. Four blocks were detected and correspond to the N, A and C domains of the pullulanase that were aligned with the N3, A, and C domains of the amylase. Superposition of the sub-fragments obtained from MOMA2 reports 352 aligned residues with a RMSD of 1.74 Å (20% of sequence identity) aligning four pairs of sub-fragments (Figure 3). Otherwise, FATCAT reports a shorter structural alignment than MOMA2 to this example (eq = 346 with a RMSD = 1.8 Å, according to

STOVCA), introducing four twists to superpose five pairs of sub-fragments with a 17.6% of sequence identity (Supplementary Figure 11). MOMA2 recognizes an invariable common core shared between the A and C domains reporting a structural alignment more compact than FATCAT, which one introduces a twist to extent the alignment of the A domains calculating an alignment with a lower percentage of sequence identity than MOMA2. This example shows that MOMA2 could be used to characterize or classify the relative movements found in flexible proteins, suggesting a new form to quantify the displacement of the internal rigid o semi-rigid domains based on the angular and distance differences observed in the Δsubmatrix.

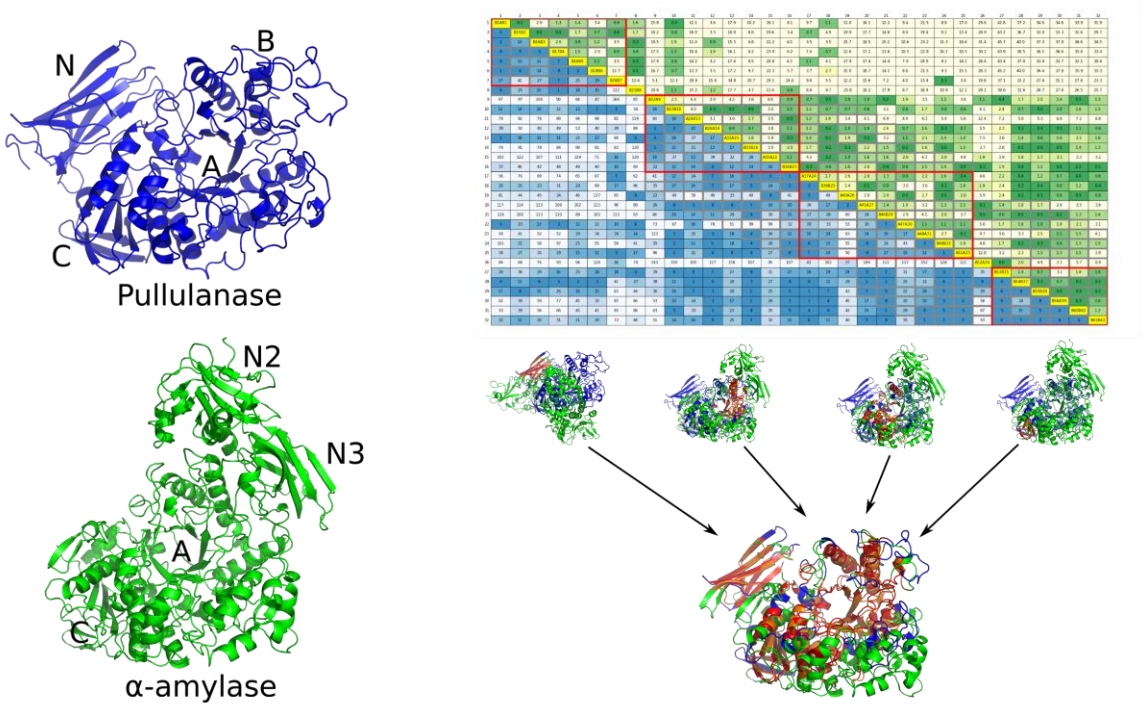


Figure 3. MOMA2 detect domain movements in remote protein structures.

Structural comparison calculated with MOMA2 between the *Klebsiella pneumoniae* pullulanase (PDB code 2fgz, chain A) and the *Thermoactinomyces vulgaris* α-amylase I (PDB code 1uh4, chain A). The figure indicates the domains found in pullulanase and α-amylase (left panel). The structure matching of the equivalent sub-fragments is highlighted with red boxes in the Δsubmatrix, while their local superpositions are shown below the Δsubmatrix (top right panel). The local superpositions are combined to report a composite solution, where the aligned residues reported by query and target structures are colored in red and orange, respectively (bottom right panel).

3.2.2.3.5. Technical limitations

On the technical aspect, among the weaknesses that have MOMA2 is the fact that it is a single chain and topology-dependent method that depends on the connectivity of SSEs. Additionally, the accuracy of MOMA2 depends directly on the methodology implemented in the assignment of the secondary structure elements to compare protein structures. Some examples show that other flexible aligners such as FATCAT or MATT report larger structural alignments than MOMA2. These methods are not limited to find equivalent pairs of residues aligning regions with loops, small α-helices, or a pair of α-helices where one of them have a kink as MOMA2 (Appendix 6.1.4, Supplementary Figure 10 and 11). MOMA2 was originally designed to report structural similarities based on the arrangement of sub-fragments composed by α-helices and β-strands with the same topological connections.

3.2.2.4. Discussion

In this study, we presented a new flexible alignment method to identify structural relationships beyond the twilight zone called MOMA2. We tested this method and other flexible computer programs available in the literature to evaluate the assumption that flexible methods do not consider the significance of sub-fragments aligned, producing valid alignments from unrelated proteins. It is the vital importance do not overestimate the similarity of unrelated pairs to establish structural relationships between related proteins that have diverged considerably during the evolution. We evaluated the behavior and performance of MOMA2 using a dataset composed of analogous domains obtained from the MALISAM dataset. We found that for a large proportion of the

analogous pairs, their structural similarity was overestimated by some flexible programs because these tools introduce many twists in the reference protein to align a pair of structures, incorporating alignment of sub-fragments composed by one or two secondary structure elements, even though some methods were optimized to introduce a minimum number of twists or were trained using comparisons between decoys. Unlike other flexible alignment programs, we found that MOMA2 does not overestimate the similarity reported between analogous proteins, and these results are attributable to the improvements incorporated in the algorithm to align matrices and the implementation of the new similarity score based on the angular and distance differences reported in the Δsubmatrix. The classification of related and unrelated multi-domain proteins showed that the modifications implemented in the alignment of matrices and the selection of significant pairs of aligned sub-fragments improve the performance of MOMA2 considerably in contrast to MOMA. The results obtained reveal that MOMA2 is a robust method to discriminate related and unrelated proteins composed by several domains using superpositions, it is more specific than MATT to detect related proteins, having an excellent early retrieval of the true positives. MOMA2 is the fastest method to perform comparisons in a benchmark set using only matrices (Table 4). However, MOMA2 is typically several times slower to calculate structural superposition from the Δsubmatrices when compared to FlexProt, FATCAT, or MATT. Despite this disadvantage, MOMA2 can quickly cover the space of known proteins to detect close and remote proteins to a query protein using matrix alignments.

Finally, we showed some of the advantages of the matrix and structural alignments of MOMA2 to recognize the largest similar part of two proteins, including

how their equivalent sub-fragments are rearranged to explore displacements and characterize domain movements. However, in some additional cases, MOMA2 shows some weaknesses that are associated with the technical aspects of the implemented method. Although other flexible aligners sometimes report a better structural alignment than MOMA2, none of these methods worked better for all cases tested, because each tool was developed with a specific purpose in mind. For example, FATCAT and FlexProt were designed for the alignment of flexible proteins where the distortions can be found within the secondary structure elements. Instead, MATT was created to calculate multiple alignments considering the flexibility of proteins. Along the same line, MOMA2 was optimized to recognize protein-folding patterns based on SSEs, which are relevant to discover structural relationships in ancient folds that are preserved in different proteins. Other methods, such as FATCAT or MATT, have been optimized to maximize the alignment length, sometimes aligning unstructured segments that are not considered by MOMA2, which are essential from a biological point of view to study the active site of an enzyme or a vital target site to develop new drugs. It is noteworthy to mention that MOMA2 defines pairs of corresponding residues within secondary structure elements such as helices and strands instead of loops, reporting excellent superpositions that are relevant to infer structural relationships.

3.2.2.5. Acknowledgements

This work has been supported by CONICYT Ph.D. fellowship, FONDECYT 1141172 and CONICYT PIA ACT1408 to FM and BFU-2016-7826-P (MINECO, SP) to DPD. Supported by the H2020-MSCA-RISE project GA No 823886. MdM to CABD

Conflict of Interest: none declared.

3.3. RELACIONES EVOLUTIVAS ENTRE LAS PROTEINAS DE CUBIERTA MEMBRANA (MC) SEGÚN SUS COMPARACIONES ESTRUCTURALES

3.3.1. Resumen

A continuación, se describen los resultados obtenidos al evaluar las similitudes estructurales entre las proteínas de cubierta de membrana con el programa MOMA2. Gracias a esta herramienta hemos desarrollado una nueva clasificación de estas proteínas según los distintos tipos de dominios que las componen, dejando de lado, la clasificación arbitraria que define a las proteínas de cubierta de membrana según su tipo de arquitectura y los elementos que comparten en común. Además, mediante esta nueva clasificación, y por primera vez, se realizó una reconstrucción filética de estas proteínas según sus comparaciones estructurales, permitiéndonos corroborar, extender y proponer nuevas relaciones entre estas proteínas incluso entre complejos en los cuales sus subunidades no han sido aún cristalizadas. Finalmente, mediante los árboles generados, hemos reformulado un escenario parsimonioso que probablemente explica en grosso modo el posible origen de los diversos complejos de cubierta de membrana que se observan en los organismos eucariotas actuales que se originaron a partir del Protocoatomer ancestral.

3.3.2. Metodología

3.3.2.1. Conjunto de estructuras de proteínas usadas para realizar las comparaciones estructurales

Para evaluar las relaciones estructurales entre las proteínas de cubierta de membrana, se obtuvieron en total de 33 estructuras de la base de datos de la Protein Data Bank (PDB, https://www.rcsb.org/ durante el 2017) que incluyen 12 cadenas

donde está presente el dominio β-propeller de una proteína de cubierta de membrana y 25 cadenas que contienen el dominio α-solenoide o SPAH (Tablas 3 y 4). Adicionalmente, se ocuparon las estructuras de la proteína tachilectina-2 y de la hemoglobina como controles negativos debido a que no comparten un origen en común con las proteínas de cubierta de membrana. La estructura de la proteína tachilectina-2 presenta un dominio β-propeller compuesto por cinco *blades* de cuatro hebras β antiparalelas, en vez de seis o siete *blades* que comúnmente presentan las proteínas de cubierta de membrana, y contiene además inserciones de α-hélices entre sus *blades* (Beisel *et al.*, 1999). En cambio, la estructura de la hemoglobina, a diferencia de los dominios SPAH de las proteínas de cubierta de membrana, posee una estructura compacta y globular compuesta principalmente por α-hélices.

Tabla 1. Conjunto de proteínas utilizadas para explorar las relaciones entre dominios β-propeller de las proteínas de cubierta de membrana.

Nombre de la proteína	Nombre del grupo o complejo	Código PDB	Cadena	Intervalo de la cadena	Largo de la proteína	Organismo
Seh1	NPC	3ewe	C	1-346	349	*Saccharomyces cerevisiae*
Sec13	COPII	3mzk	A	1-297	297	*Saccharomyces cerevisiae*
Sec31	COPII	2pm9	A	1-411	1273	*Saccharomyces cerevisiae*
Clatrina	Clatrina/adaptinas	5m5t	A	4-363	1675	*Bos taurus*
Nup133	NPC	1xks	A	75-477	1156	*Homo sapiens*
Nup170	NPC	5hax	A	74-601	1416	*Chaetomium thermophilum*
Nup157	NPC	4mhc	A	88-646	1391	*Saccharomyces cerevisiae*
Nup120	NPC	3hxr	A	1-486	1037	*Saccharomyces cerevisiae*
COPb'	COPI	2ynp	A	1-601	889	*Saccharomyces cerevisiae*
COPa	COPI	5a1u	C	1-560	1224	*Mus musculus*
IFT80	IFTB	5n4a	A	1-639	765	*Chlamydomonas reinhardtii*
Tachylectin-2	Outgroup*	1tl2	A	2-236	236	*Tachypleus tridentatus*

* La proteína Tachylectin-2 fue empleada como control negativo.
Abreviaciones: IFTB (por “Intraflagellar Transport complex B”), NPC (por “Nuclear Pore Complex”), COPI (por “coatomer I”) y COPII (por “coatomer II”).

Tabla 2. Conjunto de proteínas utilizadas para explorar las relaciones entre dominios α-solenoide de las proteínas de cubierta de membrana.

Nombre de la proteína	Nombre del grupo o complejo	Código PDB	Cadena	Intervalo de la cadena	Largo de la proteína	Organismo
Nup145C	NPC	4xmn	B	149-712	1317	*Saccharomyces cerevisiae*
Nic96	NPC	2qx5	A	186-839	839	*Saccharomyces cerevisiae*
Nup85	NPC	4xmm	D	47-744	744	*Saccharomyces cerevisiae*
Nup84	NPC	3iko	C	7-442	726	*Saccharomyces cerevisiae*
Nup120	NPC	4xmn	E	553-744	1037	*Saccharomyces cerevisiae*
Nup133	NPC	3i4r	B	518-1156	1156	*Homo sapiens*
Nup170	NPC	5haz	A	851-1416	1416	*Chaetomium thermophilum*
Nup188N	NPC	4kf7	A	1-1160	1827	*Myceliophthora thermophila*
Nup188C	NPC	4kf8	A	1445-1823	1827	*Myceliophthora thermophila*
Nup192	NPC	5hb4	B	179-1677	1756	*Chaetomium thermophilum*
COPb'	COPI	3mkq	A	586-812	889	*Saccharomyces cerevisiae*
COPa	COPI	3mkq	B	642-818	1201	*Saccharomyces cerevisiae*
COPe	COPI	3mv2	B	1-293	296	*Saccharomyces cerevisiae*
Sec31	COPII	3mzl	B	370-746	1273	*Saccharomyces cerevisiae*
Clatrina	Clatrina/Adaptinas	3lvg	A	1078-1630	1675	*Bos taurus*
AP1g	Clatrina/Adaptinas	1w63	A	1-590	822	*Mus musculus*
AP1b	Clatrina/Adaptinas	1w63	B	2-584	949	*Rattus norvegicus*
AP2a	Clatrina/Adaptinas	2jkr	A	3-623	938	*Mus musculus*
AP2b	Clatrina/Adaptinas	2jkr	B	12-582	937	*Homo sapiens*
Importina 13	NTR	2x19	B	18-954	963	*Homo sapiens*
Importina b1	NTR	2qna	A	130-876	876	*Homo sapiens*

Importina a5	NTR	2jdq	A	84-507	538	*Homo sapiens*
b-catenina	NTR	1jdh	A	135-663	781	*Homo sapiens*
COPg	COPI	5a1u	E	21-874	874	*Mus musculus*
COPb	COPI	5a1u	G	16-953	953	*Mus musculus*
Hemoglobina	Outgroup*	1buw	A	1-141	142	*Homo sapiens*

* La hemoglobina fue usada como control negativo.
Abreviaciones: NTR (por "Nuclear Transport Receptors"), NPC (por "Nuclear Pore Complex"), COPI (por "coatomer I") y COPII (por "coatomer II").

3.3.2.2. Comparaciones estructurales entre las proteínas MC

Las comparaciones estructurales entre los dominios de cada conjunto de proteínas fueron llevadas a cabo con el programa *MOMA2_pw.py* del software MOMA2 usando los parámetros por defecto. Se ejecutaron secuencialmente dos programas disponibles en el software MOMA2: *generate_p1m.py* y *get_ali.py*. El primer programa es usado para generar un script con instrucciones que se puede ejecutar directamente con el programa PyMOL para visualizar las superposiciones estructurales, mientras que el segundo programa es usado para generar un alineamiento estructural a partir los sub-fragmentos que fueron superpuestos con MOMA2. A partir de los resultados obtenidos de estos programas se reconstruyeron casi todas las figuras mostradas en este capítulo. Adicionalmente, se implementó un script en Python que usa la librería *Matplotlib* (Hunter, 2007) para generar *heatmaps* asimétricos a partir de los puntajes obtenidos de las comparaciones realizadas con MOMA2, incluyendo B_{score}, el largo del alineamiento estructural y la medida de Structural Metric Distance (SDM) que se hablará de ella más adelante (ver sección 3.3.2.6).

Finalmente, las imágenes de las superposiciones estructurales generadas para los respectivos ejemplos fueron generadas con el programa PyMOL (Schrödinger, LLC, 2015), a partir del archivo .p1m que se obtiene al comparar un par de estructuras con MOMA2.

3.3.2.3. Clasificación de los dominios según sus comparaciones estructurales.

Para la clasificación de los dominios de las proteínas de cubierta de membrana se construyeron dos redes usando el programa Cytoscape (Shannon *et al.*, 2003), a partir de tablas de diferencias que contenían las comparaciones pareadas de los dominios β-propeller y SPAH, incluyendo solamente aquellos pares que reportaron un valor de SDM menor al umbral de corte (SDM = 72 para los dominios β-propeller y de SDM = 121 para los dominios SPAH). Luego, usando el método de Community cluster (GLay) se determinaron los grupos presentes en las redes creadas (Newman and Girvan, 2004). El método GLay se basa en encontrar los grupos dentro de una red que optimiza el puntaje de modularidad usando el parámetro de "*edge betweenness*". Este parámetro se asocia a los bordes de una red cuyos valores altos sugieren que estos actúan como puentes conectando las partes pobladas, y que su eliminación puede afectar la comunicación entre los grupos de nodos por donde pasan las rutas más cortas. La eliminación de estos bordes daría como resultado la partición de la red en subredes más pequeñas densamente conectadas.

3.3.2.4. Representación gráfica de las relaciones encontradas entre las proteínas MC

En este trabajo, se emplearon los gráficos de *Circos* para representar las similitudes estructurales descubiertas con MOMA2 entre los dominios β-propeller y SPAH de las proteínas de cubierta de membrana (Krzywinski *et al.*, 2009). Para ello, se obtuvieron todas las comparaciones realizadas para cada par de dominios cuyo porcentaje de similitud relativa fuera mayor al 20%, para posteriormente utilizar el largo de sus alineamientos estructurales y sus puntajes de B_{score} para crear las conexiones que se aprecian en estos gráficos. El grosor de cada conexión refleja el largo del alineamiento estructural que fue normalizado a una escala 0 a 1 considerando el alineamiento más largo reportado entre estos dominios. En cambio, el color de cada conexión representa el valor del B_{score} normalizado que fue calculado usando como referencia los valores máximo y mínimo de los puntajes de B_{score} obtenidos a partir de las comparaciones realizadas con MOMA2. Estos valores normalizados fueron clasificados en una escala discreta de colores que va de azul a amarillo para señalar las conexiones que presentan una fuerte o débil similitud estructural según las comparaciones realizadas a nivel de sus elementos de estructura secundaria. Adicionalmente, se realizaron comparaciones de perfiles de HMM (*hidden Markov models*) a partir de las secuencias de los dominios de las proteínas MC para analizar si estas relaciones se encuentran mediante comparaciones de secuencia. Para ello, primero se crearon los perfiles de HMM realizando búsquedas con el programa *jackhmmer* (que se encuentra disponible en el software HMMER versión 3.3) contra la base de datos de secuencia de UniRef50 (usando 2 iteraciones

y e-value < 1e-5 como umbral de corte) usando las secuencias de los dominios de las proteínas MC, para luego con el programa *hhmake* de HH-suite construir los perfiles de HMM. Luego, se realizaron comparaciones de todos los perfiles HMM contra todos usando el programa *hhsearch* para recuperar los *hits* que reportaron un e-value < 1e-5. Las relaciones encontradas mediante este método fueron destacadas en azul oscuro en los gráficos de *Circos*, mientras que las relaciones descritas anteriormente en la literatura fueron destacadas en rojo oscuro.

3.3.2.5. Análisis de los motivos conservados de las proteínas MC

Para evaluar los residuos conservados entre los grupos definidos por cada dominio β-propeller, primero se obtuvieron los alineamientos múltiples a nivel de secuencia de las superposiciones estructurales generadas con MOMA2 a partir de los elementos de un grupo con respectivo su centroide. El centroide del grupo es aquel elemento que posee la menor distancia con los demás elementos de su grupo según el puntaje obtenido a partir de las comparaciones flexibles realizadas con MOMA2 (100 - Sr). Por consiguiente, se corrió una iteración del programa *jackhmmer* usando la secuencia del centroide para identificar las secuencias de sus proteínas homólogas cercanas (Johnson *et al.*, 2010). Luego, usando el servidor de CLUSTALW (https://www.genome.jp/tools-bin/clustalw) se generó un alineamiento múltiple de secuencia con las secuencias del centroide y de las proteínas homólogas (Thompson *et al.*, 1994). Posteriormente, se combinaron ambos alineamientos múltiples de forma manual usando el programa *Jalview*, considerando los segmentos alineados con respecto al centroide (Waterhouse *et al.*, 2009). Con *Jalview* se pudo filtrar las secuencias que presentaban un porcentaje de identidad de secuencia mayor

o igual a un 80%. El alineamiento múltiple resultante fue representado visualmente usando el servidor ESPript 3.0 que permite renderizar las similitudes de secuencia y la información de estructura de secundaria del centroide para propósito de análisis y publicación (Robert and Gouet, 2014). Para representar los residuos conservados en los alineamientos múltiples se usó el esquema de colores basado en las similitudes calculadas con la matriz de sustitución de Risler (Risler *et al.*, 1988), que está disponible en ESPript 3.0, donde se ocuparon umbrales de cortes de 0.7 o 0.75 para destacar los motivos encontrados entre los grupos BI y BII, o entre algunos dominios comparados, respectivamente.

3.3.2.6. Generación de los árboles filogenéticos a partir de comparaciones estructurales

Para crear los árboles filogenéticos a partir de las comparaciones estructurales, primero se realizaron comparaciones de todos contra todos de las estructuras que forman parte de conjuntos de dominios β-propeller y α-solenoide de las proteínas de cubierta de membrana usando MOMA2. Luego, por cada conjunto de dominios, se crearon 100 réplicas donde a una estructura escogida al azar se le removía un elemento de estructura secundaria para luego repetir las comparaciones pareadas de todos los dominios contra todos. De esta manera, se agregó variabilidad a los datos originales y a la vez nos permitió asegurar la fiabilidad de las similitudes estructurales calculadas con MOMA2.

Por consiguiente, se obtuvieron los puntajes de las comparaciones realizadas con los datos originales y las réplicas para construir tablas de diferencias que fueron utilizadas posteriormente para calcular los dendrogramas. Estas tablas contienen los

valores de distancia de acuerdo con la métrica descrita por Johnson *et al.* en 1990 para construir árboles filogenéticos usando alineamientos estructurales. Esta métrica llamada *Structural Distance Metric* (SDM) toma las equivalencias topológicas y los valores de RMSD de los residuos estructuralmente equivalentes para calcular su contribución en el alineamiento estructural usando la fórmula descrita a continuación:

$$SDM = -100\ln(w_1 PFTE + w_2 SRMS)$$

Donde PFTE, en este caso, está definido por el largo del alineamiento estructural al comparar un par de proteínas dividido por el máximo largo obtenido al alinear todos los pares de proteínas del conjunto, reportando en cada caso, valores que van de 0 hasta 1. En cambio, SRMS es el RMSD ponderado obtenido con MOMA2 al comparar un par de estructuras y que es convertido a un puntaje de similitud mediante la siguiente ecuación:

$$SRMS = 1 - \frac{RMSD_{pond}}{3.5}$$

donde 3.5 Å es la distancia umbral que fue usada para definir equivalencias entre un par de residuos alineados según la distancia presente entre sus carbonos α. Mientras que los pesos w_1 y w_2 son calculados a partir de las siguientes ecuaciones:

$$w_1 = [(1 - PFTE) + (1 - SRMS)]/2$$

$$w_2 = (PFTE + SRMS)/2$$

donde los pesos deben cumplir la siguiente igualdad, $w_1+w_2 = 1$ (Johnson *et al.*, 1990).

Posteriormente, las estructuras de cada conjunto o réplica son agrupadas con el programa *Neighbor* disponible en el software PHYLIP. Este programa recibe como entrada las tablas de diferencias calculadas en el paso anterior e implementa el método de agrupamiento jerárquico de UPGMA para generar los árboles a partir de los datos originales y las réplicas (Felsenstein, 1993). Por consiguiente, a partir de los 100 dendrogramas obtenidos de las réplicas se calculan los valores de soporte de los nodos de cada árbol usando el programa *nw_support* disponible en el repositorio de programas de Newick Utilities (Junier and Zdobnov, 2010). Finalmente, para representar visualmente los dendrogramas con sus valores de soporte se ocupó el servidor web de iTOL (Letunic and Bork, 2007).

3.3.2.7. Análisis de parsimonia de las proteínas MC

Para evaluar el número de transiciones evolutivas de las características observadas en los dendrogramas, se utilizó un script de R que usa la función "*phylo.signal.disc*" desarrollada por Enrico Rezende (http://grokbase.com/k-for-discrete-unordered-traits) para comparar el número mínimo de cambios de estado de los caracteres observados con respecto a una distribución generada a partir de un modelo nulo generado con el método de Maddison y Slatkin. Este modelo se genera al variar un número "n" de veces al azar los caracteres de los árboles para generar una distribución empírica de los posibles pasos evolutivos considerando el número de elementos agrupados en el dendrograma y su clasificación (Maddison and Slatkin, 1991). Este script calcula un *p-value* para el número mínimo de transiciones observadas a la vez que entrega como salida una imagen que muestra un histograma

con la distribución de las transiciones obtenidas al variar de forma azarosa los estados del dendrograma. Esta imagen indica mediante una flecha de color rojo el número mínimo de transiciones observadas para los dendrogramas analizados en la distribución de estados. Debido a que este método no es determinista, por cada conjunto de dominios se corrió 100 veces esta función para identificar el número promedio de transiciones observadas y determinar los valores de *p-value* en cada caso.

3.3.2.8. Agradecimientos

Agradecimientos al Profesor Enrico L. Rezende por facilitarnos el script en R que fue usado para el análisis de las transiciones evolutivas de los caracteres observados en los dendrogramas, además por prestarnos su apoyo en este tipo de análisis y por las múltiples discusiones científicas que tuvimos hablando acerca de este proyecto.

3.3.3. Resultados

3.3.3.1. Clasificación funcional y estructural de las proteínas MC

Como hemos visto anteriormente en la sección 1.11, las proteínas de cubierta de membrana se clasifican actualmente en dos o tres familias según la forma de su arquitectura, si es clase I y II, o si pertenece a los tipos COPI-like, COPII-like y adaptin-like. Ahora nosotros proponemos una nueva clasificación funcional de las subunidades MC considerando si estos son elementos rígidos o móviles, y si se encuentran específicamente en los complejos de cubierta de vesículas como COPI o COPII, o en el poro nuclear (Figura 14). De acuerdo con esta clasificación, las subunidades pueden ser agrupadas en cuatro clases que hemos denominado como: Cage-CV, Adaptor-CV, Cage-NPC y Adaptor-NPC. Entre estas subunidades pueden existir dos tipos de relaciones evolutivas: 1) relaciones "trans-complex" o 2) relaciones "cis-complex". Las relaciones "trans-complex" se refieren a aquellas relaciones entre los componentes estables o "*core*" de los distintos complejos, por ejemplo, entre las nucleoporinas (Nups) que forman parte del poro nuclear y entre las subunidades del complejo COPI. En cambio, las relaciones "cis-complex" se refieren a las relaciones estructurales entre los componentes estables y periféricos que conforman cada uno de los complejos. Por ejemplo, entre las Nups y las carioferinas o entre las adaptinas y la clatrina.

Actualmente es ampliamente aceptado que las proteínas que conforman parte del "*core"* del poro nuclear y el *cage* de las cubiertas de las vesículas comparten un origen en común, cuya evidencia disponible apoya la hipótesis del Protocoatomer (Devos *et al.*, 2004) (Figura 14). No obstante, actualmente no existe evidencia concluyente para afirmar que los adaptadores y los componentes del *core* del poro nuclear están

relacionados (Nup192 y Nup188 con las carioferinas, cuya relación es descrita en la hipótesis del "core/adaptor"). Todo pese a que recientes estructuras cristalográficas han sugerido esta posibilidad (Andersen *et al.*, 2013; Stuwe *et al.*, 2014) (Figura 14).

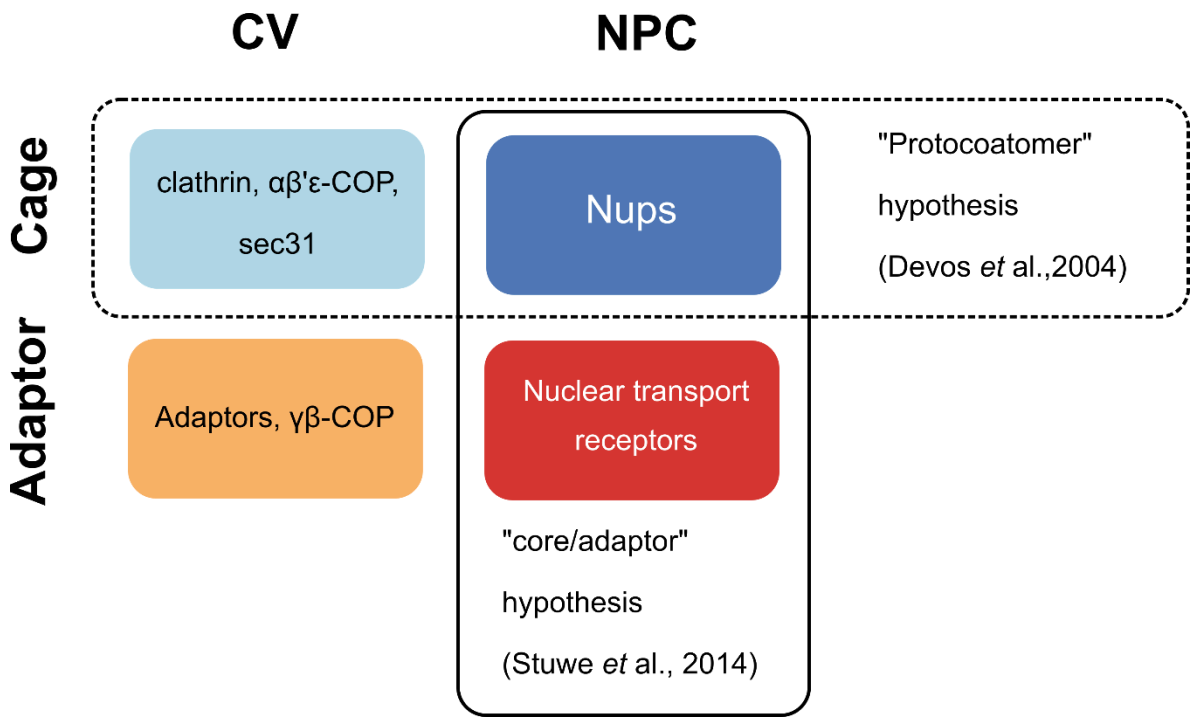


Figura 14. Clasificación funcional de las subunidades de los complejos MC.
Este esquema destaca cuatro grupos principales entre las subunidades MC donde los elementos rígidos "Cage" son mostrados en azul claro y oscuro para indicar si pertenecen a los complejos de cubierta (CV) o al poro nuclear (NPC). Las subunidades "Adaptor" se destacan con naranjo y rojo para indicar si pertenecen a los complejos de CV o el NPC, respectivamente. Las similitudes estructurales descritas en la literatura entre los grupos Cage-CV y Cage-NPC dar soporte a la hipótesis del Protocoatomer, mientras que las similitudes encontradas entre las nucleoporinas Nup188 y Nup192 con las carioferinas dan sustento a la hipótesis del "core/adaptor". Estas relaciones son destacadas mediante rectángulos curvos que agrupan los grupos señalados.

Las proteínas MC también se pueden clasificar según las similitudes encontradas entre los diferentes tipos de dominios que poseen. Como hemos visto anteriormente, estas proteínas poseen dos tipos de dominios β-propeller y SPAH,

donde algunas subunidades poseen ambos, o sólo uno de ellos. En otros casos, la región N-terminal se subdivide en dos dominios β-propeller o interactúa con un β-propeller incompleto. En cambio, algunas proteínas MC presentan dominios SPAH que poseen formas alargadas, en otros se vuelven sobre sí mismo adoptando la forma de J, o con forma de "*donuts*" o super-hélice (Figura 17). Por ello, para proporcionar más información acerca de las relaciones estructurales entre las proteínas MC se realizaron comparaciones estructurales con MOMA2 entre todos los dominios que han sido cristalizados hasta la fecha (Ver sección 3.3.2.1, Tablas 1 y 2). Estos dominios fueron extraídos de las estructuras disponibles en la base de datos de la PDB que pertenecen a 8 especies diferentes, donde la mayoría de ellas provienen de proteínas que se encuentran en las *S. cerevisiae* y *H. sapiens.* Para describir las relaciones estructurales entre los dominios analizados se utilizó una similitud reportada a partir del alineamiento de matrices (B_{score}) y las métricas derivadas de las superposiciones estructurales como el largo del alineamiento estructural o la métrica de SDM. El B_{score} da cuenta de la similitud estructural presentes entre los elementos de estructura secundaría, mientras que el largo del alineamiento da cuenta de la similitud presente entre los dominios según sus superposiciones estructurales. En cambio, la métrica SDM es una medida de distancia más precisa que se usó para generar los árboles filogenéticos según las comparaciones estructurales y para definir los grupos entre los dominios MC. Los resultados de estas comparaciones son mostrados de forma compacta mediante mapas de calor o *heatmaps*, cuyas filas y columnas fueron ordenadas usando la medida de SDM y el algoritmo de clustering UPGMA (Figura 15). En estos mapas podemos apreciar las relaciones *trans-complex*

entre los dominios β-propeller que pertenecen a las subunidades Cage-CV y Cage-NPC. Así también las relaciones *trans-complex* de los dominios SPAH entre las subunidades que pertenecen a los grupos Adaptor-CV y Adaptor-NPC, y las relaciones *cis-complex* entre las subunidades del tipo Cage-NPC y Adaptor-NPC (Figura 15A). Mediante inspección manual podemos distinguir en los *heatmaps* para las métricas de SDM al menos dos grupos bien definidos para los dominios β-propeller y dos grupos para los dominios SPAH (Figura 15B). Por ello, para definir de forma clara y precisa los grupos presentes entre los dominios MC, se realizó un análisis de redes con el programa Cytoscape considerando los grupos que se forman entre los dominios cuyas conexiones reportan distancia menor a un umbral de corte (SDM = 72 para los dominios β-propeller y SDM = 121 para los dominios SPAH). Luego usando el algoritmo *Community cluster* (GLay) identificamos tres grupos entre los dominios β-propeller que hemos denominado como BI, BII y BIII, y con respecto a los dominios SPAH, hemos identificado cuatro grupos que hemos nombrado respectivamente como AI, AII, AIII y AIV (Figura 16).

En la red formada para los dominios β-propeller, se observa que los dominios derivados de las subunidades presentes en los complejos COPI, COPII, SEA y IFT80 incluyendo también los β-propeller incompletos presentes en el poro nuclear fueron clasificados en el grupo BI. Dentro de los dominios del tipo BII, encontramos sólo aquellos que están presentes en algunas nucleoporinas del poro nuclear como Nup170, Nup107, Nup133 y Nup120. En cambio, el dominio β-propeller de Clatrina y Nup37 forman el grupo BIII, donde Nup37 es más cercano a los dominios β-propeller de las subunidades de los complejos COPI y COPII (Figura 16).

En la red formada a partir de los dominios SPAH, vemos que los dominios α-solenoide de las subunidades de COPI se conectan a la red principal a través de Nup133 (por medio de Nup170, COPγ y AP1B) formando el grupo AI. Los miembros del grupo AI presentan varios pares repetidos α-hélices de un tamaño similar que en algunos casos adoptan una forma recta (COPβ' y COPα) o forman un patrón de zigzag (Nup133). El segundo grupo llamado AII incluye todos aquellos dominios que Brohawn y colaboradores señalaron que derivan de un pliegue ancestral denominado ACE1, incluyendo en este grupo las nucleoporinas Nic96, Nup145C, Nup84 y Nup85, junto con las subunidades Sec31 y Sec16 que forman parte del complejo coatomer II (Brohawn *et al.*, 2008a; Brohawn and Schwartz, 2009) (Figura 16). Entre sus características estructurales vemos que los dominios BII se curvan sobre sí mismos adoptando una forma de J donde su extremo N-terminal interactúa con la parte central del dominio (Figura 17). Mientras que el tercer grupo llamado AIII se encuentra compuesto por los dominios SPAH presentes principalmente en los complejos adaptadores, proteínas del transporte nuclear y algunas nucleoporinas que forman parte del *scaffold* del poro nuclear. Mediante comparaciones estructurales hemos podido agrupar las nucleoporinas Nup192 y Nup188 junto con las carioferinas las cuales se ha sugerido previamente que comparten una relación evolutiva (Andersen *et al.*, 2013), incluyendo además los dominios α-solenoide de las adaptinas AP1 y AP2 que comparten una homología distante de secuencia y una similitud estructural con las subunidades del sub-complejo adaptador de COPI (COPβ y COPγ) (Faini *et al.*, 2013). Estas estructuras evaluadas con MOMA2 reportaron una fuerte similitud estructural tanto a nivel de

elementos de estructura secundaria como a nivel de residuos, agrupando también los dominios SPAH de aquellas subunidades que poseen la arquitectura β-propeller/α-solenoide como Nup170, Nup120 y Clatrina, siendo los ejemplos más divergentes de este grupo.

Finalmente, COPε se encuentra aislado de la red de los dominios SPAH constituyendo un *singleton* (Figura 16). Esta proteína es la subunidad más pequeña de COPI, cuyas repeticiones TPR (*tetratricopeptide repeat*) adoptan la forma de una super-hélice (Figura 17).

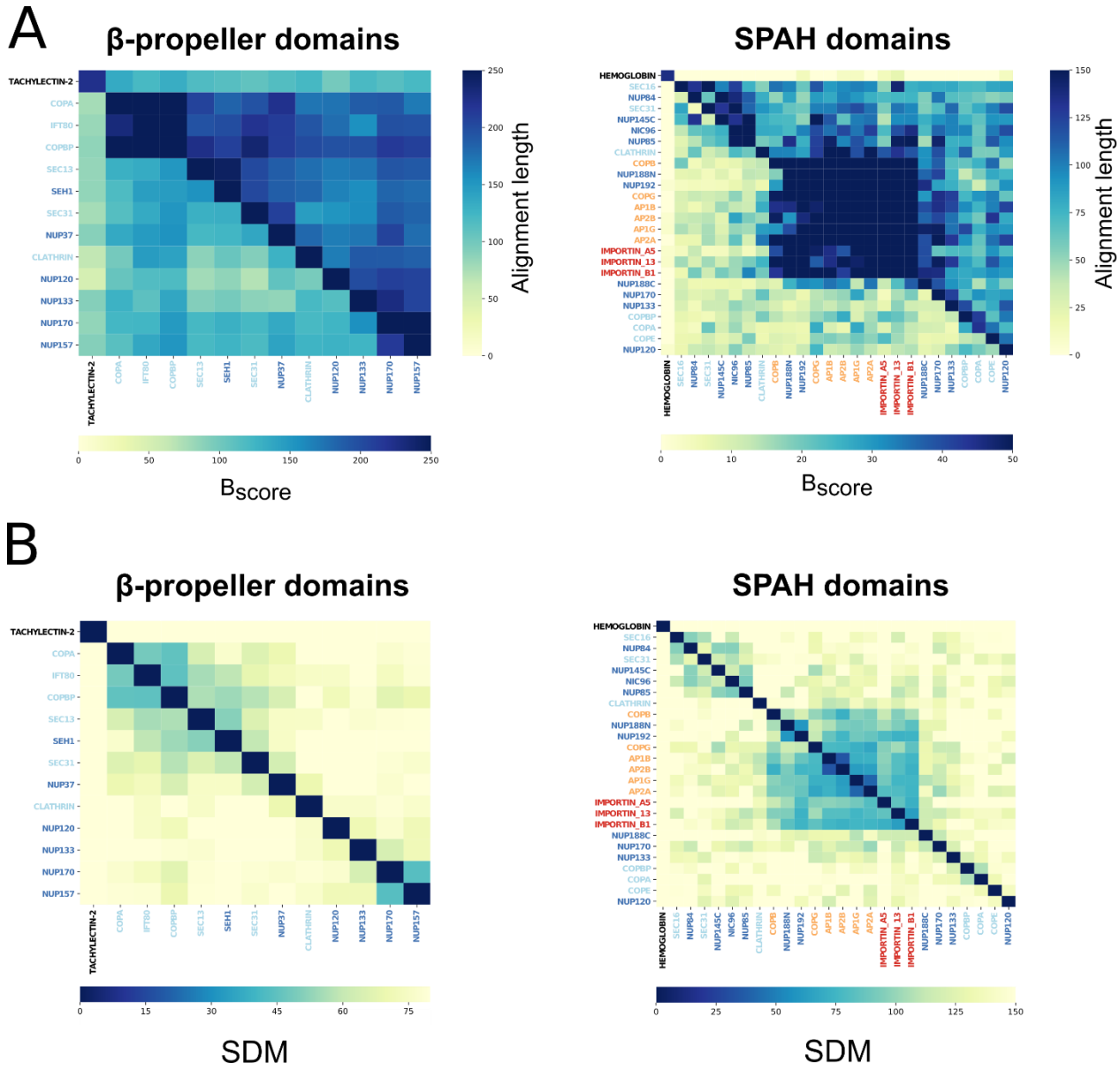


Figura 15. Similitudes estructurales entre los dominios de las subunidades rígidas y móviles entre las proteínas MC.

