Tesis de Maestría en Bioinformática

Identificación de proteínas presentes en vesículas extracelulares con potencial fusogénico basado en su similitud con fusógenos virales

Daniela Megrian

Tutores **Pablo Aguilar**

Federico Lecumberry

15 de mayo de 2017

Este trabajo fue financiado por una Beca de Posgrado Nacional de la ANII.

RESUMEN

Existe una amplia variedad de procesos biológicos donde la fusión de membranas es esencial. Las membranas celulares no fusionan espontáneamente sino que este proceso es catalizado por proteínas con capacidad fusogénica.

En este trabajo se propuso desarrollar métodos informáticos de identificación de proteínas con capacidad fusogénica. Debido a su reciente reconocimiento como proceso biológico y a su importancia en la salud humana decidimos tomar a las vesículas extracelulares (VEs) como objeto de estudio. Esto se debe a que si bien existe evidencia que sostiene a la fusión de membranas entre las VEs y la célula blanco como paso necesario para la entrega de su contenido, aún se desconoce qué proteínas llevan a cabo la fusión.

Dentro de las proteínas de fusión de membranas en el medio extracelular, las proteínas de fusión viral han sido ampliamente estudiadas, y dada su estructura tridimensional y su mecanismo molecular se han clasificado en al menos tres clases: clase I, clase II y clase III. Los fusógenos conocidos que median la fusión célula-célula presentan homología con fusógenos virales pero presentan una gran diversidad a nivel de secuencia. Por esta razón análisis que busquen similitud a nivel de secuencia no son capaces de identificar homología entre ellas. Sin embargo, se ha reportado que existen patrones de similitud a nivel de estructura secundaria dentro de cada clase de fusógenos virales.

Por esta razón, el primer paso en este trabajo fue determinar una métrica de similitud a nivel de estructura secundaria que permita clasificar fusógenos virales de acuerdo a su clase, utilizando algoritmos de aprendizaje automático. A partir de esta métrica, se clasificaron las proteínas identificadas en VEs como similares a fusógenos virales de clase I, clase II o clase III, utilizando algoritmos de aprendizaje automático del tipo one-class classification.

La métrica desarrollada utilizó como punto de partida la métrica propuesta por Przytycka et al. y permitió cumplir con los objetivos de este trabajo.

Los resultados de la clasificación de proteínas de VEs como potenciales fusógenos presentan una cantidad importante de proteínas candidatas. A pesar de que en conjunto se pudo comprobar que los clasificadores seleccionaron proteínas que tuvieran características similares en la composición de estructura secundaria con la que fueron entrenados, la mayoría de las candidatas no cumplen con características esenciales conocidas para las proteínas de fusión viral. Estas características están asociadas a su topología.

A partir del análisis homología lejana de las proteínas candidatas se identificó a ZP2 de la zona pellucida como la proteína más interesante. De acuerdo a la bibliografía se cree que ZP2 podría actuar como un receptor secundario, pero se identificaron patrones en la proteína que sugieren que podría actuar como fusógeno. Presenta una longitud similar a los fusógenos virales de clase II y la homología se identificó en una región poco conservada (respecto al dominio inmunoglobulina) de la proteína. Otro punto en común es que presenta un dominio inmunoglobulina en la porción C-terminal del ectodominio.

Será necesario hacer una búsqueda bibliográfica más a fondo sobre ZP2 con el objetivo de diseñar experimentos que permitan validar este resultado.

ÍNDICE

INTRODUCCIÓN

1. Fusión de membranas

1.1. Definición

La existencia de las membranas biológicas fue un prerrequisito para la existencia de todas las formas de vida permitiendo la separación de la naturaleza en distintos compartimientos. Estas barreras compuestas por una bicapa de fosfolípidos son continuas y estructuralmente estables pero a su vez semi-permeables. De esta forma pueden compartimentar, pero también permiten la comunicación y el transporte molecular entre los compartimientos.

Las membranas celulares encapsulan componentes en el citoplasma, permitiendo el enriquecimiento de ciertos metabolitos y restringiendo la entrada de agentes extraños. Las células eucariotas presentan además membranas intracelulares dando lugar a organelas. Estas generan subdivisiones internas que permiten la compartimentalización de procesos metabólicos y replicativos. Sin embargo estas barreras a veces deben ser perturbadas de forma de fusionarse. Como se describe más adelante, procesos esenciales para la vida celular tales como el tráfico de material e información, la reproducción sexual y el normal funcionamiento de organelas como las mitocondrias o el retículo endoplásmico precisan de la fusión de membranas.

Este proceso ocurre cuando dos membranas lipídicas se unen para formar una bicapa continua única. Independientemente de la naturaleza de las membranas a fusionarse, el proceso consta de tres etapas principales: contacto cercano de membranas, unión de capas proximales o hemifusión, y finalmente la unión de las capaz distales, lo cual produce la apertura de un poro de fusión (Chernomordik & Kozlov, 2008).

1.2. Mecanismos

La fusión comienza a partir del contacto entre dos membranas muy cercanas (Fig. 1.A), cuando una protrusión minimiza la energía de repulsión de hidratación entre sus monocapas proximales dando lugar a un contacto inmediato (Fig. 1.B). Este contacto, dado por las monocapas muy curvadas, da lugar al tallo de hemifusión (Fig. 1.C). En esta estructura las dos monocapas proximales de las membranas lipídicas de encuentran fusionadas, mientras que las monocapas internas permanecen separadas. Así, la hemifusión consiste en el mezclado de lípidos sin mezclar el contenido. Posteriormente, el intermediario de hemifusión puede sufrir una expansión simétrica axial, dando lugar a un diafragma de hemifusión (Fig. 1.D). En el diafragma las monocapas distales de las membranas forman una bicapa lipídica común. Los contenidos lumenales se mantienen separados hasta este punto, en el cual finalmente se produce la apertura de un poro en el diafragma que permite la conexión entre los compartimientos delimitados por las membranas (Fig. 1.E). El poro de fusión es una conexión entre membranas que se combinan, incluyendo las monocapas externas e internas. La formación del poro de fusión establece una conexión acuosa entre los volúmenes inicialmente separados por membranas yuxtapuestas (Chernomordik & Kozlov, 2008).

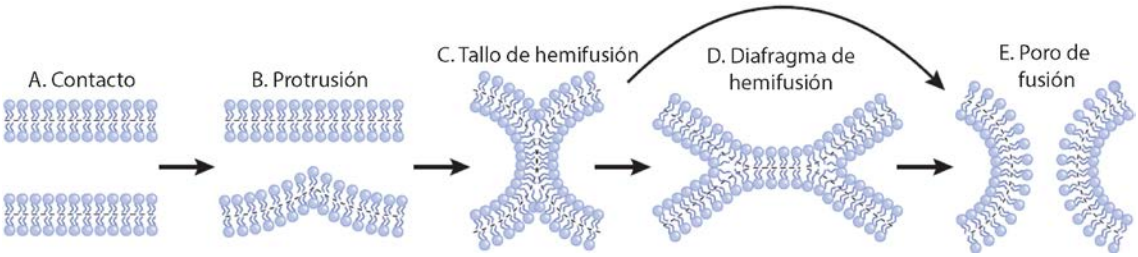


Figura 1. Fusión de membranas. Adaptada de Chernomordik, 2008. A. Contacto de pre-fusión. B. Una protrusión de membrana minimiza la energía de repulsión de hidratación entre las monocapas proximales de las membranas en contacto inmediato. C. Tallo de hemifusión con monocapas proximales fusionadas y monocapas distales sin fusionar. D. La expansión del tallo da lugar al diafragma de hemifusión. E. Formación de un poro de fusión en el diafragma, o directamente a partir del tallo.

1.3. Eventos de fusión

Se conocen diversos eventos celulares en los que es necesaria la fusión de membranas. Ocurre entre células durante la fecundación (Primakoff & Myles, 2007) y el desarrollo de tejidos sincitiales en organismos multicelulares (Rochlin, Yu, Roy, & Baylies, 2010), entre organelas tales como retículo endoplásmico y mitocondrias (Chan, 2006) y entre vesículas y organelas durante el tráfico intracelular (Chen & Scheller, 2001). La fusión también es esencial durante la invasión de virus envueltos a células hospederas (Harrison, 2015).

Mientras que la fusión de membranas es una reacción exotérmica global, es decir que es energéticamente favorable, hay varias barreras que deben superarse para alcanzar este estado. Por esta razón, las membranas celulares no fusionan espontáneamente sino que este proceso es catalizado por proteínas transmembrana de fusión denominadas fusógenos que bajan la barrera de energía libre de la reacción de rearreglo y unión de las bicapas lipídicas. Es decir que la fusión de membranas depende del acoplamiento termodinámico entre las perturbaciones en las membranas blanco y el cambio conformacional de las proteínas de fusión (Rand & Parsegian, 1984).

Si bien las identidades y mecanismos de acción de distintos fusógenos como los virales y aquellos implicados en el tráfico intracelular han sido develados e intensamente estudiados, se tiene poco conocimiento sobre las proteínas que median la fusión entre células. A continuación se describen algunos ejemplos relevantes de procesos celulares que involucran fusión de membranas. Esta descripción no pretende ser exhaustiva, por revisiones recientes referirse a (Aguilar et al., 2013), (Willkomm & Bloch, 2015),(Harrison, 2015).

1.3.1. Eventos de fusión de membranas intracelulares

1.3.1.1. Fusión de vesículas durante el tráfico intracelular

En las células eucariotas, las vesículas intracelulares se desprenden de una membrana y se fusionan con otra, transportando componentes de membrana y moléculas solubles, lo cual denominamos cargo. El lumen de cada vesícula es topológicamente equivalente a la mayoría de

los compartimentos membranosos celulares y al exterior celular. La fusión de vesículas permite entonces que el cargo no tenga que atravesar membranas durante su transporte.

Los eventos de fusión durante el tráfico intracelular son mediados por proteínas denominadas SNARE (soluble N-ethyil-maleimide-sensitive factor attachment protein receptors). Se cree que estas median la fusión de membranas en todos los pasos de la vía secretora (Jahn & Scheller, 2006).

Estas proteínas conforman una superfamilia de proteínas complementarias muy conservadas que deben estar presentes en ambas membranas destinadas a fusionar. Originalmente se planteó la distinción entre proteínas v-SNARE como aquellas ancladas a la membrana de la vesícula a fusionar o que actúa como "dador", y t-SNARE como las ancladas a la membrana blanca o "aceptor"(Sollner, Bennett, Whiteheart, Scheller, & Rothman, 1993). Sin embargo esta terminología no describe los eventos de fusión homotípica o la función de proteínas SNARE que participan en etapas de tráfico anterógrado y retrógrado (Burri et al., 2003; Dilcher et al., 2003).

Las proteínas SNARE presentan un motivo característico denominado SNARE que consiste en una región de 60 a 70 aminoácidos organizados en héptadas repetidas (HR). En sus extremos C-terminales presentan un dominio transmembrana simple unido al motivo SNARE a través de un fragmento corto (Fasshauer, 2003).

Cuando las proteínas SNARE se encuentran en forma monomérica sus motivos SNARE no presentan una estructura definida. Sin embargo, cuando son combinadas, los motivos se asocian formando complejos super-helicoidales muy estables (Fig. 2). Estos complejos están formados por cuatro alfa-hélices paralelas superenrolladas, correspondientes cada una a un motivo SNARE distinto, aunque también es posible formar un complejo menos estable a partir de otras combinaciones. Los motivos son clasificados en Qa, Qb, Qc y R-SNARE. Las proteínas con motivo R-SNARE presentan un residuo de arginina (R) en el domino SNARE. Las proteínas Q-SNARE se caracterizan por un residuo de glutamina (Q) muy conservado. Estas subfamilias se encuentran muy conservadas en la naturaleza y divergieron tempranamente en la evolución eucariota (Fasshauer, 2003).

Su región N-terminal no es tan conservada, permitiendo distintos tipos de plegamiento. Estos dominios pueden variar en longitud y están unidos al motivo SNARE por una región flexible. No se conoce con claridad la función de este dominio y su esencialidad. Se propone que puede actuar como plataforma de reclutamiento para la unión de otras proteínas necesarias para la fusión (Fasshauer, 2003).

En el ejemplo más típico de fusión mediada por SNAREs, existen tres proteínas Q-SNARE en la membrana aceptora y la proteína R-SNARE en la vesícula. Los motivos Q-SNARE en forma de complejo interactúan con el R-SNARE, resultando en la trans-oligomerización de las cuatro hélices ancladas en membranas opuestas. Este complejo en configuración trans pasa de un estado flojo a un estado apretado, es un proceso denominado de cremallera. Esto es seguido por un cambio conformacional que empuja las dos membranas hacia sí, pasando el complejo a una configuración cis, dando lugar a la formación del poro de fusión. El complejo SNARE puede ser reciclado a través de la disociación de la hélice superenrollada, proceso mediado por la proteína NSF (N-ethylmaleimide-sensitive factor)(Fasshauer, 2003).

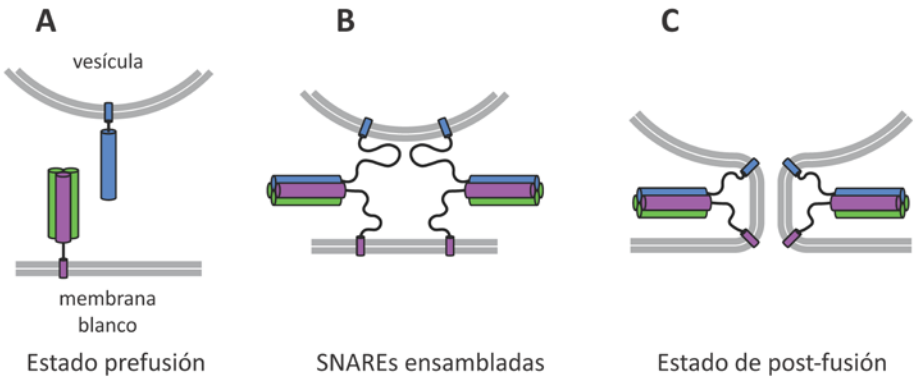


Figura 2. Mecanismo de fusión mediado por proteínas SNAREs. A. Proteínas Q-SNARE en la membrana blanco y proteína R-SNARE en la membrana blanco. B. Formación de complejo super-helicoidal. C. Formación del poro de fusión.

1.3.1.2. Fusión entre organelas

La fusión de membranas de organelas de la vía secretora y de los sistemas endosomal y lisosomal dependen de las proteínas SNARE. En la vía secretora se produce la fusión homotípica o heterotípica de membranas que conforman el retículo endoplásmico y sistema de Golgi. Por su parte, los endosomas tempranos se fusionan con endosomas tardíos o lisosomas para permitir el reciclaje o la degradación de sus contenidos. También se identificó que estas proteínas SNARE son requeridas para la fusión de vacuolas en levaduras, equivalentes de los lisosomas de mamíferos.

Sin embargo la fusión de membranas en mitocondrias presenta un mecanismo distinto. Las mitocondrias son organelas dinámicas que se fusionan y dividen de forma permanente en la mayoría de las células eucariotas (Chan, 2006). La fusión o división depende del nivel de nutrientes, el tipo de células y la función que cumplen. Su fusión consiste en la unión de las membranas exteriores y de las interiores de dos mitocondrias distintas (Fig. 3). Se identificaron algunas proteínas que controlan estos procesos las cuales corresponden a GTPasas pertenecientes a la superfamilia de las dinaminas. Estas son las mitofusinas (en particular Mfn1 y Mfn2) que participan en la fusión de la membrana exterior (Santel & Fuller, 2001) y las proteínas OPA1 y Mgm1 (Griparic, van der Wel, Orozco, Peters, & van der Bliek, 2004) que participan en la fusión de la membrana interior en mamíferos y levaduras respectivamente.

Las mitofusinas son proteínas que atraviesan la membrana mitocondrial externa dos veces, de forma que el extremo C-terminal y N-terminal están ubicados en el citoplasma. La región C-terminal de una mitofusina interacciona con otra copia de la proteína en una mitocondria adyacente. Para que ocurra la fusión es necesaria la hidrólisis de GTP mediada por el dominio GTPasa ubicada en la región N-terminal (Ishihara, Eura, & Mihara, 2004). A pesar de haber identificado estas proteínas, aún se desconoce el mecanismo por el cual la hidrólisis de GTP desencadena el cambio conformacional que produce la fusión de las membranas.

OPA1 y Mgm1 se localizan en la membrana interior y son requeridas únicamente para la fusión de estas membranas. Tampoco se conoce el mecanismo por el cual fusionan las membranas, ya que de acuerdo a los mecanismos conocidos las crestas deberían fusionarse alterando la estructura mitocondrial.

