CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

LABORATORIO NACIONAL DE GENÓMICA PARA LA BIODIVERSIDAD

Residuos clave y patrones complejos identificados en las interfaces proteína-proteína de las cápsides de virus icosaédricos en el marco de la diversidad estructural cuaternaria de proteínas

Daniel Jorge Montiel García

Director de Tesis

Dr. Mauricio Carrillo Tripp.

Comité

Dra. Laura Silva Rosales.

Dr. Francisco Barona Gómez.

Dr. Robert Winkler.

Dr. Vijay S. Reddy.

Lugares donde se llevó a cabo el trabajo computacional.

- Laboratorio de la Diversidad Biomolecular, Unidad de Genómica Avanzada (LANGEBIO) Cinvestav, Irapuato, Guanajuato, México.

- The Scripps Research Institute (TRSI), La Jolla, California, USA.

Agradecimientos.

A mi esposa, estar casado y estudiar no es cosa sencilla, se requiere de una persona que te brinde su confianza y apoyo incondicional es por eso que estoy casado con Nelly porque simplemente me ha demostrado que puede dar todo e igual seguir dando así no tenga nada, pues todo su apoyo es sin duda lo mejor que me podido recibir estos años.

Para mi madre y mis hermanos que con sus sabios consejos, me alentaron cuando mis fuerzas desfallecían.

Agradezco sinceramente a mi asesor de Tesis, Dr. Mauricio Carrillo Tripp, su esfuerzo y dedicación se ganaron mi lealtad y admiración, me siento en deuda por todo lo recibido durante el periodo de tiempo que ha durado esta Tesis Doctoral.

Agradezco al Dr. Vijay Reddy, por sus consejos en la elaboración de la Tesis, así como por permitirme disfrutar de la experiencia de conocer otras culturas.

Agradezco a los miembros del comité tutorial por las incontables horas que dedicaron a guiarme y asesorarme durante la elaboración del presente trabajo.

Agradezco al Cinvestav por brindarme el espacio para poder desarrollar el proyecto de Tesis Doctoral.

Agradezco al CONACYT por brindarme los recursos para poder alcanzar este grado académico.

Resumen

Uno de los principales objetivos de la Biología Estructural moderna es el de identificar los mecanismos moleculares que permiten que un cierto número de proteínas se agreguen y formen arreglos con una estructura cuaternaria específica que conduzca a una función particular. Se presume que la propiedad principal de las proteínas con función estructural se presenta en forma de sitios de interacción localizados en su superficie. Estos sitios favorecerían el reconocimiento molecular adecuado a través de interacciones entre residuos *clave* (RCs), guiando así el proceso de auto-ensamblaje para que se lleve a cabo de forma correcta, libre de errores. En este contexto, la cápside viral (estructura que protege el genoma viral) siendo la base de la biología mecanicista dadas sus intrigantes características estructurales, por lo que su estudio es extenso en la actualidad. A pesar de esto, aún prevalecen preguntas abiertas, cuyas respuestas ayudarían a describir las reglas fundamentales que subyacen a los procesos estructurales de ensamblaje cuaternario.

En este trabajo exploramos varias ideas relacionadas a la existencia y localización de RCs, definidos bajo un criterio de conservación estructural, usando como modelo la cápside. Iniciamos con un estudio comparativo en el que analizamos la diversidad estructural de proteínas celulares y proteínas de cápside icosaédrica (VCP) para entender las diferencias existentes entre secuencia y estructura de proteínas celulares y VCP, utilizando toda la información a nivel molecular disponible actualmente. Encontramos evidencia que sugiere que los monómeros celulares y las VCP son dos extremos opuestos en el espacio de estructura cuaternaria, donde los oligómeros celulares son un estado intermedio. Notoriamente, la interfaz proteína-proteína (el área de interacción entre 2 proteínas) es en donde se encuentra más de la mitad del total de residuos conservados en el caso de las VCP. En contraste, aproximadamente sólo el 10% de los residuos conservados se localiza en dicha región en los dímeros y oligómeros celulares. Inesperadamente, descubrimos un grupo de familias de virus cuyas VCP están evolucionado significativamente más lento que el resto de las proteínas estudiadas aquí. La relevancia biológica de esta observación será estudiada en trabajos futuros. Posteriormente, desarrollamos un método computacional para la localización de los RCs y lo aplicamos al caso de las VCP. Analizamos esta información con algoritmos de

Inteligencia Artificial para develar redes complejas formadas en la interfaz VCP-VCP. Identificamos patrones que, en general, se caracterizan por mostrar una alta concentración de los RCs alrededor de los ejes de simetría de la cápside, localizados a la misma distancia radial. Observamos diferencias particulares a cada género de virus en función de las propiedades específicas a la naturaleza físico-química de los RCs.

En su totalidad, nuestros resultados sugieren la existencia de reglas generales que se deben seguir para la correcta formación de complejos proteínicos funcionales, en donde los residuos clave deben tener un rol fundamental. Así mismo, aportamos un primer bosquejo de la forma del espacio de estructura cuaternaria, permitiendo especular sobre el origen y evolución de las proteínas con función estructural. Reportamos los resultados de este trabajo en dos publicaciones en la revista de alto impacto Journal of Structural Biology, generando información útil para la formación de nuevas hipótesis en distintas direcciones, que continuaremos investigando en trabajos posteriores.

Abstract

One of the main goals of modern Structural Biology is to identify the molecular mechanisms proteins follow in order to recognize each other and form arrangements with a particular quaternary structure leading to a specific function. One important characteristic of structural proteins is manifested in the form of signals localized at the protein surface. These signals allow molecular recognition through the interactions of *key residues* (KRs), driving an error-free assembly process. In this context, the capsid is still the base of mechanical biology, given its intriguing structural characteristics. Even though the capsid has been studied widely in past decades, there are remaining questions that could help to understand the fundamental rules behind the quaternary assembly process.

In this work, we explore the idea of the existence of KRs, defined by a structure conservation criteria, keeping the viral capsid as a model. We start with a comparative study, analyzing the structural diversity of cellular and icosahedral capsid proteins (VCP) found in nature, using all the information available at the molecular level. We found evidence suggesting that cellular monomers and capsid proteins are two opposite ends in the space of protein quaternary structure, where cell oligomers are an intermediate state. The protein-protein interface is where more than half of all conserved residues are found in the case of capsid proteins. In the other hand, only 10% of conserved residues are in that region in cellular dimers and oligomers. We discovered a group of virus families whose capsid proteins are evolving significantly slower than the rest of the proteins studied here. We will investigate the biological relevance of this observation in future works. Subsequently, we developed a computational method for the localization of KRs and applied it to the case of the VCPs. We analyzed this information with Artificial Intelligence algorithms to reveal complex networks formed in the protein-protein interface of viral capsids. We were able to identify KR patterns, characterized by high-density regions around the capsid symmetry axes and at the same radial distance. However, we observe distinct pattern differences between virus genus depending on the KRs physicochemical properties.

As a whole, our results suggest the existence of general rules that need to be followed for the proper formation of functional protein complexes where the KRs should play a

critical role. Also, we present a first sketch of the shape of the protein quaternary structure space. Our discoveries allow speculations about the origin and evolution of proteins with structural function. We reported the results of this work in two publications in the Journal of Structural Biology.

Índice

1 Introducción 13

1.1 La importancia del estudio de los virus 13

1.2 El impacto socio-económico de los virus 13

1.3 Virus 14

1.3.1 Ciclo de vida viral 14

1.3.2 Auto-ensamblaje espontáneo de proteínas de cápside 15

1.3.3 La cápside viral 16

1.3.4 Descripción de las interfaces VCP-VCP 17

1.4 Conservación de residuos en proteínas con función estructural 18

1.5 Residuos clave en las interfaces proteína-proteína 20

1.5.1 Residuos clave en virus 21

1.6 Herramientas tecnológicas. 22

1.7 Hipótesis 23

1.8 Objetivos 23

1.8.1 Objetivos específicos 23

2 Diferencias estructurales y de conservación entre proteínas celulares y proteínas de cápside 24

2.1 Metodología 25

2.1.1 Datos analizados 25

2.1.2 Similitud de estructura de proteínas 26

2.1.3 Identidad de secuencia de proteínas 26

2.1.4 Asignación de categorías estructurales 28

2.1.5 Conservación de secuencia 29

2.2 Resultados 30

2.2.1 Relación entre identidad de secuencia y similitud estructural 30

2.2.2 Distribución estructural de residuos conservados 34

2.2.3 Conservación de residuos por categoría estructural 40

3 Predicción de residuos clave de cápsides esféricas 45

3.1 Metodología 45

3.1.1 Residuos clave 45
3.1.2 Herramienta computacional para la identificación de residuos clave 47
3.2 Resultados 50
3.2.1 Identificación de residuos clave 50
3.2.2 Desplazamiento de la maya de referencia en las cápsides diferentes a T=3 52
4 Patrones de residuos clave 54
4.1 Metodología 54
4.2 Reconocimiento de patrones (PR) 54
4.2.1 Búsqueda de patrones 54
4.2.2 Clasificación de datos 55
4.2.3 Algoritmo K-Means 56
4.2.4 Preparación de los residuos clave 57
4.2.5 Descripción de los residuos clave 57
4.2.6 Agrupamiento de datos 57
4.2.7 Validación de los grupos de datos 58
4.3 Resultados **..... 59**
4.3.1 Existencia de patrones de residuos clave 59
4.3.2 Número de grupos de datos representativo de los residuos clave 60
5 Discusión 66
6 Conclusiones 72
6.1 Diferencias estructurales y de conservación entre proteínas de cápside y proteínas celulares 72
6.2 Generación de herramientas computacionales 72
6.3 Predicción de residuos clave 73
6.4 Patrones de residuos clave 74
7 Perspectivas 75
8 Bibliografía 76
9 Apéndice 81
10 Productividad científica 98

Índice de Figuras

Fig. 1 Topologías conocidas de las cápsides de virus desnudos.. 16

Fig. 2 Diagrama simplificado que ilustra las diferentes ubicaciones de los aminoácidos en la estructura terciaria y cuaternaria de complejos proteínicos.. 27

Fig. 3 Distribución de los valores del %SASA por residuo en las proteínas.. 28

Fig. 4 Correlación entre identidad en secuencia (S_G) y similitud estructural (TM-score) de proteínas.. 32

Fig. 5 Superposición de la estructura terciaria de dos casos que poseen muy baja identidad en secuencia y una alta similitud estructural..... 33

Fig. 6 Correlación entre el área de la interfaz y la superficie total de las proteínas. 36

Fig. 7 Distribución del número de residuos (primera fila) y densidad de los residuos de la superficie (segunda fila). 37

Fig. 8 Correlación entre las regiones conservadas (Sk*) y la identidad en secuencia(S_G). Cada punto en la nube representa la comparación entre un par de proteínas..... 40

Fig. 9 Distribución de probabilidad de la conservación de residuos basada en la entropía por cadena <S>..... 43

Fig. 10 Comparación estadística entre las familias de virus pertenecientes al grupo 1 y grupo 2 (G1 y G2).. 44

Fig. 11 Transformación del sistema de coordenadas cartesianas a la proyección azimutal polar ortográfica.. 46

Fig. 12 Ejemplos de CapsidMaps con los que se representan cápsides de distinto número T, (a) T=1 Satellite Tobacco Necrosis Virus, y (b) T=3 Southern Bean Mosaic Virus. . 46

Fig. 13 Fragmento de la matriz CCM con los residuos pertenecientes al género Alphanodavirus.. 49

Fig. 14 Análisis de la desviación estándar de los valores de r (barras de error).. 53

Fig. 15 Desplazamiento de la malla de referencia respecto al número T en la representación de CapsidMaps.. 53

Fig. 16 Representación gráfica de la matriz de distancias para los residuos clave de cada género en el estudio. En color azul se representan zonas de alta similitud, y las regiones en colores amarillo y rojo representan zonas de baja similitud. 59

Fig. 17 Análisis del número de grupos representativo de los residuos clave.. 61

Fig. 18 Patrones de residuos clave para los géneros estudiados. 65

Fig. 19 Porcentaje del número promedio de residuos. 66

Índice de Tablas

Tabla 1 Estadística de datos. 30

Tabla 2 Baja identidad en secuencia y alta similitud estructural. 32

Tabla 3 Prueba de Kolmogorov-Smirnov.. 33

Tabla 4 Valor de parámetros f y k del modelo exponencial. 33

Tabla 5 Área de superficie promedio de proteínas. 35

Tabla 6 Número promedio de residuos por región estructural 35

Tabla 7 Prueba T para la correlación del número de residuos. 36

Tabla 8 Porcentaje promedio de residuos conservados por categoría estructural con respecto a la identidad en secuencia. 37

Tabla 9 Conservación de residuos en proteínas.. 39

Tabla 10 Prueba estadística T para la conservación de residuos. 40

Tabla 1 Familias que pertenecen a cada uno de los grupos G…..…….…….40

Tabla 12 Distribución de los grupos de residuos clave por género.. 50

Tabla 13 Promedio de la conservación de residuos normalizada. 70

Tabla 14 Anexo: Dímeros celulares.. 76

Tabla 15 Anexo: Oligómeros celulares..... 79

Tabla 16 Anexo: Proteínas de cápside de virus icosaédricos.. 82

1 Introducción

1.1 La importancia del estudio de los virus

Dentro de la lista de problemas nacionales prioritarios publicada por el Conacyt en 2015 se pueden destacar al menos dos categorías en las que los virus tienen un impacto directo importante: enfermedades emergentes de importancia nacional, y producción de alimentos. Desde hace ya mucho tiempo se ha tomado conciencia del papel que juegan los virus en las enfermedades que aquejan a la población mundial. A lo largo de los siglos, los brotes virales han ocasionado incontables pérdidas humanas. Entre los brotes recientes más importante se encuentran el H1N1, sida, zika, y ébola. Aunque no todos los virus causan enfermedades a los seres humanos, existen virus que infectan animales y plantas, muchos de los cuales son utilizados como sustento base para nuestra alimentación. Por ejemplo, el *citrus tristeza virus* infecta a las especies del género *Citrus* (lima, limón, toronja, etc.), y el *cucumber mosaic virus* que afecta al menos 85 familias de plantas, muchas de ellas importantes para la alimentación de la población o el ganado a nivel mundial.

1.2 El impacto socio-económico de los virus

Diversos estudios realizados recientemente con el objetivo de cuantificar el impacto que causan los virus en la sociedad moderna estiman pérdidas de 200 a 800 millones de dólares por pandemia viral, sólo en los EE.UU. El impacto es significativamente más fuerte en los países en desarrollo que no cuentan con las instituciones ni la infraestructura para contener y tratar este tipo de fenómenos (Webby and Webster 2003; Fraser *et al*., 2009).

México no es la excepción al impacto de los virus. En 2009 se vivió una pandemia causada por el virus H1N1, provocando grandes pérdidas humanas y económicas. Los efectos de la pandemia propiciaron un aumento de la presión sobre la economía y agravó la crisis económica ya existente. Aunque el Banco Mundial extendió a México $25

millones en préstamos para ayuda inmediata y $180 millones más en asistencia a largo plazo, no fue suficiente para restaurar la confianza de la población e inversionistas. El peso tuvo su mayor caída en años. Durante dicha crisis, el sector ganadero experimentó fuertes pérdidas que ascendieron hasta $4.5 millones de dólares diarios a pesar de que se tomaron todas las precauciones para evitar que se extendiera la pandemia a todo el país.

1.3 Virus

Los virus son parásitos que infectan todas las formas de vida celular: eucariotas (animales vertebrados, animales invertebrados, plantas, hongos), procariotas (bacterias y arqueas) e incluso virus. Se estima que existe una mayor diversidad biológica dentro de los virus que la existente en todos los reinos juntos (Carter and Venetia, 2007). Tal diversidad se ve reflejada en el hecho de que existen virus asociados a cada uno de los grupos conocidos de organismos vivos. Estudiar a los virus en el marco de esta diversidad puede ayudar a entender mejor cuáles son los mecanismos involucrados en una gama amplia de procesos biológicos.

Básicamente, los virus están formados por un genoma, que puede ser ADN o ARN, y una capa exterior de proteínas llamada cápside. Algunos virus poseen una capa adicional de lípidos alrededor de la cubierta de proteínas, conocidos como virus envueltos. Debido a que en la actualidad existe una gran cantidad de información estructural disponible sólo para virus sin la capa de lípidos, en este trabajo se consideran únicamente los virus desnudos.

1.3.1 Ciclo de vida viral

Dado que la partícula viral, también conocida como virión, carece de la maquinaria necesaria para llevar a cabo la auto-replicación, ésta necesita infectar un tipo particular de célula (hospedero) para lograr hacer copias de sí misma. Independientemente del tipo o especie de virus, este proceso se puede generalizar definiendo seis fases principales que componen el ciclo vital de los virus: 1) reconocimiento y unión de un virión a una célula, 2) entrada del virión a la célula 3) desensamblaje del virión, liberando así el

material genético viral, 4) replicación del genoma y proteínas virales en el interior de la célula, 5) ensamblaje espontáneo de las proteínas de cápside alrededor del genoma viral dando lugar a nuevos viriones, y 6) liberación de los viriones de la célula infectada. Cada una de estas fases involucra en sí procesos biológicos complejos, y cuyos mecanismos moleculares son poco entendidos a la fecha. Este trabajo se centra en ciertos aspectos particulares involucrados en la quinta etapa del ciclo vital de los virus, i. e., cómo interactúan las proteínas que formarán la cápside a través del proceso de auto-ensamblado de virus desnudos con topologías icosaédricas.

1.3.2 Auto-ensamblaje espontáneo de proteínas de cápside

A pesar de que se han realizado diversos esfuerzos tratando de dilucidar el mecanismo molecular del auto-ensamblaje viral, este proceso aún sigue siendo poco entendido. Las proteínas de cápside tienen que encontrar la forma de interactuar entre sí dentro de la célula de manera específica con el fin de formar una macroestructura con la topología adecuada. Además, estas proteínas tienen que reconocer el genoma correcto a encapsular.

La complejidad del proceso de auto-ensamblaje se ha atacado a través de trabajos experimentales y teóricos, que como resultado han permitido describir este mecanismo en tres fases generales (Endres and Zlotnick, 2002; Keef, 2006; Rapaport, 2008). Primero, las interacciones inter-moleculares débiles son necesarias para minimizar los errores de ensamblaje y las trampas cinéticas. Las interacciones débiles contribuyen a definir el inicio de la segunda etapa, nucleación. Aquí se inicia el ensamblado, disminuyendo las trampas cinéticas de los estados intermedios debido al agotamiento de las unidades de ensamblaje. Por último, la fase cinética, donde ya existe una pequeña parte de la cápside formada la cual soportará los estados intermedios del ensamblaje ayudando así a formar la partícula completa (Zlotnick and Mukhopadhyay, 2011).

Desde el punto de vista experimental, los viriones pueden ser desensamblados en sus componentes individuales (proteínas y ácidos nucleicos) en el laboratorio. Algunos virus pueden ser re-ensamblados en viriones infecciosos a partir de los componentes purificados cuando se tienen las condiciones apropiadas de pH y fuerza iónica (Bayer *et*

al., 1968; Santi *et al.*, 2006). Esto habla de un proceso reversible, esto es, las proteínas de cápside son elementos estables que proporcionan robustez al mecanismo molecular de auto-ensamblaje.

1.3.3 La cápside viral

La función principal de la cápside es la de proteger el genoma viral. Después de salir de la célula hospedera, el virión entra en un ambiente hostil que rápidamente dañaría el genoma viral si no tuviera la protección adecuada. Los ácidos nucleicos que componen el genoma viral son susceptibles a daños físicos, tales como corte por fuerzas mecánicas, o modificación química por luz ultravioleta (luz solar), entre otros. Sin embargo, estas no son las únicas funciones de la cápside. En muchos casos, la cápside también es responsable del reconocimiento de las células hospederas correctas, estableciendo la primera interacción virión-célula.

En general, la cápside se forma a partir de un número relativamente pequeño de copias de una proteína. En algunos virus se involucra más de un tipo de proteína para formar la cápside, pero siempre es un número reducido. El plegado terciario de las proteínas de cápside (VCP) es asimétrico, pero estas interactúan entre si y se organizan para formar estructuras simétricas macro-moleculares de forma espontánea. Los virus desnudos que se han observado hasta la fecha poseen alguno de dos tipos conocidos de simetría: helicoidal o icosaédrica (Fig. 1). Ambas estructuras simétricas están formadas por una matriz de proteínas construida por copias químicamente idénticas (Carter and Venetia, 2007).

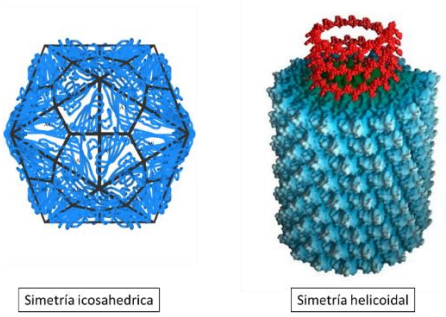


Fig. 1 Topologías conocidas de las cápsides de virus desnudos. A la izquierda se pude observar un virus con simetría icosahedrica. Las líneas negras corresponden al modelo geométrico (lattice) empleado para describir la topología de la cápside. A la derecha se puede observar un virus con simetría helicoidal.

En el caso particular de los virus icosaédricos, la cápside más simple está formada por 60 copias químicamente idénticas de una VCP. Dada la naturaleza geométrica de los icosaedros (20 caras), no es posible colocar de manera simétrica más de una proteína por cara. Por esta razón, Caspar y Klug propusieron el modelo de la cuasi-simetría, que explica la manera de colocar de manera simétrica más de una proteína por cara (Caspar and Klug, 1962). En su modelo se asume que el tamaño de las caras permanece constante y sólo el número de subunidades varia. La cara triangular del icosaedro es subdividida en pequeños triángulos (triangulación). Esta subdivisión no es arbitraria y está regida por la siguiente relación

$$T = H^2 + HK + K^2$$

donde *T* es el número de triangulación, y *H* y *K* son números enteros positivos incluyendo el cero, definidos como los ejes principales en una maya hexagonal (Prasad *et al.*, 2012). De esta forma fue posible explicar la existencia de más de una proteína por cara en cápsides icosaédricas complejas, y desde entonces se ha utilizado al número *T* como una forma de clasificar topológicamente este tipo de macro-estructuras.

1.3.4 Descripción de las interfaces VCP-VCP

Con el fin de caracterizar la organización de las proteínas de cápside y las interacciones entre ellas, se han propuesto diferentes métricas. Por ejemplo, el Q-score es una medida utilizada para describir y comparar las interfaces proteína-proteína de las subunidades de los virus icosaédricos (Damodaran *et al.*, 2002). El Q-score se puede definir como el grado de similitud entre las interacciones de las interfaces proteína-proteína, donde valores cercanos a 1 corresponden a dos interfaces similares en términos estructurales. El Q-score permite realizar una comparación cuantitativa de las interfaces, reflejando la forma y la simetría. La aplicación de este método proporciona una huella descriptiva de la arquitectura de las interfaces proteína-proteína. Por otra parte, el índice de interacción proteína-proteína (PPI) es una métrica utilizada para comparar las interacciones proteína-proteína a través de la relación que existe entre el área que se encuentra oculta al solvente con respecto al área accesible al solvente entre sus estados monómerico y

oligomerico (Shepherd *et al.*, 2006). Este índice es sensible a los cambios en la organización de la cápside y las asociaciones entre las diferentes subunidades. El PPI se utilizó para describir en detalle la cápside del virus del papiloma humano en términos de las interacciones proteína-proteína entre las subunidades y así identificar cápsides estructuralmente relacionadas. El Similarity score (S-score) es otra forma de medir y comparar de forma cuantitativa las interacciones proteína-proteína a nivel cuaternario entre un par de cápsides de virus icosaédricos (Carrillo-Tripp *et al.*, 2008). El valor de puntuación S corresponde a la fracción de lugares de interacción cuaternaria comunes entre un par de cápsides con respecto al número total de interacciones encontradas en ellas.

Cada una de las métricas mencionadas anteriormente tiene como objetivo describir de forma cuantitativa las interacciones proteína-proteína; sin embargo, cada una se especializa en medir un componente en particular. El Q-Score se usa para describir las interacciones entre proteínas en el contexto de forma y simetría, el PPI a través de la relación entre el área oculta y accesible al solvente en las diferentes conformaciones de la proteína, y por último, el S-score se centra en medir las interacciones cuaternarias.

1.4 Conservación de residuos en proteínas con función estructural

Trabajos previos que han descrito complejos celulares proteínicos sugieren que, en general, la complementariedad en las interfaces proteína-proteína se mantiene. Las interfaces se caracterizan comúnmente en términos de su área, superficie oculta al solvente y densidad de empaquetamiento. El enfoque moderno hacia las interfaces proteína-proteína establece reglas basadas en características que incluyen el tamaño, composición y tipo de residuo, hidrofobicidad y planaridad (Lawrence and Colman, 1993; Jones and Thornton, 1995).

Sin embargo, el incremento exponencial de información estructural de proteínas ha permitido revaluar los resultados obtenidos en estudios anteriores. Por ejemplo, hace un poco más de diez años se realizó un análisis de 64 interfaces proteína-proteína celulares usando métricas de conservación derivadas de dos tipos de alineamientos múltiples de secuencia, uno de ortólogos y el otro de parálogos. En ese estudio se mostró que los residuos de interfaz son más conservados que residuos en otras regiones de la superficie cuando se utiliza cualquiera de los dos tipos de alineamiento (Caffrey *et al.*, 2004). Posteriormente, un estudio de 1,812 estructuras no redundantes mostró que las interfaces de proteínas celulares difieren significativamente del core y la superficie de la proteína en términos de su composición de aminoácidos y preferencias en las interacciones entre residuos (Ofran and Rost, 2003). Más recientemente, un estudio de 2,310 proteínas celulares que forman complejos mostró que las interfaces difieren en su estructura secundaria con respecto al core y la superficie (Yan *et al.*, 2008).

La forma y el plegado terciario de las proteínas juegan un rol importante en la formación de complejos proteínicos. Es bien conocido que la secuencia primaria puede divergir más allá del punto donde es detectable la similitud a este nivel, sin embargo, la estructura terciaria puede conservar el mismo plegado. Se han realizado esfuerzos que intentan entender cuáles son los principios físico-químicos que permiten que exista una conservación estructural a pesar de la variación en secuencia. Como primera aproximación, se han realizado trabajos que buscan la clasificación del espacio de plegado de las proteínas. Por ejemplo, la clasificación estructural de proteínas (SCOP) es un sistema de clasificación jerárquica de dominios de proteína. Agrupa aquellos dominios en los que se tiene evidencia estructural, funcional y en secuencia de la existencia de un ancestro evolutivo común, a nivel de superfamilia, SF (Murzin *et al.*, 1995). Para poder comparar proteínas a lo largo del proceso evolutivo es necesario considerar los dominios (fragmentos conservados de secuencia y estructura terciaria) SF, que en principio son las unidades ancestrales fundamentales. En particular, de 560 dominios encontrados en los virus a nivel SF, más del 10% no tienen ningún pariente evolutivo estructural en los organismos celulares modernos (Abroi and Gough, 2011). Esta separación con respecto a los organismos celulares puede ser consecuencia de

que los virus evolucionan seis órdenes de magnitud más rápido, además de ser una posible consecuencia de restricciones evolutivas. Se ha propuesto que la transferencia de dominios de virus a hospederos celulares pudiese ser un mecanismo acelerado de evolución de proteínas.

Algunas características estructurales parecen ser únicas a las proteínas virales cuando son comparadas con sus contrapartes celulares. Por ejemplo, se ha observado que las proteínas virales (catalíticas, no estructurales) presentan una baja densidad de residuos, alta ocurrencia de segmentos aleatorios de asas, pequeñas regiones desordenadas y menores efectos de desestabilización cuando las mutaciones ocurren. Se piensa que estas características permiten que las proteínas virales tengan una alta flexibilidad, confiriéndoles maneras efectivas de interactuar con los componentes del hospedero, incluyendo la evasión de las contramedidas celulares tras la infección, y más recientemente, medicamentos antivirales (Tokuriki *et al.*, 2009).

Como se mencionó con anterioridad, la poca evidencia que existe sugiere que las VCP son distinguibles en términos de conservación, tanto en secuencia como en estructura terciaria, cuando se comparan con proteínas celulares, aunque esto no se ha mostrado de forma sistemática. Por ejemplo, un análisis reciente hecho sobre un set de 319 virus icosaédricos sugirió que las VCP se encuentran segregadas en el espacio de plegados con respecto a las proteínas celulares. Dicho estudio propone que el plegado de las proteínas de capside viral presenta una geometría favorable para el empaquetamiento y ensamblado de grandes complejos macro-estructurales como lo son las cápsides (Cheng and Brooks III, 2013). Subsecuentemente, otro estudio mostró que ciertos patrones globales derivados de la estructura de la cápside, tal como la densidad de empaquetamiento, son consistentes con patrones presentes en perfiles de conservación en secuencia de las VCP (Chih-Min *et al.*, 2015).

1.5 Residuos clave en las interfaces proteína-proteína

Las interfaces proteína-proteína están compuestas por dos superficies con buena complementariedad espacial y electrostática. Se asume a menudo que la energía de

unión proteína-proteína está directamente relacionada con el área que forma la interfaz. Para estudiar la contribución energética de cada residuo de la interfaz a la interacción proteína-proteína, Thorn construyó una base de datos compuesta por 2,325 mutantes puntuales de alanina de proteínas celulares, para las cuales estimaron el cambio de energía libre de unión con respecto a las proteínas silvestres (Bogan and Thorn, 1998). Los autores encontraron que la energía libre de enlace no está uniformemente distribuida sobre cada uno de los residuos que forman la interfaz. Por el contrario, se observó que existen ciertos residuos que contribuyen predominantemente a la energía libre de unión. Un estudio posterior recopiló todas las interfaces proteínicas disponibles en el PDB en 2002, y se encontró que los residuos estructuralmente conservados en las interfaces correlacionan con los residuos clave identificados a través de su contribución energética (Keskin *et al.*, 2005). Adicionalmente, se observó que los residuos clave de proteínas celulares no se encuentran localizados de manera homogénea a lo largo de toda la interfaz, sino que se encuentran concentrados en regiones con una alta concentración de residuos, formando una red compleja de interacciones.

1.5.1 Residuos clave en virus

Recientemente se propuso una metodología alternativa para predecir la localización de residuos clave en cápsides virales, definiéndolos como aquellos residuos en las interfaces VCP-VCP que se encuentran conservados en todos los niveles estructurales, i. e., desde el primario hasta el cuaternario (Carrillo-Tripp *et al.*, 2008). Utilizando este criterio, se identificaron 34 residuos clave comunes a los virus pertenecientes a la familia *Nodaviridae,* utilizando toda la información estructural disponible para esa familia en ese momento (PDBID: 2bbv, 1nov, 1f8v, fhv). Se observó una tendencia de ubicación de los residuos clave cerca de los ejes de simetría de la cápside y manteniéndose sobre un plano tangente a estos (ejes de simetría).

Siguiendo la misma metodología, posteriormente se identificaron 8 residuos clave comunes a miembros de la familia *Bromoviridae*. En este caso, adicionalmente se estimó termodinámicamente el cambio en la energía libre de unión VCP-VCP una vez que se

mutaron cada uno de los residuos clave predichos. Se observó que los cambios son drásticos en comparación con cualquier otro residuo de interfaz (Diaz-Valle *et al.*, 2014). Dichos resultados aportaron una validación teórica al método de predicción de residuos clave basada en el criterio de conservación estructural. La validación experimental se está llevando a cabo por otros miembros de nuestro grupo (Diaz-Valle *et al.*, en proceso).

1.6 Herramientas tecnológicas.

Las diversas bases de datos alrededor del mundo, que resguardan los avances y resultados biológicos obtenidos a lo largo de los años, han crecido y continúan creciendo de manera exponencial. De tal manera que las técnicas convencionales de procesamiento y análisis de datos se han vuelto inadecuadas. Para poder analizar las crecientes bases de datos se requiere de afrontar y resolver numerosos retos entre los que destacan: organización, búsqueda, almacenamiento, transferencia y visualización de datos. Para afrontar los retos derivados del análisis de estas extensas bases de datos, ha nacido una nueva área de investigación llamada “BigData”. La cual puede definirse como el análisis de colecciones de datos, cuyos tamaños van más allá de la capacidad de las herramientas de software comúnmente usadas para capturar, procesar y manejar los datos dentro de un tiempo tolerable (Snijders, 2012).

Las herramientas de visualización involucran la creación y el estudio de la representación visual de los datos. En otras palabras información que ha sido resumida en un esquema, incluyendo atributos y variables para las unidades de información. Esta área es de suma importancia para el campo de la biología, debido a que requiere de poder sintetizar un esquema una cantidad significativa de variables. La biología ha observado múltiples avances en este campo entre los cuales podemos destacar los siguientes: CapsidMaps, Arboles filogenéticos, Grafos, Ontologías, solo por mencionar algunas.

Las diversas herramientas anteriormente mencionadas se deben aplicar sin olvidar los principios planteados por la ingeniería de software. La cual establece las guías requeridas para construir software de calidad. En el contexto biológico el software construido siguiendo los criterios de calidad, permite el mantenimiento, expansión y colaboración de las herramientas informáticas. De lo contrario es muy probable que las

herramientas informáticas construidas sean herramientas de un solo uso, permitiendo responder solo una o un pequeño conjunto de preguntas.

1.7 Hipótesis

Existen residuos clave en las interfaces VCP-VCP de la cápside, distinguibles por su conservación en secuencia y estructura cuaternaria, distribuidos de forma no-aleatoria y organizados de tal manera que se forman patrones característicos.

1.8 Objetivos

Estimar la naturaleza única de la VCP en comparación con proteínas celulares en el contexto de la conservación de residuos a distintos niveles estructurales. Así mismo, identificar la localización de los residuos *clave* en la interfaz VCP-VCP y buscar patrones característicos a cada familia o género de virus icosaédricos a través del análisis de la diversidad estructural encontrada en la cápside a nivel cuaternario.

1.8.1 Objetivos específicos

- Estudiar la correlación existente entre la conservación de secuencia y la similitud del plegado de proteínas analizando la diversidad estructural conocida.
- Realizar un análisis sistemático de toda la información estructural de proteínas de cápside disponible a la fecha, e identificar la localización de residuos *clave* en la interfaz VCP-VCP de cápsides icosaédricas, con el uso de una herramienta computacional que utilice paradigmas del área de BigData.
- Develar la existencia de patrones complejos de residuos *clave* en las interfaces de la cápside de los virus icosahedricos, a través de algoritmos de Inteligencia Artificial.

2 Diferencias estructurales y de conservación entre proteínas celulares y proteínas de cápside

Es ampliamente aceptado que la similitud estructural entre dos proteínas está directamente relacionado con la identidad en secuencia. En este capítulo mostramos resultados comparativos del análisis de la correlación, distribución y niveles de variación de los residuos conservados en la estructura de las proteínas celulares y VCPs. Para esto, utilizamos toda la información estructural disponible de alta resolución disponible a la fecha. Los residuos se categorizaron en función de distintas regiones estructurales. En todos los complejos analizados, los residuos encontrados en el núcleo de las proteínas son los que se encuentran más conservados, seguido por los residuos en las interfaces proteína-proteína. Los residuos expuestos al solvente muestran grandes variaciones en secuencia. Nuestros resultados sugieren que los monómeros celulares y las proteínas de cápside pueden ser dos extremos en el espacio de interacciones cuaternarias, en donde los dímeros y oligómeros celulares se encuentran como estados intermedios. Adicionalmente, encontramos un grupo de familias de virus icosaédricos cuyas proteínas de cápside parecen estar evolucionando a un ritmo más lento que el resto de los complejos analizados. Los resultados de este capítulo fueron publicados en (Montiel-García *et al.*, 2016).

2.1 Metodología

2.1.1 Datos analizados

Se recopilaron todos los datos disponibles de estructuras tridimensionales y las secuencias correspondientes de proteínas celulares y proteínas de cápside icosaédricas. Los datos recopilados se dividieron en cuatro grupos. Para los complejos celulares, que llamaremos de forma general *n-meros*, se incluyeron monómeros (primer grupo) (n=1), dímeros (n=2, Apéndice Tabla 15, segundo grupo), y oligómeros (n=3, 4, 5, 6, 8, 10,12 y 22, Apéndice Tabla 16, tercer grupo). El cuarto grupo está compuesto por proteínas de cápside las cuales abarcan 36 géneros de 21 familias diferentes de virus icosaédricos, de acuerdo a la clasificación propuesta por el comité internacional de taxonomía de virus (ICTV) (Fauquet *et al.*, 2005). Este último dataset toma en cuenta un rango amplio de números T, representativo de la diversidad topológica conocida a la fecha de virus icosaédricos: 1, 2, 3, 4, 7d, 7l, pT3 (Apéndice Tabla 17, cuarto grupo).

Para garantizar una alta calidad de la información analizada, los criterios usados para filtrar los datos incluyeron el que las estructuras tridimensionales hubieran sido determinadas por difracción de rayos X (con una resolución <= 4Å), tuviesen cadenas polipeptídicas con secuencias completas y consistentes (sin asas faltantes o fragmentos mal anotados), y cadenas largas (> 65 residuos). Las coordenadas atómicas de las proteínas celulares fueron extraídas del Protein Data Bank (PDB) (Berman and Westbrook, 2000), y las de las proteínas de cápside fueron tomadas del Virus Particle Explorer Database (VIPERdb) (Carrillo-Tripp *et al.*, 2009). Siguiendo con la definición estándar, el termino *protómero* denota una cadena única en un complejo multimérico (Valdar and Thornton, 2001). Por lo tanto, cada homó-mero se representa por sólo un protómero, mientras que los heteró-meros se representan por dos o más protómeros. Así, los grupos de datos que analizamos en este trabajo están compuestos por 5,087 monómeros celulares, 51 dímeros celulares (representados por 57 protómeros), 65 oligómeros celulares (representados por 101 protómeros) y 212 proteínas de capside (representadas por 293 protómeros).

2.1.2 Similitud de estructura de proteínas

La métrica Root Mean Square Deviation (RMSD) ha sido el estándar para medir la similitud estructural entre dos proteínas desde los inicios de la biología estructural. Sin embargo, otras métricas propuestas recientemente proveen una mejor cuantificación. Tal es el caso del TM-Score (Zhang and Skolnick, 2005). El TM-Score presenta ventajas significativas sobre el RMSD. Los valores de TM-Score están acotados en el intervalo (0,1], donde 1 representa dos estructuras idénticas (equivalente a un RMSD de 0). Además, el TM-Score es independiente del tamaño de las proteínas.