Los *heatmaps* muestran los puntajes de similitud estructural obtenidos al comparar las matrices (B_{score}, triángulo inferior) y el número total de residuos alineados (*alignment length*, triángulo superior) de las superposiciones estructurales (A), incluyendo también los valores obtenidos con la métrica de distancia SDM (B) a partir de las comparaciones realizadas entre los dominios β-propeller y SPAH con MOMA2. Estos valores son representados en una escala de amarillo a azul indicando una similitud estructural débil a fuerte (o una distancia estructural cercana o distante) según los puntajes obtenidos de sus comparaciones estructurales. Las etiquetas de las proteínas analizadas son coloreadas en azul claro y oscuro para indicar que pertenecen a los grupos Cage-CV y Cage-NPC, y coloreadas en naranjo y rojo para indicar que pertenecen a los grupos Adaptor-CV y Adaptor-NPC, respectivamente. Los códigos PDB de las estructuras analizadas en los *heatmaps* son descritos en la Tablas 1 y 2 de la sección 3.3.2.1.

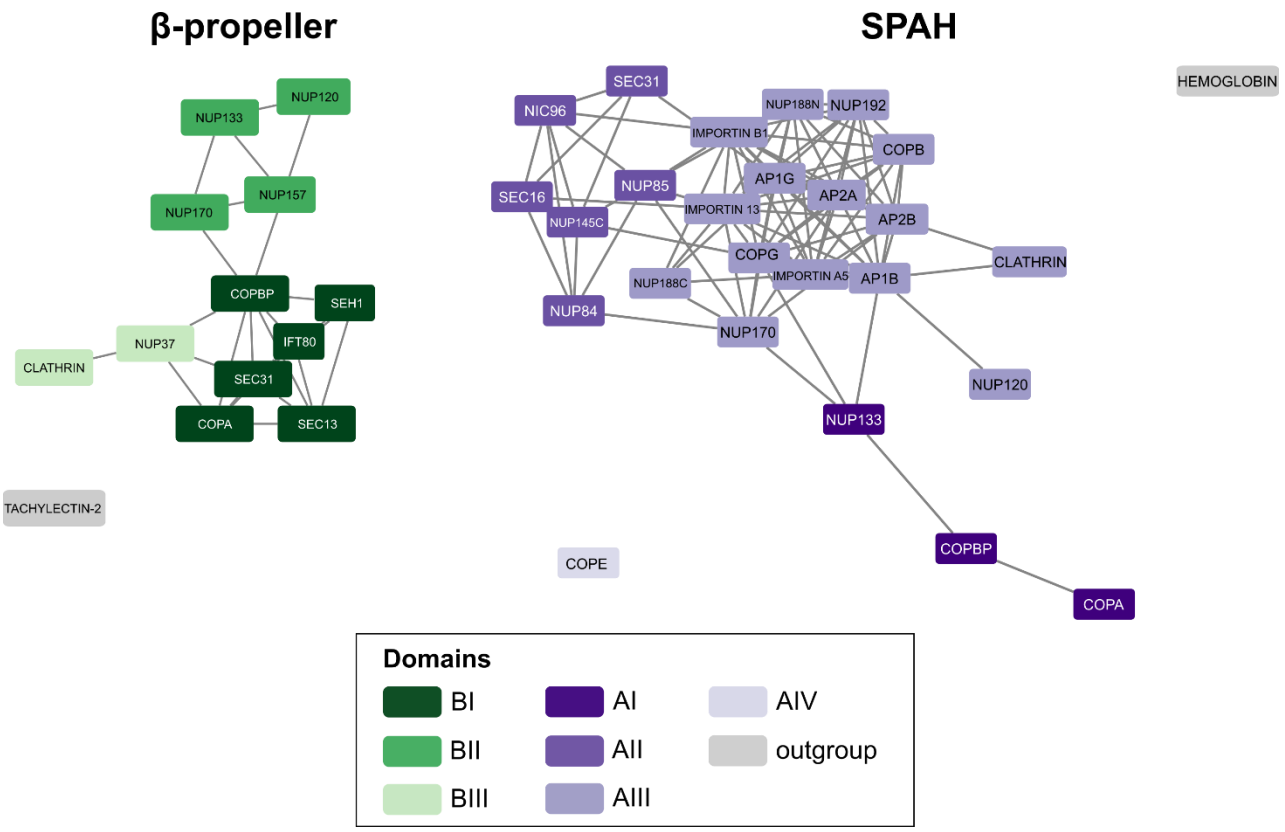


Figura 16. Clasificación estructural de los dominios presentes entre las proteínas MC. La figura muestra dos redes generadas con Cytoscape que describen las conexiones estructurales encontradas entre los dominios β-propeller y SPAH según sus valores de SDM. Los grupos encontrados entre los dominios β-propeller fueron descritos usando un gradiente de verdes, en cambio, los grupos entre los dominios SPAH fueron descritos usando un gradiente de morados. Las proteínas Tachilectina-2 y hemoglobina fueron ocupadas como controles negativos en estas redes siendo indicadas en gris, respectivamente.

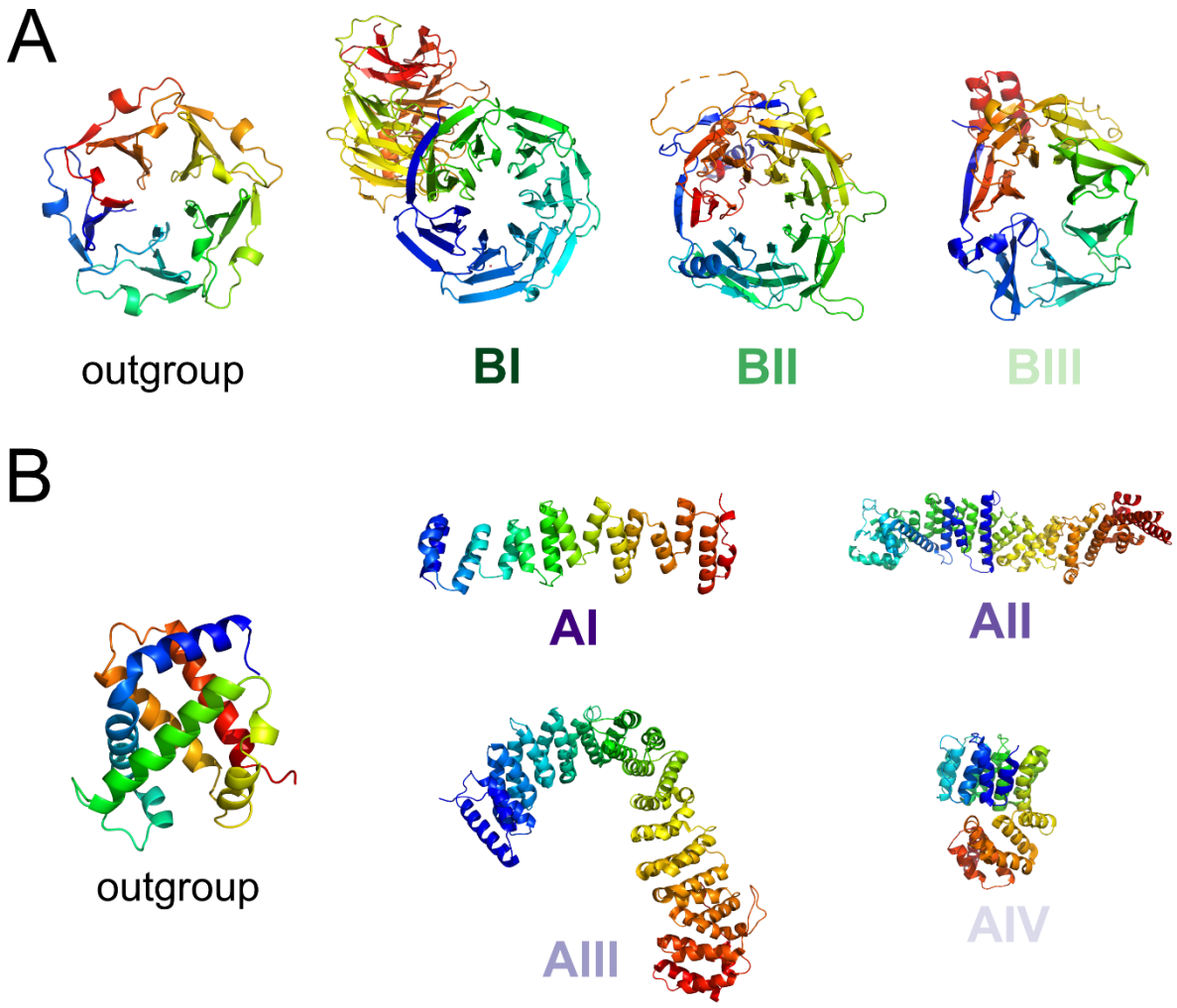


Figura 17. Representantes de los diferentes tipos de dominios encontrados en las proteínas MC.

Los representantes de los grupos encontrados entre los dominios β-propeller (A) corresponden a los dominios presentes en las proteínas COPβ', Nup170 y Clatrina (códigos PDB 2ynp, 5hax y 5m5t), incluyendo a parte el dominio irregular de la Tachilectina-2 que se usó como control negativo (código PDB 1tl2). Entre los dominios SPAH (B), los representantes de cada grupo corresponden a los dominios α-solenoides presentes en las proteínas COPβ', Nic96, AP1γ y COPε (códigos PDB 3mkq, 2qx5, 1w63 y 3mv2), incluyendo a parte la hemoglobina (1buw) que fue usada como control negativo. Estas estructuras fueron coloreadas en *rainbow* con el programa PyMOL para indicar las posiciones de inicio y término de sus estructuras.

3.3.3.2. Análisis estadístico de las proteínas MC

Los resultados obtenidos de las comparaciones realizadas entre los dominios MC muestran que los dominios β-propeller son estructuralmente muy similares entre sí a pesar de que comparten una baja señal de secuencia entre ellos. El análisis estadístico de las comparaciones pareadas muestra que los dominios β-propeller estos poseen una alta similitud estructural reportando porcentajes de similitud relativa y de *overlap* estructural promedio entre 46% a un 56% a pesar de que estos reportan porcentajes de identidad de secuencia promedio menores al 15%, indicando que, a nivel de sus secuencias aminoacídicas, estos dominios han divergido considerablemente (Tabla 3). Los dominios del tipo BI poseen un mayor porcentaje de *overlap* estructural promedio que los dominios del tipo BII, mientras que los porcentajes de similitud relativa junto con sus desviaciones estándar no muestran una enorme diferencia entre ellos, sugiriendo de que las distribuciones de estos valores son relativamente similares (Tabla 3). También, se puede apreciar que los dominios del grupo BI poseen un mayor porcentaje de identidad promedio en comparación con los dominios del grupo BII mostrando a la vez un menor grado de dispersión, sugiriendo que hay una mayor divergencia a nivel de secuencia entre los dominios β-propeller de las nucleoporinas que pertenecen a la clase BII que entre los dominios de la clase BI que pertenecen a subunidades de diferentes complejos (Tabla 3).

Tabla 3. Estadísticas de las comparaciones estructurales realizadas entre los dominios β-propeller.

Grupos	BI	BII	Todos
N	6	4	12
%id seq.	20.07±4.3	14.88±7.6	13.37±5.6
%Sr	52.47±11.0	54.53±10.5	47.05±9.2
%SO	65.16±7.5	58.24±10.5	55.34±9.8

En cada celda se muestra el valor promedio de cada puntaje ± su desviación estándar.
N: número de elementos, id seq.: identidad de secuencia, Sr: similitud relativa y SO: estructural *overlap*

Con respecto a los dominios SPAH, solamente hemos realizado una comparación estadística de las similitudes reportadas para los miembros de los grupos AI, AII y AIII descartando el grupo AIV debido a que cuenta con solo un elemento (Tabla 4). Estos resultados indican que todos los dominios SPAH poseen un bajo porcentaje promedio de identidad de secuencia, aproximadamente de un 10% mostrando una baja dispersión. Por otro lado, los dominios SPAH dan cuenta de una mayor variación estructural reportando porcentajes promedio de similitud relativa y de *overlap* estructural aproximadamente de 23% y 30% respectivamente (Tabla 4). Estos resultados muestran que los dominios SPAH poseen una baja similitud a nivel de secuencia y estructura en comparación con los dominios β-propeller. Analizando de forma individual cada grupo de dominio vemos que los dominios del tipo AI poseen un mayor porcentaje de *overlap* estructural promedio que los dominios del tipo AII y AIII (SO promedio = 55% aprox.), esto se puede explicar debido a que este grupo solamente cuenta con tres elementos y el mayor

porcentaje de *overlap* está dado por la comparación entre los dominios COPβ' y COPα que reportan una fuerte similitud estructural. Mientras que los dominios del tipo AII muestran un menor grado de dispersión de su valor de similitud relativa promedio con respecto a los grupos AI y AIII (Tabla 4). También, se puede apreciar que los dominios del grupo AII y AIII donde ambos reportan un porcentaje de identidad promedio similar pero el grupo AIII muestra una mayor dispersión (Tabla 4).

Tabla 4. Estadísticas de las comparaciones estructurales realizadas entre los dominios SPAH.

Grupos	AI	AII	AIII	Todos
n	3	6	15	25
%id seq.	11.79±5.4	12.50±3.7	12.02±8.8	10.23±5.9
%Sr	38.67±19.5	28.73±8.5	28.37±13.9	22.67±11.1
%SO	55.00±14.3	33.70±9.4	34.75±14.1	30.00±12.6

En cada celda se muestra el valor promedio de cada puntaje ± su desviación estándar.
N: número de elementos, id seq.: identidad de secuencia, Sr: similitud relativa y SO: estructural *overlap.*

3.3.3.3. Motivos conservados entre las proteínas MC

A pesar de que las proteínas de cubierta de membrana poseen bajos porcentajes de identidad, aún es posible identificar motivos conservados entre ellos, especialmente a nivel de sus dominios β-propeller que son los más conservados estructuralmente. Para identificar estas regiones en los grupos BI y BII, hemos construido alineamientos múltiples de secuencias considerando los alineamientos pareados de los dominios de cada grupo con su respectivo centroide incluyendo también información adicional de las secuencias de aquellas proteínas que presentan homología de secuencia con el centroide (Figuras 18 y 20). De manera parecida, se realizó el mismo tipo de análisis para el grupo BIII, que cuenta únicamente con los dominios de Nup37 y Clatrina, pero enriqueciendo el alineamiento estructural producido con MOMA2 con las secuencias de las proteínas homólogas a estas dos proteínas (Figura 22).

El alineamiento múltiple generado a partir de los dominios β-propeller del tipo BI reveló la presencia de varias regiones conservadas en las hebras β que forman parte del segundo *blade* hasta el séptimo, exceptuando las proteínas Seh1 y Sec13 que poseen seis *blades* en lugar de siete (Figura 19). Estos motivos están constituidos principalmente por grupos de aminoácidos hidrofóbicos como Val, Leu, Ile, Trp y Ala, cuyas cadenas laterales se alejan de las estructuras de las hojas β para interactuar con las cadenas laterales de los aminoácidos presentes en los *blades* cercanos (Figuras 18 y 19). El alineamiento múltiple de secuencias adicionalmente señala la presencia de columnas de aminoácidos cargados o polares como Lys, Thr, Ser y Asp, cuyas cadenas se orientan al exterior o al interior del anillo formando por las *blades*

(Figuras 18 y 19). Mediante estas comparaciones se aprecia que la mayoría de los residuos conservados poseen una función propiamente estructural manteniendo la forma compacta y circular que poseen los dominios del tipo BI. Interesantemente, en el alineamiento múltiple de secuencias de los miembros del grupo BI, se distingue la conservación de residuos idénticos compuestos principalmente por histidinas en una sola columna del alineamiento múltiple. Estas histidinas se encuentran en la misma posición en las estructuras de todos los miembros del grupo BI. Estos residuos conservados están presentes en los *loops* que conectan la cuarta hebra del segundo *blade* con la primera hebra del tercer *blade* en los dominios β-propeller de COPβ', Sec31 y IFT80 (Figura 19). Por otra parte, en los dominios β-propeller incompletos con seis *blades* como la proteína Seh1, la histidina conservada se encuentra entre el *loop* que conecta el primer *blade* con el segundo *blade* (Figura 19). Las cadenas laterales de estos residuos están orientadas hacia los grupos hidroxilos de un par de serinas que se encuentran en algunos dominios BI o hacia el par formado por el grupo hidroxilo de la serina con el grupo carboxilo de la glicina (Figura 19). Estos residuos se encuentran en la segunda hebra del tercer *blade* (COPβ', COPα, IFT80, Sec31) o segundo *blade* (solo en Seh1 y Sec13) de los dominios del grupo BI. En el alineamiento múltiple de secuencia, estos residuos se mapean en las columnas conservadas que aparecen debajo de la hebra β11 del primer subdominio de COPβ' (Figura 18). Cabe la posibilidad que estos residuos sirvan para mantener el pliegue del dominio apoyando la posibilidad de homología entre estas proteínas.

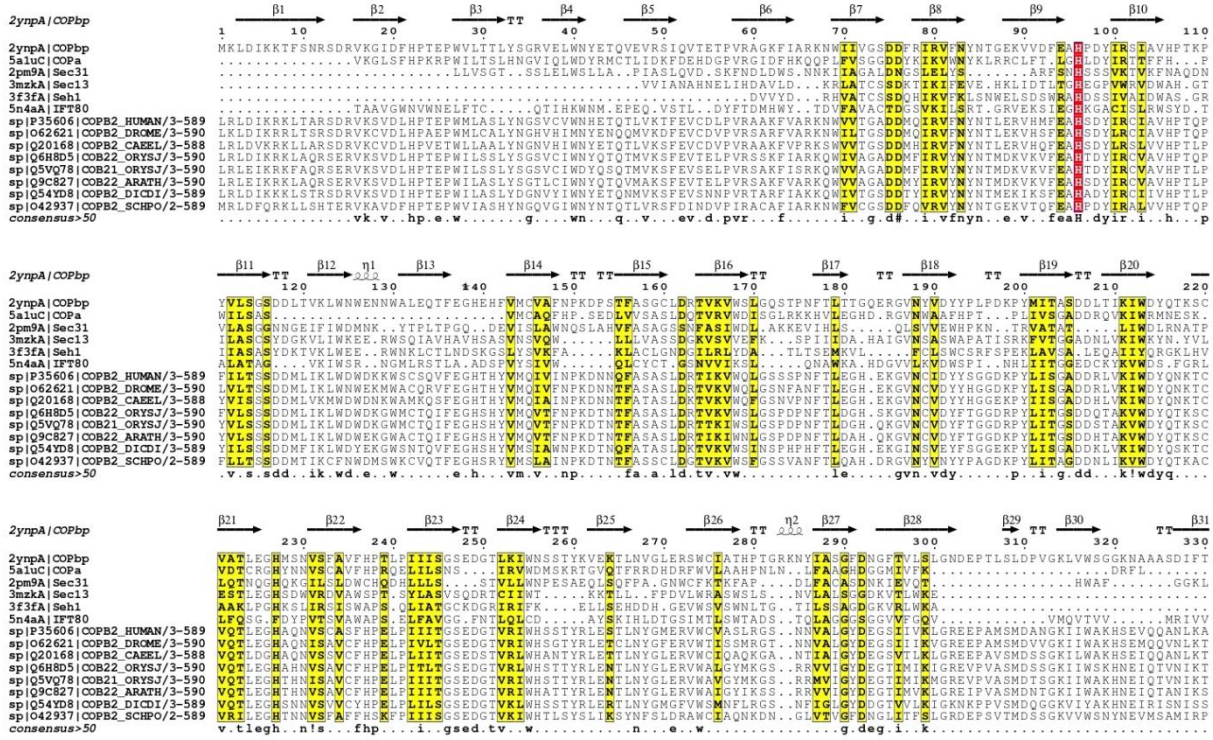


Figura 18. Alineamiento múltiple de los dominios del tipo BI según sus comparaciones estructurales.

Alineamiento múltiple generado a partir de las superposiciones estructurales realizadas contra el dominio β-propeller de COPβ' (el primer subdominio) a la vez que se alinearon las secuencias de sus proteínas homólogas. En el alineamiento múltiple, los residuos conservados en las columnas son destacados en amarillo y los residuos idénticos en rojo. En la parte superior de esta figura, se ilustran los elementos de estructura secundaria del centroide del grupo BI. Esta figura fue generada con el servidor de ESPript 3.0 usando el esquema de similitud de Risler y el umbral definido por defecto (Global score >= 0.7) para identificar los motivos conservados.

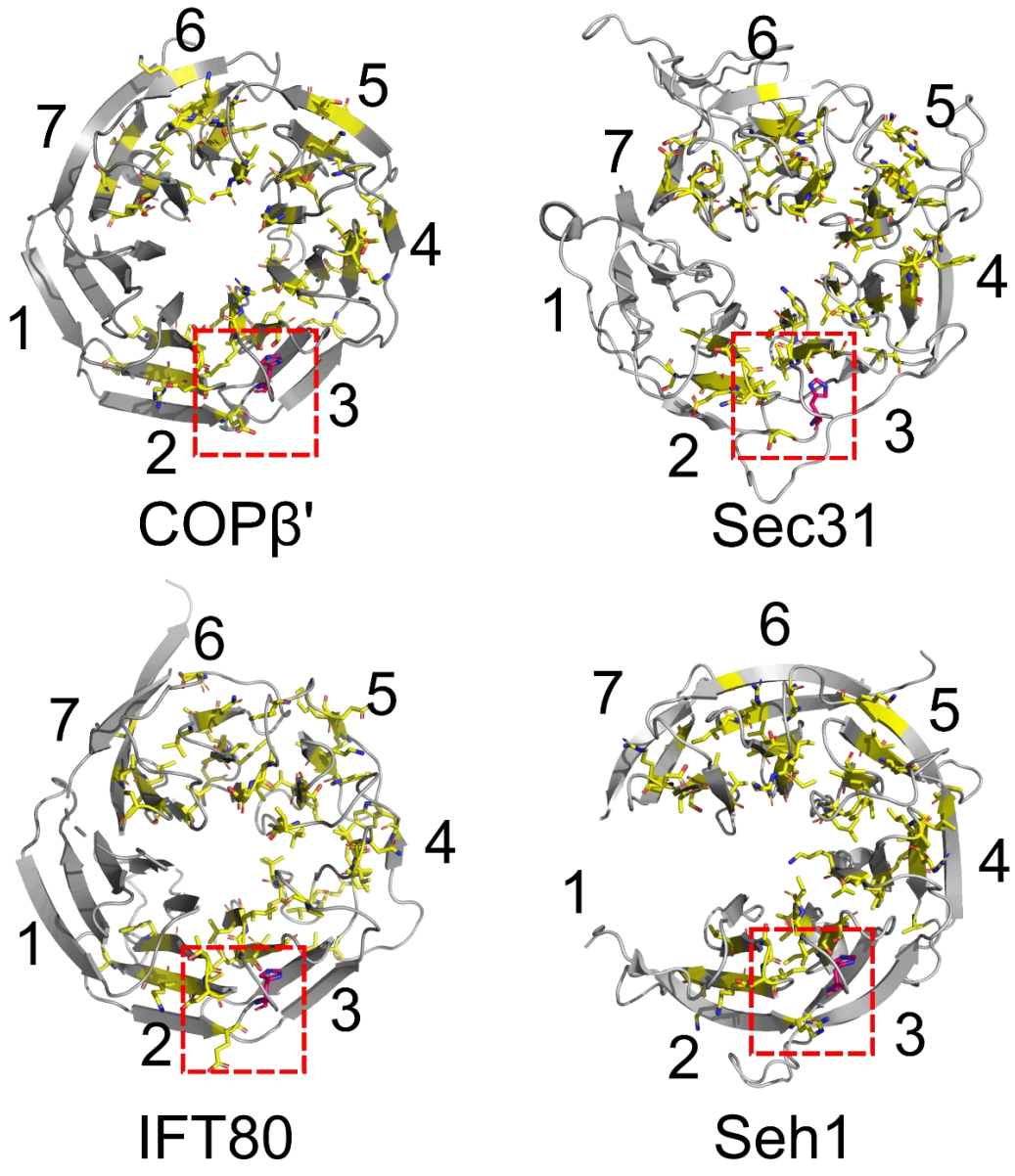


Figura 19. Residuos estructuralmente conservados entre los dominios β-propeller del tipo BI.

Los dominios β-propeller descritos en esta figura pertenecen a los complejos COPI, COPII, IFT y SEA que corresponden a las subunidades COPβ', Sec31, IFT80 y Seh1 (códigos PDB 2ynp, 2pm9, 5n4a y 3ewe). Cada dominio fue representado en *cartoon* con el color gris y los *blades* son renumerados considerando el centroide del grupo. Los residuos conservados fueron representados en *stick*, usando los esquemas de colores amarillo y rosado fuerte para destacar los residuos similares e idénticos según el alineamiento múltiple de secuencias entre los miembros del grupo BI. Los recuadros con líneas segmentadas muestran la posición de la histidina conservada y los residuos cercanos con los cuales interactúa. Las figuras fueron creadas usando PyMOL.

Repitiendo el mismo análisis para los dominios del grupo BII, el alineamiento múltiple de secuencias mostró que estos dominios al igual que los dominios del grupo BI poseen varias regiones con residuos principalmente hidrofóbicos como Leu, Ile, Val, Ala, Gly y Phe, y sólo algunas columnas presentan la conservación de algunos residuos cargados como Ser, Thr o Glu (Figura 20). Estos residuos conservados se encuentran en casi todos los *blades* con excepción el sexto *blade*, encontrándose presentes en las hebras β internas de cada *blade* (Figura 21). El mapeo de estos residuos señala que los residuos hidrofóbicos interactúan entre ellos entre sus cadenas laterales entre las *blades* continuas de forma similar a lo visto en los dominios del tipo BI (Figura 21). Estos residuos posiblemente cumplen una función estructural, manteniendo la forma compacta y rígida que poseen los dominios β-propeller. Los residuos cargados se encontraron en las hebras β6, β18, β19 y β23 del dominio β-propeller de Nup170 (Figura 20).

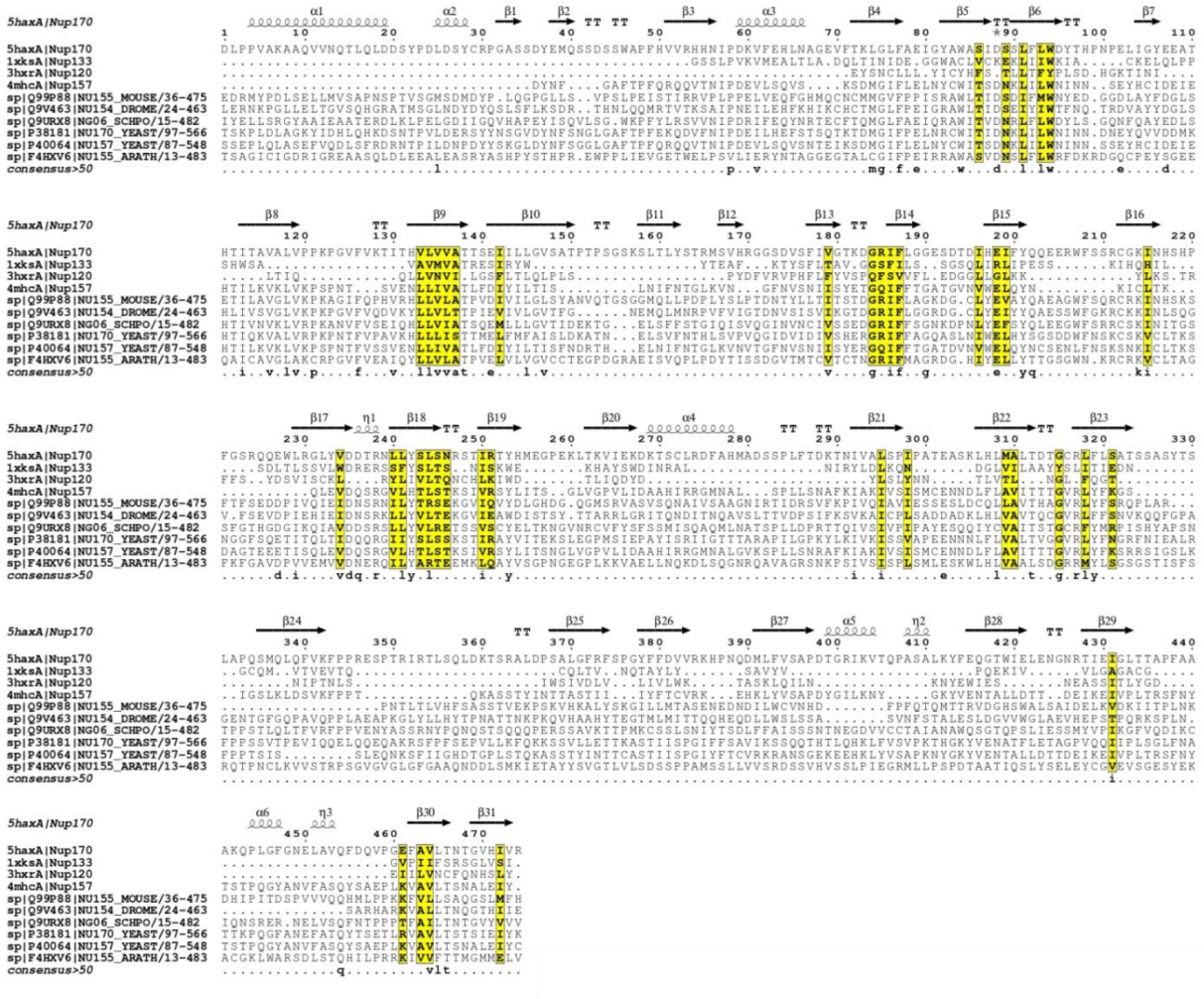


Figura 20. Alineamiento múltiple de los dominios del tipo BII según sus comparaciones estructurales.

El alineamiento múltiple fue generado de las comparaciones estructurales entre los dominios BII con respecto al dominio β-propeller de Nup170, alineando a la vez, las secuencias de sus proteínas homólogas. Los residuos conservados son destacados en amarillo mientras que los elementos de estructura secundaria del dominio β-propeller de Nup170 son indicados en la parte superior del alineamiento múltiple usando flechas para señalar las hebras y resortes para las hélices. Esta figura fue generada con el servidor de ESPript 3.0 usando la matriz de puntaje de Risler y un umbral de corte definido por defecto (Global score = 0.7) para indicar los residuos conservados.

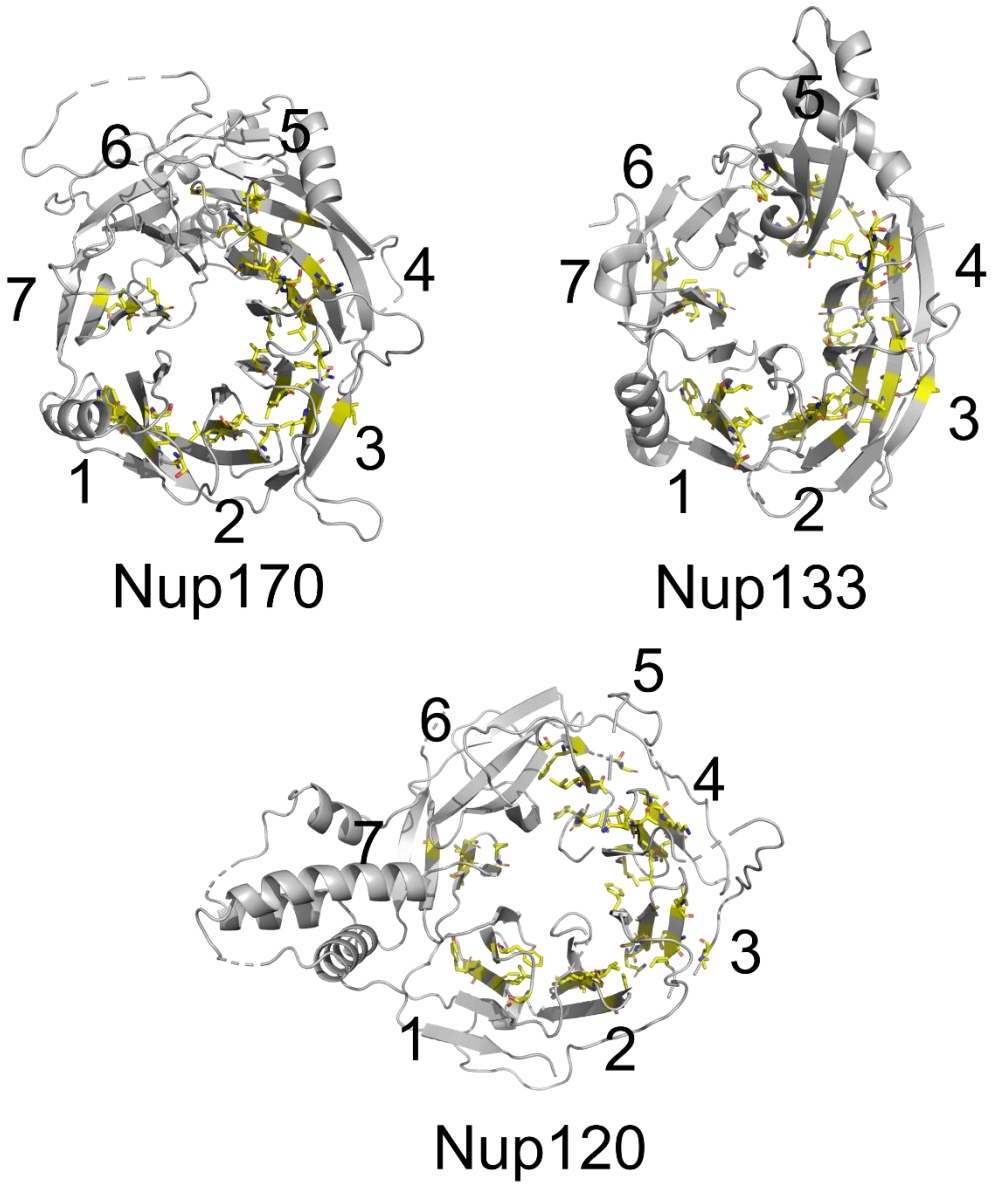


Figura 21. Residuos estructuralmente conservados entre los miembros del grupo BII. La figura muestra los residuos conservados presentes en los dominios Nup170, Nup133 y Nup120 (códigos PDB 5hax, 1xks y 3hxr, respectivamente) según el alineamiento múltiple de los miembros del grupo BII donde cada dominio fue representado en *cartoon* renumerando los *blades* que los componen. Los residuos conservados fueron representados en *stick* usando el esquema de color amarillo y en color gris se destacaron aquellos residuos que no están conservados. Esta figura fue creada usando el programa PyMOL.

Entre los dominios del tipo BIII, los residuos conservados se encuentran presentes en todos los *blades* donde la mayoría están entre los *blades* primero y cuarto (Figura 22). Estos motivos en su mayoría están compuestos por aminoácidos hidrofóbicos que interactúan entre sí entre los *blades* primero y séptimo, segundo y tercero, y entre el tercero y cuarto, respectivamente (Figura 23). También, se encontraron algunas columnas de residuos cargados en el alineamiento múltiple de secuencias incluyendo principalmente Gln y Arg (Figura 22).

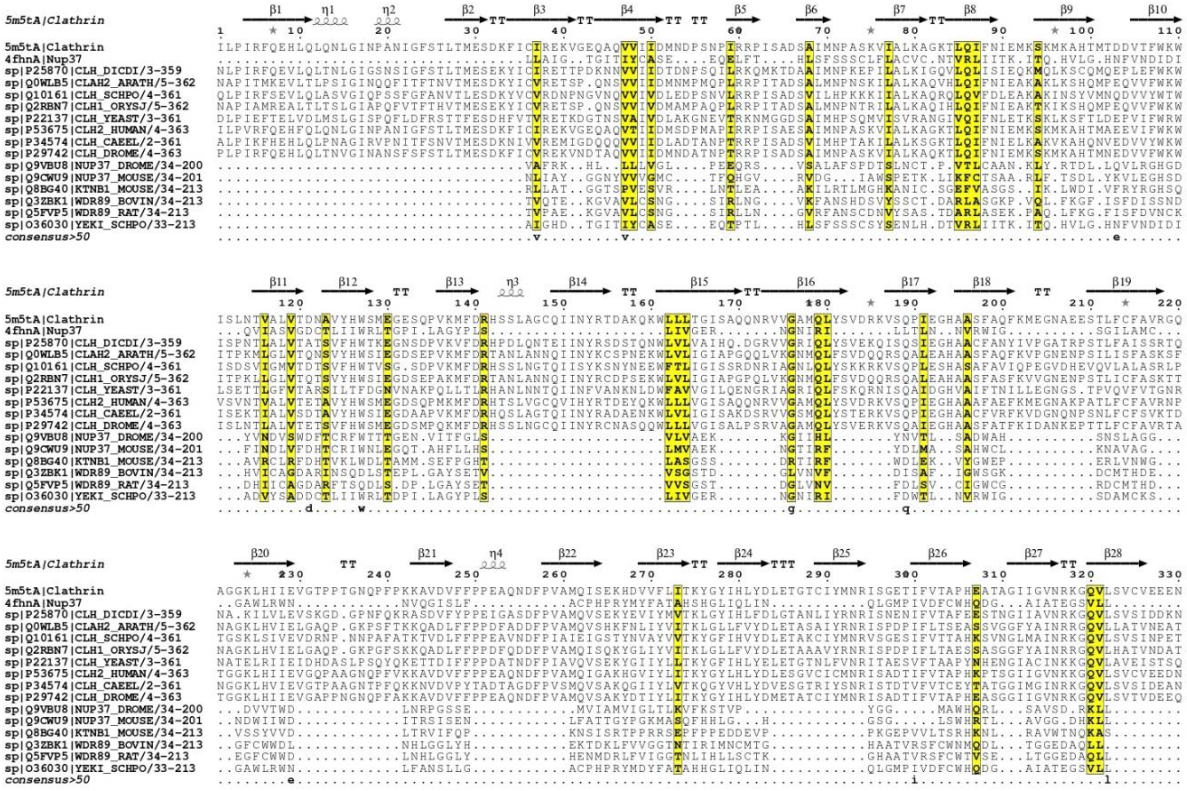


Figura 22. Alineamiento múltiple de los dominios del tipo BIII según sus comparaciones estructurales.

El alineamiento múltiple fue generado de la comparación estructural entre los dominios β-propeller de Nup37 y Clatrina, alineando a la vez, las secuencias de sus proteínas homólogas. Los residuos conservados son destacados en amarillo mientras que los elementos de estructura secundaria de Clatrina son indicados en la parte superior del alineamiento múltiple usando flechas para señalar las hebras y resortes para las hélices. Esta figura fue generada con el servidor de ESPript 3.0 usando la matriz de puntaje de Risler y un umbral de corte definido por defecto (Global score = 0.7) para indicar los residuos conservados.

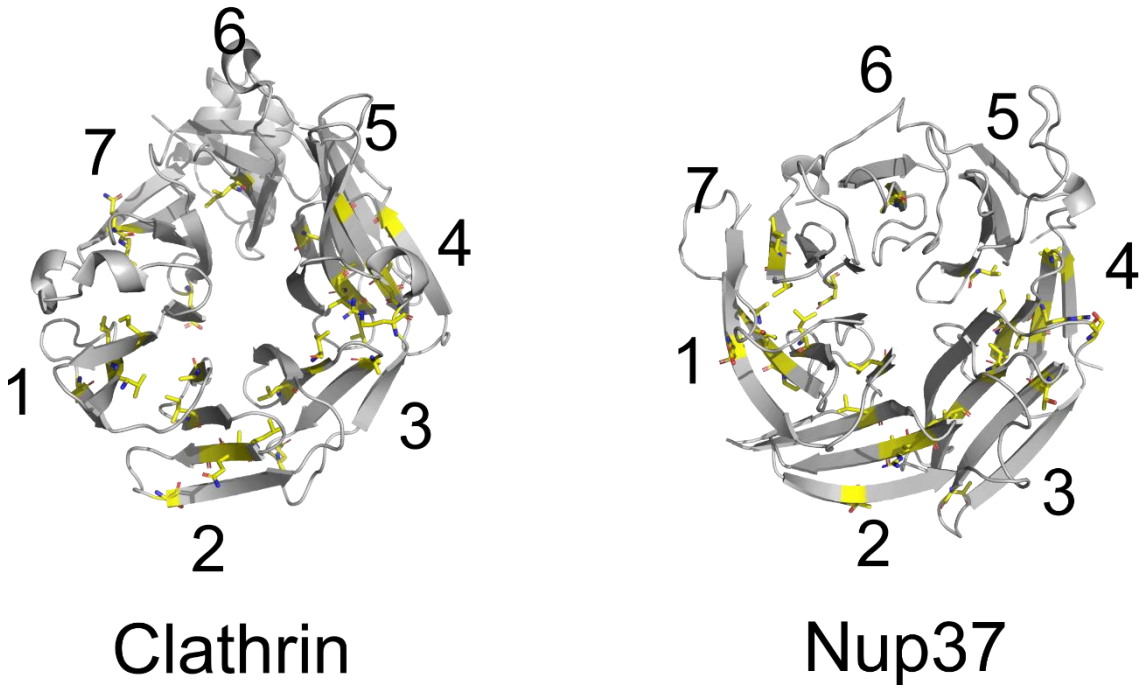


Figura 23. Residuos estructuralmente conservados entre los miembros del grupo BIII. La figura muestra los residuos conservados presentes entre los dominios Clatrina y Nup37 (códigos PDB 5m5t y 4fhn, respectivamente) según el alineamiento múltiple de los miembros del grupo BIII, donde cada dominio fue representado en *cartoon* renumerando los *blades* que los componen. Los residuos conservados fueron representados en *stick* usando el esquema de color amarillo y en color gris se destacaron aquellos residuos que no están conservados. Esta figura fue creada usando el programa PyMOL.