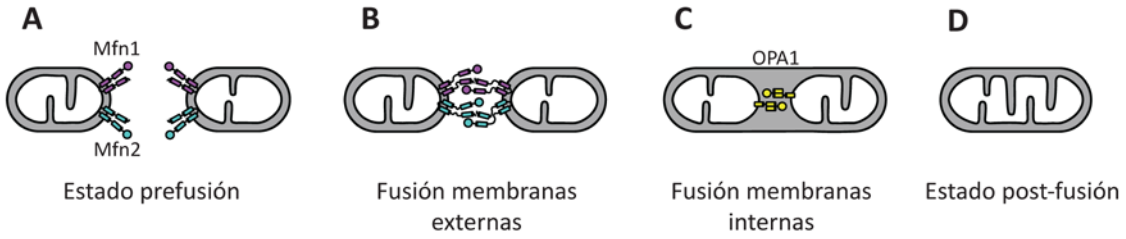


Figura 3. Fusión de mitocondrias. A. Proteínas mitofusinas Mfn1 y Mfn2 presentes en la membrana externa de ambas mitocondrias a fusionar. B. Interacción de mitofusinas de mitocondrias adyacentes. C. Interacción de proteínas OPA1 de las membranas internas de las mitocondrias a fusionar. D. Mitocondria generada a partir de la fusión de dos mitocondrias.

De manera similar a lo que ocurre entre mitocondrias, GTPasas pertenecientes a la superfamilia de las dinaminas llamadas atlastinas median la fusión homotípica entre membranas del retículo endoplásmico (Orso et al., 2009).

1.3.2. Eventos de fusión de membranas en el medio extracelular

Dentro de los eventos de fusión de membranas que se producen en el medio extracelular podemos distinguir al menos cuatro situaciones: la fusión entre un virus envuelto y su célula blanco, la fusión entre dos células, la fusión de una célula consigo misma (auto-fusión) y la fusión entre una vesícula extracelular (VE) y su célula blanco. Los mecanismos de fusión mediados por virus han sido ampliamente estudiados, mientras que se conoce muy poco sobre los mecanismos involucrados en la fusión entre células, en la auto-fusión y prácticamente nada sobre la fusión de VEs.

1.3.2.1. Fusión virus-célula hospedera

Los virus envueltos presentan una bicapa lipídica que les permite infectar a su hospedero a través de la fusión de sus membranas. Esta envoltura, obtenida de la membrana de la célula hospedera anterior, contiene a la partícula viral hasta que es liberada en el citoplasma de la célula blanco comenzando un nuevo ciclo de infección.

A pesar de que las proteínas de fusión viral conocidas presentan una alta variedad, todas ellas catalizan la fusión de forma básicamente igual (Fig. 4). De acuerdo a las proteínas de fusión conocidas y estudiadas hasta este momento, el modelo aceptado actualmente para explicar la fusión es el modelo de tallo (Jahn & Sudhof, 1999). En este modelo, los fusógenos se presentan en una conformación metaestable, también llamada de pre-fusión, hasta que se dan interacciones clave con una membrana blanco. Aquí se producen rearreglos estructurales que, en el caso de los fusógenos virales llevan a la exposición de una región hidrofóbica distintiva denominada péptido o loop de fusión. Este se inserta en la membrana blanco dando lugar a una estructura extendida intermedia, formando un puente. Este estado extendido es también conocido como intermediario de pre-horquilla. Finalmente esta estructura colapsa en una conformación de post-fusión de tipo *hairpin* u horquilla que acerca ambas membranas. De esta forma, el péptido de fusión y el dominio transmembrana de la proteína de fusión se acercan, forzando a las dos bicapas lipídicas a ubicarse en una posición cercana. Se conocen formas de pre-fusión de fusógenos virales para las cuales varía la cantidad de oligómeros que la componen,

pero todas las estructuras de post-fusión virales conocidas al momento presentan una conformación trimérica plegada en forma de horquilla.

La energía liberada por el plegamiento de la proteína a su conformación de menor energía es utilizada para el acercamiento de las membranas en el sitio de unión, y la deshidratación del área delimitada por las mismas.

Este modelo mecanístico es notablemente similar al propuesto para la fusión de membranas mediada por los complejos SNARE (ver sección 1.3.1.1 y Fig. 2). En las proteínas SNARE, la interacción entre proteínas de fusión se realiza en ambas membranas (acción bilateral), también involucrando la transición de una forma extendida a una plegada tipo horquilla (Itakura, Kishi-Itakura, & Mizushima, 2012). Las proteínas fusogénicas virales conocidas son todas glicoproteínas transmembrana tipo I. Las mismas están ancladas a la membrana de la envoltura viral por uno o dos dominios transmembrana alfa-hélice en su región C-terminal y exponen al medio extracelular el llamado ectodominio N-terminal, el cual establecerá contacto con la membrana blanco. Este contacto depende de un segmento hidrofóbico del ectodominio, denominado loop o péptido de fusión ubicado en la región más distal a la membrana del virus. En algunas proteínas de fusión viral de clase I el péptido de fusión se ubica en la posición N-terminal, en otras en una posición más interna en la secuencia, insertándose en la membrana posiblemente en forma de loop.

Varias proteínas de fusión corresponden al fragmento C-terminal de una proteína precursora, por lo que el inicio del mecanismo de fusión requiere que se libere de su fragmento N-terminal. El fragmento N-terminal generalmente corresponde a un dominio de unión con un receptor presente en la célula blanco (Harrison, 2008).

La fusión de virus envueltos con su membrana blanco puede darse en la membrana celular o en una locación intracelular si el virus es internalizado por endocitosis. Cuando la fusión se da con la membrana celular, la interacción entre el virus y el receptor está dada a pH neutro. Sin embargo, en la fusión dependiente de la internalización del virus es requerida la exposición a pH ácido en organelos de la vía endocítica. La fusión debe ser desencadenada por la interacción de la proteína de fusión o una proteína de unión con su receptor y/o correceptor, aunque en algunos casos es suficiente la exposición a pH ácido para desencadenar la fusión (Bouvier & Palese, 2008).

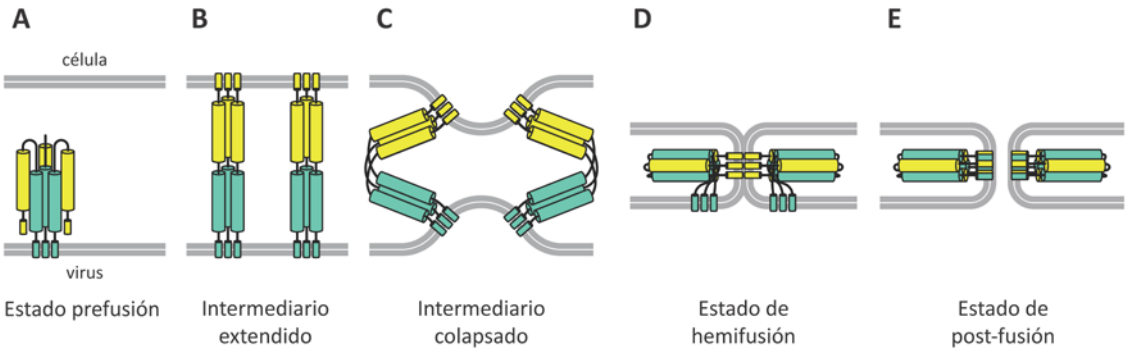


Figura 4. Estados intermedios de fusión de membranas. A. Proteína de fusión en conformación metaestable de pre-fusión. B. Extensión e interacción de la proteína de fusión con la membrana blanco. C. Colapso de la estructura intermedia y acercamiento de las membranas. D. Unión de monocapas proximales de las membranas opuestas. E. Formación del poro de fusión.

Se han descripto tres clases principales de proteínas de fusión viral denominadas clase I, clase II y clase III, siendo estas últimas las menos caracterizadas estructuralmente (Harrison, 2015). Sin embargo las proteínas de fusión viral de muchos virus no han podido ser asignadas a estas clases ya sea porque no se ha determinado su identidad o porque sus características estructurales y funcionales no se ajustan a las clases descriptas.

1.3.2.1.1. Fusógenos virales de clase I

Las proteínas de fusión viral de clase I forman un trímero de alfa-hélices que conforman una superhélice enrollada muy estable. Esta estructura es característica de los fusógenos virales de clase I, siendo una herramienta para definir si un fusógeno pertenece a esta clase o no.

El fusógeno del virus de la Influenza es el fusógeno viral de clase I mejor caracterizado ya que la estructura cristalográfica de su conformación de pre-fusión, post-fusión y de su precursor han sido estudiadas durante más de 30 años (Harrison, 2008). La proteína HA del virus de la Influenza es sintetizada como un precursor, el cual es procesado por proteasas de la célula hospedera para producir dos subunidades unidas por puentes-disulfuro. La subunidad HA1 consta de 328 aminoácidos y tiene la capacidad de unión a su receptor. La proteína HA2 contiene el dominio transmembrana, tiene 221 aminoácidos y es la responsable de la actividad de fusión del virus. Esta última se mantiene en una conformación metaestable en la superficie del virus mientras no existe contacto de su péptido de fusión ubicado en la posición N-terminal con la membrana blanco (Fig. 5.A). En esta conformación el péptido de fusión se encuentra hundido en el trímero. La fusión se desencadena por una disminución del pH, desestabilizando los contactos del trímero y proyectando el péptido de fusión hacia su membrana blanco. El péptido de fusión corresponde a una región apolar de la proteína, relativamente rica en residuos de glicina y alanina.

Las estructuras post-fusión (Fig. 5.B) conocidas demuestran la existencia de un plegamiento tipo horquilla del trímero de fusógenos quedando tanto los péptidos de fusión como los dominios transmembrana muy cerca unos de otros. Otra característica distintiva de las proteínas de fusión de clase I es una región central compuesta por una estructura denominada hélice superenrollada. Esta región está caracterizada por la presencia de un motivo de héptadas repetidas que favorece la formación de dicha estructura. Este motivo está compuesto por siete aminoácidos que aparecen repetidos en la proteína, siendo el primer y cuarto aminoácido de la héptada típicamente hidrofóbicos.

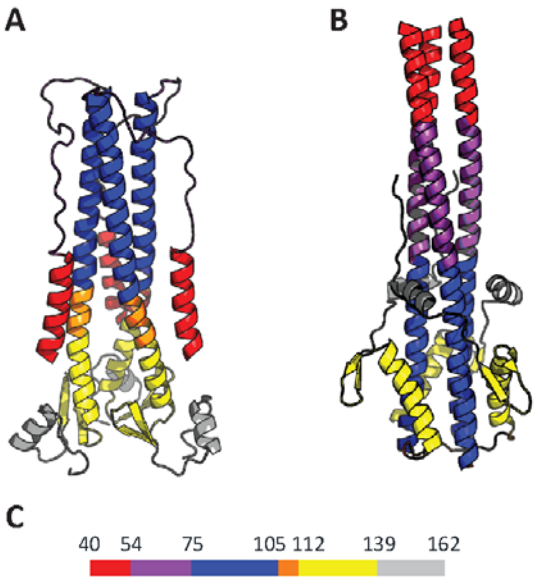


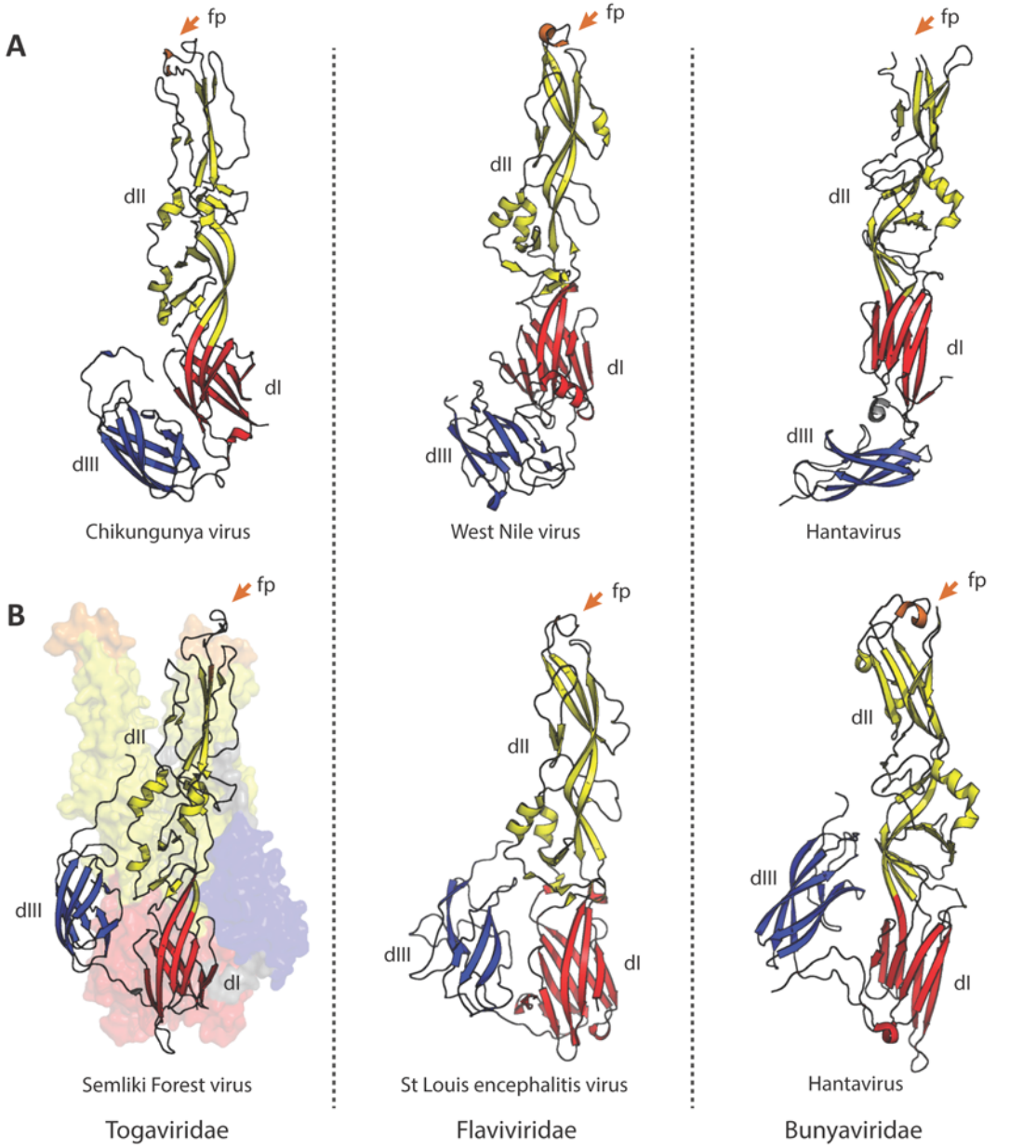
Figura 5. Estructuras del fusógeno viral de clase I HA2. A. Estructura de prefusión. B. Estructura de post-fusión. C. Representación lineal de los dominios. Los colores indican dominios que cambian su conformación entre las conformaciones de prefusión y post-fusión.

1.3.2.1.2 Fusógenos virales de clase II

Las proteínas de fusión de los virus miembros de las familias Togaviridae y Flaviviridae son ejemplos de fusógenos virales de clase II (Kielian, 2006). Las estructuras secundarias y terciarias de estas proteínas son significativamente similares sugiriendo una ancestría común, a pesar de que no es posible identificar una similitud a nivel de secuencia aminoacídica. La estructura tridimensional de estas proteínas es radicalmente distinta de la descripta para los fusógenos virales de clase I.

Las proteínas de clase II están ancladas a la envoltura viral a través de un dominio transmembrana ubicado en su región C-terminal, el cual está unido al ectodominio por una región flexible. El ectodominio consiste de tres dominios globulares bien definidos (Fig. 6). El dominio I contiene la región N-terminal de la proteína y consiste de un barril-beta. El dominio II está organizado principalmente en hojas-beta y contiene el loop de fusión. El dominio III se ubica en el extremo opuesto del dominio I presenta un dominio tipo inmunoglobulina que contiene los epítopes responsables para el tropismo celular y neutralización de anticuerpos. A diferencia de lo que sucede para los fusógenos virales de clase I, los ectodominios existen como monómeros o dímeros en su conformación de pre-fusión, pero en su conformación de post-fusión siempre existen en forma de trímeros. Esta conformación en forma de homotrímero es más estable y es requerida para que se produzca la fusión.

El mecanismo de fusión se desencadena por la disminución de pH en un compartimiento endosomal, el cual afecta los contactos internos de la proteína, exponiendo el loop de fusión. En ese momento el mismo puede insertarse en la membrana de su hospedero. Luego la proteína se pliega sobre si misma, acercando su loop de fusión a su dominio transmembrana, resultando en la fusión de las membranas. Este plegamiento es similar al producido en los fusógenos virales de clase I a pesar de la diferencia en la arquitectura de sus proteínas.