En el presente trabajo, empleamos la métrica TM-Score para evaluar la similitud entre proteínas. Elegimos al TM-Score debido a que presenta las siguientes características: es una métrica con bajo costo computacional, lo que la hace altamente atractiva para ser implementada en análisis masivos de datos (BigData), y posee una sensibilidad mayor a otras métricas comúnmente usadas para evaluar similitud estructural. Una característica importante del TM-Score es que valores >0.5 indican que las dos estructuras comparadas tienen el mismo plegado (Xu and Zhang, 2010). La herramienta computacional TM-Align (Zhang and Skolnick, 2005) identifica la mejor superposición estructural entre un par de proteínas y en esta configuración calcula el TM-Score. El alineamiento global de aminoácidos entre las proteínas fue derivada de dicha superposición estructural óptima.

2.1.3 Identidad de secuencia de proteínas

La identidad en secuencia (S_G) se define como la fracción de residuos idénticos (I_G) con respecto al número total de residuos alineados (A_G) entre las secuencias comparadas, esto es $S_G = \frac{I_G}{A_G}$. Para distinguir la ubicación de los residuos conservados a lo largo de la estructura de la proteína definimos tres categorías: interfaz proteína-proteína (IN), core (CO) y superficie accesible al solvente (S). Estas regiones se ilustran en la Fig. 2 para el caso de monómeros, dímeros, y oligómeros o VCPs. Así mismo, definimos un índice de identidad de secuencia por categoría,

$$S_K^* = \frac{I_K}{I_G}(K = IN, CO, S),$$

para cuantificar el porcentaje de conservación encontrado en diferentes regiones de la proteína. Mapeamos la localización de los aminoácidos respecto las diferentes categorías estructurales después de realizar el alineamiento de secuencias a partir del alineamiento tridimensional. Bajo este concepto, I_K es el número de residuos conservados en cada categoría estructural. Este procedimiento identifica también aquellos residuos conservados que no pertenecen a la misma categoría estructural en ambas secuencias del par de proteínas analizadas (Fig. 2D). A dichos residuos les llamamos huérfanos (ORPH) en el contexto de este trabajo.

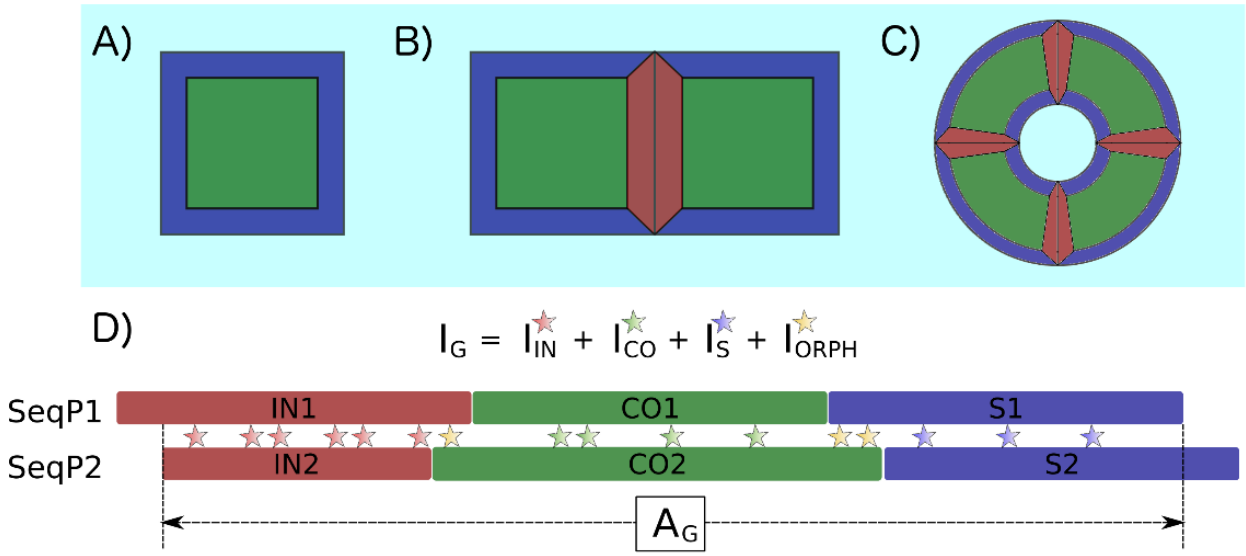


Fig. 2 Diagrama simplificado que ilustra las diferentes ubicaciones de los aminoácidos en la estructura terciaria y cuaternaria de complejos proteínicos. Los aminoácidos encontrados en la superficie accesible al solvente, S, están marcados en color azul, los aminoácidos del núcleo de la proteína, CO, en verde, y los aminoácidos de la interfaz proteína-proteína, IN, en rojo. La distribución se ilustra para: (A) un monómero, (B) un dímero, y (C) un oligómero o cápside. Para estudiar el nivel de conservación de los aminoácidos en las ubicaciones antes mencionadas, se generó un alineamiento en secuencia a partir del alineamiento estructural para cada par de proteínas (P1 vs P2). (D) Los residuos conservados, I_G, son identificados y etiquetados de acuerdo a su ubicación en la estructura de la proteína.

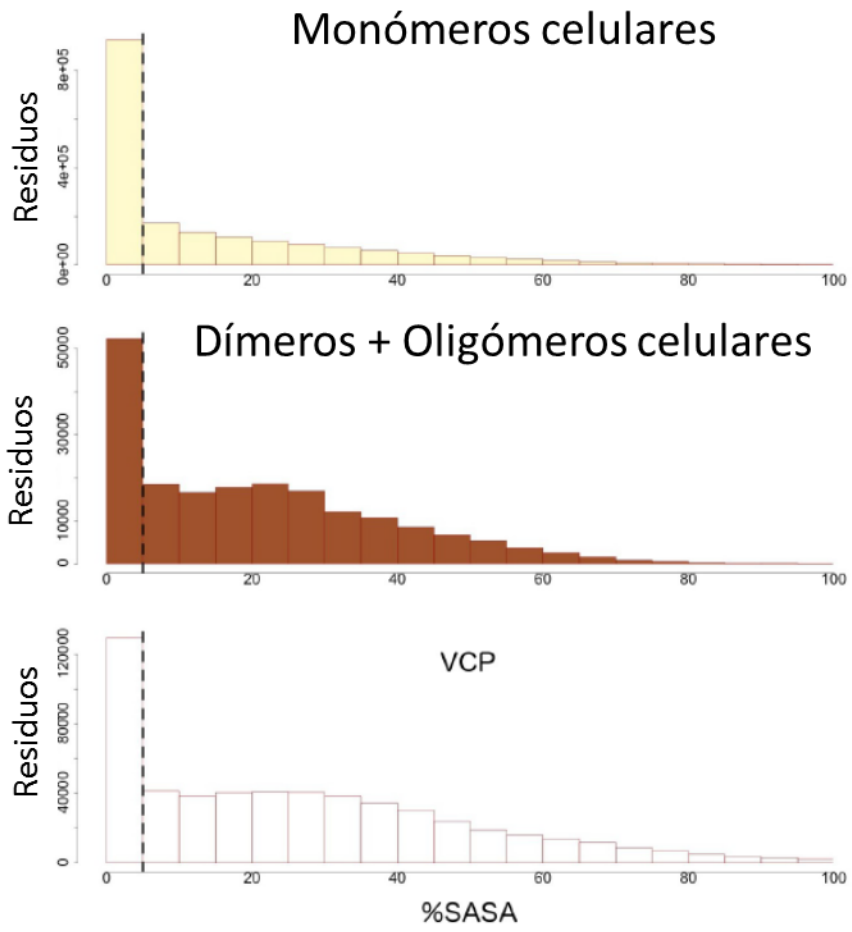


Fig. 3 Distribución de los valores del %SASA por residuo en las proteínas. La línea punteada vertical muestra el valor del umbral utilizado como criterio para identificar a los residuos del core (<5%), superficie (>5%).

2.1.4 Asignación de categorías estructurales

La clasificación estructural toma en cuenta la estructura terciaria y cuaternaria de los complejos proteínicos. Esto distingue cuándo los residuos están en la interfaz proteína-proteína o en la superficie accesible al solvente (Fig. 2 A-C). En este trabajo se utilizó el mismo criterio de clasificación estructural para las proteínas celulares y de cápside. Los residuos de interfaz se definen como todos aquellos que tienen al menos un contacto cercano con una proteína vecina. Para identificarlos se utilizó el método de distancias inter-residuo, la cual es una estrategia para localizar residuos en contacto entre dos moléculas cercanas, AB. En el caso de las interacciones proteína-proteína, la definición de contactos utilizada fue obtenida de (Damodaran *et al.*, 2002) la cual provee una

descripción de la presencia/ausencia de interacciones entre residuos en la interfaz AB. Este método requiere el cálculo de la distancia entre cada residuo de la proteína A y cada residuo de la proteína B. Los pares de residuos que se encuentran a una distancia menor a un radio de corte definido son identificados como residuos parte de la interfaz del complejo proteínico AB. Este método funciona bien cuando las posiciones atómicas del complejo A y B son conocidas. Debido a que la información estructural de las moléculas del solvente no se encuentran disponibles, no es posible utilizar este método para diferenciar los residuos del core y la superficie accesible al solvente. Dada la restricción planteada con anterioridad, se propuso utilizar el concepto de área accesible al solvente (SASA) como criterio de asignación para core y superficie. Los residuos del core se encuentran ocultos al solvente dentro de la proteína, mientras que los residuos accesibles al solvente pueden interactuar directamente con el ambiente. De acuerdo a lo anterior, se puede hacer uso del nivel de exposición al solvente como criterio para distinguir los residuos que no pertenecen a la interfaz entre core o superficie. Este método permite distinguir cuáles residuos tienen suficiente área accesible al solvente para ser considerados parte de la superficie de la proteína. Así, calculamos el valor de SASA por residuo para las proteínas celulares y proteínas de cápside (Fig. 3). Encontramos un pico en el intervalo [0,5] del área relativa de la superficie accesible al solvente (%SASA) para proteínas celulares y cápside. Asumimos que los residuos en este rango son proteínas de core y, por el contrario, residuos con %SASA > 5% son residuos de superficie. Los valores de SASA fueron calculados usando la librería de PDBASA (Shrake and Rupley, 1973).

2.1.5 Conservación de secuencia

La conservación en secuencia está relacionada a la variabilidad de los residuos en la posición i *-ésima* de la secuencia en una cadena, medida por la entropía de Shannon,

$$S(i) = -\sum_k p_k \ln p_k$$

donde $p_k = \frac{n_k}{N}$ es la frecuencia de un residuo de tipo k y n_k es la fracción de secuencias que tienen el residuo de tipo k en la posición i *-ésima* en un alineamiento múltiple (MSA)

de secuencias. S(i) varía entre 0, en posiciones altamente conservadas, y aproximadamente 3, cuando cualquier tipo de aminoácido es igualmente encontrado en la posición i*-ésima* del MSA. Adicionalmente, la entropía normalizada se define como $s(i) = \frac{S(i)}{< S >}$, donde $< S >$ es el valor promedio de $S(i)$ de todos los residuos de la cadena de polipéptidos. Los valores de $S(i)$ de cada protómero fueron obtenidos de la base de datos HSSP (estructuras de proteínas derivadas por homología) (Touw *et al.*, 2015).

2.2 Resultados

2.2.1 Relación entre identidad de secuencia y similitud estructural

La década de los 80's marcó el inicio de una explosión en la cantidad de información estructural generada a nivel molecular en el campo de la biología. A pesar de esto, en esa época aún eran escasas las estructuras de alta resolución de proteínas celulares. En 1986 el grupo de Lesk analizó la secuencia y estructura de 32 pares de proteínas celulares homólogas (Chotia and Lesk, 1986). Los hallazgos de su estudio mostraron que los cambios estructurales estaban directamente relacionados con la extensión en los cambios de la secuencia. Siguiendo la misma estrategia metodológica, en este trabajo realizamos un análisis comparativo por pares de estructuras y secuencias de todos los protómeros que componen nuestros cuatro grupos de datos. De forma análoga al trabajo de Lezk, comparamos la divergencia estructural, definida como (*1-TM-Score*), en función de la fracción de residuos mutados (*1-*S_G). Nuestros resultados se muestran en la Fig. 4. Los umbrales característicos para TM-Score y S_G producen cuatro sectores en el plano cartesiano. El sector I contiene pares de proteínas homólogas con diferencias en el plegado terciario. El sector II contiene pares de proteínas no-homólogas con diferencias en el plegado terciario. El sector III contiene pares de proteínas no-homólogas con el mismo plegado terciario. El sector IV contiene pares de proteínas homólogas con

el mismo plegado terciario. El número total de protómeros analizados se reporta en la Tabla 2. A primera vista pareciera que las nubes de puntos tienen la misma distribución. Sin embargo, un análisis de densidad mostró diferencias entre todos los grupos de datos, aunque el contraste más marcado se observa entre las proteínas celulares y las de cápside. La mayoría de los pares caen en el sector II para las proteínas celulares, de acuerdo a observaciones previas (Sander and Schneider, 1991; Rost, 1999), mientras que los pares de proteínas de cápside están distribuidas en los sectores II, III y IV de forma un tanto uniforme.

Tabla 2 Número total de pares de proteínas analizados para cada conjunto de datos por sectores. El porcentaje de pares en cada sector con respecto al total se muestra entre paréntesis. Las proteínas homólogas (pares con identidad en secuencia $S_G > 0.3$*) se encuentran en los sectores I y IV.*

	Total de pares	Sector I	Sector II	Sector III	Sector IV
Monómeros	12,936,241	103 (0.001%)	12,822,459 (99.10%)	105,459 (0.82%)	8,220 (0.064%)
Monómeros RND	38,809	0 (0.0%)	38,480 (99.15%)	304 (0.78%)	25 (0.064%)
Dímeros + Oligómeros	1,596+5,050= 8,290	8 (0.097%)	8,037 (97.00%)	133 (1.60%)	112 (1.35%)
VCP	42,778	0 (0.0%)	14,454 (33.79%)	12,348 (28.87%)	15,976 (37.35%)

El porcentaje de comparaciones en el sector III es bajo para las proteínas celulares, sin embargo, encontramos ejemplos donde el plegado de las proteínas está altamente conservado a pesar de una gran divergencia en secuencia. Las proteínas que componen cada uno de estos pares atípicos tienen diferente función y provienen de un organismo distinto, como se muestra en la

Tabla 3

Sorpresivamente, los casos extremos que identificamos en este estudio no involucran el plegado ubicuo barril beta (TIM), sino son proteínas con plegados beta-propeller o beta-trifoil (clasificación CATH 2.130 y 2.80, respectivamente) como se aprecia en la Fig. 5.

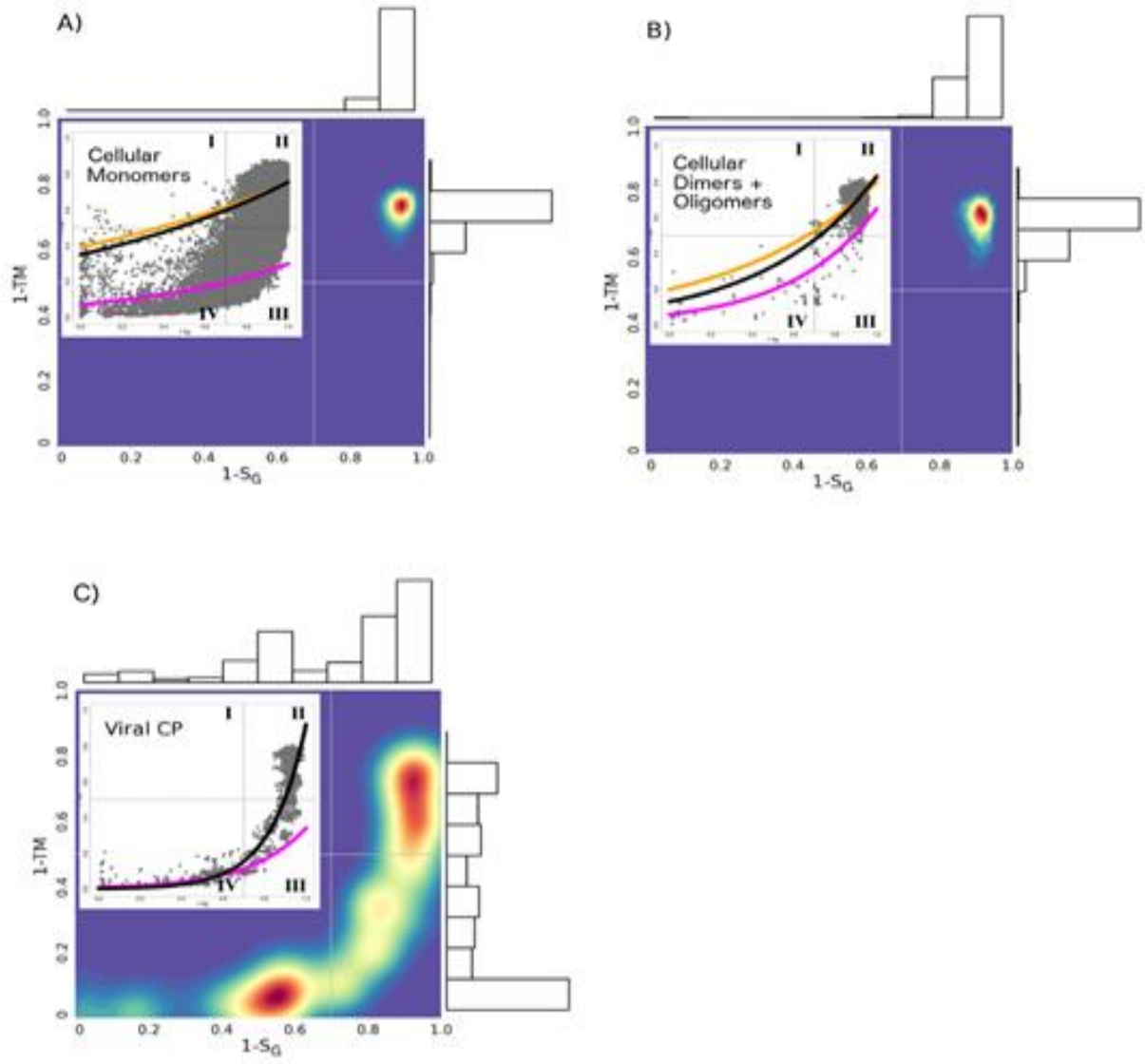


Fig. 4 Correlación entre identidad en secuencia (S_G) y similitud estructural (TM-score) de proteínas. De manera equivalente, se muestra la fracción de residuos mutados (1-S_G) y la divergencia estructural (1-TM). (A) Análisis de todas las combinaciones por pares de monómeros celulares, (B) dímeros celulares y oligómeros, (C) proteínas de cápside de virus icosaédricos. Se muestra el estimador kernel normal bivariante en dos dimensiones, evaluado en una malla de 300 puntos en cada dirección. Las regiones de baja densidad están coloreadas en purpura, mientras que las regiones de alta densidad son rojas. Insertos: cada punto en la nube, representa un par único de proteínas (gris). Se realizó un ajuste utilizando el modelo exponencial para los siguientes escenarios: en negro ($1-S_G$*) > 0, en magenta (*$1-S_G$*) < 0.7 y en naranja (*$1-S_G$*)>0.7.*

Tabla 3 Ejemplos de casos de comparaciones de proteínas con baja identidad en secuencia y alta similitud estructural. Se muestra el código PDB, función, organismo, y su clasificación de plegado (CATH código de superfamilia), la identidad en secuencia global, y la semejanza de estructura terciaria (TM-Score).

PDB1	PDB2	S_G	TM-score
2Z2N: Lyase; S. aureus (2.130.10.10)	**4J0W**: RNA Binding Protein; H. sapiens (2.130)	9.30%	0.82
2HES: Biosynthetic Protein; S. cerevisiae (2.130.10.10)	**2Z2N**: Lyase; S. aureus (2.130.10.10)	5.10%	0.83
2Z2N: Lyase; S. aureus (2.130.10.10)	**3FM0**: Biosynthetic Protein; H. sapiens (2.130.10.10)	6.70%	0.83
3VGZ: Protein Binding; E. coli (2.130)	**4J87**: Protein Transport; S. pombe (2.130)	9.60%	0.82
2Z2N: Lyase; S. aureus (2.130.10.10)	**4J87**: Protein Transport; S. pombe (2.130)	7.40%	0.83
3PG0: De Novo Protein; Artificial gene (2.80.10.50)	**4OEE**: Fibroblast growth factor; H. sapiens (2.80)	8.90%	0.81
4DEN: Antiviral Protein; Actinomycete (2.80.10.50)	**4OEE**: Fibroblast growth factor; H. sapiens (2.80)	9.00%	0.83

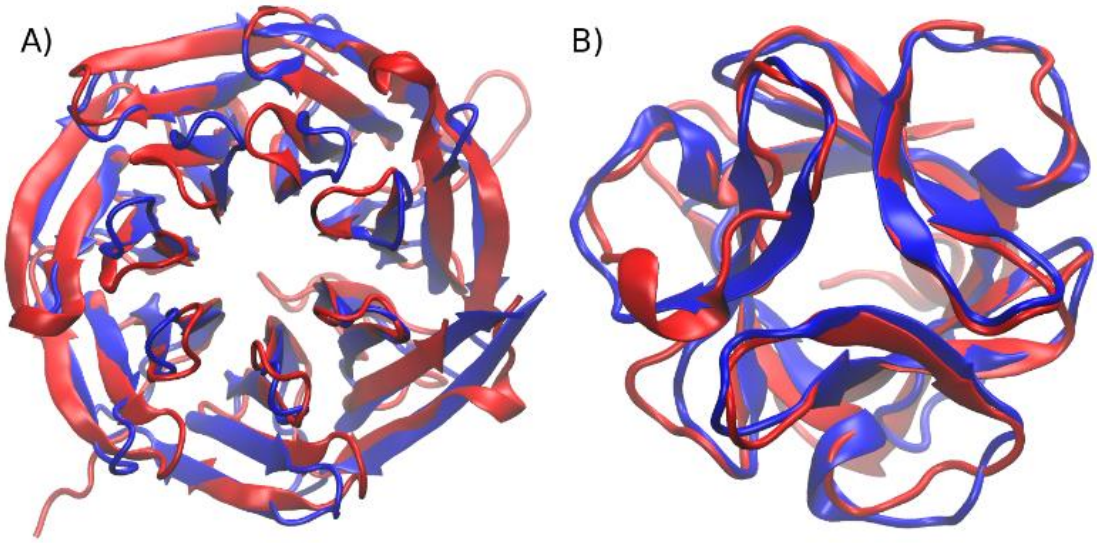


Fig. 5 Superposición de la estructura terciaria de dos casos que poseen muy baja identidad en secuencia y una alta similitud estructural A) azul PDBID:2Z2N y rojo PDBID:4JOW B) azul PDBID:4DEN y rojo PDBID:4OEE.

Para asegurar que los métodos estadísticos que empleamos en este trabajo fueran confiables y que no se introdujera un sesgo artificial en los resultados debido a un desbalance de datos, construimos una muestra de prueba, MonomersRND, seleccionando aleatoriamente un pequeño porcentaje de pares del conjunto de monómeros celulares. El tamaño de este nuevo conjunto de datos es similar al del conjunto de las cápsides virales. La Tabla 4 muestra los resultados del análisis Kolmogorov-Smirnov, prueba estadística diseñada para determinar si dos grupos de datos provienen de la misma distribución de probabilidad. Como era de esperarse, la

prueba indicó que el conjunto de datos MonomersRND sigue la misma distribución que los monómeros celulares. Sin embargo, los otros grupos de datos siguen distribuciones distintas, incluso cuando se comparan monómeros celulares con *n-meros* celulares.

Tabla 4 Valor D de la prueba estadística no paramétrica de Kolmogorov-Smirnov (valor-P). Se estimó de forma independiente para la identidad en secuencia (S_G, diagonal superior) y para la similitud estructural (TM-Score, diagonal inferior). Un valor-P < 0.05 significa que los datos comparados no provienen de la misma distribución.

TM-score \ S_G	Monómeros	Monómeros RND	Dímeros+Oligómeros	VCP
Monómeros	--	0.002 (0.816)*	0.377 (<2.2e-16)	0.569 (<2.2e-16)
Monómeros RND	0.002 (0.601)*	--	0.378 (<2.2e-16)	0.569 (<2.2e-16)
Dímeros+Oligómeros	0.102 (<2.2e-16)	0.102 (<2.2e-16)	--	0.526 (<2.2e-16)
VCP	0.756 (<2.2e-16)	0.756 (<2.2e-16)	0.692 (<2.2e-16)	--

**valor-P* ≥ 0.05

En su trabajo, (Chotia and Lesk, 1986) propusieron un modelo exponencial que ajustaron a los datos que analizaron para describir la relación entre identidad de secuencia y similitud estructural. Aquí, realizamos un ajuste a nuestros datos empleando dicho modelo, expresado como:

$$(1-TMScore) \sim f * e^{k(1-S_G)}$$

La Tabla 5 muestra los valores calculados para los coeficientes de proporcionalidad *f* y *k*, considerando tres escenarios distintos: proteínas homólogas, proteínas no-homólogas y todo el rango de S_G. Las diferencias del ajuste entre los diferentes escenarios y grupos de datos es mostrado en la Fig. 4 (recuadros).

Tabla 5 Estimación de los parámetros f y k para el modelo exponencial usando el algoritmo de Gauss-Newton.

	$(1-S_G) > 0$		$(1-S_G) < 0.7$		$(1-S_G) > 0.7$	
	f	*k*	*f*	*k*	*f*	*k*
Monómeros	0.351	0.768	0.072	1.429	0.393	0.647
Dímeros+Oligómeros	0.130	1.854	0.060	2.392	0.990	1.390
VCP	0.003	5.912	0.009	3.612	0.003	5.900

2.2.2 Distribución estructural de residuos conservados

En primera instancia, calculamos el área de la superficie accesible al solvente de todos los protómeros pertenecientes a los grupos de proteínas celulares y proteínas de

cápside. La Fig. 6 muestra la distribución encontrada para estos valores en cada grupo de datos, así como la correlación entre el área de interfaz y de superficie total. Encontramos una correlación positiva débil en el caso de las proteínas celulares. Lo opuesto se observa en las proteínas de cápside, para las cuales la correlación es fuerte. En promedio, las proteínas celulares tienden a tener un área de interfaz más pequeña al de las proteínas de cápside (Fig. 6). La diferencia más significativa se encontró en la región de interfaz. En el caso de las proteínas de cápside el 74% de la superficie total pertenece a la interfaz, mientras que en los dímeros celulares es sólo de 10% y de 20% para los oligómeros. Este resultado está directamente relacionado con el número de residuos que componen cada una de las categorías estructurales. En promedio, los monómeros y oligómeros celulares poseen el mismo número total de residuos que las proteínas de cápside (Fig. 7). El número total de residuos es significativamente menor para los dímeros celulares (Tabla 7 y Tabla 8). Más de la mitad del número total de residuos de las proteínas de cápsides forman parte de la interfaz, mientras que sólo un porcentaje pequeño del total de residuos de los dímeros y oligómeros celulares forma parte de esta categoría estructural. En el caso de los monómeros, 40% del número total de residuos forman parte del core de la proteína. Este porcentaje se mantiene en los dímeros y oligómeros, pero se reduce a la mitad en las proteínas de cápside. El 60% restante del total de residuos se encuentran en la superficie expuesta al solvente en el caso de los monómeros. Este porcentaje se ve reducido en ~10% en el caso de los dímeros y oligómeros celulares para hacer lugar a la región de interfaz. Las proteínas de cápside tienen un porcentaje de residuos expuestos al solvente menor, en promedio sólo se encuentra el 26% del total de residuos. Conociendo el área y el número de residuos en la superficie expuesta al solvente y en la interfaz, estimamos la distribución de la densidad de residuos de superficie σ. La distribución de la densidad de residuos en la región de interfaz es la misma, en promedio, para las proteínas celulares y de cápside Tabla 7. Sin embargo, σ es significativamente más alta en la región accesible al solvente en el caso de monómeros celulares en comparación con los dímeros y oligómeros. Interesantemente, las proteínas de cápside tienen el mismo valor de σ en la superficie accesible al solvente que los monómeros celulares.

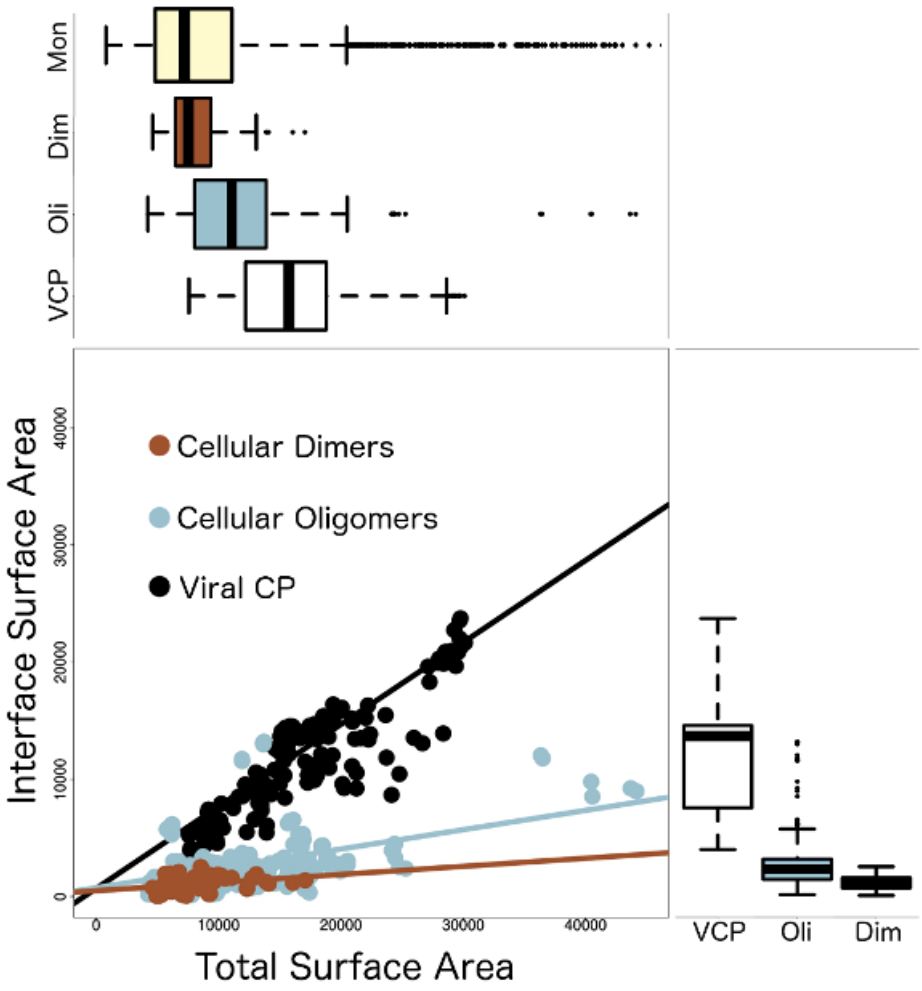


Fig. 6 Correlación entre el área de la interfaz y la superficie total de las proteínas. Se calculó independientemente para monómeros celulares (Mon), dímeros celulares (Dim), oligómeros celulares (Oli), y proteínas de cápside viral (VCP). Para los monómeros celulares únicamente se analizó el área total de la superficie. Los valores del área están en Å².

Tabla 6 Área de superficie promedio <SA> en Å²[porcentaje del área total], y la desviación estándar (SD). Estimaciones realizadas de forma independiente para los monómeros celulares, dímeros celulares, oligómeros celulares y proteínas de cápside viral (VCPs). Correlación de Pearson, r, estimada a partir del área de superficie y la interfaz proteína-proteína. Estimación de la densidad promedio de residuos en la superficie <σ>, para las interfaces proteína-proteína y el área accesible al solvente (SAS).

	Total		Interface		Correlación	<σ>	
	<SA>	SD	<SA>	SD	r	Interface	SAS
Monómeros	9,169	7,110	0 [00%]	0	--	0	16
Dímeros	8,318	2,597	1,073 [13%]	583	0.31	12	12
Oligómeros	11,670	5,411	2,680 [23%]	2,141	0.42	13	13
VCPs	16,535	5,823	12,260 [74%]	4,551	0.90	13	17

Tabla 7 Número promedio de residuos <NR> [porcentaje del total] y desviación estándar (SD) de las proteínas en las diferentes regiones estructurales Estimaciones independientes para los monómeros celulares, dímeros celulares, oligómeros celulares y proteínas de cápside viral.

	Total		Interface		Core		Superficie	
	<NR>	SD	<NR>	SD	<NR>	SD	<NR>	SD
Monómeros	245	135	--	--	97 [40%]	77	148 [60%]	70
Dímeros	178	83	13 [07%]	7	78 [44%]	55	87 [49%]	33
Oligómeros	253	120	35 [14%]	35	102 [40%]	65	116 [46%]	65
VCPs	281	119	158 [56%]	63	51 [18%]	36	72 [26%]	47

Tabla 8 Valores-P de la prueba estadística T para la correlación del número de residuos. Valores bajo la diagonal prueban las diferencias en la densidad de residuos en la superficie accesible al solvente (SAS), mientras que valores arriba de la diagonal prueban las diferencias en la densidad de residuos de la interfaz proteína-proteína. Un valor-P > 0.05 muestra que no existe una diferencia estadística significativa entre los valores promedio (fondo verde: diferencia en medias igual a 0 con un intervalo de confianza del 95%).

SAS \ Interface	Monómeros	Dímeros	Oligómeros	VCPs
Monómeros	NA	0.000	0.000	0.000
Dímeros	3.35E-012	0.722	0.190	0.095
Oligómeros	< 2.2e-16	0.134	0.966	0.983
VCPs	0.213	2.76E-009	2.17E-008	2.06E-008

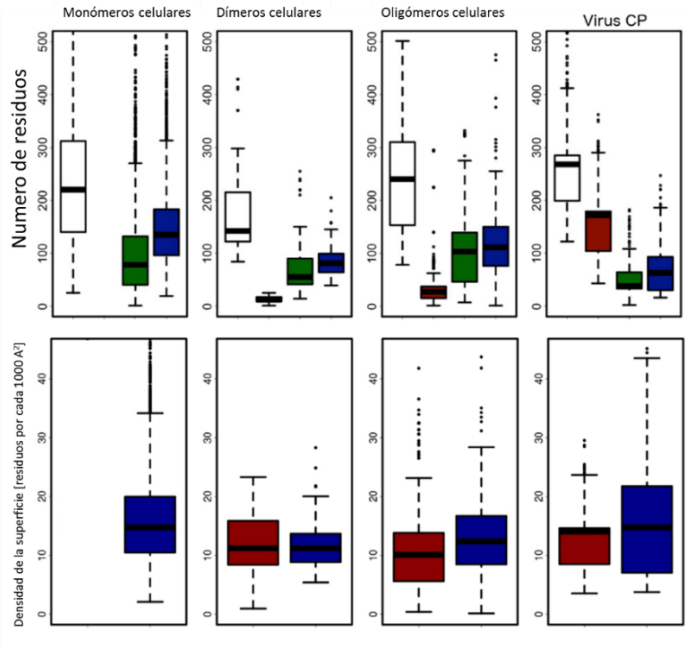


Fig. 7 Distribución del número de residuos (primera fila) y densidad de los residuos de la superficie (segunda fila). Estimadas de manera independiente para los monómeros celulares (primera columna), dímeros celulares (segunda columna), oligómeros celulares (tercera columna), y proteínas de cápside (cuarta columna). El total se muestra con barras blancas, la interfaz proteína-proteína se muestra en rojo, el core en verde y la zona accesible al solvente en azul.

Por otro lado, la identidad en secuencia S_G sólo brinda un panorama general acerca de la similitud entre dos secuencias de aminoácidos mediante la estimación del porcentaje relativo de residuos que son idénticos en el tipo y posición en el alineamiento global. Para investigar en profundidad cómo se encuentran localizados los residuos conservados en la estructura terciaria de la proteína, se analizó su distribución en las diferentes categorías estructurales. Lo anterior se realizó mediante la estimación del índice de identidad en secuencia por categoría estructural, S_k^*, para todos los grupos de datos. La Fig. 8 muestra la correlación y distribución de S_k^* en función de S_G, independientemente calculados para las regiones de interfaz, core y superficie. Los valores promedio se encuentran reportados en la Tabla 9. Ya que conocemos el tamaño de las diferentes categorías estructurales, no es sorprendente que más de la mitad de los residuos conservados de las proteínas de cápside se encuentren en la región de interfaz. En contraste, las proteínas celulares poseen un número pequeño de residuos conservados en en esta región. Por otra parte, aún cuando el número de residuos es mayor en la superficie accesible al solvente comparado con el core, se encontró una cantidad mayor de residuos conservados en el core de la proteína que en la superficie en todos los casos.

Tabla 9 Porcentaje promedio de residuos conservados en las diferentes categorías estructurales <Sk>, desviación estándar SD y correlación de Pearson r con respecto a la identidad en secuencia (S_G). Estimaciones realizadas de manera independiente para n pares de proteínas, en dímeros + oligómeros celulares y proteínas de cápside, para $S_G>0$, $S_G>0.3$, $S_G<0.3$*

	Celular Dimeros + Oligomeros				Viral CP			
	n	<S_k*>	SD	*r*	*n*	<S_k*>	SD	*r*
	INTERFACE							
S_G < 30%	3,832	14	6	-0.64	2,433	42	0.23	0.67
S_G > 30%	59	9	4	0.13	1,453	66	0.12	-0.34
S_G > 0	3,891	13	6	-0.42	3,886	52	0.22	0.5
	CORE							
S_G < 30%	3,832	41	18	-0.04	2,433	23	13	-0.34
S_G > 30%	59	32	1	-0.43	1,453	15	6	0.13
S_G > 0	3,891	41	18	-0.08	3,886	20	11	-0.34

	AREA ACCESIBLE AL SOLVENTE							
S_G < 30%	3,832	30	15	-0.06	2,433	20	11	-0.36
S_G > 30%	59	38	14	0.32	1,453	8	6	0.55
S_G > 0	3,891	30	15	0.05	3,886	12	12	-0.16
	ORPHANS							
S_G < 30%	3,832	15	--	--	2,433	15	--	--
S_G > 30%	59	21	--	--	1,453	11	--	--
S_G > 0	3,891	16	--	--	3,886	16	--	--

Además de considerar el rango completo de S_G, establecimos un umbral del 30% de identidad en secuencia para diferenciar entre proteínas no-homólogas y homólogas. Se observó una correlación débil entre S_k^* y S_G para todos los casos, excepto en la región de interfaz de las proteínas celulares no-homólogas, las cuales presentan una fuerte correlación negativa, y las proteínas de cápside no-homólogas que tienen una fuerte correlación positiva. Existen variaciones en la cantidad promedio de residuos conservados en las diferentes categorías estructurales. Esto se observa tanto para las proteínas celulares como para las proteínas de cápside, sin embargo, el comportamiento entre ambas es diferente. En el caso de los *n-meros* S_k^* muestra valores superiores para proteínas no-homólogas en la interfaz y el core con respecto a las proteínas homólogas. Un comportamiento opuesto se observa en la superficie accesible al solvente. En el caso de las proteínas de cápside, S_k^* es superior para las proteínas no-homólogas en el core y la superficie accesible al solvente con respecto a las proteínas homólogas. Lo opuesto se observa en la interfaz. Encontramos otra diferencia interesante en el comportamiento observado entre *n-meros* celulares y las proteínas de cápside. La cantidad de residuos conservados en la categoría de huérfanos (ORPH) se incrementa para las proteínas homólogas con respecto a las no-homólogas en los *n-meros* celulares. Un comportamiento opuesto se observa en las proteínas de cápside (Tabla 9).