Por consiguiente, un análisis más detallado de la conservación de secuencia presente en los dominios β-propeller de las proteínas de cubierta de membrana fue derivado a partir de sus superposiciones estructurales (Figura 24). Este alineamiento múltiple de secuencias muestra la conservación local de ocho motivos que se pueden encontrar solos o en grupos de tres residuos continuos en la cadena. Estos motivos son en su mayoría hidrofóbicos cuyos residuos se mapean en los *blades* segundo, tercero, quinto y sexto, con excepción de una sola columna que presenta principalmente aminoácidos cargados (entre ellos se encuentran mayormente Arg seguidos por Ser) que se mapearon en la hebra β16 del cuarto *blade* en el primer subdominio de COPβ' (Figura 24). Interesantemente, el alineamiento múltiple de las estructuras muestra una cercanía de las cadenas laterales entre los residuos hidrófobos conservados presentes entre el segundo y tercer *blade* del primer dominio de COPβ' que también están presentes en otros dominios β-propeller entre las proteínas de cubierta de membrana (Figura 25). Estos motivos de secuencia se encuentran presentes solamente en las hebras centrales de los *blades* que componen los dominios β-propeller, posiblemente estos residuos son necesarios para mantener sus estructuras compactas y globulares, lo que podría constituir un vestigio remanente de su dominio ancestral común.

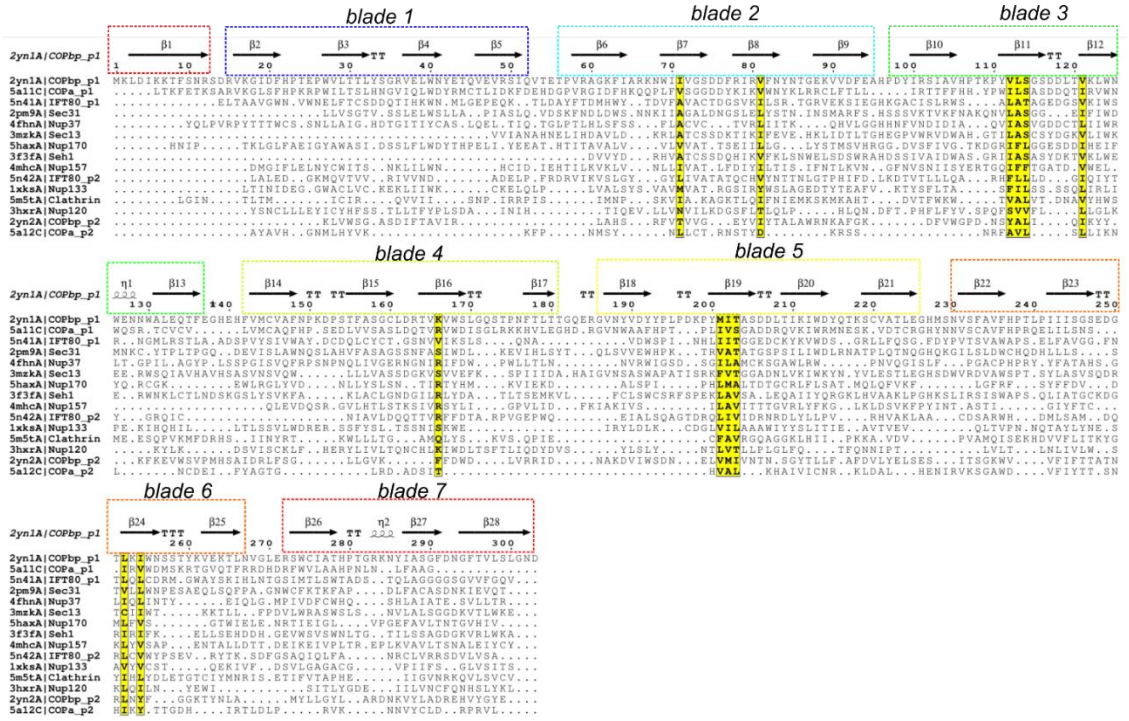


Figura 24. Vestigios del dominio β-propeller ancestral conservados en las proteínas MC. Alineamiento múltiple de los dominios β-propeller de las proteínas de cubierta de membrana de acuerdo con las superposiciones pareadas calculadas con MOMA2 con respecto al centroide (primer β-propeller de COPβ'). En la parte superior del alineamiento múltiple se destacan los elementos de estructura secundaria presentes en el centroide donde las hebras β son representadas con flechas y las hélices con resortes. Además, se indican mediante rectángulos con líneas segmentadas, los elementos de estructura secundaria que forman parte de los *blades.* Los residuos conservados son destacados en amarillo en el alineamiento múltiple. Esta figura fue generada en el servidor de ESPript 3.0 usando la matriz de puntaje de Risler y un umbral de corte definido por defecto.

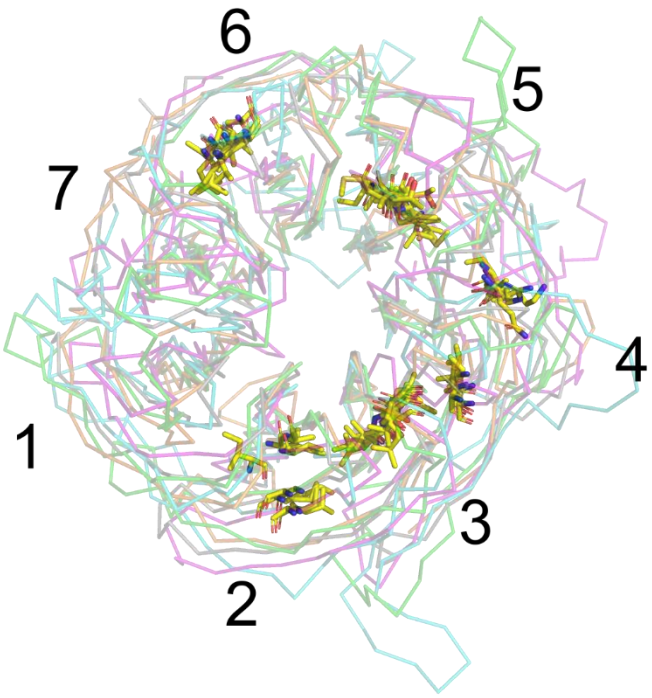


Figura 25. Conservación estructural de los dominios β-propeller en las proteínas MC. La figura muestra la superposición de los dominios β-propeller del tipo BI y BII según las comparaciones realizadas con MOMA2. Las proteínas están representadas en su esqueleto de la cadena principal. Las estructuras de los dominios de Sec31, Nup170, Clatrina, IFT80 y COPβ' (códigos PDB 2pm9, 5hax, 5m5t, 5n4a y 2ynp) fueron destacados con los colores verde, celeste, rosado, naranjo y gris, respectivamente. Los motivos de secuencia de estas proteínas según el alineamiento múltiple de la Figura 21 fueron destacados en la superposición óptima. Estos motivos conservados fueron representados en *stick* usando un esquema de color amarillo.

3.3.3.4. Exploración de las relaciones estructurales entre las proteínas MC

A fin de explorar las relaciones estructurales presentes entre las proteínas de cubierta de membrana, hemos representado gráficamente sus similitudes por medio de gráficos circulares conocidos como *Circos plots*, considerando simultáneamente los *matches* entre sus dominios β-propeller y SPAH (Figuras 26, 31, 32, 34 y 36)(Krzywinski *et al.*, 2009). Estos gráficos destacan todas las comparaciones que reportan un porcentaje de similitud relativa mayor al 20%, señalando además si estas conexiones han sido descritas anteriormente en el estado del arte a través de superposiciones rígidas o si pueden ser detectadas por medio de comparaciones de perfiles de HMMs (mediante el uso del programa HHsearch) (Brohawn *et al.*, 2008a; Sampathkumar *et al.*, 2013; Stuwe *et al.*, 2014; Söding *et al.*, 2005). Con respecto a sus comparaciones estructurales, hemos considerado el largo del alineamiento estructural (eq) y el B_{score} para indicar si las relaciones encontradas son débiles o fuertes, señalando también si estas conexiones son nuevas con respecto a lo que ya se conoce.

Mediante los gráficos de *Circos*, hemos estudiado las relaciones descritas dentro de cada tipo de subunidad (por ejemplo, entre los dominios de las subunidades del grupo Cage-CV) o entre los distintos tipos de subunidades de acuerdo con su clasificación funcional (entre los dominios de las subunidades Cage-NPC y Adaptor-NPC). A continuación, exploraremos las relaciones intra e intergrupales encontradas entre los dominios MC, señalando como MOMA2 nos ha permitido confirmar, extender y descubrir nuevas relaciones entre las proteínas MC.

3.3.3.4.1. Relaciones intragrupales entre las proteínas MC

Entre las relaciones intragrupales podemos destacar aquellas entre los dominios de las subunidades del tipo "Cage" que se encuentran presentes en los complejos de cubierta de vesículas y en el poro nuclear. En los *Circos plots* podemos apreciar que tanto HHsearch como MOMA2 detectan mayor parte de las conexiones presentes entre las subunidades de los complejos COPI, COPII, IFT y Clatrina tanto a nivel de sus dominios β-propeller como SPAH (Figura 26A). Las comparaciones calculadas con MOMA2 muestran que existen fuerte similitudes estructurales entre los dominios β-propeller cuyas subunidades pertenecen al grupo Cage-CV, por ejemplo, se observa las fuertes conexiones presentes entre el dominio β-propeller de IFT80 con los dominios de COPα y COPβ', o entre el dominio β-propeller de Sec31 con los dominios de Clatrina, COPα, COPβ', y IFT80. Mientras que a nivel de los dominios SPAH, tanto MOMA2 como HHsearch detectaron las relaciones presentes entre Sec16 y Sec31, y entre COPα y COPβ' que han sido anteriormente descritas en la literatura (Lee and Goldberg, 2010; Schlacht and Dacks, 2015). Por otro lado, MOMA2 pudo detectar algunas similitudes débiles a nivel de los dominios SPAH de la que podemos destacar la conexión encontrada entre COPα y Sec31.

Dentro de las subunidades del tipo "Cage-NPC", hemos confirmado relaciones señaladas anteriormente entre algunas nucleoporinas, destacando la relaciones evolutivas descritas para Nup192 y Nup188, o entre las Nups del tipo COPII-like (Stuwe *et al.*, 2014; Brohawn *et al.*, 2008a; Brohawn and Schwartz, 2009), donde en algunos casos nosotros las pudimos confirmar a través de las

comparaciones de los perfiles de HMM (Figura 26B). Es importante señalar que, por medio de MOMA2, hemos podido extender y descubrir nuevas relaciones estructurales entre algunas subunidades del poro nuclear que no han sido descritas antes. Por ejemplo, las similitudes estructurales entre los dominios β-propeller presentes en las subunidades que conforman el poro nuclear. Mediante HHsearch, hemos detectado solamente la relación que posee Seh1 y Nup37, que de acuerdo con nuestra clasificación de estos dominios estos pertenecen a los grupos BI y BIII. Mediante MOMA2, hemos podido extender dicha relación a los dominios β-propeller de tipo BII señalando que estos dominios analizados probablemente divergieron de un dominio ancestral del tipo BI (Figuras 26B y 27).

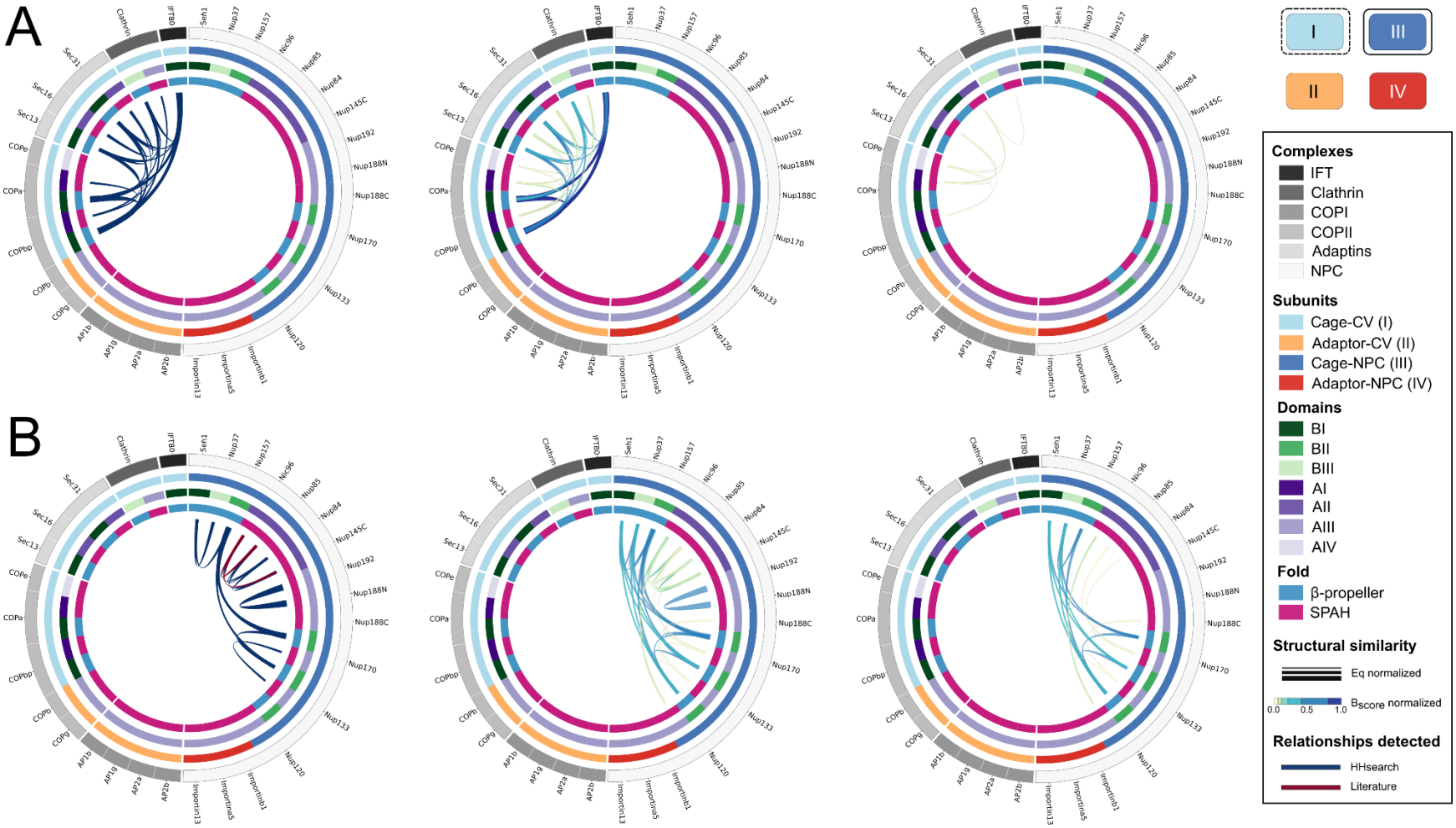


Figura 26. Relaciones intragrupales entre las subunidades del tipo "Cage".
Esta figura muestra las relaciones intragrupales entre las subunidades del tipo “Cage-CV” (A) y “Cage-NPC” (B). Cada parte consta de tres gráficos circulares donde el gráfico de la izquierda da a conocer las relaciones que se han descrito en la literatura o que pueden ser inferidas de las comparaciones realizadas con HHsearch (e-value < 1e-5). El gráfico del medio da a conocer las conexiones encontradas

con MOMA2 según el largo del alineamiento (Eq) y la similitud estructural obtenida de las comparaciones de las matrices SSE (B_{score}). Finalmente, el gráfico de la derecha da a conocer las relaciones nuevas encontradas con MOMA2. Las conexiones reportadas por MOMA2 son representadas por el grosor y el color del enlace, donde una fuerte conexión estructural es representada con una línea gruesa de color azul y una débil conexión es indicada con una línea delgada de color amarillo. Las subunidades son clasificadas según al complejo que pertenecen (Complexes), la clasificación funcional de sus subunidades (Subunits), la clasificación estructural de sus dominios (Domains) y la forma de su pliegue (Fold).

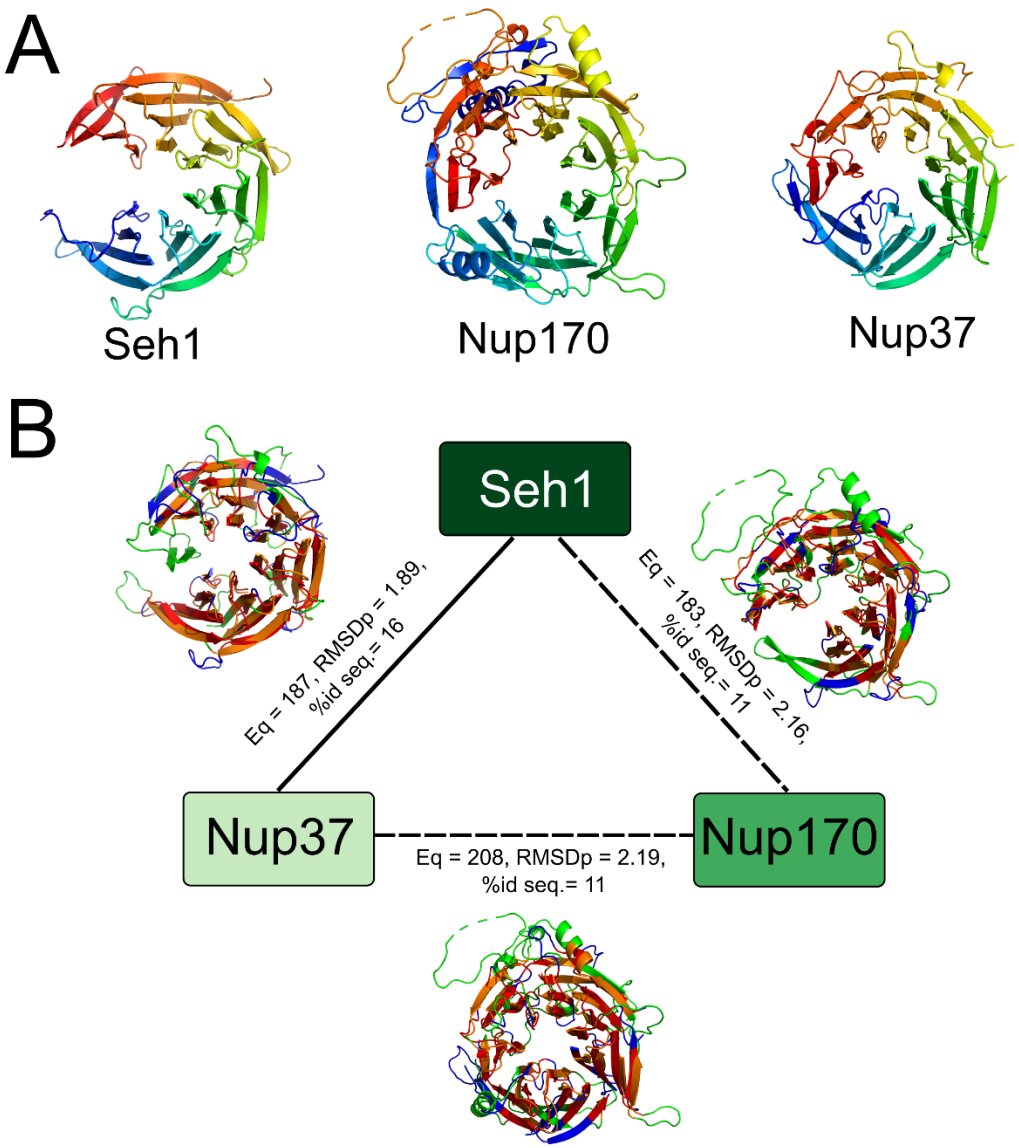


Figura 27. Extensión de las relaciones encontradas entre los dominios β-propeller del poro nuclear.

Extensión de las conexiones presentes entre los dominios β-propeller de Seh1, Nup170 y Nup37 (códigos PDB 3f3f, 5hax y 4fhn) cuyas estructuras están representadas en *cartoon* y coloreadas en *rainbow* (A). En el diagrama central se señala las relaciones encontradas entre estas subunidades pertenecen a los grupos BI, BII y BIII (B). En el diagrama se destaca la conexión presente entre Seh1 y Nup37 es detectada con HHsearch. En cambio, las conexiones entre Seh1 y Nup170, y entre Nup170 y Nup37 donde fueron detectadas con MOMA2. Las superposiciones estructurales muestran los pares de residuos equivalentes entre ambas estructuras en rojo y anaranjado, cuyos residuos no alineados son mostrados en azul y verde en las proteínas *query* y *target*, respectivamente. Las figuras de las estructuras y las superposiciones fueron generadas con PyMOL.

Por otro lado, también nos ha permitido derivar relaciones nuevas entre subunidades cuyo dominio SPAH adopta un pliegue único a diferencia de otras subunidades que conforman el *scaffold* del poro nuclear, como es el caso de la nucleoporina Nup120. Típicamente, las nucleoporinas Nic96, Nup145C, Nup85, y Nup84 poseen una terminación N-terminal que es seguida de un dominio helicoidal tripartito compartido con las subunidades Sec31 y Sec16 del complejo COPII conocido como *ancestral coatomer element 1* (ACE1) (Brohawn *et al.*, 2008a). Los dominios ACE1 poseen tres elementos denominados corona, tronco y cola que adoptan un patrón J-like, zigzagueando hacia un lado del tronco, haciendo un giro en U dentro de la corona y cayendo luego en el lado opuesto del tronco terminando en la cola. Hasta ahora estos elementos no habían podido ser superpuestos correctamente con las herramientas actuales, pero gracias al uso de MOMA2, hemos podido superponer estos elementos compartidos corroborando lo descrito por Brohawn y colaboradores en 2008 (Figura 28). Mediante las comparaciones realizadas podemos destacar que el tronco es el módulo estructural más conservado entre los dominios AII (según nuestra clasificación), alineándose parte de éste en todas las comparaciones realizadas con respecto a la proteína Nic96. En cambio, podemos apreciar que solamente en algunas superposiciones se alinea parte de la cola. Por ejemplo, en las comparaciones calculadas entre Nic96 vs. Nup85, o entre Nic96 vs. Sec31. Por el contrario, en la mayoría de estas comparaciones estructurales solamente se alineó un pequeño sub-fragmento de la corona, con excepción de la superposición reportada entre Nic96 vs. Sec31.

En cambio, en el caso de Nup120, su estructura cristalográfica revela que esta proteína adopta una topología significativamente diferente a las proteínas ACE1, donde un pequeño apéndice de 4-hélices de su dominio helicoidal se inserta en el dominio β-propeller, mientras que la parte central de su dominio SPAH es tanto ancha como larga siendo diferente a los dominios elongados de las proteínas ACE1. Sin embargo, usando MOMA2 podemos apreciar que Nup120 guarda una similitud estructural con Nic96 a pesar de que en la literatura se sugiere Nup120 carece de similitud estructural con otras nucleoporinas (Leksa *et al.*, 2009). De hecho, en la superposición reportada se alinearon los tres elementos del dominio ACE1 de Nic96 con el extremo C-terminal de Nup120 (Eq = 120, $RMSD_{pond}$ = 2.07 y % id seq. = 7%), superponiendo tres pares de sub-fragmentos de 3, 6 y 4 pares de elementos de estructura secundaria de largo (Figura 29). Mientras que su extremo N-terminal fue la única parte que no se superpuso (Figura 29). Por consiguiente, el alineamiento múltiple de secuencia de los fragmentos superpuestos señala la conservación de aminoácidos en su mayoría hidrofóbicos que se encuentran en las hélices α9, α11, α13, y α19 de Nup120, y cuyas cadenas laterales interactúan con las hélices de las hebras continuas (Figura 30). Estos resultados sugieren que tanto que los dominios SPAH de las nucleoporinas del tipo COPII-like y Nup120 derivan posiblemente de un pliegue ancestral donde el dominio helicoidal de Nup120 se especializó adoptando una topología única.

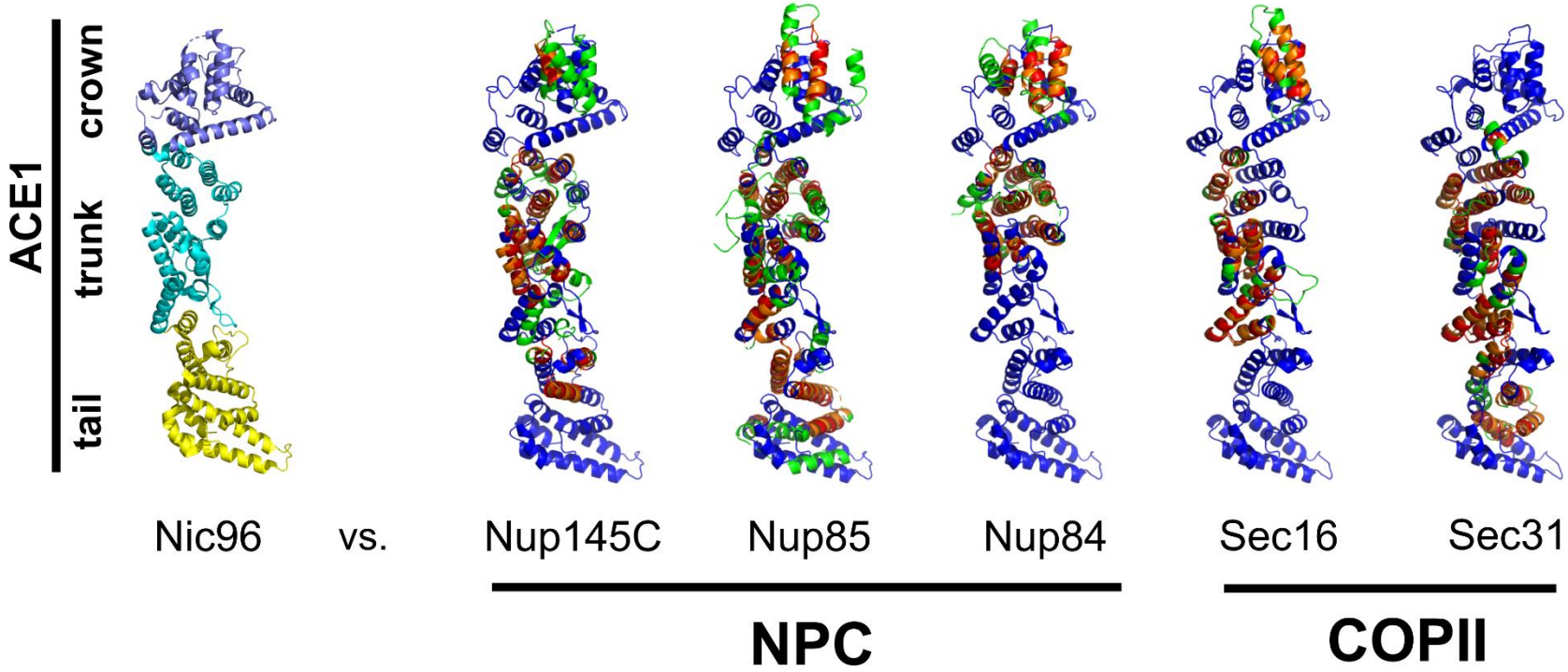


Figura 28. Superposición estructural de los módulos ACE1 en los dominios del tipo AII.
Considerando la descripción realizada por Brohawn *et al.* 2008 de los módulos ACE1, se destacó los módulos corona, tronco y cola presentes en el dominio α-solenoide de Nic96 con los colores violeta, celeste y amarillo, respectivamente (código PDB 2qx5). Las estructuras que fueron alineadas con Nic96 corresponden a los dominios SPAH de las proteínas Nup85, Sec31, Nup84, Nup145C y Sec16 (códigos PDB 4xmm, 3mzl, 3iko, 4xmn y 3mzk, respectivamente). Las superposiciones estructurales muestran los residuos alineados de las estructuras *query* y *target* con los colores rojo y naranjo, respectivamente. Los residuos no alineados son representados en azul (*query*) y verde (*targets*).

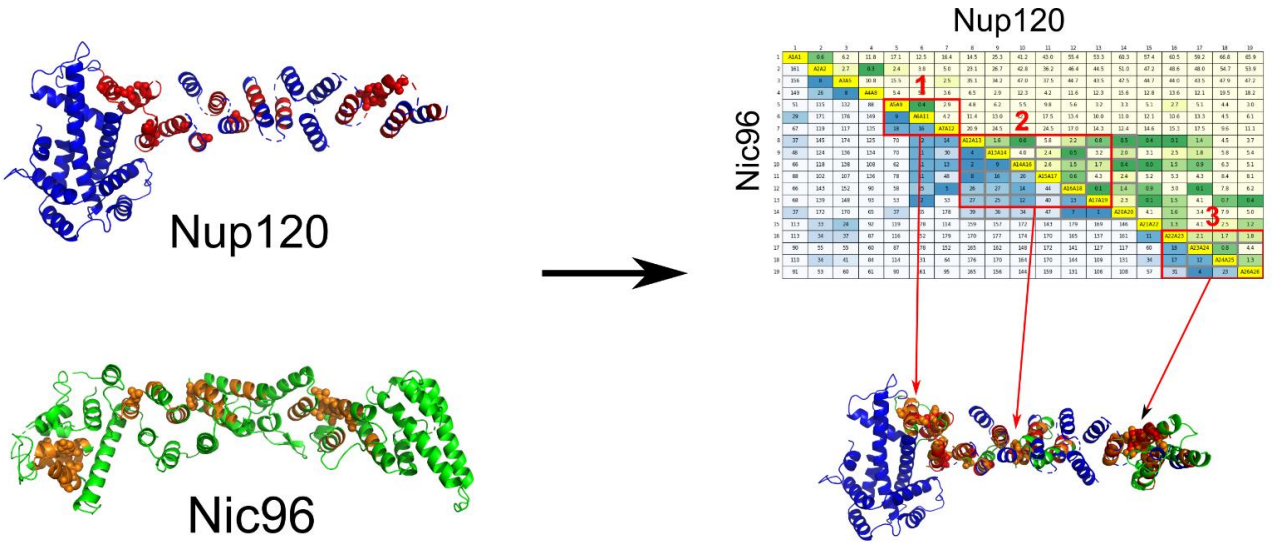
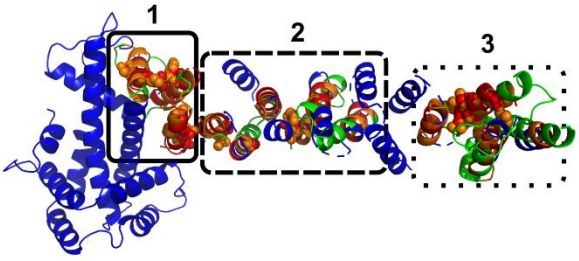


Figura 29. Superposición estructural de los dominios SPAH de Nup120 y Nic96.
Superposición estructural calculada con MOMA2 entre Nup120 (código PDB 4xmn, E) y Nic96 (código PDB 2qx5, A). Los residuos estructuralmente equivalentes en ambas estructuras son señalados en rojo y naranjo para las estructuras *query* y *target*, mientras que los residuos no alineados son descritos en azul y verde, respectivamente. La matriz de diferencia reporta cuales fueron los pares de sub-fragmentos seleccionados (rectángulos en color rojo), cuya superposición estructural se muestra debajo de la matriz. Finalmente, los residuos conservados entre ambas estructuras son representados en *sphere*.


1
2
3

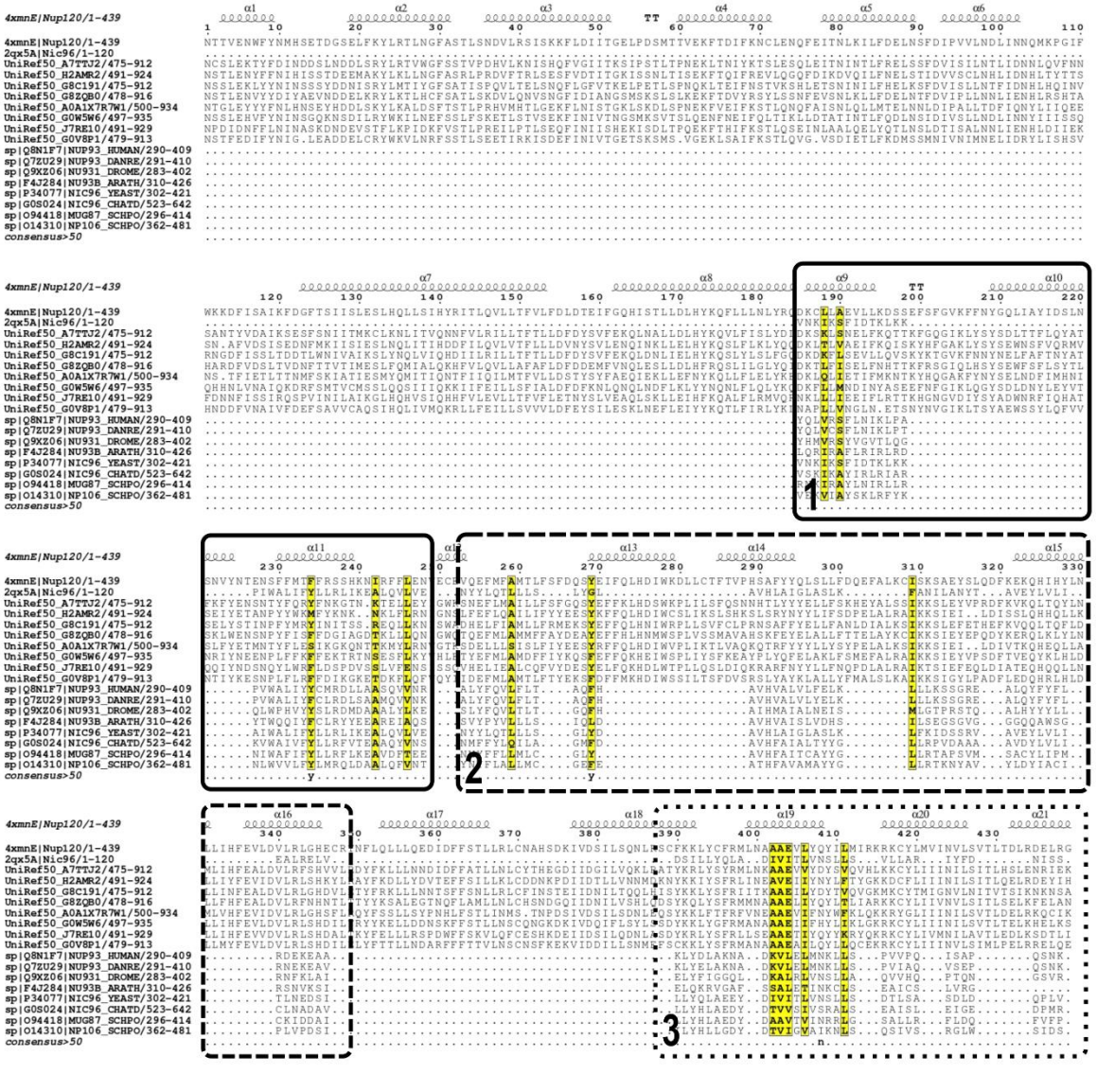

4xmnE|Nup120/1-439
4xmnE|Nup120/1-439
2qx5A|Nic96/1-120
UniRef50_A7TTJ2/475-912
UniRef50_H2AMR2/491-924
UniRef50_G8C191/475-912
UniRef50_G8ZQB0/478-916
UniRef50_A0A1X7R7W1/500-934
UniRef50_G0W5W6/497-935
UniRef50_J7RE10/491-929
UniRef50_G0V8P1/479-913
sp|Q8N1F7|NUP93_HUMAN/290-409
sp|Q7ZU29|NUP93_DANRE/291-410
sp|Q9XZ06|NU931_DROME/283-402
sp|F4J284|NU93B_ARATH/310-426
sp|P34077|NIC96_YEAST/302-421
sp|G0S024|NIC96_CHATD/523-642
sp|O94418|MUG87_SCHPO/296-414
sp|O14310|NP106_SCHPO/362-481
consensus>50
1
2
3

Figura 30. Vestigios del pliegue ACE1 en Nup120.
En la parte superior de la imagen se muestra la superposición estructural de Nup120 y Nic96 destacando las tres regiones superpuestas donde se alinearon los elementos corona, tronco y cola de Nic96 (rectángulos con línea continua, segmentada y punteada, respectivamente), señalando los residuos estructuralmente equivalentes con los colores rojo y naranjo, y aquellos conservados usando la representación *sphere*. En la parte inferior se muestra el alineamiento múltiple de secuencias generado a partir de la comparación estructural de los dominios Nup120 y Nic96, alineando a la vez, las secuencias de sus proteínas homólogas. Los residuos conservados son destacados en amarillo denotando con rectángulos las regiones superpuestas. Esta figura fue generada con el servidor de ESPript 3.0 usando la matriz de puntaje de Risler y un umbral de corte definido por defecto (Global score = 0.75) para indicar los residuos conservados.

Por otra parte, los gráficos de *Circos* muestran que MOMA2 confirma la mayoría de las conexiones intragrupales entre las subunidades del tipo "Adaptor" (Figura 31). Por ejemplo, tanto MOMA2 como HHsearch destacan las relaciones evolutivas reportadas anteriormente para las adaptinas AP1 y AP2 con las subunidades adaptadoras del complejo COPI (COPβ y COPγ), de los cuales se han mencionado que los adaptadores tetraméricos de Clatrina comparten una homología distante de secuencia y similitud estructural con las subunidades del sub-complejo de COPI (Faini *et al.*, 2013). En cambio, el caso de las importinas podemos destacar que sólo MOMA2 pudo reportar una fuerte similitud estructural entre las carioferinas Importina 13 y Importina a5 a diferencia de HHsearch (Eq = 233, Similitud relativa = 36%, $RMSD_{pond}$ = 2.1 Å y % id seq. = 13%), donde Importina 13 es un receptor de transporte relacionado a las carioferinas β mientras que la Importina a5 pertenece a la familia de las carioferinas α (Mingot *et al.*, 2001, 13; OKA and YONEDA, 2018).

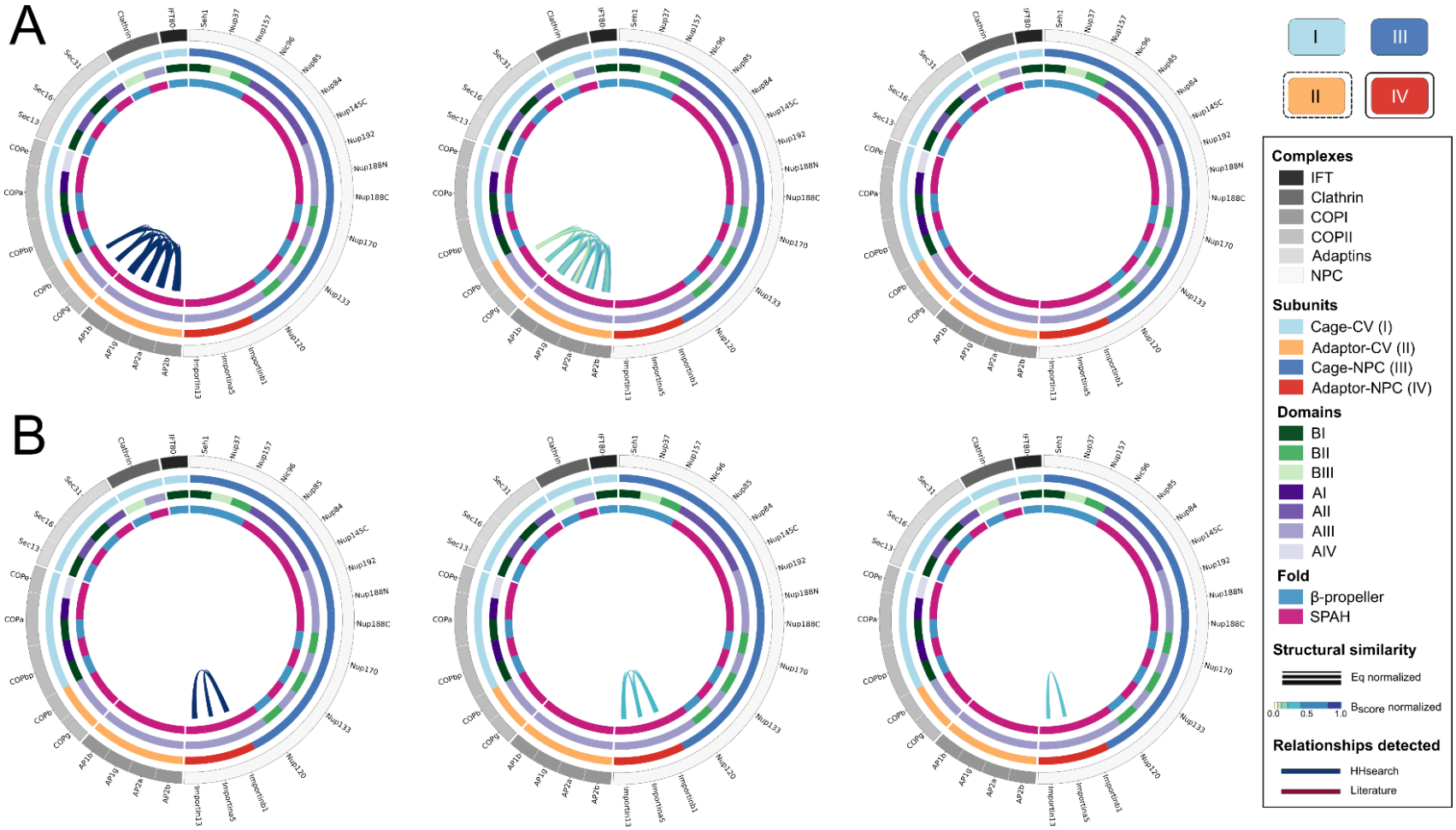


Figura 31. Relaciones intragrupales entre las subunidades del tipo "Adaptor".
Esta figura muestra las relaciones intragrupales entre las subunidades del tipo “Adaptor-CV” (A) y “Adaptor-NPC” (B). Cada parte consta de tres gráficos circulares donde el gráfico de la izquierda da a conocer las relaciones que se han descrito en la literatura o que pueden ser inferidas de las comparaciones realizadas con HHsearch (e-value < 1e-5). El gráfico del medio da a conocer las conexiones

encontradas con MOMA2 según el largo del alineamiento (Eq) y la similitud estructural obtenida de las comparaciones de las matrices SSE (B_{score}). Finalmente, el gráfico de la derecha da a conocer las relaciones nuevas encontradas con MOMA2. Las conexiones reportadas por MOMA2 son representadas por el grosor y el color del enlace, donde una fuerte conexión estructural es representada con una línea gruesa de color azul y una débil conexión es indicada con una línea delgada de color amarillo. Las subunidades son clasificadas según al complejo que pertenecen (Complexes), la clasificación funcional de sus subunidades (Subunits), la clasificación estructural de sus dominios (Domains) y la forma de su pliegue (Fold).