A
fp
dII
dI
dIII
Chikungunya virus
fp
dII
dI
dIII
West Nile virus
fp
dII
dI
dIII
Hantavirus
B
fp
dII
dIII
dI
Semliki Forest virus
fp
dII
dIII
dI
St Louis encephalitis virus
fp
dII
dIII
dI
Hantavirus
Togaviridae
Flaviviridae
Bunyaviridae

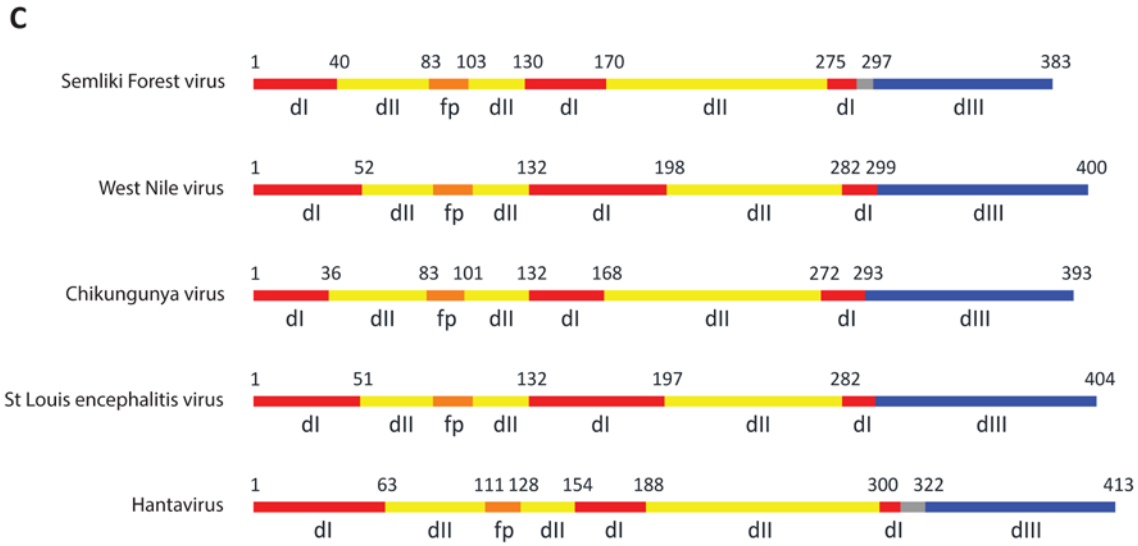

C
Semliki Forest virus
1 40 83 103 130 170 275 297 383
dI dII fp dII dI dII dI dIII
West Nile virus
1 52 132 198 282 299 400
dI dII fp dII dI dII dI dIII
Chikungunya virus
1 36 83 101 132 168 272 293 393
dI dII fp dII dI dII dI dIII
St Louis encephalitis virus
1 51 132 197 282 404
dI dII fp dII dI dII dI dIII
Hantavirus
1 63 111 128 154 188 300 322 413
dI dII fp dII dI dII dI dIII

Figura 6. Estructuras de fusógenos virales de clase II de familias Togaviridae, Flaviviridae y Bunyaviridae. A. Estructuras de prefusión. Para Semliki Forest virus se muestra la superficie tridimensional de los otros monómeros que conforman el trímero. B. Estructuras de post-fusión. C. Distribución de los dominios en la secuencia viral de cada virus representado en las partes A y B. El dominio dI se indica en color rojo, el dominio dII en amarillo, el dominio dIII en azul. El péptido de fusión se indica en color anaranjado y con una flecha en las partes A y B. Poner menos, con un ejemplo pre- y post- fusion para ilustrar el mecanismo propuesto.

1.3.2.1.3 Fusógenos virales de clase III

Al igual que lo descripto para los fusógenos virales de clase I y clase II esta clase de fusógenos no comparte conservación a nivel de secuencia aunque sí comparte su organización tridimensional (Y. Xu, Rahman, Othman, Hu, & Huang, 2012). A su vez, esta estructura es característica de estos fusógenos y difiere de la de los fusógenos de clase I y clase II. En particular, sus ectodominios son más extensos, presentan un único dominio transmembrana, tienen cinco dominios compuestos tanto por hélice alfa como por hoja beta y presenta una estructura similar a la hélice superenrollada de los fusógenos de clase I ubicada centralmente en la proteína. Al menos el fusógeno viral de clase III del virus de la estomatitis vesicular viral existe como un trímero tanto en su estado pre- como post-fusión. Al igual que en los fusógenos de clase I y clase II presenta un péptido de fusión que no se encuentra expuesto en la superficie en su estado de pre-fusión. También en este caso, las conformaciones de las estructuras pre- y post-fusión apoyan el mecanismo de plegado tipo horquilla (Fig. 7).

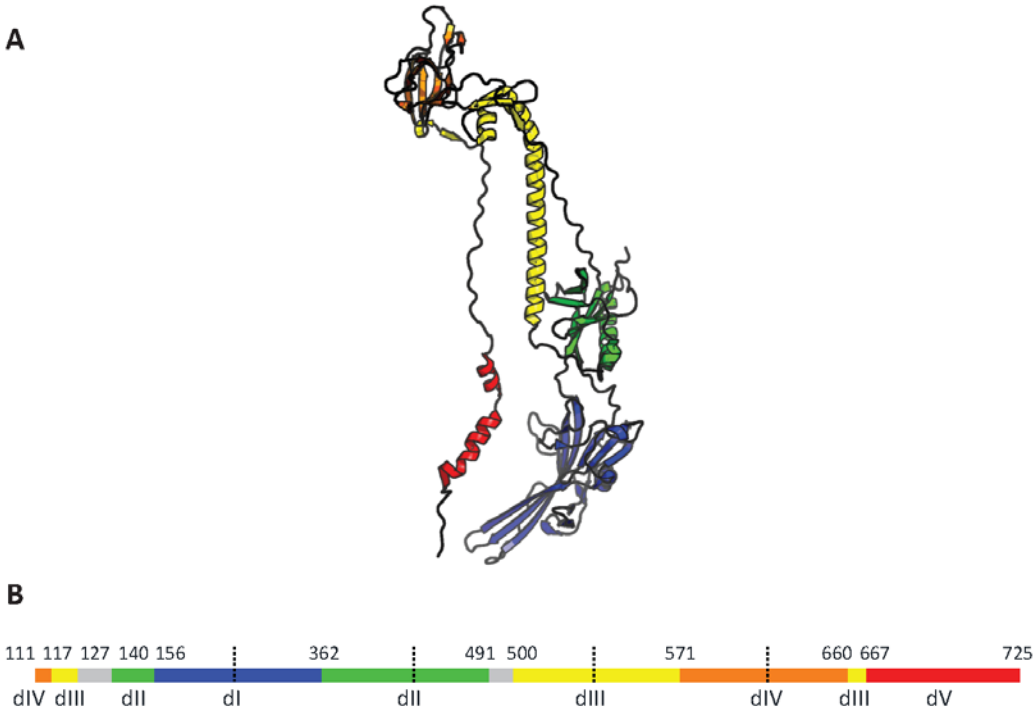


Figura 7. Estructura del fusógeno viral de clase III gB del virus herpes simplex. A. Estructura tridimensional del ectodominio en su conformación de post-fusión. B. Respresentación lineal de los dominios de la proteína. Las líneas punteadas representan el corte de la longitud total del dominio para su representación compacta.

1.3.2.1.4 Proteínas FAST

Las proteínas FAST (fusion-associated small transmembrane) de reovirus son las únicas proteínas de fusión viral conocidas en virus no envueltos. Sin embargo estas proteínas no inducen la fusión entre el virus y su célula blanco, por lo que no tienen implicancia en la entrada del virus a la célula, si no que inducen la fusión entre células infectadas formando un sincitio. Al igual de lo que sucede con los fusógenos virales descriptos previamente, es suficiente que la proteína FAST esté presente solamente en una de las dos membranas a fusionar. Sin embargo, estos fusógenos virales son significativamente más pequeños, presentando ectodominios de entre 19 y 37 aminoácidos. El mencanismo de fusión propuesto difiere al postulado para los fusógenos virales clase I, II y III. Las proteínas FAST pueden funcionar como fusógenos autónomos para inducir la unión e bicapas lipídidas artificiales, sin embargo, su estructura rudimentaria sugiere que un cofactor celular estaría involucrado en el complejo proceso de la sincitiogénesis. Se identificó a la anexina A1 como cofactor involucrado en la expansión del poro postfusión. (Ciechonska, Key, & Duncan, 2014).

1.3.2.2 Fusión célula-célula

La mayoría de las células de un organismo se mantienen mononucleadas durante su existencia, sin embargo algunas son sometidas a procesos de fusión muy regulados y restringidos a ciertos tipos celulares. La fusión celular permite la fertilización y el desarrollo de tejidos sincitiales como el músculo esquelético y la placenta en mamíferos o tejidos epidérmicos en nemátodos. También se discute el papel de la fusión célula-célula en la reparación de tejidos y en el desarrollo de procesos carcinogénicos. A pesar de conocerse diversos eventos de fusión célula-célula y de ser estudiada desde hace más de 100 años existe poca información sobre las proteínas que median el mecanismo de este tipo de fusión de membranas.

Dos familias independientes de proteínas de fusión fueron identificadas y caracterizadas. Una de ellas, las sincitinas, está compuesta por diversas proteínas que comparten un origen retroviral (Gong et al., 2005). La otra corresponde a las proteínas EFF-1 y AFF-1 de *Caenorhabditis elegans*, que comparten similitud estructural con proteínas de fusión viral de clase II (Perez-Vargas et al., 2014).

1.3.2.2.1 Fusión mediada por sincitinas

Durante la formación de la placenta intervienen las proteínas denominadas sincitinas. Durante una etapa temprana de la gestación un conjunto de células mononucleares de la capa externa del blastocisto denominado trofoblasto, se fusionan para formar una capa continua multinucleada denominada sincitiotrofoblasto. Esta capa es necesaria para la implantación y mantenimiento del embrión en desarrollo, siendo responsable de varias funciones llevadas a cabo por la placenta, como el transporte de oxígeno, nutrientes, tolerancia inmunitaria y producción de hormonas (Burton & Fowden, 2015).

La fusión del trofoblasto se produce a través del mismo mecanismo conocido para la fusión de virus envueltos y sus células blanco. El gen de la sincitina es un elemento viral endógeno proveniente del retrovirus HERV-W correspondiente al gen env que codifica una proteína de 538 residuos (S. Mi et al., 2000).Estudios filogenéticos indican que las sincitinas son fusógenos de origen retroviral relativamente modernos, los cuales fueron “asimilados” por mamíferos repetidas veces a lo largo de la evolución (Cornelis et al., 2015). El alineamiento de la estructura de sincitina con la proteína gp160 de VIH comparte un perfil estructural similar (Fig. 8). Además, el alineamiento de sus secuencias refleja una alta similitud a nivel de aminoácidos.

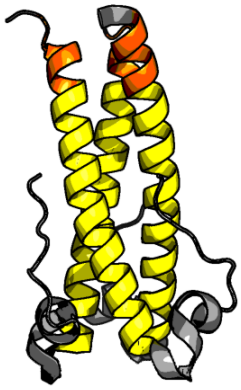


Figura 8. Estructura del dominio central de la proteína Sincitina-2 humana. Este se organiza de forma similar a los fusógenos virales de clase I. Los colores se corresponden con los de la Fig. 5.

También comparte características estructurales particulares con fusógenos virales de clase I. Presenta un péptido de fusión en el extremo N-terminal de la proteína transmembrana, el cual es rico en residuos de glicina. La proteína también presenta una estructura secundaria principalmente organizada en hélice-alfa y existe como trímero ensamblado en la membrana viral. A diferencia de la proteína HA del virus de la Influenza, las sincitinas median la fusión a pH neutro, de forma similar a la proteína gp160 de VIH. El mecanismo de fusión está dado por los mismos pasos propuestos para la fusión virus-célula.

1.3.2.2.2 Fusión mediada por EFF-1 y AFF-1

La segunda familia de proteínas de fusión célula-célula conocida corresponde a las proteínas de la denominada familia de fusión (FF). Estas proteínas son responsables de diversas y numerosas fusiones durante el desarrollo de *Caenorhabditis elegans*. Las proteínas identificadas incluyen EFF-1 (epithelial fusion failure 1) involucrada en la formación de hipodermis, vulva y faringe, y AFF-1 (anchor-cell fusión failure 1), requerida para la formación del himen, músculos faríngeos y sincitio epidérmico en nematodos (Perez-Vargas et al., 2014). Ambas son proteínas transmembrana de tipo I (el extremo C-terminal se encuentra en el citoplasma) y adoptan una estructura trimérica similar a la conocida para proteínas de fusión viral de clase II en su conformación de postfusión (Fig. 9). Sin embargo el loop de fusión es reemplazado por un segmento hidrofílico. Se demostró que bloqueando la trimerización de EFF-1 se bloquea la fusión de membranas. Sin embargo se demostró que para la fusión mediada por estas proteínas es necesaria su presencia en ambas células a fusionar. Esto sugiere, que a pesar de ser estructuralmente similares a los fusógenos virales clase II, existe también una similitud con el mecanismo descripto para las proteínas SNARE, donde es necesaria una interacción bilateral para llevar a cabo la fusión.

De forma de verificar la capacidad fusogénica de estas proteínas, se comprobó que se forma un sincitio cuando se expresan ectópicamente en células de mamíferos e insectos (Avinoam et al., 2011). También se demostró que partículas de rhabdovirus pueden ser pseudotipadas con AFF-1 reemplazando la proteína de fusión viral en su superficie. Como era esperado, las partículas pseudotipadas requieren también la presencia de AFF-1 o EFF-1 en las células blanco para poder fusionar.

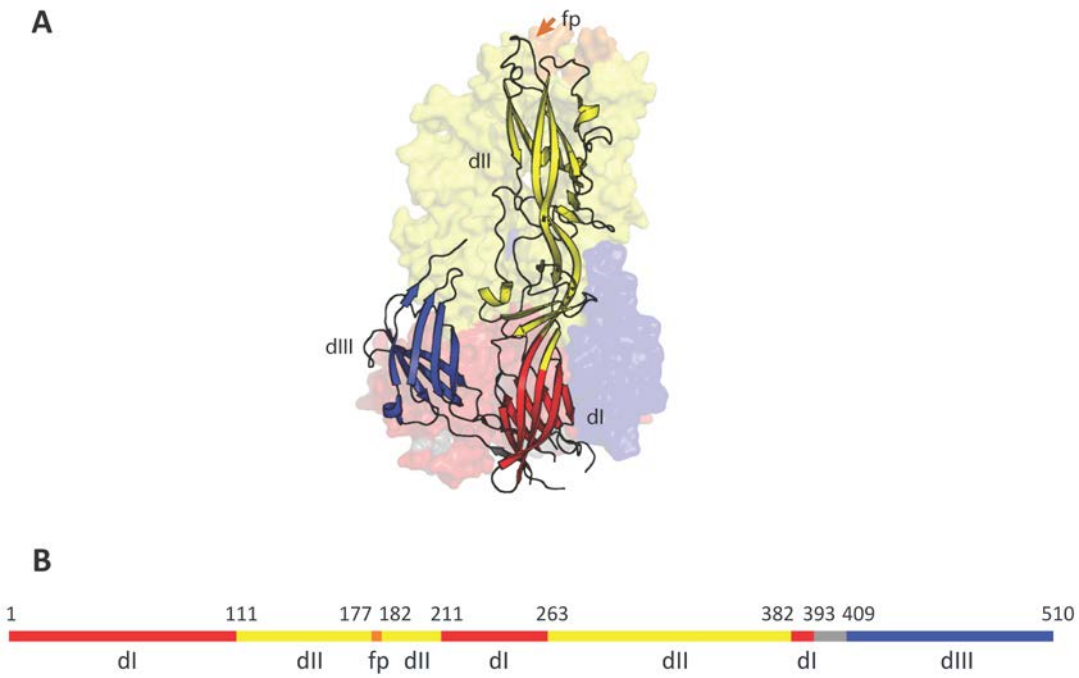


Figura 9. Estructura de la proteína EFF-1 de C. elegans en su conformación post-fusión (Perez-Vargas et al., 2014). A. Estructura tridimensional de la estructura de la proteína. Se indican los dominios dI (rojo), dII (amarillo) y dIII (azul), y el segmento hidrofílico (fp). Con los mismos colores se representan las otras subunidades que conforman el trímero. B. Representación lineal de la proteína. Se indican las posiciones de inicio y fin de los dominios.