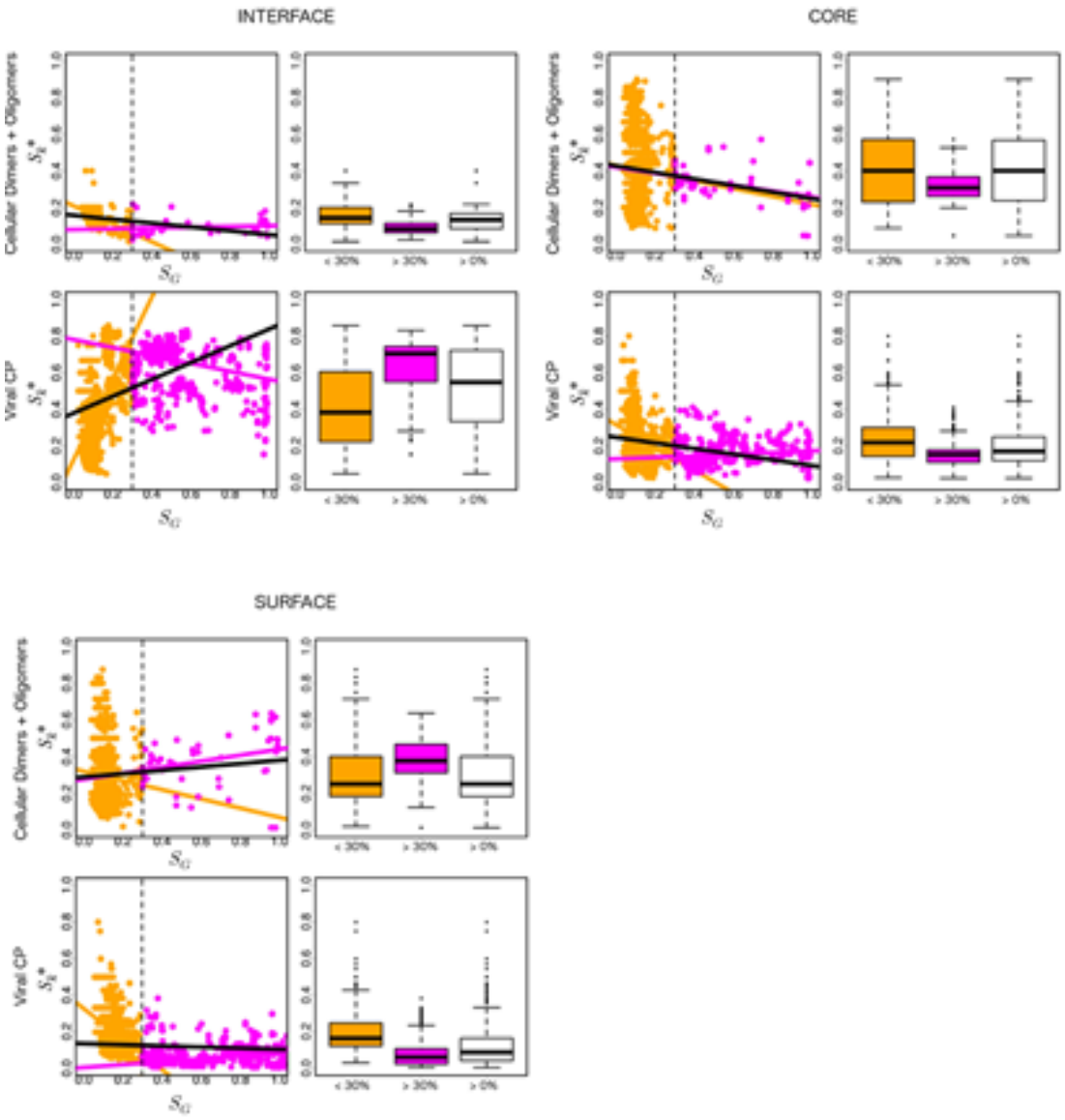


Fig. 8 Correlación entre las regiones conservadas (Sk) y la identidad en secuencia(S_G). Cada punto en la nube representa la comparación entre un par de proteínas. La regresión lineal y el análisis estadístico se muestra para pares con S_G >0 (blanco y negro), S_G >0.3 (magenta) y S_G < 0.3 (naranja).*

2.2.3 Conservación de residuos por categoría estructural

Por último, estimamos la conservación basada en el concepto de entropía de Shannon para cada residuo, $S(i)$, para cada protómero en nuestros grupos de datos. La Tabla 10 muestra el promedio de conservación de residuos <*S*> y el valor normalizado <*s*> calculado por categoría estructural. En el caso de las proteínas celulares, el valor de los

residuos conservados normalizado no cambia significativamente con respecto al valor no normalizado. En general, las regiones de core e interfaz están significativamente más conservadas que la región de la superficie expuesta, observándose mayor preferencia de conservación por el core, especialmente en oligómeros celulares (Tabla 11).

Tabla 10 Conservación de residuos, en promedio, basada en el cálculo de la entropía <S>, normalizada por la conservación de residuos, <s>. La desviación estándar está indicada entre paréntesis.

	Monómeros		Dímeros		Oligómeros		VCPs		
	<s>	<S>	<s>	<S>	<s>	<S>	<s>†	$<S>_{G1}$††	$<S>_{G2}$††
Interface	--	--	0.91 (0.27)	0.98 (0.44)	0.94 (0.14)	0.92 (0.34)	0.96 (0.13)	0.41 (0.17)	0.99 (0.16)
Core	0.78 (0.10)	0.83 (0.31)	0.83 (0.15)	0.84 (0.26)	0.77 (0.09)	0.75 (0.26)	0.75 (0.12)	0.28 (0.09)	0.84 (0.10)
SAS	1.23 (0.10)	1.29 (0.41)	1.26 (0.18)	1.28 (0.37)	1.29 (0.13)	1.23 (0.34)	1.29 (0.12)	0.56 (0.21)	1.24 (0.18)

† Valores considerando una sola distribución de probabilidad. †† Valores considerando dos distribuciones de probabilidad independientes.

Tabla 11 Valores de la prueba estadística T para los resultados obtenidos en la tabla anterior. Comparaciones realizadas entre la interfaz proteína-proteína (I), core (C), y superficie accesible al solvente (S). Un valor > 0.05 muestra que no existen diferencias estadísticas significativas entre los valores promedio comparados (Elementos resaltados en verde: diferencia de medias igual a 0 con un intervalo de confianza del 95%).

	Monómeros		Dímeros		Oligómeros		VCPs		
	<s>	<S>	<s>	<S>	<s>	<S>	<s>	$<S>_{G1}$	$<S>_{G2}$
I vs C	NA	NA	0.1	0.1	1.00E-13	1.00E-03	< 2.00E-16	5.00E-06	2.00E-05
I vs S	NA	NA	2.00E-08	4.00E-03	< 2.00E-16	6.00E-07	< 2.00E-16	7.00E-05	2.00E-07
C vs S	< 2.00E-16	< 2.00E-16	4.00E-16	4.00E-07	< 2.00E-16	7.00E-16	< 2.00E-16	2.00E-13	1.00E-15

Por otra parte, inesperadamente se encontró que la distribución de probabilidad de <*S*> para las proteínas de cápside es bimodal, como se muestra en la Fig. 9. Nombramos estas dos distribuciones G1 y G2. Es interesante observar que el valor promedio de los residuos conservados por categoría estructural en el caso de G2 se comporta de forma similar a las proteínas celulares. Sin embargo, G1 parece estar evolucionando a una velocidad menor. Este resultado es muy intrigante, y amerita un estudio profundo, por lo que se propone como un trabajo futuro. Aún así, realizamos un análisis comparativo preliminar cuyos resultados se presentan en la Tabla 12.

Tabla 12 Familias que pertenecen a cada uno de los grupos G. Se presenta Familia, numero T, Genoma y Hospedero.

G1				G2			
Familia	T	Genoma	Hospedero	Familia	T	Genoma	Hospedero
Adenoviridae	1	dsDNA	Vertebrados	Dicistroviridae	p3	ssRNA(+)	Invertebrados
Birnaviridae	1	dsRNA	Peces, aves e insectos	Picornaviridae	p3	ssRNA(+)	Vertebrados
Bromoviridae	3	ssRNA(+)	Plantas	Siphoviridae	7l	dsDNA	Bacterias y arqueobacterias
Caliciviridae	3	ssRNA(+)	Vertebrados	Sobemoviridae	3	ssRNA(+)	Plantas
Comoviridae	p3	ssRNA(+)	Plantas	Togaviridae	4	ssRNA(+)	Humanos, mamíferos, aves y mosquitos
Hepadnaviridae	4	dsDNA-RT	Humanos, simios y aves	Tombusviridae	3	ssRNA(+)	Plantas
Hepeviridae	1	ssRNA(+)	Humanos, cerdos, jabalíes, monos, roedores y aves	Tymoviridae	3	ssRNA(+)	Plantas
Leviviridae	3	ssRNA(+)	Enterobacterias, caulobacter, pseudomonas, acinetobacter				
Microviridae	1	ssDNA	Enterobacterias y spiroplasmas				
Nodaviridae	3	ssRNA(+)	Vertebrados e Invertebrados				
Parvoviridae	1	ssDNA	Vertebrados e Insectos				
Polyomaviridae	7d	dsDNA	Mamíferos y aves				
Sobemoviridae	3	ssRNA(+)	Plantas				
Tetraviridae	4	ssRNA(+)	Mariposas y polillas				

No existe una correlación aparente con el número T, tipo de genoma, ni tipo de hospedero. Ambos grupos tienen en promedio la misma superficie de interfaz y área expuesta al solvente y la misma cantidad de residuos en la región de la interfaz. Sin embargo, una diferencia significativa se encuentra en el número de residuos que componen el core y la superficie expuesta al solvente, donde G1 posee al menos el doble que G2 (Fig. 10).

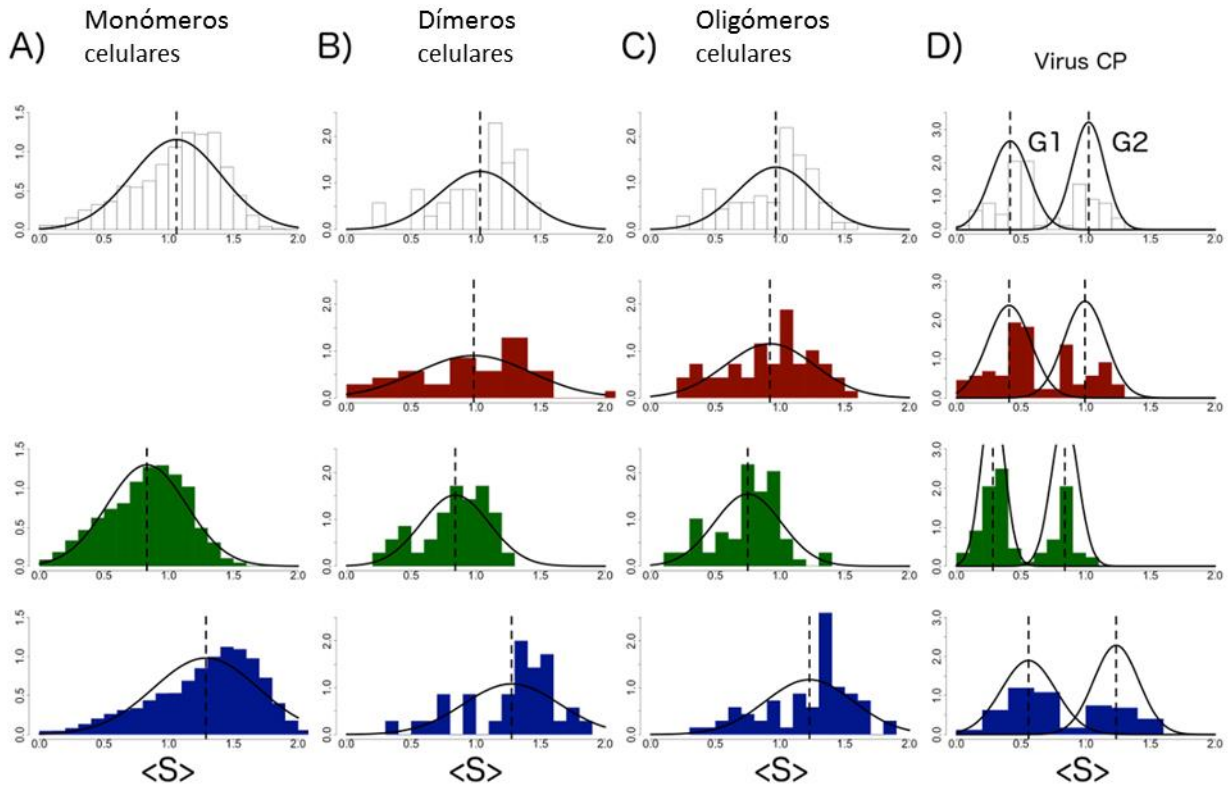


Fig. 9 Distribución de probabilidad de la conservación de residuos basada en la entropía por cadena <S>. Conservación promedio de residuos calculada para toda la proteína (blanco), la interfaz proteína-proteína (rojo), el core (verde), y la superficie accesible al solvente (azul). Se muestra una distribución normal (negro) con la misma media (línea vertical punteada).

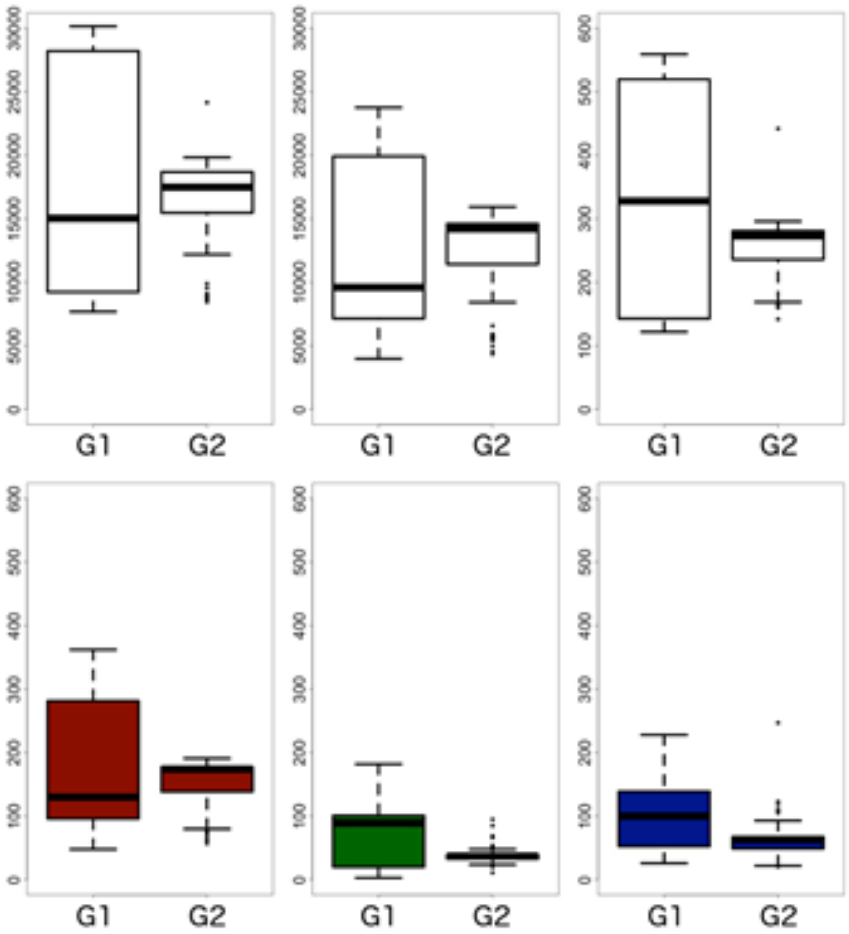


Fig. 10 Comparación estadística entre las familias de virus pertenecientes al grupo 1 y grupo 2 (G1 y G2). Fila superior de izquierda a derecha: caracterización del área total accesible al solvente, área de interfaz, y número de residuos por cadena. La fila inferior muestra el número de residuos en la interfaz proteína-proteína (rojo), core (verde), y área accesible al solvente (azul).

3 Predicción de residuos clave de cápsides esféricas

En el capítulo anterior describimos la existencia de zonas estructurales más conservadas que otras en el contexto de la diversidad estructural de proteínas, encontrando diferencias significativas entre proteínas celulares y VCPs. A continuación integramos dichos resultados con el conocimiento previo de la existencia de un subconjunto de residuos de interfaz VCP-VCP conservados en todos los niveles estructurales. Estos residuos les llamamos residuos clave, los cuales presumimos son esenciales para el correcto auto-ensamblaje de la cápside. Información sobre su localización es importante para lograr entender mejor cómo funciona el mecanismo molecular involucrado en la formación de macro-estructuras complejas. Previamente, los residuos clave sólo habían sido predichos para dos familias de virus; *Bromoviridae* y *Nodaviridae*. En esos casos, la predicción de residuos clave se realizó de manera manual. En este capítulo describimos el desarrollo metodológico para la predicción sistemática de residuos clave de forma automatizada mediante herramientas computacionales. Así mismo, reportamos los resultados de la predicción de residuos clave realizada para todas las familias de virus desnudos icosaédricos conocidos, a través de un análisis masivo (BigData). Parte de los resultados de este capítulo fueron publicados en (Carrillo-Tripp *et al.*, 2015).

3.1 Metodología

3.1.1 Residuos clave

En el capítulo anterior se mostró que la interfaz VCP-VCP contiene la mayoría de los residuos conservados. Aquí utilizaremos la definición de residuo clave propuesta en (Carrillo-Tripp *et al.*, 2008). Esta definición se basa en la existencia de residuos conservados en todos los niveles estructurales de la proteína. Así, los residuos clave se identificaron para todas las familias de virus icosaédricos con información estructural

disponible al momento del estudio. Las secuencias y estructuras de las VCPs fueron tomadas de VIPERdb (Carrillo-Tripp *et al.*, 2009).

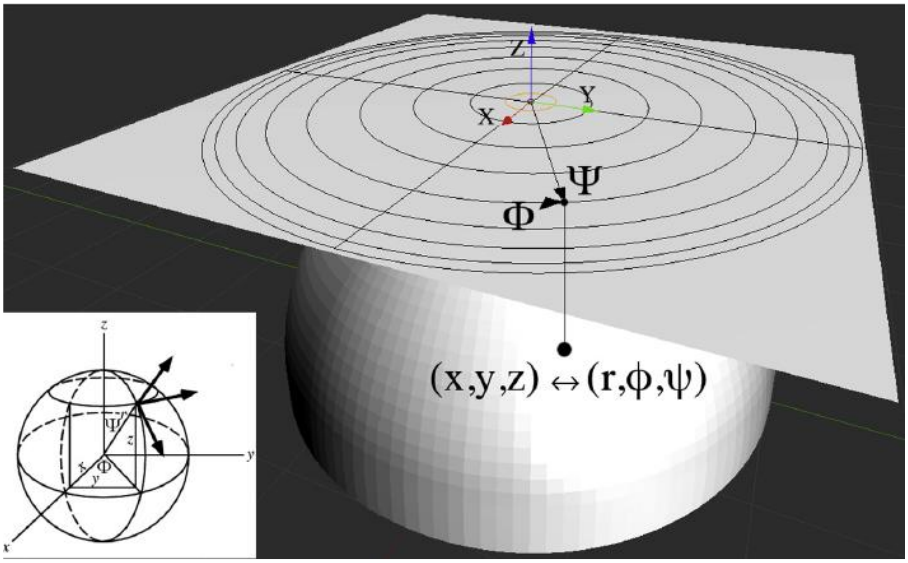


Fig. 11 Transformación del sistema de coordenadas cartesianas a la proyección azimutal polar ortográfica. Se muestra la equivalencia entre el espacio de coordenadas cartesiano (x,y,z) y el espacio de coordenadas esféricas (r,Φ,Ψ) y la posterior proyección sobre el plano colocado sobre el polo norte de la esfera unitaria.

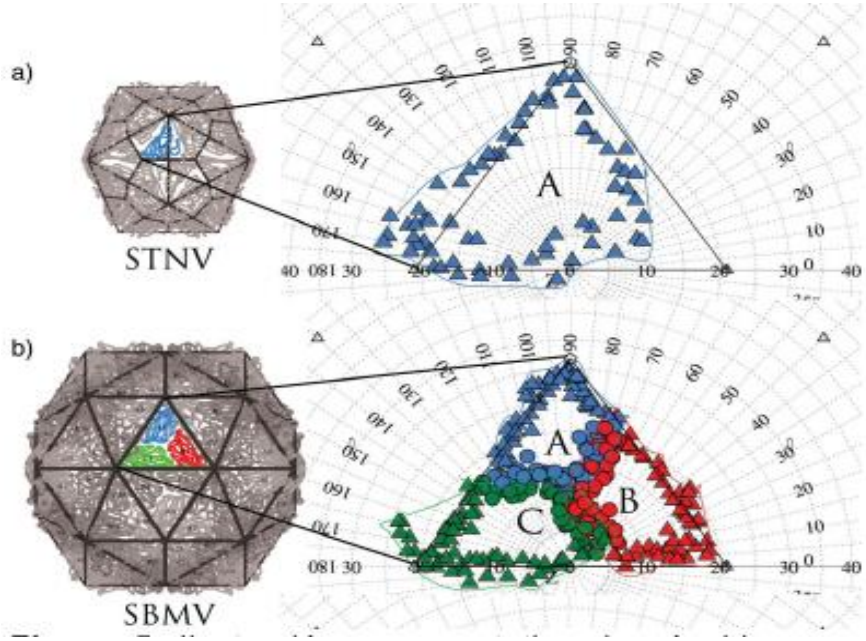


Fig. 12 Ejemplos de CapsidMaps con los que se representan cápsides de distinto número T, (a) T=1 Satellite Tobacco Necrosis Virus, y (b) T=3 Southern Bean Mosaic Virus. Se muestran las diferentes subunidades que conforman la celda unitaria: A (azul), B (rojo), C (verde).

Para identificar los residuos conservados en los cuatro niveles estructurales comunes a un cierto grupo de virus, se generaron mapas de dos dimensiones que llamamos CapsidMaps. Esta metodología fue implementada dentro del portal científico de referencia mundial en el campo de la virología estructural VIPERdb, aprovechando la tecnología ofrecida por Google en términos de librerías web para la generación de mapas geográficos (Carrillo-Tripp *et al.*, 2015). Los CapsidMaps son representaciones de las cápsides utilizando proyecciones sobre una esfera unitaria y mapeadas a un plano tangencial utilizando la transformación azimutal ortográfica (Fig. 11). Dada la naturaleza esférica de las cápsides icosaédricas es posible representar las coordenadas cartesianas (*x, y, z*) del centro de masa de cada residuo en coordenadas esféricas (*r, Φ, Ψ*), donde *r* es la magnitud del vector *R* que define el centro de masa de cada residuo, *Φ* es el ángulo entre el eje *X* y la proyección del vector *R* en el plano *XY*, y *Ψ* es el ángulo entre el eje *Z* y el vector *R*. Para reducir los mapas de tres dimensiones a dos, cada vector *R* es transformado a un vector unitario, dejando todos los residuos sobre la superficie de la esfera de radio uno. De esta manera, se crea un mapa en dos dimensiones de las cápsides icosaédricas (Fig. 12). Esta reducción de dimensiones permite observar de manera simplificada las interacciones cuaternarias entre los residuos de las subunidades que forman la cápside. Para efectos de la búsqueda de conservación estructural se generó un CapsidMap de los residuos de interfaz para cada virus incluido en este estudio. Los CapsidMaps pertenecientes a un mismo género viral (acorde a la clasificación propuesta por el ICTV) fueron superpuestos. De esta manera se identificó un subconjunto de residuos conservados en los cuatro niveles estructurales comunes a todos los miembros de un género particular de virus icosaédricos (residuos clave).

3.1.2 Herramienta computacional para la identificación de residuos clave

Identificar los residuos conservados comunes al superponer un conjunto de mapas bidimencionales es una tarea relativamente simple para un ser humano, ya que uno es capaz de generalizar y establecer un criterio para apreciar cuándo un conjunto de puntos

se encuentran en posiciones lo suficientemente cercanas. Sin embargo, esto no es trivial cuando se busca automatizar esta tarea por medio de un algoritmo computacional. Para que un algoritmo pueda determinar si dos residuos están en la misma posición se utilizan como entrada únicamente los valores numéricos de las posiciones espaciales. Por ejemplo, la posición [2.01, 3, 5] no es la misma que [2.02, 3, 5] para un algoritmo computacional. Por esta razón, desarrollamos un algoritmo que pudiese establecer un criterio de evaluación, similar a como lo hace una persona, i. e., permitiendo un cierto margen de error. Los puntos mencionados anteriormente se deben considerar al momento de desarrollar una herramienta sistemática para la identificación de los residuos clave existentes en el universo estructural de cápsides icosaédricas. El tener una herramienta de este tipo permitirá extender y refinar los resultados de este estudio conforme crezca la cantidad de información disponible en el futuro.

Para desarrollar la herramienta, primero fue necesario traducir los CapsidMaps de su formato visual a una representación matemática (matricial). Como producto de esto, se obtuvo una matriz $M \, x \, L$, donde la dimensión *M* representa el rango del ángulo *Φ* (0-360) y la dimensión *L* representa el rango del ángulo *Ψ* (0-180). Cada elemento de la matriz representa una celda con una superficie uniforme en el espacio *Φ*-*Ψ*. Las matrices producto de los CapsidMaps de todas las cápsides pertenecientes a un mismo género se sobreponen para formar así una matriz concentradora a la que llamaremos *CCM*. Se empleó un algoritmo de convolución que incluye una función de lógica difusa (Este tipo de lógica toma dos valores aleatorios, pero contextualizados y referidos entre sí. Así, por ejemplo, una persona que mida dos metros es claramente una persona alta, si previamente se ha tomado el valor de persona baja y se ha establecido en un metro.) para emular el criterio humano en la identificación de residuos conservados aplicado a la matriz *CCM* Fig. 13.

Fig. 13 Fragmento de la matriz CCM con los residuos pertenecientes al género Alphanodavirus. Las etiquetas en rojo son residuos clave en los que todos ellos se encuentran exactamente en la misma posición, las etiquetas verdes y azules son residuos clave que se encuentran distribuidos en dos o más posiciones cercanas y son detectados por la función difusa. Las etiquetas en negro representan los residuos que no son clave en la interfaz VCP-VCP.

3.2 Resultados

3.2.1 Identificación de residuos clave

Realizamos un análisis sistemático para la identificación de residuos clave utilizando toda la información estructural de cápsides icosaédricas disponible al momento del estudio. Antes de esto, fue necesario establecer valores óptimos para los parámetros de sensibilidad y así evaluar el desempeño de la herramienta. Para evaluar la herramienta de predicción automatizada se procedió a identificar los residuos clave encontrados en la literatura. Primero los identificados en la familia *Nodaviridae* (Carrillo-Tripp *et al.*, 2009), y luego los encontrados en la familia *Bromoviridae* (Díaz-Valle *et al.*, 2014). Al momento que se realizó el presente estudio ya se contaba con 7 estructuras adicionales para la familia *Nodaviridae*, en comparación con las 4 que se usaron en la predicción original. Nuestros resultados son congruentes con el estudio original. Se identificaron 36 residuos de interfaz como residuos clave. Todos los residuos identificados mantienen las mismas características de distribución estructural, i. e., concentrados cerca de los ejes de simetría y localizados en un mismo plano. Con la parametrización obtenida en la identificación de residuos clave de la familia *Nodaviridae* se identificaron los pertenecientes a la familia *Bromoviridae*, la cual está compuesta por dos géneros (*Bromovirus, Cucumovirus*). Encontramos 73 residuos clave para el género *Bromovirus* y 63 para el género *Cucumovirus*. Esto es congruente con la literatura con los resultados previos.

El análisis sistemático de la diversidad estructural de las cápsides icosaédricas permitió la identificación de un total de 1,659 residuos clave, que se encuentran distribuidos a lo largo de 16 familias y 21 géneros. La Tabla 13 muestra la distribución de residuos clave por género.

Tabla 13 Distribución de los 21 grupos de residuos clave encontrados por género. Se muestra la familia y género, el número T, el número de estructuras disponibles para el género (#Estructuras), el número de residuos clave encontrados (#residuos clave), número promedio de residuos por estructura (#Promedio de residuos) y la desviación estándar entre paréntesis.

					#Promedio de residuos
Familia	**Genero**	**Numero T**	**#Estructuras**	**#residuos clave**	**(Desv. Est.)**
Birnaviridae	Avibirnavirus	1	3	101	423.0 (3.56)
Bromoviridae	Bromovirus	3	3	73	487.0 (11.43)
Bromoviridae	Cucumovirus	3	2	63	539.0 (1.00)
Comoviridae	Comovirus	pT3	5	58	551.4 (8.24)
Comoviridae	Nepovirus	pT3	2	3	508.5 (4.50)
Dicistroviridae	Cripavirus	pT3	2	218	820.5 (33.50)
Hepadnaviridae	Orthohepadnavirus	4	2	109	579.0 (9.00)
Hepeviridae	Hepevirus	1	2	75	470.0 (2.00)
Microviridae	Microvirus	1	6	47	824.7 (278.80)
Nodaviridae	Alphanodavirus	3	11	36	990.0 (21.24)
Parvoviridae	Densovirus	1	2	7	413.5 (1.50)
Parvoviridae	Dependovirus	1	15	73	519.3 (1.58)
Parvoviridae	Parvovirus	1	15	4	545.1 (7.74)
Picornaviridae	Aphthovirus	pT3	4	62	673.0 (10.34)
Picornaviridae	Cardiovirus	pT3	4	209	800.2 (20.96)
Picornaviridae	Enterovirus	pT3	27	33	839.9 (19.56)
Polyomaviridae	Polyomavirus	7d	3	337	2119.7 (44.31)
Siphoviridae	Lambda-like viruses	7l	5	75	1927.8 (27.20)
Sobemoviridae	Sobemovirus	3	8	8	604.1 (17.58)
Tombusviridae	Carmovirus	3	2	30	862.5 (60.50)
Tymoviridae	Tymovirus	3	8	38	541.1 (10.09)

Se puede observar que existen géneros en los que se identificó un número atípico de residuos clave, i. e., más de 100 o menos de 10. Los géneros con más de 100 residuos clave son *Avibirnavirus*, *Cripavirus*, *Orthohepadnavirus*, *Cardiovirus*, *Polyomavirus*. A través un análisis puntual, encontramos que el gran número de residuos clave en el género *Avibirnavirus* se debe a que dos de las tres estructuras corresponden al mismo virus (*Bursal Disease Virus*). Esto explica por qué aproximadamente el 25% de los residuos de interfaz en ese género se identificaron como residuos clave. Lo mismo se observa para los géneros *Orthohepadnavirus* (*Human Hepatitis B*) y *Cardiovirus* (*Encephalomyocarditis Virus*). En el caso del género *Cripavirus* existe una estructura para el *Cricket Paralysis Virus* (1b35) y otra para el *Triatoma Virus* (3nap). Ambas

cápsides poseen un S-Score de 0.77, valor indicativo de su alta similitud a nivel cuaternario, razón por la cual el 26% de los residuos de este género fueron identificados como residuos clave. Para el género *Polyomavirus* existen dos estructuras del *Murine Polyoma Virus* y una del *Simian Virus*. Ambos virus tienen un gran parecido a nivel cuaternario con un S-Score de 0.87. Las VCPs poseen un número alto de residuos (2,119 en promedio) por lo que los 337 residuos clave identificados representan sólo el 15% de los residuos. El hecho de que para este género existan dos estructuras del mismo virus no afecta el proceso de identificación de residuos clave. El único requisito para el uso de nuestra herramienta es la existencia de al menos dos estructuras de cápside distintas por género.

De manera similar, es necesario entender por qué hay géneros con menos de 10 residuos clave, como los *Nepovirus*, *Densovirus*, *Parvovirus*,y *Sobemovirus*. En el caso del género *Nepovirus* la estructura del *Fanleaf Virus* está incompleta, razón por la cual se encuentra una baja cantidad de residuos clave. En el caso de los *Densovirus* existen dos estructuras pertenecientes al *Bombyx mori virus* y al *Galleria mellonella virus*, respectivamente, las cuales presentan una baja similitud cuaternaria con un S-Score de 0.38. Esto explica el porqué de la baja cantidad de residuos clave. Un caso similar se observa en los géneros *Parvovirus* y *Sobemovirus*.

3.2.2 Desplazamiento de la maya de referencia en las cápsides diferentes a T=3

Para evaluar la validez de la aproximación que hacen los CapsidMaps al pasar de un espacio de tres dimensiones a dos, se calculó la desviación estándar de la magnitud del vector *R* (*r*) para los residuos clave. La aproximación de la esfera unitaria empleada por los CapsidMaps es buena puesto que las desviaciones estándar de *r* son en su mayoría muy pequeñas Fig. 14. Por otro lado, se observó que existe un desfase en la celda unitaria con respecto al lattice para todos los virus con un número T distinto a 3 Fig. 14. Probablemente se requiera diseñar un lattice particular para cada número T y de esta manera asegurar el alineamiento correcto de la celda unitaria.

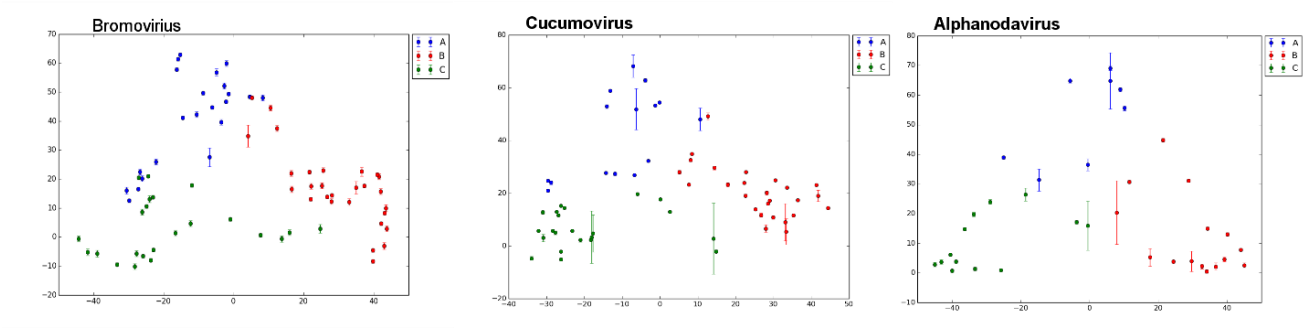


Fig. 14 Análisis de la desviación estándar de los valores de r (barras de error). Los diferentes colores representan cada una de las tres subunidades.

La metodología empleada para identificar los residuos clave requiere que todas las estructuras estén en la misma orientación para generar apropiadamente los CapsidMaps. Pese a los esfuerzos del equipo de trabajo del VIPERdb, aún no es posible hacer esto para todos los virus. Las cápsides con T=3 tienen las mejores propiedades de orientación (Fig. 15), razón por la cual en el siguiente capítulo se analizarán únicamente virus icosaédricos con topología T=3.

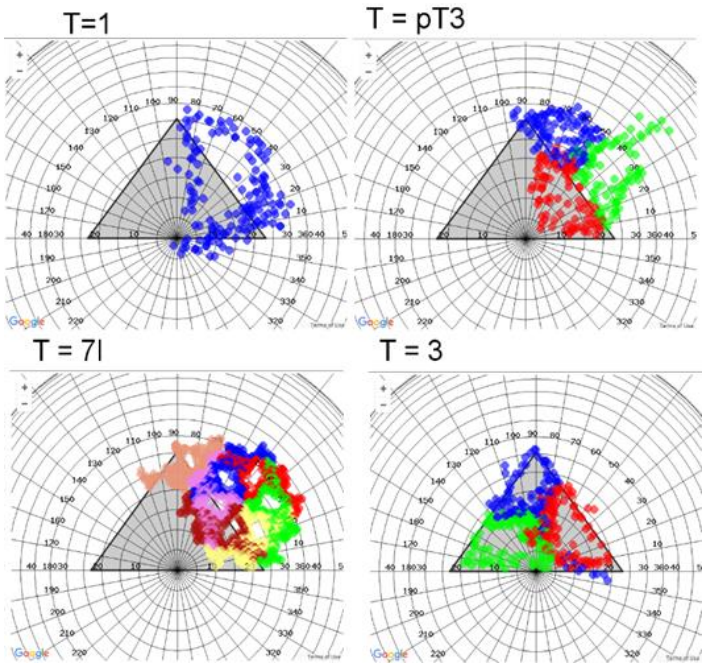


Fig. 15 Desplazamiento de la maya de referencia respecto al número T en la representación de CapsidMaps. La mayoría de los residuos deberían quedar dentro de la referencia geométrica (triángulo negro), tal y como se muestra para el caso de los virus con número T=3. Cada color representa una subunidad independiente en la celda unitaria.