3.3.3.4.2. Relaciones intergrupales entre las proteínas MC

Entre las relaciones intergrupales descritas en los gráficos de *Circos* podemos destacar las conexiones presentes entre los tipos “Cage-CV” y “Cage-NPC”, donde algunas de ellas pueden ser detectadas a través de HHsearch y mientras que otras han sido descritas previamente comparando sus estructuras dando sustento a la hipótesis del Protocoatomer (Figura 32A). Gracias a las comparaciones calculadas con MOMA2, hemos descubierto más relaciones relevantes entre en estos dos tipos de subunidades donde podemos destacar, en primer lugar, las fuertes similitudes estructurales presentes entre los dominios β-propeller del tipo BI con los dominios del tipo BII, por ejemplo, entre Clatrina y Nup157 o entre las nucleoporinas Nup170, Nup133 y Nup120 con los dominios de Clatrina, COPα, COPβ’ y IFT80. En segundo lugar, podemos distinguir relaciones estructurales a nivel de sus dominios α-solenoide que no habían sido descritas previamente, por ejemplo, entre Nup85 y COPε, Nup120 con COPα y COPβ’, y entre Nup170 con Clatrina. Aunque los dominios de Nup170 y Clatrina comparten un bajo porcentaje de identidad de secuencia (debajo de un 10%), este par posee una fuerte similitud estructural a nivel de sus dominios β-propeller reportando un porcentaje de similitud relativa cercano al 45% (Eq = 190, $RMSD_{pond}$ = 2.06 Å), mostrando una fuerte conservación de los ángulos diedros entre sus pares de elementos de estructura secundaria (Figura 33A). Mientras que, a nivel de sus dominios SPAH, el alineamiento de matrices detectó tres bloques conservados entre los dominios de Clatrina y Nup170, reportando un

total de 188 pares de residuos equivalentes con un $RMSD_{pond}$ igual a 2.36 Å y un 8% de identidad de secuencia (Figura 33B). A pesar de la forma alargada del dominio SPAH de Clatrina es diferente al dominio de Nup170, igualmente se encontraron sub-fragmentos equivalentes que reportaron mínimas diferencias angulares y de distancia (bloques de compuestos por 3, 5, 7 pares alineados de SSEs), superponiendo en total 15 pares de hélices. Estos resultados señalan que Nup170 muestra una similitud cercana a Clatrina, a diferencia de lo descrito anteriormente por Whittle and Schwartz en 2009, cuyas búsquedas realizadas con métodos de comparación rígida como VAST y DALI no retornaron alineamientos con más de seis pares de hélices de largo.

Por consiguiente, las fuertes conexiones encontradas con MOMA2 entre los dominios de las subunidades de “Cage-NPC” y “Adaptor-NPC” corroboran las similitudes estructurales descritas anteriormente por Andersen y colaboradores en 2013 para las nucleoporinas Nup192 y Nup188, y las carioferinas α y β (Figura 32B). Además, se puede apreciar una relación estructural cercana a nivel de los pares SSEs entre las nucleoporinas Nup85 y Nup170 con Importina a1 (carioferina α), lo cual, no se había registrado anteriormente.

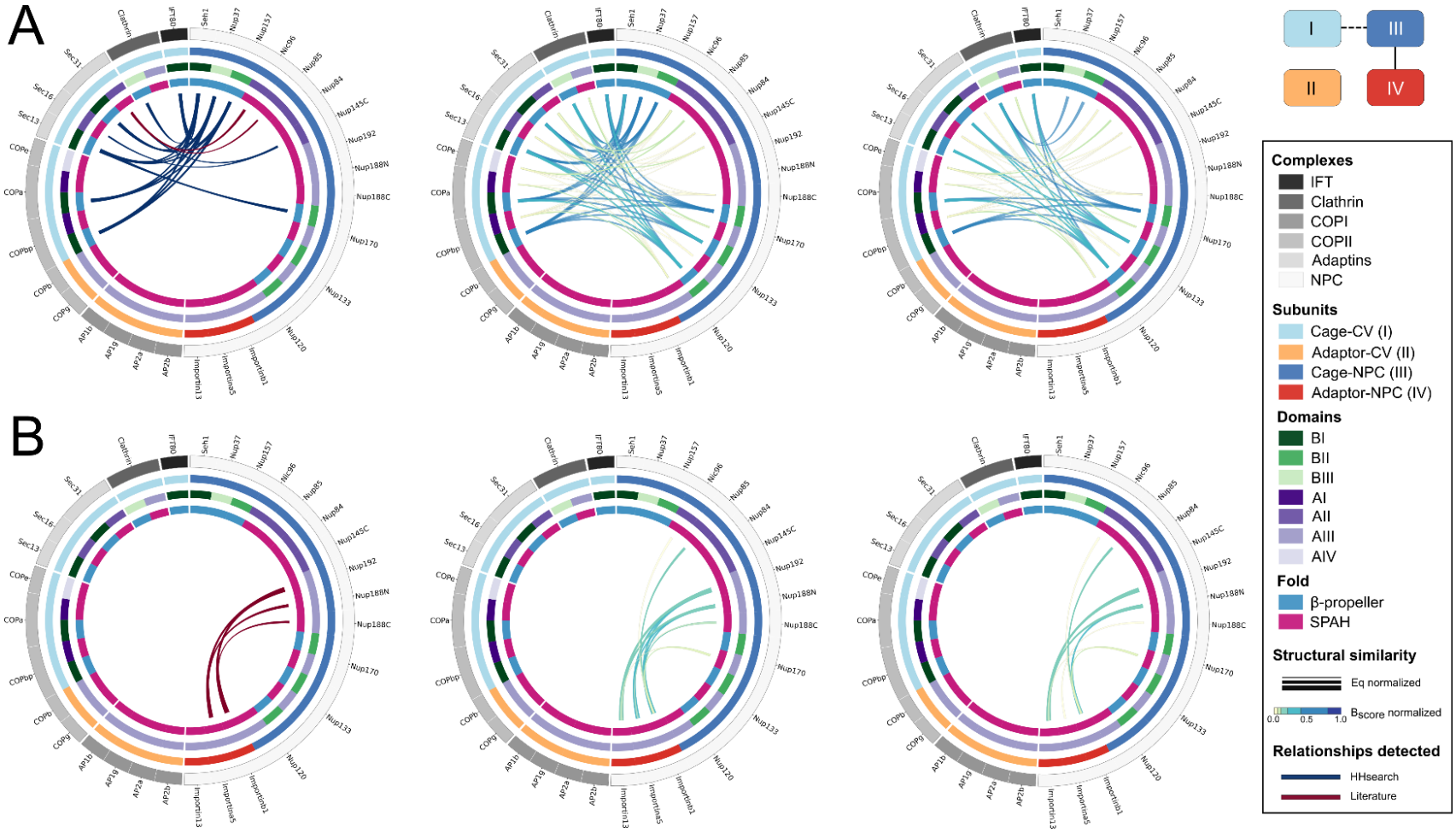


Figura 32. Extensión de las relaciones estructurales conocidas entre las proteínas MC.
Esta figura muestra las relaciones intergrupales entre las subunidades "Cage-CV" y "Cage-NPC" (A), y entre "Cage-NPC" y "Adaptor-NPC" (B). Cada parte consta de tres gráficos circulares donde el gráfico de la izquierda da a conocer las relaciones que se han descrito en la literatura o que pueden ser inferidas de las comparaciones realizadas con HHsearch (e-value < 1e-5). El gráfico del medio da a

conocer las conexiones encontradas con MOMA2 según el largo del alineamiento (Eq) y la similitud estructural obtenida de las comparaciones de las matrices SSE (B_{score}). Finalmente, el gráfico de la derecha da a conocer las relaciones nuevas encontradas con MOMA2. Las conexiones reportadas por MOMA2 son representadas por el grosor y el color del enlace, donde una fuerte conexión estructural es representada con una línea gruesa de color azul y una débil conexión es indicada con una línea delgada de color amarillo. Las subunidades son clasificadas según al complejo que pertenecen (Complexes), la clasificación funcional de sus subunidades (Subunits), la clasificación estructural de sus dominios (Domains) y la forma de su pliegue (Fold).

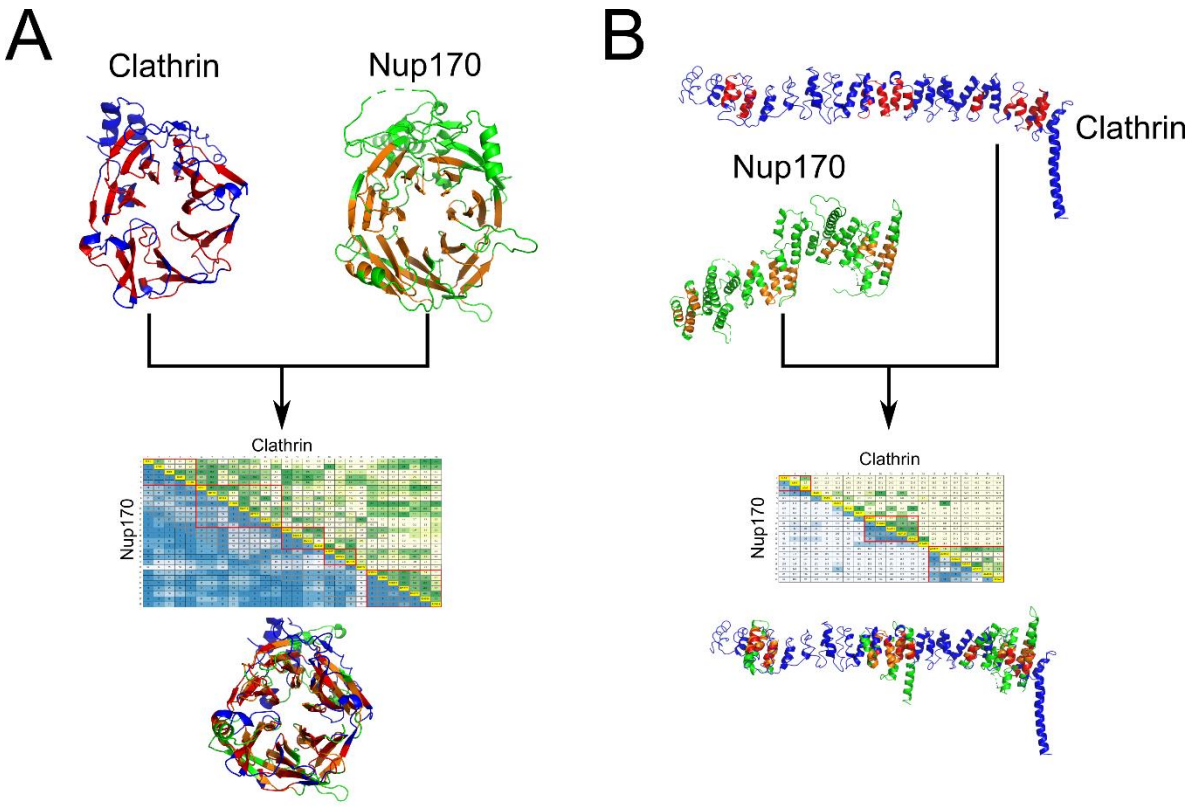


Figura 33. Similitud estructural cercana entre los dominios de Clatrina y Nup170. Superposiciones estructurales de los dominios β-propeller (A) y SPAH (B) de Clatrina (códigos PDB 5m5t y 3lvg) y Nup170 (códigos PDB 5haz y 5hax) según los alineamientos de MOMA2. Los residuos estructuralmente equivalentes son señalados en rojo y naranjo para las estructuras *query* y *target*, mientras que los residuos no alineados son descritos en azul y rojo, respectivamente. La matriz de diferencia reporta cuales fueron los pares de sub-fragmentos seleccionados (rectángulos en color rojo), cuya superposición global se muestra debajo de la matriz.

Por consiguiente, las fuertes conexiones descritas con MOMA2 entre las subunidades del tipo “Adaptor-CV” con las subunidades “Cage-NPC” y “Adaptor-NPC” nos ha permitido extender las similitudes descritas parcialmente con los métodos actuales de comparación de secuencias (HHsearch) y de estructuras (Dali, CE y Multiprot usado en las superposiciones descritas por Sampathkumar et *al.*, 2013). Estas similitudes encontradas nos han permitido conectar las nucleoporinas Nup192 y Nup188 con las adaptinas y a vez las carioferinas con las adaptinas dentro del grupo AIII (Figura 34). De hecho, podemos observar la fuerte conexión estructural presente entre estas familias de proteínas al visualizar las superposiciones flexibles calculadas entre la Importina 13 con las proteínas Nup192 y AP1g (Figura 35). Anteriormente, Stuwe y colaboradores usando DALI reportaron que la carioferina α de levadura es parecida estructuralmente al dominio ARM de Nup192, de manera similar nosotros encontramos usando MOMA2, una relación estructural cercana entre la carioferina α de humano con la proteína Nup192 de *Chaetomium thermophilum*, superponiendo principalmente el dominio ARM de Nup192 (Figura 35). Por consiguiente, solamente con MOMA2 podemos apreciar la fuerte similitud estructural que poseen estos dominios con Importina 13 (Figura 34B), donde ambas estructuras reportan alineamientos con un porcentaje de *overlap* estructural igual al 31% para Nup192 y de un 49% para AP2γ a pesar de que poseen un porcentaje de identidad de secuencia menor a un 12% (Figura 35). Es importante señalar que se han encontrado otras nucleoporinas que han mostrado una fuerte similitud estructural a nivel de sus elementos de estructura secundaria

con respecto a las adaptinas, destacando entre estos ejemplos las conexiones encontradas entre Nic96 con las subunidades de AP2 y entre Nup170 con AP1γ (Figura 34A).

Mediante los gráficos de *Circos* también se observa que existen relaciones estructurales débiles entre los dominios SPAH de las adaptinas con aquellos dominios presentes en el *cage* de los complejos COPI y COPII, destacándose solamente la relación distante entre el dominio α-solenoide de Clatrina con los dominios α-solenoide de las adaptinas (Figura 36A). En este caso en particular, la estructura de Clatrina de *Bos taurus* se superpuso con la subunidad AP2α de *Mus musculus* alineando 149 pares de residuos en total con un $RMSD_{pond}$ de 2.39 Å (reportando un 9% de identidad de secuencia). Esta superposición estructural muestra el alineamiento local de cuatro pares de sub-fragmentos donde los bloques 2 y 3 comparten una baja diferencia angular con el cuarto bloque (Figura 37). Aunque estas estructuras poseen dominios SPAH que se pliegan de forma distinta (donde Clatrina tiene un dominio alargado, en cambio AP2α se curva en la mitad), sus dominios presentan motivos conservados de residuos hidrofóbicos conservados cuyas cadenas laterales interactúan con las hélices circundantes y se encuentran presentes en las hélices α5-α10, α15, α17, α19-α23 y α27-α29 de AP2α (Figura 38).

Mientras que las relaciones reportadas entre los dominios α-solenoide de los grupos “Cage-CV” y “Adaptor-NPC” son casi inexistentes, indicando que las subunidades a las cuales pertenecen han divergido considerablemente durante la evolución de estos complejos por lo cual es difícil establecer si existe una conexión significativa entre ellos (Figura 36B).

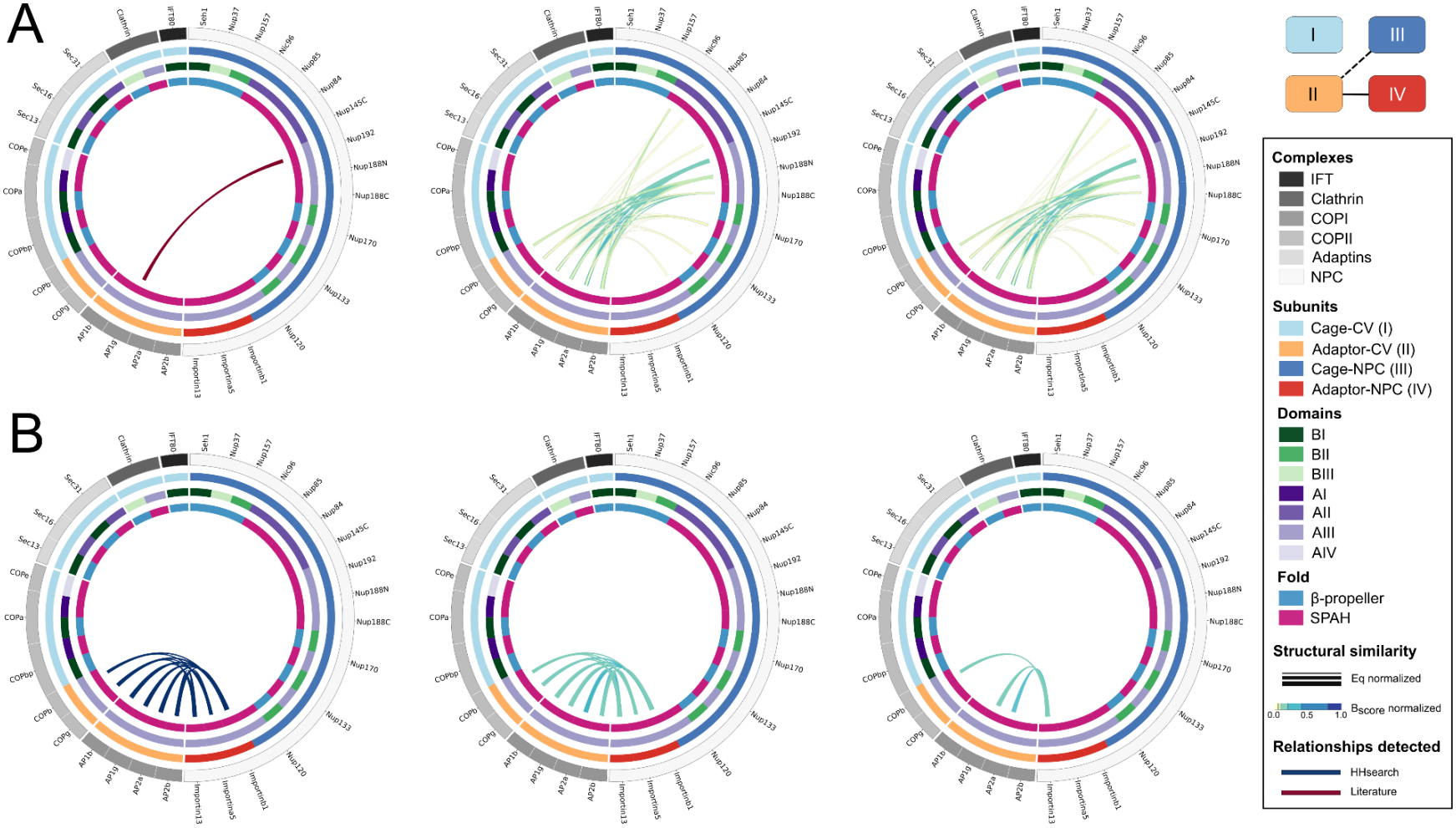


Figura 34. Relaciones intergrupales de las subunidades "Adaptor-CV" con las nucleoporinas y carioferinas.
Esta figura muestra las relaciones intergrupales entre las subunidades “Adaptor-CV” con las subunidades “Cage-NPC” (A) y “Adaptor-NPC” (B). Cada parte consta de tres gráficos circulares donde el gráfico de la izquierda da a conocer las relaciones que se han descrito en la literatura (rojo oscuro) o que pueden ser inferidas de las comparaciones realizadas con HHsearch (e-value < 1e-5, azul oscuro). El

gráfico del medio da a conocer las conexiones encontradas con MOMA2 según el largo del alineamiento (Eq) y la similitud estructural obtenida de las comparaciones de las matrices SSE (B_{score}). Finalmente, el gráfico de la derecha da a conocer las relaciones nuevas encontradas con MOMA2. Las conexiones reportadas por MOMA2 son representadas por el grosor y el color del enlace, donde una fuerte conexión estructural es representada con una línea gruesa de color azul y una débil conexión es indicada con una línea delgada de color amarillo. Las subunidades son clasificadas según al complejo que pertenecen (Complexes), la clasificación funcional de sus subunidades (Subunits), la clasificación estructural de sus dominios (Domains) y la forma de su pliegue (Fold).

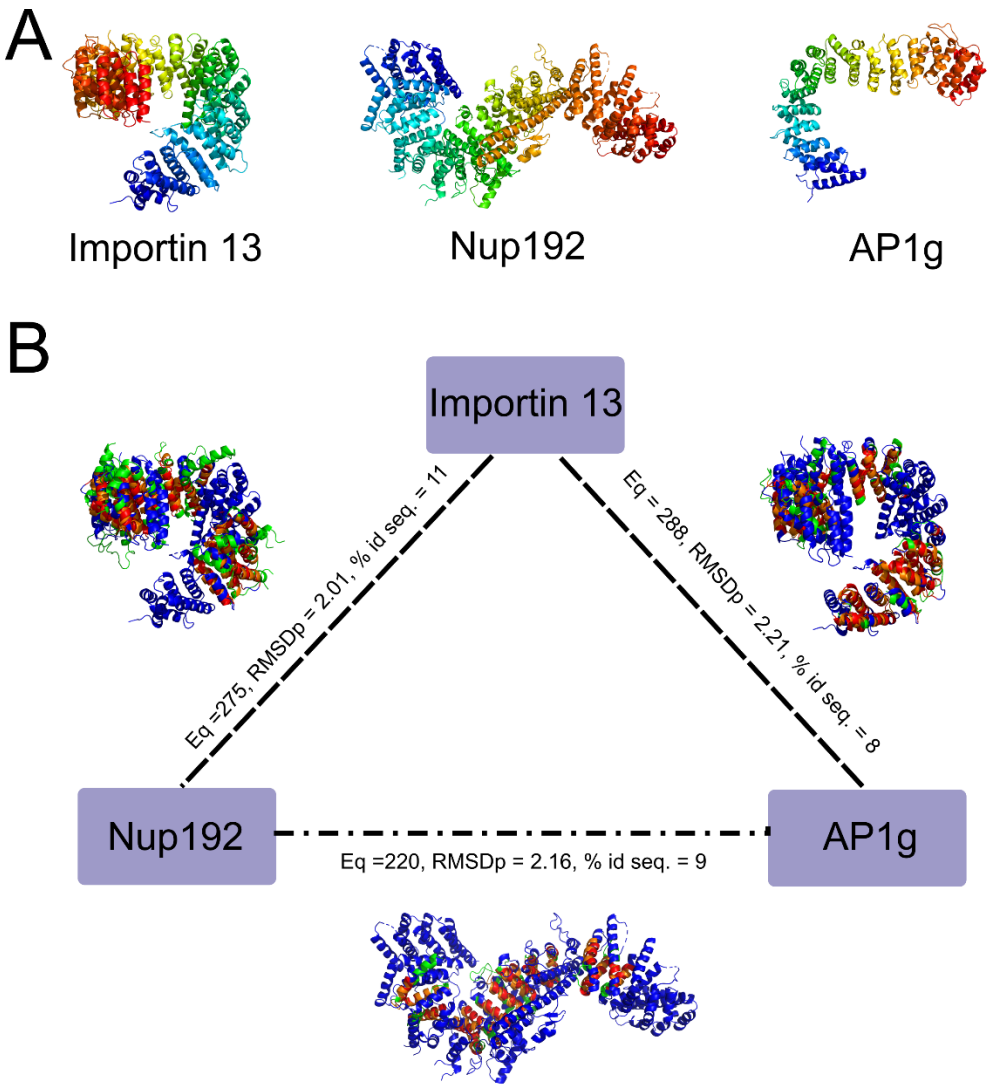


Figura 35. Ejemplo de la extensión de las relaciones presentes entre las nucleoporinas, carioferinas y adaptinas.

Comparación estructural de los dominios SPAH de Importina 13, AP1g y Nup192 (códigos PDB 2x19 cadena B, 1w63 cadena A y 5hb4 cadena B) (A) cuyas superposiciones fueron calculadas con MOMA2 (B). En la parte A se muestra las estructuras usadas en el análisis en representación *cartoon* y coloreadas en *rainbow*. Mientras que en la parte B se muestran las superposiciones obtenidas destacando si sus conexiones son detectadas solamente con MOMA2 (líneas discontinuas) o han sido descritas con anterioridad en la literatura (línea punto y trazo). El color morado de cada bloque señala que los dominios comparados pertenecen al mismo grupo (AIII). Los residuos estructuralmente equivalentes son descritos en rojo y naranjo para las proteínas *query* y *target*, cuyos residuos no alineados son representados en azul y verde, respectivamente. Las figuras de las estructuras fueron generadas con PyMOL.

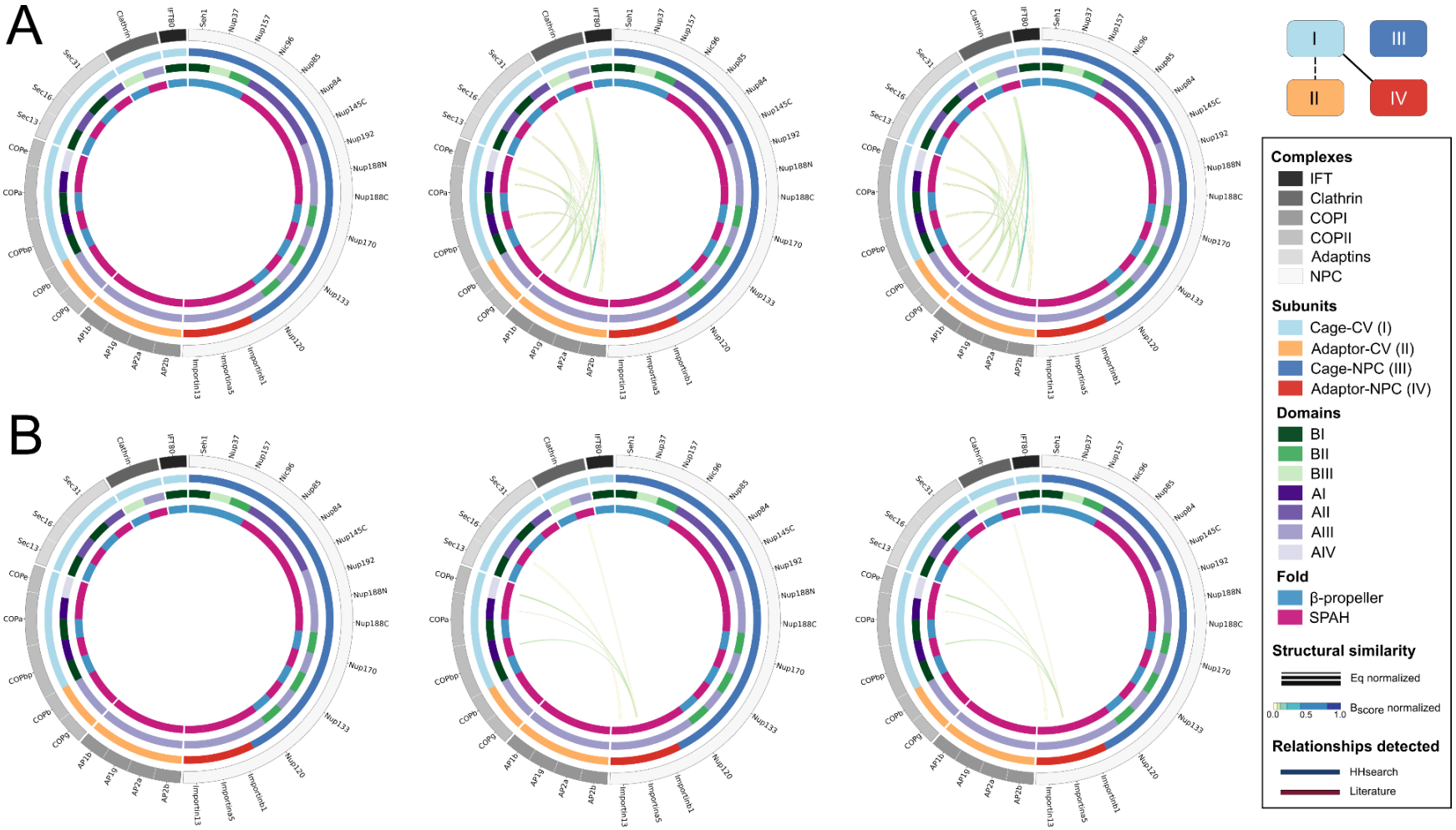


Figura 36. Relaciones intergrupales distantes entre las proteínas MC.
Esta figura muestra las relaciones intergrupales entre las subunidades "Cage-CV" con las subunidades "Adaptor-CV" (A) y "Adaptor-NPC" (B). Cada parte consta de tres gráficos circulares donde el gráfico de la izquierda da a conocer las relaciones que se han descrito en la literatura (rojo oscuro) o que pueden ser inferidas de las comparaciones realizadas con HHsearch (e-value < 1e-5, azul oscuro). El

gráfico del medio da a conocer las conexiones encontradas con MOMA2 según el largo del alineamiento (Eq) y la similitud estructural obtenida de las comparaciones de las matrices SSE (B_{score}). Finalmente, el gráfico de la derecha da a conocer las relaciones nuevas encontradas con MOMA2. Las conexiones reportadas por MOMA2 son representadas por el grosor y el color del enlace, donde una fuerte conexión estructural es representada con una línea gruesa de color azul y una débil conexión es indicada con una línea delgada de color amarillo. Las subunidades son clasificadas según al complejo que pertenecen (Complexes), la clasificación funcional de sus subunidades (Subunits), la clasificación estructural de sus dominios (Domains) y la forma de su pliegue (Fold).

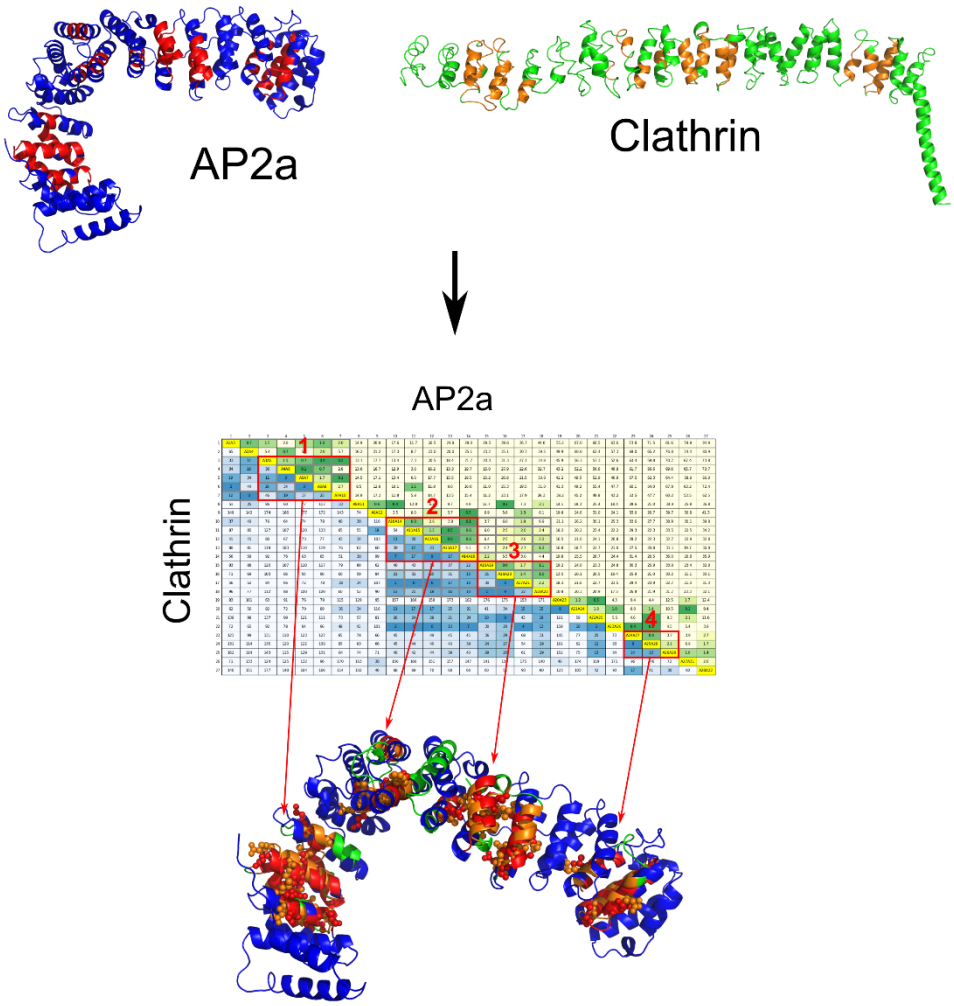


Figura 37. Similitud estructural cercana entre los dominios SPAH de AP2α y Clatrina. Superposición de los dominios SPAH de AP2α (códigos PDB 2jkr, cadena A) y Clatrina (códigos PDB 3lvg, cadena A) según el alineamiento de MOMA2. Los residuos estructuralmente equivalentes son señalados en rojo y naranjo para las estructuras *query* y *target*, mientras que los residuos no alineados son descritos en azul y rojo, respectivamente. La matriz de diferencia reporta cuales fueron los pares de sub-fragmentos seleccionados (rectángulos en color rojo), cuya superposición global se muestra debajo de la matriz.

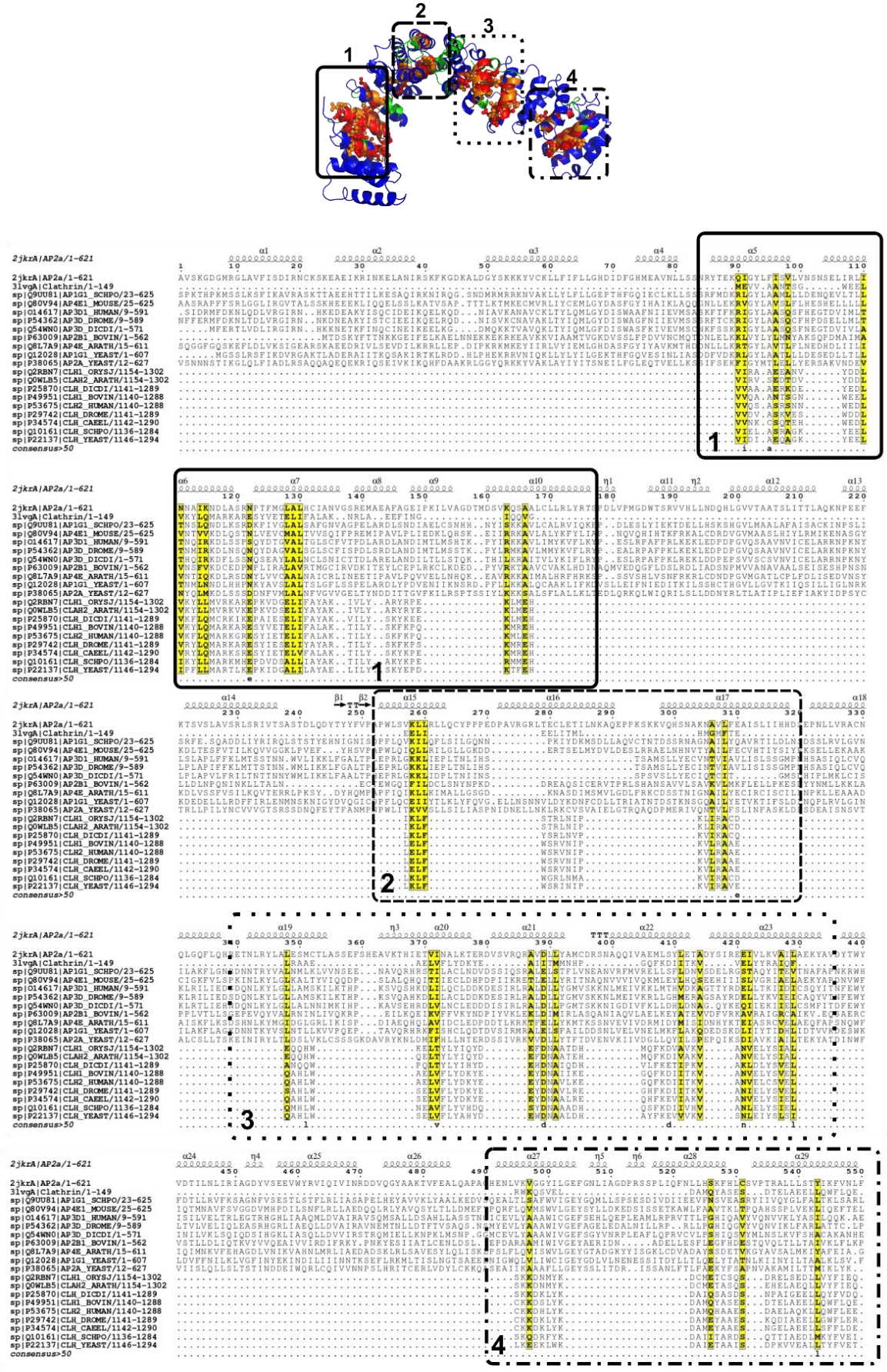

1
2
3
4

Figura 38. Residuos conservados presentes entre los dominios SPAH de AP2α y Clatrina. En la parte superior de la imagen se muestra la superposición estructural de AP2α y Clatrina destacando las regiones superpuestas (rectángulos con línea continua, segmentada, punteada, y punto-trazo, respectivamente), señalando los residuos estructuralmente equivalentes con los colores rojo y naranjo y aquellos conservados usando la representación sphere. En la parte inferior se muestra el alineamiento múltiple de secuencias generado a partir de la comparación estructural de los dominios AP2α y Clatrina, alineando a la vez, las secuencias de sus proteínas homólogas. Los residuos conservados son destacados en amarillo denotando con rectángulos las regiones superpuestas. Esta figura fue generada con el servidor de ESPript 3.0 usando la matriz de puntaje de Risler y un umbral de corte definido por defecto (Global score = 0.75) para indicar los residuos conservados.

3.3.3.5. Árboles filogenéticos derivados de las relaciones encontradas

Hasta la fecha, las proteínas de cubierta de membrana se han clasificado en dos o tres familias considerando solamente las características cualitativas que estas poseen, las cuales incluyen el tipo de arquitectura que presentan, las similitudes estructurales parciales entre algunos de sus miembros y la presencia de elementos compartidos que participan en el ensamblaje de sus complejos (Field *et al.*, 2011; Schlacht and Dacks, 2015; Field and Rout, 2019; Sampathkumar *et al.*, 2013). Estas semejanzas han permitido sugerir que las proteínas de cubierta de las vesículas junto con el poro nuclear comparten un ancestro en común, y a la vez, estas observaciones ha llevado a varios investigadores a plantear escenarios alternativos que establecen conexiones tentativas entre estas proteínas por medio de árboles cualitativos (Schlacht and Dacks, 2015; Rout and Field, 2017). Por otra parte, las proteínas MC poseen bajos porcentajes de similitud entre ellas por lo que es difícil establecer árboles filogenéticos efectivos con los métodos tradicionales (basados en alineamientos múltiple de secuencias) por lo que es necesario probar otro enfoque para poder evaluar las conexiones presentes entre las proteínas MC. Por ello, se utilizó la información derivada de las comparaciones estructurales para construir un

árbol filogenético basado principalmente en las similitudes de las superposiciones reportadas para sus dominios β-propeller y SPAH.

Para llevar a cabo esta tarea, primero fue necesario comparar estructuralmente los dominios cristalizados hasta la fecha para las proteínas MC con MOMA2, realizando comparaciones de todos contra todos para cada conjunto de dominios. Posteriormente, se utilizó los valores de SDM calculados a partir de estos alineamientos y el método de agrupamiento jerárquico UPGMA para construir dos árboles filogenéticos, uno para los dominios β-propeller y otro para los dominios SPAH. Para calcular estos árboles también se incluyeron en el análisis dos proteínas no relacionadas evolutivamente con las proteínas MC que poseen una composición similar de elementos de estructura secundaría a las proteínas MC. Como la Tachilectina-2 que posee un dominio β-propeller irregular y la hemoglobina-α que es una proteína *all*-α que adopta una forma globular, usando ambas estructuras controles negativos para los árboles generados. Segundo, para asegurarse que las relaciones observadas en los árboles creados son fiables se generaron al azar 100 réplicas de los datos originales para construir 100 árboles distintos que son utilizados posteriormente para calcular los valores de soporte a cada uno de los nodos en los árboles originales obtenidos para cada conjunto de dominios (ver la sección 3.3.4.6). De esta manera, las ramas de cada dendrograma estructural cuenta con valores de soporte que simulan de forma parecida a los valores de *bootstrapping* que aparecen en los árboles generados a partir de los alineamientos de secuencias, pero que, en nuestro caso, estos valores indican el número de veces donde una rama o un grupo de elementos conectados aparece en las réplicas calculadas a pesar de la variabilidad

agregada. Por ejemplo, una conexión en el árbol es considerada fiable si esta reporta un valor de soporte mayor a 50 señalando que existe suficiente evidencia para asegurar la relación observada, en cambio, si este valor es menor a 50 no existe suficiente soporte para asegurar la conexión observada en el dendrograma. Por otro lado, a medida que este valor es más cercano a 100, se establece que la conexión posee un alto grado de fiabilidad, sugiriendo una fuerte conexión entre los elementos agrupados.