1.3.2.2.3 Fusión de gametos

La creación de un nuevo individuo de una especie de mamífero depende de la fusión de la membrana de dos células haploides, el gameto femenino y el masculino, para formar un cigoto. A pesar de la importancia de la fusión de gametos se conoce muy poco acerca de los mecanismos moleculares de su fusión.

En ratón se identificaron dos genes esenciales para la fusión, Izumo1 en el espermatozoide y Cd9 en el ovocito. Sin embargo no se pudo comprobar que estas tengan función fusogénica.

Más recientemente se identificó a la proteína HAP2 como candidata a ser el fusógeno involucrado en la fusión de gametos (Fedry et al., 2017; Valansi et al., 2017). Se pudo comprobar que la proteína HAP2 es suficiente para promover la fusión célula-célula en mamíferos. Además, esta proteína está presente en la mayoría de los taxones eucariotas (animales, plantas y protistas), sugiriendo que muchos organismos eucariotas comparten un mecanismo de fusión de gametos común.

Se ha reportado que la hemifusión y la fusión completa dependen de que HAP2 esté presente en ambas células. Además EFF-1 puede sustituir a HAP2 en una de las membranas a fusionar, indicando que comparten un mecanismo de fusión similar. Sin embargo otros estudios sugieren que en algunos organismos es suficiente que esté presente en uno de los dos gametos, sugiriendo la posibilidad de que funcione de forma similar a un fusógeno viral. Al igual que EFF-1, HAP2 es homóloga a las proteínas de fusión viral de clase II. Fedry et al. sostienen que la organización topológica idéntica de elementos de estructura secundaria, terciaria y cuaternaria

de HAP2 con los fusógenos virales de clase II solamente se pueden explicar postulando un ancestro común. Una vez que surge una proteína requerida para una función compleja como la fusión de membranas, su gen es utilizado sistemáticamente, generalmente transferido por vía horizontal. También es posible una transferencia que involucre retrotranscripción, seguida de la integración al genoma.

1.3.2.2.4 Desarrollo y reparación de músculo

Durante el desarrollo, los mioblastos están comprometidos como precursores de células musculares. Estos se alinean juntos y se fusionan para formar miotubos multinucleados característicos del tejido muscular. El mecanismo de fusión comienza una vez que los mioblastos dejan de proliferar abandonando el ciclo celular debido a la ausencia de factores de crecimiento. En este momento comienzan a secretar fibronectina sobre su matriz extracelular, permitiendo la unión de integrinas que permiten la adhesión a la matriz. Posteriormente se produce el reconocimiento entre mioblastos, llevando al alineamiento de los mismos en cadenas. Finalmente se produce la fusión de las membranas, aunque no se conoce qué proteína lleva a cabo el mecanismo. Este proceso forma un gran sincitio capaz de producir fibras musculares necesarias para el control del movimiento (Rochlin et al., 2010).

La fusión de mioblastos también ocurre durante la regeneración muscular, producida por daño sobre el tejido muscular. Esta situación activa células satélite que sufren procesos de división asimétrica para generar mioblastos. Al igual que sucede durante el desarrollo, la maquinaria que media el mecanismo de fusión es desconocida. Recientemente se ha identificado un par de proteínas que podrían promover la fusión de mioblastos en ratón, Myomixer y Myomaker (Bi et al., 2017). La expresión de Myomixer coincide con la diferenciación de mioblastos y es esencial para la fusión y formación del músculo esquelético durante la embriogénesis. Myomixer se localiza en la membrana plasmática donde promueve la fusión de mioblastos y se asocia con Myomaker, una proteína fusogénica de membrana.

1.3.2.2.5 Desarrollo de hueso

El remodelado óseo es un proceso que involucra la fusión celular que no está limitado al desarrollo embrionario. Este proceso depende de la reabsorción de tejido óseo mediada por células multinucleadas denominadas osteoclastos (Ishii & Saeki, 2008). La función de estas células es entonces crucial para el mantenimiento de la homeostasis ósea.

Los osteoclastos dependen de la fusión de progenitores hematopoyéticos, como son los monocitos o macrófagos. El proceso por el cual se forma un osteoclasto requiere la migración celular, quimiotaxis, reconocimiento célula-célula, anclaje y finalmente fusión de membranas. Las proteínas implicadas y el mecanismo por el cual fusionan aún no se conocen.

1.3.2.2.6 Fusión en cáncer

Aunque está en fuerte discusión, existen algunos reportes que vinculan la fusión célula-célula con distintos tipos de cáncer. Los macrófagos también pueden fusionar con células somáticas para promover la reparación tisular, y con células tumorales para promover la metástasis (Pawelek & Chakraborty, 2008). Así, se ha propuesto que la fusión celular puede tener un papel importante en el desarrollo del cáncer, promoviendo su iniciación, progresión, mantenimiento y metástasis. La combinación de dos tipos celulares distintos en una misma célula puede otorgarle mecanismos que aumenten su agresividad y potencial de metástasis, por ejemplo

permitiéndole adquirir una capacidad migratoria. Tampoco aquí se ha identificado la maquinaria proteica que media, si existe, la fusión entre macrófagos y células tumorales.

1.3.2.3 Vesículas extracelulares

Las vesículas extracelulares (VEs) son vesículas esféricas rodeadas de membrana liberadas por las células tanto eucariotas como procariotas para mediar el transporte molecular o intermediar la comunicación intercelular. Aunque no se ha alcanzado un consenso, es aceptado que existen distintos tipos, caracterizados por sus características fisicoquímicas, origen subcelular y vía de liberación (Thery, Ostrowski, & Segura, 2009). Una de estas clasificaciones propone la distinción entre exosomas, microvesículas, partículas membranosas y vesículas apoptóticas (van der Pol, Boing, Harrison, Sturk, & Nieuwland, 2012).

Los exosomas pueden ser identificados en la mayoría de los fluidos biológicos, fracciones de fluidos biológicos y medios de cultivo de células (Caby, Lankar, Vincendeau-Scherrer, Raposo, & Bonnerot, 2005; Keller, Sanderson, Stoeck, & Altevogt, 2006; Pisitkun, Shen, & Knepper, 2004; Vella, Greenwood, Cappai, Scheerlinck, & Hill, 2008). Su diámetro varía entre 30 y 100 nm y su membrana fosfolipídica presenta niveles relativamente altos de colesterol, esfingomielina y ceramidas, presentando además balsas lipídicas (Mathivanan, Ji, & Simpson, 2010; Simons & Raposo, 2009; Thery et al., 2009; Wubbolts et al., 2003). Las proteínas en su membrana presentan la misma orientación que en la célula y su contenido está caracterizado por proteínas involucradas en el transporte y fusión. Los exosomas pueden liberarse a partir de vesículas intraluminales formadas en el cuerpo multivesicular. Este último puede fusionarse con lisosomas para la degradación de su contenido o con la membrana celular para secretar sus vesículas intraluminales como exosomas (Fig. 10). Existe también una vía directa de formación de exosomas, en la cual se liberan vesículas que se geman directamente de la membrana celular hacia el exterior (Booth et al., 2006; Fang et al., 2007; Lenassi et al., 2010). Estos exosomas son indistinguibles de los generados por la vía indirecta.

Las microvesículas o micropartículas son liberadas por las células durante procesos de estrés celular. Aunque se ha reportado que estas vesículas tienen un diámetro mayor a los exosomas es posible que el rango de diámetros se solape (Fig. 10). Tampoco existen otros criterios absolutos para distinguirlas de exosomas.

Las partículas membranosas o prominosomas son liberadas únicamente por células epiteliales y presentan dos rangos de diámetro, uno similar a los exosomas y otro de alrededor de 600 nm. A diferencia de exosomas y microvesículas, es posible identificarlas por el marcador CD133 (Marzesco et al., 2005).

Las vesículas o cuerpos apoptóticos son liberadas por células sufriendo apoptosis. Su mayor diferencia con los otros tipos de vesículas es su mayor tamaño, variando entre 1 y 5 um de diámetro, aunque su densidad es similar a la de exosomas (Hristov, Erl, Linder, & Weber, 2004; Kerr, Wyllie, & Currie, 1972; Thery et al., 2001; Turiak et al., 2011). Éstas presentan su superficie fosfatidilserina, necesaria para su reconocimiento por receptores fagocíticos. Son liberadas por un proceso de “blebbing”, en el cual se forman protrusiones en la superficie celular que luego son liberadas, de forma similar a lo que sucede para las microvesículas.

Desde un punto de vista topológico-celular, la fusión entre VEs y la célula blanco es equivalente a la fusión entre células, o entre una célula y un virus con cubierta membranosa. Sin embargo los mecanismos por los cuales fusionan permanecen inciertos.

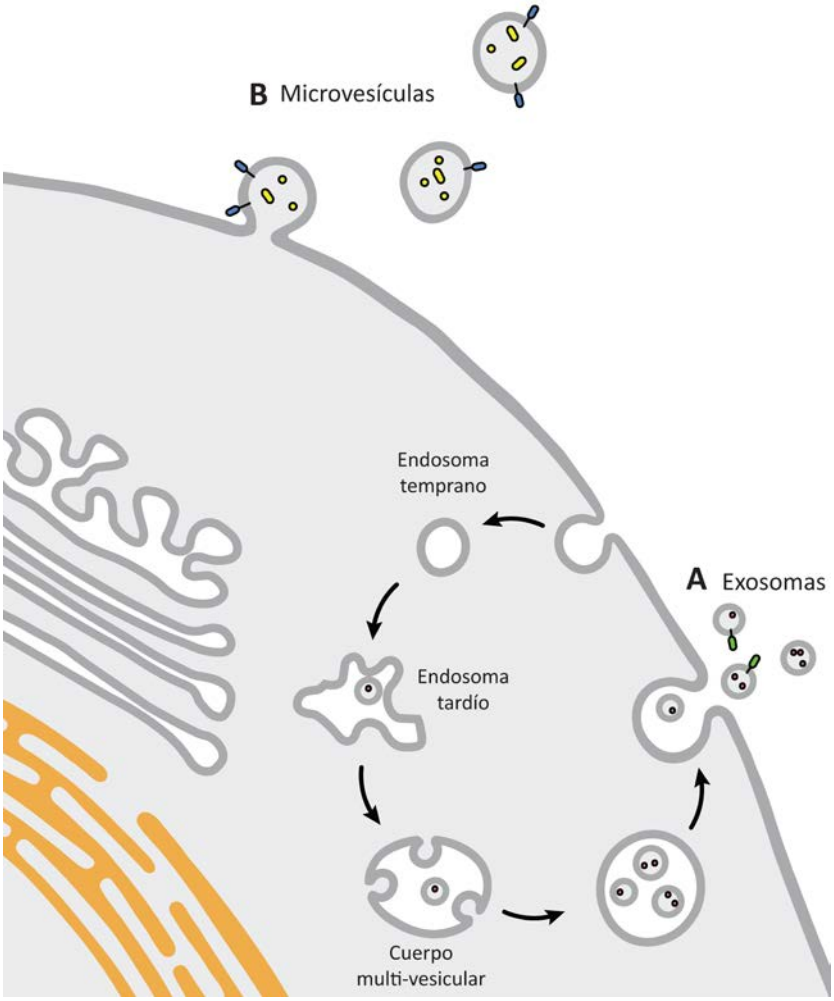


Figura 10. Liberación de VEs. A. Liberación de exosomas a partir de vesículas intraluminales formadas en el cuerpo multivesicular. B. Liberación de microvesículas por protrusión.

Cada vez existe más evidencia de que las VEs tienen funciones especializadas y recientemente han sido reconocidas como un mecanismo de señalización intercelular. Es de particular interés entender los mecanismos por los cuales las VEs entregan el mensaje dado que tienen gran potencial para ser usados en terapia.

Las células tumorales liberan exosomas que tienen la propiedad de modular el sistema inmune. Por esta razón es de interés explicar qué proteínas median la fusión de membranas, ya que pueden ser utilizadas como blancos en terapia. Un estudio reciente observó la inhibición de la liberación de exosomas por células tumorales luego del tratamiento con dimethyl amiloride, aumentando así la eficacia antitumoral de ciertas drogas quimioterapéuticas (Chalmin et al., 2010).

Por otra parte, la incubación de células presentadoras de antígeno con péptidos tumorales genera la liberación de exosomas que son capaces de producir la supresión del crecimiento de células tumorales mediada por células T citotóxicas (Andre et al., 2002; Zitvogel et al., 1998).

También se ha estudiado la eficiencia de las VEs como vehículos capaces de pasar ciertas barreras farmacológicas como la hematoencefálica. Esta particularidad resulta especialmente interesante en el tratamiento de enfermedades neurológicas. Se ha visto que esta función es efectiva para el transporte de ARN de interferencia en dosis terapéuticas, de manera no inmunogénica y a tejidos específicos (van den Boorn, Schlee, Coch, & Hartmann, 2011). Se ha observado el potencial del transporte de ARNsi mediado por exosomas a través de la inhibición de más del 50% de la expresión y producción de BACE1, el cual es un blanco terapéutico en la enfermedad de Alzheimer (Alvarez-Erviti et al., 2011).

1.4 Consideraciones evolutivas de las proteínas de fusión de membranas

La mayor parte del conocimiento sobre las estructuras y mecanismos de las proteínas de fusión de membranas proviene del estudio de proteínas de fusión viral. Como ya se describió, a pesar de su diversidad, la mayor parte de las proteínas de fusión viral conocidas se pueden clasificar en tres clases estructuralmente distintas. Algunas de estas proteínas tienen una estrecha relación con proteínas de fusión de membranas celulares. Por ejemplo, las células del trofoblasto utilizan fusógenos de clase I derivadas de retrovirus endógenos para formar el sincitiotrofoblasto. Por su parte, las proteínas SNARE presentan características generales similares a proteínas de fusión viral de clase I, siendo este probablemente un ejemplo de evolución convergente. En ambos casos se forma un conjunto de alfa-hélices que acerca membranas opuestas.

Por otra lado, las estructuras tridimensionales de las proteínas HAP2 y EFF-1 son idénticas a las de las proteínas de fusión de clase II. Esta similitud es más fácilmente explicada por la existencia de un ancestro común. Resulta difícil imaginar cómo la evolución convergente podría resultar en un nivel de similitud estructural tan alto entre estas proteínas de fusión célula-célula y proteínas de fusión viral. Así surge la incógnita de si los eucariotas capturaron estas proteínas desde un gen viral o ciertos virus las adquirieron de un progenitor eucariota.

Centrándonos en la proteína HAP2, la segunda hipótesis es soportada por el hecho que HAP2 está presente en la mayoría de los taxa eucariotas salvo fungi, pero incluyendo protozoarios, plantas y amebas. Por esta razón, su distribución es consistente con su presencia en el último ancestro común de todos los eucariotas. Así es que HAP2 podría ser una proteína de fusión de gametos antigua que fue reutilizada por un virus eucariota primitivo para la entrada a su célula hospedera. Alternativamente, es posible que HAP2 se originara en un virus y su gen correspondiente haya sido obtenido por un organismo eucariota en la evolución temprana.

Independientemente del mecanismo de obtención, los datos estructurales demuestran un nexo evolutivo no ambiguo entre proteínas de fusión. Estos ilustran cambios genéticos extensivos entre virus y células a lo largo de la evolución y la sorprendente adaptación de una proteína para mantener la misma función mientras adopta otro modo de acción.

OBJETIVOS

Existe una amplia variedad de procesos biológicos donde la fusión de membranas es esencial. Las membranas celulares no fusionan espontáneamente sino que este proceso es catalizado por proteínas con capacidad fusogénica. A diferencia de lo que ocurre con los procesos de fusión de membranas intracelulares, las proteínas de fusión que median la fusión de membranas en el medio extracelular, a excepción de la fusión virus-célula blanco, son casi desconocidos.

Este trabajo de Maestría se propuso desarrollar métodos informáticos de identificación de proteínas con capacidad fusogénica.

Debido a su reciente reconocimiento como proceso biológico y a su importancia en la salud humana decidimos tomar a las VEs como objeto de estudio. Esto se debe a que si bien existe evidencia que sostiene a la fusión de membranas entre las VEs y la célula blanco como paso necesario para la entrega de su contenido (Prada & Meldolesi, 2016), aún se desconoce qué proteínas llevan a cabo la fusión.

Así, el **objetivo general** de este trabajo es identificar proteínas presentes en VEs con potencial fusogénico.

Dentro de las proteínas de fusión de membranas extracelulares, las proteínas de fusión viral han sido ampliamente estudiadas, y dada su estructura tridimensional y su mecanismo molecular se han clasificado en al menos tres clases: clase I, clase II y clase III. Además, los fusógenos conocidos que median la fusión entre células presentan homología con fusógenos virales.