4 Patrones de residuos clave

Los avances en las Ciencias Computacionales en áreas como aprendizaje máquina, minería de datos y búsqueda de patrones, ofrecen alternativas sistemáticas para analizar grandes cantidades de información y transformar esta en conocimiento. En este capítulo se describe el análisis que llevamos a cabo sobre los residuos clave predichos para virus icosaédricos, a través de la aplicación de un algoritmo de búsqueda de patrones. Este algoritmo permite identificar las relaciones existentes en un espacio n-dimensional. Cada residuo clave fue descrito en términos de sus características físico-químicas utilizando un descriptor de diez dimensiones. Encontramos patrones particulares comunes a cada género viral. Esto sugiere la existencia de un mecanismo de auto-ensamblaje común a todos los virus pertenecientes a un género particular, aunque con posibles variaciones entre familias. Los resultados de este capítulo serán publicados en un artículo científico posteriormente (en proceso de escritura).

4.1 Metodología

4.2 Reconocimiento de patrones (PR)

4.2.1 Búsqueda de patrones

Los avances tecnológicos en el campo de la biología estructural han dado como resultado bastas cantidades de información que ha sido almacenada en bases de datos especializadas. Sin embargo, la velocidad con la que se genera la información supera la capacidad para analizarla generalmente, por lo que en las últimas décadas se ha recurrido a técnicas de reconocimiento de patrones y aprendizaje automático desarrolladas en el área de la Inteligencia Artificial de las Ciencias Computacionales. Estas técnicas emplean conceptos matemáticos, estadísticos y computacionales para identificar relaciones complejas entre conjuntos de datos, y ya han probado ser sumamente útiles al ser aplicados en el análisis de datos biológicos (Bioinformática). Muchos de los descubrimientos recientes en Biología se han dado gracias a la

implementación y uso de algoritmos de aprendizaje automático, tales como redes neuronales, métodos de reconocimiento de patrones, y métodos basados en reglas matemáticas, además de métodos probabilísticos como las cadenas ocultas de Markov (Tiwari *et al.*, 1997; Zhang and Nei, 1997; Salz-Berg *et al.*, 1998; Lukashin and Borodovsky, 1998).

4.2.2 Clasificación de datos

Organizar datos en grupos coherentes bajo un cierto criterio es una de las formas fundamentales para la generación de conocimiento cuando se analizan estos. Por ejemplo, los científicos recurren a clasificar los organismos para su entendimiento en grupos organizados por dominio, reino, clase, etc. El reconocimiento de patrones es el desarrollo y estudio de métodos y algoritmos para el agrupamiento de objetos acorde a alguna medida de sus características intrínsecas o similitud. Un ejemplo de reconocimiento de patrones es la clasificación. La clasificación pretende asignar cada valor de entrada a un conjunto dado de clases. Los algoritmos de clasificación se dividen en dos vertientes principales, i. e., supervisados y no-supervisados. Los algoritmos supervisados son aquellos algoritmos en los que se tiene una fase de entrenamiento, esto quiere decir que se le presenta al algoritmo un set de datos para los cuales ya se tiene identificado cuál es la clase correcta a la que pertenece cada elemento. Posterior a la fase de entrenamiento se presentan datos nuevos y estos se agrupan en función del conocimiento adquirido previamente por el algoritmo. En los algoritmos no-supervisados, también conocidos como algoritmos de agrupamiento, no se tiene información previa de la categoría correcta a la que pertenece cada elemento del set de datos. Su funcionamiento se basa en la división del espacio *n* dimensional en *m* partes (grupos). Cada uno de los elementos pertenecientes a un mismo grupo posee una similitud mayor con los miembros de su grupo en relación a miembros de otros grupos.

4.2.3 Algoritmo K-Means

El algoritmo más popular de agrupamiento es conocido como *K-means*. Su nombre está ligado a su historia, y fue independientemente descubierto en diferentes campos científicos (Steinhaus, 1956; Ball and Hall, 1965; MacQueen, 1967; Lloyd, 1982). Aunque el algoritmo fue propuesto por primera vez hace más de 50 años, es hoy uno de los algoritmos más ampliamente usados para realizar agrupamiento. Las razones principales de su éxito son su facilidad de implementación, simplicidad y eficiencia.

Así, sea $X = \{x_i\}, i = 1,..., n$ el conjunto de puntos *n* dimensionales para ser agrupados en un conjunto de *K* grupos, $C = \{c_k, k = 1,..., K\}$. El algoritmo K-means encuentra la partición en la que el error cuadrado es minimizado. Sea μ_k la media del grupo c_k. El error cuadrado entre μ_k y los puntos en el grupo c_k es definido como:

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

El objetivo del algoritmo K-means es minimizar la suma del error cuadrado a través de los *k* grupos,

$$J(c_k) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

Minimizar la función objetivo del K-means es un problema NP-duro cuando *K=2* (Drineas *et al.*, 2004). Esto significa que el K-means sólo converge en mínimos locales. Un estudio subsecuente mostró que existe una alta probabilidad de que el algoritmo converja en el óptimo global cuando los clústeres están claramente separados (Meila, 2006). El algoritmo empieza con una partición inicial de k grupos y asigna patrones a los grupos con el fin de reducir la sumatoria del error cuadrado. Posteriormente genera una partición nueva asignando cada patrón al grupo cuyo centroide sea el más cercano, y así se generan nuevos centroides y se repiten los pasos anteriores hasta encontrar la partición que minimice la suma de errores cuadrados.

4.2.4 Preparación de los residuos clave

Para entender si existe alguna preferencia en la organización de los residuos clave en las interfaces de las subunidades que conforman las cápsides de virus icosaédricos, se realizó un proceso de búsqueda de patrones. Dado el volumen de información, para este fin se empleó el algoritmo de clustering K-means el cual permite identificar patrones no obvios. El set de datos utilizado para este fin está compuesto por todos los juegos de residuos clave de cápsides icosahedricas con topología T=3. En total analizamos 4 familias que conforman 5 grupos de residuos clave (uno por cada género). Un aspecto importante es el hecho de que en los datos analizados se incluyó el caso de la familia *Bromoviridae*, para la cual se identificaron los residuos clave de dos géneros distintos, *Bromovirus* y *Cucumovirus*, los cuales ofrecen un modelo de estudio para los patrones obtenidos inter-género.

4.2.5 Descripción de los residuos clave

Existen múltiples estrategias para describir a los aminoácidos acorde a sus distintas características y propiedades físico-químicas (por ejemplo, hidrófobos, polares, cargados, SASA, etc.). En general, se ha observado que existe una tendencia de conservación acorde a las características físico-químicas (Zimmerman *et al.*, 1968). De esta forma, cada residuo identificado como residuo clave en este trabajo fue descrito tomando en cuenta la posición de su centro de masa en coordenadas cartesianas (x, y, z) y esféricas (r, Φ, Ψ), así como algunas de sus características físico-químicas (hidrofobicidad, polaridad, carga positiva, y carga negativa). Esto implica que por cada residuo clave existen 10 parámetros para describirlo en el espacio multi-dimensional.

4.2.6 Agrupamiento de datos

Antes de realizar la agrupación de los datos se calculó la matriz de distancias. Esta matriz tiene como objetivo identificar la existencia de grupos en el conjunto de datos. La matriz se compone de la distancia euclidiana por pares entre todos los elementos, definida como:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

La diagonal principal siempre será 0. El parámetro más importante para el algoritmo K-means es el número de grupos que se desea encontrar (*k*). Este parámetro debe ser seleccionado de manera que represente adecuadamente los datos para maximizar la eficiencia de la agrupación. No existe una técnica exacta para la determinación del número de grupos que representan un conjunto de datos, por lo que en el marco de este trabajo se utilizó una combinación de dos técnicas: suma de errores cuadrados (SSE) y el coeficiente de Silhouette. La aplicación de esta combinación de técnicas ofrece una buena aproximación del parámetro k óptimo. Para la agrupación y la posterior búsqueda de patrones, se realizó la descripción de todos los residuos predichos como residuos clave organizados por género. Como salida se obtuvieron cinco juegos de grupos, uno por cada género utilizado.

4.2.7 Validación de los grupos de datos

Antes de poder sacar conclusiones acerca de los grupos de datos, es necesario evaluar la calidad de estos. Para determinar la no aleatoriedad de los mismos, así como su importancia, se repitió la aplicación del algoritmo de agrupación k-means 100 veces para cada uno de los cinco grupos de residuos clave (uno por género). Posteriormente, se evaluó la ubicación de los centroides en cada una de las iteraciones, así como la integridad de los grupos a través del algoritmo SSE. Para estimar la importancia y consistencia de cada grupo individual se usó el coeficiente de Silhouette. Los valores del coeficiente de Silhouette se encuentran dentro del rango [-1, 1], donde valores cercanos a -1 significan que los grupos son efecto de la aleatoriedad y los elementos no guardan ninguna relación entre sí. Valores iguales a 0 significan que los grupos están traslapados. Finalmente, valores cercanos a 1 significan que los elementos que componen los grupos son muy similares entre sí.

4.3 Resultados

4.3.1 **Existencia de patrones de residuos clave**

La prueba de la existencia de patrones formados por los residuos clave está dada por la matriz de distancias, la cual es el resultado del cálculo de las distancias euclidianas entre todos los elementos. Reportamos los valores de distancias usando un gradiente de color que va del azul obscuro, para distancias muy cercanas a cero, a rojo, para distancias significativamente grandes. Esta técnica permite observar la existencia de similitudes en el universo de los datos. La aparición de zonas con alto grado de similitud se observa en la Fig. 16. Zonas en color azul guardan alta relación en alguna parte del espacio *n*-dimensional. Por otro lado, se observan zonas que no guardan ninguna relación entre ellas representadas por los bloques en tonos amarillos y rojos. Estos resultados sugieren la existencia de patrones definidos, puesto que se observa la aparición de bloques con alto grado de similitud entre sí, así como bloques con baja similitud. Esto muestra la existencia de patrones asociados a los residuos clave (redes complejas). Como puede observarse en la Fig. 16, este comportamiento está presente en los cinco géneros analizados en este estudio.

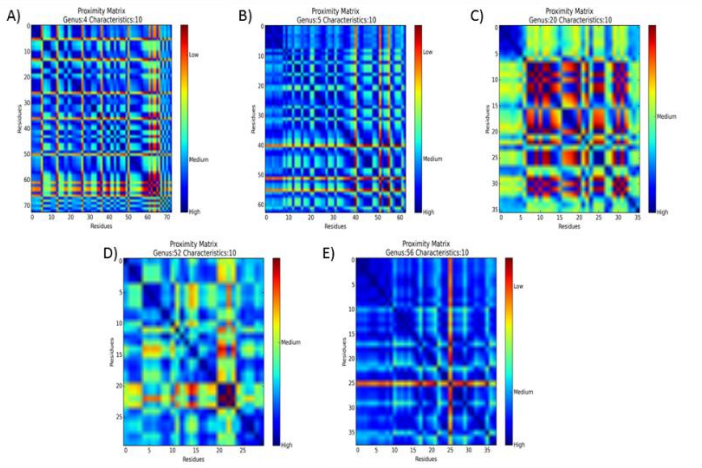


Fig. 16 Representación gráfica de la matriz de distancias para los residuos clave de cada género en el estudio. En color azul se representan zonas de alta similitud, y las regiones en colores amarillo y rojo representan zonas de baja similitud. Esta gráfica muestra una pre-visualización de lo que pudiesen encontrar los algoritmos de agrupación (Genero: (A) 4-Bromovirus [Bromoviridae], (B) 5-Cucumovirus [Bromoviridae], (C) 20-Alphanodavirus [Nodaviridae], (D) 52-Carmovirus [Tombusviridae], (E) 56-Tymovirus [Tymoviridae]).

4.3.2 Número de grupos de datos representativo de los residuos clave

Determinar el valor óptimo para el parámetro k del algoritmo de clasificación es sin duda uno de los retos más grandes en la búsqueda de patrones. En este parámetro radica la calidad de los grupos obtenidos. El mecanismo empleado para determinar este parámetro es generalmente parte de conocimiento *a priori* de los datos. Sin embargo, en nuestro caso no se tienen ningún conocimiento de la cantidad de los grupos representativos para cada grupo de residuos clave. Por esta razón, fue necesario el uso integrado de dos técnicas cuantitativas, i. e., SSE y el coeficiente de Silhouette, para determinar de forma sistemática el valor óptimo de *k*. Los valores para el parámetro k que probamos fueron entre 2 y 14 grupos. Los resultados obtenidos se muestran en la Fig. 17A. En el caso del algoritmo SSE observamos una disminución del error conforme aumenta el número de grupos, adquiriendo un comportamiento asintótico a partir de 4 grupos. Conforme continúa aumentando el número de grupos, continúa disminuyendo el valor del SSE hasta llegar a 0, cuando el número de grupos es igual al número de puntos. A partir de k=4, la disminución de SSE conforme se aumenta el número de grupos se reduce. Para entender la significancia estadística de cada uno de los valores de k evaluados, se calculó el coeficiente de Silhouette para cada valor de k empleado, teniendo como resultado lo mostrado en la Fig. 17B. Se puede observar que los valores de k para los cuales los grupos formados tienen mayor significancia para el conjunto de datos es con k=4. Combinando los resultados del SSE y los coeficientes de Silhouette se agruparon todos los residuos clave utilizando un valor de k=4 como parámetro para el K-means, valor que produce el menor error y la mejor calidad de los grupos.

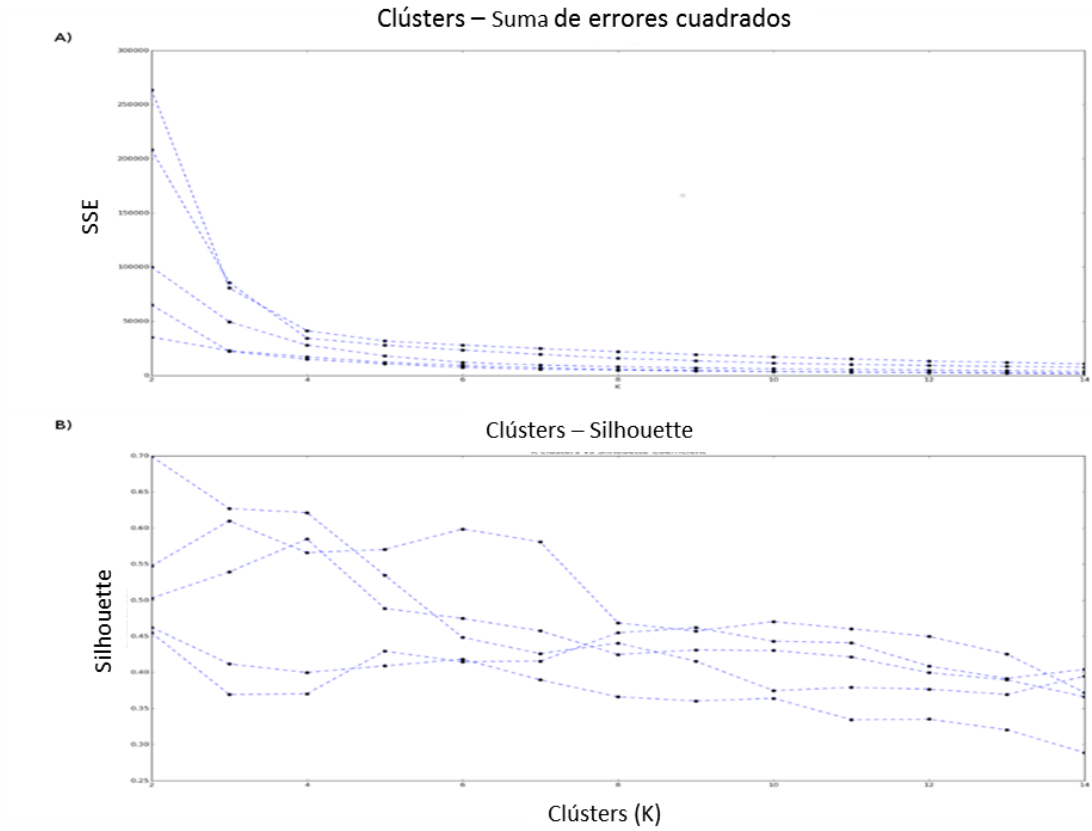
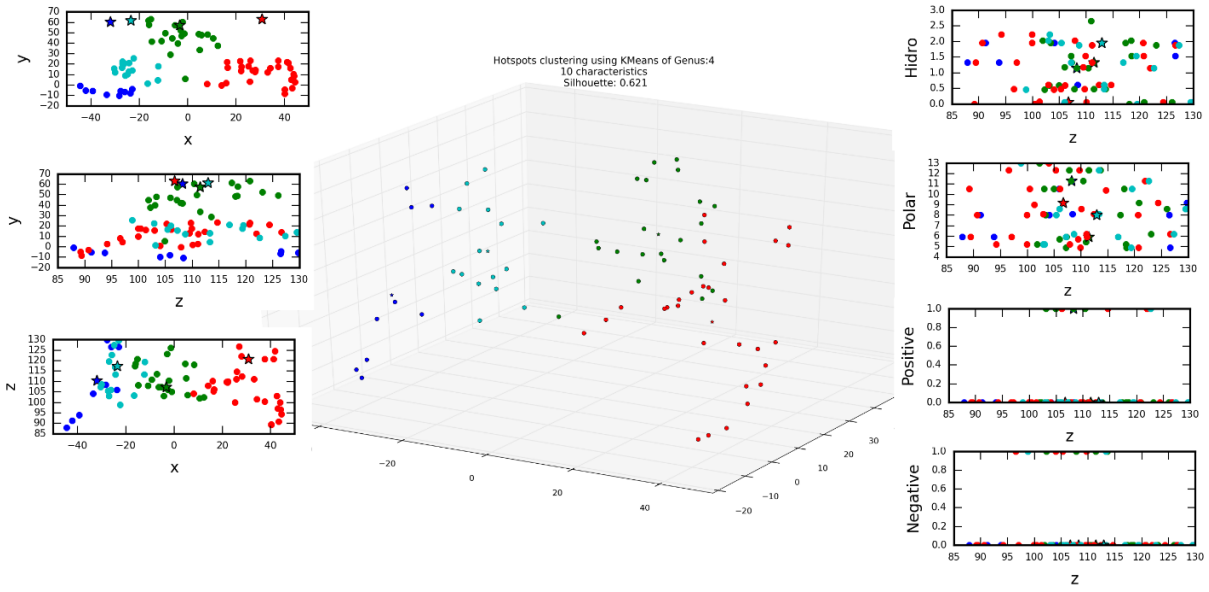


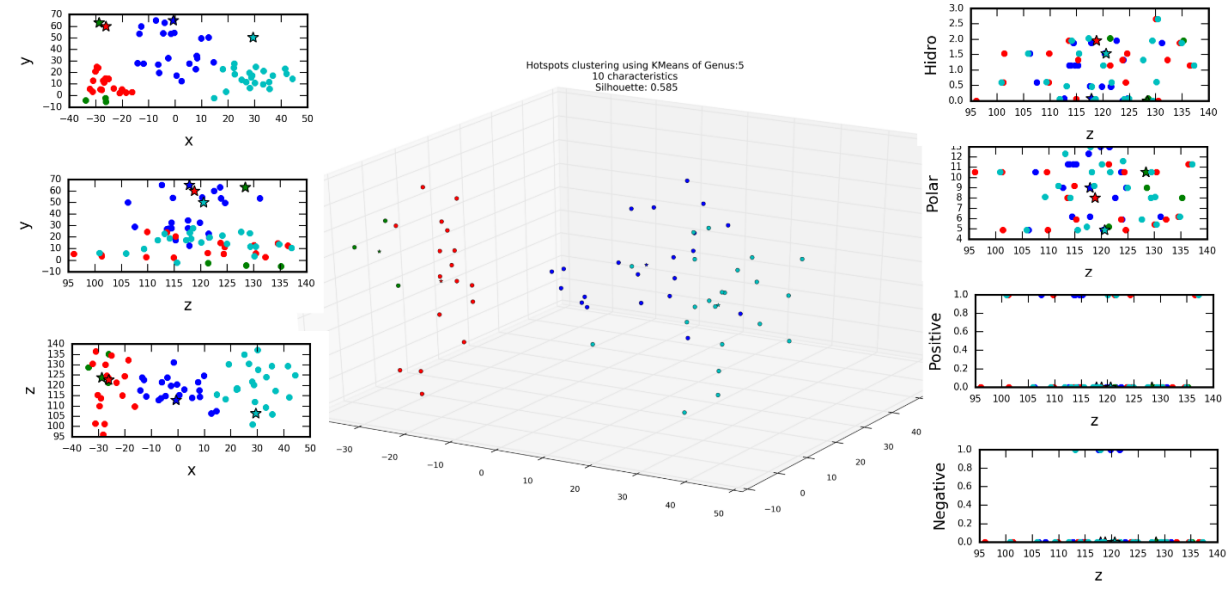
Fig. 17 Análisis del número de grupos representativo de los residuos clave. En ambas gráficas el eje X representa el número de grupos, mientras que el eje Y muestra la suma de errores cuadrados (A), y el coeficiente de Silhouette (B).

Todos los virus icosaédricos incluidos en esta parte del análisis son de topología T=3. Esto significa que se encuentran tres proteínas químicamente idénticas dentro de la celda unitaria que describe la cápside (en el marco del presente estudio cada VCP perteneciente a la celda unitaria se distinguirá con las letras A, B y C respectivamente). 60 copias de la celda unitaria forman la cápside completa cuando se realizan transformaciones de rotación y traslación. Los resultados del proceso de agrupación se pueden apreciar en la Fig. 18. Como se mencionó anteriormente, el valor utilizado para el parámetro k para formar los grupos fue 4, por lo que en la figura aparecen cuatro colores (rojo, azul, verde y magenta), cada uno de los cuales identifica a cada grupo encontrado. La asignación de los colores a cada grupo no tiene ningún significado más allá de distinguir un grupo de otro, y no guardan ninguna relación entre un género y otro. Para todos los géneros virales estudiados se encontraron patrones característicos particulares a cada uno de ellos. Se observa que los patrones están organizados en

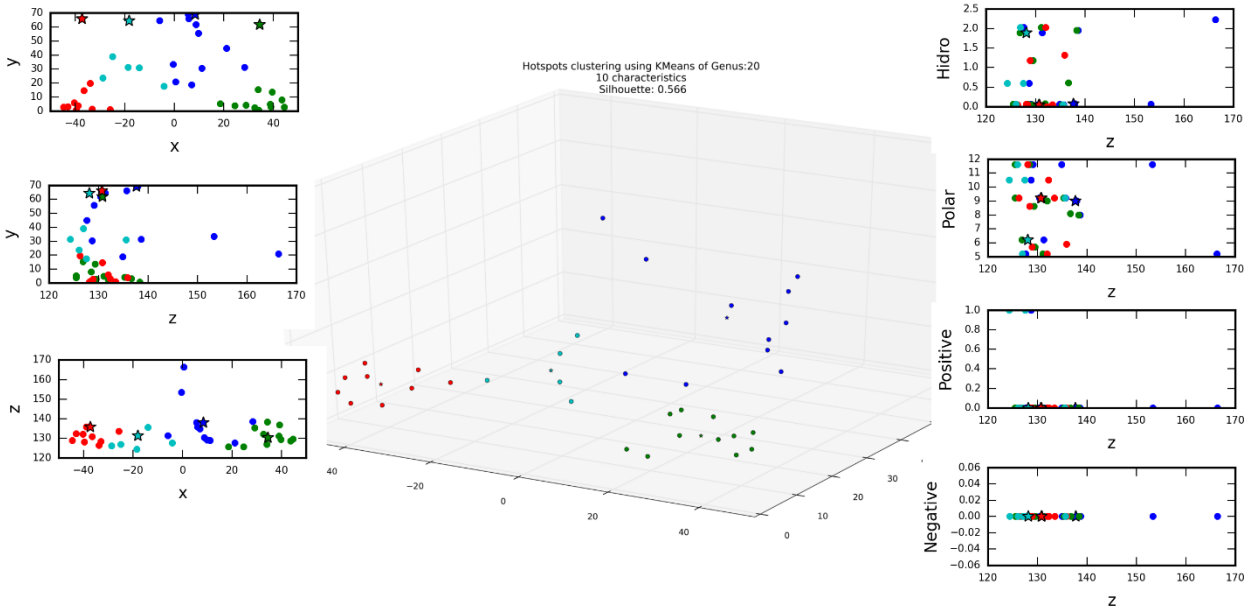
zonas estructurales bien definidas, específicamente cerca de los ejes de simetría de la cápside. Para el género *Bromovirus* la interfaz de la subunidad A queda agrupada en dos grupos: en verde la interfaz AB y en cian la interfaz AC. Los clústeres rojo y azul son fragmentos de la subunidad B y C, probablemente encargados de iniciar la formación de los ejes de simetría. El grupo azul muestra una tendencia hacia la hidrofobicidad y no posee residuos con carga. El resto de los grupos muestran una tendencia hacia la polaridad y poseen algunos residuos con carga tanto positiva como negativa. El género *Cucumovirus* resalta la importancia del eje de cuasi-simetría, como puede observarse en el grupo azul. Los grupos rojo, verde y cian se forman cerca de los ejes de simetría. Es interesante observar que en el caso del género *Bromovirus* las interacciones AB y AC se encuentran en distintos grupos. Especulamos que esto se debe a que dichas interfaces se forman en distintos instantes de tiempo, caso contrario a lo observado en el género *Cucumovirus* donde las interacciones entre las subunides A, B y C se encuentran en un mismo grupo. Los residuos clave del género *Alphanodavirus* exhiben un comportamiento distinto al resto de los géneros estudiados. Como ya se había reportado, los residuos clave se encuentran localizados sobre un mismo plano (Carrillo-Tripp *et al.*, 2008). Es necesario realizar estudios futuros para entender cuál es la implicación de este comportamiento y si esto se relaciona con el proceso de auto-ensamblaje. El género *Carmovirus* agrupa las interacciones entre las subunidades en dos clústeres, rojo y azul respectivamente. Los clústeres verde y cian se encuentran en la interacción BC.



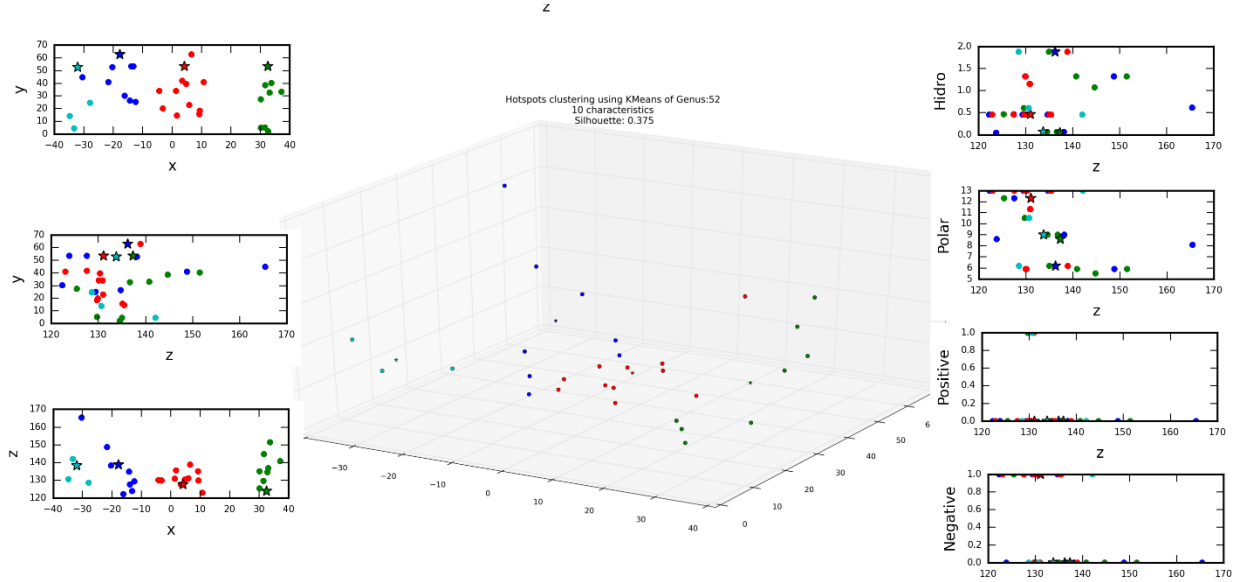
a) *Bromoviridae*



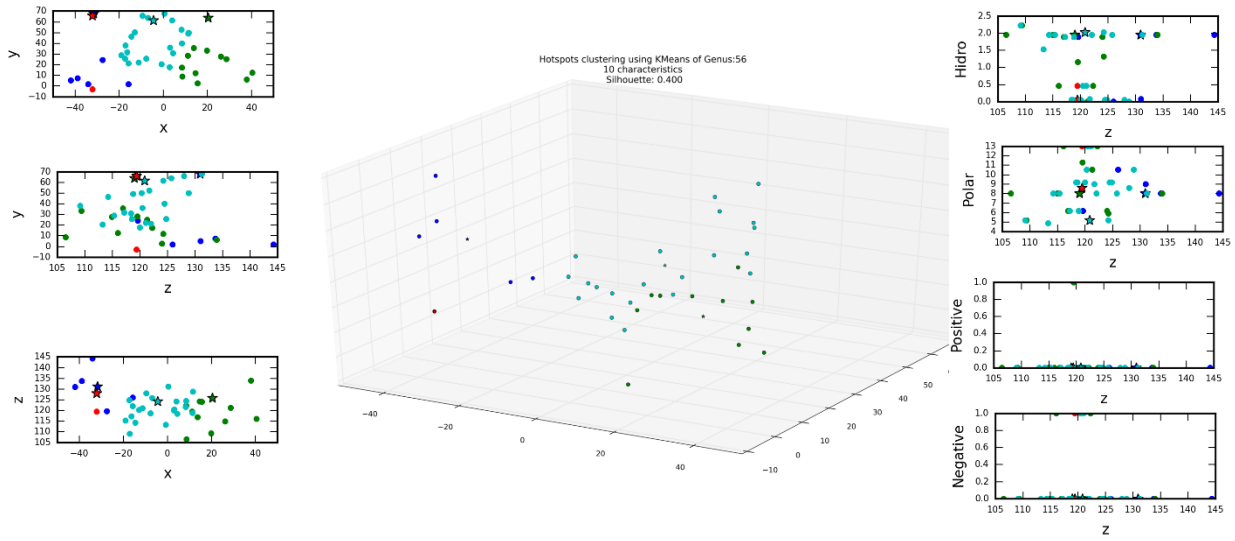
b) *Cucumovirus*



c) *Alphanodavirus*



d) *Carmovirus*



e) *Tymovirus*

Fig. 18 Patrones de residuos clave para los géneros estudiados. En cada panel se muestra la ubicación espacial de los grupos de datos (clústers) particulares a cada género y la correlación por pares de las características más relevantes. Los colores solo representan la agrupación de los datos en un clúster particular, su significado se limita a distinguir un clúster de otro. Insets: Visualización de los clúster para un par de variables dado (las estrellas representan los centroides de cada clúster)

5 Discusión

En el presente trabajo reportamos hallazgos que resaltan las diferencias entre las proteínas celulares y las proteínas de cápside con respecto a la ubicación, cantidad y nivel de conservación de los residuos en la estructura terciaria. Analizamos y comparamos cuatro grupos de datos: monómeros celulares, dímeros celulares, oligómeros celulares y proteínas de cápside icosaédrica. Desde una perspectiva general, los monómeros celulares y las proteínas de cápside parecen ser dos extremos en el espacio a nivel cuaternario al que está sujeta la diversidad estructural encontrada en la naturaleza. Los dímeros celulares y los oligómeros son estados intermedios. Estos resultados se encuentran resumidos en la Fig. 19.

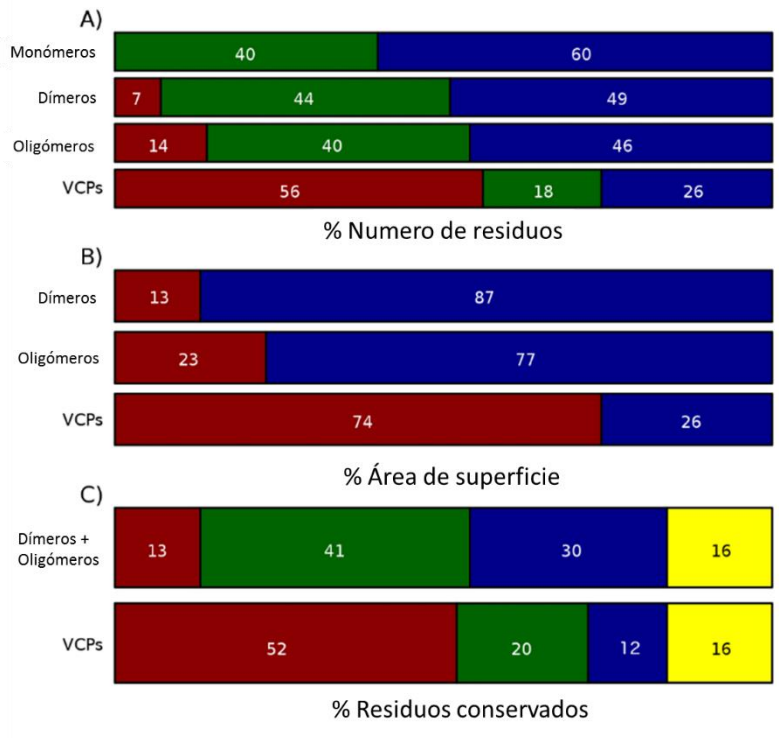


Fig. 19 Porcentaje del número promedio de residuos (A), porcentaje del área de superficie (B), porcentaje de residuos conservados (C), pertenecientes a cada categoría estructural. En rojo se muestra la interfaz proteína-proteína, en verde el core de la proteína, en azul la superficie accesible al solvente, y en amarillo los huérfanos.

Encontramos que la correlación entre la identidad en secuencia y la conservación estructural terciaria sigue un comportamiento exponencial en todos los casos, concomitante con el observado por (Chotia and Lesk 1986). Sin embargo, se observan variaciones en la distribución en el espacio (S_G, TM -score). Los monómeros celulares presentan una alta concentración en la zona de proteínas no-homologas con diferente plegado. Por otra parte, las proteínas de cápside se encuentran distribuidas en tres regiones principales que abarcan todo el rango de homología en secuencia y similitud estructural. Dímeros y oligómeros celulares presentan un comportamiento similar a los monómeros, sin embargo, la forma de su distribución sugiere que son un intermediario entre los monómeros celulares y las proteínas de cápside icosaédrica.

Adicionalmente, realizamos un análisis detallado para entender mejor la correlación entre la identidad en secuencia global y la similitud estructural para el caso de las proteínas de cápside. Observamos que la mayoría de los pares que tienen valores de TM-Score > 0.7 son proteínas de cápside que pertenecen a la misma familia. Por el contrario, los pares que tienen valores de TM-Score < 0.7 pertenecen en su mayoría a comparaciones entre miembros de diferentes familias. Es interesante señalar que la clasificación taxonómica propuesta por el ICTV no incluye información estructural explícita como criterio para la clasificación de los virus. Aun así, nuestros resultados son congruentes con dicha clasificación. Una observación que despertó nuestro interés es la existencia de una cantidad considerable de proteínas de cápside cuyos valores de TM-Score se encuentran entre 0.5 y 0.7. Estos pares pertenecen a comparaciones inter-familia que poseen una identidad en secuencia global por debajo del 20%. Esta observación sugiere una divergencia importante durante la evolución de los virus. Especulamos que las cápsides de virus icosaédricos han evolucionado a partir de un conjunto pequeño de proteínas estructurales ancestrales. Una hipótesis similar resultado de un análisis evolutivo extenso fue propuesta recientemente por (Nasir and Caetano-Anolles, 2015).

Aún cuando algunas de las comparaciones fueron realizadas entre proteínas de cápside de virus miembros del mismo género con identidades en secuencia cercanas al 90%, estos virus muestran diferentes propiedades virológicas/serológicas. Aún más

importante, virus miembros de la misma familia pueden causar enfermedades distintas (por ejemplo, polio vs. resfriado común). Esta es la razón por la que las comparaciones entre proteínas homólogas de cápside consideradas en este estudio son apropiadas.

Es importante resaltar que a pesar de la divergencia en secuencia, el plegado de las VCPs persiste (e. g., jelly-roll B-barrel) y la gran mayoría de los residuos conservados permanecen en las interfaces VCP-VCP. Esta observación es consistente con estudios previos realizados en los que se emplean diferentes enfoques de análisis y se resalta la evolución y la conservación de las estructuras. Por ejemplo, (Abroi and Gough, 2011) propusieron que la viroesfera podría ser el motor de la génesis de las estructuras de proteínas celulares. (Cheng and Brooks III, 2013) mencionan que la relación estructural, y no la funcional, encontrada en algunas clases de proteínas celulares modernas fue el resultado de interacciones genéticas ancestrales entre los virus y sus hospederos. En este sentido, nuestros resultados contribuyen como un indicador más que las VCP son un buen modelo para investigar la co-evolución de la secuencia y la estructura terciaria/cuaternaria en un contexto de alta presión de selección.

Encontramos que la distribución de los residuos conservados en la estructura terciaria es diferente entre las proteínas celulares y las proteínas de cápside. En promedio, 41% de los residuos conservados en los dímeros y oligómeros celulares se encuentran en el core, seguidos de un 30% en la superficie expuesta al solvente, y un 13% en la interfaz. En el caso de las proteínas de cápside, más de la mitad de los residuos conservados se localizan en la interfaz, seguidos de un 20% en el core y solo el 12% en la superficie expuesta al solvente. Cerca de un 15% de los residuos conservados no pudieron ser ubicados en ninguna categoría estructural (orphans) en todos los casos. Observamos que las diferencias en la localización de los residuos conservados entre las proteínas celulares y de cápside están directamente relacionadas con el número de residuos que componen las diferentes categorías estructurales definidas en este trabajo. Esta observación sugiere que los residuos conservados están distribuidos de forma homogénea a lo largo de toda la estructura de las proteínas en todos los casos. Desde un punto de vista global, nuestros resultados indican que los dímeros y oligómeros celulares poseen características estructurales de composición, distribución y variación intermedias entre las observadas en monómeros celulares y proteínas de cápside.

Debido a que las VCP requieren auto-ensamblarse para formar las cápsides virales, ¾ de la superficie está dedicada a formar la región de la interfaz, con la mitad del total de residuos conservados en dicha región. Estos hallazgos concuerdan con evidencia observada con anterioridad. Por ejemplo, hay evidencia que las mutaciones en la VCP tienen lugar preferentemente en la superficie accesible por el solvente, seguramente como un mecanismo para contrarrestar la respuesta inmunológica de la célula infectada (Jameson *et al.*, 1985; Kanda *et al.*, 1986; Vitiello *et al.*, 2005), mostrando una selección positiva en su evolución (Esteves, *et al.* 2008).