Gracias a este nuevo modelo cuantitativo de las relaciones entre las proteínas MC podemos apreciar que los dominios β-propeller poseen una mayor fiabilidad en sus conexiones que los dominios SPAH (Figura 39). La mayoría de los valores de soporte de los nodos del dendrograma de los dominios β-propeller son mayores o iguales a 70, sugiriendo que las relaciones establecidas entre estos dominios son bastantemente consistentes. Solamente hubo un nodo que presentó un VS < 70, el cual está presente en la rama que conecta IFT80 con COPβ' señalando una fiabilidad moderada (VS = 63). Esto es debido a que, en las otras réplicas, IFT80 agrupaba con COPα o con la rama conformada por COPα y COPβ'. Sin embargo, la rama que agrupa estos elementos reporta una alta fiabilidad confirmando la relación evolutiva que ha sido descrita anteriormente para esta subunidad del complejo IFT con las subunidades del *cage* de COPI (Dam *et al.*, 2013b).

Es importante señalar que los dominios que hemos clasificados como BI y BII que están presentes en dos ramas separadas del árbol de los dominios β-propeller se conectándose sus integrantes entre sí con una alta fiabilidad. De hecho, las relaciones cercanas presentes en el árbol entre algunos dominios β-propeller como

Nup170 *C. thermophilum* con Nup157 de *S. cerevisiae* (dominios BII), o entre Sec13 de *S. cerevisiae* con su parálogo Seh1 (dominios BI) que mostraron el máximo valor de soporte, confirmando las relaciones descritas para estas proteínas en el estado del arte (Whittle and Schwartz, 2009; Algret *et al.*, 2014). Lo interesante de este análisis es que se aprecia una fuerte conexión entre la nucleoporina Nup37 con el dominio β-propeller de Sec31, ambos pertenecientes a los dominios del tipo BIII y BI que mostraron una conexión estructural cercana reportando una alta fiabilidad (VS = 88). También, podemos destacar la conexión cercana de los dominios β-propeller que pertenecen a las nucleoporinas Nup170, Nup133, y Nup120, cuyas conexiones fueron encontradas en más del 90% de las réplicas agrupándose también el dominio de Clatrina que pertenece a los dominios del tipo BIII siendo el elemento más distante de esta rama. Estas conexiones nos sugieren la posibilidad de que todos estos dominios probablemente evolucionaron y divergieron de un dominio β-propeller ancestral con un plegamiento similar al tipo BI.

Por otro lado, en el árbol de dominios SPAH se observa que los dominios de los tipos AI, AII, y AIII muestran fuertes conexiones intragrupales a diferencia de los dominios β-propeller que se destacan por sus fuertes conexiones intergrupales. Esto se aprecia al observar los nodos que conectan las ramas que agrupan los dominios AI, AII y AIII, que muestran débiles valores de soporte al conectar estas ramas indicando que no existe evidencia suficiente para afirmar que un tipo de dominio SPAH derivó de otro. Además, se observa que las proteínas como COPα y COPβ', Nup188 y Nup192, Importina 13 y Importina β1, y las subunidades de las adaptinas AP1 y AP2 reportan entre ellas valores de soporte iguales y cercanos a

100, corroborando las relaciones descritas anteriormente para estas proteínas (Lee and Goldberg, 2010; Mingot *et al.*, 2001; Andersen *et al.*, 2013, 188; Boehm and Bonifacino, 2001).

A nivel de los dominios SPAH, encontramos que algunas nucleoporinas del poro nuclear muestran similitudes estructurales cercanas con las subunidades de los complejos Clatrina/adaptinas, COPI y COPII. Por ejemplo, el extremo C-terminal de Nup120 mediante un alineamiento flexible se encontró más cercano estructuralmente al dominio α-solenoide de COPε a pesar de ambos dominios a pertenecen a dos grupos distintos (AIII y AIV). En cambio, el dominio α-solenoide de Nup133 muestra un fuerte valor de soporte junto con las dos subunidades de COPI, siendo clasificados juntos en el grupo AI. Adicionalmente, se observa que las subunidades descritas por Brohawn y colaboradores en 2008 presentan una relación cercana a los complejos COPII, mostrando una conexión cercana de Nup84 con Sec16, y de Nup85, Nic96 y Nup145C con Sec31, respectivamente. Estos dominios se agruparon en la misma rama en el 93% de las réplicas reportando una fuerte fiabilidad. Mientras que los resultados obtenidos de las comparaciones estructurales y de las conexiones reportadas en los árboles de los dominios SPAH, muestran que el dominio de Nup170 junto con parte del extremo C-terminal de Nup188 se muestran estructuralmente más cercanas a las subunidades parecidas a las adaptinas agrupándose en el 91% de las veces con la subrama conformada por los dominios descritos anteriormente del tipo *adaptin-like*. Mientras que los dominios SPAH de Nup188 (externo N-terminal) y Nup192 denotan una fuerte conexión estructural con las carioferinas y las adaptinas como se había reportado a través de los gráficos de

Circos. Sin embargo, estos dos últimos grupos poseen un soporte moderado entre ellos. Posiblemente en las réplicas restantes, Nup192 y Nup188N se conectan directamente algunas veces a cada una de estas subramas, eso puede explicar porque la conexión formada por las adaptinas e importinas reporta un valor de soporte igual a 67. En la rama que agrupa los dominios del tipo AIII también se puede distinguir la similitud estructural cercana entre las adaptinas de COPI con las adaptinas AP1 y AP2, cuyas relaciones pueden ser detectadas mediante comparaciones de secuencias, agrupando los dominios SPAH de las subunidades de los complejos AP1 y AP2 con el dominio de COPγ, mientras que el dominio de COPβ que fue uno de los elementos más divergentes dentro de los dominios AIII. El agrupamiento de estos dominios nos señala que el dominio de Clatrina es cercano a los dominios SPAH de las adaptinas, siendo el integrante más distante del grupo AIII cuya conexión muestra un fuerte valor de soporte con la rama que agrupa las nucleoporinas Nup192 y Nup188, carioferinas y adaptinas.

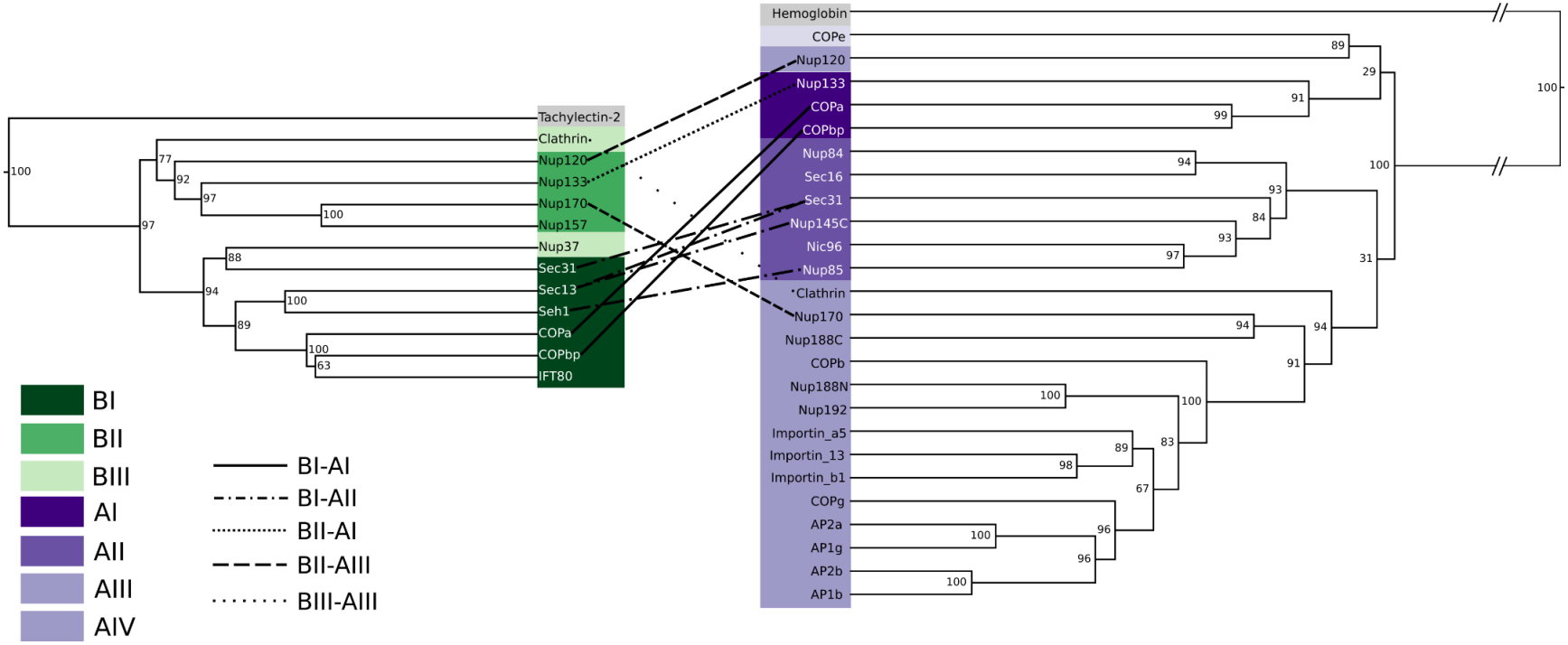


Figura 39. Modelo de las relaciones estructurales entre las proteínas MC según sus dominios.
Los dendrogramas fueron creados a partir de las comparaciones estructurales realizadas con MOMA2 entre los subconjuntos de los dominios β-propeller y SPAH, usando el algoritmo de agrupamiento UPGMA y la métrica SDM. Los valores descritos en los nodos corresponden a los valores de soporte (VS) calculados para cada rama, mientras que los distintos tipos de dominios encontrados entre las proteínas de cubierta de membrana son destacados con los colores descritos en la leyenda. En cambio, los diferentes tipos de líneas que conectan las etiquetas de ambos dendrogramas pareados indican la interacción de dos proteínas distintas que se complementan adoptando una arquitectura β-propeller/SPAH o la presencia de ambos dominios de la misma proteína en ambos árboles. Finalmente, las proteínas tachilectina-2 y la hemoglobina fueron usadas como controles negativos.

Por último, la conexión menos evidente en el árbol de los dominios SPAH fue aquella descubierta entre la subunidad COPε con la nucleoporina Nup120 que mostró un fuerte valor de soporte (VS = 89), a pesar de que ambas proteínas muestran un plegamiento estructural diferente. En el caso de Nup120 su extremo C-terminal presenta un arreglo de hélices junto con su dominio β-propeller formando un pliegue único, mientras que su extremo C-terminal está constituido por pares de hélices en tándem que se extienden interactuando con los extremos C-terminales de las nucleoporinas Nup145C y Nup85 formando el *triskelion* del complejo Y (Stuwe *et al.*, 2015). Por otro lado, la subunidad COPε adopta un pliegue helicoidal que encapsula una horquilla β que sobresale del extremo C-terminal de COPα (Hsia and Hoelz, 2010). A pesar de estas diferencias, por medio del alineamiento de matrices de MOMA2 identificamos la presencia de tres pares de sub-fragmentos que presentan bajas diferencias angulares y de distancia a nivel de sus pares de SSEs (Figura 40). Estos pares a la vez constituyen una superposición global que reporta un porcentaje de *overlap* estructural de un 38%, alineándose en total 111 residuos equivalentes con un $RMSD_{pond}$ igual a 2.21 Å y mostrando un 12% identidad de secuencia. Aunque a través de comparaciones estructurales flexibles se conectaron estos dominios divergentes en el árbol de los dominios SPAH, los resultados de este análisis señalan que aún no contamos con el soporte suficiente para afirmar con seguridad que el dominio SPAH de Nup120 comparte una relación distante con las otras subunidades del complejo COPI a pesar de que esta rama si se conecta con

aquella conformada solamente por Nup133, COPα y COPβ' cuyo VS fue igual a 29.

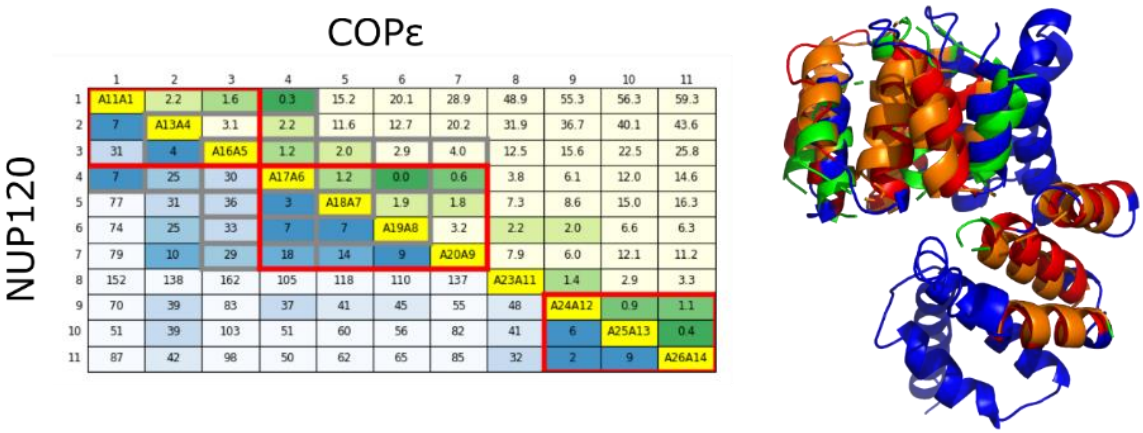


Figura 40. Superposición estructural de COPε y Nup120.
Esta figura muestra el alineamiento estructural calculado con MOMA2 entre el dominio α-solenoide de COPε (código PDB 3mv2, B) y Nup120 (código PDB 4xmn, E). El alineamiento de matrices destaca las diferencias angulares y de distancia entre los pares de elementos de estructura secundaria alineados, indicando con bloques rojos los sub-fragmentos equivalentes. Mientras que la superposición estructural muestra en rojo y naranjo los residuos estructuralmente equivalentes entre las estructuras *query* (azul, COPε) y *target* (verde, Nup120).

3.3.3.5.1. Diferentes tipos de arquitecturas presentes entre los complejos MC

El siguiente paso de nuestro análisis fue enriquecer con información adicional las conexiones comunes presentes en ambos dendrogramas pareados obtenidos a partir de cada conjunto de dominios. Para ello, se emplearon distintos tipos de líneas para conectar las etiquetas de aquellos dominios que pertenecen a una misma proteína (como COPβ', COPα, Sec31, Nup120, Nup133, Nup170 y Clatrina) o para conectar los dominios de las proteínas que interactúan entre sí adoptando una arquitectura del tipo β-propeller/SPAH (como Seh1-Nup85, Sec13-Sec31 y Sec13-Nup145C) (Figura 39). Estas conexiones nos permitieron identificar cinco arquitecturas del tipo β-propeller/SPAH entre las proteínas MC, las cuales son descritas a continuación:

a) La combinación BI-AI que está presente en las subunidades que componen el *cage* del complejo COPI (COPα y COPβ').
b) La combinación BI-AII que se encuentra presente en las subunidades de los complejos COPII y las nucleoporinas del complejo Y, donde algunas nucleoporinas adoptan esta arquitectura al interactuar una proteína que posee solamente un dominio β-propeller incompleto inestable que interactúa en *trans* con los dominios α-solenoide de algunas nucleoporinas, por medio de un *blade* aislado.
c) La combinación BII-AI se encuentra presente solamente en la proteína Nup133.

d) La combinación BII-AIII se encuentra presente en las nucleoporinas Nup120 y Nup170.

e) La combinación BIII-AIII se encontró principalmente en la proteína Clatrina.

Al analizar estas combinaciones podemos señalar también que hay otras subunidades dentro de los complejos MC que poseen solamente un solo dominio β-propeller o SPAH que interactúan en conjunto con las proteínas que poseen la arquitectura β-propeller/SPAH, destacando por ejemplo, la subunidad COPε que interactúa con el extremo C-terminal de COPα (Lee and Goldberg, 2010). En el poro nuclear encontramos la proteína Nup37 que es un dominio β-propeller completo del tipo BIII que ha sido cristalizado junto con Nup120, interactuando directamente con esta nucleoporina para mantener la asociación entre las nucleoporinas transmembrana y aquellas que conforman el complejo Y (Bilokapic and Schwartz, 2012). Por otro lado, las carioferinas, adaptinas y algunas nucleoporinas en cambio poseen solamente un dominio AIII y, a que diferencia de las nucleoporinas COPII-*like*, estas subunidades no poseen o interactúan en *trans* con algún dominio β-propeller incompleto.

En total hemos descrito nueve tipos diferentes de combinaciones de dominios, incluyendo las combinaciones β-propeller/SPAH, sólo β-propeller o sólo SPAH, las cuales están presentes en los principales complejos MC de la siguiente manera (Figura 39):

- Complejos Clatrina/Adaptinas: BIII-AIII, AIII
- Complejos COPI: BI-AI, AIII, AIV

- Complejos COPII: BI-AII, AII, BI
- Poro nuclear: BII-AI, BII-AIII, BI, AII, AIII, BIII
- Carioferinas αβ: AIII

Interesantemente, podemos destacar que el *scaffold* del poro nuclear está compuesto por varias proteínas que presentan una mezcla de dominios aislados y algunas combinaciones cuyos tipos también están presentes en los principales complejos de cubierta de vesículas, como el par BI-AII (COPII), o los dominios aislados BI (presentes en todos los complejos CV), AII (COPII), AIII (adaptinas). Adicionalmente, en el poro nuclear podemos encontrar combinaciones de únicas de dominios como BII-AI o BII-AIII, cuyos dominios BII presentan inserciones de hélices entre sus *blades* donde sus dominios de SPAH son parecidos a las subunidades de los complejos COPI y adaptinas. Otras subunidades presentes en el poro nuclear se complementan entre sí adoptando una arquitectura β-propeller/SPAH del tipo COPII-like (o BI-AII con el nuevo esquema de clasificación), como Seh1-Nup85 y Sec13-Nup145C. Estas combinaciones de dominios denominados “fósiles moleculares” nos sugieren que la interacción de dos o tres tipos diferentes de arquitecturas dio origen a un poro temprano, que luego mediante eventos de duplicación y divergencia, éste fue evolucionando hasta los complejos modernos que hoy conocemos en los organismos eucariotas (Field and Rout, 2019).

Por otra parte, a partir de las conexiones pareadas entre estos árboles podemos especular acerca de la presencia de algunas combinaciones de los dominios β-propeller/SPAH en otros complejos relacionados en este último

tiempo con las proteínas de cubierta de membrana, como los complejos IFT y SEA (Rout and Field, 2017). Por ejemplo, las comparaciones estructurales realizadas para el dominio β-propeller de la proteína IFT80 muestran que esta subunidad de complejo IFT es estructuralmente más cercano a las subunidades del complejo COPI (Figura 39). Sin embargo, su dominio α-solenoide no ha sido cristalizado completamente, de hecho, la estructura disponible de la proteína IF80 (PDB: 5n4a) presenta apenas dos pares de hélices las cuales no son suficientes para evaluar su similitud estructural con los demás dominios α-solenoide. Pero considerando las posiciones que adoptan los dominios α-solenoide de COPβ' y COPα en el árbol de dominios SPAH se puede inferir que la proteína IFT80 cuenta, al igual que estas proteínas, con un dominio del tipo AI, si consideramos también las predicciones de estructura secundaria descritas en la literatura que sugieren que la proteína IFT80 es estructuralmente similar a estas subunidades (Dam *et al.*, 2013b). Dam y colaboradores señalaron también que algunas subunidades del complejo IFT como TTC21, IFT88, TTC26, BBS4 y BBS8 adoptan un dominio estructural similar a COPε, por lo que también podemos suponer que este complejo cuenta con dominios del tipo AIV (Dam *et al.*, 2013b). En el caso de los complejos SEA, la presencia de la nucleoporina Seh1 sugiere que existe una conexión evolutiva entre el poro nuclear y los complejos SEA (Dokudovskaya *et al.*, 2011). En los dendrogramas calculados se aprecia que este β-propeller incompleto interactúa con Nup85 adoptando una combinación BI-AII, y tomando en cuenta esto, es posible sugerir que la proteína Seh1 interactúe

del mismo modo con otras subunidades SPAH del complejo SEA adoptando una conformación del tipo BI-AII (Figura 39).

3.3.3.6. Escenarios posibles de la evolución de las proteínas MC

Considerando el hecho de que las proteínas de cubierta de membrana evolucionaron a partir de un ancestro común, algunos artículos han planteado ciertos escenarios que tratan de explicar el posible origen de los complejos de cubierta de vesículas y su conexión con el origen del poro nuclear (Devos *et al.*, 2004; Field and Dacks, 2009; Sampathkumar *et al.*, 2013; Rout and Field, 2017; Field and Rout, 2019). Entre los escenarios planteados podemos destacar lo que sugirieron Sampathkumar y colaboradores en 2013, los cuales describieron dos posibles escenarios que pueden explicar la aparición de los actuales complejos MC según las arquitecturas que presentan las subunidades que conforman estos complejos. Uno de estos escenarios plantea que el origen de estos complejos se debió a la reintegración de complejos separados derivados del protocoatomer en complejos híbridos. En cambio, el segundo escenario plantea que al menos tres familias de proteínas derivadas del protocoatomer evolucionaron en conjunto a partir de un solo complejo ancestral, y que luego por eventos de duplicación y divergencia se formaron los complejos actuales (Sampathkumar *et al.*, 2013).

Para analizar estos posibles escenarios, se realizó una reconstrucción evolutiva de los distintos tipos de combinaciones de dominios presentes en las proteínas de cubierta de membrana, considerando como punto de partida, los dendrogramas obtenidos a partir de sus comparaciones estructurales y las distintas arquitecturas que poseen de acuerdo con la nueva clasificación que hemos

desarrollado. Para llevar a cabo este análisis, se ocupó el método de Maddison y Slatkin (2013) que considera que los rasgos o las características observadas en un árbol filogenético tienen igual costo de pasar de un estado a otro (en términos de "eventos evolutivos"). Este método se encarga de evaluar el número de transiciones de los caracteres observados en un dendrograma con respecto a una distribución de cambios originados de manera aleatoria mediante un modelo nulo (ver sección 3.3.2.7). El método de Maddison y Slatkin se ocupó específicamente para evaluar el número de transiciones de las distintas arquitecturas observadas en las proteínas alineadas considerando el agrupamiento de sus dominios β-propeller y SPAH. Para ello, se corrieron 100 réplicas por cada árbol de dominios debido a que este modelo genera distribuciones aleatorias en cada repetición para determinar finalmente el número promedio de transiciones observadas.

Los resultados obtenidos de este análisis indican que, independientemente del tipo de árbol generado, el número de transiciones observadas en las 100 repeticiones calculadas fue en todos los casos menor que el promedio de transiciones definidas al azar por los modelos nulos. En el árbol de los dominios β-propeller se observó que el número promedio de transiciones requeridas para explicar las combinaciones de dominios presentes en las proteínas agrupadas fue igual a 7 mostrando valores de *P-value* entre 0.003 y 0.016. En cambio, el número promedio de transiciones observadas en el árbol de los dominios SPAH fue igual a 9 mostrando *P-values* ≤ 0.001 (ejemplos de un par de estas réplicas se muestran en la Figura 40). Estos resultados sugieren la posibilidad de que el origen de los distintos tipos de arquitecturas que observamos actualmente en las proteínas de cubierta de membrana

se debió a la transición a partir de un número menor de cambios evolutivos o transiciones de un complejo ancestral que poseía al menos dos o tres tipos de subunidades que mostraban arquitecturas diferentes del tipo β-propeller/SPAH. Este resultado apoya el segundo escenario parsimonioso planteado por Sampathkumar y colaboradores que sugieren que tres tipos de arquitecturas evolucionaron a partir de solo complejo ancestral que divergió y se especializó en los complejos actuales, reteniendo varios aspectos de las tres arquitecturas originales (Sampathkumar *et al.*, 2013).

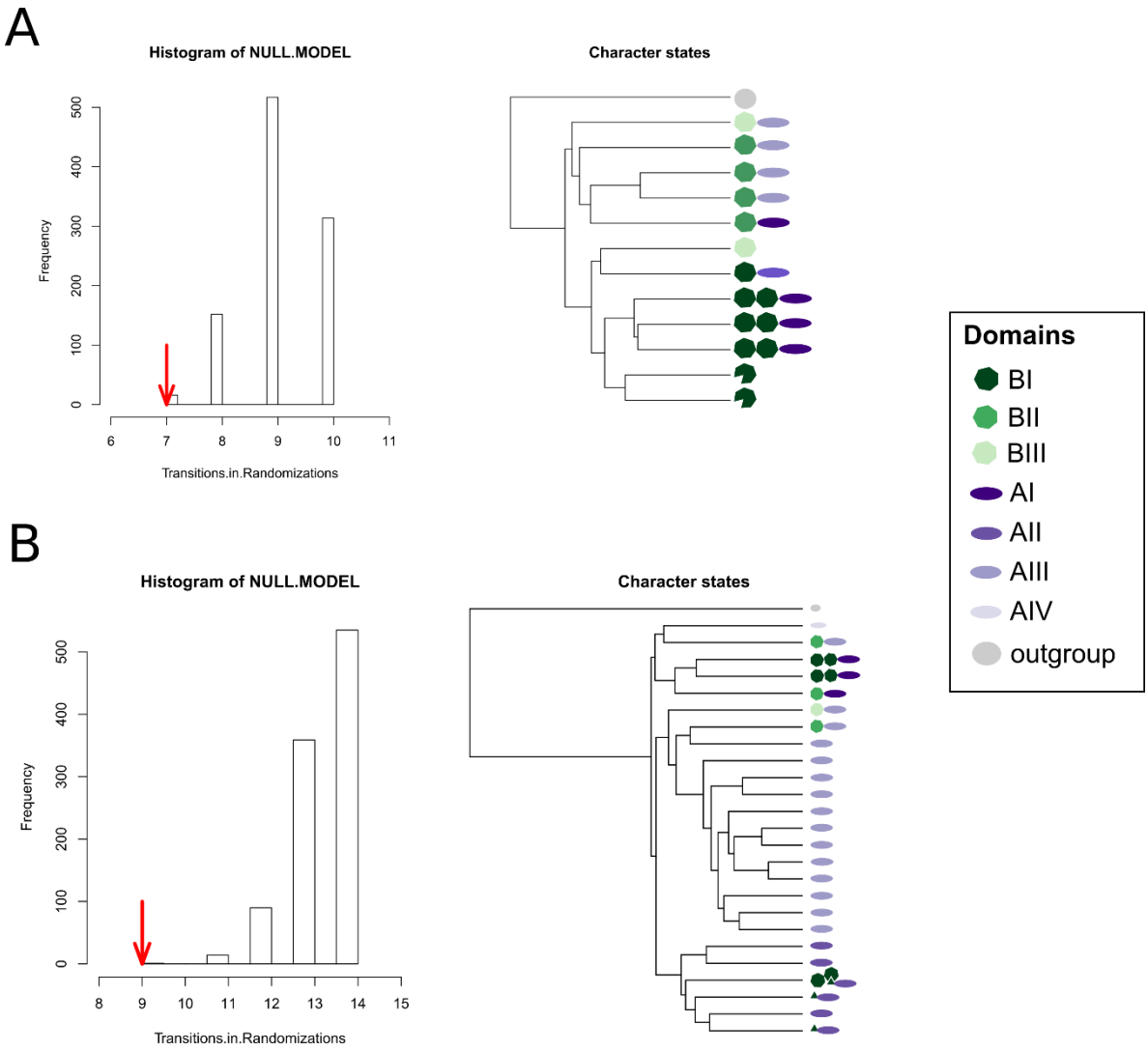


Figura 41. Evaluación de las transiciones de las distintas arquitecturas observadas entre las proteínas MC.

Cada histograma es una réplica generada con el método de Maddison y Slatkin que evalúa el número de transiciones observadas (flecha roja) con respecto a una distribución de transiciones aleatorias de los caracteres que son calculadas a partir de un modelo nulo. En estos ejemplos se consideró las combinaciones de dominios de los elementos agrupados a partir de las comparaciones realizadas entre sus dominios β-propeller (A) y entre sus dominios SPAH (B). En cada árbol, específicamente cada combinación es representada mediante hexágonos verdes que representan los distintos tipos de dominios β-propeller y elipses moradas que representan respectivamente los diferentes tipos de dominios SPAH.

3.3.4. Discusión

3.3.4.1. Clasificación funcional y estructural de las proteínas MC

En este trabajo, hemos propuesto dos nuevos esquemas para la clasificación de las proteínas de cubierta de membrana, uno de ellos corresponde a un esquema funcional que clasifica estas proteínas en cuatro grupos según el tipo de subunidad que posee el complejo, mientras que el otro esquema es propiamente estructural clasificando el pliegue de sus dominios en siete grupos, respectivamente.

El primer esquema nos ofrece una ventaja al clasificar las subunidades de forma funcional en las clases “Cage-CV”, “Adaptor-CV”, “Cage-NPC” y “Adaptor-NPC”, porque nos permite hacer un paralelismo de los elementos móviles y rígidos presentes en los complejos de cubierta de vesícula *versus* aquellos presentes en el poro nuclear. Considerando que el poro nuclear posee nucleoporinas que comparten un patrón de plegamiento común con complejos de cubierta de membrana y probablemente comparten un origen común con Clatrina, COPI y COPII, este tipo de diagrama nos permitió explorar las relaciones distantes que existen entre sus subunidades que no son detectadas a partir de comparaciones de secuencia primaria, pero que pueden ser reveladas usando métodos sensibles de comparación estructural flexible como MOMA2.

Mientras que el segundo esquema se apoya completamente en las comparaciones obtenidas con MOMA2 entre los dominios que componen estas proteínas. Esta nueva clasificación agrupa los dominios β-propeller de las proteínas MC en los tipos BI, BII y BIII y sus dominios SPAH se agrupan respectivamente en

los tipos AI, AII, AIII y AIV. Esta nueva clasificación puede parecer contraria al criterio actual que se basa en el tipo de arquitectura que poseen principalmente los complejos COPI y COPII, incluyendo en algunos casos el tipo de pliegue que adoptan las adaptinas (Rout and Field, 2017; Sampathkumar *et al.*, 2013; Field and Rout, 2019). Sin embargo, al igual que la clasificación descrita en el estado del arte, la clasificación que proponemos agrupa cercanamente aquellas subunidades relacionadas evolutivamente entre los diferentes complejos, por ejemplo, los dominios SPAH de algunas nucleoporinas como Nic96, Nup145C, Nup84 y Nup85 junto con algunas subunidades del complejo coatomer II que poseen una arquitectura del tipo II según Field y Rout, sus dominios fueron clasificadas con el nuevo esquema dentro del mismo grupo, en el tipo AII. Otro ejemplo de esto lo constituyen las nucleoporinas Nup192 y Nup188 que adoptan un pliegue similar a las adaptinas (*adaptin-like*) según Sampathkumar y colaboradores, y del mismo modo mediante los resultados obtenidos de sus comparaciones estructurales flexibles hemos clasificado estas nucleoporinas dentro del grupo AIII junto con las subunidades adaptadoras del complejo coatomer I, carioferinas, y las adaptinas AP1 y AP2 (Field and Rout, 2019; Beck *et al.*, 2018).

Sin embargo, este nuevo esquema estructural no agrupa algunos dominios dentro de un mismo grupo considerando que algunas nucleoporinas como Nup170, Nup133, y Nup120 se han clasificado en el estado del arte de acuerdo al tipo de arquitectura que poseen dentro del tipo I o COPI-*like* (Field and Rout, 2019). Pero es importante señalar que la clasificación que hemos desarrollado se basa en los tipos de dominios que posee una subunidad de manera individual según sus

comparaciones estructurales, en lugar de considerar como los dominios de estas subunidades interactúan entre sí parecido a lo que se aprecia en los complejos COPI y COPII. Aunque con el nuevo esquema sí se observa una cercanía estructural de una de estas nucleoporinas con algunas subunidades que forman parte del *cage* en el complejo COPI, como el dominio SPAH de Nup133 que se agrupo junto a los dominios de las subunidades COPβ' y COPα conformando el grupo AI, mientras que Nup170 al ser clasificado junto con COPγ y COPβ dentro del grupo AIII, sugiere que su dominio α-solenoide es más parecido al plegamiento estructural que adoptan las adaptinas.

Pese a que los dominios β-propeller de las proteínas de cubierta de membrana son estructuralmente parecidos entre sí y, de cierto modo, menos flexibles que los dominios α-solenoide, hemos podido distinguir tres tipos distintos mediante alineamientos estructurales flexibles (Field *et al.*, 2011). Hemos encontrado que los dominios β-propeller de tipo BI están presentes en los complejos de COPI, COPII, IFT y en el *scaffold* del poro nuclear, mientras que el tipo BII es exclusivo a algunas nucleoporinas, mientras que Clatrina y Nup37 conforman el grupo BIII. La principal diferencia que separa a ambos tipos es la presencia de hélices intercaladas entre algunas *blades* que es característico de los dominios de tipo BII y no se encuentra presente en los dominios del tipo BI.

Por otra parte, mediante los alineamientos múltiple de secuencias basados en comparaciones estructurales, hemos descubierto motivos conservados a nivel de secuencia principalmente entre los tipos BI, BII y BIII donde la fuerte similitud estructural nos permitió suponer que aún quedan vestigios de tipos de residuos que

probablemente estuvieron presentes en el dominio β-propeller del protocoatomer ancestral a pesar de que estas estructuras han divergido considerablemente durante la evolución. Sin embargo, dicho análisis no pudo llevarse a cabo para los dominios SPAH incluyendo los tipos AI, AII, AIII y AIV debido a que estos dominios son estructuralmente más divergentes que los dominios β-propeller presentando generalmente una débil o casi nula señal a nivel de sus secuencias (menor a 15% de identidad de secuencia en la mayoría de los casos). Estos resultados nos indica que, aunque las comparaciones a nivel de secuencia son insuficientes para establecer relaciones entre estas proteínas, al usar la información derivada de sus estructuras si nos es posible descubrir motivos remanentes presentes entre sus dominios que probablemente sean relevantes para mantener su forma característica, especialmente si hablamos de los dominios β-propeller cuya forma ha permanecido con menores cambios de forma que los dominios SPAH.

3.3.4.2. Relaciones estructurales entre las proteínas MC

Las comparaciones estructurales no solamente nos han proporcionado reorganizar el esquema de clasificación actual de las proteínas de cubierta de membrana, sino que además nos han permitido explorar ampliamente las relaciones estructurales que existen entre estas proteínas, confirmando relaciones evolutivas descritas anteriormente, extendiendo las relaciones ya conocidas y encontrando nuevas.

Mediante los gráficos de *Circos*, hemos podido explorar estas relaciones considerando de forma simultánea las similitudes estructurales de las proteínas de cubierta de membrana a través de los *matches* locales de sus dominios β-propeller y

SPAH (Figuras 26, 31, 32, 34 y 36), construyendo una imagen global de las relaciones presentes entre estas proteínas a pesar de que algunas de ellas han sido cristalizadas completamente hasta la fecha (Tablas 3 y 4).

Las comparaciones realizadas confirman la fuerte similitud estructural que poseen los dominios β-propeller de las subunidades que forman parte del *cage* en los complejos de cubierta de vesículas con los dominios de las nucleoporinas que conforman el *scaffold* del poro nuclear. Igualmente, estos resultados corroboran las similitudes estructurales remotas descritas entre los dominios α-solenoide de algunas nucleoporinas que adoptan una arquitectura COPII-like (como Nic96 o Nup145C) con subunidades del complejo COPII, o entre algunas Nups conocidas como *adaptin-like* (que incluye Nup188 y Nup192) con las carioferinas. Las similitudes encontradas dan soporte a que estas proteínas comparten un ancestro en común (Protocoatomer hipótesis), y que existe una relación evolutiva entre algunas nucleoporinas con los receptores de transporte nuclear (la hipótesis "*core/adaptor*"). Además, los gráficos de *Circos* nos permitieron extender las relaciones descritas para los receptores del transporte nuclear y las nucleoporinas Nup192 y Nup188 con los dominios adaptadores de COPI y las adaptinas AP1 y AP2, señalando que no solamente existen subunidades con una arquitectura tipo I parecida a los complejos del tipo COPI, como se ha postulado en estos años, sino que es posible que haya un tercer grupo del tipo *adaptin-like* (Sampathkumar *et al.*, 2013; Field and Rout, 2019).

Las comparaciones flexibles de estas proteínas con MOMA2 ha permitido dilucidar también nuevas relaciones como la conexión presente entre Nup120 con Nic96, que conecta dos tipos dominios que poseen un plegamiento distinto de sus

dominios SPAH. Sin embargo, aunque el extremo N-terminal del dominio helicoidal de Nup120 adopta un pliegue único con su dominio β-propeller, aún podemos distinguir en su extremo C-terminal muestra una cercanía con los bloques *crown*, *trunk* y *tail* de los dominios ACE1, cuyo representante es Nic96. Por otra parte, podemos destacar la relación cercana de Nup170 con Clatrina, tanto a nivel de su dominio β-propeller como SPAH y también de la Importina 13 (receptor de Importina β) con la nucleoporina Nup192 y la adaptina AP1γ.

En los árboles generados mediante comparaciones estructurales de los dominios MC (Figura 39), nos señalan el agrupamiento cercano de las subunidades relacionadas evolutivamente dentro de los complejos de cubierta de membrana cuyas comparaciones estructurales y de secuencias dan pistas de que comparten un origen evolutivo común. La gran mayoría de los ejemplos descritos que reportan fuertes conexiones estructurales también se han corroborado de forma simultánea con las comparaciones realizadas entre sus perfiles de HMM. Podemos mencionar entre estos ejemplos, la conexión cercana con una alta fidelidad entre las proteínas Sec13 y Seh1 dentro del árbol de los dominios β-propeller, o entre los dominios β-propeller de Nup170 junto con su parálogo Nup157 y su homólogo remoto Nup133 (Seo *et al.*, 2013; Debler *et al.*, 2008; Whittle and Schwartz, 2009, 133). También, podemos dar cuenta de la conexión cercana descrita en los árboles para los dominios de las subunidades del complejo COPI como COPβ’ y COPα, cuyos alineamientos estructurales y de secuencia señalan que han divergido de un ancestro en común (Lee and Goldberg, 2010). Es importante mencionar que no solamente confirmamos las relaciones encontradas entre COPα y COPβ’, sino que también extendimos sus

conexiones con la subunidad IFT80 del complejo IFT, confirmando las similitudes encontradas por van Dam y colaboradores en 2013. Estos investigadores a través de búsquedas entre perfiles de HMM señalaron la relación evolutiva de varias subunidades del complejo IFT con aquellas del complejo COPI, indicando que IFT80 es parecido COPβ' (Dam *et al.*, 2013b). Las comparaciones estructurales también confirmaron las relaciones distantes descritas anteriormente entre las subunidades de las adaptinas AP1 y AP2 con las subunidades adaptadoras COPγ y COPβ del complejo COPI, las cuales son ampliamente aceptadas que comparten una homología remota que puede ser detectada a nivel de secuencia (Faini *et al.*, 2013).

Las conexiones observadas entre los árboles de los dominios β-propeller y SPAH también confirmaron relaciones que se han sugerido anteriormente para las proteínas de cubierta de membrana a través de predicciones de elementos de estructura secundaria y comparaciones estructurales (Devos *et al.*, 2004; Brohawn *et al.*, 2008a). Por ejemplo, el agrupamiento de las nucleoporinas que han sido descritas anteriormente que poseen una arquitectura del tipo I o del tipo II en diferentes ramas del árbol de dominios SPAH (Beck *et al.*, 2018), y que nosotros hemos reclasificado sus dominios como AI y AII, respectivamente (Beck *et al.*, 2018). Además, mediante comparaciones flexibles constatamos las similitudes reportadas por otros métodos de comparación estructural como DALI que señalan que las nucleoporinas Nup192 y Nup188 son más cercanas a las carioferinas, pero que en nuestro caso reportamos porcentajes de cobertura estructural mayores a los registrados en la literatura (SO > 10%)(Sampathkumar *et al.*, 2013; Stuwe *et al.*, 2014).

No solamente se reafirmaron varias de las relaciones descritas en la literatura a partir del árbol de los dominios SPAH, sino que también extendimos algunas de ellas. Un claro ejemplo es la fuerte similitud estructural observada entre algunas nucleoporinas y las carioferinas con las subunidades adaptadoras de clatrina y COPI. Estas proteínas poseen dominios que adoptan un pliegue similar a una super-hélice conformada por varios pares de α-hélices en tándem por lo que algunos investigadores han clasificado estos dominios dentro de un mismo tipo (Beck *et al.*, 2018; Sampathkumar *et al.*, 2013). Hasta la fecha solamente se ha estudiado en profundidad la similitud estructural y funcional entre las nucleoporinas Nup188 y Nup192 junto con las carioferinas, dando pistas de que probablemente estas proteínas co-evolucionaron junto con el complejo del poro nuclear, pero no se ha descrito ningún artículo que abarque el estudio de las similitudes estructurales que poseen Nup192 y Nup188 con las adaptinas, y viceversa entre las adaptinas y las carioferinas como se llevó a cabo en este trabajo.

Finalmente, hemos encontrado nuevas relaciones entre algunas proteínas de cubierta de membrana según las comparaciones estructurales de sus dominios. En el árbol de los dominios β-propeller podemos apreciar en una misma rama la conexión de la nucleoporina Nup37 con los dominios del tipo BI, indicando que esta proteína es muy parecida a los dominios β-propeller presentes en las subunidades de los complejos COPI y COPII, a diferencia de otras Nups que presentan inserciones de hélices entre sus *blades* (Figura 39). En cambio, en el árbol de dominios SPAH se aprecia la agrupación de Clatrina y Nup170 con los dominios con un plegamiento parecido a las adaptinas (Figura 39). Aunque en la literatura solamente se ha hecho

mención brevemente de que Nup170 es más cercano estructuralmente a Clatrina (Sampathkumar *et al.*, 2013), mediante nuestras comparaciones estructurales hemos encontrado una nueva relación remota entre Clatrina y AP2α que nos permite extender nuestra visión de las relaciones distantes que guardan estos dominios con las proteínas del tipo *adaptin-like* (Figuras 33 y 37).