Las proteínas de fusión viral y las proteínas homólogas conocidas que catalizan la fusión célula-célula presentan una gran diversidad a nivel de secuencia. Por esta razón análisis que busquen similitud a nivel de secuencia no son capaces de identificar homología entre ellas. Sin embargo, se ha reportado que existen patrones de similitud a nivel de estructura secundaria dentro de cada clase de fusógenos virales.

Por estas razones, los **objetivos específicos** de este trabajo de maestría son:

- Determinar una métrica de similitud a nivel de estructura secundaria que permita clasificar fusógenos virales de acuerdo a su clase, utilizando algoritmos de aprendizaje automático.
- A partir de esta métrica, clasificar las proteínas identificadas en VEs como similares a fusógenos virales de clase I, clase II o clase III, utilizando algoritmos de aprendizaje automático del tipo one-class.

A continuación se describe brevemente distintas metodologías utilizadas para la búsqueda de homología entre proteínas, las cuales son de interés para el desarrollo de este trabajo.

METODOLOGÍAS

2 Métodos de búsqueda de homología

Esta sección está basada en el capítulo 3 Pevsner et al. (Pevsner, 2015). Una de las preguntas más básicas sobre genes o proteínas es si estos están relacionados con algún otro gen o proteína. Dado que este trabajo se basa en la búsqueda de proteínas se hará referencia a la búsqueda de homología de proteínas. La similitud de dos proteínas a nivel de secuencia sugiere que estas son homólogas y también sugiere que pueden tener funciones en común. Dos proteínas son homólogas si comparten un ancestro evolutivo en común. La homología entre proteínas generalmente implica una estructura tridimensional relacionada. Al analizar muchas secuencias de proteínas es posible identificar dominios o motivos que son compartidos dentro de un grupo de moléculas. Estos análisis de la relación de proteínas son llevados a cabo alineando secuencias.

Cuando dos secuencias son homólogas, sus secuencias de aminoácidos generalmente comparten una identidad significativa. Mientras que la homología es una inferencia cualitativa (las secuencias son homólogas o no lo son), la identidad y la similitud son cantidades que describen qué tan emparentadas son las secuencias. Es importante resaltar que dos moléculas pueden ser homólogas sin compartir una identidad significativa. Por lo tanto el propósito de un alineamiento pareado es averiguar el grado de similitud y la posibilidad de existir homología entre dos moléculas. Finalmente, la evidencia más importante para determinar si dos proteínas son homólogas se obtiene del análisis estructural combinado con análisis evolutivos.

2.1 Alineamientos pareados

Es posible evaluar la relación entre dos proteínas realizando un alineamiento pareado. Los alineamientos pareados son una herramienta útil para identificar mutaciones que ocurrieron durante la evolución y causaron divergencia en las secuencias de las secuencias en cuestión. Las mutaciones más comunes son las sustituciones, inserciones y deleciones.

Existen diversos algoritmos que permiten alinear secuencias de proteínas. Éstos, por ejemplo, tienen en cuenta la sustitución o alineamiento de aminoácidos que son similares pero no idénticos, la incorporación de gaps en el alineamiento para describir deleciones o inserciones en las secuencias y si permite alineamientos locales o globales.

Las inserciones y deleciones se denominan gaps en un alineamiento. El sistema más simple de penalización de gap es el lineal, donde la incorporación de un gap siempre tiene la misma penalización. Un sistema típico de penalización de gap es el de tipo afín. En este hay dos tipos de penalidad de gap, uno es la penalidad de crear un gap (-a), y otro es la de extenderlo (-b). Así, si el gap se extiende por k aminoácidos se asigna una penalidad de –(a+bk).

Las sustituciones (o *mismatch*) ocurren cuando una mutación resulta en un cambio en el codón que codifica un aminoácido, dando lugar a otro. Esto resulta en el alineamiento de dos aminoácidos que no son idénticos. Un par de aminoácidos similares están relacionados a nivel estructural o funcional. Por ejemplo, por ser básicos, ácidos, hidroxilados o hidrofóbicos. Al alinear dos aminoácidos distintos pero con características similares estamos realizando una sustitución conservativa.

Por esta razón, cuando se alinean dos proteínas es necesario contar con un sistema para puntuar tanto un match (alineamiento de dos aminoácidos idénticos) como un mismatch. La mayoría de

los métodos de búsqueda de homología a nivel de aminoácidos hoy en día utilizan sistemas que de alguna forma están relacionados con las matrices PAM (*point accepted mutation*) o las BLOSUM (blocks sustitution matrix). Estas se generaron a partir alineamientos múltiples (MSA) de proteínas.

La mejor forma de determinar los límites de la detección de los alineamientos pareados es a través de tests estadísticos que evalúan la probabilidad de encontrar un match por azar. También es útil comparar el porcentaje de identidad de dos secuencias versus su distancia evolutiva.

Existen dos grandes tipos de alineamiento: global y local. Un alineamiento global, como el método de Needleman-Wunsch (Needleman & Wunsch, 1970), contiene la secuencia completa de cada proteína. Un alineamiento local, como el método de Smith-Waterman (Smith & Waterman, 1981), se enfoca en regiones de alta similitud entre dos secuencias. Los dos métodos citados garantizan encontrar una o más soluciones óptimas para el alineamiento de dos proteínas. Existen también algoritmos de búsqueda rápida como BLAST y FASTA que representan una forma simplificada y veloz de alineamiento local.

2.1.1 Algoritmo de Needleman-Wunsch

El algoritmo de Needleman-Wunsch compara dos secuencias en una matriz bidimensional (Fig. 11). Cualquier match o mismatch entre dos secuencias se representa como un camino diagonal, sin embargo el score que se le asigna se ajusta de acuerdo a un sistema de score. En consecuencia, un alineamiento perfecto entre dos secuencias idénticas se representa como una línea diagonal desde el extremo superior izquierdo al extremo inferior derecho. Los gaps son representados en esta matriz utilizando caminos horizontales o verticales. Un gap en la primera secuencia se representa como una línea vertical, mientras que un gap en la segunda secuencia se representa como una línea horizontal.

El algoritmo de Needleman-Wunsch es un ejemplo de programación dinámica (Sedgewick, 1988). Esto significa que el camino óptimo es detectado extendiendo de forma incremental subcaminos óptimos, es decir, realizando una serie de decisiones en cada paso del alineamiento que corresponden a cuál par de aminoácidos corresponde al mejor score.

2.1.2 Algoritmo de Smith-Waterman

El algoritmo de alineamiento local de Smith-Waterman es el método más riguroso por el cual se pueden alinear subsets de dos secuencias de proteínas. El alineamiento local es extremadamente útil en una variedad de aplicaciones, como la búsqueda en una base de datos en la cual queremos alinear dominios de proteínas pero no secuencias enteras.

Un algoritmo de alineamiento local de secuencias se parece al del alineamiento global en que dos proteínas están organizadas en una matriz y se busca un camino óptimo a través de la diagonal. Sin embargo no hay penalidad por comenzar el alineamiento en una posición interna, y el alineamiento no se extiende a los extremos de las dos secuencias necesariamente.

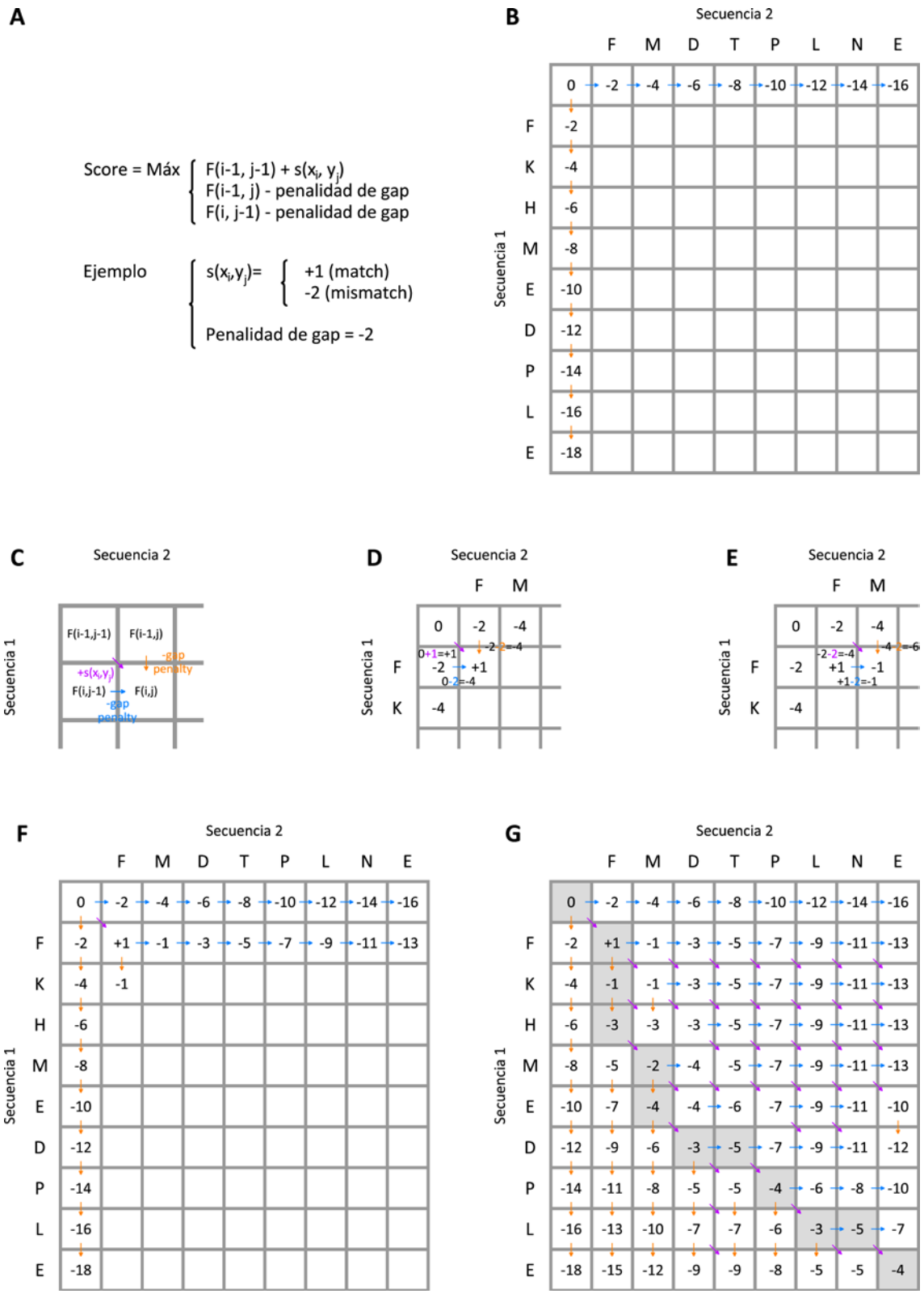


Figura 11. Alineamiento pareado de dos secuencias de aminoácidos utilizando el algoritmo de programación dinámica de Needleman-Wunsch (1970) para un alineamiento global. Figura adaptada a partir de Pevsner et al. (Pevsner, 2015). A. El score de cada celda (F(i,j)) se determina como el valor máximo de tres operaciones: la suma del valor de la celda superior izquierda (F(i-1, j-1)) más el valor de match o mismatch según corresponda ($s(x_i, y_j)$), la suma del valor de la celda de

la izquierda (F(i-1, j)) más la penalidad de gap, o la suma del valor de la celda de arriba (F(i, j-1)) más la penalidad de gap. Entonces, en cada celda el score se asigna utilizando el algoritmo recursivo que identifica el score más alto de los tres cálculos y se indica con una flecha qué camino dio lugar a ese score. El sistema de score en este ejemplo utiliza +1 para un match, -2 para un mismatch y -2 para penalizar un gap. B. Para las secuencias de largo m y n se genera una matriz de dimensiones m+1 por n+1 y se incorporan las penalidades de gap en la primera fila y primera columna. Como ya se describió, a cada posición de gap le corresponde un score de -2. C. Luego se rellena el resto de la matriz de forma ordenada teniendo en cuenta el sistema de score descripto en A. D. Para calcular el score en la celda de la segunda fila y segunda columna se toma el máximo de los tres scores: +1, -4 y -4. Este mejor score (+1) sigue el camino de la flecha indicada en violeta, y se mantiene la información de cuál es el mejor camino que resulta del score de cada celda, de forma de poder reconstruir luego el alineamiento pareado. E. Para calcular el score en la segunda fila, tercera columna, se toma nuevamente el máximo de los tres scores: -4, -1, -6. El mejor score sigue la flecha azul desde la izquierda. F. Luego se sigue rellenando los scores a través de la primera fila de la matriz. G. La matriz completa incluye el score final del alineamiento óptimo (-4) el cual corresponde a la celda inferior derecha, correspondiente al extremo c-terminal de cada proteína. El alineamiento se determina con un procedimiento de tipo trace-back. El camino creado siguiendo las flechas desde esta celda corresponde al o los alineamientos con score máximo. En gris se indica el camino óptimo. Reconstruyendo el alineamiento se obtiene: FKHMED-PL-E
F—M-DTPLNE

Las reglas para rellenar la matriz difieren ligeramente de las de Needleman-Wunch. Una de estas diferencias es que la primera fila y la primera columna se rellenan con ceros y no con la penalidad de gap. Además no se permiten valores negativos en la matriz, por lo que si se obtiene un score negativo para una celda, este valor se sustituye por un cero. El alineamiento óptimo puede comenzar y terminar en cualquier posición de la matriz. Por esta razón, se identifica el valor máximo en la matriz y este va a corresponder al extremo C-terminal del alineamiento. El procedimiento de trace-back comienza en esa posición y sigue el camino hasta que alcanza un celda con el valor cero. Esta celda corresponde al comienzo del alineamiento.

2.1.3 Métodos heurísticos

Mientras que el algoritmo de Smith-Waterman garantiza encontrar el o los alineamientos óptimos entre dos secuencias, tiene una gran desventaja que es ser muy lento. Para alineamientos pareados la velocidad no suele ser un gran problema. Pero cuando el algoritmo de alineamiento pareado se aplica al problema de comparar una secuencia (o query) contra una base de datos entera la velocidad del algoritmo se transforma en un problema significativo y puede variar en varios órdenes de magnitud. Se han desarrollado dos algoritmos de alineamiento local populares que son alternativas rápidas a Smith-Waterman: FASTA (Pearson & Lipman, 1988) y BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990). El tiempo de cálculo menor se debe a que FASTA y BLAST restringen la búsqueda escaneando la base de datos para matches probables antes de hacer más alineamientos rigurosos. Estos algoritmos heurísticos

sacrifican algo de sensibilidad por velocidad, por ende no garantizan encontrar alineamientos óptimos.

El algoritmo FASTA presenta cuatro pasos principales. Primero genera una tabla de búsqueda que contiene pequeños fragmentos de aminoácidos de una base de datos. El tamaño de estos fragmentos está determinado por un parámetro llamado k-tupla. Si se define k-tupla=3, entonces la secuencia query se examina en bloques de tres aminoácidos contra matches de tres aminoácidos encontrados en la tabla de búsqueda. FASTA identifica los diez segmentos de máximo score que alinean para una k-tupla dada. Estas 10 regiones alineadas se vuelven a alinear, permitiendo reemplazos conservativos, utilizando una matriz de sustitución como la PAM250. Estas regiones de score alto se unen si son parte de una misma proteína. Finalmente se realiza un alineamiento global o local en estas secuencias de alto score, optimizando el alineamiento de la secuencia query con los mejores matches de la base de datos. El algoritmo de programación dinámica entonces es aplicado a la búsqueda de la base de datos en una cantidad limitada, permitiendo a FASTA devolver resultados muy rápido ya que evalúa solamente una porción de los alineamientos potenciales.

Por su parte, el algoritmo BLAST se introdujo como una herramienta de búsqueda de alineamiento local que identifica alineamientos entre la secuencia query y una base de datos de secuencias y luego extiende el match en cualquier dirección. El resultado de la búsqueda consiste tanto en secuencias muy relacionadas de la base de datos como también en secuencias relacionadas marginalmente, junto con un esquema de puntuación para describir el grado de relación entre el query y cada hit con la base de datos. BLAST compila una lista de palabras de longitud fija w que son derivadas de la secuencia query. Entonces BLAST primero compila una lista preliminar de alineamientos pareados entre el query y la base de datos a los que llama pares de palabras. El algoritmo escanea la base de datos buscando pares que cumplan con un score umbral T. Cuando ocurre, ese hit se extiende utilizando un alineamiento con y sin gaps. BLAST extiende los pares de palabras para encontrar aquellos que superen un score umbral S a partir del cual se reporten los hits. Los scores se calculan a partir de una matriz de score (como la BLOSUM62) junto con las penalidades de gap. Finalmente se hace un procedimiento de trace-back para asignar la ubicación de las inserciones, deleciones y mismatches.