La distribución de los residuos conservados en las diferentes zonas de las proteínas celulares es parecida en todos los *n-meros*. En promedio, el core es el más conservado, seguido por la región de la interfaz. Existe una gran variación de secuencia en la superficie expuesta al solvente de todas las proteínas celulares. Este comportamiento se observa también en el caso de las VCP. Adicionalmente, identificamos dos grupos de familias de virus que se comportan de forma distinta en términos de variaciones de secuencia. Uno de estos grupos, G2, tiene valores de conservación de residuos muy similares a los de las proteínas celulares. Sin embargo, el segundo grupo, G1, tiene variaciones menores en la mutación de residuos (Tabla 14), a pesar de que las diferencias relativas entre categorías estructurales se mantienen iguales. De forma preliminar, encontramos que una diferencia significativa entre G1 y G2 es el número de residuos que forman el core y la superficie expuesta al solvente, donde el segundo muestra valores más bajos. (Bahadur and Janin, 2008) analizaron la conservación de residuos de 32 virus icosaédricos y reportaron los valores normalizados de la entropía de Shannon, <*s*>, para la interfaz, core y superficie (0.9, 0.7 y 1.6 respectivamente). Nosotros pudimos reproducir esos mismos resultados cuando consideramos la existencia de una sola distribución de probabilidad de <*S*>, como puede observarse en la Tabla 10. La aproximación del grupo de Janin, y el tamaño reducido de datos que emplearon, imposibilitaron que ellos encontraran las dos distribuciones que nosotros vemos. La razón y el significado de la existencia de estos dos grupos de familias de virus no es obvia, y se requiere realizar más investigación en el futuro.

Tabla 14 Promedio de la conservación de residuos basada en la entropía <S>, y conservación de residuos normalizada <s>. Comparación entre monómeros celulares (M), dímeros celulares (D), oligómeros celulares (O), proteínas de cápside viral grupo 1 (G1), y proteínas de cápside viral grupo 2 (G2). Prueba T de dos colas, valores > 0.05 muestran que no existe diferencia significativa entre las distribuciones de valores promedio comparados (valores resaltados en verde: diferencia de medias igual a 0 con un intervalo de confianza del 95%)

	Promedio por cadena		
	<S>		p-value
M vs D	1.06	1.03	6.34E-01
M vs O	1.06	0.97	1.27E-02
M vs G1	1.06	0.42	< 2.2e-16
M vs G2	1.06	1.02	1.03E-01
D vs O	1.03	0.97	3.09E-01
D vs G1	1.03	0.42	9.55E-14
D vs G2	1.03	1.02	8.54E-01
O vs G1	0.97	0.42	< 2.2e-16
O vs G2	0.97	1.02	1.86E-01
G1 vs G2	0.42	1.02	< 2.2e-16

El criterio de alta resolución que utilizamos para filtrar la información estructural en la construcción de nuestros grupos de datos confiere una alta confiabilidad a los resultados obtenidos. De esta forma, los grupos de datos que analizamos son una muestra representativa de la diversidad de las proteínas existente en la naturaleza. En el presente trabajo incluimos toda la información estructural disponible a la fecha. Estos análisis serán extendidos y refinados en el futuro, conforme aumente la cantidad de información estructural disponible. Por ejemplo, otras topologías de cápsides virales como las helicoidales, no fueron incluidas dado el limitado número de estructuras disponibles, sin embargo pudiesen ser consideradas en estudios posteriores.

El análisis sistemático que realizamos de forma masiva utilizando información relacionada a la estructura de proteínas de cápside con el objetivo de predecir los residuos clave presentó desafíos importantes al momento de diseñar el algoritmo computacional que realizaría esta tarea. La implementación final será descrita en un artículo especializado en ingeniería de software. El reto más importante fue la

implementación de un umbral “inteligente”. Para la solución de este problema se utilizó una técnica llamada lógica difusa. Esta técnica permite generar funciones de membresía que logran determinar el grado de pertenencia de un elemento a un conjunto de datos dado. La aplicación de esta técnica puede servir en muchos otros problemas a los que se enfrentan los biólogos en general. En nuestro caso particular, la técnica permitió determinar de forma confiable si dos residuos están estructuralmente alineados cuando se comparan dos o más CapsidMaps.

Nuestros resultados de la búsqueda de patrones muestran que, en general, los residuos clave no están distribuidos de manera aleatoria en la interfaz VCP-VCP. Aparentemente, las características físico-químicas de dichos residuos juegan un papel importante para la formación y estabilidad de las cápsides icosaédricas, aunque esto amerita estudios experimentales futuros. Es ampliamente aceptado que el auto-ensamblaje de las cápsides virales es un proceso que ocurre de manera cooperativa. Sin embargo, en este proceso deben existir elementos que guíen el ensamblaje. Estos elementos se conocen como interruptores conformacionales (Wood, 1979). Es muy probable que los residuos clave jueguen este papel en el caso de las cápsides, debido a su alto nivel de conservación. La confirmación de que los residuos clave son los interruptores conformacionales que guían el auto-ensamblaje escapa del alcance del presente estudio, y deberá ser analizada con más detalle en futuros estudios.

Por último, el uso de algoritmos del área de la Inteligencia Artificial generó evidencia que sugiere que no existe un mecanismo general de auto-ensamblaje común a todos los géneros de virus icosaédricos. Aparentemente, los mecanismos moleculares son particulares a cada género de virus. Esto es un resultado relevante en varios campos de la ciencia, con aplicaciones potenciales en las áreas biotecnológicas, agrícolas y de salud.

6 Conclusiones

En general, los residuos se encuentran más conservados en el núcleo de las proteínas, seguido por la interfaz proteína-proteína. En la superficie de los complejos proteínicos existe una mayor variación, tanto en el caso de proteínas celulares como proteínas de cápside. Por otro lado, nuestros análisis estadísticos sugieren que existe un grupo de familias de virus que evolucionan más lentamente que el resto de los complejos proteínicos estudiados en este trabajo. Realizamos la predicción de residuos clave en la interfaz VCP-VCP para todas las familias de virus icosaédricos con información estructural disponible. Encontramos patrones específicos formados por los residuos clave, lo que sugiere la existencia de un mecanismo molecular de ensamblaje específico a cada género. Las implicaciones particulares de nuestros resultados se describen a continuación.

6.1 Diferencias estructurales y de conservación entre proteínas de cápside y proteínas celulares

El trabajo realizado en esta área extiende y complementa los resultados previamente reportados. Se detectaron dos grupos de familias virales que evolucionan de manera diferente. El análisis estadístico realizado en las colecciones de datos estructurales de alta resolución permitió resaltar las diferencias importantes entre proteínas celulares y proteínas de cápside. Nuestros resultados sugieren que los monómeros y las proteínas de cápside son dos extremos en el espacio cuaternario de complejos proteínicos en el que los dímeros y oligómeros celulares se encuentran como un estado intermedio.

6.2 Generación de herramientas computacionales

Las herramientas desarrolladas en el presente trabajo fueron elementos indispensables para la obtención de los resultados. Fue necesario desarrollar una plataforma robusta, escalable y eficiente que permitiera el análisis masivo de los datos (BigData), así como la integración rápida de nuevos módulos especializados. Para este fin, se tomó como

base la herramienta CapsidMaps (Carrillo-Tripp *et al.*, 2015), una plataforma computacional diseñada específicamente para la visualización de datos estructurales pertenecientes a cápsides icosaédricas, empleando las librerías de desarrollo web de Google. Para tal fin, aplicamos un proceso de Ingeniería de Software, desarrollando una poderosa herramienta para analizar los oligómeros formados por proteínas de cápside icosaédricas. El núcleo de procesamiento de esta herramienta se convirtió en una pieza invaluable para la identificación posterior de los residuos clave.

Logramos que la herramienta CapsidMaps fuera eficiente en un ambiente remoto (plataforma web). Para esto fue necesario romper la creencia errónea de que un análisis que requiera de cálculos computacionales complejos, o el procesamiento de datos en una cantidad significativa, únicamente se puede optimizar agregando elementos nuevos de hardware. Obtuvimos tiempos menores de cómputo diseñando y escribiendo las herramientas con las instrucciones adecuadas que permiten una integración completa con el hardware disponible de forma automatizada. Aumentamos así la velocidad con la que se realizaron los cálculos computacionales. De esta forma, logramos reducir los tiempos de cómputo de meses a minutos, sin agregar ningún elemento de hardware nuevo. En resumen, las herramientas que desarrollamos permitieron probar la hipótesis aquí planteada en tiempos prácticos.

6.3 Predicción de residuos clave

Realizamos la predicción de residuos clave para todas las familias de virus icosaédricos disponibles en VIPERdb que contaran con al menos dos miembros (especies), utilizando las herramientas computacionales que desarrollamos. Hasta antes de este trabajo sólo se conocía la localización de residuos clave en las familias *Nodaviridae* y *Bromoviridae*. Debido a las limitaciones propias impuestas por el *lattice* de las familias con número T distinto a 3 en la generación de los CapsidMaps, sólo fue posible analizar las cápsides con topología T=3 para la búsqueda de patrones.

6.4 Patrones de residuos clave

Nuestros resultados sugieren la existencia de patrones de residuos clave en las interfaces de las cápsides icosaédricas, particulares a cada género estudiado. Los patrones encontrados parecen no tener preferencia hacia una subunidad en especial. Sin embargo, los residuos clave se concentran cerca de los ejes de simetría de la cápside de forma general, quizás por la importancia que jueguen estas posiciones durante el proceso de auto-ensamblaje. Los patrones que encontramos son distintos entre un género y otro, lo que pudiese sugerir la existencia de más de un mecanismo molecular de auto-ensamblaje. Los algoritmos de clusterización que empleamos mostraron ser una metodología con mucho potencial para el análisis semi-automático de datos biológicos. Es necesario continuar trabajando en la optimización de los protocolos de análisis para hacerla más eficiente.

7 Perspectivas

La realización de este trabajo permitió responder preguntas fundamentales concernientes a la naturaleza estructural de las proteínas en general, y en particular a las proteínas de cápsides icosaédricas. Sin embargo, nuestros resultados generan nuevas ideas y cuestionamientos que podrán ser contestados a través de futuras investigaciones en las siguientes líneas propuestas a continuación.

- Estudio de familias de virus con valores pequeños en variación de secuencia (evolución lenta).
- Utilizar las VCP como modelo para estudiar la co-evolución de la secuencia y la estructura terciaria/cuaternaria en un contexto de alta presión de selección.
- Estudiar el posible rol de los residuos clave como interruptores moleculares.
- Realizar análisis de dinámica molecular para entender el impacto y la función de cada patrón de residuos clave durante el proceso de auto-ensamblaje de la cápside.

8 Bibliografía

Abroi, A, and J Gough. 2011. “Are viruses a source of new protein folds for organisms? - Virosphere structure space and evolution.” *Bioessays* 33: 623-635.

Bahadur, R P, and J Janin. 2008. “Residue conservation in viral capsid assembly Proteins.” *Proteins* 71: 407-414.

Ball, Geoffrey H, and David J Hall. 1965. *ISODATA, A Novel Method of Data Analysis and Pattern Classification.* Menlo: Stanford Research Institute.

Bayer, M.E., Blumberg, B.S., Werner, B., 1968, “Particles associated with Australian antingen in the sera of patients with leukaemia, Down’s Syndrome and hepatitis.” *Nature* 218(5146):1057-9.

Berman, H M, J Westbrook, Z Feng, G T Gilliland, N Bhat, H Weissig, I Shindyalov, y P E Bourne. 2000. “The Protein Data Bank.” *Nucleic Acids Res* 28: 235-242.

Bogan, Andrew A, y Kurt S Thorn. 1998. “Anatomy of Hot Spots in Protein Interfaces.” *J. Mol. Biol.* 280: 1-9.

Caffrey, D R, S Somaroo, J D Hughes, J Mintseris, y E S Huang. 2004. “Are protein–protein interfaces more conserved in sequence than the rest of the protein surface?” *Protein Sci* 13 (1): 190-202.

Carrillo-Tripp, Mauricio, Charles L. Brooks III, and Vijay S. Reddy. 2008. "A novel method to map and compare protein-protein interactions in spherical viral capsids." *Proteins* 73 (3): 644-655.

Carrillo-Tripp, Mauricio, Craig M. Shepherd, Ian A. Borelli, Sangita Venkataraman, Padmaja Natarajan, John E. Johnson, Charles L. Brooks, y Vijay S. Reddy. 2009. “VIPERdb2: an enhanced and web API enabled relational database for structural virology.” *Nucleic Acids Res.* 37 (Database issue): D436-D442.

Carrillo-Tripp, Mauricio, Daniel Jorge Montiel-Garcia, C L Brooks III, y Vijay S Reddy. 2015. “CapsidMaps: protein-protein interaction pattern discovery platform for the structural analysis of virus capsids using Google Maps.” *J Struct Biol.* 47-55.

Carter, John B, y Saunders A. Venetia. 2007. *Virology Principles and Applications.* Trento: John Wiley.

Cheng, S, y C L Brooks III. 2013. “Viral Capsid Proteins Are Segregated in Structural Fold Space.” *PLoS Comput Biol* 9: e1002905.

Chih-Min, C, H Yu-Wen, H Tsun-Tsao, S Chung-Shiuan, y H Jenn-Kang. 2015. “Sequence Conservation, Radial Distance and Packing Density in Spherical Viral Capsids.” *PLoS ONE* 9: e0132234.

Chotia, C, y A.M Lesk. 1986. “The relation between the divergence of sequence and structure in proteins.” *EMBO J* 5: 823-826.

Damodaran, K, Vijay S Reddy, Jhon E Jhonson, y Charles L Brooks III. 2002. “A General Method to Quantify Quasi-equivalence in Icosahedral Viruses.” *J. Mol. Biol.* 324: 723-737.

Diaz-Valle, Armando, Gabriela Chavez-Calvillo, y Mauricio Carrillo-Tripp. 2014. “in silico Binding Free Energy Characterization of Cowpea Chlorotic Mottle Virus Coat Protein Homodimer Variants.” *Advances in Computational Biology* 232: 21-28.

Drineas, P, A Frieze, R Kannan, S Vempala, y V Vinay. 2004. “Clustering Large Graphs via the Singular Value Decomposition.” *Machine Learning* 56: 9-33.

Esteves, P.J, J. Abrantes, M. Carneiro, A Muller, G Thompson, y W Van der Loo. 2008. “Detection of positive selection in the major capsid protein VP60 of the rabbit haemorrhagic disease virus (RHDV).” *Virus Res* 137: 253-256.

Fauquet, C, M.A Mayo, J Maniloff, U Desselberger, y L.A Ball. 2005. *Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses.* Elsevier Academic Press.

Fraser, Cristophe, Christl Donnelly, Simon Cauchemez, William P Hanage, Maria D Van Kerkhove, Deirdre Hollingsworth, Jaime Griffin, y otros. 2009. “Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings.” *Science* 324 (5934): 1557–1561.

Holm, L, y C Sander. 1994. “Structural similarity of plant chitinase and lysozymes from animals and phage. An evolutionary connection.” *FEBS Lett* 340: 129-132.

Jameson, B.A, J Bonin, E Wimmer, y O.M Kew. 1985. “Natural variants of the Sabin type 1 vaccine strain of poliovirus and correlation with a poliovirus neutralization site.” *Virology* 143: 337-341.

Jones, S, y J M Thornton. 1995. “Protein-Protein interactions: a review of protein dimer structures.” *Prog. Biophys. Molec.* 63: 31-65.

Kanda, T, A Furuno, y K Yoshiike. 1986. “Mutation in the VP-1 gene is responsible for the extended host range of a monkey B-lymphotropic papovavirus mutant capable of growing in T-lymphoblastoid cells.” *J. Virol.* 59: 531-534.

Keskin, Ozlem, Buyong Ma, y Ruth Nussinov. 2005. “Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues.” *J. Mol. Biol.* 345: 1281-1294.

Lawrence, M C, y P M Colman. 1993. “Shape complementarity at protein/protein interfaces.” *J. Mol. Biol.* 234: 946-950.

Lloyd, Stuart P. 1982. “Least Squares Quantization in PCM.” *IEEE TRANSACTIONS ON INFORMATION THEORY* 28 (2): 129-137.

Lukashin, Alexander V, y Mark Borodovsky. 1998. “GeneMark.hmm: New solutions for gene finding.” *Nucleic Acids Res.* 26 (4): 1107-1115.

MacQueen, J. 1967. “Some methods for classification and analysis of multivariate observations.” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press. 281-297.

Meila, Marina. 2006. “The uniqueness of a good optimum for K-means.” *ICML '06 Proceedings of the 23rd international conference on Machine learning.* New York: ACM. 625-632.

Murzin, A G, S E Brenner, T Hubbard, y C Chothia. 1995. “SCOP: a structural classification of proteins database for the investigation of sequences and structures.” *J. Mol. Biol.* 536-40.

Nasir, A, y G Caetano-Anolles. 2015. “A phylogenomic data-driven exploration of viral origins and evolution.” *Sci. Adv.* 1: e1500527.

Ofran, Y, y B Rost. 2003. “Analysing Six Types of Protein–Protein Interfaces.” *J. Mol. Biol.* 325: 377-378.

Prasad B.V., Schmid M.F. 2012. “Principles of virus structural organization” *Adv Exp Med Biol.* 726:17-47.

Rost, B. 1999. “Twilight zone of proteins sequence alignments.” *Protein Eng.* 12: 85-94.

Salz-Berg, Steven, David Searls, y Simon Kasif. 1998. *Computational Methods in Molecular Biology.* Amsterdam: Elsevier Press.

Sander, C, y R Schneider. 1991. “Database of homology-derived protein structures and the structural meaning of sequence alignment.” *Proteins: Structure, Function and Genetics* 9: 56-68.

Santi, L., Huang, Z., Mason, H, 2006 “Virus like particles production in green plants” *Methods* 40(1):66-76.

Shepherd, Craig M , Ian A Borelli, Gabriel Lander, Padmaja Natarajan, Vinay Siddavanahalli, Chandrajit Bajaj, John E Johnson, Charles Brooks, and Vijay S Reddy. 2006. "VIPERdb: a relational database for structural virology." *Nucleic acids research* 34 (1): D386-D389.

Shrake, A, y J A Rupley. 1973. “Environment and exposure to solvent of protein atoms Lysozyme and insulin.” *J. Mol. Biol.* 79: 351-364.

Snijders, C.; Matzat, U.; Reips, U.-D. 2012. "'Big Data': Big gaps of knowledge in the field of Internet". *International Journal of Internet Science*. 7: 1–5.

Steinhaus, Hugo. 1956. “Sur la division des corp materiels en parties.” *Bull. Acad. Polon. Sci* 1: 801-804.

Tiwari, S, S Ramachandran, A Bhattacharya, S Bhattacharya, y R Ramaswamy. 1997. “Prediction of probable genes by Fourier analysis of genomic sequences.” *Comput Appl Biosci* 13 (3): 263-70.

Tokuriki, N, C J Oldfield, V N Uversky, I G Berezovsky, y D S Tawfik. 2009. “Do viral proteins possess unique biophysical features?” *Trends in Biochemical Sciences* 34: 53-59.

Touw, W.G, C Baakman, J Black, T.A.H Te Beek, E Krieger, R.P Joosten, y G Vriend. 2015. “A series of PDB related databases for everyday needs.” *Nucleic Acids Research* 43: D364-D368.

Valdar, W.S.J, y J.M Thornton. 2001. “Protein-protein interfaces: Analysis of amino acid conservation in homodimers.” *Proteins: structure,function, and genetics* 42: 108-124.

Vitiello, C.L, C.R Merril, y S Adhya. 2005. “An amino acid substitution in a capsid protein enhances phage survival in mouse circulatory system more than a 1000-fold.” *Virus Res* 114: 101-103.

Webby, Richard J., and Robert G Webster. 2003. "Are We Ready for Pandemic Influenza?" *Science* 1519-1522.

Wood, W. 1979 “Bacteriophage T4 assembly and the morphogenesis of subcellular structure.” *Harvey Lect.* 73:203-213.

Xu, Jinrui, y Yang Zhang. 2010. «How significant is a protein structure similarity with TM-score = 0.5?” *Bioinformatics* 26 (7): 889-895.

Yan, C., F. Wu, D Dobbs, y V Honabar. 2008. “Characterization of Protein–Protein Interfaces.” *Protein* 27: 59-70.

Zhang, J, y M Nei. 1997. “Accuracies of ancestral amino acid sequences inferred by parsimony, likelihood, and distance methods.” *J Mol Evol* 44: 139-146.

Zhang, Y., y J. Skolnick. 2005. “TM-align: a protein structure alignment algorithm based on the TM-score.” *Nucleic acids research* 33 (7): 2302-2309.

Zimmerman, J.M., Eliezer, N., Simha, R. 1968 “The characterization of amino acid sequences in proteins by statistical methods” *Jounrnal of theoretical biology* 21(2):170-201

9 Apéndice

Tabla 15 Dímeros celulares. Se muestra: código PDB, nombre del complejo proteínico, resolución atómica del modelo (Res), número de residuos por subunidad (#aa), y estequiometría del complejo. La clasificación estructural SCOP fue utilizada para agrupar a los dímeros en familias.

PDB	Homo-Dímeros celulares	Res [Å]	#aa	Estequiometría
All alpha				
2IJK	Structure of a rom protein dimer	2	63	A2
1O3U	Crystal structure of an hepn domain protein (tm0613) from thermotoga maritima	2	135	A2
1GLQ	1.8 angstroms molecular structure of mouse liver class pi glutathione s-transferase complexed with s-(p-nitrobenzyl)glutathione and other inhibitors	2	209	A2
3WRP	Flexibility of the dna-binding domains of trp repressor	2	108	A2
1IZM	Structure of ygfb from haemophilus (hi081)	2	189	A2
2O8P	Crystal structure of ywhb- homologue of 4-oxalocrotonate tautomerase	2	61	A2
1ALL	Allophycocyanin	2	160	A2
All beta				
1VH4	Crystal structure of a stabilizer of iron transporter	2	435	A2
1LCL	Crystal structure of human charcot-leyden crystal protein, an eosinophil lysophospholipase	2	142	A2
1W9A	Crystal structure of rv1155 from mycobacterium tuberculosis	2	147	A2
1VHY	Crystal structure of haemophilus influenzae protein hi0303, pfam duf558	2	257	A2
1QOU	Cen (centroradialis) protein from antirrhinum	2	121	A2
1VSC	The crystal structure of an n-terminal two-domain fragment of vascular cell adhesion molecule 1 (vcam-1): a cyclic peptide based on the domain 1 c-d loop can inhibit vcam-1-alpha 4 integrin interaction.	2	196	A2
1ELP	Gamma-d crystallin structure	2	173	A2
1LVE	Structure of the variable domain of human immunoglobulin k-4 light chain len	2	122	A2

1BNC	Three-dimensional structure of the biotin carboxylase subunit of acetyl-coa carboxylase	2	449	A2
Alpha and mainly parallel beta sheets (a/b)				
1LUC	Bacterial luciferase	2	355	A2
1O6D	Structural analysis of a set of proteins resulting from a bacterial genomics project.	2	163	A2
1PDO	Phosphoenolpyruvate-dependent phosphotransferase system	2	135	A2
1VIC	Crystal structure of cmp-kdo synthetase	2	262	A2
1VGT	Crystal structure of 4-diphosphocytidyl-2c-methyl-d-erythritol synthase	2	238	A2
1K3O	Crystal structure analysis of apo glutathione s-transferase	2	221	A2
1O4W	Crystal structure of a pin (pilt n-terminus) domain containing protein (af0591) from archaeoglobus fulgidus	2	147	A2
1VGY	Crystal structure of succinyl diaminopimelate desuccinylase	2	293	A2
1VHX	Crystal structure of putative holliday junction resolvase	2	150	A2
1HUR	Human adp-ribosylation factor 1 complexed with gdp, full length non-myristoylated	2	180	A2
1DPG	Glucose 6-phosphate dehydrogenase from leuconostoc mesenteroides	2	485	A2
1SFL	Ferric soybean leghemoglobin complexed with nicotinate	2	143	A2
1U0S	The crystal structure of the snake venom toxin convulxin	2	135	A2
Alpha and mainly antiparallel beta sheets (a+b)				
3BZT	Crystal structural of the mutated p263a escu c-terminal domain	2	137	A2
1QTO	Crystal structure of a bleomycin resistance determinant from bleomycin-producing streptomyces verticillus	2	122	A2
1QVE	Crystal structure of the truncated k122-4 pilin from pseudomonas aeruginosa	2	126	A2
2Z1E	Crystal structure of hype from thermococcus kodakaraensis	2	338	A2
2NML	Crystal structure of hef2/erh	2	104	A2
1J27	Crystal structure of a hypothetical protein, tt1725, from thermus thermophilus hb8	2	102	A2
1J8B	Structure of ybab from haemophilus influenzae (hi0442), a protein of unknown function	2	112	A2

1AH6	Structure of the tetragonal form of the n-terminal domain of the yeast hsp90 chaperone	2	220	A2
2G3T	Crsytal structure of human spermidine/spermine n1-acetyltransferase (hssat)	2	179	A2
2NWV	Crystal structure of xisi protein-like (yp_323822.1) from anabaena variabilis atcc 29413	2	114	A2
1O50	Crystal structure of a cbs domain-containing protein (tm0935) from thermotoga maritima	2	157	A2
1GFL	Structure of green fluorescent protein	2	238	A2
1IXL	Crystal structure of uncharacterized protein ph1136 from pyrococcus horikoshii	2	131	A2
1DAA	Crystallographic structure of d-amino acid aminotransferase complexed with pyridoxal-5'-phosphate	2	282	A2
2J7Z	Crystal structure of recombinant human stromal cell-derived factor-1alpha	2	68	A2
1VL4	Crystal structure of a putative modulator of dna gyrase (pmba) from thermotoga maritima at 1.95 a resolution reveals a new fold.	2	447	A2
	Hetero-dímeros celulares			
2XKN	Crystal structure of the fab fragment of the anti-egfr antibody 7a7	1	223-216	AB
4TRP	Crystal structure of monoclonal antibody against neuroblastoma associated antigen.	1	214-220	AB
3KDM	Crystal structure of human anti-steroid fab 5f2 in complex with testosterone	2	225-218	AB
2PCD	Structure of protocatechuate 3,4-dioxygenase from pseudomonas aeruginosa at 2.15 angstroms resolution	2	200-238	AB
1HCG	Structure of human des(1-45) factor xa at 2.2 angstroms resolution	2	241-51	AB
1TCR	Murine t-cell antigen receptor 2c clone	3	202-237	AB

Tabla 16 Oligómeros celulares. Se muestra: código PDB, nombre del complejo proteínico, resolución atómica del modelo (Res), estequiometría de la proteína, simetría del complejo, y número de cadenas en la unidad biológica (n-mero).

PDB	Oligómeros celulares	Res [Å]	Estequiometría	Simetría	n-mer
3X30	Crystal structure of metallo-beta-lactamase from thermotoga maritima	2	A3	C3	3

1WDC	Scallop myosin regulatory domain	2	ABC	A	3
1TCO	Ternary complex of a calcineurin a fragment, calcineurin b, fkbp12 and the immunosuppressant drug fk506 (tacrolimus)	3	ABC	A	3
3HPL	Kcsa e71h-f103a mutant in the closed state	3	ABC	A	3
5D98	Influenza c virus rna-dependent rna polymerase - space group p43212	4	ABC	A	3
1UBS	Tryptophan synthase (e.c.4.2.1.20) with a mutation of lys 87->thr in the b subunit and in the presence of ligand l-serine	2	A2B2	C2	4
1GHD	Crystal structure of the glutaryl-7-aminocephalosporanic acid acylase by mad phasing	2	A2B2	C2	4
1SCU	The crystal structure of succinyl-coa synthetase from escherichia coli at 2.5 angstroms resolution	3	A2B2	C2	4
1EFU	Elongation factor complex ef-tu/ef-ts from escherichia coli	3	A2B2	C2	4
1N9P	Crystal structure of the cytoplasmic domain of g-protein activated inward rectifier potassium channel 1	2	A4	C4	4
3E83	Crystal structure of the the open nak channel pore	2	A4	C4	4
1ZSX	Crystal structure of human potassium channel kv beta-subunit (kcnab2)	2	A4	C4	4
1FRP	Crystal structure of fructose-1,6-bisphosphatase complexed with fructose-2,6-bisphosphate, amp and zn2+ at 2.0 angstroms resolution. Aspects of synergism between inhibitors	2	A4	D2	4
1FUQ	Fumarase with bound 3-trimethylsilylsuccinic acid	2	A4	D2	4
1GP1	The refined structure of the selenoenzyme glutathione peroxidase at 0.2-nm resolution	2	A4	D2	4
1IDS	X-ray structure analysis of the iron-dependent superoxide dismutase from mycobacterium tuberculosis at 2.0 angstroms resolutions reveals novel dimer-dimer interactions	2	A4	D2	4
1GPM	Escherichia coli gmp synthetase complexed with amp and pyrophosphate	2	A4	D2	4
1HYH	Crystal structure of l-2-hydroxyisocaproate dehydrogenase from lactobacillus confusus at 2.2 angstroms resolution-an example of strong asymmetry between subunits	2	A4	D2	4
2QKS	Crystal structure of a kir3.1-prokaryotic kir channel chimera	2	A4	C4	4

1MAS	Purine nucleoside hydrolase	3	A4	D2	4
4HBN	Crystal structure of the human hcn4 channel c-terminus carrying the s672r mutation	3	A4	C4	4
4KL1	Hcn4 cnbd in complex with cgmp	3	A4	C4	4
2WLK	Structure of the atp-sensitive inward rectifier potassium channel from magnetospirillum magnetotacticum	3	A4	C4	4
5DEN	The first structure of a full-length mammalian phenylalanine hydroxylase reveals the architecture of an auto-inhibited tetramer	3	A4	D2	4
2X6A	Potassium channel from magnetospirillum magnetotacticum	3	A4	C4	4
3K06	Crystal structure of cng mimicking nak mutant, nak-ntpp, k+ complex	2	A4	C4	4
4PDM	Crystal structure of k+ selective nak mutant in rubidium	2	A4	C4	4
3EAU	Voltage-dependent k+ channel beta subunit in complex with cortisone	2	A4	C4	4
1MLD	Refined structure of mitochondrial malate dehydrogenase from porcine heart and the consensus structure for dicarboxylic acid oxidoreductases	2	A4	D2	4
5A1G	The structure of human mat2a in complex with s-adenosylethionine and ppnp.	2	A4	D2	4
2PTM	Structure and rearrangements in the carboxy-terminal region of spih channels	2	A4	C4	4
4LP8	A novel open-state crystal structure of the prokaryotic inward rectifier kirbac3.1	2	A4	C4	4
1P7B	Crystal structure of an inward rectifier potassium channel	4	A4	C4	4
2WLL	Potassium channel from burkholderia pseudomallei	4	A4	C4	4
3DMM	Crystal structure of the cd8 alpha beta/h-2dd complex	3	ABCD	A	4
3JCF	Cryo-em structure of the magnesium channel cora in the closed symmetric magnesium-bound state	4	A5	C5	5
3VR2	Crystal structure of nucleotide-free a3b3 complex from enterococcus hirae v-atpase [ea3b3]	3	A3B3	C3	6
1ECP	Purine nucleoside phosphorylase	2	A6	D3	6
5CGZ	Crystal structure of galb, the 4-carboxy-2-hydroxymuconate hydratase, from pseuodomonas putida kt2440	2	A6	D3	6

4R9O	Crystal structure of putative aldo/keto reductase from salmonella enterica	2	A6	D3	6
1EXB	Structure of the cytoplasmic beta subunit-t1 assembly of voltage-dependent k channels	2	A4B4	C4	8
1LEH	Leucine dehydrogenase from bacillus sphaericus	2	A8	D4	8
2BL2	The membrane rotor of the v-type atpase from enterococcus hirae	2	A10	C10	10
4KGV	The r state structure of e. coli atcase with atp bound	2	A6B6	D3	12
4FYV	Aspartate transcarbamoylase complexed with dctp	2	A6B6	D3	12
4BJH	Crystal structure of the aquifex reactor complex formed by dihydroorotase (h180a, h232a) with dihydroorotate and aspartate transcarbamoylase with n-(phosphonacetyl)-l-aspartate (pala)	2	A6B6	D3	12
3D6N	Crystal structure of aquifex dihydroorotase activated by aspartate transcarbamoylase	2	A6B6	D3	12
2H3E	Structure of wild-type e. coli aspartate transcarbamoylase in the presence of n-phosphonacetyl-l-isoasparagine at 2.3a resolution	2	A6B6	D3	12
9ATC	Atcase y165f mutant	2	A6B6	D3	12
8ATC	Complex of n-phosphonacetyl-l-aspartate with aspartate carbamoyltransferase. X-ray refinement, analysis of conformational changes and catalytic and allosteric mechanisms	3	A6B6	D3	12
5AA5	Actinobacterial-type nife-hydrogenase from ralstonia eutropha h16 at 2.85 angstrom resolution	3	A6B6	D3	12
2BE9	Crystal structure of the ctp-liganded (t-state) aspartate transcarbamoylase from the extremely thermophilic archaeon sulfolobus acidocaldarius	3	A6B6	D3	12
1PG5	Crystal structure of the unligated (t-state) aspartate transcarbamoylase from the extremely thermophilic archaeon sulfolobus acidocaldarius	3	A6B6	D3	12
1HQ6	Structure of pyruvoyl-dependent histidine decarboxylase at ph 8	3	A6B6	D3	12
3D7S	Crystal structure of wild-type e. Coli asparate transcarbamoylase at ph 8.5 at 2.80 a resolution	3	A6B6	D3	12
1CMK	Crystal structures of the myristylated catalytic subunit of camp-dependent protein kinase reveal open and closed conformations	3	A6B6	D3	12

5FM6	Double-heterohexameric rings of full-length rvb1(adp)rvb2(apo)	3	A6B6	D3	12
2ATC	Crystal and molecular structures of native and ctp-liganded aspartate carbamoyltransferase from escherichia coli	3	A6B6	D3	12
2UV8	Crystal structure of yeast fatty acid synthase with stalled acyl carrier protein at 3.1 angstrom resolution	3	A6B6	D3	12
2V5H	Controlling the storage of nitrogen as arginine: the complex of pii and acetylglutamate kinase from synechococcus elongatus pcc 7942	3	A6B6	D3	12
4USJ	N-acetylglutamate kinase from arabidopsis thaliana in complex with pii from chlamydomonas reinhardtii	3	A6B6	D3	12
2BE7	Crystal structure of the unliganded (t-state) aspartate transcarbamoylase of the psychrophilic bacterium moritella profunda	3	A6B6	D3	12
5FMW	The poly-c9 component of the complement membrane attack complex	2	A22	C22	22
3E1L	Crystal structure of e. coli bacterioferritin (bfr) soaked in phosphate with an alternative conformation of the unoccupied ferroxidase centre (apo-bfr ii).	3	A24	O	24
1IES	Tetragonal crystal structure of native horse spleen ferritin	3	A24	O	24

Tabla 17 Proteínas de cápside viral. Código PDB, nombre y genero del virus, número de residuos por subunidad (#aa), y número T. La taxonomía propuesta por el ICTV fue usada para agrupar los virus en familias y géneros.