3.3.4.3. Arquitecturas encontradas entre las proteínas MC

Al comparar los dominios de las subunidades de los complejos de cubierta de membrana, notamos la presencia de alrededor de cinco tipos diferentes de combinaciones regulares de los dominios β-propeller/SPAH, al igual que cuatro tipos de subunidades que presentan un sólo dominio β-propeller o SPAH que están presentes en los complejos COPI, COPII, IFT, clatrina/adaptinas y el poro nuclear.

En los complejos de cubierta de vesículas encontramos pocas combinaciones de dominios diferentes a diferencia del poro nuclear, donde podemos ver la presencia compartida de dominios β-propeller del tipo BI entre los complejos COPI y COPII, o la presencia compartida de dominios SPAH del tipo AIII entre COPI y los complejos clatrina/adaptinas. Enfocándonos solamente en el poro nuclear, podemos apreciar que este complejo se encuentra compuesto por una mezcla de subunidades que poseen diferentes tipos de combinaciones de dominios sugiriendo probablemente que el poro nuclear evolucionó a partir de una amalgama de proteínas que poseían arquitecturas del tipo COPI-*like*, COPII-*like* y *adaptin-like* (Field and Rout, 2019; Sampathkumar *et al.*, 2013; Beck *et al.*, 2018). Especialmente podemos apreciar combinaciones del tipo BI-AII, o dominios del tipo BI, AI, AIII en las subunidades que constituyen el *scaffold* del poro nuclear. Solamente a nivel de sus

dominios SPAH, encontramos algunas nucleoporinas consideradas del tipo COPI-like o tipo I como Nup133, Nup170, Nup188 y Nup192 que sus dominios SPAH han sido clasificados junto a los componentes de los complejos COPI y Clatrina/adaptinas en los grupos AI y AIII con la clasificación que hemos propuesto. Mientras que las nucleoporinas que adoptan un plegamiento similar a Sec31 como Nic96, Nup145C, Nup84 y Nup85 consideradas en la literatura como subunidades del tipo II o COP-like, cuyos dominios SPAH han sido clasificados dentro del tipo AII.

Además, el poro nuclear presenta algunas innovaciones con respecto a las combinaciones observadas en los complejos de cubierta de vesículas. De hecho, las nucleoporinas Nup120, Nup133 y Nup170 presentan dominios β-propeller (que hemos denominado BII) que presentan varias decoraciones o inserciones de pequeños grupos de hélices entre algunos *blades* a diferencia de los dominios β-propeller de tipo BI. Podemos destacar, por ejemplo, la topología única de la proteína Nup120, donde se observa una fusión entre el dominio β-propeller y parte del dominio α-solenoide que no se aprecia en los dominios β-propeller de otras nucleoporinas conocidas o en las subunidades presentes en los complejos de cubierta de vesículas (Leksa *et al.*, 2009).

Por último, aunque en el poro nuclear no hay nucleoporinas donde el dominio del tipo BI aparezca fusionado con el dominio del tipo AII, como el caso de Sec31, sí se aprecia que algunas nucleoporinas cuyos dominios SPAH son parecidos a Sec31 y Sec16 junto con los dominios β-propeller incompletos como Sec13 o Seh1 forman

heterodímeros que reflejan el mismo tipo de arquitectura que poseen las subunidades de los complejos COPII.

3.3.4.4. Escenario evolutivo que dio origen a las proteínas MC

Al evaluar el número de transiciones que dieron lugar a los distintos tipos de combinaciones observadas en las proteínas MC eucariotas, nuestros resultados apuntan a que el origen de estas proteínas se debió a un escenario más parsimonioso donde al menos tres o dos tipos diferentes de arquitecturas evolucionaron juntas a partir de un solo complejo ancestral, el cual divergió y se especializó en los complejos que observamos actualmente en los eucariotas modernos (Sampathkumar *et al.*, 2013).

Considerando esto, hemos propuesto una modificación al escenario plausible establecido por Sampathkumar y colaboradores según los resultados obtenidos al comparar los diferentes tipos de dominios presentes en las proteínas de cubierta de membrana (Figura 42). En este escenario modificado planteamos que los complejos actuales probablemente evolucionaron a partir de un último complejo común que al menos poseía tres o dos tipos de subunidades que presentaban combinaciones del tipo BI-AI, BI-AII y sólo AIII, porque al menos uno sólo de sus dominios se encuentra en la mayoría de los complejos conocidos que derivan del protocoatomer. Posiblemente los diferentes tipos de arquitecturas que presentan actualmente las proteínas MC se pueden explicar a través de eventos de duplicación de sus dominios, pérdida secundaría y divergencia, los cuales permitieron que estos complejos fueran evolucionando, dando lugar a los diferentes tipos que conocemos actualmente. Esta modificación al escenario propuesto por Sampathkumar y colaboradores en 2013,

nos sugiere en primera instancia que tipos de dominios inicialmente dieron origen a los demás tipos presentes en las proteínas MC, pero no establece cuales fueron los pasos evolutivos que dieron origen a estas proteínas. Pero con los antecedentes descritos en el estado del arte y los resultados obtenidos a partir de este trabajo podemos inferir intuitivamente estos posibles pasos.

Entre estos antecedentes tenemos que algunos tipos dominios probablemente evolucionaron a partir de otros. A través de las comparaciones estructurales realizadas con MOMA2 podemos suponer que los dominios BII y BIII, que se encuentran presentes en las nucleoporinas Nup170, Nup133, Nup120, Nup37, probablemente divergieron de un dominio β-propeller ancestral del tipo BI (Figuras 24 y 39). Mientras que los dominios del tipo BI se encuentran presentes tanto en las proteínas de cubierta de vesículas como el poro nuclear, permitiéndonos conectar todos estos complejos entre sí. Sin embargo, entre los dominios SPAH, esta relación de qué tipo de dominio provino de quién no es tan evidente considerando el análisis filogenético estructural que generamos en este trabajo (Figura 39). Por otro lado, la presencia de los dominios del tipo AIII presentes tanto en complejos clatrina/adaptinas, COPI y en algunas nucleoporinas del *scaffold* del poro nuclear (Figuras 34 y 39), nos revela la conexión profunda que presenta el poro nuclear con estos complejos. Al igual que los dominios AII de algunas nucleoporinas que en la literatura se define este tipo de pliegue como ACE1 *fold*, conecta también el poro nuclear con los complejos COPII (Figuras 28 y 39) (Brohawn *et al.*, 2008b). Por otra parte, los complejos COPII no presentan dominios del tipo AIII porque muy probablemente mediante un reemplazo secundario, estos complejos adoptaron

proteínas con un plegamiento diferente a los dominios SPAH para cumplir la misma función que las adaptinas (Sec23-Sec24) antes de la aparición de LECA (Schlacht and Dacks, 2015). Además, gracias a la clasificación de los dominios extendemos la visión de los tipos de dominios que presentan algunos complejos relacionados recientemente con las proteínas de cubierta de membrana (Figura 42). Por ejemplo, considerando las subunidades que tienen en común los complejos SEA con el poro nuclear y los complejos COPII (proteínas Sec13 y Seh1), proponemos que este complejo presenta probablemente subunidades que poseen o interactúen entre sí adoptando combinaciones de dominios del tipo BI-AII (Algret *et al.*, 2014). Igualmente, las comparaciones realizadas con la única subunidad del complejo IFT que ha sido cristalizada y las predicciones anteriores de elementos de estructura secundaria nos lleva a plantear que los complejos IFT poseen subunidades que adoptan combinaciones de dominios del tipo BI-AI, o sólo AIII al igual que los complejos COPI (Dam *et al.*, 2013b). Como actualmente no contamos con estructuras de los complejos HOPS/CORVET por lo que aún no podemos inferir con seguridad la presencia de una combinación específica de dominios β-propeller/SPAH.

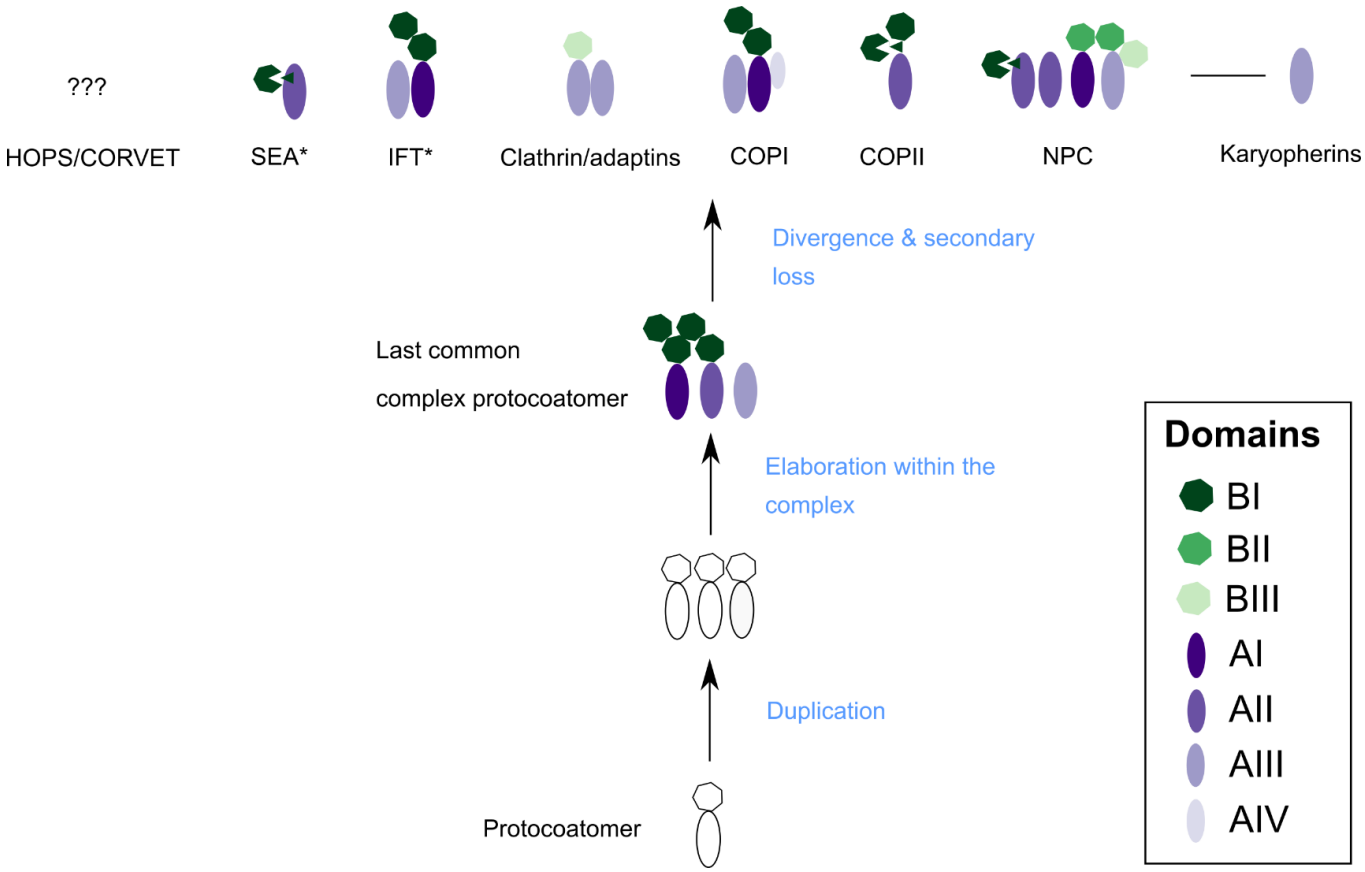


Figura 42. Combinaciones de dominios que dieron origen a las arquitecturas actuales presentes los complejos MC.
Esta figura es una modificación de la idea propuesta inicialmente por Sampathkumar et al., 2013 que considera los tipos diferentes de combinaciones de dominios encontrados en los complejos de cubierta de membrana según las comparaciones estructurales. Los heptágonos representan los dominios β-propeller de las proteínas de cubierta de membrana mientras que las elipses representan sus dominios SPAH. Finalmente, los asteriscos en algunas etiquetas señalan que probablemente el complejo señalado tenga esa posible configuración, pero en el caso de los complejos HOPS/CORVET aún es incierto.

En este trabajo planteamos un nuevo escenario parsimonioso de los posibles pasos evolutivos presentes entre las proteínas de cubierta de membrana (Figura 43). En el escenario que plateamos señalamos que el protocoatomer ancestral se duplico inicialmente donde las subunidades resultantes eran combinaciones de los dominios del tipo BI-AI, luego una de estas subunidades sufrió una duplicación de su dominio β-propeller cuya duplicación aún está presente en algunas subunidades de los complejos COPI y IFT. Como resultado de estas duplicaciones, se formó un complejo ancestral a partir del cual derivaron respectivamente los complejos del tipo pre-COPI y pre-COPII. El complejo pre-COPI probablemente este compuesto por un coatomer que poseía un dominio β-propeller del tipo BI seguido de un dominio SPAH del tipo AI, asociado otro coatomer donde el dominio BI ha divergido al tipo BIII seguido de un dominio del tipo AIII que posiblemente divergió de un tipo AI. El coatomer BIII-AIII probablemente perdió su dominio β-propeller y su dominio SPAH se convirtió en un dominio adaptador. Este dominio adaptador es posiblemente el ancestro común que conecta cercanamente los complejos AP con los complejos adaptadores de COPI, y posee una relación estructural distante con las carioferinas y las nucleoporinas del tipo *adaptin-like.* Por el otro, tenemos un complejo ancestral del tipo pre-COPII que probablemente perdió el coatomer asociado que tenía una función del tipo adaptador siendo reemplazado por otro tipo de proteínas que adopto ese rol (el sub-complejo Sec23-Sec24). Además, el segundo dominio β-propeller de la subunidad pre-COPII probablemente se escindió dando origen a la proteína Sec13, mientras que su dominio SPAH se curvo interactuando sobre sí mismo dando lugar a la aparición del *fold* ACE1. Luego los complejos derivados de estos complejos ancestrales se combinaron entre sí dando origen de forma tardía al poro

nuclear, donde podemos encontrar subunidades que presentan tipos de dominios que están también presentes en los complejos COPI, COPII, y clatrina/adaptinas. Finalmente, la presencia de algunos elementos comunes entre los complejos COPI y SEA sugiere que estos derivaron probablemente de un ancestro en común. Mientras que las comparaciones de HMM descritas en la literatura y las comparaciones realizadas con MOMA2 sugieren que los complejos COPI y IFT probablemente divergieron de un ancestro en común del tipo pre-COPI.

Sin embargo, las mezclas de combinaciones de los dominios presentes en los complejos analizados pueden apuntar a que hubo posiblemente la reintegración de complejos individuales en varios pasos o transiciones dando origen a las arquitecturas que se observan actualmente. Por ejemplo, en el *scaffold* del poro nuclear existe una variedad de proteínas MC con combinaciones diferentes de dominios β-propeller/α-solenoide en lugar de los pocos tipos que presentan algunas subunidades de los complejos COPI, COPII o clatrina/adaptinas. Aunque las comparaciones estructurales y el análisis realizado con el modelo de Maddison y Slatkin nos dan pistas de que posiblemente hubo un escenario plausible que dio lugar a la aparición de las proteínas de cubierta de membrana, igualmente es necesario realizar un estudio a profundidad para poder descartar el escenario alternativo planteado por Sampathkumar y colaboradores en 2013. Posiblemente la cristalización de los dominios β-propeller y SPAH de las subunidades de otros complejos asociados recientemente a las proteínas de cubierta de membrana, además de la obtención de estructuras con una mejor resolución, nos permitirán evaluar en profundidad estos escenarios planteados y extender las relaciones estructurales descritas hasta ahora.

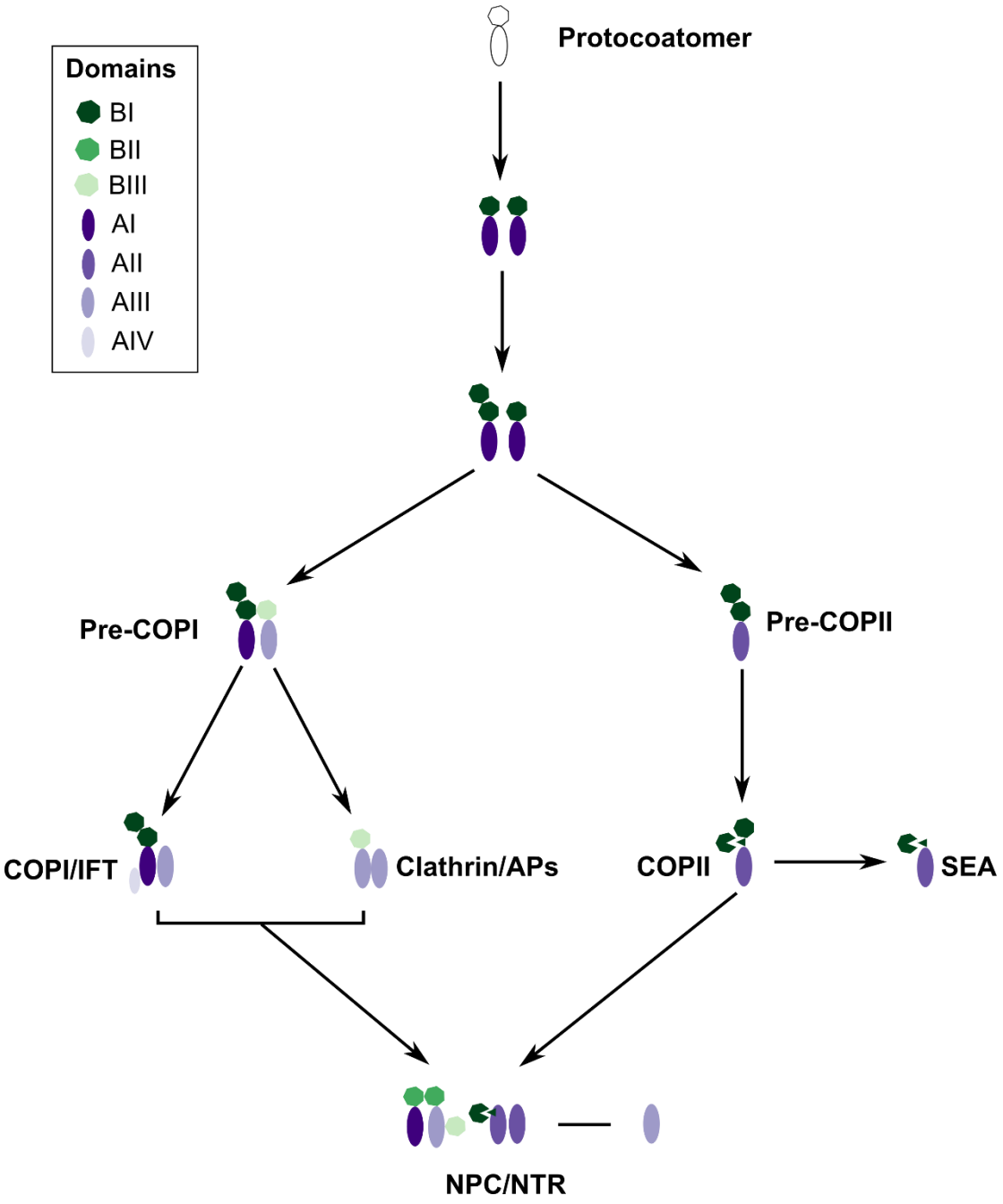


Figura 43. Posibles pasos evolutivos que dieron origen a las proteínas de cubierta de membrana.
Este esquema da a conocer un escenario parsimonioso que sugiere los posibles pasos evolutivos que dieron origen a los complejos actuales a partir del protocoatomer. Los heptágonos representan los dominios β-propeller de las proteínas de cubierta de membrana mientras que las elipses representan sus dominios SPAH.

3.4. CÓMO USAR MOMA2 PARA COMPARAR PROTEÍNAS

3.4.1. Modo de instalación

El repositorio de MOMA2 se puede ejecutar en todos los sistemas operativos que sean compatibles con el programa Docker. Este programa es un software de código abierto que permite encapsular varios programas en una especie de contenedor virtual que puede ejecutarse en diferentes entornos operativos incluyendo Windows, Linux y Mac (https://www.docker.com/). Este contenedor en vez de crear una máquina virtual para ejecutar las aplicaciones solamente cuenta con los procesos mínimos necesarios para ejecutar las aplicaciones encapsuladas. Para instalarlo en una computadora, es primordial revisar los requisitos mínimos necesarios para ejecutar esta aplicación los cuales se encuentran disponibles en https://docs.docker.com/get-docker/.

Después de haber instalado Docker en la computadora, se debe realizar un "*pull*" de la imagen de MOMA2. Para realizar este paso, es necesario abrir la terminal de Docker para escribir el siguiente comando:

```
$ docker pull fggutierrez2018/moma2:latest
```

Este comando permite descargar la imagen del repositorio de MOMA2 que se encuentra disponible la página web de "hub docker" (https://hub.docker.com/). Luego, es necesario crear una nueva carpeta para unir este directorio con el directorio presente en el contenedor de MOMA2 (cuya ruta es "/home/momatools/data/"). En este directorio se pondrán las estructuras que deseamos superponer con MOMA2. Pero, para montar el directorio de trabajo se debe ingresar el siguiente comando en una terminal (sin cerrarla para poder ejecutar posteriormente los scripts del repositorio, Figura 44):

```
$docker run -it -v
/path/of/working/folder/data:/home/momatools/data/
fggutierrez2018/moma2
```

Al ejecutar este comando, nos encontraremos dentro del entorno virtual de Docker que nos permitirá después ejecutar los scripts del repositorio de MOMA2, los cuales se encuentran en la ruta "/home/momatools/scripts".

Finalmente, se requiere crear dos carpetas dentro del directorio de trabajo que se llamen "input" y "output", porque de esta manera, los scripts de MOMA2 cargarán como entrada los archivos ".pdb" que se encuentren en la carpeta "input" y retornarán como salida, las comparaciones pareadas con los alineamientos estructurales dentro de la carpeta "output". A continuación, se mostrarán algunos ejemplos de cómo se ejecutan estos scripts en la consola.

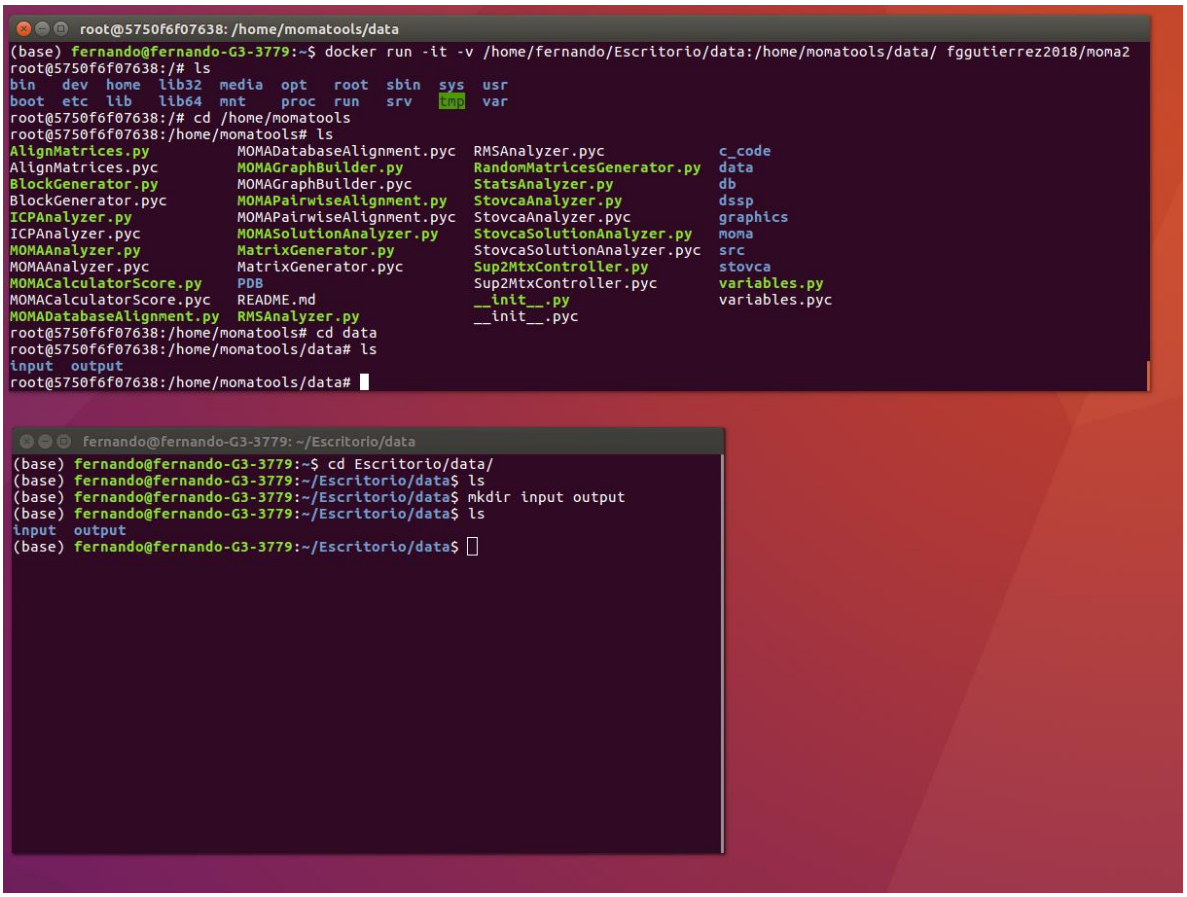


Figura 44. Ejemplo de la ejecución del repositorio de MOMA2 en Ubuntu.
Para correr MOMA2 en Docker, se debe abrir un terminal para ejecutar la imagen del repositorio (terminal en la parte superior). También, se deben crear dos carpetas llamadas “input” y “output” en la carpeta de trabajo, debido a que, en estas carpetas, MOMA2 recibirá de entrada los archivos PDB que deseamos comparar y nos entregará como salida los alineamientos estructurales y/o las búsquedas que hayamos realizado (terminal en la parte inferior).

3.4.2. Comparaciones estructurales entre proteínas

Para realizar comparaciones pareadas de estructuras se debe ejecutar el script *MOMA2_pw.py* que se encuentra disponible en la ruta "/home/momatools/src". Este programa codifica las estructuras de las proteínas en matrices de elementos de estructura secundaria para posteriormente generar un alineamiento a nivel de matrices a partir del cual se calculará una superposición estructural. En los parámetros de entrada, los usuarios podrán decidir si usan el programa DSSP o KAKSI, o ambos a la vez para asignar los elementos de estructura secundaria del par a analizar para construir las matrices. Mientras que, de manera interna, este script comparará varias veces las matrices generadas probando diferentes esquemas de puntaje para obtener un ranking de los mejores alineamientos según sus B_{scores} (Figura 45). Posteriormente, el script calculará por defecto las superposiciones flexibles para los mejores cinco alineamientos de matrices seleccionando finalmente aquel que reporta el mayor número de pares de residuos equivalentes (Figura 45). El mejor alineamiento estructural posteriormente se guarda en un directorio dentro de la ruta "./output/pairwise_alignments".

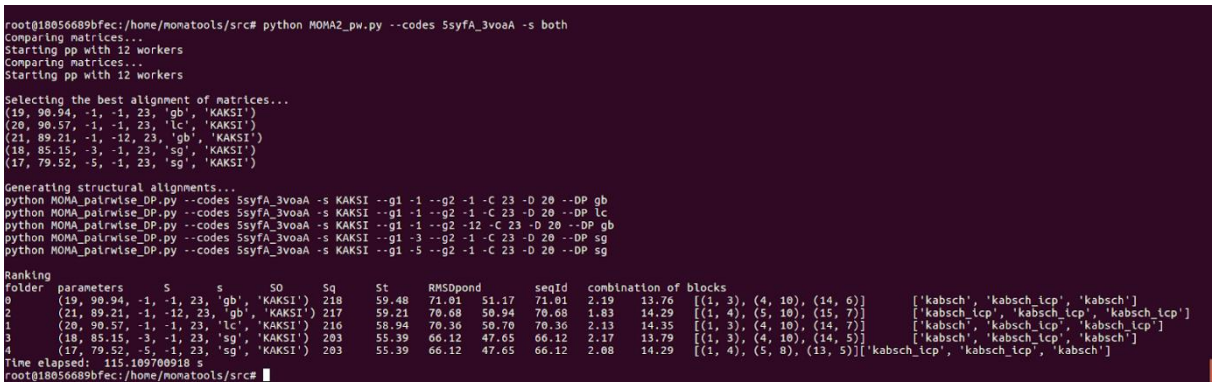


Figura 45. Salida que entrega el script MOMA2_pw.py al comparar un par de estructuras. Captura de pantalla de la consola virtual de Docker en donde se ejecuta el script MOMA2_pw.py, donde a modo de ejemplo, se comparó la proteína Tubulina B (PDB: 5syf, cadena A) contra FtsZ (PDB: 3voa, cadena A). Por defecto, este programa encuentra los primeros cinco mejores alineamientos a nivel de matriz para calcular sus superposiciones estructurales. En este caso, el mejor alineamiento reportó 218 pares de residuos equivalentes con $RMSD_{pond}$ igual a 2.19 Å y con un 13.76% de identidad de secuencia.

Este directorio cuenta a la vez con varias subcarpetas que contienen los siguientes archivos (Figura 46):

- La sesión de PyMOL que muestra el mejor alineamiento estructural se encuentra disponible en la carpeta “best_combination” y tiene el nombre “best_combination.p1m”.
- Los alineamientos estructurales locales calculados entre los sub-fragmentos equivalentes se encuentran en las carpetas “blocks” y “blocks_icp”.
- La imagen del alineamiento a nivel de matriz calculado entre las estructuras comparadas se indica en el archivo “blocks_selected.png”.
- Los resultados de las comparaciones estructurales generados a partir del análisis de STOVCA se describen en el archivo “stovca_analysis.log”.

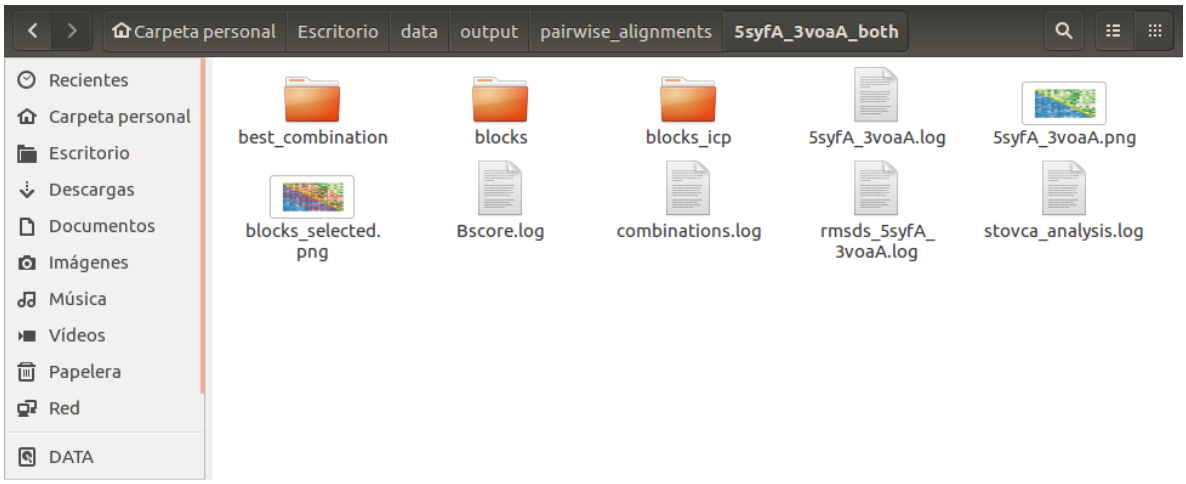


Figura 46. Estructura de la salida que entrega el script MOMA2_pw.py al comparar un par de estructuras.
Siguiendo con el ejemplo anterior, esta es la salida que entrega MOMA2_pw.py cuando compara la proteína Tubulina B (PDB: 5syf, cadena A) contra FtsZ (PDB: 3voa, cadena A).

Otros scripts adicionales como *get_ali.py* y *generate_p1m.py*, que están presentes en la ruta "/home/momatools/src" del container, nos permiten calcular un alineamiento de secuencia de acuerdo con los residuos equivalentes encontrados en la superposición estructural, además de visualizar esta superposición con el programa PyMOL. El script *get_ali.py* permite extraer el alineamiento a nivel de secuencia de los sub-fragmentos encontrados equivalentes mediante el siguiente comando:

```
$ python get_ali.py [nombre de la carpeta solución]
```

El nombre de la carpeta solución se refiere a la ruta en donde se guardó el directorio que entregó el script *MOMA2_pw.py*. El script *get_ali.py* entrega como salida un archivo llamado "blocks_ali.txt" que se almacena en la carpeta solución y que muestra los pares de residuos equivalentes (eq) presentes en los sub-fragmentos alineados. Este archivo de salida señala por

medio de asteriscos los pares de residuos idénticos y con los símbolos de suma aquellos pares que son similares estructuralmente según su composición química (Figura 47).

Figura 47. Formato del archivo de salida que entrega el script get_ali.py.
Este archivo de texto muestra los alineamientos locales a nivel de secuencia de los sub-fragmentos superpuestos según la mejor combinación encontrada. Cada sub-fragmento es indicado por un código de letras y números que señalan a qué archivo PDB y cadena pertenecen, seguido de dos números separados por guiones bajos que señalan las posiciones de inicio y el largo de cada bloque. Finalmente, las letras q y t presentes en las etiquetas de los bloques indican que estos pertenecen a las estructuras *query* o *target*, respectivamente.

En cambio, el script *generate_p1m.py* se ejecuta de la siguiente forma:

```
$ python generate_p1m.py [nombre de la carpeta solución]
```

Este script crea un archivo llamado “best_alignment.p1m” en la carpeta “best_combination” para visualizar con el programa PyMOL, el mejor alineamiento estructural obtenido con MOMA2 (Figura 48). Este archivo contiene los alineamientos locales de los sub-fragmentos equivalentes que son representados en una solución global (Figura 48A), mostrando además la

estructura de la proteína *query* sin fragmentar que se mantiene estática a medida que se rota y traslada la estructura *target* según los *matches* locales de los sub-fragmentos (Figura 48B). Esto último ofrece una enorme ventaja para estudiar los cambios conformacionales y los movimientos de cuerpo rígido en las proteínas flexibles. Por otro lado, los residuos equivalentes señalados en las superposiciones estructurales son coloreados con el esquema descrito por Sippl para el programa TOPMATCH (Sippl and Wiederstein, 2008), donde los residuos no alineados de las proteínas *query* y *target* son indicados con los colores azul y verde, mientras que sus residuos equivalentes son destacados en rojo y naranjo, respectivamente (Figura 48).

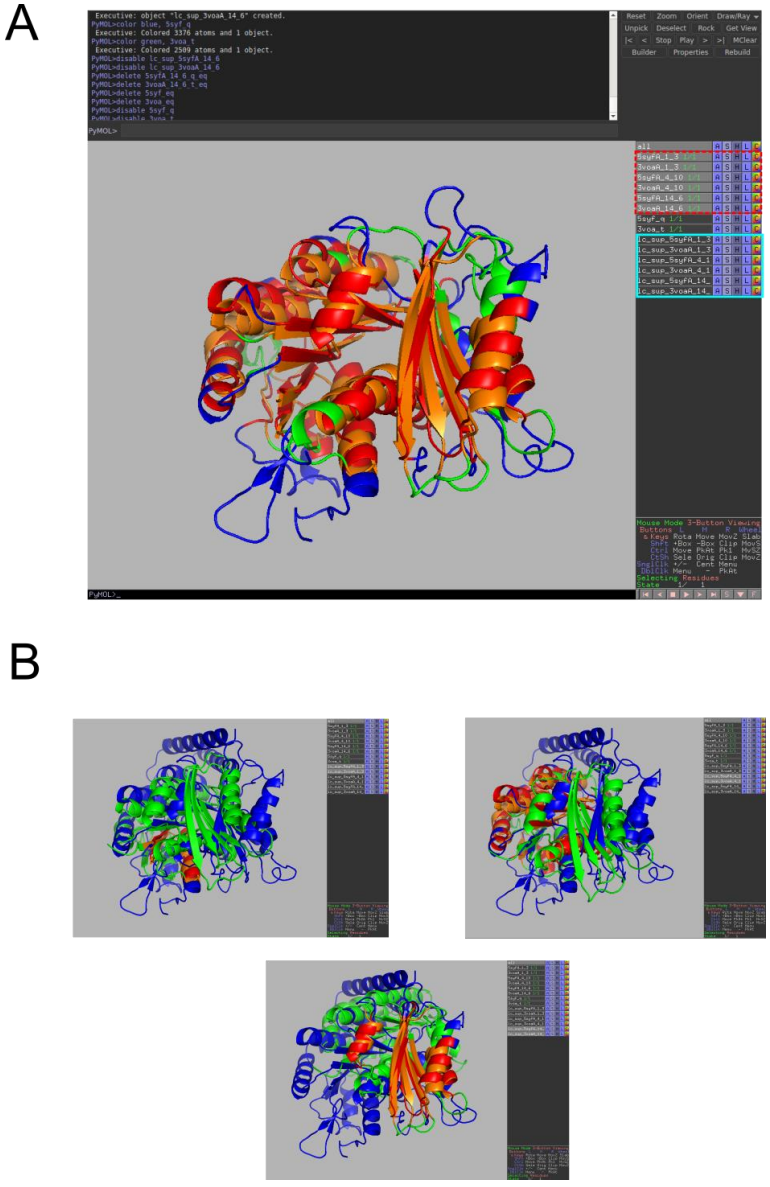


Figura 48. Ejemplo de la sesión de PyMOL generada con el script generate_p1m.py.
En la sesión de PyMOL se muestra la solución global conformada por los alineamientos locales de los sub-fragmentos según la mejor combinación encontrada (A), incluyendo también, las tres superposiciones generadas de ambas cadenas a partir de los *matches* locales de sus sub-fragmentos equivalentes (B). En el panel derecho de PyMOL se señalan mediante el recuadro rojo con líneas segmentadas los sub-fragmentos alineados que fueron seleccionados de la matriz de diferencia, mientras que las superposiciones alternativas de las cadenas se indican en el recuadro de color celeste. Los residuos no alineados son coloreados en azul y verde según la estructura a la que pertenecen (*query* y *target*), y los residuos estructuralmente equivalentes con resaltados en rojo y naranjo, respectivamente.

3.4.3. Búsquedas contra la base de datos

El script *MOMA2_db.py* permite realizar comparaciones estructurales contra una base de datos de matrices, la cual fue creada a partir de 27,500 cadenas de estructuras no redundantes según las listas de PISCES (Wang and Dunbrack, 2005). Para correr búsquedas contra la base de datos de matrices, primero debe estar disponible la estructura de la proteína *query* en la carpeta "/home/momatools/data/input/" para posteriormente ejecutar este script con los siguientes parámetros:

```
$ python MOMA2_db.py --code [código PDB + id de la cadena]
--cpu=[número de hilos] -s [método de asignación de SSE]
```

Este script realiza en paralelo varias búsquedas en contra la base de datos, probando a la vez distintas combinaciones de parámetros que han sido calibradas por defecto. El usuario puede escoger que listas de parámetros usar al ejecutar este programa o establecer su propio criterio de penalización de los *gaps*. De las búsquedas realizadas contra la base de datos como salida se obtienen varios archivos comprimidos que contienen los resultados de las búsquedas realizadas que se almacenan automáticamente en la carpeta "../data/output/comparisons". Luego mediante el uso de otro script llamado *generate_ranking.py*, se analizan estos archivos para extraer los alineamientos significativos de las comparaciones realizadas contra la base de datos. Este script calcula un *p-value* por cada comparación de matrices según la posición de su valor de B_{score} en una distribución de puntajes obtenida a partir de comparaciones al azar entre pares de proteínas no relacionadas. Para ejecutar este script se debe escribir el siguiente comando:

```
$ python generate_ranking.py --pval [cut-off] [código PDB +
id de la cadena]
```

Como salida, se reporta un archivo de texto que contiene un ranking de los *hits* encontrados que reportan un *p-value* menor al umbral de corte definido por el usuario, ordenados de mayor a menor según sus valores de B_{score} (Figura 49). Cada *hit* incluye una descripción con el nombre de la proteína cristalizada, la especie y el reino a la cual pertenece incluyendo los parámetros que dieron lugar a su Δsubmatriz. Estos parámetros pueden ser usados para reconstruir las superposiciones estructurales de los *hits* seleccionados por el usuario utilizando posteriormente el script *MOMA2_pairwise_DP.py*.

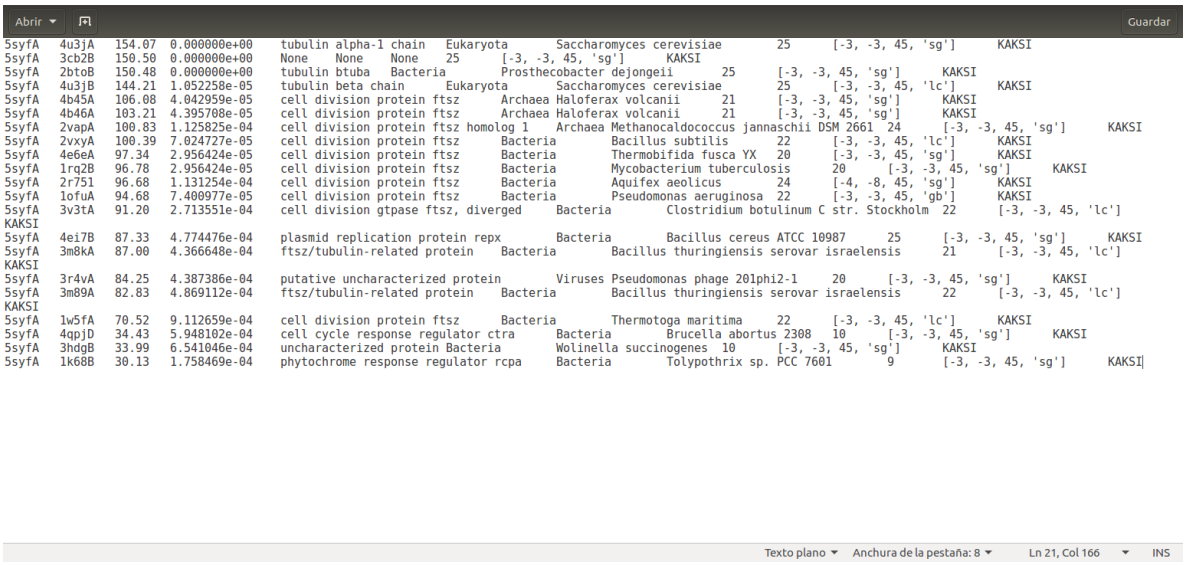


Figura 49. Ejemplo de la salida que entrega el script generate_ranking.py.
En este ejemplo se empleó como estructura *query* la cadena de la tubulina α (PDB 5syf, cadena A) que se comparó contra una base de datos curada de la PDB (PDB90). Usando un umbral de p-value igual a 0.001, el archivo que entrega el programa muestra un ranking de los hits obtenidos y ordenados según sus puntajes de B_{score}.