En general, con el fin de asignar anotaciones funcionales a los genes se suele buscar homología o similitud de secuencia utilizando algoritmos como BLAST o FASTA. Un método más sensible que estos es PSI-BLAST (Altschul et al., 1997), que realiza búsquedas iterativas utilizando un perfil de secuencia obtenido a partir de un alineamiento múltiple generado en la iteración anterior. Otros métodos de búsqueda de homología son aquellos que buscan contra bases de datos de dominios, como InterPro (Finn et al., 2017), Pfam (Finn et al., 2016) o PROSITE (Sigrist et al., 2002).

Aunque la homología de secuencia es una forma genuina de inferir funcionalidad teniendo en cuenta la evolución, no es siempre correcta esta inferencia. Además, una cantidad considerable de genes secuenciados en distintos genomas aún permanecen no anotados. Por esta razón se han propuesto métodos alternativos para mejorar la predicción funcional de proteínas.

2.1.4 Predicción de la estructura de proteínas

La similitud estructural entre proteínas puede surgir a diferentes niveles de organización estructural. Pueden ser locales, involucrando algunos pocos elementos de estructura secundaria, o global, extendiéndose a la estructura terciaria o cuaternaria global. La similitud estructural puede indicar una relación biológicamente relevante entre proteínas, aportando información sobre su función y evolución.

Existen distintas formas de predecir la estructura de una proteína, aquellos que se centran en la estructura secundaria y aquellos que predicen la estructura terciaria.

2.1.4.1 Predicción de la estructura terciaria

Dentro de estos, para una proteína que comparte una similitud sustancial (más de un 30%) a otras proteínas de estructura conocida se aplica el modelado por homología. También para proteínas que comparten plegamiento pero no necesariamente son homólogas se utiliza el threading, es decir la predicción de una estructura a partir del alineamiento con una proteínas *target* para las que se conoce su estructura. Aquellas proteínas que son análogas (relacionadas por evolución convergente en lugar de homología) pueden ser estudiadas de esta forma. Finalmente, para proteínas que no presentan homología identificable (o analogía) a proteínas de estructura conocida se utilizan métodos de tipo ab initio.

2.1.4.2 Predicción de la estructura secundaria

Las proteínas tienden a organizarse con sus aminoácidos hidrofóbicos en el interior y los hidrofílicos expuestos a la superficie. Este núcleo hidrofóbico es producido a pesar de la naturaleza altamente polar del esqueleto peptídico de una proteína. La forma más común que tiene una proteína para resolver este problema es organizar sus aminoácidos interiores en estructura secundarias consistentes en hélices alfa y hojas beta (Fig. 12). Linus Pauling y Rober Corey (Pauling, Corey, & Branson, 1951) predijeron estas estructuras a partir de estudios sobre principios químicos de hemoglobina, keratina y otros pétptidos y proteínas. Estos modelos fueron posteriormente confirmados por cristalografía de rayos X. Estas estructuras secundarias consisten en patrones de interacciones entre aminoácidos en las cuales los grupos amino y carboxilo de la cadena principal forman puentes de hidrógeno. Se conocen tres tipos de hélices: las hélices α que tiene 3.6 aminoácidos por giro y representan el ∼97% de todas las hélices, las hélices 3_{10} que tiene 3 aminoácidos por giro (por lo cual está más empaquetada) y las hélices π que ocurren con poca frecuencia y tienen 4.4 aminoácidos por giro. Las hojas beta están formadas por cadenas beta compuestas por dos a cinco aminoácidos. Se organizan en orientaciones paralelas o antiparalelas que tienen distintos patrones de puente de hidrógeno. Las hojas beta tienen propiedades de orden superior, incluyendo la formación de barriles o sánguches y motivos de estructura super secundaria, como loops beta-alfa-beta y barriles alfa/beta. Las proteínas comúnmente contienen combinaciones de tanto hélice alfa como hoja beta.

Se han desarrollado múltiples métodos para predecir la estructura secundaria de una proteína a partir de su secuencia de aminoácidos. Los más recientes utilizan alineamientos múltiples, estructuras previamente resueltas y la aplicación de aproximaciones de machine-learning como redes neuronales.

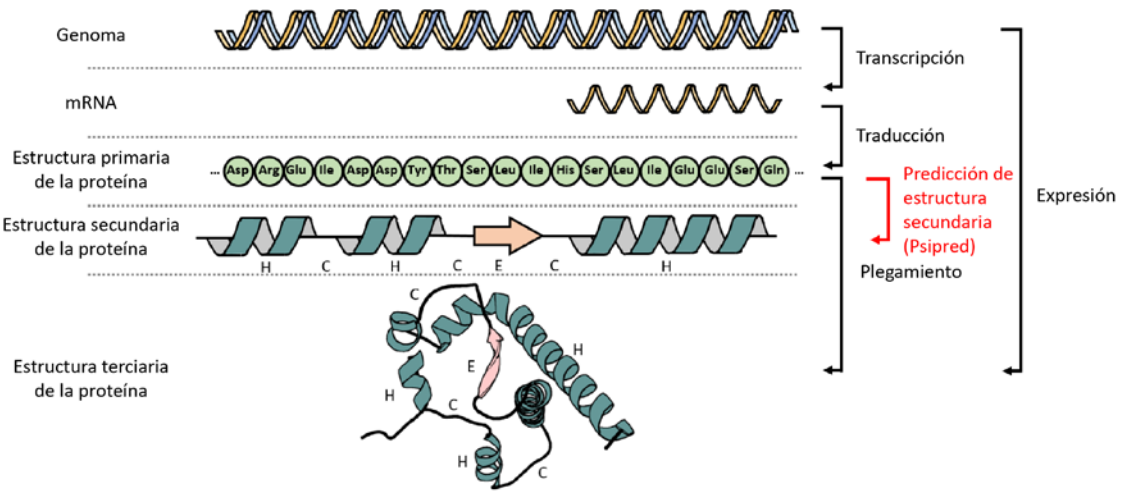


Figura 12. Contexto de predicción de la estructura secundaria de una proteína. La expresión génica es el proceso en el cual la información contenida en el genoma se utiliza para la síntesis directa de proteínas. La proteína se pliega en una molécula tridimensional funcional en dos niveles: estructura secundaria y estructura terciaria.

Uno de los métodos más utilizados es PSIPRED (Jones, 1999). El algoritmo de PSIPRED consta de tres etapas: la generación de un perfil de secuencia, la predicción de la estructura secundaria inicial y finalmente el filtrado de la estructura predicha. Para la primera parte se utilizan perfiles intermedios generados por PSI-BLAST como input de una red neuronal. A partir del perfil se genera una matriz de 20xM elementos, donde M es la longitud de la secuencia original y cada elemento representa el logaritmo de la probabiliad de la sustitución de un aminoácido particular en esa posición de la secuencia. PSIPRED luego utiliza una red de tipo feed-forward back-propagation con una única capa oculta para la predicción.

La búsqueda de alineamiento de estructuras secundarias de proteínas tomó fuerza con el surgimiento de herramientas confiables de predicción de estructuras secundarias de proteínas a partir de sus secuencias de amino ácidos. Estas herramientas devuelven, para cada amino ácido, un carácter H, E o C correspondiente a la estructura más probable para esa posición (Fig. 10). H corresponde a hélice-alfa, E a hoja-beta y C a loops o estructuras desorganizadas.

2.1.4.2.1 Métodos y aplicaciones

Sobre el método propuesto por Przytycka et al. (Przytycka, Aurora, & Rose, 1999) denominado SSEA (Secondary Structure Element Alignment) se ha basado gran parte de la literatura posterior relativa al alineamiento de estructuras secundarias. En este método la estructura secundaria de cada proteína se representa como una secuencia resumida y ordenada de los caracteres H, E y C. Los caracteres repetidos consecutivos se colapsan en lo que denominaremos elemento, y se almacena la longitud del elemento (Fig. 13).

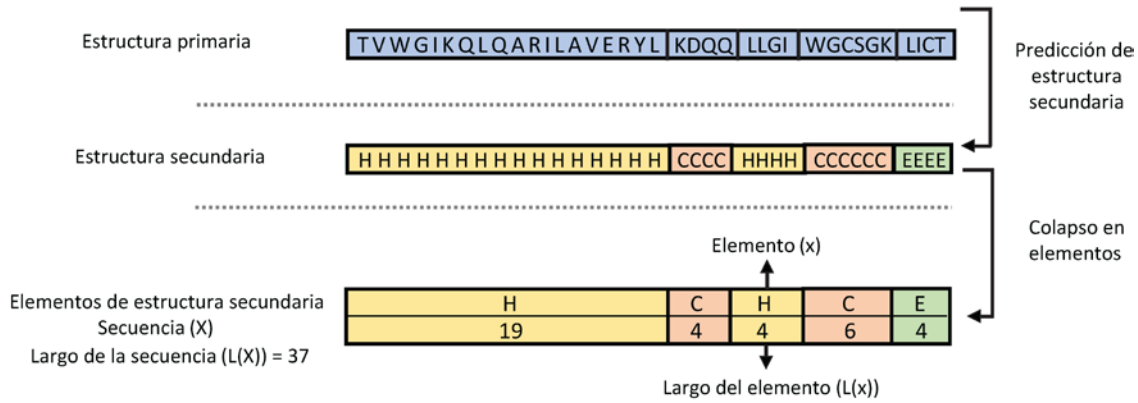


Figura 13. Codificación en elementos de estructura secundaria. El método de Przytycka implica la predicción de la estructura secundaria de las proteínas a partir de su secuencia de aminoácidos y luego el colapso en elementos de estructura secundaria.

El algoritmo que proponen Przytycka et al. es análogo al algortimo de alineamiento global basado en programación dinámica de Needleman-Wunsch, pero utilizando otro sistema de scores. Para alinear dos elementos x e y, el score se define como:

$$S(x,y)\begin{cases} min(L(x),L(y)) & if\ (x=y) \\ \frac{1}{2}min(L(x),L(y)) & if(x=\{H,E\}\ and\ y=C)\ or\ (x=C\ and\ y=\{H,E\}) \\ 0 & if(x=H\ and\ y=E)\ or\ (x=E\ and\ y=H) \end{cases}$$

Siendo L(x) y L(y) el largo del elemento x e y, respectivamente.

El score de similitud final se normaliza por el promedio los largos de las secuencias del par de proteínas. Este score final (S(X,Y)) vale entre 0 y 1, y cuanto mayor es el valor mayor es la similitud entre esas dos proteínas según su estructura secundaria.

Przytycka et al. proponen y aplican esta métrica para generar un árbol taxonómico a través de un algoritmo de clustering. El árbol generado a partir de esta métrica fue comparado con árboles generados con métodos que incluyen más información y la organización taxonómica fue concordante. Por lo tanto concluyen que la estructura secundaria permite clasificar proteínas automáticamente. Casi al mismo tiempo Xu et al (H. Xu, Aurora, Rose, & White, 1999). utilizan una métrica similar a la propuesta por Przytycka et al. para identificar dos enzimas en Archaea, que no habían podido ser identificadas por otros métodos. También realizan alineamientos utilizando un algoritmo de programación dinámica análogo al propuesto por Needleman y Wunsch, pero no colapsan los caracteres consecutivos.

McGuffin et al. (McGuffin & Jones, 2002) proponen que la predicción de la estructura secundaria de proteínas y el alineamiento de sus elementos permite distinguir homólogos lejanos de mejor forma que los métodos basados en la secuencia de amino ácidos. La capacidad de identificar homólogos lejanos a partir del alineamiento de elementos de la estructura secundaria fue evaluada también por Zhang et al. (Zhang, Kochhar, & Grigorov, 2005). La identificación de homología lejana se realizó con un método basado en Support Vector Machines (SVM) y

compararon distintas métricas. La clasificación a partir del alineamiento de estructuras secundarias obtuvo uno de los mejores valores de exactitud.

Si et al. (Si, Yan, Wang, Zhang, & Su, 2009) aplican el método para identificar proteínas que presentan un plegamiento tridimensional muy conservado denominado barril-TIM (triosa fosfato isomerasa). El alineamiento de elementos de estructura secundaria en conjunto con otros descriptores (basados en alineamientos de secuencia y dominios) permiten identificar este dominio en el proteoma de Bacillus subtilis con un 99 % de confianza utilizando SVM.

El método SSEA también se aplicó satisfactoriamente para la predicción de proteínas de membrana externa de bacterias por Yan et al. (Yan, Chen, & Zhang, 2011). La discriminación de estas proteínas frente a otros tipos de proteínas no había sido satisfactoria hasta ese momento. Se propuso la utilización del método propuesto por Przytycka et al. dada la variabilidad en determinadas regiones de la secuencia de amino ácidos y el conocimiento sobre la conservación de la estructura secundaria de estas proteínas. Más recientemente Ni et al. (Ni & Zou, 2014) desarrollaron un método basado en SVM. A partir de la métrica propuesta anteriormente, desarrolla una función de kernel que permite clasificar proteínas de membrana externa con una exactitud del 97.7%.

3. Técnicas de aprendizaje automático

La mayoría de las aplicaciones de reconocimiento de patrones o aprendizaje automático se basan en métodos de clasificación o regresión. En los métodos de regresión, se pretende encontrar una descripción funcional de los datos, en general con el objetivo de predecir valores para un nuevo input. En particular, la regresión lineal (en la cual la función es lineal en las variables del input) es la más popular y más estudiada forma de regresión. Por ejemplo, si creemos que el largo de un organismo varía linealmente con su edad o peso, tomamos medidas de la edad, el peso y la longitud de muchos ejemplares y utilizamos regresión lineal para medir los coeficientes. Por otra parte, la clasificación tradicional multiclase pretende clasificar un elemento desconocido dentro de una de varias clases predefinidas (dos en el caso más simple de clasificación binaria). Los límites de decisión están definidos por los elementos de entrenamiento (o patrones) de cada clase. Algunos problemas identificados en los métodos de clasificación convencional multiclase son la estimación del error de clasificación, la medida de la complejidad de una solución, la “maldición de la dimensionalidad” y la generalización del método de clasificación. La mayoría de los clasificadores convencionales funcionan correctamente trabajando con clases igualmente balanceadas y su desempeño disminuye cuando una clase está subrepresentada o ausente. Este último corresponde al caso extremo en el que los elementos a clasificar pueden pertenecer o no a las clases definidas por el set de entrenamiento.

3.1. Función de One-Class Classification

Existe una extensión del problema de clasificación denominada “Data Domain Description” o “One-Class Classification” (OCC). El problema de clasificación en una sola clase positiva es diferente a la clasificación multiclases ya que en OCC, la clase positiva está muy bien muestreada mientras que la clase negativa está ausente, poco muestreada o mal definida. En OCC, una de las clases (definida como la clase positiva) está bien caracterizada por el set de entrenamiento.

Para la otra clase (negativa), no existen entradas o son muy pocas, o no forman una muestra estadísticamente representativa de la clase negativa. La tarea en OCC es entrenar un modelo con objetos de una clase positiva y clasificar otro conjunto de objetos, de forma que acepte la mayor cantidad de objetos posibles de la clase positiva, mientras que minimiza la chance de aceptar objetos no positivos, o outliers. En particular, cuando se trata de métodos de barrera se define una barrera de clasificación alrededor de la clase positiva y no es trivial decidir qué tan justa debería ajustarse la barrera en cada una de las direcciones alrededor de los datos basándose en solamente una clase. Tampoco es trivial decidir qué atributos deberían ser utilizados para encontrar la mejor separación de elementos de la clase positiva y *outliers*. En particular, cuando la barrera de los datos es extensa y no convexa, la cantidad de elementos del set de entrenamiento debe ser muy grande para obtener una solución satisfactoria. Por lo tanto, se espera que los algoritmos de OCC requieran un buen muestreo del dominio y un mayor número de elementos de entrenamiento con respecto a los algoritmos de clasificación multiclase convencionales. Los problemas identificados en los clasificadores multiclase también aparecen en OCC y pueden incluso presentar mayores inconvenientes.

3.2. Usos de One-Class Classification

En OCC la tarea no es distinguir entre clases de objetos como en los problemas de clasificación o producir la salida deseada para cada elemento de entrada como en los problemas de regresión, sino que es dar una descripción de un set de datos. Esta descripción debería abarcar toda la clase representada por el set de entrenamiento, e idealmente debería rechazar todos los otros elementos posibles en el espacio de elementos. Por esta razón OCC es utilizado para detección de *outliers* o *novelty detection*, es decir, la detección de objetos que difieren significativamente del set de datos de entrenamiento. La ausencia de muestras negativas puede deberse al alto costo de generarlas o a la baja frecuencia a la que ocurren eventos anormales. Por ejemplo, OCC puede ser relevante detectando fallas en máquinas. Un clasificador debería detectar cuándo una máquina está mostrando un comportamiento anormal o defectuoso. Las medidas del funcionamiento normal de la máquina (datos de entrenamiento compuestos por la denominada clase positiva) son fáciles de obtener. Por otra parte, la mayoría de los defectos podrían no haber ocurrido en la realidad, por lo que es difícil o incluso imposible obtener una clase negativa bien muestreada. Además, podríamos no querer esperar hasta que los defectos ocurran, ya que podrían involucrar un alto costo, malfuncionamiento de la máquina o riesgo para los operadores humanos.