PDB	Virus	Genero	#aa	T-Num
FV01: *Adenoviridae*				
1X9P	Human adenovirus 2 penton base	*Mastadenovirus*	569	1
1X9T	Human adenovirus 2 penton base in complex with an ad2 n-terminal fibre peptide	*Mastadenovirus*	569	1
FV02: *Birnaviridae*				
2DF7	Crystal structure of infectious bursal disease virus vp2 sub particle	*Avibirnavirus*	430	1
3IDE	Structure of ipnv subviral particle	*Aquabirnavirus*	428	1
FV03: *Bromoviridae*				

1CWP	Cowpea chlorotic mottle virus (ccmv)	*Bromovirus*	204	3
1JS9	Brome mosaic virus	*Bromovirus*	189	3
1ZA7	The crystal structure of salt stable cowpea cholorotic mottle virus at 2.7 angstroms resolution.	*Bromovirus*	190	3
1F15	Cucumber mosaic virus (cmv)	*Cucumovirus*	218	3
1LAJ	Tomato aspermy virus	*Cucumovirus*	217	3
FV04: *Caliciviridae*				
3M8L	Crystal structure analysis of the feline calicivirus capsid protein	*Calicivirus*	662	3
1IHM	Crystal structure analysis of norwalk virus capsid	*Norovirus*	520	3
2GH8	X-ray structure of a native calicivirus	*Vesivirus*	703	3
FV05: *Comoviridae*				
1BMV	Protein-/rna$ interactions in an icosahedral virus at 3.0 angstroms resolution	*Comovirus*		pT3
1NY7	Cowpea mosaic virus	*Comovirus*	369	pT3
1PGL	Bean pod mottle virus, middle component	*Comovirus*	370	pT3
1PGW	Bean pod mottle virus, top component	*Comovirus*	370	pT3
2BFU	X-ray structure of cpmv top component	*Comovirus*	369	pT3
RCMV	Red clover mottle virus	*Comovirus*	368	pT3
1A6C	Tobacco ringspot virus	*Nepovirus*	513	pT3
2Y7T	X-ray structure of the grapevine fanleaf virus	*Nepovirus*	504	pT3
FV06: *Dicistroviridae*				
1B35	Cricket paralysis virus	*Cripavirus*	282	pT3
FV07: *Hepadnaviridae*				
1QGT	Human hepatitis b viral capsid	*Orthohepadnavirus*	143	4
2G33	Human hepatitis b virus t=4 capsid, strain adyw	*Orthohepadnavirus*	148	4
2G34	Human hepatitis b virus t=4 capsid strain adyw complexed with assembly effector hap1	*Orthohepadnavirus*	147	4

FV08: *Hepevirida*e				
2ZTN	Hepatitis e virus orf2 (genotype 3)	*Hepevirus*	606	1
FV09: *Levivirida*e				
1AQ3	Bacteriophage ms2 mutant (t59s)/rna operator	*Levivirus*	129	3
1AQ4	Bacteriophage ms2 mutant (t45a)/rna operator	*Levivirus*	129	3
1DZS	Bacteriophage ms2/rna hairpin (4one-5)	*Levivirus*	129	3
1E7X	Bacteriophage ms2/rna hairpin (2one-5)	*Levivirus*	129	3
1GKV	Bacteriophage ms2/rna hairpin (c-7)	*Levivirus*	129	3
1GKW	Bacteriophage ms2/rna hairpin (g-10)	*Levivirus*	129	3
1KUO	Bacteriophage ms2/rna hairpin (c-10)	*Levivirus*	129	3
1U1Y	Bacteriophage ms2/rna aptamer f5 (2-aminopurine at -10 pos.)	*Levivirus*	129	3
1ZDH	Bacteriophage ms2/19 nt. Ms2 rna fragment	*Levivirus*	129	3
1ZDI	Bacteriophage ms2/19 nt. Ms2 rna fragment	*Levivirus*	129	3
1ZDJ	Bacteriophage ms2/8 nt. Ms2 rna fragment	*Levivirus*	129	3
1ZDK	Bacteriophage ms2/23 nt. Ms2 rna fragment	*Levivirus*	129	3
2B2D	Rna stemloop operator from bacteriophage qbeta complexed with an n87s,e89k mutant ms2 capsid	*Levivirus*	413	3
2B2E	Rna stemloop from bacteriophage ms2 complexed with an n87s,e89k mutant ms2 capsid	*Levivirus*	416	3
2B2G	Ms2 wild-type rna stemloop complexed with an n87s mutant ms2 capsid	*Levivirus*	416	3
2BU1	Bacteriophage ms2/rna hairpin (5bru-5) complex	*Levivirus*	129	3
2C50	Ms2-rna hairpin (a-5) complex	*Levivirus*	129	3
2C51	Ms2-rna hairpin (g-5) complex	*Levivirus*	129	3
2W4Y	Caulobacter bacteriophage 5 - virus-like particle	*Levivirus*	122	3
2W4Z	Caulobacter bacteriophage 5	*Levivirus*	122	3
5MSF	Bacteriophage ms2/rna aptamer f5	*Levivirus*	129	3

6MSF	Bacteriophage ms2/rna aptamer f6	*Levivirus*	129	3
7MSF	Bacteriophage ms2/rna aptamer f7	*Levivirus*	129	3
FV10: *Microviridae*				
1AL0	Procapsid of bacteriophage phix174	*Microvirus*	421	1
1CD3	Procapsid of bacteriophage phix174	*Microvirus*	426	1
1GFF	Bacteriophage g4	*Microvirus*	426	1
1M06	Bacteriophage α3	*Microvirus*	431	1
1RB8	φ-x 174 dna binding protein j	*Microvirus*	431	1
2BPA	Bacteriophage φ-x 174	*Microvirus*	426	1
FV11: *Nodaviridae*				
1F8V	Pariacoto virus (pav)	*Alphanodavirus*	401	3
1NOV	Nodamura virus	*Alphanodavirus*	375	3
2BBV	Black beetle virus (bbv)	*Alphanodavirus*	379	3
2Q23	Crystal structure of flock house n363t mutant	*Alphanodavirus*	415	3
2Q25	Flock house virus coat protein d75n mutant	*Alphanodavirus*	414	3
2Q26	Fhv virus like particle	*Alphanodavirus*	381	3
3LOB	Crystal structure of flock house virus calcium mutant	*Alphanodavirus*	500	3
4FSJ	Crystal structure of the virus like particle of flock house	*Alphanodavirus*	382	3
4FTB	Crystal structure of the authentic flock house virus particl	*Alphanodavirus*	384	3
4FTE	Crystal structure of the d75n mutant capsid of flock house virus	*Alphanodavirus*	384	3
4FTS	Crystal structure of the n363t mutant of the flock house vir	*Alphanodavirus*	384	3
FV12: *Partitiviridae*				
3ES5	Crystal structure of partitivirus (psv-f)	*Parititivirus*	420	2
3IYM	Backbone trace of the capsid protein dimer of a fungal partitivirus from electron cryomicroscopy and	*Parititivirus*	434	2

Fv13: *parvoviridae*				
3N7X	Crystal structure of penaeus stylirostris densovirus capsid	*Brevidensovirus*	329	1
1DNV	Parvovirus (densovirus) from galleria mellonella	*Densovirus*	436	1
3P0S	Crystal structure of bombyx mori densovirus 1 capsid	*Densovirus*	454	1
1LP3	Adeno-associated virus	*Dependovirus*	598	1
2G8G	Structurally mapping the diverse phenotype of adeno-associated virus serotype 4	*Dependovirus*	734	1
2QA0	Structure of adeno-associated virus serotype 8	*Dependovirus*	827	1
3J1Q	Structure of aav-dj, a retargeted gene therapy vector: cryo-electron microscopy at 4.5a resolution	*Dependovirus*	737	1
3KIC	Crystal structure of adeno-associated virus serotype 3b	*Dependovirus*	736	1
3KIE	Crystal structure of adeno-associated virus serotype 3b	*Dependovirus*	736	1
3NG9	Structure to function correlations for adeno-associated viru 1	*Dependovirus*	736	1
3NTT	Structural insights of adeno-associated virus 5. A gene ther for cystic fibrosis	*Dependovirus*	724	1
3OAH	Structural characterization of the dual glycan binding adeno-associated virus serotype 6	*Dependovirus*	736	1
3RA2	Structural studies of aav8 capsid transitions associated wit endosomal trafficking	*Dependovirus*	738	1
3RA4	Structural studies of aav8 capsid transitions associated wit endosomal trafficking	*Dependovirus*	738	1
3RA8	Structural studies of aav8 capsid transitions associated wit endosomal trafficking	*Dependovirus*	738	1
3RA9	Structural studies of aav8 capsid transitions associated wit endosomal trafficking	*Dependovirus*	738	1
3RAA	Structural studies of aav8 capsid transitions associated with endosomal trafficking	*Dependovirus*	738	1
3TSX	Structure-function analysis of receptor-binding in adeno-ass virus serotype 6 (aav-6)	*Dependovirus*	736	1
3UX1	Structural characterization of adeno-associated virus seroty	*Dependovirus*	736	1
1C8D	Canine panleukopenia virus	*Parvovirus*	584	1
1C8E	Feline panleukopenia virus	*Parvovirus*	584	1
1C8F	Feline panleukopenia virus	*Parvovirus*	584	1
1C8G	Feline panleukopenia virus	*Parvovirus*	584	1

1C8H	Canine parvovirus strain d (ph 5.5)	*Parvovirus*	584	1
1FPV	Feline panleukopenia virus	*Parvovirus*	584	1
1IJS	Canine parvovirus strain d, mutant a300d	*Parvovirus*	584	1
1K3V	Porcine parvovirus	*Parvovirus*	579	1
1MVM	Mice minute virus (mvm), strain i	*Parvovirus*	587	1
1P5W	Canine parvovirus	*Parvovirus*	584	1
1P5Y	Canine parvovirus	*Parvovirus*	584	1
1S58	B19 parvovirus capsid	*Parvovirus*	554	1
1Z14	Structural determinants of tissue tropism and in vivo pathogenicity for the parvovirus minute virus	*Parvovirus*	587	1
2CAS	Canine parvovirus	*Parvovirus*	584	1
4DPV	Canine parvovirus strain d	*Parvovirus*	584	1
Fv14: *picornaviridae*				
1BBT	Foot and mouth disease virus (fmdv)	*Aphthovirus*	220	pT3
1FMD	Foot and mouth disease virus type c-s8c1	*Aphthovirus*	220	pT3
1FOD	Foot and mouth disease virus (reduced)	*Aphthovirus*	220	pT3
1QQP	Foot and mouth disease virus	*Aphthovirus*	220	pT3
1MEC	Mengo encephalomyocarditis virus (ph 4.6)	*Cardiovirus*	274	pT3
1TME	Theiler's murine encephalomyelitis virus (da strain)	*Cardiovirus*	266	pT3
1TMF	Theiler's murine encephalomyelitis virus (bean strain)	*Cardiovirus*	276	pT3
2MEV	Mengo encephalomyocarditis virus	*Cardiovirus*	268	pT3
1AL2	Poliovirus (type 1, mahoney strain) mutant v1160i	*Enterovirus*	302	pT3
1AR6	Poliovirus (type 1, mahoney strain) mutant v1160i, p1095s	*Enterovirus*	302	pT3
1AR7	Poliovirus (type 1, mahoney strain) mutant p1095s, h2142y	*Enterovirus*	302	pT3
1AR8	Poliovirus (type 1, mahoney strain) mutant p1095s	*Enterovirus*	302	pT3

1AR9	Poliovirus (type 1, mahoney strain) mutant h2142y	*Enterovirus*	302	pT3
1ASJ	Poliovirus (type 1, mahoney strain)	*Enterovirus*	302	pT3
1BEV	Bovine enterovirus vg-5-27	*Enterovirus*	281	pT3
1COV	Coxsackievirus b3 coat protein	*Enterovirus*	281	pT3
1D4M	Coxsackievirus a9	*Enterovirus*	284	pT3
1EAH	Poliovirus type2/sch48973 complex	*Enterovirus*	301	pT3
1EV1	Echovirus 1 (farouk strain)	*Enterovirus*	281	pT3
1HXS	Poliovirus (mahoney strain)	*Enterovirus*	302	pT3
1MQT	Swine vesicular disease virus	*Enterovirus*	283	pT3
1OOP	Swine vesicular disease virus	*Enterovirus*	283	pT3
1PIV	Poliovirus (type 3, sabin strain)/win51711 complex	*Enterovirus*	302	pT3
1PO1	Poliovirus (type 1, mahoney strain)/r80633 complex	*Enterovirus*	302	pT3
1PO2	Poliovirus (type 1, mahoney strain)/r77975 complex	*Enterovirus*	302	pT3
1POV	Poliovirus type 1 (mahoney strain) empty capsid	*Enterovirus*	341	pT3
1PVC	Poliovirus type 3 (sabin strain)	*Enterovirus*	302	pT3
1VBA	Poliovirus (type 3, sabin strain)/r78206 complex	*Enterovirus*	302	pT3
1VBB	Poliovirus (type3, sabin strain)/r80633 complex	*Enterovirus*	302	pT3
1VBC	Poliovirus (type3, sabin strain)/r77975 complex	*Enterovirus*	302	pT3
1VBD	Poliovirus (type 1, mahoney strain)/r78206 complex	*Enterovirus*	302	pT3
1VBE	Poliovirus (type 3, sabin strain) mutant f124l, f134l/r78206 complex	*Enterovirus*	302	pT3
1Z7S	The crystal structure of coxsackievirus a21	*Enterovirus*	298	pT3
2PLV	Poliovirus type 1 (mahoney strain)	*Enterovirus*	302	pT3
3EPC	Cryoem structure of poliovirus receptor bound to poliovirus	*Enterovirus*	302	pT3
1AYM	Human rhinovirus 16	*Rhinovirus*	285	pT3

1AYN	Human rhinovirus 16	*Rhinovirus*	285	pT3
1C8M	Human rhinovirus 16/pleconaril complex	*Rhinovirus*	285	pT3
1FPN	Human rhinovirus 2	*Rhinovirus*	283	pT3
1HRI	Human rhinovirus 14/sch38057 complex	*Rhinovirus*	289	pT3
1HRV	Human rhinovirus 14/sdz 35-682 complex	*Rhinovirus*	289	pT3
1NA1	Human rhinovirus 14/pleconaril complex	*Rhinovirus*	289	pT3
1NCQ	Human rhinovirus 14/pleconaril complex	*Rhinovirus*	289	pT3
1NCR	Human rhinovirus 16/pleconaril complex	*Rhinovirus*	285	pT3
1ND2	Human rhinovirus 16	*Rhinovirus*	285	pT3
1ND3	Human rhinovirus 16/pleconaril complex	*Rhinovirus*	285	pT3
1QJU	Human rhinovirus 16/antiviral c complex	*Rhinovirus*	285	pT3
1QJX	Human rhinovirus 16/antiviral c complex	*Rhinovirus*	285	pT3
1QJY	Human rhinovirus 16/antiviral c complex	*Rhinovirus*	285	pT3
1R08	Human rhinovirus 14/winviii complex	*Rhinovirus*	289	pT3
1R09	Human rhinovirus 14/winr61837 complex	*Rhinovirus*	289	pT3
1R1A	Human rhinovirus serotype 1a	*Rhinovirus*	287	pT3
1RHI	Human rhinovirus 3	*Rhinovirus*	288	pT3
1RMU	Human rhinovirus 14 mutant c199y	*Rhinovirus*	289	pT3
1RUC	Human rhinovirus 14 mutant n1105s/win52035 complex	*Rhinovirus*	289	pT3
1RUD	Human rhinovirus 14 mutant n1105s/win52035 complex	*Rhinovirus*	289	pT3
1RUE	Human rhinovirus 14 mutant n1219a/win52035 complex	*Rhinovirus*	289	pT3
1RUF	Human rhinovirus 14 mutant n1219a	*Rhinovirus*	289	pT3
1RUG	Human rhinovirus 14 mutant n1219s/win52035 complex	*Rhinovirus*	289	pT3
1RUH	Human rhinovirus 14 mutant n1219s/win52084 complex	*Rhinovirus*	289	pT3

1RUI	Human rhinovirus mutant s1223g/win52084 complex	*Rhinovirus*	289	pT3
1RUJ	Human rhinovirus 14 mutant s1223g	*Rhinovirus*	289	pT3
1RVF	Human rhinovirus 14/fab complex	*Rhinovirus*	289	pT3
1V9U	Human rhinovirus 2/receptor fragment complex	*Rhinovirus*	283	pT3
1VRH	Human rhinovirus/sdz 880-061 complex	*Rhinovirus*	289	pT3
2HWB	Human rhinovirus 14 mutant i2170v/win65291 complex	*Rhinovirus*	289	pT3
2HWC	Human rhinovirus 14 mutant i2170v/win54954 complex	*Rhinovirus*	289	pT3
2HWD	Human rhinovirus 1a/win56291 complex	*Rhinovirus*	287	pT3
2HWE	Human rhinovirus 1a/win54954 complex	*Rhinovirus*	287	pT3
2HWF	Human rhinovirus 1a/r61837 complex	*Rhinovirus*	287	pT3
2R04	Human rhinovirus 14/winiv complex	*Rhinovirus*	289	pT3
2R06	Human rhinovirus 14/winvi complex	*Rhinovirus*	289	pT3
2R07	Human rhinovirus 14/winvii complex	*Rhinovirus*	289	pT3
2RM2	Human rhinovirus 14/winii(sr) complex	*Rhinovirus*	289	pT3
2RMU	Human rhinovirus 14 mutant v188l	*Rhinovirus*	289	pT3
2RR1	Human rhinovirus 14/wini(r) complex	*Rhinovirus*	289	pT3
2RS1	Human rhinovirus 14/wini(s) complex	*Rhinovirus*	289	pT3
2RS3	Human rhinovirus 14/winiii(s) complex	*Rhinovirus*	289	pT3
2RS5	Human rhinovirus 14/winv(s) complex	*Rhinovirus*	289	pT3
3DPR	Human rhinovirus 2 bound to a concatamer of the vldl receptor module v3	*Rhinovirus*	283	pT3
4RHV	Human rhinovirus serotype 14	*Rhinovirus*	289	pT3
3CJI	Structure of seneca valley virus-001	*Senecavirus*	279	pT3
Fv15: *polyomaviridae*				
1SID	Murine polyoma virus	*Polyomavirus*	383	7d

1SIE	Murine polyoma virus	*Polyomavirus*	383	7d
1SVA	Simian virus 40	*Polyomavirus*	361	7d
Fv16: *siphoviridae*				
1OHG	Bacteriophage hk97 mature empty capsid	*Lambda-like viruses*	384	7l
2FRP	Bacteriophage hk97 expansion intermediate iv	*Lambda-like viruses*	383	7l
2FS3	Bacteriophage hk97 k169y head i	*Lambda-like viruses*	383	7l
2FSY	Bacteriophage hk97 pepsin-treated expansion intermediate iv	*Lambda-like viruses*	383	7l
2FT1	Bacteriophage hk97 head ii	*Lambda-like viruses*	383	7l
2GP1	Bacteriophage hk97 prohead ii crystal structure	*Lambda-like viruses*	383	7l
3DDX	Hk97 bacteriophage capsid expansion intermediate-ii model	*Lambda-like viruses*	383	7l
3P8Q	Hk97 prohead i encapsidating inactive virally encoded protease	*Lambda-like viruses*	374	7l
3QPR	Hk97 prohead i encapsidating inactive virally encoded protease	*Lambda-like viruses*	383	7l
Fv17: *sobemoviridae*				
1F2N	Rice yellow mottle virus	*Sobemovirus*	238	3
1NG0	Cocksfoot mottle virus	*Sobemovirus*	253	3
1SMV	Sesbania mosaic virus (smv)	*Sobemovirus*	260	3
1X33	T=3 recombinant capsid of semv cp	*Sobemovirus*	268	3
1X35	Recombinant t=3 capsid of a site specific mutant of semv cp	*Sobemovirus*	268	3
2IZW	Crystal structure of ryegrass mottle virus	*Sobemovirus*	235	3
2VQ0	Capsid structure of sesbania mosaic virus coat protein deletion mutant rcp(delta 48 to 59)	*Sobemovirus*	268	3
4SBV	Southern bean mosaic virus (sbmv)	*Sobemovirus*	260	3
Fv18: *tetraviridae*				
2QQP	Crystal structure of authentic providence virus	*Betatetravirus*	623	4
1OHF	Nudaurelia capensis ω virus	*Omegatetravirus*	641	4

3S6P	Crystal structure of helicoverpa armigera stunt virus	*Omegatetravirus*	647	4
Fv19: *togaviridae*				
3J0C	Models of e1, e2 and cp of venezuelan equine encephalitis virus tc-83 strain	*Alphavirus*	442	4
3J0F	Sindbis virion	*Alphavirus*	439	4
Fv20: *tombusviridae*				
1OPO	Carnation mottle virus	*Carmovirus*	348	3
2ZAH	X-ray structure of melon necrotic spot virus	*Carmovirus*	390	3
1C8N	Tobacco necrosis virus (tnv)	*Necrovirus*	276	3
2TBV	Tomato bushy stunt virus (tbsv)	*Tombusvirus*	387	3
FV21: *Tymoviridae*				
1AUY	Turnip yellow mosaic virus (tymv)	*Tymovirus*	189	3
1DDL	Desmodium yellow mottle virus (dymv)	*Tymovirus*	188	3
1E57	Physalis mottle virus	*Tymovirus*	188	3
1QJZ	Physalis mottle virus (phmv)	*Tymovirus*	188	3
1W39	Turnip yellow mosaic virus (artificial top component)	*Tymovirus*	189	3
2FZ1	Structure of empty head turnip yellow mosaic virus (atc) at 100 k	*Tymovirus*	189	3
2FZ2	Structure of turnip yellow mosaic virus at 100 k	*Tymovirus*	189	3
2XPJ	Crystal structure of physalis mottle virus with intact ordered rna	*Tymovirus*	188	3

10 Productividad científica.

Journal of Structural Biology 190 (2015) 47–55

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

CapsidMaps: Protein–protein interaction pattern discovery platform for the structural analysis of virus capsids using Google Maps

Mauricio Carrillo-Tripp [a,*], Daniel Jorge Montiel-García [a], Charles L. Brooks III [c,d,e], Vijay S. Reddy [b]

[a] Biomolecular Diversity Laboratory, Unidad de Genómica Avanzada (Langebio) CINVESTAV, Irapuato, Mexico
[b] Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA
[c] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[d] Department of Chemistry, University of Michigan, Ann Arbor, MI, USA
[e] Department of Biophysics, University of Michigan, Ann Arbor, MI, USA

ARTICLE INFO

Article history:
Received 17 December 2014
Received in revised form 20 January 2015
Accepted 10 February 2015
Available online 16 February 2015

Keywords:
Viral capsids
Quaternary interactions
Web interface
Heat maps
Similarity score

ABSTRACT

Structural analysis and visualization of protein–protein interactions is a challenging task since it is difficult to appreciate easily the extent of all contacts made by the residues forming the interfaces. In the case of viruses, structural analysis becomes even more demanding because several interfaces coexist and, in most cases, these are formed by hundreds of contacting residues that belong to multiple interacting coat proteins. CapsidMaps is an interactive analysis and visualization tool that is designed to benefit the structural virology community. Developed as an improved extension of the φ–ψ Explorer, here we describe the details of its design and implementation. We present results of analysis of a spherical virus to showcase the features and utility of the new tool. CapsidMaps also facilitates the comparison of quaternary interactions between two spherical virus particles by computing a similarity (S)-score. The tool can also be used to identify residues that are solvent exposed and in the process of locating antigenic epitope regions as well as residues forming the inside surface of the capsid that interact with the nucleic acid genome. CapsidMaps is part of the VIPERdb Science Gateway, and is freely available as a web-based and cross-browser compliant application at http://viperdb.scripps.edu.

1. Introduction

Viruses are macromolecular machines made up of nucleo-protein components. Multiple copies of one or a few kinds of coat proteins (CPs) must assemble into symmetric closed shells (capsids), encapsulating their own genome in order to form a functional viral particle. In this respect, viral capsids provide an excellent resource for studying protein–protein interaction mechanisms in homo-oligomers that form symmetric closed shells as well as protein–nucleic acid interactions. In the case of virus capsids that self-assemble, this occurs rapidly and spontaneously with a high degree of fidelity. It is reasonable to assume that the self-assembly process require "molecular recognition signals" that are encoded into sequence and structure of the coat proteins in order to form the final virus particle. These molecular recognition signals are important for the initial nucleation and subsequent assembly. Identification of these signals could be the key to understand viral assembly, and possibly, macromolecular assembly in general. Nature and extent of these interactions between the coat protein subunits determine the size and the robustness of the capsid. Furthermore, analysis of protein–protein interactions at the CP subunit interfaces may provide insights into different assembly mechanisms involved in spontaneous self-assembly vs. those that might require auxiliary proteins.

Here, we refer to an "interaction pattern" as the specific spatial arrangement of CP subunit interface residues for a particular spherical capsid. Interface residues are defined by a distance based criteria, calculated for all residues in all subunits in an assembled capsid. Once identified, each residue forming the interaction pattern can be further characterized by the number of close contacts it makes with neighboring subunits, its buried surface area, or its sequence identity among members of the same virus family. In addition, the triangulation number (T) is a parameter used to describe how the CPs arrange among themselves in spherical viruses that display icosahedral symmetry. It usually refers to the number of unique/distinct structural environments, or often the

* Corresponding author at: Laboratorio de la Diversidad Biomolecular, Unidad de Genómica Avanzada, Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional, Km 9.6 Libramiento Norte Carretera a León, 36821 Irapuato, Gto., Mexico.
E-mail address: trippm@langebio.cinvestav.mx (M. Carrillo-Tripp).

http://dx.doi.org/10.1016/j.jsb.2015.02.003
1047-8477/

number of coat protein subunits present in one of the icosahedral asymmetric units (IAUs, each one being 1/60th of an icosahedron), as described by Caspar and Klug (Caspar and Klug, 1962). A number of icosahedral capsids have been structurally characterized at high resolution, displaying a diversity of geometric architectures having T numbers equal to 1, 3, 4, 7, 13 and above (see Fig. 1 for an example of the first three).

The simplest spherical viruses that display icosahedral symmetry are built by 60 copies of the same coat protein coming from a single gene product and occupy identical environments. More complex and larger capsids are composed of 60xT CPs (e.g., 120, 180, 240, 420, etc.), although 120 subunit capsids (e.g., inner capsids of reoviruses) do not conform to canonical Caspar and Klug formulation of quasi-equivalence (Caspar and Klug, 1962). In the cases where $T > 1$, the CPs that occupy an IAU are arranged in quasi-equivalent environments. Interestingly, even within a category of capsids having the same T number (e.g., $T = 3$), there could be significant variations in subunit associations when forming the respective capsids (Damodaran et al., 2002).

To date, there are 271 unique three dimensional capsid structures known at atomic resolution, from 38 virus families and 73 genera. This dataset is curated and deposited at VIPERdb, a Science Gateway specialized in structural virology (Carrillo-Tripp et al., 2009). VIPERdb is supported by a relational database with a web interface, specifically designed for the analysis and visualization of icosahedral virus capsid structures. VIPERdb has become the global reference in the field as a comprehensive resource for the virology community, with an emphasis on the description and comparison of derived data from structural and computational analyses of virus capsids.

A novel methodology to map the location of residues involved in subunit–subunit interactions has been previously described (Carrillo-Tripp et al., 2008). These normalized diagrams proved to be useful as roadmaps for the visualization of the density, distribution and characteristics of the residues at the sub-unit interfaces. A preliminary implementation of this methodology was reported in the framework of VIPERdb, namely, the φ–ψ Explorer (Carrillo-Tripp et al., 2009). Although basic functionality was achieved, several efficiency improvements and analysis additions have been implemented since. Evolution of web technologies in recent years has allowed the development of CapsidMaps, an improved and updated extension to the φ–ψ Explorer visualization and analysis capabilities. Here we describe the design and implementation specific details of CapsidMaps, and present results of the analysis of a spherical virus to showcase the power of the new tool.

2. Materials and methods

2.1. The φ–ψ space representation

Azimuthal polar orthographic diagrams, or APODs, are simplified two dimensional representations of a three dimensional macromolecular virus structure (Carrillo-Tripp et al., 2008). APODs are made by projecting the capsid protein residues onto a unit sphere and then mapping to a plane using a mathematical transformation. Given the spherical nature of icosahedral virus capsids, the three dimensional Cartesian coordinates $R(x, y, z)$ of the center of mass of each residue can also be represented in Spherical coordinates $R(r, \varphi, \psi)$ using the set of equations

$$r = \sqrt{x^2 + y^2 + z^2} \tag{1}$$

$$\varphi = \arctan(y/x) \tag{2}$$

$$\psi = \arccos(z/r), \tag{3}$$

where r is the magnitude of the vector R pointing to the center of mass of a protein residue, φ is the azimuthal angle between the X axis and the projection of R into the XY plane (analogous to a longitude), and ψ is the polar angle between the Z axis and R (analogous to a latitude). Each vector R is first scaled into a unit vector, leaving all points lying on the surface of a unit sphere. A two dimensional map can be created by making an azimuthal polar orthographic projection onto a tangential plane to the sphere. This map represents the view of the sphere seen from the top of one of the poles, with the φ angle starting at 0° at the positive X axis growing counterclockwise up to 360° after a full turn, and the ψ angle starting at 0° at the center of the map growing in concentric circles up to 90° at the sphere's equator (Fig. 2).

This method provides a unique advantage by mapping residue locations onto the same central IAU representation, irrespective of the capsid size or T-number, since all asymmetric units are equivalent to each other, allowing the comparison between different viruses. However, the comparison of two viruses within a particular family and with the same T number gives most meaningful results. Furthermore, this methodology allows to selectively map residues located only at specific protein regions, i.e., at the inter-subunit interfaces, the core of the proteins, the protein–nucleic acid interface or the solvent exposed residues. This provides a way to compare interactions of two or more spherical capsids, as well as to quantitatively estimate and visually assess the extent

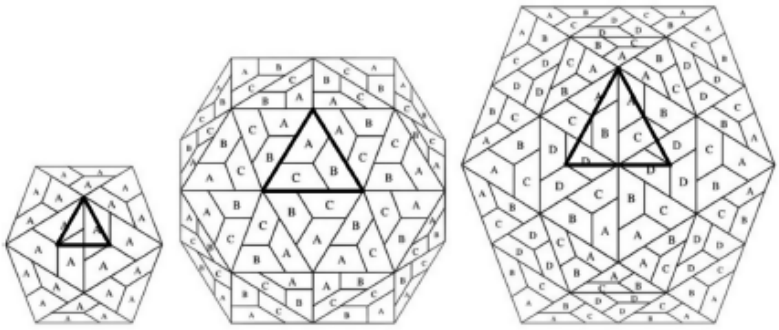


Fig. 1. Schematic representation of $T = 1$, $T = 3$, and $T = 4$ icosahedral lattices. Each trapezoid corresponds to an independent subunit. In the case of $T = 1$ capsids, all the subunits occupy an equivalent environment, namely A, whereas in $T > 1$ capsids, subunits occupy three, four or more distinct environments (in this case A, B, C, or D). The central icosahedral asymmetric unit is highlighted in each case. Adapted from Reddy and Johnson (2005).

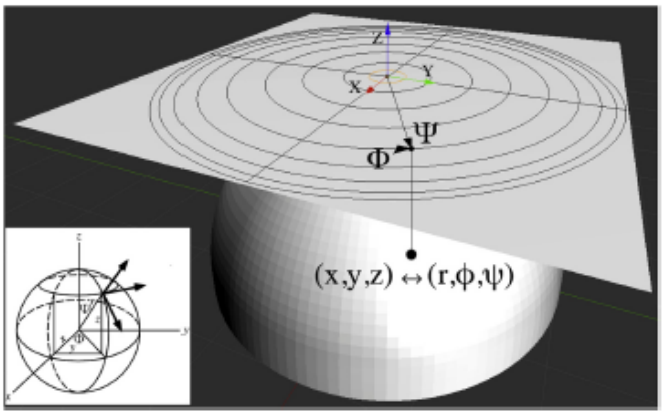


Fig. 2. Coordinate system transformation and polar azimuthal orthographic projection of the center of mass of a protein residue. Every coat protein residue is represented as a point in space (black dots). Cartesian coordinates (x, y, z) are transformed into an spherical system (r, φ, ψ), setting, leaving all residues lying on the surface of a unit sphere. An orthogonal projection (parallel to the Z axis) onto a tangential plane through the North Pole is then applied, ending up with a two dimensional φ–ψ polar map representation of the relative position of all coat protein residues.

of quaternary similarities. This method also highlights the density and distribution of protein–protein interactions with respect to the icosahedral symmetry axes in spherical space. Employing this methodology, answers to specific questions like *"what residues in the Black Beetle Virus are conserved among all members of the Nodavirus family and have contacts near the 5-fold symmetry axis?"*, or *"what is the most abundant charged amino acid type on the outside surface among all the members of the Leviviridae family?"*, can be easily found.

2.2. Similarity score

Quantifying the similarity between the quaternary structure present in an IAU of two different spherical capsids is useful when trying to compare and contrast protein–protein interfaces. The *S*-score is a metric of similarity between the subunit interfaces of two viruses taking into account the subunit quaternary contacts (Carrillo-Tripp et al., 2008). It is defined as twice the ratio of common locations of interactions normalized by the total number of points of interactions in both the capsids that are being compared. The value of the *S*-score for a pair of viruses with identical subunit interfaces will be 1, decreasing up to 0 as the comparison diverges. In this way, the *S*-score is a valuable metric to quantitatively assess the similarity in quaternary interactions and subunit environments between two capsids. The *S*-score accurately identifies quaternary structure similarity, in comparison to other metrics which fail to discern details in the subunit interfaces (RMSD or TM-score). An analysis of spherical viral capsids in terms of the APODs and *S*-scores shows that the intra family or genus pair comparisons have *S*-scores values in the range of 1.0–0.6. On the other hand, inter family capsid pairs, which exhibit similar subunit folds and same capsid architecture (*T*-number) show *S*-scores values in the range of 0.6–0.3. Furthermore, capsid pairs with different *T* numbers will produce *S*-scores values below 0.3 (Carrillo-Tripp et al., 2008).

2.3. CapsidMaps: polar φ–ψ maps implementation

Interactive tools developed for web browsers in recent years can be used to create interactive maps, instead of using static images. Mapping of residues of spherical viruses onto polar φ–ψ maps is analogous to mapping cities onto a two dimensional longitude–latitude world map. This naturally led to exploit existing application programming interfaces (APIs) used to interact with geographic data through a web browser, and use it to display polar φ–ψ maps of the locations of residues of the icosahedral virus capsid proteins. The Google Maps API allows developers to embed maps into web pages in several ways, for either simple use or extensive customization. With a few modifications to the libraries, Google Maps were adapted to the structural virology needs, taking advantage of its capabilities to develop the CapsidMaps tool, in this way offering an interactive way of displaying residue locations in φ–ψ space, as previously described (Carrillo-Tripp et al., 2009).

In its present optimized form, the CapsidMaps application is composed of different layers. An schematic of its architecture can be found in Fig. 3. The VIPER relational data base is located at the foundation, with a server-side VIPERdb API layer functioning as a bridge for the client-side scripts to access data, which in turn uses the Google Maps API library to interactively build and display the maps. A description of the improved implementation follows.

2.4. Server-side API

Extensive or complex processing that requires execution of several SQL statements has been implemented into stored procedures, therefore optimizing data extraction and avoiding overhead and network traffic. With the use of the object oriented programing paradigm (OOP), all searches are made by a *searchObj()* class instance. In this model, specific queries are stored in the data base, so there is no need to send big requests. Instead, the interface makes calls to the right procedure. There are other class instances in charge of the communication between the core API and the data base, data manipulation, XML formatting and transferring of data. The use of OOP also gives the advantage of having a scalable and easily maintainable code.

2.5. Client-side component

This is the interface the browser downloads so the user can interact with the platform. It is responsible for all communications between the client and the server, interpretation of the XML

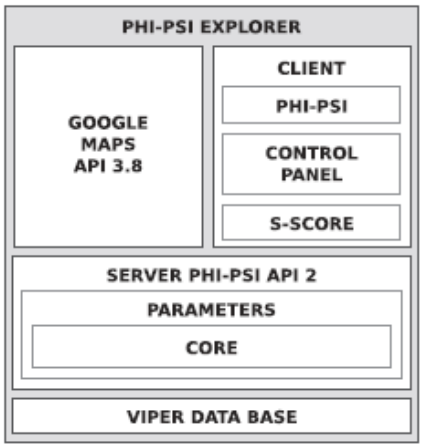


Fig.3. The CapsidMaps tool integrates several layers. The VIPER data base is at the foundation of the platform. The client layer communicates with the server layer to extract specific information from the data base. The client then displays the data dynamically using the Google Maps libraries.

formatted replies, manipulation of the φ–ψ maps through a control panel and the automatized calculation of the *S*-score.

2.6. Google Maps API

A brief summary of the library's features related to the development of CapsidMaps is discussed here. Please refer to the Google Maps JavaScript API Developer's Guide (https://developers.google.com/maps/documentation/javascript/tutorial) for more details. An upgrade to the Google Maps JavaScript API v3 was done with respect to the previous φ–ψ Explorer application. This represents several major improvements in terms of functionality, efficiency and usability. The new release supports the display and management of three custom map types: (i) standard tile sets consisting of images which collectively constitute full maps, (ii) image tile overlays that display on top of existing base map types, and (iii) non-image map types which allows to manipulate the display of map information at its most fundamental level. CapsidMaps implements the later. The custom map relies on creating a class that implements a *MapType* interface. This interface specifies certain properties and methods that allow the API to initiate requests to the custom map type when it determines that it needs to display data within the current view port and zoom level. The client has to handle these requests to decide which data to load. Classes inheriting the *MapType* interface require several methods to be implemented and their properties defined and populated by the developer.

In order for the CapsidMaps application to display a capsid residue in the map (the screen), φ and ψ values have to be translated into a "world" coordinate system. This translation is accomplished using a map projection. Google Maps uses the Mercator projection to create maps from geographic data and convert events on the map into geographic coordinates. Hence, a custom projection method was written for CapsidMaps based on the *google.maps.Projection* interface, in order to display capsid structural data in a map. A *Projection* implementation must provide a bi-directional mapping mechanism, i.e., it has to have a definition on how to translate from capsid coordinates (φ–ψ) to the *Projection*'s world coordinate system, and vice versa. Google Maps assumes that projections are rectilinear. Each *Projection* provides two methods which translate between these two coordinate systems, allowing to convert between capsid and world coordinates:

- The *Projection.fromLatLngToPoint()* method converts a φ–ψ pair into a world coordinate. This method is used to position overlays on the map (and to position the map itself).
- The *Projection.fromPointToLatLng()* method converts a world coordinate into φ–ψ values. This method is used to convert events such as clicks that happen on the map into capsid coordinates.

In order to calculate pixel coordinates in a convenient way, the Google API assumes that a map at zoom level 0 is a single tile of the base tile size. Thus, world coordinates are defined relative to pixel coordinates, using the projection to convert φ and ψ values to pixel positions on this base tile. In this way, a world coordinate is independent of the current zoom level. World coordinates in Google Maps are measured from the projection's origin (upper left corner) and increase in the *X* direction towards the right, and increase in the *Y* direction downwards. In other words, for zoom level 0 the pixel coordinates are equal to the world coordinates. The Maps API constructs a viewport given the zoom level center of the map and the size of the containing Document Object Model (DOM) element, usually defined with an HTML * * tag, and translates this bounding box into pixel coordinates. At the end, the usable world coordinate space in the viewport has an area of ({0–256}, {0–256}) pixels.

Google Maps API V3 implements many new features and code optimizations in comparison to previous versions. Now it implements the Model-View-Controller architecture. The number of markers the application can handle has increased significantly through the use of the *markerCluster* class object, having the additional benefit to create density heat-maps (discussed later). Also, the Google API design is simpler, allowing the programmer to write less Javascript code for the client-side layer, decreasing network traffic.

3. Results and discussion

3.1. CapsidMaps features

The Black Beetle Virus (PDBID = 2bbv, *T* = 3) is chosen as an example to showcase the CapsidMaps application and its new functionalities. Once a particular virus is selected on VIPERdb, the polar CapsidMap can be accessed from the *Info-Page*, under the corresponding tab on the left menu. When the browser opens the *Info-Page*, all the data needed for the generation of a polar φ–ψ map is retrieved from the VIPERdb server through an application programming interface function call and delivered to the client in XML structured format, asynchronously updated using the AJAX technique. The CapsidMap is then populated with one marker for each protein residue returned, displaying them on a polar grid, with the origin at the center of the viewport. Following the φ–ψ Explorer definition, the φ angle grows counterclockwise from the right (0–360°) and the ψ angle (0–90°) grows radially outward (Fig. 2). In addition to the position coordinates (r, φ, ψ), the structural information associated with each residue namely, amino acid type, sequence id, parent subunit, secondary structure, number of neighbor interactions, family sequence identity and solvent accessible surface area. The user interface is composed by the interactive polar φ–ψ map on the main display area, with a control panel on the right. Because of the features available in the Google Maps API, the map itself transparently integrates all the interactive features that a regular Google map has, including zooming in or out, and dragging the display with the mouse in any direction. The control panel is composed of several sections, each one providing a

specific set of options. On the *View* section the user can quickly switch from the current view to either the default view (zero level zoom) or the Q3F (quasi 3-fold) view, which is a zoomed in view of the central IAU down the quasi 3-fold axis of the reference asymmetric unit of the capsid. The *Show and Hide* section offers the option to independently display (on by default) or hide any of the individual elements on the map: coat protein subunits, the icosahedral asymmetric unit (represented by a triangular frame), or the heat-map. In addition, the user can select which specific group of residues to display for further analysis: interface, core, outside surface or inside surface. By default, all amino acid types are displayed, however, the user can select and show one particular type from the drop-down select box in this section (Res. Type). In addition, the user can download relevant data of the currently displayed residues of interest on the CapsidMaps interface at any point during this selection process, through a function call on the *Residue Info* section. A separate page will open, showing a compiled list of residues with its corresponding information. This functionality is helpful when a deeper analysis of a capsid is needed, or a detailed comparison between two or more capsids is desired.