3.5. PLANES A FUTURO

Después de la defensa del proyecto de tesis, se pretende publicar dos artículos derivados de este trabajo. El primer artículo consistirá en promocionar la nueva herramienta generada para la comparación flexible de estructuras de proteínas, donde MOMA2 estará disponible de forma gratuita como un repositorio de Docker que podrá ser usado en cualquier sistema operativo. Más adelante, se planteará la posibilidad de crear un servidor web que podrá permitir otros usuarios (que no son afines al uso de la terminal para ejecutar comandos) realizar búsquedas contra las bases de datos matrices que serán creadas a partir de diversos conjuntos curados de estructuras no redundantes. El segundo artículo consistirá en publicar los resultados obtenidos de las comparaciones estructurales realizadas con MOMA2 entre las proteínas de cubierta de membrana donde pudimos corroborar, extender y descubrir nuevas relaciones estructurales, abarcando la nueva clasificación de las proteínas MC según sus dominios, los nuevos árboles filogenéticos creados a partir de las comparaciones de sus dominios y los nuevos pasos evolutivos sugeridos usando un modelo parsimonioso. Además, a medida que el número de estructuras cristalizadas aumente describiendo otros complejos asociados a las proteínas de cubierta de membrana, podremos usar nuevamente esta herramienta para extender las relaciones encontradas hasta ahora obteniendo una visión más amplia de la evolución de estas proteínas. Finalmente, MOMA2 posee el potencial para explorar las relaciones evolutivas distantes de otros complejos asociados al sistema de endomembranas. Por ejemplo, esta herramienta se ha empleado recientemente para evaluar las similitudes estructurales entre las proteínas CATCH con la finalidad de estudiar su homología y evolución modular (Santana-Molina *et al.*, 2021). Esto nos demuestra el enorme potencial que posee MOMA2 para descubrir relaciones distantes

entre proteínas divergentes, para estudiar la flexibilidad y los movimientos modulares de los dominios, y finalmente para clasificar las proteínas en diferentes familias o tipos de plegamientos estructurales.

CONCLUSIONES

i. MOMA (MOrphing & MAtching) es un algoritmo de comparación estructural basado en el arreglo espacial de los elementos de estructura secundaria que es eficiente para reconocer intuitivamente los pares de sub-fragmentos equivalentes cuando se usa para comparar pares de proteínas distantemente relacionadas.

ii. La comparación de matrices de elementos de estructura secundaria posee un excelente rendimiento para clasificar grupos de proteínas dentro de una misma familia o tipo de plegamiento, siendo también el método más rápido de los métodos analizados en este trabajo para realizar búsquedas contra las bases de datos de estructuras.

iii. Sin embargo, MOMA no cuenta con un método eficiente y robusto que permita determinar la mejor combinación de sub-fragmentos para obtener un alineamiento compuesto, un método que permita distinguir de forma precisa los pares de SSEs equivalentes, y representar la mejor solución en una superposición estructural.

iv. MOMA2 es una nueva versión del algoritmo de MOMA, que posee una característica dual, donde primero determina un alineamiento de matrices de SSE para evaluar la similitud estructural de un par de proteínas, para luego identificar la mejor combinación de pares sub-fragmentos relacionados con el fin de generar una superposición estructural flexible.

v. MOMA2 posee un excelente rendimiento para discriminar pares de proteínas multidominio relacionadas de aquellas que no están relacionadas a partir de sus

superposiciones flexibles a diferencia de otros programas de alineamiento estructural flexible.

vi. Sin embargo, el tiempo de cálculo de las superposiciones estructurales de MOMA2 es lento comparado con otros programas de alineamiento flexible, sacrificando poder de cómputo para determinar superposiciones precisas.

vii. MOMA2 puede ser empleado para explorar relaciones estructurales remotas entre proteínas relacionadas que presentan una gran variación estructural, para clasificar familias de proteínas, y para estudiar el desplazamiento de los sub-fragmentos rígidos en proteínas flexibles.

viii. Las comparaciones estructurales realizadas con MOMA2 entre las proteínas de cubierta de membrana nos han permitido crear un nuevo esquema de clasificación basándose en la similitud estructural de sus dominios.

ix. De acuerdo con esta nueva clasificación estructural, los dominios β-propeller se dividen en tres tipos denominados BI, BII y BIII, en cambio, los dominios SPAH se dividen en cuatro tipos denominados respectivamente como AI, AII, AIII y AIV.

x. Las superposiciones estructurales y los análisis de sus motivos de secuencia sugieren que los dominios β-propeller del tipo BI, BII, y BIII posiblemente derivaron de un dominio β-propeller ancestral parecido al tipo BI.

xi. Las comparaciones realizadas con MOMA2, nos ha permitido confirmar, extender y determinar nuevas relaciones estructurales entre las proteínas de cubierta de membrana.

    a. Hemos corroborado las relaciones estructurales entre Sec31 y Sec16 con las Nups descritas en la literatura como COPII-like, entre COPβ’ con Nup133

descrita como COPI-like, la cercanía estructural de las nucleoporinas Nup192 y Nup188 con las carioferinas, o la fuerte similitud estructural entre las adaptinas con los complejos adaptadores de COPI.

b. Hemos extendido las relaciones estructurales entre los dominios β-propeller de algunas subunidades que forman parte del poro nuclear y que están clasificadas en tres tipos diferentes (BI, BII y BIII).

c. Hemos extendido las relaciones encontradas entre las nucleoporinas Nup188, Nup192 y las carioferinas con las adaptinas de Clatrina y COPI, sustentando la idea de que entre las proteínas MC existe un tercer tipo de proteínas MC denominado *adaptin-like*.

d. Hemos encontrado una relación distante entre el *fold* ACE1 de Nic96 con el extremo C-terminal de Nup120, cuyo dominio SPAH se ha especializado formando un nuevo fold en su dominio β-propeller.

e. Las comparaciones estructurales nos han permitido descubrir nuevas relaciones estructurales distantes entre algunas subunidades del poro nuclear con los complejos de cubierta de vesículas, por ejemplo, entre Clatrina y Nup170, o entre COPβ' y Nup133.

f. Nos ha permitido encontrar relaciones estructurales entre las subunidades del tipo *cage* con el tipo *adaptador* entre los complejos de vesícula, por ejemplo, entre Clatrina y AP2a.

xii. Hemos desarrollado el primer modelo filogenético que explora, de forma cuantitativa y paralela, las relaciones estructurales presentes en las proteínas de cubierta de membrana a través de sus dominios β-propeller y SPAH.

xiii. El dendrograma de los dominios β-propeller nos sugiere que todos estos dominios surgieron de un β-propeller ancestral posiblemente similar al tipo BI.

xiv. El dendrograma de los dominios SPAH muestra al menos tres tipos de dominios que presentan fuertes relaciones intragrupales a diferencia de sus relaciones intergrupales, sugiriendo que estos dominios evolucionaron al menos a partir de tres tipos diferentes de dominios SPAH.

xv. De acuerdo con este modelo, hemos podido inferir 9 combinaciones diferentes de los dominios presentes en las proteínas de cubierta de membrana, incluyendo combinaciones de los dominios β-propeller/SPAH, sólo β-propeller y sólo SPAH.

xvi. El *scaffold* poro nuclear está compuesto por una amalgama de combinaciones de dominios cuyos tipos se encuentran también presentes en otros complejos de cubierta de membrana, sugiriendo que en su formación fue producto de la iteración inicial de dos o tres tipos de arquitecturas diferentes que dieron origen a un poro temprano, y luego por eventos de duplicación y divergencia éste fue evolucionado al complejo moderno que hoy observamos.

xvii. El *scaffold* del poro nuclear presenta además combinaciones únicas de dominios como BII-AI, BIII y BII-AIII, así como la ausencia de subunidades del tipo BI-AI y AI.

xviii. Gracias a estas comparaciones podemos inferir la presencia de algunas combinaciones de dominios presentes en otros complejos de cubierta de membrana que actualmente no se han cristalizado completamente, como la presencia de combinaciones BI-AI y AIII en el complejo IFT, o la presencia de las combinaciones BI-AII y BI en los complejos SEA.

xix. De acuerdo con la reconstrucción evolutiva de los dominios observados en los dendrogramas generados, se sugiere que estos complejos evolucionaron posiblemente a partir de un escenario parsimonioso que plantea que al menos tres o dos familias de proteínas derivadas del protocoatomer evolucionaron a partir de un solo complejo ancestral, luego por eventos de duplicación y divergencia, se formaron los complejos actuales.

xx. Mediante estas comparaciones nos es posible especular que el complejo ancestral que dio origen a las proteínas de cubierta de membrana estaba compuesto posiblemente por subunidades que presentaban combinaciones de dominios del tipo BI-AI. Después los tipos BI-AII y AIII fueron apareciendo eventualmente a medida que las subunidades del tipo BI-AI fueron divergiendo.

xxi. Posiblemente con la cristalización de las subunidades de los otros complejos asociados en estos últimos años a las proteínas de cubierta de membrana y la obtención de un mayor número de estructuras con una mejor resolución y para una mayor cantidad de especies diferentes, nos permitirá evaluar en profundidad los escenarios planteados para estas proteínas y extender las relaciones estructurales descritas hasta ahora.

REFERENCIAS

Abe,A. *et al.* (2004) Complex structures of Thermoactinomyces vulgaris R-47 alpha-amylase 1 with malto-oligosaccharides demonstrate the role of domain N acting as a starch-binding domain. *J. Mol. Biol.*, **335**, 811–822.

Algret,R. *et al.* (2014) Molecular Architecture and Function of the SEA Complex, a Modulator of the TORC1 Pathway. *Molecular & Cellular Proteomics*, **13**, 2855–2870.

Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends in Biochemical Sciences*, **23**, 444–447.

Andersen,K.R. *et al.* (2013) Scaffold nucleoporins Nup188 and Nup192 share structural and functional properties with nuclear transport receptors. *Elife*, **2**, e00745.

Beck,M. *et al.* (2018) From the resolution revolution to evolution: structural insights into the evolutionary relationships between vesicle coats and the nuclear pore. *Current Opinion in Structural Biology*, **52**, 32–40.

Beisel,H.G. *et al.* (1999) Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab Tachypleus tridentatus. *EMBO J.*, **18**, 2313–2322.

Besl,P. J. and McKay,N.D. (1992) A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **14**, 239–256.

Besl,Paul J. and McKay,N.D. (1992) Method for registration of 3-D shapes. In, *Sensor Fusion IV: Control Paradigms and Data Structures*. International Society for Optics and Photonics, pp. 586–607.

Bilokapic,S. and Schwartz,T.U. (2012) Molecular basis for Nup37 and ELY5/ELYS recruitment to the nuclear pore complex. *PNAS*, **109**, 15241–15246.

Bliven,S. and Prlić,A. (2012) Circular Permutation in Proteins. *PLOS Computational Biology*, **8**, e1002445.

Boehm,M. and Bonifacino,J.S. (2001) Adaptins. *MBoC*, **12**, 2907–2920.

Bonifacino,J.S. and Glick,B.S. (2004) The Mechanisms of Vesicle Budding and Fusion. *Cell*, **116**, 153–166.

Bourne,P.E. and Shindyalov,I.N. (2005) Structure Comparison and Alignment. In, *Structural Bioinformatics*. John Wiley & Sons, Ltd, pp. 321–337.

Brohawn,S.G. *et al.* (2008a) Structural Evidence for Common Ancestry of the Nuclear Pore Complex and Vesicle Coats. *Science*, **322**, 1369–1373.

Brohawn,S.G. *et al.* (2008b) Structural evidence for common ancestry of the nuclear pore complex and vesicle coats. *Science*, **322**, 1369–1373.

Brohawn,S.G. and Schwartz,T.U. (2009) Molecular architecture of the Nup84–Nup145C–Sec13 edge element in the nuclear pore complex lattice. *Nature Structural & Molecular Biology*, **16**, 1173–1177.

Burkowski,F.J. (2008) Structural bioinformatics: an algorithmic approach CRC Press.

Carpentier,M. *et al.* (2005) YAKUSA: A fast structural database scanning method. *Proteins: Structure, Function, and Bioinformatics*, **61**, 137–151.

Cheng,H. *et al.* (2008) MALISAM: a database of structurally analogous motifs in proteins. *Nucleic Acids Res*, **36**, D211–D217.

Dacks,J.B. *et al.* (2009) Evolution of specificity in the eukaryotic endomembrane system. *The International Journal of Biochemistry & Cell Biology*, **41**, 330–340.

Dacks,J.B. and Doolittle,W.F. (2002) Novel syntaxin gene sequences from Giardia, Trypanosoma and algae: implications for the ancient evolution of the eukaryotic endomembrane system. *Journal of Cell Science*, **115**, 1635–1642.

Dacks,J.B. and Field,M.C. (2007) Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. *Journal of Cell Science*, **120**, 2977–2985.
Dacks,J.B. and Robinson,M.S. (2017) Outerwear through the ages: evolutionary cell biology of vesicle coats. *Current Opinion in Cell Biology*, **47**, 108–116.
Dam,T.J.P. van *et al.* (2013a) Evolution of modular intraflagellar transport from a coatomer-like progenitor. *PNAS*, **110**, 6943–6948.
Dam,T.J.P. van *et al.* (2013b) Evolution of modular intraflagellar transport from a coatomer-like progenitor. *PNAS*, **110**, 6943–6948.
Darwin,C. (2004) On the origin of species, 1859 Routledge.
Debler,E.W. *et al.* (2008) A Fence-like Coat for the Nuclear Pore Membrane. *Molecular Cell*, **32**, 815–826.
Devos,D. *et al.* (2004) Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture. *PLOS Biology*, **2**, e380.
Devos,D.P. (2013) Structural Aspects of MC Proteins of PVC Superphylum Members. In, Fuerst,J.A. (ed), *Planctomycetes: Cell Structure, Origins and Biology*. Humana Press, Totowa, NJ, pp. 77–87.
Dokudovskaya,S. *et al.* (2011) A Conserved Coatomer-related Complex Containing Sec13 and Seh1 Dynamically Associates With the Vacuole in Saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, **10**, M110.006478.
Eddy,S. and Wheeler,T. (2007) HMMER-biosequence analysis using profile hidden Markov models. *URL http://hmmer. janelia. org*.
Elias,M. *et al.* (2012) Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. *J Cell Sci*, **125**, 2500–2508.
van den Ent,F. *et al.* (2001) Prokaryotic origin of the actin cytoskeleton. *Nature*, **413**, 39.
Faini,M. *et al.* (2013) Vesicle coats: structure, function, and general principles of assembly. *Trends in Cell Biology*, **23**, 279–288.
Felsenstein,J. (1993) PHYLIP (phylogeny inference package), version 3.5 c Joseph Felsenstein.
Field,M.C. *et al.* (2011) On a bender—BARs, ESCRTs, COPs, and finally getting your coat. *The Journal of Cell Biology*, **193**, 963–972.
Field,M.C. and Dacks,J.B. (2009) First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. *Current Opinion in Cell Biology*, **21**, 4–13.
Field,M.C. and Rout,M.P. (2019) Pore timing: the evolutionary origins of the nucleus and nuclear pore complex. *F1000Res*, **8**.
Fox,N.K. *et al.* (2013) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research*, **42**, D304–D309.
Freund,Y. and Schapire,R.E. (1999) Large Margin Classification Using the Perceptron Algorithm. *Machine Learning*, **37**, 277–296.
Gibrat,J.-F. *et al.* (1996) Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, **6**, 377–385.
González-Sánchez,J.C. *et al.* (2015) A multi-functional tubulovesicular network as the ancestral eukaryotic endomembrane system. *Biology (Basel)*, **4**, 264–281.
Guerler,A. and Knapp,E.-W. (2008) Novel protein folds and their nonsequential structural analogs. *Protein Sci*, **17**, 1374–1382.
Gutiérrez,F.I. *et al.* (2016) Efficient and automated large-scale detection of structural relationships in proteins with a flexible aligner. *BMC bioinformatics*, **17**, 20.
Hasegawa,H. and Holm,L. (2009) Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, **19**, 341–348.

Hayama,R. *et al.* (2017) The nuclear pore complex core scaffold and permeability barrier: variations of a common theme. *Current Opinion in Cell Biology*, **46**, 110–118.
Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, **20**, 478–480.
Hsia,K.-C. and Hoelz,A. (2010) Crystal structure of α-COP in complex with ϵ-COP provides insight into the architecture of the COPI vesicular coat. *Proc Natl Acad Sci U S A*, **107**, 11271–11276.
Hunter,J.D. (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, **9**, 90–95.
Jékely,G. (2007) Origin of Eukaryotic Endomembranes: A Critical Evaluation of Different Model Scenarios. In, *Eukaryotic Membranes and Cytoskeleton: Origins and Evolution*, Advances in Experimental Medicine and Biology. Springer New York, New York, NY, pp. 38–51.
Jékely,G. (2003) Small GTPases and the evolution of the eukaryotic cell. *BioEssays*, **25**, 1129–1138.
Jékely,G. and Arendt,D. (2006) Evolution of intraflagellar transport from coated vesicles and autogenous origin of the eukaryotic cilium. *BioEssays*, **28**, 191–198.
Johnson,L.S. *et al.* (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.
Johnson,M.S. *et al.* (1990) [42] Phylogenetic relationships from three-dimensional protein structures.
Junier,T. and Zdobnov,E.M. (2010) The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics*, **26**, 1669–1670.
Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, **32**, 922–923.
Kabsch,W. and Sander,C. (1983) DSSP: definition of secondary structure of proteins given a set of 3D coordinates. *Biopolymers*, **22**, 2577–2637.
Kamat,A.P. and Lesk,A.M. (2007) Contact patterns between helices and strands of sheet define protein folding patterns. *Proteins: Structure, Function, and Bioinformatics*, **66**, 869–876.
Kee,H.L. *et al.* (2012) A size-exclusion permeability barrier and nucleoporins characterize a ciliary pore complex that regulates transport into cilia. *Nature Cell Biology*, **14**, 431–437.
Kim,S.J. *et al.* (2018) Integrative structure and functional anatomy of a nuclear pore complex. *Nature*, **555**, 475–482.
Konagurthu,A.S. *et al.* (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.
Konagurthu,A.S. *et al.* (2008) Structural search and retrieval using a tableau representation of protein folding patterns. *Bioinformatics*, **24**, 645–651.
Koonin,E.V. and Galperin,M. (2013) Sequence — Evolution — Function: Computational Approaches in Comparative Genomics Springer Science & Business Media.
Krzywinski,M.I. *et al.* (2009) Circos: An information aesthetic for comparative genomics. *Genome Res.*
Lee,C. and Goldberg,J. (2010) Structure of Coatomer Cage Proteins and the Relationship among COPI, COPII, and Clathrin Vesicle Coats. *Cell*, **142**, 123–132.
Leksa,N.C. *et al.* (2009) The structure of the scaffold nucleoporin Nup120 reveals a new and unexpected domain architecture. *Structure*, **17**, 1082–1091.
Lesk,A.M. (1995) Systematic representation of protein folding patterns. *Journal of molecular graphics*, **13**, 159–164.
Lesk,A.M. and Chothia,C. (1980) How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *Journal of Molecular Biology*, **136**, 225–270.
Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

Maddison,W.P. and Slatkin,M. (1991) Null Models for the Number of Evolutionary Steps in a Character on a Phylogenetic Tree. *Evolution*, **45**, 1184–1197.
Malod-Dognin,N. and Pržulj,N. (2014) GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity. *Bioinformatics*, **30**, 1259–1265.
Martin,J. *et al.* (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC structural biology*, **5**, 17.
Menke,M. *et al.* (2008) Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology*, **4**, e10.
Mikami,B. *et al.* (2006) Crystal Structure of Pullulanase: Evidence for Parallel Binding of Oligosaccharides in the Active Site. *Journal of Molecular Biology*, **359**, 690–707.
Mingot,J.-M. *et al.* (2001) Importin 13: a novel mediator of nuclear import and export. *EMBO J*, **20**, 3685–3694.
Mosca,R. *et al.* (2008) Alignment of protein structures in the presence of domain motions. *BMC Bioinformatics*, **9**, 352.
Murzin,A.G. (1998) How far divergent evolution goes in proteins. *Current Opinion in Structural Biology*, **8**, 380–387.
Newman,M.E.J. and Girvan,M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, 026113.
OKA,M. and YONEDA,Y. (2018) Importin α: functions as a nuclear transport factor and beyond. *Proc Jpn Acad Ser B Phys Biol Sci*, **94**, 259–274.
Orengo,C.A. and Taylor,W.R. (1996) [36] SSAP: Sequential structure alignment program for protein structure comparison. In, *Methods in Enzymology*, Computer Methods for Macromolecular Sequence Analysis. Academic Press, pp. 617–635.
Patthy,L. (2009) Protein evolution John Wiley & Sons.
Pedregosa,F. *et al.* (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research*, **12**, 2825–2830.
Risler,J.L. *et al.* (1988) Amino acid substitutions in structurally related proteins a pattern recognition approach: Determination of a new and efficient scoring matrix. *Journal of Molecular Biology*, **204**, 1019–1029.
Robert,X. and Gouet,P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res*, **42**, W320–W324.
Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng Des Sel*, **12**, 85–94.
Rout,M.P. and Field,M.C. (2017) The Evolution of Organellar Coat Complexes and Organization of the Eukaryotic Cell. *Annu. Rev. Biochem.*, **86**, 637–657.
Salem,S. *et al.* (2010) FlexSnap: Flexible Non-sequential Protein Structure Alignment. *Algorithms Mol Biol*, **5**, 12.
Sampathkumar,P. *et al.* (2013) Structure, Dynamics, Evolution, and Function of a Major Scaffold Component in the Nuclear Pore Complex. *Structure*, **21**, 560–571.
Santana-Molina,C. *et al.* (2021) Homology and Modular Evolution of CATCHR at the Origin of the Eukaryotic Endomembrane System. *Genome Biology and Evolution*, **13**.
Schlacht,A. and Dacks,J.B. (2015) Unexpected Ancient Paralogs and an Evolutionary Model for the COPII Coat Complex. *Genome Biol Evol*, **7**, 1098–1109.
Schrödinger, LLC (2015) The PyMOL Molecular Graphics System, Version 1.8.
Seo,H.-S. *et al.* (2013) Structure and nucleic acid binding activity of the nucleoporin Nup157. *Proc Natl Acad Sci U S A*, **110**, 16450–16455.
Shannon,P. *et al.* (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, **13**, 2498–2504.

Shatsky,M. *et al.* (2004) FlexProt: Alignment of Flexible Protein Structures Without a Predefinition of Hinge Regions. *Journal of Computational Biology*, **11**, 83–106.
Sippl,M.J. and Wiederstein,M. (2008) A note on difficult structure alignment problems. *Bioinformatics*, **24**, 426–427.
Slater,A.W. *et al.* (2012) Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. *Bioinformatics*, **29**, 47–53.
Söding,J. *et al.* (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*, **33**, W244–W248.
Spang,A. *et al.* (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, **521**, 173–179.
Stivala,A. *et al.* (2009) Tableau-based protein substructure search using quadratic programming. *BMC Bioinformatics*, **10**, 153.
Stuwe,T. *et al.* (2015) Architecture of the nuclear pore complex coat. *Science*, **347**, 1148–1152.
Stuwe,T. *et al.* (2014) Evidence for an evolutionary relationship between the large adaptor nucleoporin Nup192 and karyopherins. *PNAS*, **111**, 2530–2535.
Terashi,G. and Takeda-Shitaka,M. (2015) CAB-Align: A Flexible Protein Structure Alignment Method Based on the Residue-Residue Contact Area. *PLOS ONE*, **10**, e0141440.
Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–4680.
Vergara,I.A. *et al.* (2008) StAR: a simple tool for the statistical comparison of ROC curves. *BMC bioinformatics*, **9**, 265.
Walker,G. *et al.* (2011) Eukaryotic systematics: a user's guide for cell biologists and parasitologists. *Parasitology*, **138**, 1638–1663.
Wang,G. and Dunbrack,R.L.,Jr (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic acids research*, **33**, W94–W98.
Waterhouse,A.M. *et al.* (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
Whittle,J.R.R. and Schwartz,T.U. (2009) Architectural nucleoporins Nup157/170 and Nup133 are structurally related and descend from a second ancestral element. *J. Biol. Chem.*, **284**, 28442–28452.
Wiederstein,M. *et al.* (2014) Structure-based characterization of multiprotein complexes. *Structure*, **22**, 1063–1070.
wwPDB consortium (2018) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, **47**, D520–D528.
Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**, ii246–ii255.
Zaremba-Niedzwiedzka,K. *et al.* (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, **541**, 353–358.
Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, **33**, 2302–2309.
Zhang,Z.H. *et al.* (2010) Reduced representation of protein structure: implications on efficiency and scope of detection of structural similarity. *BMC Bioinformatics*, **11**, 155.
Zmasek,C.M. and Godzik,A. (2012) This Déjà Vu Feeling—Analysis of Multidomain Protein Evolution in Eukaryotic Genomes. *PLOS Computational Biology*, **8**, e1002701.

ANEXOS

6.1. Supplementary material: "MOMA2: Improving structural similarity detection beyond the twilight zone"

6.1.1. Problems found with MOMA

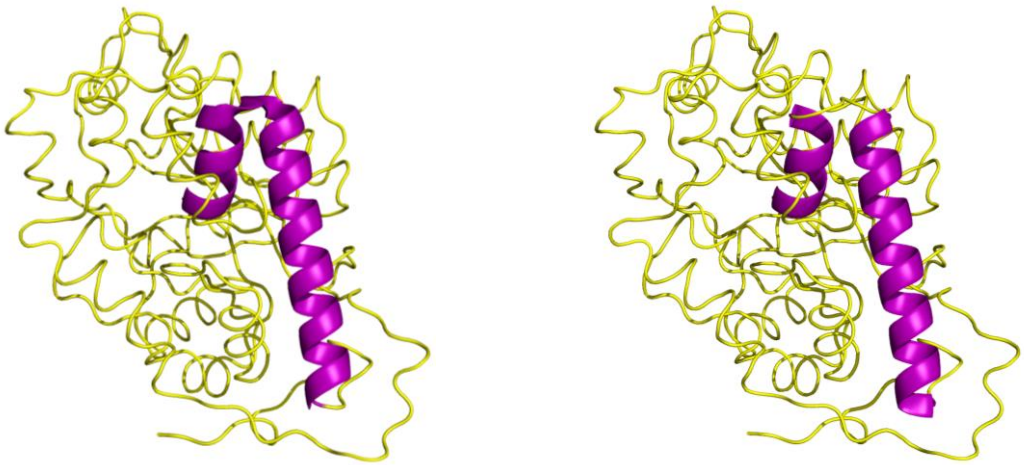
The MOMA algorithm encodes protein structures as matrices of secondary structure elements (SSEs) to compare each pair to evaluate the structural similarity between a pair of structures (Gutierrez et *al.*, 2016). Also, we demonstrated that this algorithm is a practical and fast to assess structural similarities between distantly related proteins, but we have also identified six primary sources of error that we addressed to solve them, and which are described as follows:

1. SSE assignment
2. Internal angles
3. Local vs. global alignment
4. Similarity score based on a matrix alignment
5. Extraction of significant local matches
6. Structural superpositions

We addressed these issues as follows:

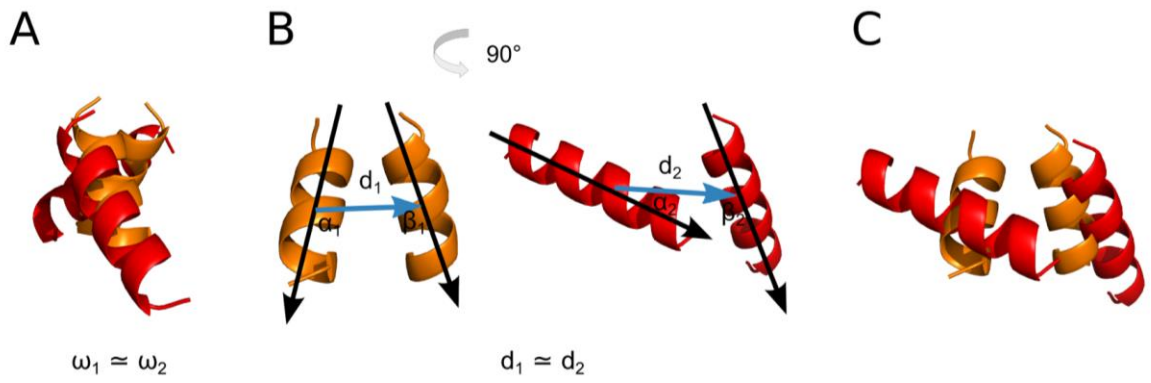
1. **SSE assignment.** The assignment of the SSEs has a substantial impact on the similarity reported from the alignment of the matrices. Instead, the calculation of the angular and distance differences between the secondary structural elements is mainly affected by their correct assignment. In a large group of analyzed cases, DSSP works well to assign their precise start and end positions of the secondary structure elements. However, in

some cases, DSSP cannot recognize torsion points or breaks presented on α-helices or β-strands, treating them as a single unit instead of assigning two individual elements (Figure S1). To counterweight this issue, MOMA2 now includes the KAKSI program as an alternative to identify the secondary structure elements in the protein structures. Comparisons performed with KAKSI showed that assigns less curved or kinked α-helices than other methods as DSSP or Stride, proving better assignments in these difficult examples (Figure S1).



Supplementary Figure 1. Example of disagreement between DSSP and KAKSI.
Secondary structure element assignation in L(+)-mandelate dehydrogenase from Pseudomonas putida (PDB code 1p4c chain A). DSSP assigns one single helix from 308 to 340. Instead, KAKSI assigns two helices from 308 to 315 and 320 to 341. The divergent assignments are drawn in cartoon representation and highlighted in purple. Images generated with PyMOL.

2. **Internal angles.** We also observed the incorrect identification of some equivalent SSE pairs in the alignments reported initially with MOMA. Although some SSEs pairs showed similar dihedral angles and distances, these elements were incorrectly superposed at the residue level because their internal angles have changed considerably (Fig. S2). In MOMA2, we define a new variable called "internal angle" as the angle formed by one of the vectors that represents a secondary structure element (a black arrow) which intersects with the vector that joins their centroids (blue arrow). These angles we named respectively as α and β angles.



Supplementary Figure 2. Consideration of internal angles to select equivalent SSE pairs. Each pair of secondary structure elements is represented using vectors (black arrows). MOMA calculates the dihedral angles (ω_1, ω_2) between each SSE vector with their respective centroid vector which is indicated by blue arrows (d_1, d_2). In this case, both pairs of secondary structure elements (colored by red and orange) have similar dihedral angles and distances (A and B). However, these pair are not well superposed because their internal angles (α_1, α_2, β_1, β_2) present large angular variations. As a result, with the new scoring function, these secondary structure element pairs are not considered equivalent (C).

The internal angles are evaluated in the scoring function to align the rows of SSE matrices. The scoring function can be determined with the following formula (Equation 1):

$$f(E_iE_j, E_kE_l, \omega_{ij}, \omega_{kl}, d_{ij}, d_{kl}, \alpha_{ij}, \alpha_{kl}, \beta_{ij}, \beta_{kl}) = \quad (1)$$

$$-C, E_iE_j \neq E_kE_l \quad \text{i.}$$

$$-C, \Delta\omega > 2C \quad \text{ii.}$$

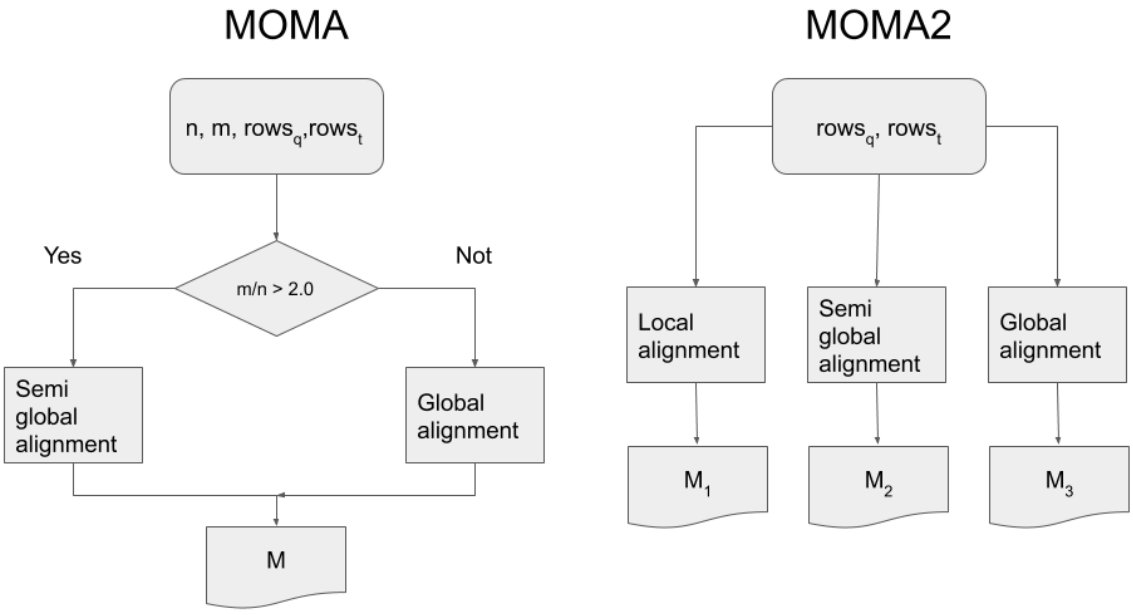
$$0, d_{ij} > D \; or \; d_{kl} > D \quad \text{iii.}$$

$$-C, \Delta\alpha > 22.5º \; or \; \Delta\beta > 22.5º \quad \text{iv.}$$

$$C - \Delta\omega, otherwise \quad \text{v.}$$

where E_iE_j and E_kE_l are two pairs of SSEs derived from the query and target structures respectively, whose dihedral and internal angles are defined as ω_{ij}, ω_{kl}, α_{ij}, α_{kl}, β_{ij}, and β_{kl}, and their distances are indicated as d_{ij} and d_{kl}, respectively. The selection of the equivalent SSE pairs is subjected to several restrictions; for example, the constraint "i" ensures that the SSE pairs must be composed of the same type of secondary elements. Moreover, constraints "ii" and "iii" ensure that their dihedral angles should be lower than the C constant, and their distances should be lower to the D constant. The constraint "iv" is included in the MOMA2 scoring function to identify the equivalent SSE pairs whose angular difference of their inner angles must be lower than 22.5°. This new restriction ensures to recognize locally SSE pairs with lower structural differences (Figure S2). If the value Δω overpasses these constraints, the option "v" calculates a score based on the angular variation of the dihedral angles.

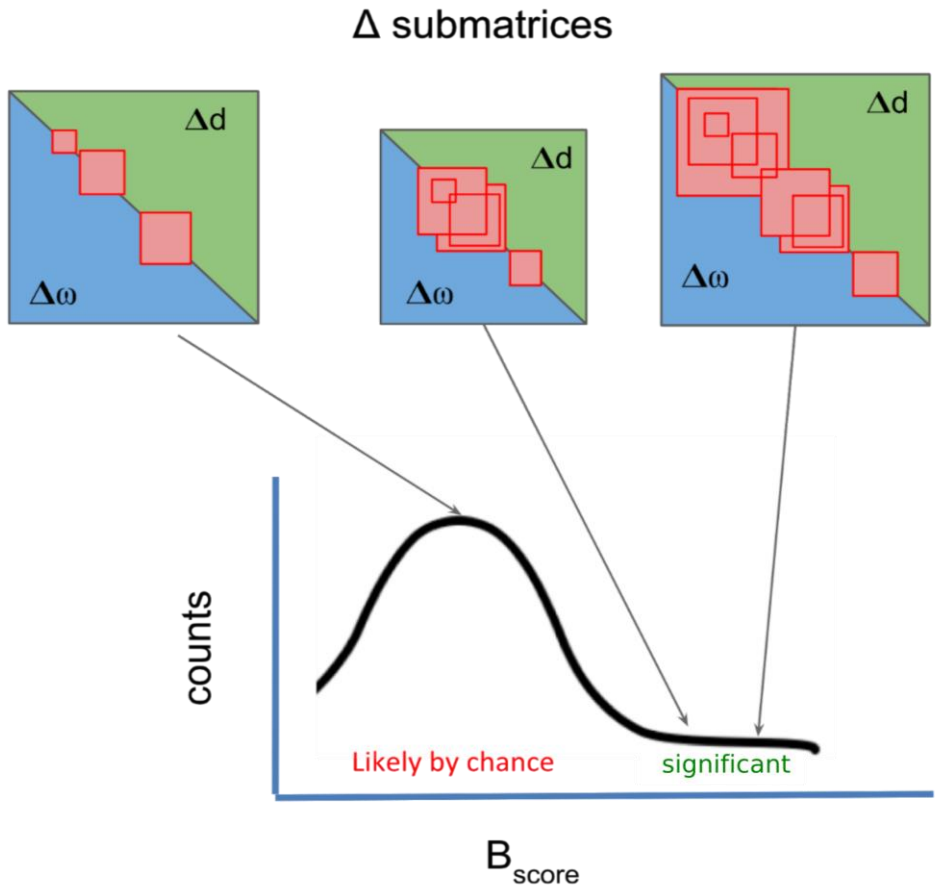
3. **Local vs. global alignment.** Initially, the MOMA algorithm was designed to select a global or semi-global alignment to align the rows of the input matrices based on their sizes. In some cases, forcing a global comparison when the analyzed protein structures only share a small similar region in common. New modifications were implemented in the first step of the dynamic programming to get accurate matrix alignments (Figure S3). While MOMA2 aligns the rows of the matrices using three dynamic programming algorithms generating three scoring matrices as an output instead of a one scoring matrix as MOMA. Then, these scoring matrices are used to perform three local alignments of the sequences derived from the secondary structure elements through a second step of dynamic programming. Considering these local alignments, the initial matrices are resized to construct 2D difference matrices called Δsubmatrices because these matrices contain the angular and distance differences calculated between the secondary structure pairs that were evaluated with MOMA2.



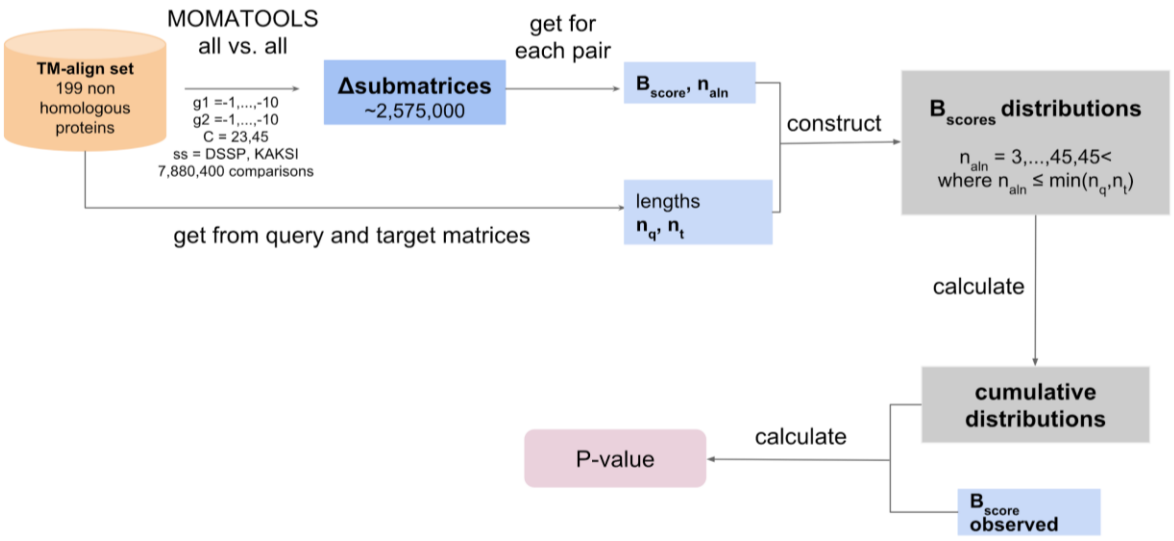
Supplementary Figure 3. Modification implemented in the alignment of SSE matrices. MOMA selects according to the sizes of the input matrices (n and m secondary structure elements) the algorithm which is used to align the matrices (Left). MOMA2 aligns the rows of the matrices ($rows_q$, $rows_t$) using three dynamic programming algorithms (local, semi-global and global alignment), generating three matrices' scores instead of one (Right). The optimal scores reported by NxN alignments of the rows are used to create the scoring matrices (M, M_1, M_2, and M_3).

4. **Similarity score based on a matrix alignment.** To evaluate the similarity reported in the Δsubmatrices, we create a new score (called "S score") based on the angular differences instead of considering their distance values (Gutierrez et *al.*, 2016). We noted in several matrix alignments that some blocks had SSE pairs with small angular differences and large distance differences. As a result of this, the sub-fragments obtained from these blocks were not well superposed with MOMA. To solve this problem, MOMA2 implements a new similarity score called B_{score} that reports a structural similarity based on the number of aligned sub-fragments overlapped with lower angular

or distance differences (Figure S4). Finally, we implemented a p-value to evaluate the significance of these scores based on the distribution of B_{scores} calculated from comparisons performed between non-homologous pairs of proteins (Figure S5).