Otro ejemplo es el diagnóstico automático de una enfermedad. Es relativamente simple compilar datos positivos (todos los pacientes que se conoce que tienen una enfermedad común) pero los datos negativos pueden ser difíciles de obtener dado que no se puede asumir que los otros pacientes en la base de datos son casos negativos si nunca fueron evaluados, y esas evaluaciones pueden ser costosas.

También permite clasificar proteínas a partir de dominios conservados. Dado un dominio conservado identificado en la secuencia de aminoácidos o en la estructura en cierto conjunto de proteínas, es posible utilizar este conjunto como set de entrenamiento de un clasificador OCC. No es posible definir una clase negativa y resolverlo como un problema multiclase ya que la diversidad de la clase negativa es enorme.

3.3. Métodos de One-Class Classification

Los OCC pueden ser divididos en cinco categorías: probabilísticos, basados en distancia, basados en reconstrucción, basados en dominios e information-theoretic techniques (Pimentel, Clifton, Clifton, & Tarassenko, 2014). Los métodos OCC difieren en la definición de la distancia o similitud, y en la optimización del umbral respecto al set de entrenamiento. En este trabajo se exploraron las primeras cuatro categorías.

3.3.1 Detección probabilística

Las aproximaciones probabilísticas asumen que los datos de entrenamiento están generados por una distribución de probabilidad subyacente que puede ser estimada. A la distribución resultante se le puede luego asignar un umbral para definir los límites de la clase positiva en el espacio de elementos y evaluar si un elemento del conjunto de test pertenece a la misma distribución o no. Estos métodos asumen que las regiones de baja densidad en el set de entrenamiento indican que esas áreas tienen una baja probabilidad de contener elementos positivos. Cuando se optimiza un umbral se encuentra un volumen mínimo para el modelo de densidad de probabilidad dado. Por construcción, solamente las áreas de gran densidad se incluyen y las regiones poco representadas no son incluidas en la descripción. El umbral se determina utilizando la distribución objetivo empírica.

Las aproximaciones probabilísticas tienen un buen soporte matemático y pueden detectar novedad de forma eficiente si se obtiene una estimación precisa de la función de densidad de la probabilidad. Además, una vez construido el modelo se requiere una cantidad mínima de información para representarlo, por lo que no es necesario almacenar todo el set de datos utilizado para el entrenamiento. Los métodos probabilísticos también se conocen como métodos transparentes ya que sus *outputs* pueden ser analizados usando técnicas numéricas estándar. De todas formas, el desempeño de estas aproximaciones es limitado cuando el tamaño del set de entrenamiento es muy pequeño, en particular en espacios de dimensión moderadamente alta debido a la denominada “maldición de la dimensionalidad”.

Dentro de la detección probabilística existen dos grandes grupos de aproximaciones: las paramétricas (asumen un número finito de parámetros) y las no paramétrica (asumen que la distribución de los datos no puede ser definida a partir de un conjunto finito de parámetros).

3.3.1.1 Aproximaciones paramétricas

Estas aproximaciones asumen que el conjunto de entrenamiento está generado por una distribución paramétrica subyacente con parámetros estimados a partir de los datos de entrenamiento, donde el conjunto de parámetros es finito. Se puede asumir una variedad de distribuciones, como la de Gauss o la de Poisson, y existen diversos métodos para evaluar elementos del conjunto de test.

3.3.1.1.1. Gaussian model

La distribución más comúnmente utilizada para variables continuas es la gaussiana. Este método asume que los datos de entrenamiento están distribuidos de acuerdo a la distribución normal.

Los parámetros son estimados a partir de los datos de entrenamiento utilizando la estimación de máxima verosimilitud (MLE) para la cual existe una solución analítica cerrada para la distribución gaussiana.

3.3.1.1.2. Gaussian mixture models

La distribución gaussiana asume un modelo muy específico de los datos. Este debe ser unimodal y convexo lo cual para la mayoría de los sets de datos no se cumple. Para obtener un método de densidad más flexible, la distribución gaussiana puede ser extendida a una mezcla de gaussianas. Esta es una combinación lineal de distribuciones normales. Tiene un sesgo menor que la distribución gaussiana simple, pero requiere muchos más datos para estimar los parámetros del modelo. En general se utiliza una cantidad de kernels menor que el número de elementos en el set de entrenamiento. Cuando el número de gaussianas se define de antemano por el usuario, las medias y las covarianzas de las gaussianas individuales pueden ser estimadas eficazmente con una rutina de esperanza-maximización (EM) (Figueiredo, 2002).

3.3.1.2. Aproximaciones no paramétricas

Estas aproximaciones no asumen que la estructura de un modelo es fija, sino que el modelo crece en tamaño cuanto sea necesario para ajustarse y acomodarse a la complejidad de los datos. La técnica estadística no paramétrica más simple es el uso de histogramas, los cuales muestran frecuencias tabuladas de forma gráfica. El algoritmo define típicamente una medida de la distancia entre un nuevo elemento de entrenamiento y el modelo basado en un histograma para determinar si es un outlier o no.

3.3.1.2.1 Estimador de Parzen

La estimación de la densidad es generalmente una mezcla de kernels gaussianos centrados en cada uno de los elementos de entrenamiento (Fig. 14). El estimador de Parzen suele ser una extensión de la mezcla de gaussianas: ubica un kernel (típicamente gaussiano) en cada punto de los datos y suma las contribuciones locales desde cada kernel. Entonces, el entrenamiento de la densidad de Parzen consiste en la determinación de un único parámetro, el ancho óptimo del kernel. Este es optimizado utilizando máxima verosimilitud. Dado que se estima únicamente un parámetro el modelado de los datos es muy débil, como en un modelo no paramétrico. Una buena descripción depende totalmente de la representatividad del set de entrenamiento. El costo computacional para entrenar un estimador de la densidad de Parzen es relativamente bajo.

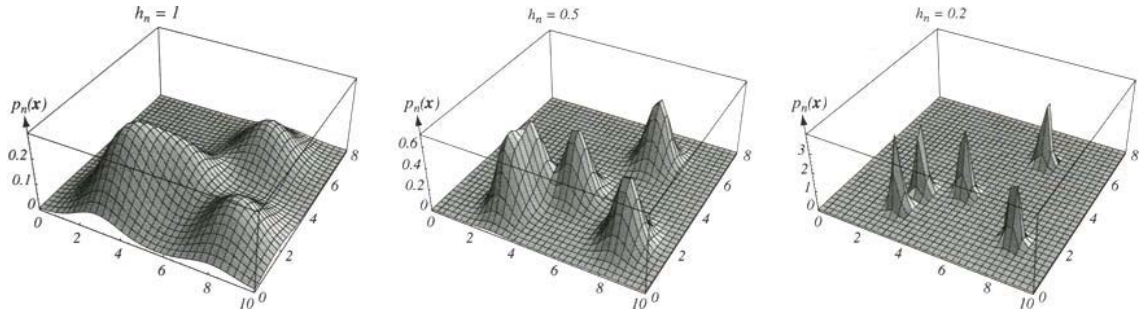


Figura 14. Ejemplos de ventanas de Parzen circulares en dos dimensiones utilizando distintos valores de ventana para cinco muestras. Tomada de Duda et al. (Duda, Hart, & Stork, 2001).

3.3.2 Detección basada en distancia

Cuando se dispone de una cantidad limitada de datos, uno debería evitar resolver un problema más general, como la estimación de una distribución de probabilidad, como paso intermedio para resolver el problema original de clasificación. Para resolver este problema más general se podrían necesitar más datos que para resolver el problema original. Esto significa que estimar la densidad completa de los datos para un OCC podría ser muy demandante cuando solamente se necesita un límite. En los métodos basados en distancia, por lo tanto, solamente se optimiza un límite cerrado alrededor del set objetivo. Aunque el volumen no siempre es minimizado activamente en los métodos basados en distancia, la mayoría de los métodos tiene un sesgo importante hacia la solución de volumen mínimo. Qué tan pequeño es el volumen depende del ajuste del método a los datos.

Dado que estos métodos dependen de forma importante en las distancias entre los objetos, tienden a ser sensibles al escalado de las características. Por otro lado, el número de objetos que se requiere es más pequeño que para los métodos basados en probabilidad, aunque es importante definir correctamente las distancias.

El output de estos métodos no puede ser interpretado como una probabilidad. El modelo es entrenado de forma que una fracción del set objetivo es rechazada, por lo que se obtiene un umbral para esta tasa de rechazo. Al disminuir el umbral el modelo se ajusta más a los datos, pero no garantiza que se capturen las áreas de alta densidad. Al cambiar dicha fracción se podría requerir volver a entrenar el método.

Generalmente, en sets de alta dimensión es computacionalmente costoso calcular la distancia entre elementos y como resultado estas técnicas no tienen escalabilidad. Los métodos basados en clustering son capaces de ser usados en modelos incrementales, ya que se pueden incorporar nuevos elementos al sistema y testearlos para identificar novedad. Se han desarrollado nuevas técnicas para optimizar el proceso de detección de novedad y reducir la complejidad temporal respecto al tamaño de los datos. De todas formas estas técnicas sufren de tener que elegir un valor apropiado de tamaño o número de clusters y también son susceptibles a la “maldición de la dimensionalidad”.

3.3.2.1 Métodos basados en vecinos más cercanos

Estos están entre los métodos más comúnmente utilizados para detectar novedad. La aproximación k-vecinos más cercanos (k-NN) se basa en la premisa que los elementos de la clase positiva tienen vecinos cercanos en el set de entrenamiento, mientras que aquellos que no pertenecen a la clase positiva se ubican lejos de estos elementos. Un punto se define como outlier si está ubicado lejos de sus vecinos (Fig. 15). Al aumentar k disminuye la sensibilidad local del método, pero hace que el método sea menos sensible al ruido. La distancia euclídea es una opción popular para atributos continuos univariados o multivariados, pero se pueden utilizar otras medidas como la de Mahalanobis, medidas adaptadas al problema, o donde se introduce el conocimiento previo del problema.

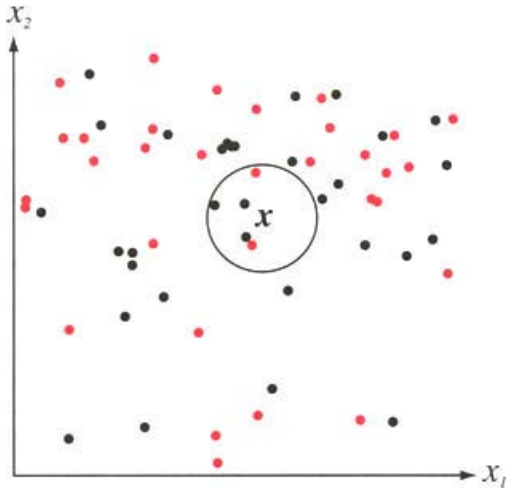


Figura 15. Ejemplo de k-vecinos más cercanos. La búsqueda de los k-vecinos más cercanos comienza con el punto x. Se aumenta el tamaño de una región circular centrada en este hasta que se encierran k puntos de entrenamiento. Se etiqueta al punto x de acuerdo a la mayoría de los puntos de entrenamiento encerrados. En este caso k=5 por lo que el punto x se debería etiquetar en la categoría de los puntos de color negro. Tomada de Duda et al. (Duda et al., 2001).

3.3.2.2 Métodos basados en clustering

En estos métodos generalmente la clase positiva está caracterizada por un número más o menos pequeño de elementos prototipo en el espacio de datos. La distancia mínima de un elemento del conjunto de test al prototipo más cercano se utiliza generalmente para cuantificar anormalidad.

3.3.2.2.1 k-means

También conocido como algoritmo de Lloyd generalizado (Lloyd, 1982). Es probablemente el método más popular de clustering de datos estructurados dada la simplicidad de implementación. El primer paso en el algoritmo es seleccionar un conjunto de centros de cluster inicial. Una de las formas de hacerlo es seleccionando k centros de cluster aleatoriamente. Luego se calculan las distancias entre estos centros y cada elemento en el set de entrenamiento, y se identifica aquellos puntos que están más cerca de cada centro de cluster. Los correspondientes centros de cluster son trasladados al centroide de estos puntos más cercanos y se repite el procedimiento. Esto puede representarse como la asignación de elementos de entrenamiento de acuerdo a un diagrama de Voronoi generado por los centros de cluster (Fig. 16). El algoritmo converge cuando los centros de cluster no se mueven de una iteración a la siguiente. En general se utiliza este método en etapas de preprocesamiento.

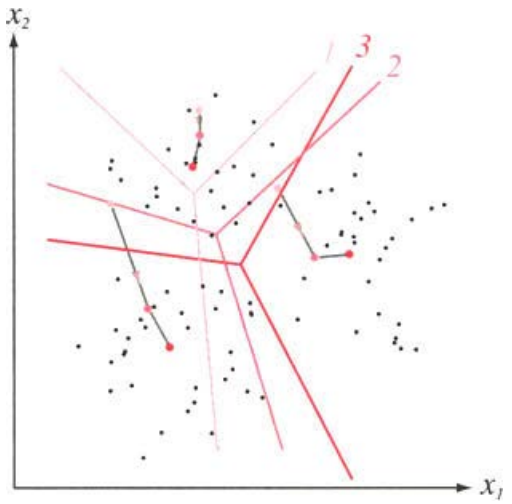


Figura 16. Ejemplo de k-means. Trayectorias de las medias del procedimiento de clustering aplicado a datos en dos dimensiones. Se seleccionó un set inicial (rosado más claro) de tres centroides (correspondientes a tres clusters). Cada punto del set de datos se asigna al cluster correspondiente al centroide más cercano. Se actualiza la posición de los centroides como la media de los puntos asignados a cada cluster y se repite el algoritmo dos veces más (rosado medio y rosado fuerte). Se muestran los diagramas de Voronoi, donde las medias corresponden a los centros de las celdas de Voronoi. En este caso la convergencia de alcanzó en tres iteraciones. Tomada de Duda et al. (Duda et al., 2001).

3.3.2.2.2 Minimum spanning tree

El método se basa en una representación en forma de grafo de la clase positiva para capturar la estructura subyacente de los datos, ajustando un Minimum Spanning Tree (MST) sobre el set de entrenamiento (Fig. 17). Este método realiza una clasificación basada en similitud con la clase positiva, dado que calcula la distancia desde un elemento del set de test a su arista más cercana (Graham & Hell, 1985). Se demostró que el método funciona bien cuando el tamaño de los datos es pequeño y de alta dimensionalidad.

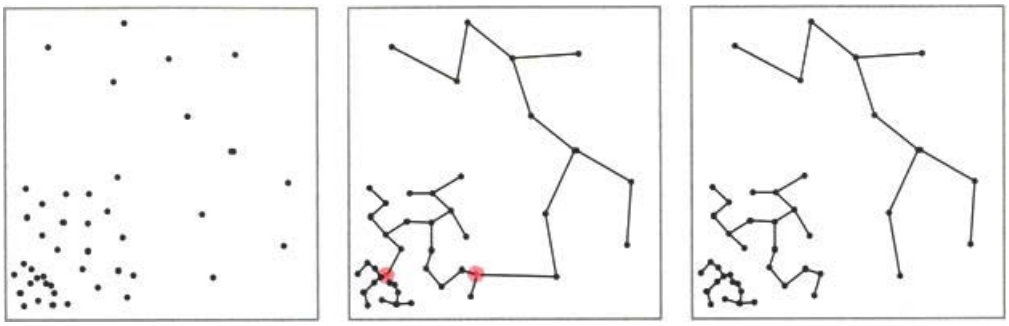


Figura 17. Ejemplo de minimum spanning tree. Los datos originales de muestran a la izquierda y su minimum spanning tree en el medio. En cada nodo, las aristas son de aproximadamente el mismo largo, con excepción de los marcados en rojo. Cuando se eliminan esas aristas inconsistentes, se producen tres clusters como se muestra a la derecha. Tomada de Duda et al. (Duda et al., 2001).