3.2. Interface, core and surface protein residues

All residues in the coat protein are grouped into four categories based on their three dimensional spatial location: interface, core, exterior surface or interior surface (Fig. 4).

The interface residues are displayed by default, the trail of interface residues form a distinctive interaction pattern that is unique to a specific virus or family of viruses. Any of the other three groups (core, exterior or interior surfaces) can be independently loaded into the CapsidMap by clicking on the corresponding control. Also by default, the markers are colored according to the parent subunit the residues belong to, using a standard color scheme. The IAU is depicted by a triangular frame, where the upper vertex represents the 5-fold axis and the other two vertices correspond to 3-fold axes of symmetry with a 2-fold axis located in between. The

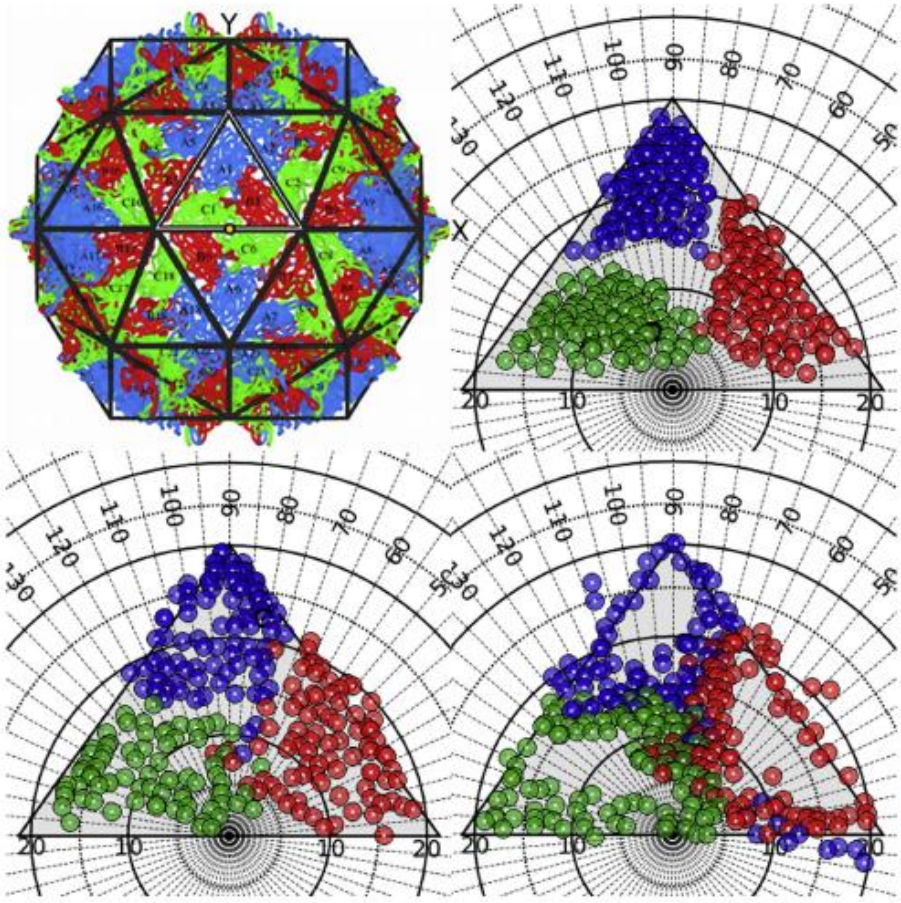


Fig.4. Black Beetle Virus spherical capsid (PDBID = 2bbv, $T = 3$). (Top left) Full capsid structure formed by 180 independent coat protein subunits, color coded according to their specific environment using a standard scheme; A in blue, B in red and C in green. The central icosahedral asymmetric unit (CIAU) is highlighted by a white triangle, showing where the North Pole would be in the unit sphere by a yellow dot. In this representation the positive X axis points right, and the positive Y axis points up. (Top right and Clockwise) Polar φ–ψ maps of the CIAU (represented by a triangular black frame), showing different regions of the coat proteins: core, interface and solvent exposed. Symbols are colored according to the subunit they belong to.

advantage of displaying structural data in this way is to facilitate the comparison of the capsid maps of different capsids. The core residues cluster together with high density, while the interface residues that are responsible for holding the capsid together surround the boundary of individual CPs. The localizations of the interface residues along the periphery of the subunits show the differences between the distinct quasi-equivalent environments. Furthermore, by visually inspecting the solvent exposed residues on the capsid one can search and identify the locations of antigenic epitopes and their relative proximity to other surface exposed regions.

3.3. Heat-maps

As mentioned previously, Google Maps API v3 has the ability to create heat-maps, according to the local marker density (Fig. 5).

This is a helpful representation for the analysis of viral capsids quaternary structure, since it gives information on where the highly populated interaction regions are located in the protein–protein interfaces (interaction patterns). For example, it is easy to visually identify important residue clusters which could be involved in the generation of nascent oligomers (dimers, trimers, etc.) that will eventually give rise to the whole macromolecular structure. Depiction of the heat-maps in the reference IAU frame could provide a novel and powerful way to study the self-assembly mechanism of viral particles in the future.

3.4. Sequence identity, degree of residue interaction, and buried surface area

The section *Color by* in the control panel offers the option to change the color-coding of the markers (Fig. 6).

The color-code currently being used is displayed at the far left, under the info-box. The default is to color the markers depending on the parent subunit the residues belong to and the local quaternary environment they are in. However, there are three other options to color-code the markers depending on the sequence identity, solvent accessible surface area (SASA) or buried surface area (BSA), or number of neighboring interactions each residue makes (when the interface residues are displayed). Sequence

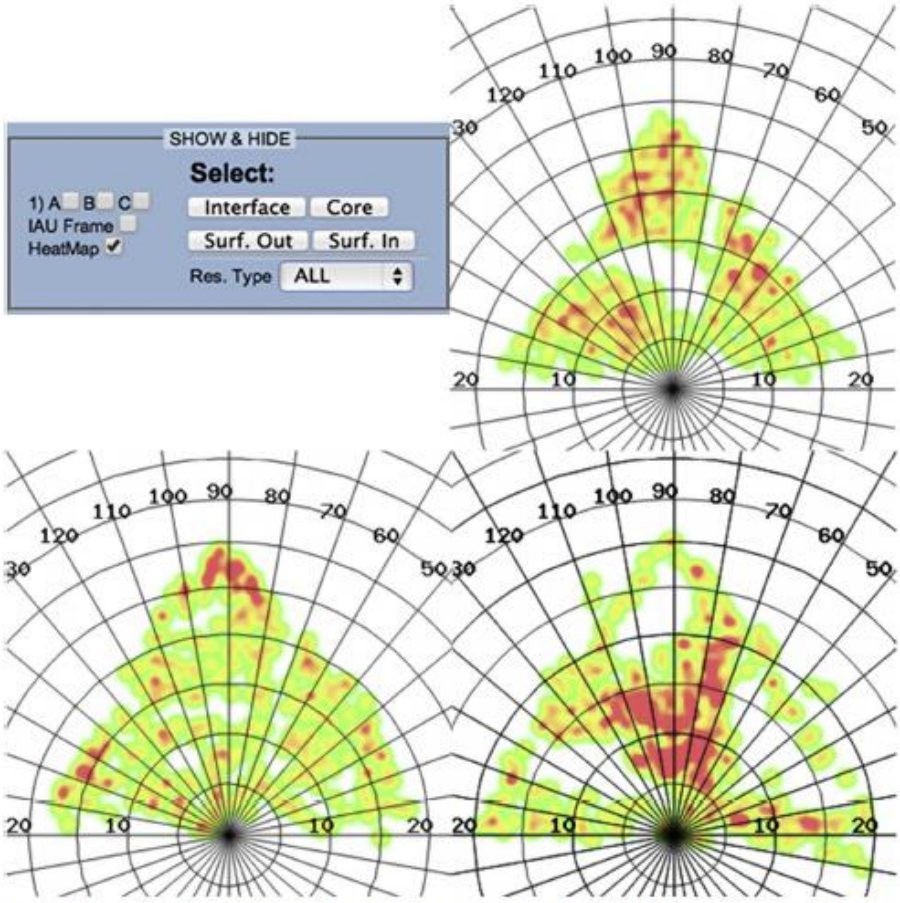


Fig.5. Black Beetle Virus heat-maps, color coded according to the local residue density; areas of higher density are colored red and areas of lower density appear green. (Top left) Control panel with options selected to hide all subunits and the CIAU, leaving the heat-map layer on. (Top right and Clockwise) φ–ψ heat-maps showing different regions of the coat proteins: core, interface and solvent exposed.

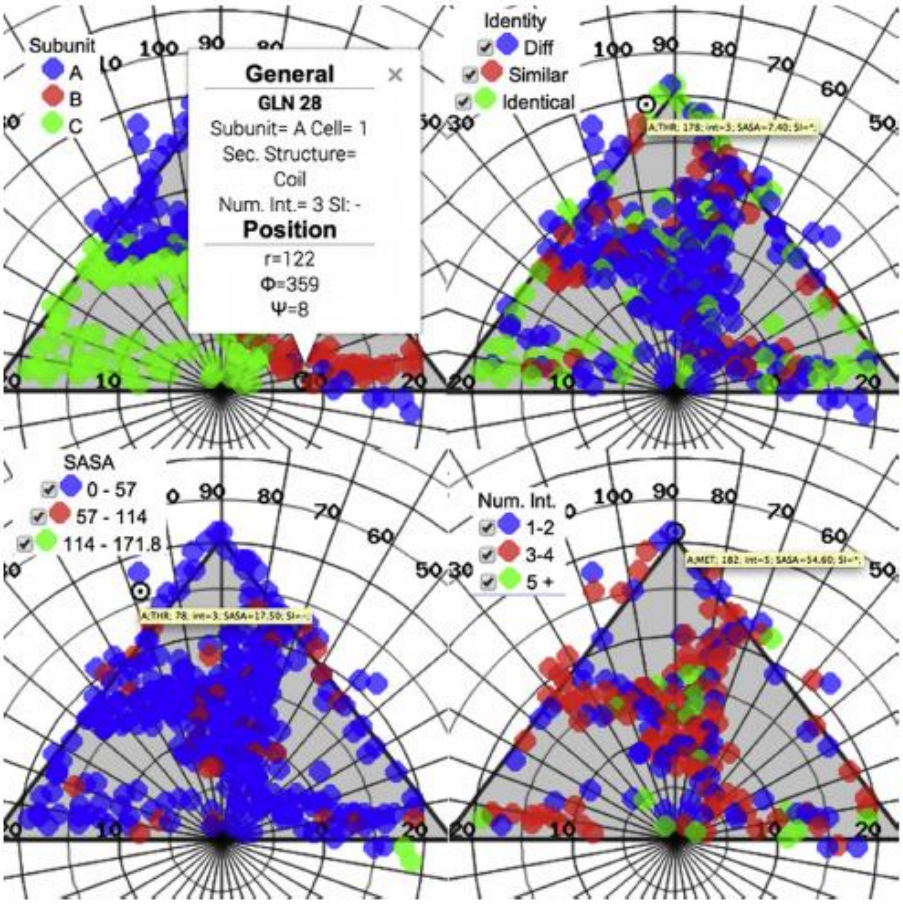


Fig.6. Black Beetle Virus interface residues, color coded according to four different criteria: (Top left and Clockwise) capsid environment type, family sequence conservation, number of interaction contacts per residue, and solvent accessible surface area. When clicking on a particular marker, a balloon shows general information about the corresponding residue, and hovering over it shows a synthesis of the data.

identity, i.e., the a.a. sequence conservation among members of the corresponding virus family, has been previously calculated through an intra-family multiple sequence alignment (IFMSA). The IFMSA calculation is based on the three dimensional structure multiple alignment between all structurally characterized members of a particular family. The results have been added to the VIPER database, from where they can be retrieved via an API procedure call. All the IFMSAs were computed with the package T-Coffee (Notredame et al., 2000), using a consensus of several pairwise structure alignment profiles built with SAP (Taylor et al., 1994), MUSTANG (Konagurthu et al., 2006) and TM-align (Zhang and Skolnick, 2005). The precomputed IFMSAs can also be downloaded from the VIPERdb server in clustalw format through a link in the *Biodata* tab. The power of this type of representation can be appreciated in the case of the Black Beetle Virus, where it can be seen that there is a high degree of sequence conservation near the axes of symmetry, implying a common assembly mechanism shared by all members of the Nodavirus family. The number of interactions of each interface residue makes with the neighboring subunits is calculated when the virus entry is deposited in VIPERdb, storing the results in the data base for later retrieval through a procedure call. The data set can be visualized by clicking on the corresponding button on the control panel. This information is useful to identify the location of residues that contribute the most in the network formed at the quaternary interfaces. In the case of the Black Beetle Virus, the distribution of these residues is not uniform along the borders of the coat proteins, but are rather concentrated in specific regions. Having a large number of interactions might imply that these residues could be responsible for the stability of the viral particle, pinpointing a putative target for capsid disruption.

Other quantities calculated when a new entry is deposited in VIPERdb are the related SASA and BSA. The later is used to estimate subunit association energies by multiplying its value by some solvation parameters (Horton and Lewis, 1992). This is done for each residue in the IAU, regardless of its classification inside the coat protein, and considering all neighboring subunits in the whole capsid. These values, first described by Lee and Richards (1971), and Rashin (1984), measure the degree of exposure to the solvent, or genome, a specific residue has (in $Å^2$). The classic view of the meaning of such values is that the amount a residue is buried

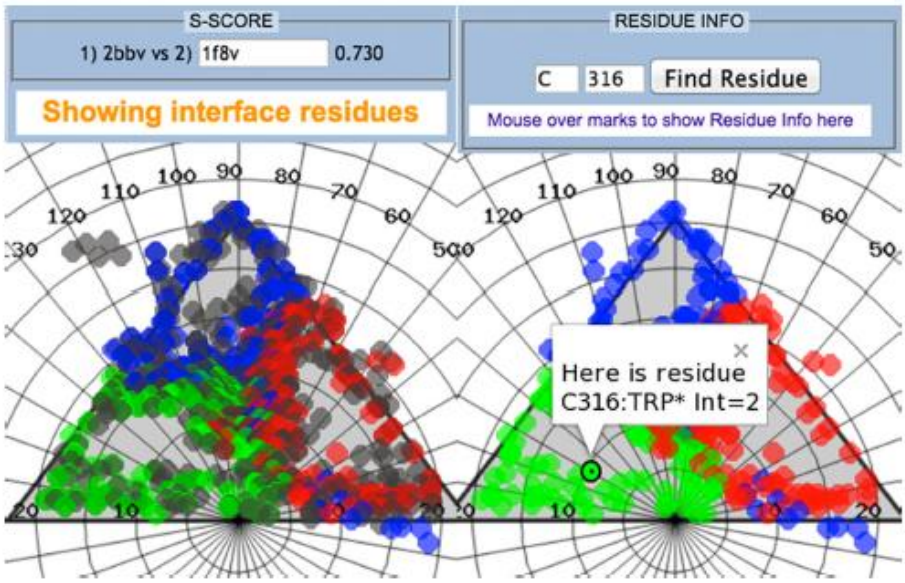


Fig.7. Black Beetle Virus interface residues, color coded according to the local subunit environment. (left) The application has the ability to calculate the *S*-score between two capsids. Once a second virus is selected by its PDBID, the metrics value is displayed in the control panel. Additionally, residues of the second virus are also displayed on the map (gray color), allowing for the visual assessment of where main quaternary differences might be. (Right) Another added feature is the ability to search for a specific residue by specifying a subunit and a sequence id.

inside a protein, or a protein–protein interface, is important (the former being related to the stability of the protein fold and the latter to the formation of oligomers), since it has been shown that their values correlate to the folding or binding free energy (Miller et al., 1987; Janin et al., 1988; Noskov and Lim, 2001). This can clearly be observed in the case of the Black Beetle Virus, where the majority of the interface residues have SASA values close to zero, and equivalently large BSA values, implying an energetically stable capsid. Finally, and as discussed earlier, the CapsidMaps inherit all dynamic features that Google Maps have. Clicking on an individual marker will pop up an Info-Window with relevant structural information about the corresponding coat protein residue: amino acid type, residue sequence number, type of secondary structure it belongs to, number of interactions made (if the residue belongs to an interface), residue BSA or SASA value (if the residue belongs to the core/interface or the exposed surface, respectively) and the residue exact position in spherical coordinates.

3.5. Quantitative comparisons: S-score

In order to measure the degree of similitude between the interaction patterns of two viruses (or the same virus in two different conditions if the structural data is available), the calculation of the *S*-score was implemented. The estimation of the *S*-score requires the manipulation of big matrices (Carrillo-Tripp et al., 2008), and even though its value is just the mathematical dot product between two of them, it can be a cumbersome computation. The CapsidMaps application facilitates this, by providing an interface to select a second virus entry on VIPERdb through the *S-score* section of the control panel. After the *S*-score has been calculated, its value is shown. Additionally, the residues of the second virus are displayed on the map, superimposed on top of the residues of the first virus. In order to make a clear distinction between the two data sets, markers of the second virus are colored gray. This provides a good way to quantitatively assess the quaternary similarity between two capsids (regardless of family or *T* number), but also to visually identify regions in space where differences might occur (Fig. 7).

3.6. Searching for specific residues

Clicking or hovering over a marker will display information related to the corresponding residue in two ways: as a small balloon on the map or on an info-box on the *Residue Info* section of the control panel. Searching for a specific residue, however, can be difficult when the map is populated with a high number of markers. Therefore, the *Residue Info* section offers the option to quickly find a residue by entering the name of the parent subunit and the residue sequence id. When the user clicks on the corresponding button, an Info-Window (balloon) will pop up on the map, showing the location of such residue (Fig. 7).

4. Conclusions

Here we show the use of the CapsidMaps application to visualize spherical virus structure derived data, proving to be a powerful way to analyze these macromolecular protein assemblies. This hybrid web application is a good example of how the unconventional use and integration of unrelated tools and datasets can break old paradigms, opening the door to further improve the way research is done. The case study used clearly demonstrates that the application is able to provide a novel approach that facilitates the study of complex systems. To the best of our knowledge, CapsidMaps is the first application of its kind in the field of structural virology. In conclusion, CapsidMaps is a sophisticated visualization and analysis tool which will potentially enhance our understanding of complex structural data sets, making it easier to discover

patterns and relationships and to form hypotheses in the realm of structural virology.

Acknowledgements

This work was supported by the Consejo Nacional de Ciencia y Tecnología de México [grant number 132376 to M.C.-T.] and the USA National Institutes of Health through the Center for Multi-Scale Modeling for Structural Biology [grant number RR012255 to C. L. B. III and V. R.].

References

Carrillo-Tripp, M., Brooks, C.L., Reddy, V.S., 2008. A novel method to map and compare protein–protein interactions in spherical viral capsids. Proteins 73 (3), 644–655.

Carrillo-Tripp, M., Shepherd, C.M., Borelli, I.A., Venkataraman, S., Natarajan, P., Johnson, J.E., Brooks, C.L., Reddy, V.S., 2009. VIPERdb2: an enhanced and web API enabled relational database for structural virology. Nucleic Acids Res. 37 (Database issue), D436–D442.

Caspar, Dt., Klug, A., 1962. Physical Principles in the Construction of Regular Viruses, 1st ed., vol. 27. Cold Spring Harbor Laboratory.

Damodaran, K., Reddy, V.S., Johnson, J.E., Brooks III, C.L., 2002. A general method to quantify quasi-equivalence in icosahedral viruses. Mol. Biol. 324, 723–737.

Horton, N., Lewis, M., 1992. Calculation of free energy of association for protein complexes. Protein Sci. 1 (1), 169–181.

Janin, J., Miller, S., Chothia, C., 1988. Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. 204 (1), 155–164.

Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.M., 2006. MUSTANG: a multiple structural alignment algorithm. Proteins: Struct., Funct., Bioinf. 64 (3), 559–574.

Lee, B., Richards, F.M., 1971. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55 (3), 379–383.

Miller, S., Lesk, A.M., Janin, J., Chothia, C., 1987. The accessible surface area and stability of oligomeric proteins. Nature 328 (6133), 834–836.

Noskov, S.Y., Lim, C., 2001. Free energy decomposition of protein-protein interactions. Biophys. J. 81 (2), 737–750.

Notredame, C., Higgins, D.G., Heringa, J., 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302 (1), 205–217.

Rashin, A.A., 1984. Buried surface area, conformational entropy, and protein stability. Biopolymers 23 (8), 1605–1620.

Reddy, V.S., Johnson, J.E., 2005. Structure-derived insights into virus assembly. Adv. Virus Res. 64 (1), 45–68.

Taylor, W.R., Flores, T.P., Orengo, C.A., 1994. Multiple protein structure alignment. Protein Sci. 3 (10), 1858–1870.

Zhang, Y., Skolnick, J., 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33 (7), 2302–2309.

Journal of Structural Biology 196 (2016) 299–308

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Structure based sequence analysis of viral and cellular protein assemblies

Daniel J. Montiel-García [a], Ranjan V. Mannige [b,1], Vijay S. Reddy [b], Mauricio Carrillo-Tripp [a,2,*]

[a] Biomolecular Diversity Laboratory, Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional, Mexico
[b] Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA

ARTICLE INFO

Article history:
Received 27 June 2016
Accepted 18 July 2016
Available online 29 July 2016

Keywords:
Viral capsid proteins
Sequence conservation
Protein–protein interactions
Subunit interface
Bioinformatics
Protein complexes

ABSTRACT

It is well accepted that, in general, protein structural similarity is strongly related to the amino acid sequence identity. To analyze in great detail the correlation, distribution and variation levels of conserved residues in the protein structure, we analyzed all available high-resolution structural data of 5245 cellular complex-forming proteins and 293 spherical virus capsid proteins (VCPs). We categorized and compare them in terms of protein structural regions. In all cases, the buried core residues are the most conserved, followed by the residues at the protein–protein interfaces. The solvent-exposed surface shows greater sequence variations. Our results provide evidence that cellular monomers and VCPs could be two extremes in the quaternary structural space, with cellular dimers and oligomers in between. Moreover, based on statistical analysis, we detected a distinct group of icosahedral virus families whose capsid proteins seem to evolve much slower than the rest of the protein complexes analyzed in this work.

1. Introduction

It has been half a century since early efforts started to describe protein-protein complexes of the cell (Chotia, 1974; Richards, 1974; Chotia and Janin, 1975). Such works showed that the shape complementarity of cellular protein interfaces could be used to characterize the interactions regarding the size of buried surface, paucity of buried water molecules and packing density of interface atoms. The modern view of cellular protein interfaces has produced a rule base of characteristics that include size, residue composition, hydrophobicity, and planarity (Lawrence and Colman, 1993; Jones and Thornton, 1995).

The exponential increase of available structural information has enabled us to revisit and improve on past results. For example, studies of non-redundant datasets of a couple of thousand protein structures have shown that cellular protein interfaces differ in amino acid composition, residue–residue preferences between interactions, and secondary structure, from those of surface and core residues (Ofran and Rost, 2003; Yan et al., 2008). In general, studies use datasets containing complexes from different species. However, focus on a single species was approached by analyzing all the available data on structural complexes from the yeast *Saccharomyces cerevisiae* (Talavera et al., 2011). It was found that, as previously seen, there is a significant contribution of main-chain atoms to protein-protein contacts and the type of interaction seems to depend on both amino acid side chain and secondary structure type involved at the contact. Cellular homo and hetero-complexes showed no clear distinction. Interestingly, there seem to be no significant differences between the interface regions and the rest of the surface from a thermodynamic standpoint regarding the solvation energy.

Just like the cellular proteins that have a structural function, the virus capsid proteins (VCPs) also present intriguing features. In the case of icosahedral viruses, at least 60 copies of a type of VCP must self-assemble into symmetric closed protein complexes in the form of spherical shells (capsids) that encapsulate the viral genome (Cann, 2005). The capsids display a defined size and structural architecture depending on the type of virus (Caspar and Klug, 1962). A detailed description of the molecular specificity, recognition and self-assembly properties of the VCPs remain elusive. Such molecular mechanisms still need to be well understood, in comparison to cellular proteins (Janin et al., 2008). Recently, the geometric and physical-chemical properties of a set of 49 icosahedral virus capsids were analyzed and compared with the interfaces of cellular protein–protein complexes. It seems that small capsid

* Corresponding author.
E-mail addresses: mauricio.carrillo@cinvestav.mx, mauricio.carrillo@cimat.mx (M. Carrillo-Tripp).
[1] Present address: Theory of Nanostructured Materials facility, Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.
[2] Present address: Ciencias de la Computación, Centro de Investigación en Matemáticas, A.C., Guanajuato 36000, Mexico.

http://dx.doi.org/10.1016/j.jsb.2016.07.013
1047-8477/

interfaces are loosely packed, like crystal contacts, whereas the larger interfaces are close-packed, as in cellular homodimers (Bahadur et al., 2007). Also, a statistical analysis of a set of 319 icosahedral viruses showed that VCPs exhibit an apparent segregation in structural fold space (Cheng and Brooks, 2013). It was suggested that the unique folds of VCPs present a favorable geometry to allow adequate packing and assembly into the right architecture. Such structural folds might be under particular constraints during evolution by the requirement of the assembled cage-like structure, as opposed to their surface chemistry. Furthermore, some structural characteristics seem to be also unique to non-capsid viral proteins. When compared to their cellular counterparts, they show lower contact densities, higher occurrence of random coil segments, shorter disordered regions, and less destabilizing effects when mutations happen (Tokuriki et al., 2009).

Even though proteins can diverge beyond the point where there is no detectable sequence similarity, the packing of the tertiary structure can maintain similar folds. Efforts have been made to understand the underlying principles of structural conservation during protein sequence evolution. Earlier works have analyzed the relationship between the divergence of sequence and the three-dimensional structure of cellular proteins (Chotia and Lesk, 1986), and the relation between the sequence identity and structure similarity to the alignment length (Sander and Schneider, 1991; Rost, 1999). An alternative approach came in the form of the classification of the protein fold space. One example is SCOP, an expert-based hierarchical classification of protein structures (Murzin et al., 1995). SCOP groups together those domains that have structural, functional, and sequence evidence for a common evolutionary ancestor at the superfamily (SF) level. In particular, out of the 560 SCOP v1.73 protein domains found in viruses, >10% do not have any structural or evolutionary relatives in modern cellular organisms at the SF level (Abroi and Gough, 2011).

In general, the variations on the level of residue conservation at different locations in the protein structure is still a controversial subject, being inconclusive or even contradictory. For instance, Grishin and Phillips, 1994 concluded that interface and core residues are not well conserved and evolve nearly as rapidly as the overall protein sequence, after analyzing 135 sequences and 16 structures of five cellular oligomeric enzyme families. Valdar and Thornton, 2001 concluded that interface residues are more conserved than expected for a random distribution, after analyzing 195 sequences representative of six cellular homodimer families. Caffrey et al., 2004 found that the interface is rarely more conserved than the surface, after analyzing 64 homologous cellular dimers. Guharoy and Chakrabarti, 2005 concluded that the average conservation at the central region of the protein-protein interface is higher than its surroundings, after analyzing 122 cellular homodimers. In the case of VCPs, Bahadur and Janin, 2008 concluded that the core and interface residues are better conserved than the chain average, after analyzing 32 icosahedral viruses. Subsequently, Chih-Min et al., 2015 concluded that some global patterns derived from the capsid structure, like the residue packing density, are consistent with those present in VCP sequence conservation profiles. Overall, it is a plausible idea that the fact that all previous analyses have been performed on families of homologous proteins using small data sets could bias the results. The particular role of the VCP in a structural and evolutionary context needs further investigation, and an extensive comparison to cellular proteins is in order.

In this work, we further investigate and highlight the differences between cellular and icosahedral capsid proteins. We address three particular questions. First, what is the correlation between the conservation of sequence and the similarity in tertiary and quaternary structures? Second, how are the conserved residues distributed in the protein structure? And third, what is the variation on the level of residue conservation at different locations in the protein structure? We analyzed all the available high-resolution structural data and compared cellular protein *n*-mers with icosahedral VCPs.

2. Materials and methods

2.1. Datasets

In this work, we analyzed all the data available on the three-dimensional structure and sequence of cellular and icosahedral capsid proteins, grouped in four independent datasets. In the case of cellular *n*-mer complexes, we included monomers ($n = 1$. Data not shown), dimers ($n = 2$. Table S1), and higher order oligomers ($n = 3, 4, 5, 6, 8, 10, 12, 22$, and 24. Table S2). In the case of VCPs, we included icosahedral viruses belonging to 36 different genera from 21 different families, according to the classification proposed by the International Committee on Taxonomy of Viruses (ICTV, Fauquet et al., 2005). This dataset spans a broad range of icosahedral triangulation numbers (T = 1, 2, 3, 4, 7d, 7 L, pT3. Table S3).

In all cases, the basic criteria used to choose structures was to have available data determined by X-ray crystallography (resolution $\leqslant 4$ Å), consistent polypeptide chain sequence (no missing loops or fragment miss-annotations), and long chains (>65 residues). Atomic coordinates were obtained from the Protein Data Bank (Berman et al., 2000), in the case of cellular proteins, and from the Virus Particle Explorer database (Carrillo-Tripp et al., 2009), in the case of VCPs. Following (Valdar and Thornton, 2001), the term *protomer* denotes a unique polypeptide chain of a multimeric complex. Hence, homomers will be represented by one protomer, whereas heteromers will be represented by two or more protomers. Hence, our datasets consisted of 5087 cellular monomers, 51 cellular dimers (represented by 57 protomers), 65 cellular oligomers (represented by 101 protomers), and 212 VCPs (represented by 293 protomers).

2.2. Structure similarity

The root mean square deviation (RMSD) has been the standard way to measure structural similarity. However, other metrics provide a better quantification, like the TM-score (Zhang and Skolnick, 2004). The TM-score presents several advantages over the RMSD.

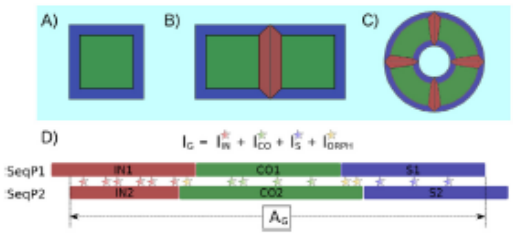


Fig. 1. Simplified diagrams depicting the three different locations of amino acids in the protein tertiary and quaternary structure. Amino acids are located at the protein's solvent accessible surface (S, in blue), at the protein core (CO, in green), or at the protein-protein interface (IN, in red). Schematics shown for a monomer (A), a dimer (B), and an oligomer or virus capsid proteins (C), immersed in a solvent (light blue). In order to study the distribution and level of conservation of amino acids in these locations, a sequence alignment is derived from a 3D structural alignment for a pair of proteins (P1 vs. P2). Then, conserved residues (identical amino acids, indicated by stars) are identified and labeled according to their location in the protein structure (D). In the case where no location correspondence is found between the two proteins for a conserved residue, those are categorized as orphans (ORPH, in yellow). *A* represents the number of aligned residues, and *I* is the number of identical residues.

TM-score values are bound to the interval (0, 1], with 1 being two identical structures (equivalent to an RMSD value of 0). The TM-score is independent of protein size, and it weighs a close match stronger than a distant one. Based on Bayesian theory, it was shown that a TM-score value >0.5 indicates that the two structures have the same fold/topology (Xu and Zhang, 2010). The TM-align tool identifies the best structural superposition between a pair of proteins and calculates the TM-score (Zhang and Skolnick, 2005). Based on the optimal structural superposition, the amino acid sequence alignment between the two proteins was derived.

2.3. Sequence identity

The sequence identity (S_G) considers the fraction of identical residues (I_G) from the total number of aligned residues (A_G) in a sequence alignment, i.e., $S_G = I_G/A_G$. To distinguish the location of the conserved residues in the protein structure, we use three different categories: protein-protein interface (IN), protein core (CO), and solvent accessible surface (S), depicted in Fig. 1. We define a sequence identity index per location category, $S_k^* = I_k/I_G$ (k = IN, CO or S), to quantify the relative percentage of amino acid conservation found in the different regions of the protein. Based on the protein structure superposition, regions specific to each location category were identified in the amino acid sequence alignment. Hence, I_k is the number of conserved residues found in each category region. This procedure excludes certain conserved residues that do not match a common structural location in both sequences based on the above classification, as can be seen from Fig. 1D. We define such residues as orphans (ORPH) in the context of this work.

2.4. Structural categories

The structural classification takes into account the tertiary and quaternary complex structure, i.e., whether the residues are at the protein-protein interface, the core, or on the solvent accessible surface (Fig. 1). We used the same criteria for cellular and capsid proteins. Interface residues are those having at least one close contact with a neighboring protein. The residue type specific cut-offs method is a proper strategy to identify contacting residues between two closely interacting molecules, AB. In the case of protein-protein interactions, the definition of contacts we used (Damodaran et al., 2002) provides a true description of the presence/absence of inter-residue interactions at the AB interface. In the cut-offs method, one calculates the distance between every residue of protein A versus every residue in protein B. Those residue pairs $R_i^A - R_j^B$ being at a shorter distance than the corresponding residue type specific cut-off value are identified as interface residues in the protein complex AB. This approach works well when the atomic positions of both A and B are known. Because the solvent molecules are missing in the data, the lack of atomic information prevents the use of this method to distinguish core from surface residues. The next best approach is to consider the solvent accessible surface area (SASA). Core residues are buried in the protein, whereas solvent accessible residues can interact directly with the environment. It makes sense to use the level of exposure to the solvent as the criteria to group the non-interface residues into core or surface. The SASA method is useful to distinguish which residues have enough area accessible to the solvent to be considered to be at the surface of the protein. A comparison we made between the two methods showed that the SASA approach underestimates the number of interface residues by 10%, on average, when compared to the distance based approach. We performed an extensive examination looking into the distribution of SASA values per residue in proteins, independently done for each dataset (Fig. S1). We found a peak at the [0, 5] interval of the relative accessible surface area (%SASA) in both cellular and capsid proteins. We assume residues in this range are the protein's core residues. Correspondingly, residues with %SASA > 5% are the surface residues. SASA values were computed using the PDBASA library (Shrake and Rupley, 1973).

2.5. Sequence conservation

The sequence conservation is related to the residue variability at each sequence position i of a polypeptide chain, measured by the Shannon entropy,

$$S(i) = -\Sigma_k p_k \ln p_k$$

where $p_k = n_k/N$ is the frequency of residue type k, and n_k is the fraction of sequences having the residue type k at position i on a

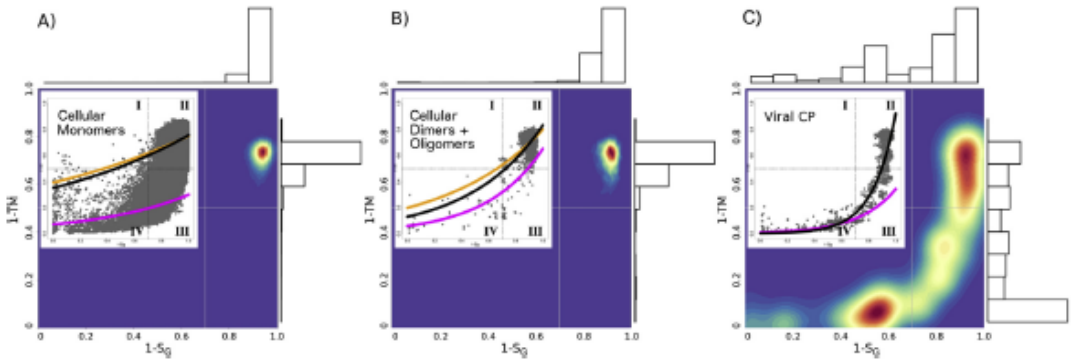


Fig. 2. Correlation between the sequence identity and the structure similarity in proteins. Equivalently, shown is the fraction of mutated residues ($1 - S_C$) and the structure divergence ($1 - TM$). All-vs-all protein pairs analysis of cellular monomers (A), cellular dimers plus oligomers (B), and icosahedral virus capsid proteins (C). Two-dimensional axis-aligned bivariate normal kernel density estimation, evaluated on a square grid of 300 points in each direction. Low density regions are shown in purple, whereas high density regions are red. Probability densities of ($1 - S_C$) and (1-TM) are shown in horizontal and vertical histograms, respectively. Insets: each point in the cloud represents a unique protein pair (gray). Fits to the exponential model are shown for the set of pairs with ($1 - S_C$) > 0 (black), ($1 - S_C$) < 0.7 (magenta), and ($1 - S_C$) > 0.7 (orange). Thresholds are defined as: ($1 - TM$) < 0.5 are pairs with the same protein fold, and ($1 - S_C$) < 0.7 are pairs of homologous proteins. These thresholds produce four sectors of the Cartesian plane, indicated by the roman numerals I, II, III, and IV.

multiple sequence alignment (MSA) of N sequences. $S(i)$ varies between 0 at positions fully conserved, and approximately 3 at positions where all residue types are equally found in the MSA. A normalized entropy can be defined as $s(i) = S(i)/\langle S \rangle$, where $\langle S \rangle$ is the average value of $S(i)$ over all the residues of the polypeptide chain. Values of $S(i)$ for each protomer in our datasets were extracted from the HSSP database (Touw et al., 2015). All statistical tests and plots were carried out using the R package and libraries therein.

3. Results

3.1. Relation between sequence identity and structure similarity

Chotia and Lesk, 1986 analyzed 32 pairs of homologous cellular protein structures. They found that the extent of the structural changes was directly related to the extent of the sequence changes. Following the same strategy, we performed a pair-wise comparative analysis of the structure and sequence of all protomers in each of our datasets. Following their work, we plot the structure divergence (1 − TM-score) as a function of the fraction of mutated residues ($1 - S_G$), where S_G is the global sequence identity, shown in Fig. 2 (insets). Thresholds on the TM-score and S_G produce four sectors in the Cartesian plane. Sector I contains pairs of homologous proteins having a different tertiary fold. Sector II contains pairs of non-homologous proteins with a different tertiary fold. Sector III contains pairs of non-homologous proteins having the same tertiary fold. Sector IV contains pairs of homologous proteins having the same tertiary fold. The total number of protomer pairs analyzed in this work are listed in Table 1. At first glance, it appears that the cloud of points behaves the same in all cases. However, a density analysis reveals differences among all datasets, although more contrasting between cellular proteins and VCPs. Most pairs lie in sector II in the former case (Sander and Schneider, 1991; Rost, 1999), whereas pairs are similarly distributed between sectors II, III, and IV in the later case. Furthermore, ~30% of the characterized VCPs pairs have TM-score values >0.9, notwithstanding the sequence identity value [20–99%].