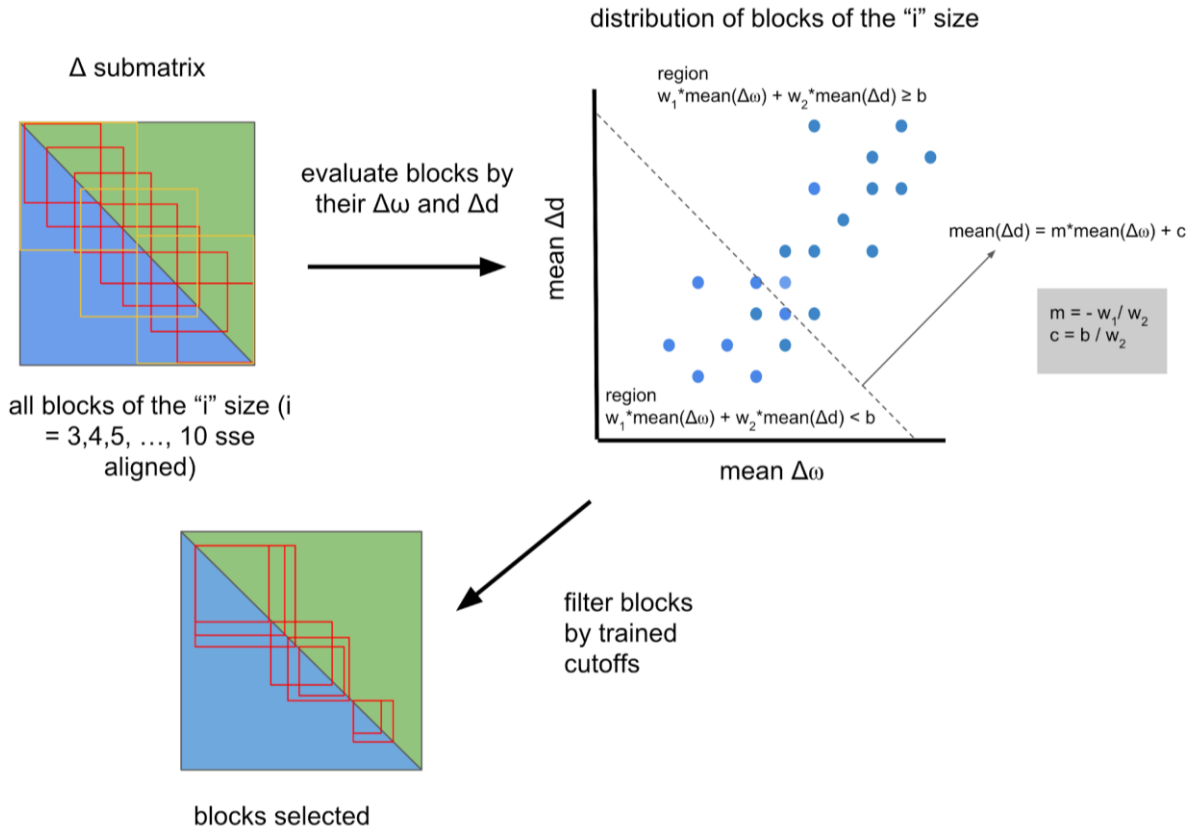
Supplementary Figure 4. A new similarity score derived from matrix alignments.
The figure shows that a region in a Δsubmatrix covered by many overlapping blocks indicates a high B_{score} suggesting that proteins aligned are structurally very similar. However, a region covered by a few small blocks which are not overlapped indicates a low B_{score}, suggesting that the alignment reported is not significant and probably the proteins compared with MOMA2 are not related.



Supplementary Figure 5. Flowchart to obtain a P-value from a matrix alignment. A set of 199 nonhomologous proteins from TM-align set (https://zhanglab.ccmb.med.umich.edu/TM-align/benchmark/) was used to determine the p-values from query searches performed against a database of matrices. From ~2,575,000 matrix alignments calculated with the new version of MOMA (using different combinations of parameters), several distributions of their B score values were generated. The p-value is the probability that the B_{score} calculated for a NxN Δsubmatrix is greater than or equal to the B_{score} observed from the distributions of the matrix alignments calculated from unrelated pairs of proteins whose dimensions are lower or equal to n. If this probability is less than a significance level of $\alpha = 0.05$, this alignment is considered significant.

5. **Extraction of significant local matches.** MOMA automatically detects partial structural matches and tolerates a substantial structural variation of the domains. Unlike MOMA, rigid aligners cannot detect these changes reporting lower structural similarity from these examples (Gutierrez et al., 2016). However, the MOMA algorithm is unable to quantify the significance of these local matches found. This problem emerged from alignments generated with MOMA that reported a combination of aligned fragments composed by few pairs aligned of secondary structure elements, which probably adopted a similar spatial orientation by chance. We implemented the Perceptron method to train a list of thresholds to select significant local matches from a matrix alignment (Table S1

and Figure S6). A perceptron is a linear classifier that uses a decision boundary to determine the significant blocks from a Δsubmatrix. Each block is represented by a 2-dimensional vector considering the values of their angular and distance mean differences; this simplification can represent each block of the Δsubmatrix as a point on a 2D-graph (Figure S6). Using the weights (w1, w2) and bias (b) values trained from the perceptron algorithm from a trial dataset, we determined the slope "m" and the constant "c" of the decision boundaries which are used to select the significant blocks in the structures compared with MOMA2 (Figure S6 and Table S1).



Supplementary Figure 6. Selection of significant blocks from a matrix alignment using the list of cut-offs trained with the Perceptron.
A list of cut-offs was calculated with the Perceptron algorithm using the angular and distance mean differences obtained from a trained dataset of blocks. MOMA2 selected a list of significant blocks whose mean values are lower than the thresholds trained with the Perceptron method.

Supplementary Table 1. List of cut-offs parameters obtained by the Perceptron analysis.

block length	w1	w2	b	slope	constant	P	N
3	-0.041	-0.452	1.92	-0.09	4.25	20670	59592
4	-0.051	-0.753	3.49	-0.068	4.632	18206	49622
5	-0.073	-0.916	4.942	-0.079	5.398	15801	40140
6	-0.077	-0.92	5.384	-0.083	5.852	13506	31338
7	-0.052	-0.831	4.686	-0.063	5.636	11362	23405
8	-0.066	-0.916	5.316	-0.072	5.801	9410	16569
9	-0.072	-0.765	4.852	-0.094	6.343	7647	11065
10	-0.1	-1.081	6.508	-0.092	6.02	6135	7011

The table shows the weights given to the average values of the angular and distance differences (w_1, w_2) and the bias values (b) reported by each block. The size of the selected blocks ranges from 3 to 10 pairs of aligned SSEs. The list of cut-offs was trained using a dataset composed of 102,737 blocks obtained from alignments of domains classified into the same superfamily (P) and 238,742 blocks obtained from alignments of domains classified into different folds (N), according to the SCOP classification. Using this information, we calculated the decision boundary lines defined by the equation of the line (slope, c constant where the y-axis cuts the line) to select the significant blocks.

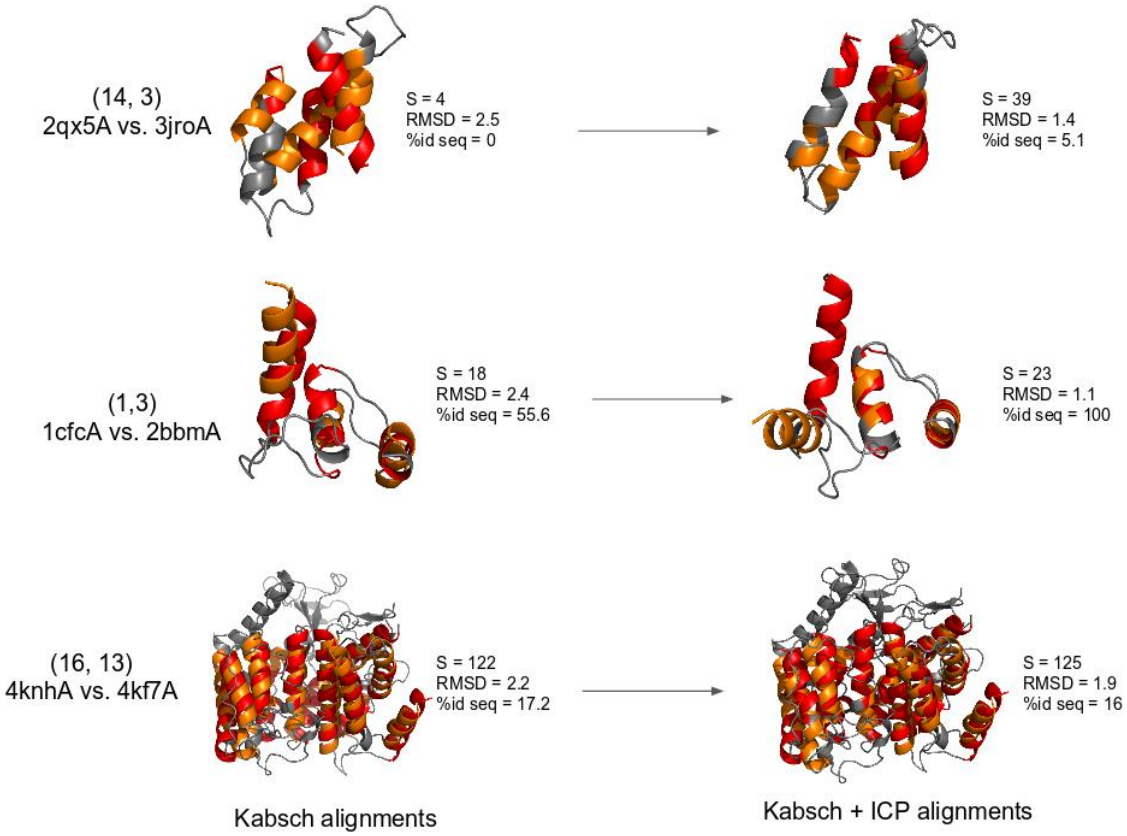
To evaluate the composition of the selected blocks, MOMA2 analyses which blocks have a considerable structural divergence counting the number of positions were their SSE pairs have high angular or distance differences, although these blocks were selected with the first group of filters. To obtain effective superpositions of the selected sub-fragment pairs, it is necessary to limit the number of local SSE pairs with a high structural divergence, tolerating to a certain extent, the local variation of several positions according to the size of the blocks (Table S2).

Supplementary Table 2. The number of positions tolerated with large angular, or distance differences based on their internal composition of the blocks.

n_{sse}	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
n_a	1	1	2	3	4	6	7	9
n_d	1	1	2	3	4	6	7	9

This table shows the number of positions tolerated with an angular difference greater than 45° (n_a) or a distance difference greater than 5 Å (n_d) for blocks whose sizes range from 3 to 10 pairs of SSEs.

6. **Structural superpositions.** Initially, MOMA used SSE vectors to superpose a pair of structures, which are not very precise to superpose remote proteins. To solve this issue, we implemented in MOMA2 a modified version of the Iterative Closest Point (ICP) algorithm to improve the superpositions obtained with the Kabsch algorithm to report better superpositions of the aligned sub-fragments (Figure S7). The implementation of the ICP algorithm allowed to identify equivalent residues presented in the aligned sub-fragments, which can be represented in a sequence alignment. Each pair of aligned sub-fragments is now evaluated with the STOVCA program to calculate their structural local similarities and their scores are used to calculate the percentage of relative similarity and structural overlap, the percentage of identical sequence residues and a weighted RMSD value for each combination of aligned sub-fragments (Figure S8).



Supplementary Figure 7. ICP algorithm refines the alignment of the equivalent sub-fragments.

Initially, the sub-fragments obtained from these three blocks are superposed with the Kabsch algorithm. Then, the obtained local superpositions are refined with the ICP algorithm. The aligned sub-fragments correspond to the structures of Nic96 (PDB code 2qx5, chain A), Nup145C (PDB code 3jro, chain A), Calmodulins (PDB codes 1cfc, chain A and 2bbm, chain A), Nup188 (PDB code 4kf7, chain A), and Nup192 (PDB code 4knh, chain A). Bock starts and lengths are indicated above the PDB codes in between parenthesis. The secondary structure elements found equivalent in the query and target structures are represented in red and orange, respectively. Where "S" is the number of aligned positions, "RMSD" is the root-mean-square deviation of atomic positions and "%id seq." is the percentage of identical aligned residues. These images were generated with PyMOL.

$$S = \sum_{i}^{l} eq_i \quad (1)$$

$$Sr = \frac{2\sum_{i}^{l} eq_i}{L_q + L_t} \quad (2)$$

$$SO = \frac{\sum_{i}^{l} eq_i}{min(L_q, L_t)} \quad (3)$$

$$RMSD_{pond} = \frac{\sum_{i}^{l} RMSD_i eq_i}{\sum_{i}^{l} eq_i} \quad (4)$$

$$\%\ id\ seq. = 100\frac{2\sum_{i}^{l} id_i}{L_q + L_t} \quad (5)$$

Supplementary Figure 8. Scores reported by MOMA2 from each superposition generated by a combination of local matches.
Where id_i is the number of identical residues aligned over the eq_i aligned residues from the "i" pair of aligned fragments, respectively; L_q and L_t are the total length of residues of the query and target structures; $RMSD_i$ is the root-mean-square deviation of atomic positions derived from the "i" pair of aligned fragments. These parameters are used to calculate the total number of aligned residues (S), the percentage of relative similarity (Sr), structural overlap (SO), the weighted mean of the RMSD ($RMSD_{pond}$), and the percentage of sequence identity of the query and target in the equivalent regions, according to Sippl 2008.

6.1.2. Comparisons performed with the MALISAM dataset

Supplementary Table 3. Summary of the comparisons of 92 pairs of analogous domains using four flexible aligners.

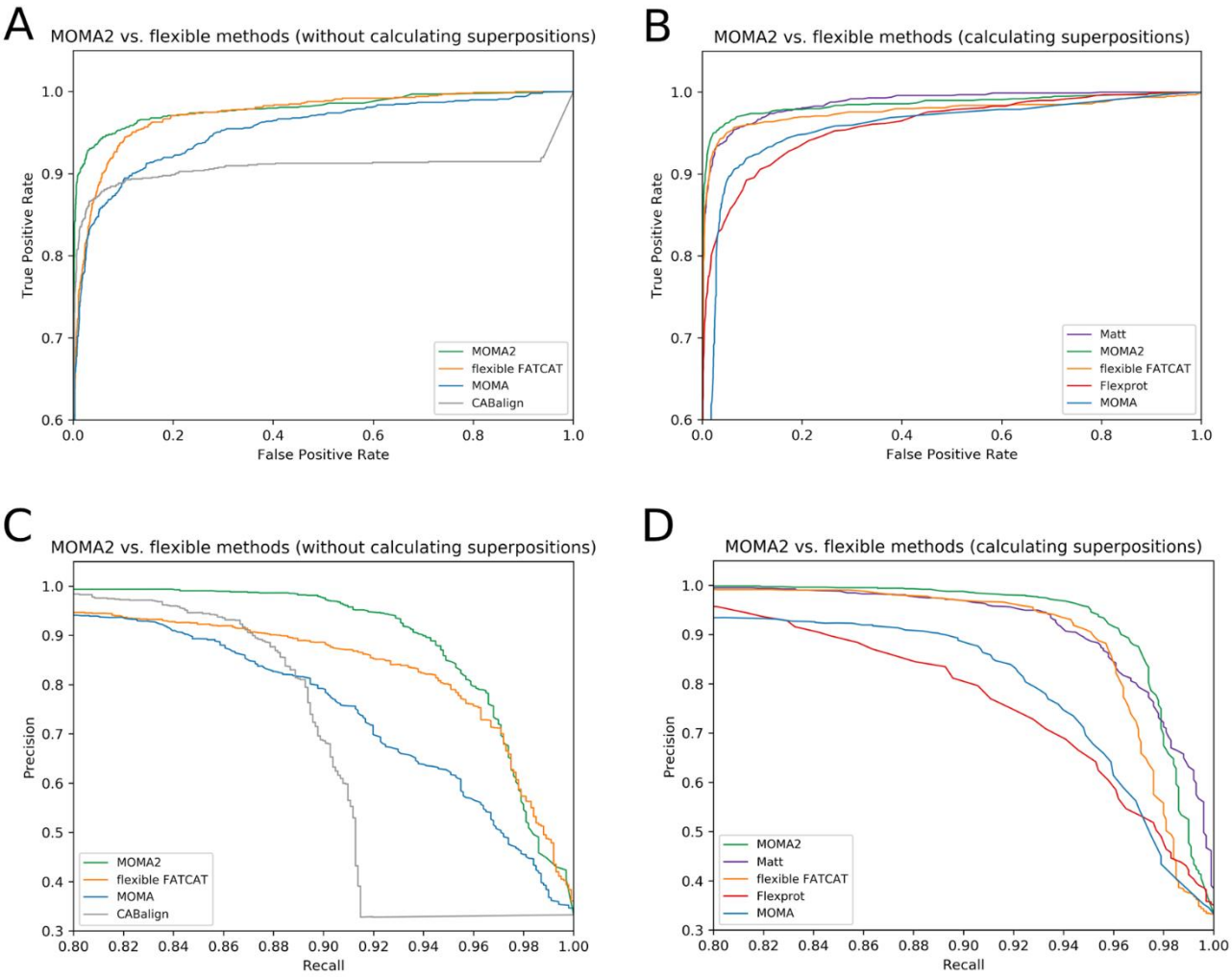
Domain 1	Domain 2	MALISAM	MOMA2	FATCAT	FLEXPROT	MATT	TM	FAST	DALI	TOPMATCH
d1a2za_	d1ghha_	45	40	44	64	52	35	47	49	43
d1b05a_	d1lc0a2	43	20	39	55	33	36	39	36	26
d1b05a_	d1nkia_	42	38	42	48	44	42	42	41	42
d1b0pa2	d1k0da2	29	28	43	56	34	32	33	29	31
d1b62a2	d1i6pa_	50	28	40	57	56	42	42	46	48
d1ce8a6	d1di2a_	37	32	37	47	35	36	37	36	37
d1chma1	d1g62a_	23	25	22	36	35	28	29	18	21
d1cs1a_	d1q8ia1	48	39	48	53	49	47	48	47	48
d1dc1a_	d1nf2a_	43	42	46	60	52	44	43	43	44
d1dc1a_	d1rblm_	34	35	23	56	36	23	33	35	24
d1dmza_	d1hdha_	42	40	39	52	44	41	38	41	40
d1dzka_	d1qd1b2	33	29	42	55	55	20	33	26	31
d1e3pa3	d1gpqa_	54	49	57	71	66	50	48	51	53
d1efdn_	d1r0va3	35	34	35	40	35	38	38	40	30
d1fbxa_	d1o08a_	36	0	35	54	31	38	33	37	16
d1fy2a_	d1qmha1	44	42	43	51	39	40	42	40	42
d1g99a1	d1gqyb3	41	19	41	50	37	40	40	39	41
d1g99a1	d1izna_	55	45	58	69	63	55	55	59	61
d1gc5a_	d1omha_	51	27	46	75	73	47	38	51	53
d1gx3a_	d1epaa_	46	45	53	78	56	47	46	47	12
d1gx3a_	d1pbya5	39	50	48	60	61	44	37	45	27
d1h0hb_	d1hz6b_	39	33	40	54	46	36	44	40	40
d1h6da2	d1kjka_	30	19	34	35	29	28	31	29	24
d1hbnb2	d1aisb2	47	42	54	72	41	43	46	46	47
d1hi9a_	d1rpua_	49	43	54	64	41	46	45	49	21
d1hrda1	d1rp3a3	47	0	57	71	58	35	43	57	48
d1htjf_	d1gnla_	50	42	47	57	48	47	33	45	47
d1htjf_	d1nhua_	38	39	52	64	59	42	34	40	55
d1i9ea_	d1nppa1	32	32	31	48	47	28	32	29	29
d1ihoa_	d1ep7a_	43	41	45	48	51	42	40	45	43
d1is8a_	d1umwa1	42	28	41	57	39	30	38	38	39
d1j22a_	d1m2aa_	47	44	44	53	51	47	39	50	48
d1jdbc2	d1r9ca_	15	18	25	44	29	21	17	26	32
d1jnwa_	d1ew4a_	42	31	40	57	45	38	35	39	43
d1miwa2	d1gyfa_	29	27	31	43	34	32	33	23	24
d1n7va_	d1izna_	47	29	50	55	48	29	42	47	43
d1nf2a_	d1ptma_	42	39	36	48	39	40	37	41	44
d1nkta3	d1ldja1	39	0	42	48	42	42	15	44	43
d1o17d2	d1ir6a_	55	44	58	71	60	55	45	56	55
d1ogqa_	d1fjgc1	31	10	29	50	39	25	29	30	27
d1ogsa2	d1dq3a4	31	30	32	46	36	32	26	26	31
d1ogsa2	d1is1a_	37	36	32	45	36	38	23	38	32
d1p5gx1	d1xppa_	45	27	45	59	55	39	43	47	42
d1piea1	d1nxha_	36	0	55	57	44	37	37	35	38

(Continue Supplementary Table 3)

d1piea1	d1q0qa1	48	40	50	55	47	48	38	50	49
d1qd1b1	d1itwa_	35	31	47	64	54	9	48	42	27
d1qd1b1	d1n62c2	36	32	31	40	37	37	40	37	36
d1qh4a2	d1ekga_	43	28	68	73	49	41	43	42	49
d1qnia1	d1k4za_	33	17	34	42	37	32	34	34	34
d1slca_	d1sq9a_	27	50	41	48	53	31	44	38	20
d1st6a8	d1gvna_	54	45	57	76	58	46	38	51	61
d1tv8a_	d1nbwb_	48	41	46	62	51	43	43	48	44
d1tv8a_	d1t3ea3	61	56	61	72	57	59	57	56	61
d1uoua1	d1fs1b1	30	33	55	58	51	35	45	42	34
d1uyra1	d1vkka_	72	69	64	80	81	69	74	70	68
d1uyrb1	d1a79a1	51	33	53	63	57	41	50	35	46
d1vaoa2	d1no5a_	23	38	37	51	35	36	40	37	36
d1vk3a3	d1w2ia_	37	28	33	50	33	34	37	40	36
d2csta_	d1j4wa1	49	45	49	60	44	48	45	45	50
d2gaw.1	d1nppa1	41	27	28	53	36	23	40	32	25
d2onea1	d1hz6b_	34	29	36	52	44	37	35	38	35
d3pmga1	d1e8ca1	38	36	38	43	37	38	41	40	39
d6mhta_	d1vm0b_	48	43	38	61	47	43	49	46	43
d1a05a_	d1dgsa3	37	32	45	59	25	40	40	34	31
d1a05a_	d1j71a_	41	29	38	43	39	36	41	39	41
d1a05a_	d1rblm_	35	36	34	51	53	35	41	37	19
d1aa7a_	d1b68a_	54	30	54	57	36	42	52	55	50
d1aa7a_	d1qkra_	57	33	56	74	65	41	52	56	46
d1aora2	d1dkza2	38	30	39	58	27	30	31	38	11
d1b6ra1	d1k8kd1	41	21	41	52	35	37	35	41	39
d1b6ra1	d1v97a4	43	23	45	56	51	37	35	42	43
d1eg3a1	d1paqa_	44	0	48	62	44	39	27	46	33
d1gxja_	d1p5dx3	37	34	39	52	47	36	34	34	38
d1h5wa_	d2cbla2	66	45	83	93	79	70	67	73	25
d1izna_	d1keka2	32	2	36	59	27	38	24	31	32
d1knza_	d1rp3a3	44	18	45	69	69	56	36	38	36
d1knza_	d1vola1	57	35	68	69	61	55	56	57	59
d1m9na2	d1j27a_	50	55	48	69	61	50	48	46	49
d1mk5a_	d1efub2	49	53	49	53	64	53	53	54	48
d1okra_	d1v9da_	26	36	53	60	37	29	20	40	25
d1omoa_	d1i7b.1	51	37	56	66	57	56	39	57	49
d1izna_	d1gkza1	38	3	50	61	31	38	34	37	16
d1ps6a_	d1f9za_	36	33	35	48	42	34	31	39	29
d1r2ja2	d1ecmb_	35	0	39	73	66	50	46	37	55
d1rwha2	d1n67a2	34	19	34	65	47	24	24	50	22
d1smta_	d1b7ta4	42	0	63	77	49	40	18	43	26
d1st6a2	d1jqna_	69	39	79	105	84	74	62	80	42
d1lq7a_	d1gqea_	56	0	61	51	58	59	50	61	58
d1oi2a_	d1uw1a_	32	24	28	46	32	17	29	28	26
d1pqsa_	d1i7la2	41	35	37	52	53	37	45	39	39
d1qysa_	d1i7b.1	67	59	58	72	70	55	63	67	67
d1uw1a_	d1q9ua_	35	30	32	53	33	32	13	26	34

The number of aligned positions reported from superpositions generated with MALISAM, MOMA2, FATCAT, FlexProt, MATT, TM, FAST, DALI, and TOPMATCH were calculated with the STOVCA program using a distance cut-off of 3.5 Å to select equivalent Cα atoms. If the number of aligned residues reported is greater than the number of aligned residues registered by MALISAM, these values are highlighted in red.

6.1.3. Benchmark tests



Supplementary Figure 9. The comparison of the performance between MOMA2 and other flexible alignment tools using ROC and Precision-Recall curves.
The performance of the flexible aligners in the identification of related and unrelated proteins was evaluated using a multi-domain benchmark dataset composed of 997 pairs of proteins whose domains almost share the same superfamily (related pairs), and 1998 pairs of proteins whose domain pairs do not share the same fold classification (unrelated pairs). We tested the scores obtained from the MOMA2, MOMA, flexible FATCAT, and CAB-align to classify these proteins without calculating structural superpositions (A and C). Also, we tested the number of

aligned residues calculated with the STOVCA program from superpositions generated with MOMA2, MOMA, flexible FATCAT, FlexProt, and MATT to classify related and unrelated proteins (B and D).

Supplementary Table 4.Performance of the methods in the classification of related and unrelated proteins according to their similarity scores.

Methods	MOMA2	flexible FATCAT	MOMA	CABalign
MOMA2	**MOMA2**	0.0067	0.0268	0.072
flexible FATCAT	0.047	**flexible FATCAT**	0.02	0.0653
MOMA	< 0.001	< 0.001	**MOMA**	0.0453
CABalign	< 0.001	< 0.001	< 0.001	**CABalign**

Significance level (alpha): 0.05
Upper triangle: AUC differences
Bottom triangle: p-values
p-value < 0.05
p-value > 0.05

Supplementary Table 5. Performance of the methods in the classification of related and unrelated proteins according to the number of aligned residues.

Methods	Matt	MOMA2	flexible FATCAT	FlexProt	MOMA
Matt	**Matt**	0.0013	0.0103	0.0280	0.0326
MOMA2	0.6418	**MOMA2**	0.0090	0.0267	0.0313
flexible FATCAT	0.0023	0.0242	**flexible FATCAT**	0.0177	0.0223
FlexProt	< 0.001	< 0.001	< 0.001	**FlexProt**	0.0046
MOMA	< 0.001	< 0.001	< 0.001	0.3950	**MOMA**

Significance level (alpha): 0.05
Upper triangle: AUC differences
Bottom triangle: p-values
p-value < 0.05
p-value > 0.05

6.1.4. Additional examples

Here we present four examples of flexible protein comparisons that illustrate the power of MOMA2 to superpose individually two or more domains in proteins with overall structural distortions. Additionally, we observed some limitations associated with the methodology implemented.

The first example corresponds to the comparison between the tissue factor (PDB code 1a21, chain A), and the growth hormone-binding protein (PDB code 1hwg, chain C) (Figure S10A). According to SCOP classification, both proteins are grouped in the fibronectin type III family, and they are organized into two β-sheet domains connected by an unstructured linker. The rigid alignment obtained by TOPMATCH reports the lowest number of aligned positions as expected. FATCAT introduces one twist to align both domains which result in a partial superposition of the N-terminal domains. Instead, MOMA2 detects a longer structural alignment than FATCAT (eq = 122 and RMSD = 1.7 Å), detecting equivalences among the N-terminal and C-terminal domains introducing two pairs of aligned sub-fragments. However, the superposition generated with Matt reports the highest number of aligned residues of the four tested methods (eq = 143 and RMSD = 1.3 Å), aligning the small helices (composed of three residues) present between both domains. Even though the superposition generated with MOMA2 detects the domains present in these proteins, MOMA2 cannot align these small helices. This is a technical limitation of the MOMA2 algorithm which only aligns pairs of α-helices or β-strands whose sizes are longer or equal than 4 or 3 residues, respectively.

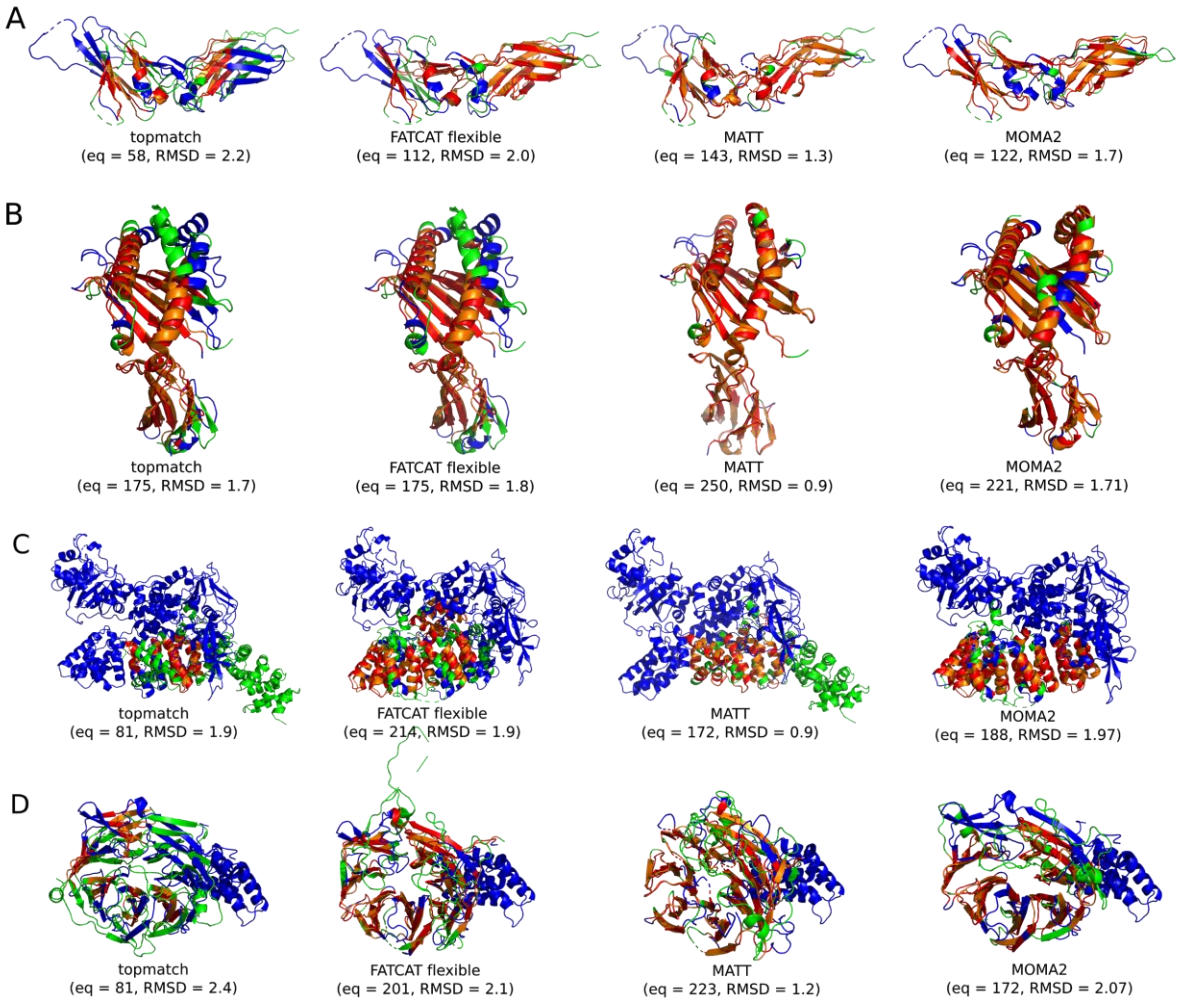
In the second example, the histocompatibility antigen (PDB code 2clr, chain A) was compared to the neonatal FC receptor (PDB code 3fru, chain A) (Figure S10B). According to SCOP classification, both proteins have two structurally similar domains. The first domain

belongs to the family of MHC antigen-recognition of type I which is followed by a second domain that belongs to the family of C1-set domains. On one hand, without introducing any twists, the FATCAT alignment is very similar to the one of TOPMATCH, with slight differences in the residues aligned. On the other hand, the alignment calculated with Matt reports the most significant structure alignment with 250 aligned residues with an RMSD of 0.9 Å. On the other side, MOMA2 reports a better superposition than TOPMATCH and FATCAT, introducing two rigid-body movements to align these structures. One hinge was detected between the domains, while the second hinge was located inside of the MHC antigen-recognition domain. Compared to Matt, MOMA2 cannot align one of the α-helices of the Histocompatibility antigen associated with the recognition of the antigen with the two α-helices of the Neonatal FC receptor. The MOMA2 algorithm cannot align two helices with a one helix because one technical limitation of this method based on the alignment of matrices is the assignation of equivalences SSE-to-SSE. Other methods like FATCAT typically use the alpha carbons to align a pair of structures.

The third example is the comparison between the elongation factor eEF3 (PDB code 2iw3, chain A), and the Plakophilin (PDB code 1xm9, A) (Figure S10C). The elongation factor eEF3 consists of a HEAT repeat domain, which is associated with the mechanism of transfer RNA release from the E-Site. Their α-helices are stacked constructing a spring-like helicoidal structure. Instead, the Plakophilin protein is composed of nine ARM repeats and its function is associated with cell adhesion. Both structures are alpha-solenoid domains formed by pairs of α-helices. The TOPMATCH alignment reports the lowest number of aligned positions, aligning only some pairs of helices. Otherwise, the flexible alignment obtained by FATCAT has the highest number of residues aligned, followed by MOMA2 and MATT. Although FATCAT

superposes a higher number of residues than MOMA2, introducing five twists (eq = 214 and RMSD = 1.9 Å), the superposition calculated with MOMA2 is more compacted with a similar RMSD than FATCAT introducing only three pairs of aligned sub-fragments (eq = 188 and RMSD = 1.97 Å). FATCAT reports the highest number of residues aligned because this method superimposes small pairs of sub-fragments composed by one or two pairs of secondary structure elements.

Finally, the fourth example is a comparison of the β-propeller domain of the Clathrin heavy chain against the β-propeller domain of Nup159 (Figure S10D). Despite a lack of detectable sequence similarity between the coated vesicle proteins and nucleoporins, it is now generally accepted that these proteins have a common origin (Devos *et al.*, 2004). These domains have the same common fold composed of seven blades of four or three beta-sheets. However, these structures have a considerable local variation among their blades which are difficult to align with a rigid aligner as TOPMATCH. Instead, FATCAT and MATT report longer structural alignments than MOMA2, because these methods introduce many twists and short segments to superpose blades (Figure S10D). The superposition obtained with MOMA2 consists of three pairs of aligned segments which are superposed six of the seven-blade pairs (Figure S10D).



Supplementary Figure 10. Examples of structural superpositions calculated with MOMA2 and other structural aligners.

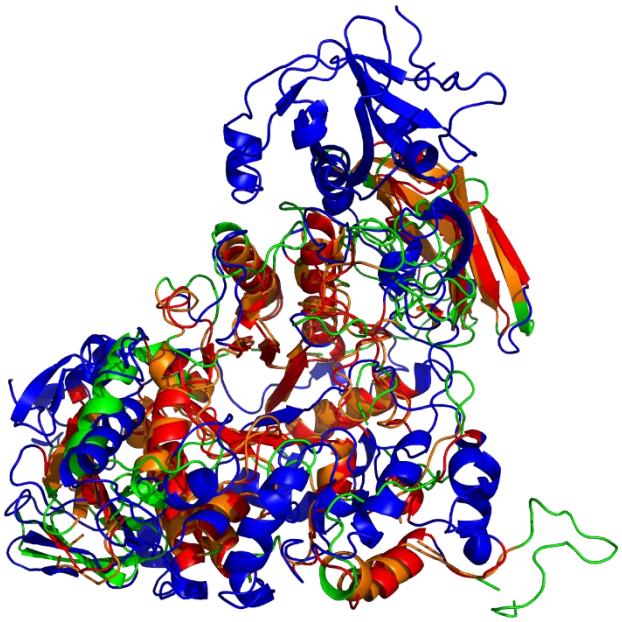
(A) Alignment of the Tissue factor (PDB code 1a21, chain A) and the Growth hormone-binding protein (PDB code 1hwg, chain C)

(B) Alignment of the Histocompatibility antigen (PDB code 2clr, chain A) and the Neonatal FC receptor (PDB code 3fru, chain A)

(C) Structural alignment between the Elongation factor 3 (PDB code 2iw3, chain A; Heat repeat) and the Plakophilin (PDB code 1xm9, chain A; ARM repeat)

(D) Structural alignment between the Clathrin heavy chain (PDB code 1bpo, chain A) and the Nup159 (PDB code 1xip, chain A)

Unaligned residues are shown in blue (query) and green (target), whilst structural matches of the aligned residues derived from query and target structures are colored in red and orange, respectively. The number of aligned residues (eq) and the Root Mean Square Deviation (RMSD) for each superposition was calculated with the STOVCA program.



Supplementary Figure 11. Flexible structural alignment calculated between the pullulanase and the α-amylase.
The superposition was calculated with FATCAT between the ***Klebsiella pneumoniae*** pullulanase (PDB code 2fgz, chain A) and the ***Thermoactinomyces vulgaris*** α-amylase I (PDB code 1uh4, chain A). Equivalent residues at 3.5 Å were colored in red and orange according to STOVCA in the query (blue) and target (green) structures, respectively.

6.1.5. Pseudocodes

6.1.5.1. Iterative algorithm to extract a list of combinations of blocks

INPUT
blocks = ((start$_1$, length$_1$, da$_1$, dd$_1$), …, (start$_n$, length$_n$, da$_n$, dd$_n$))

GET_LIST_OF_TUPLES(blocks)
```
1   i ← 1
2   n ← LENGTH(blocks)
3   T ← ()
4   while i ≤ n
5           start ← blocks_{i,1}
6           length ← blocks_{i,2}
7           positions ← RANGE(start, start+length, 1)
8           da ← blocks_{i,3}
9           dd ← blocks_{i,4}
10          insert (start, length, positions, da, dd) in T
11  return T
```

GET_COMBINATION(T)
```
1   C ← ()
2   positions_covered ←T_{1,3}
3   for i ← 2 to LENGTH(T)
4           start ← T_{i,1}
5           length ← T_{i,2}
6           positions ← T_{i,3}
7           da ← T_{i,4}
8           dd ← T_{i,5}
9           if INTERSECTION(positions, positions_covered) = ø then
10                  positions_covered ← positions_covered U positions
11                  insert (start, length, positions, da, dd) in C
12  da_max = 0
13  dd_max = 0
14  block_selected = ()
15  for i ← 1 to LENGTH(C)
16          da ← C_{i,4}
17          dd ← C_{i,5}
18          if da > da_max or dd > dd_max then
19                  block_selected ← C_i
20                  if da > da_max then
21                          da_max ← da
22                  if dd > dd_max then
23                          dd_max ← dd
```

```
24 remove block_selected from T
25 return C, T

GET_COMBINATIONS(blocks)
1   T ← GET_LIST_OF_TUPLES(blocks)
2   T ← SORT(T)
3   LC ← ()
4   i ← 0
5   while LENGTH(T) < 0
6           C, T ← GET_COMBINATION_OF_BLOCKS(T)
7           insert C in LC
8   LC ← SORT(LC)
9   return LC
```

OUTPUT
LC is the list of combinations of blocks

6.1.5.2. Pseudocode of the Iterative Closest Point (ICP) algorithm

INPUT
C_q and C_t which correspond to the Cα coordinates from query and target local segments.
n is the number of iterations.

```
ICP(Cq, Ct, n)
1   i ← 1
2   initial_Ct ← Ct
3   eqmax, rmsdmin ← EVALUATE_ALIGNMENT(Cq, Ct)
4   eqmin ← eqmax
5   rmsdmax ← rmsdmin
6   while i ≤ n
7           sampleq ← GET_RANDOM_SAMPLE(Cq)
8           samplet ← GET_RANDOM_SAMPLE(Ct)
9           pairsq,t ← FIND_NEAREST_NEIGHBORS(sampleq, samplet)
10          sample_filteredq, sample_filteredt ←
REJECT_BAD_PAIRS(sampleq, samplet, pairsq,t)
11          previous_Ct ← Ct
12          T ← FIND_OPTIMAL_TRANSFORMATION(sample_filteredq,
sample_filteredt)
13          Ct  ← APPLY_TRANSFORMATION(Ct, T)
14          eq, rmsd  ← EVALUATE_ALIGNMENT(Cq, Ct)
15          if rmsd < rmsdmin and eq > eqmax and eq > eqmin then
16                  eqmax  ← eq
17                  rmsdmin  ←  rmsd
```

```
18          else
19                  if eq > eq_max and eq > eq_min and rmsd_min/rmsd ≥ 0.8 then
20                          eq_max ← eq
21                  else if rmsd < rmsd_min and eq > eq_min then
22                          rmsd_min ← rmsd
23                  else
24                          C_t ← previous_C_t
25 T ← FIND_OPTIMAL_TRANSFORMATION(initial_C_t, C_t)
26 return T
```

OUTPUT

T is the transformation matrix which is used to rotate and translate a target segment to a query segment.