Even though the percentage of pairs in sector III is low in the case of cellular proteins (<1%), there are several examples where

Table 3
Nonlinear weighted least-squares estimates for the parameters of the exponential model (1-TM-score) ~ f*exp(k*(1-S_G)) using the Gauss-Newton algorithm. Fits to each dataset for scenarios (1-S_G) > 0 (black), (1-S_G) < 0.7 (magenta), and (1-S_G) > 0.7 (orange) are shown in Fig. 2.

	(1-S_G) > 0		(1-S_G) < 0.7		(1-S_G) > 0.7	
	f	k	f	k	f	k
Monomer	0.351	0.768	0.072	1.429	0.393	0.647
Dimers + Oligomers	0.130	1.854	0.060	2.392	0.990	1.390
VCP	0.003	5.912	0.009	3.612	0.003	5.900

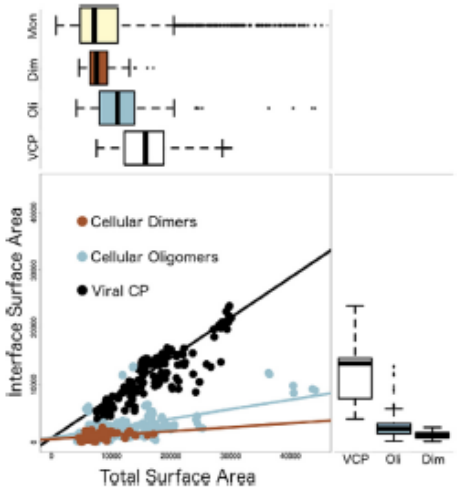


Fig. 3. Correlation between interface surface area and total surface area in proteins, independently estimated for cellular monomers (Mon), cellular dimers (Dim), cellular oligomers (Oli), and icosahedral virus capsid proteins (VCP). In the case of cellular monomers, only the total surface area is analyzed. Linear regression is shown for each set. Statistics summaries (median, first - third quartiles, minimum - maximum values, and outliers) are depicted with boxplots for the interface (right) and the total surface area (top). Area values are in units of Å^2.

Table 1
Count of total protein pairs in each dataset, dissected into sectors (as defined in Fig. 2). Percentage of pairs in each sector in parentheses. Homologous proteins (pairs with a sequence identity S_G > 0.3) are found in sectors I and IV.

	Total pairs	Sector I	Sector II	Sector III	Sector IV
Monomer	12,936,241	103 (0.001%)	12,822,459 (99.10%)	105,459 (0.82%)	8220 (0.064%)
MonomerRND	38,809	0 (0.0%)	38,480 (99.15%)	304 (0.78%)	25 (0.064%)
Dimers + Oligomers	1596 + 5050 = 8290	8 (0.097%)	8037 (97.00%)	133 (1.60%)	112 (1.35%)
VCP	42,778	0 (0.0%)	14,454 (33.79%)	12,348 (28.87%)	15,976 (37.35%)

Table 2
Non-parametric two-sample Kolmogorov-Smirnov test D value (p-value), independently estimated for the total sequence identity (S_G, above the diagonal) and for the structure similarity (TM-score, below the diagonal) distributions of each analyzed dataset. A p-value < 0.05 means that the compared datasets do not come from the same distribution.

TM-score/S_G	Monomer	MonomerRND	Dimers + Oligomers	VCP
Monomer	–	0.002 (0.816)*	0.377 (<2.2e−16)	0.569 (<2.2e−16)
MonomerRND	0.002 (0.601)*	–	0.378 (<2.2e−16)	0.569 (<2.2e−16)
Dimers + Oligomers	0.102 (<2.2e−16)	0.102 (<2.2e−16)	–	0.526 (<2.2e−16)
VCP	0.756 (<2.2e−16)	0.756 (<2.2e−16)	0.692 (<2.2e−16)	–

* p-value ≥ 0.05.

the protein fold is extremely conserved, in spite of a high sequence divergence and different function and organism source, as previously observed (e. g. Holm and Sander, 1994). Surprisingly, the extreme cases found in this study do not involve the ubiquitous Beta Barrel fold (CATH classification 2.40), but the Mainly Beta 7 Propeller and Trefoil fold (CATH classification 2.130 and 2.80 respectively), as shown in Table S4 and Fig. S2.

To test for a sampling imbalance, we constructed a fifth dataset, MonomerRND, by randomly selecting a small percentage of pairs from the cellular monomers dataset. The size of this new dataset is similar to that of the VCPs dataset. Table 2 shows the Kolmogorov-Smirnov test results when comparing all datasets, independently done for the TM-score and S_G. As expected, the test indicates that the MonomerRND dataset follows the same distribution as the cellular monomers dataset. However, all other datasets follow different distributions, even when comparing cellular monomers with higher order cellular *n*-mers.

Chothia and Lesk, 1986 proposed an exponential model to describe the relation between sequence identity and structure similarity. Here, we fit our data to such model, expressed as

$$(1 - \text{TM-score}) \sim f^{*} \exp(k^{*}(1 - S_G))$$

Table 3 shows the values of the computed coefficients of proportionality f and k, considering three different scenarios, i.e., homologous proteins, non-homologous proteins, and the whole S_G range. The differences between sequence scenarios and datasets are also illustrated in Fig. 2.

Table 4
Average protein surface area <*SA*>, in units of Å^2 [percentage of total], with standard deviation *SD*, estimated for the cellular monomers, cellular dimers, cellular oligomers, and icosahedral virus capsid proteins (VCPs) datasets. Values estimated for the total surface and the protein-protein interface with Pearson's correlation *r*. Estimation of the average surface density <*σ*>, in units of residues per 1000 Å^2, for the protein-protein interface and the solvent accessible surface (SAS).

	Total		Interfaces		Correlation	<σ>	
	<SA>	SD	<SA>	SD	r	Interface	SAS
Monomers	9169	7110	0 [00%]	0	–	0	16
Dimers	8318	2597	1073 [13%]	583	0.31	12	12
Oligomers	11,670	5411	2680 [23%]	2141	0.42	13	13
VCPs	16,535	5823	12,260 [74%]	4551	0.90	13	17

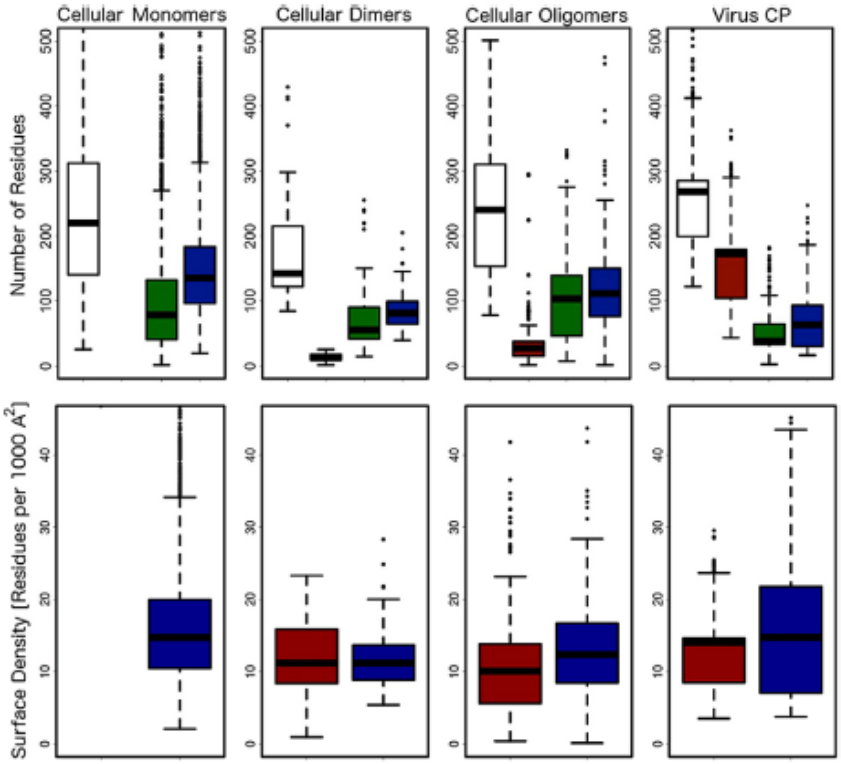


Fig. 4. Distribution of the number of residues (top row) and surface residue density (bottom row) in proteins, independently estimated for cellular monomers (first column), cellular dimers (second column), cellular oligomers (third column), and icosahedral virus capsid proteins (fourth column). Statistics summaries (median, first – third quartiles, minimum – maximum values, and outliers) are depicted with boxplots for total counts (empty), protein-protein interface (red), protein core (green), and proteins solvent accessible surface (blue).

3.2. Distribution of conserved residues

We estimated the protomer's surface area of cellular proteins and VCPs. Fig. 3 shows the distribution of values in each dataset and the correlation between the interface and the total surface area. We found a weak positive correlation in the case of cellular proteins, as opposed to VCPs, for which the correlation is strong. On average, cellular proteins tend to have a smaller total and interface surface area than VCPs (Table 4 and S5). A remarkable difference is that the interface region of VCPs takes 74% of the total surface, whereas for cellular dimers is only around 10%, and around 20% for higher *n*-mers. This observation is directly related to the number of residues comprising the structural categories. On average, cellular monomers and oligomers have the same total number of residues as VCPs (Fig. 4). This value is significantly smaller for cellular dimers (Table 5 and S6). More than half of the total residues make the interfaces of VCPs, whereas a small percentage are interface residues in cellular oligomers. In the case of cellular monomers, 40% of the total residues form the core of the protein. This amount is conserved in cellular dimers and oligomers but is reduced by half in VCPs. The other 60% of the total residues are exposed at the surface of cellular monomers. This amount is reduced by ~10% in the case of cellular dimers and oligomers in order to make the interface region. VCPs present a significantly smaller number of solvent exposed residues, being only about 26% of the total residues on average. Having the values of the surface area and the number of exposed and interface residues, we estimated the surface residue density distribution, σ. The residue density at the interface regions is the same, on average, in both cellular proteins and VCPs (Table 4 and S7). However, σ is significantly higher at the solvent accessible region for cellular monomers than it is for dimers and oligomers. Interestingly enough, VCPs have the same σ value at the solvent accessible area as cellular monomers.

S_G only gives an overall similarity between two amino acid sequences, estimating the relative percentage of residues that are identical in type of amino acid and position in the global alignment. To further investigate how the conserved residues are localized in the protein tertiary structure, we analyzed the distribution of the conserved residues on the different structural categories by estimating the sequence identity index per location category, S^*_k, in all datasets. Fig. 5 shows the correlation and distribution of S^*_k as a function of S_G, independently calculated for the interface, core and surface protein regions. The average values are reported in Table 6. Knowing how the size of the different structural categories contrast between cellular proteins and VCPs, it is not surprising to find that more than half of the conserved residues are in the interface region in the later case. The amount of conserved residues at the interface is low in cellular *n*-mers, given that this is a small region in such proteins. Interestingly, even though the number of residues is higher at the solvent accessible surface compared with

Table 5
Average number of residues <*NR*> [percentage of total], with standard deviation *SD*, in proteins at different structure locations independently estimated for cellular monomers, cellular dimers, cellular oligomers, and icosahedral virus capsid proteins (VCPs).

	Total		Interface		Core		Surface	
	<NR>	SD	<NR>	SD	<NR>	SD	<NR>	SD
Monomers	245	135	–	–	97 [40%]	77	148 [60%]	70
Dimers	178	83	13 [07%]	7	78 [44%]	55	87 [49%]	33
Oligomers	253	120	35 [14%]	35	102 [40%]	65	116 [46%]	65
VCPs	281	119	158 [56%]	63	51 [18%]	36	72 [26%]	47

Table 6
Average percentage of conserved residues at different structure locations $<S^*_k>$, with standard deviation *SD* and Pearson's correlation *r* with respect to the total sequence identity (S_G). Independent estimations made by the analysis of *n* pairs in the cellular dimers plus oligomers and icosahedral virus capsid proteins datasets, for $S_G > 0$, $S_G > 0.3$, or $S_G < 0.3$.

	Cellular dimers + Oligomers				Viral CP			
	n	$<S^*_k>$	SD	*r*	*n*	$<S^*_k>$	SD	*r*
	Interface							
$S_G < 30\%$	3832	14	6	−0.64	2433	42	0.23	0.67
$S_G > 30\%$	59	9	4	0.13	1453	66	0.12	−0.34
$S_G > 0$	3891	13	6	−0.42	3886	52	0.22	0.5
	CORE							
$S_G < 30\%$	3832	41	18	−0.04	2433	23	13	−0.34
$S_G > 30\%$	59	32	1	−0.43	1453	15	6	0.13
$S_G > 0$	3891	41	18	−0.08	3886	20	11	−0.34
	Solvent accessible surface							
$S_G < 30\%$	3832	30	15	−0.06	2433	20	11	−0.36
$S_G > 30\%$	59	38	14	0.32	1453	8	6	0.55
$S_G > 0$	3891	30	15	0.05	3886	12	12	−0.16
	Orphans							
$S_G < 30\%$	3832	15	–	–	2433	–	–	–
$S_G > 30\%$	59	21	–	–	1453	11	–	–
$S_G > 0$	3891	16	–	–	3886	16	–	–

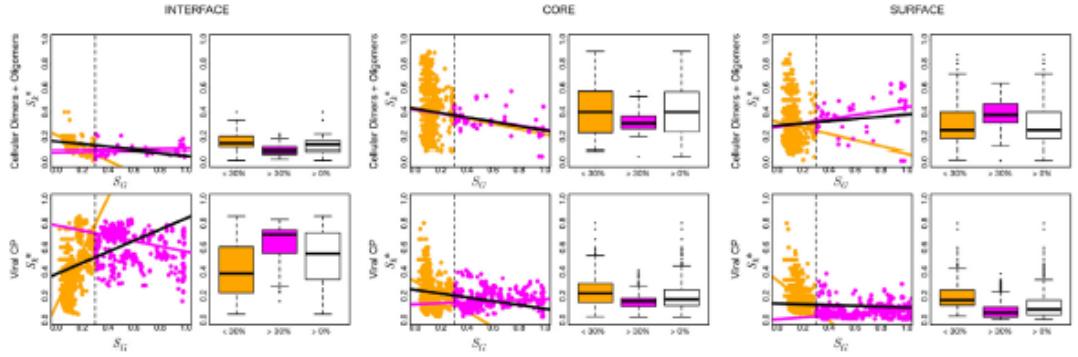


Fig. 5. Correlation between the specific location of conserved residues (S^*_k) and the total sequence identity (S_G) in proteins, at the protein–protein interface, protein core, and proteins solvent accessible surface, independently estimated for cellular dimers plus oligomers (top rows), and icosahedral virus capsid proteins (bottom rows). Each point in the cloud represents a pair of polypeptide chains. Linear regression and statistical analysis are shown for pairs with $S_G > 0$ (black, white), $S_G > 0.3$ (magenta), and $S_G < 0.3$ (orange). Statistics (median, first – third quartiles, minimum – maximum values, and outliers) are summarized with boxplots.

those at the proteiń s buried core (Fig. 4), the amount of conserved residues is significantly higher at the later in both cellular proteins and VCPs.

In addition to considering the whole S_G range to perform the analyses, we also used a 30% sequence identity arbitrary threshold to compare non-homologous vs. homologous protein scenarios. The correlation of S_k^* with respect to S_G is weak to nonlinear in all cases, except for the interface region of non-homologous cellular proteins, which have a strong negative correlation, and non-homologous VCPs, having a strong positive correlation. There is a variation in the average amount of conserved residues at all structure categories for cellular proteins and VCPs, although the relative behavior is different. In the case of cellular *n*-mers, S_k^* is higher for non-homologous proteins at the interface and core than homologous proteins. The opposite is seen at the solvent exposed surface. In the case of VCPs, S_k^* is higher for non-homologous proteins at the core and the solvent exposed surface than homologous proteins. The opposite is seen at the interface. We found another interesting difference between cellular *n*-mers and VCPs. Whereas the amount of the percentage of conserved residues in the orphan category increases for homologous proteins with respect to non-homologous in cellular *n*-mers, the opposite is seen in VCPs (Table 6).

3.3. Residue conservation in the structural categories

We estimated the entropy-based conservation by residue $S(i)$ for each protomer in our datasets. Table 7 shows the average residue conservation <S> and the normalized value <s> calculated by structural category. In the case of cellular proteins, the normalized residue conservation does not change much from the non-normalized residue conservation since <S> for the whole polypeptide chain is approximately 1. In general, the core and interface regions are significantly more conserved than the solvent exposed surface, with some preference for the buried volume, specially at cellular oligomers (Table S8).

On the other hand, we found that the probability distribution of <S> for the VCPs dataset is bimodal, as illustrated in Fig. 6. We

Table 7
Average entropy-based residue conservation estimation per polypeptide chain, <S>, and normalized residue conservation, <s>, in proteins, independently estimated for cellular monomers, cellular dimers, cellular oligomers, and icosahedral virus capsid proteins (VCPs). Average residue conservation estimated for the protein-protein interface, protein core, and proteiń s solvent accessible surface (SAS). Standard deviation indicated in parentheses.

	Monomers		Dimers		Oligomers		VCPs		
	<s>	<S>	<s>	<S>	<s>	<S>	<s>†	$\langle S\rangle_{G1}$ ††	$\langle S\rangle_{G2}$ ††
Interface	–	–	0.91 (0.27)	0.98 (0.44)	0.94 (0.14)	0.92 (0.34)	0.96 (0.13)	0.41 (0.17)	0.99 (0.16)
Core	0.78 (0.10)	0.83 (0.31)	0.83 (0.15)	0.84 (0.26)	0.77 (0.09)	0.75 (0.26)	0.75 (0.12)	0.28 (0.09)	0.84 (0.10)
SAS	1.23 (0.10)	1.29 (0.41)	1.26 (0.18)	1.28 (0.37)	1.29 (0.13)	1.23 (0.34)	1.29 (0.12)	0.56 (0.21)	1.24 (0.18)

† Values considering one single probability distribution.
†† Values assuming two independent probability distributions.

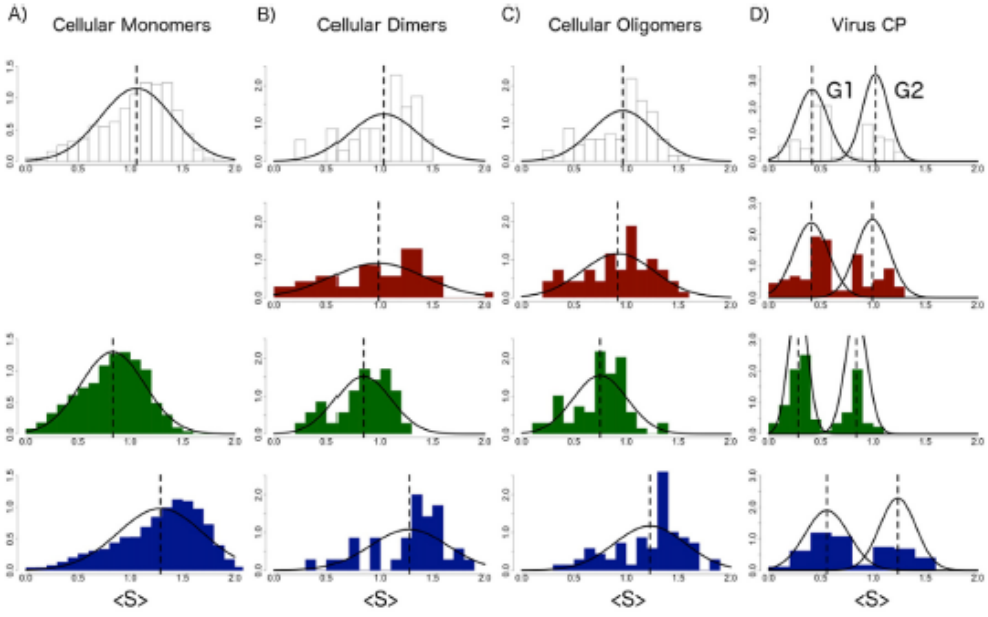


Fig. 6. Probability distribution of the average entropy-based residue conservation estimation per polypeptide chain, <S>, in proteins, independently estimated for cellular monomers (A), cellular dimers (B), cellular oligomers (C), and icosahedral virus capsid proteins (D). Average residue conservation estimated for the whole polypeptide chain (white), protein-protein interface (red), protein core (green), and protein's solvent accessible surface (blue). Normal distribution with the same mean (vertical dashed line) and standard deviation of the probability distribution is shown for each case.

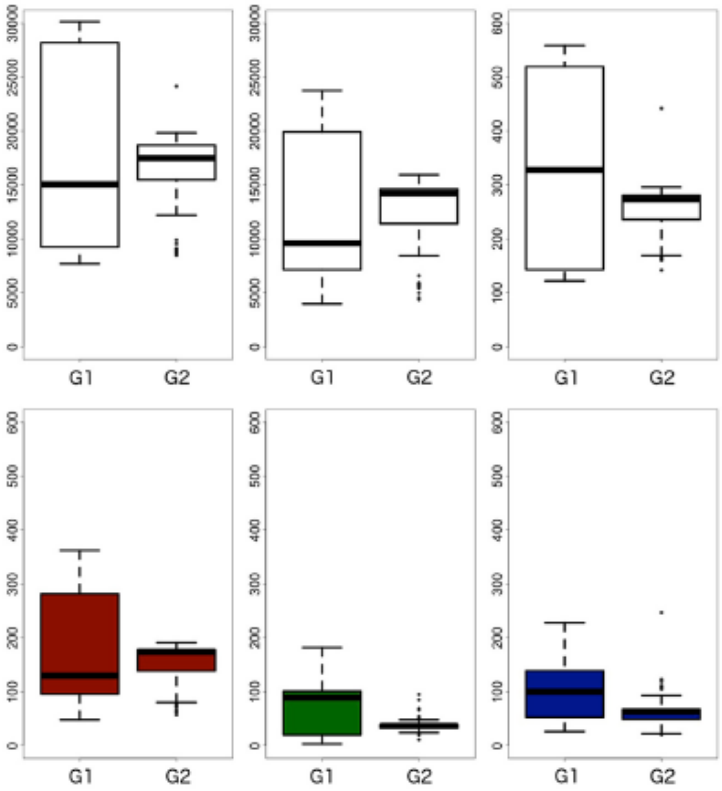


Fig. 7. Statistical comparison between virus families group 1 and group 2 (G1 and G2 respectively). Characterization of (top row, from left to right) total solvent accessible surface area (SASA, in Å^2), interface SASA, total number of residues per polypeptide chain, (bottom row, from left to right) number of residues at the protein–protein interface (red), protein core (green), proteins solvent accessible surface (blue). Statistics (median, first – third quartiles, minimum – maximum values, and outliers) are summarized with boxplots.

labeled the distinct distributions as G1 and G2. It is interesting that the average residue conservation per structure category of G2 behaves as the cellular proteins. However, G1 seems to be evolving much slower. We identified the virus families that belong to each group as: Adenoviridae (T = 1), Birnaviridae (T = 1), Bromoviridae (T = 3), Caliciviridae (T = 3), Comoviridae (T = pT3), Hepadnaviridae (T = 4), Hepeviridae (T = 1), Leviviridae (T = 3), Microviridae (T = 1), Nodaviridae (T = 3), Parvoviridae (T = 1), Polyomaviridae (T = 7d), Sobemoviridae (T = 3), and Tetraviridae (T = 4) in G1, and Dicistroviridae (T = p3), Picornaviridae (T = pT3), Siphoviridae (T = 7 L), Sobemoviridae (T = 3), Togaviridae (T = 4), Tombusviridae (T = 3), and Tymoviridae (T = 3) in G2. There is no obvious correlation with the T number. Both groups have, on average, the same total and interface solvent accessible surface area, and the same number of residues making the interface region. However, one significant difference is in the number of residues comprising the core and the exposed surface, with G1 having almost twice as many as G2 (Fig. 7 and Table S9).

4. Discussion

Our findings reveal several differences between cellular protein oligomers (*n*-mers) and icosahedral viral capsid proteins regarding the location, amount, and level of conservation of residues in the tertiary structure. We analyzed and compared four datasets, namely, cellular monomers, cellular dimers, cellular oligomers, and VCPs. Overall, cellular monomers and VCPs seem to be two extremes in the quaternary structural diversity found in nature, with the cellular dimers and oligomers as an intermediate state. Our main results are summarized in Fig. 8.

The correlation between the sequence identity and the tertiary structure conservation seems to follow an exponential model in all cases (Chothia and Lesk, 1986). However, their distribution in [S_G, TM-score] space is quite different (Fig. 2). Cellular monomers are concentrated in a high density, non-homologous, different fold region. On the other hand, VCPs are evenly distributed in three main regions along the whole range of sequence homology and tertiary structure similarity. Cellular dimers and oligomers, although similar to cellular monomers, show a distinct distribution that seems to be in between cellular monomers and VCPs.

We performed a deeper examination of the correlation between protein global sequence identity (S_G) and structural similarity (TM-score) for the case of VCPs. A large majority of pairs having TM-score values >0.7 are capsid proteins that belong to the same virus family. Likewise, most pairs having TM-score values <0.7 are capsid proteins that belong to different virus families. Of note, ICTV taxonomic classification of viruses does not explicitly include structural information as criteria to group viruses. Interestingly,

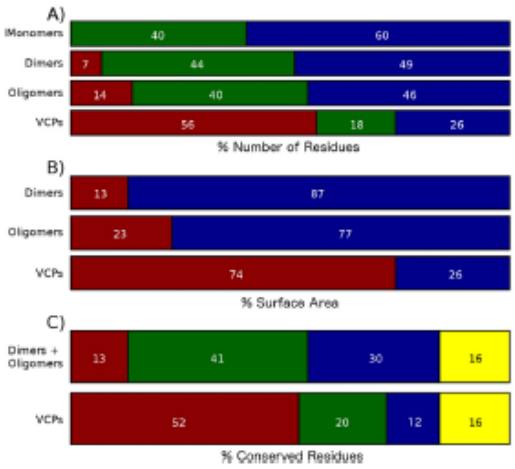


Fig. 8. Average values for the percentage of number of residues (A), the percentage of surface area (B), and the percentage of conserved residues (C) in proteins, at the protein-protein interface (red), protein core (green), protein's solvent accessible surface (blue), and orphans (yellow), independently estimated for cellular monomers, cellular dimers, cellular oligomers, and icosahedral virus capsid proteins (VCPs).

there is a considerable proportion of VCP pairs in the TM-score range between 0.5 and 0.7. They are inter-family viruses that have a global sequence identity below 20%. This observation suggests a substantial sequence divergence during virus evolution. We speculate that virus families have evolved from a few common ancestors, a hypothesis also proposed recently by Nasir and Caetano-Anollés, 2015. Even though some of the comparisons were made between homologous VCPs within the same family, sometimes with sequence identities reaching 90%, these viruses are distinct and display unique virological/serological properties. Importantly, different viruses within the same family can cause distinct diseases (e.g., polio vs. common cold). Hence, we believe that the comparisons made in this study between homologous VCPs are appropriate.

It is important to remark that despite this sequence divergence, the protein fold (e.g., jelly-roll β-barrel) persists and that the majority of conserved residues remain at the protein-protein interfaces. This observation is in agreement with previous conclusions drawn from different analysis approaches regarding evolution and structure conservation. Abroi and Gough, 2011, suggested that the virosphere could be an engine for the genesis of cellular protein structures. Cheng and Brooks, 2013, mention that the structural, but not functional, close relationship found between some classes of modern cellular proteins and VCPs resulted from ancient genetic interactions between viruses and their hosts. In this sense, our results are yet another indication that VCPs are a good model for further investigation on sequence divergence due to pressures rendered by the host immune system.

The distribution of conserved residues in the protein tertiary structure is also different between cellular proteins and VCPs. On average, 41% of the conserved residues in cellular dimers and oligomers are found at the buried core, followed by 30% at the exposed surface, and 13% at the interface region. However, more than half of the total conserved residues of VCPs are found at the interface, followed by 20% at the core, and only 12% at the surface. About 15% of the conserved residues could not be matched to a common structural category and were classified as orphans (ORPH). We found that the distribution of conserved residues is different in cellular proteins compared to VCPs. This observation is correlated to the number of residues making the different protein structure categories. This result suggests that the conserved residues are evenly distributed over the whole protein structure in all cases. Again, we can see that the cellular dimers and oligomers have intermediate values between those of cellular monomers and VCPs (Fig. 4).

Because icosahedral VCPs self-assemble with multiple neighbors to form capsid shells, three-quarters of the protein surface is used to make the protein-protein interfaces, with half of the total conserved residues at this region. This finding is in agreement with evidence showing that mutations in virus capsids mostly take place at the solvent accessible surface, as a way of countering/evading host immune responses (e.g. Jameson et al., 1985; Kanda et al., 1986; Yang et al., 2005; Vitiello et al., 2005) in an evolutionary positive selection manner (Esteves et al., 2008). Of note, a perfect correlation can be seen between the residue structure classification and conservation analysis done in this work with the level of conservation and structural features of spherical capsids recently reported by Chih-Min et al., 2015 (Fig. S3).

The distribution of the variation on the level of residue conservation in different structural categories in the protein is similar in all cellular protein *n*-mers. On average, the buried core is the most conserved, closely followed by the interface region (Grishin and Phillips, 1994; Valdar and Thornton, 2001; Guharoy and Chakrabarti, 2005). There is a greater sequence variation at the solvent exposed surface in all cellular proteins (Caffrey et al., 2004). This relation is also true for some VCPs. We clearly identified two distinct groups of virus families that behave differently concerning sequence variation (see *Residue conservation in the structural categories* in the Results section). One of these groups, G2, has residue conservation average values very similar to those of cellular proteins. However, the second group, G1, seems to have significantly lower residue variations (Table S10), although the relative differences between structure categories remain the same as in all other cases. We found that a difference between G1 and G2 is the number of residues making the buried core and the solvent exposed surface, with the later having significantly lower values. Bahadur and Janin, 2008 analyzed the residue conservation of 32 icosahedral viruses and reported normalized values, <s>, for the interface, core, and surface of 0.9, 0.7 and 1.6, respectively. We can reproduce those results if we assume a single probability distribution of <S>, as can be seen in Table 7 and Fig. S4. Their approach and small dataset precluded the realization that their results were the average of two distinct distributions. The reason and the meaning of the existence of these two groups of virus families deserve further investigation.

The high-resolution cutoff criteria used for assembling our datasets provides high confidence in the reported results. Hence, these datasets are a good representative sample of nature's protein diversity. Our analyses can readily be extended later as more structural data becomes available. In this work, we have included all the icosahedral virus structures available to date. Other viral capsid topologies, such as helicoidal viruses, were not included due to a limited small number of structures available, but will be considered in future studies.

5. Conclusions

Our work extends and complements results previously reported. We find a general agreement regarding sequence variations occurring at different regions of the tertiary structure of cellular proteins and spherical virus capsid proteins. However, we could detect two distinct virus family groups with seemingly

different evolution rates. The robust statistical analyses performed on high-resolution structural datasets had the power to highlight important differences between cellular proteins and spherical VCPs. Our results provide evidence that cellular monomers and VCPs are two extremes in the quaternary space of protein complexes, with the cellular dimers and oligomers being an intermediate state.

Acknowledgements

We acknowledge the thoughtful suggestions of the editor and reviewers which greatly improved this article. V.S.R and M.C.T. would also like to acknowledge the discussions with Professor Charles L. Brooks III at the inception of this work. M.C.T. would like to thank Dr. Johan Van Horebeek from the Computer Science Department, CIMAT, México, for his helpful advice on the statistical methodology. This work was supported by the USA National Institutes of Health (NIH) to the center of Multi-scale modeling tools for structural biology (MMTSB) grant number RR012255 to V.S.R., the Mexican Consejo Nacional de Ciencia y Tecnología (Conacyt) grant number 132376 to M.C.T., and the 2013 Fulbright-García Robles funding to D.J.M.-G and M.C.T. by the USA J. William Fulbright Scholarship Board.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2016.07.013.

References

Abroi, A., Gough, J., 2011. Are viruses a source of new protein folds for organisms? – virosphere structure space and evolution. BioEssays 33, 626–635.
Bahadur, R.P., Janin, J., 2008. Residue conservation in viral capsid assembly. Proteins 71, 407–414.
Bahadur, R.P., Rodier, F., Janin, J., 2007. A dissection of the protein–protein interfaces in icosahedral virus capsids. J. Mol. Biol. 367, 574–590.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G.T., Bhat, N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.
Caffrey, D.R., Somaroo, S., Huges, J.D., Mintseris, J., Huang, E.S., 2004. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. 13, 190–202.
Cann, J.A., 2005. Principles of Molecular Virology, fourth ed. Elsevier Academic Press.
Carrillo-Tripp, M., Shepherd, C.M., Borelli, I.A., Venkataraman, S., Natarajan, P., Johnson, J.E., Brooks, C.L., Reddy, V.S., 2009. VIPERdb2: an enhanced and web API enabled relational database for structural virology. Nucleic Acids Res. 37, D436–D442.
Caspar, D.T., Klug, A., 1962. Physical principles in the construction of regular viruses, first ed., 27. Press, Cold Spring Harbor Laboratory.
Cheng, S., Brooks III, C.L., 2013. Viral capsid proteins are segregated in structural fold space. PLoS Comput. Biol. 9, e1002905.
Chih-Min, C., Yu-Wen, H., Tsun-Tsao, H., Chung-Shiuan, S., Jenn-Kang, H., 2015. Sequence conservation, radial distance and packing density in spherical viral capsids. PLoS ONE 10, e0132234.
Chotia, C., 1974. Hydrophobic bonding and accessible surface area in proteins. Nature 254, 338–339.
Chotia, C., Janin, J., 1975. Principles of protein–protein recognition. Nature 256, 705–708.
Chotia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5, 823–826.
Damodaran, K., Reddy, V.S., Johnson, J.E., Brooks III, C.L., 2002. A general method to quantify quasi-equivalence in icosahedral viruses. Mol. Biol. 324, 723–737.
Esteves, P.J., Abrantes, J., Carneiro, M., Müller, A., Thompson, G., van der Loo, W., 2008. Detection of positive selection in the major capsid protein VP60 of the rabbit haemorrhagic disease virus (RHDV). Virus Res. 137, 253–256.
Fauquet, C., Mayo, M.A., Maniloff, J., Desselberger, U., Ball, L.A., 2005. Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press.
Grishin, N.V., Phillips, M.A., 1994. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. Protein Sci. 3, 2455–2458.
Guharoy, M., Chakrabarti, P., 2005. Conservation and relative importance of residues across protein-protein interfaces. PNAS 102, 15447–15452.
Holm, L., Sander, C., 1994. Structural similarity of plant chitinase and lysozymes from animals and phage. An evolutionary connection. FEBS lett. 340, 129–132.
Jameson, B.A., Bonin, J., Wimmer, E., Kew, O.M., 1985. Natural variants of the Sabin type 1 vaccine strain of poliovirus and correlation with a poliovirus neutralization site. Virology 143, 337–341.
Janin, J., Bahadur, R.P., Chakrabarti, P., 2008. Protein-protein interaction and quaternary structure. Quat. Rev. Biophys. 41, 133–180.
Jones, S., Thornton, J.M., 1995. Protein-protein interactions: a review of protein dimer structures. Prog. Biophys. Mol. Biol. 63, 31–65.
Kanda, T., Furuno, A., Yoshiike, K., 1986. Mutation in the VP-1 gene is responsible for the extended host range of a monkey B-lymphotropic papovavirus mutant capable of growing in T-lymphoblastoid cells. J. Virol. 59, 531–534.
Lawrence, M.C., Colman, P.M., 1993. Shape complementarity at protein/protein interfaces. J. Mol. Biol. 234, 946–950.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
Nasir, A., Caetano-Anollés, G., 2015. A phylogenomic data-driven exploration of viral origins and evolution. Sci. Adv. 1, e1500527.
Ofran, Y., Rost, B., 2003. Analysing six types of protein–protein interfaces. J. Mol. Biol. 325, 377–387.
Richards, F.M., 1974. The interpretation of protein structures: total volume, group volume distributions and packing density. J. Mol. Biol. 82, 1–14.
Rost, B., 1999. Twilight zone of proteins sequence alignments. Protein Eng. 12, 85–94.
Sander, C., Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: structure. Funct. Genet. 9, 56–68.
Shrake, A., Rupley, J.A., 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79, 351–364.
Talavera, D., Robertson, D.L., Lovell, S.C., 2011. Characterization of protein–protein interaction interfaces from a single species. PLoS One 6, e21053.
Tokuriki, N., Oldfield, C.J., Uversky, V.N., Berezovsky, I.G., Tawfik, D.S., 2009. Do viral proteins possess unique biophysical features? Trends Biochem. Sci. 34, 53–59.
Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., Joosten, R.P., Vriend, G., 2015. A series of PDB related databases for everyday needs. Nucleic Acids Res. 43 (Database issue), D364–D368.
Valdar, W.S.J., Thornton, J.M., 2001. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins Struct. Funct. Genet. 42, 108–124.
Vitiello, C.L., Merril, C.R., Adhya, S., 2005. An amino acid substitution in a capsid protein enhances phage survival in mouse circulatory system more than a 1000-fold. Virus Res. 114, 101–103.
Xu, J., Zhang, Y., 2010. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26, 889–895.
Yan, C., Wu, F., Dobbs, D., Honavar, V., 2008. Characterization of protein–protein interfaces. Protein 27, 59–70.
Yang, R., Wheeler, C.M., Chen, X., Uematsu, S., Takeda, K., Akira, S., Pastrana, D.V., Viscidi, R.P., Roden, R.B.S., 2005. Papillomavirus capsid mutation to escape dendritic cell-dependent innate immunity in cervical cancer. J. Virol. 79, 6741–6750.
Zhang, Y., Skolnick, J., 2004. Scoring function for automated assessment of protein structure template quality. Proteins: Struct. Funct. Bioinf. 57, 702–710.
Zhang, Y., Skolnick, J., 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309.