3.3.3 Detección basada en reconstrucción

Son una clase de métodos muy flexibles que son entrenados para modelar la distribución de los datos subyacente sin suposiciones a priori sobre las propiedades de los datos. Un ejemplo son las redes neuronales que requieren la optimización de un número predefinido de parámetros que definen la estructura del modelo, y su desempeño puede ser muy sensible a estos parámetros. Por lo tanto, puede ser muy difícil entrenar en espacios de dimensión muy alta. Además, las redes que usan algoritmos constructivos, en los cuales a la estructura del modelo se le permite crecer, sufren del problema adicional de tener que seleccionar el método más efectivo de entrenamiento para permitir la integración de nuevas unidades en la estructura del modelo existente, y un criterio de parada apropiado (cuándo dejar de agregar nuevas unidades).

3.3.3.1 Autoencoder neural networks

El autoencoder tradicional es una red neuronal que intenta reproducir su input, por lo que el output esperado es el propio input, es decir que intenta aprender una aproximación de la función identidad (Fig. 18). De manera más formal, un autoencoder toma cada vector input x y lo mapea a una representación oculta y. Esta representación, a veces denominada latente, se mapea de vuelta a un vector reconstruido $\hat{x}$. La idea básica es que el autoencoder está construido de tal forma que el mapeo de x en y revela una estructura esencial en cada vector x que de otra forma no es obvia. Por ejemplo, si el autoencoder tiene menos unidades ocultas que unidades de input, debe encontrar una representación que reduce el input de tal forma que puede ser reconstruido eficientemente. La presentación reducida tiene menor dimensionalidad que el input y representa una abstracción del input.

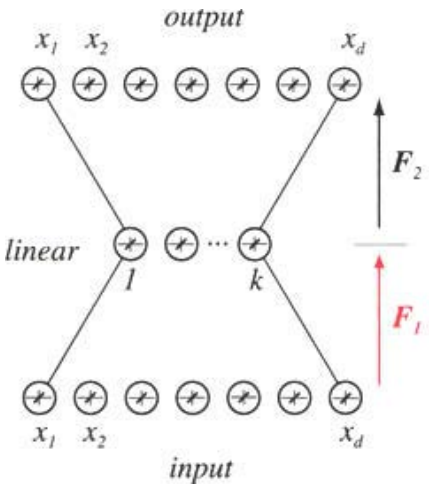


Figura 18. Ejemplo de red neuronal de tres capas con unidades ocultas lineales, entrenada para ser un autoenconder. Tomada de Duda et al. (Duda et al., 2001).

3.3.4 Detección basada en dominios

Estos métodos requieren la creación de un límite basado en la estructura del set de entrenamiento. Estos métodos son típicamente insensibles al muestreo específico y densidad de la clase positiva ya que describen el límite de la clase positiva, o el dominio, y no la densidad de clase. La pertenencia de elementos desconocidos a la clase es determinada por su ubicación respecto al límite. Como los métodos SVM de dos clases, el método one-class SVM determina la ubicación del límite de novedad utilizando solamente aquellos datos que se ubican más cerca del mismo en el espacio transformado (vectores de soporte). Todos los otros datos del set de entrenamiento (aquellos que no son vectores de soporte) no son considerados al establecer un límite de novedad. Por lo tanto, la distribución de los datos en el set de entrenamiento no es considerada lo cual es visto como "no resolver un problema más general de lo que es necesario".

SVM es una técnica popular para formar límites de decisión que separan los datos en distintas clases. La SVM original es una red que es idealmente adecuada para clasificación binaria de datos que son linealmente separables. Una particularidad del método es el uso de funciones kernel, que proyectan el problema en un espacio de características de mayor dimensión (incluso infinito). Esto permite la búsqueda del hiperplano que proporciona la máxima separación entre dos clases en el espacio transformado. Los elementos de entrenamiento que se ubican cerca de la frontera definiendo este margen separador son llamados vectores de soporte (Fig. 19).

Desde la introducción de la idea original se han propuesto diversas modificaciones y mejoras. Se han propuesto dos aproximaciones para detectar novedad, Support Vector Data Description (SVDD) y One-Class SVM (OCSVM).

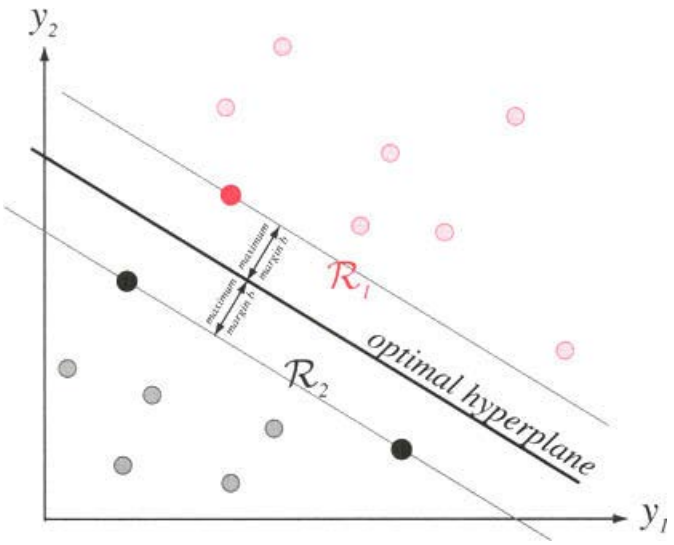


Figura 19. Ejemplo de SVM. El entrenamiento con una support vector machine consiste en identificar el hiperplano óptimo, es decir el que presente una distancia máxima a sus patrones de entrenamiento más cercanos. Los vectores de soporte son aquellos patrones que se separan una distancia b del hiperplano. Los tres vectores de soporte se indican como puntos sólidos. Tomada de Duda et al. (Duda et al., 2001).

Estas aproximaciones determinan la ubicación del límite de novedad utilizando solamente aquellos datos que se ubican más cerca y no se basan en propiedades de la distribución de los datos en el set de entrenamiento. Una desventaja de estos métodos es la complejidad asociada con el cálculo de funciones de kernel. Aunque se han propuesto algunas extensiones para superar este problema, la elección de la función de kernel apropiada también puede ser problemática. Además, no es fácil seleccionar valores para los parámetros que controlan el tamaño de la región límite.

3.3.4.1 One-class support vector machine

La idea de OCSVM fue propuesta por Schölkopf et al. (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt, 2000). El objetivo de esta aproximación es encontrar un hiperplano que separe los elementos de entrenamiento del origen con un determinado umbral. Para manejar problemas no lineales, se puede utilizar un kernel de forma de mapear los elementos en un nuevo espacio de características y luego clasificar los elementos transformados con margen máximo. Esta aproximación requiere fijar a priori el porcentaje de elementos positivos que se permitirá que caiga fuera de la descripción de clase positiva. Esto hace al método más tolerante a outliers en el set de entrenamiento. De todas formas, el establecimiento de este parámetro influye fuertemente en el desempeño de este método.

3.3.4.2 Support vector data description

El método SVDD propuesto por Tax y Duin (Tax & Duin, 2004), define el límite de novedad como una hiperesfera (en lugar de un hiperplano) con el mínimo volumen que comprende todos o la mayoría de los elementos del set de entrenamiento. SVDD automáticamente optimiza los parámetros del modelo utilizando datos sin etiquetar generados artificialmente distribuidos en una hiperesfera alrededor de la clase positiva. Esto causa que el método pueda manejar aplicaciones que involucran alta-dimensionalidad. La detección de novedad se determina si un elemento del conjunto de test se encuentra dentro de la hiperesfera. De forma de lidiar con el problema de que los datos transformados no están distribuidos de forma esférica, Campbell y Bennett (Campbell, 2001) utilizan distintos kernels con métodos de optimización de programación lineal, en lugar de las aproximaciones de programación cuadráticas típicamente utilizadas en SVM.

SVDD separa los datos de interés de distintas clases ubicando una hiperesfera alrededor de la clase de objetos que son representados por el set de entrenamiento de todos los otros posibles objetos en el espacio de objetos. La hiperesfera está definida por un centro y un radio. Se puede agregar un umbral para permitir que el modelo de hiperesfera rechace una fracción de los objetos de entrenamiento, lo cual disminuye el volumen de la hiperesfera. Las barreras de la hiperesfera pueden hacerse más flexibles introduciendo funciones de kernel de ancho definido por el usuario.

3.4 Combinación de clasificadores

Como en los problemas de clasificación multiclase, un OCC puede no capturar todas las características de los datos. Sin embargo, utilizando solamente el mejor clasificador y descartando los clasificadores con desempeños pobre se podría estar eliminando información valiosa. Una solución viable para mejorar el desempeño de distintos clasificadores, que pueden diferir en la complejidad o en el algoritmo de entrenamiento subyacente utilizando para construirlo, es la combinación de un conjunto de clasificadores. Esto puede servir para mejorar el desempeño y también la robustez de clasificación. Los clasificadores son comúnmente combinados, por ejemplo, para proveer una decisión combinada promediando las probabilidades posteriores estimadas.

3.5 Paquete Dd_tools

El paquete data description toolbox (Dd_tools) (Tax, 2015) contiene herramientas, clasificadores y funciones de evaluación aplicadas a OCC. Este paquete es una extensión del paquete PRTools (Duin et al., 2007) de Matlab. Entre sus funciones se encuentran la distribución Gaussiana (gauss_dd), la mezcla de Gaussianas (mog_dd), el estimador de densidad de Parzen (parzen_dd), las redes neuronales autoencoder (autoenc_dd), el método k-means (kmeans_dd), el método mínimum spanning tree (mst_dd), el método de vecinos más cercanos (knndd), el método de support vector data description (svdd), y el método de programación lineal (lpdd). En el anexo se detallan las herramientas.

OBJETIVO 1

4 Desarrollo de una métrica de similitud entre estructuras secundarias.

Siguiendo la hipótesis que plantea que proteína fusogénicas presentes en VEs pueden tener similitudes estructurales con fusógenos conocidos, el primer objetivo de este trabajo fue desarrollar una métrica de similitud con el fin de clasificar proteínas de fusión viral en clase I y clase II para así poder identificar nuevas proteínas fusogénicas que se enmarquen en estas dos clases. En la Fig. A1 se esquematiza todo el procedimiento así como los resultados de este objetivo.

Dado que toda la bibliografía disponible que aborda el problema de la búsqueda de similitud de estructuras secundarias se basa en la métrica propuesta por Przytycka et al., utilizaremos esta métrica como punto de partida.

Las proteínas de fusión viral se sintetizan como precursores inactivos (VFPs) que en determinadas condiciones se escinden o clivan, liberando una proteína transmembrana con capacidad fusogénica. La proteína de fusión consta de una región extracelular o ectodominio, la cual lleva a cabo el mecanismo de fusión, un dominio transmembrana que permite su anclaje a la membrana de la célula a la que pertenece, y un dominio intracelular que tiene funciones asociadas a la señalización intracelular. El ectodominio de la proteína fusogénica, al cual denominaremos VF, es la región de la proteína que presenta una estructura conservada respecto de su clase (Fig. 20). Por esta razón, todos los modelos implementados en este trabajo fueron entrenados con los ectodominios de los fusógenos virales.

Consideramos que las proteínas candidatas a tener capacidad fusogénica en otros modelos biológicos podrían también presentarse como parte de un precursor. Por esta razón además de evaluar cómo funciona la métrica para clasificar un conjunto de VFs de clase I y clase II, también evaluamos cómo funciona la métrica para clasificar un conjunto de VFPs de clase I y clase II. Dada la reducida cantidad de secuencias de fusógenos virales de clase III disponibles se decidió no trabajar con ellas ya que las clases estarían muy desbalanceadas, influyendo en el desempeño de los clasificadores.

A partir de esta métrica se evaluó la clasificación de proteínas en dos etapas principales. La primera etapa se basa en la clasificación de fusógenos virales (VFs) en clase I o clase II con el fin evaluar una métrica análoga a la propuesta por Przytycka et al. que funcione bien para los datos. Para esta clasificación trabajamos con un clasificador de tipo SVM y con k-Nearest Neighbors (k-NN). El conjunto de entrenamiento lo definimos como un conjunto de VFs de clase I y clase II.

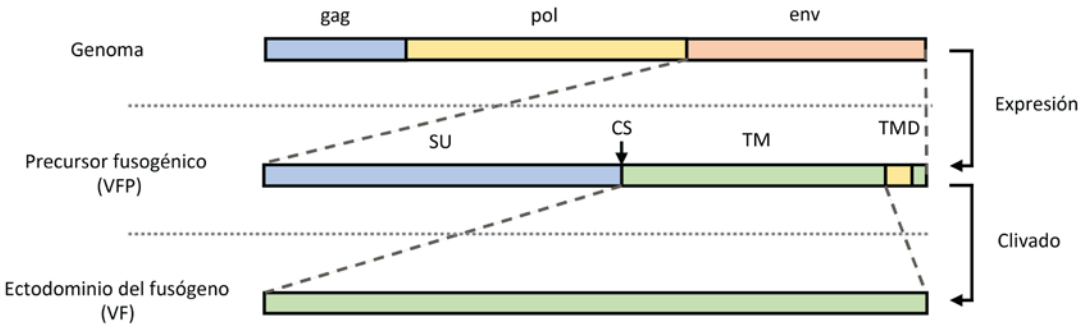


Figura 20. Ejemplo de obtención del fusógeno para un retrovirus. El gen env codifica el precursor fusogénico, que en determinadas condiciones se cliva para dar lugar a dos proteínas, una de ellas el fusógeno. En este trabajo nos referiremos al fusógeno o VF como la porción extracelular de la proteína de fusión.

La segunda etapa consiste en entrenar un clasificador del tipo One-class SVM (OC-SVM) con VFs de clase I y clasificar un conjunto de VFs como clase I (clase positiva) o clase II (clase negativa). Se pretende evaluar el método con el fin de considerar su aplicación para la clasificación de proteínas de otros modelos biológicos en proteínas con capacidad fusogénica (clase positiva) y no fusogénica (clase negativa).

4.1 Tratamiento de datos de fusógenos virales clase I y II

Como primer paso se obtuvieron las secuencias representativas de VFs y VFPs de clase I y clase II con las que posteriormente se evaluó la métrica. A cada una de estas secuencias se les calculó su estructura secundaria y se colapsó la misma según el algoritmo de Przytycka et al. Finalmente a cada par de secuencias se les aplicó el algoritmo SSEA modificado en este trabajo para obtener una representación de todas las similitudes entre las secuencias analizadas.

4.1.1 Definición de región fusogénica

Se obtuvieron las secuencias de aminoácidos de las VFPs depositadas en la base de datos pública UniProt (The UniProt, 2017). Se seleccionaron aquellas proteínas etiquetadas según UniProt como fusógeno viral de clase I o fusógeno viral de clase II. Se obtuvieron 27846 secuencias de clase I y 1800 de clase II, de largos variables de hasta 691 aminoácidos. Se extrajo la región fusogénica de cada proteína a partir de las anotaciones presentes en UniProt, y de aquí en adelante se trabajó con la VFP y el VF en paralelo.

4.1.2 Clustering

Se tenía conocimiento de la redundancia de secuencias en UniProt. Esto se observa particularmente para la proteína Hemaglutinina, fusógeno de clase I del virus de la Influenza, de la cual se encuentran depositadas secuencias de miles de cepas. Por esta razón, como paso previo a la predicción de estructuras secundarias se clusterizaron las secuencias al 99 % de identidad con la herramienta CD-HIT [Fu, 2012]. Aplicando este algoritmo se obtuvieron 1769 secuencias representativas de fusógenos virales de clase I y 1103 de clase II. Se seleccionaron al azar 100 VFPs de clase I y 100 VFPs de clase II (set de test VFP), y también a sus correspondientes 100 VFs de clase I y 100 VFs de clase II (set de test VF). Además se seleccionaron al azar otras 100 VFs de clase I y 100 VFs de clase II (set de entrenamiento).

4.1.3 MSA y predicción de estructura secundaria

Se calcularon las predicciones de estructura secundaria de estas 600 secuencias. Esto se realizó con la herramienta Psipred, a través del paquete HHSuite (Soding, 2005).

4.1.4 Cálculo de distancias

Para medir la similitud entre dos proteínas nos basamos en el método descripto por Przytycka et al. en la sección 2.1.4.2.1. Dado que trabajamos con VFs, pero VFPs cuya homología con VFs está reducida a una fracción, esperaríamos que un alineamiento de tipo global fallara. Por esta razón, como primera modificación al algoritmo, proponemos aplicar un algoritmo de alineamiento local, análogo al de Smith-Waterman, que permita clasificar las VFPs. Se utilizan algoritmos de alineamiento local cuando las secuencias a alinear tienen longitudes distintas y se buscan regiones de similitud. Por lo tanto, para calcular la métrica de similitud realizamos alineamientos de las representaciones de las estructuras secundarias de las proteínas (VFs y VFPs) de a pares, aplicando el método descripto por Przytycka et al., pero sustituyendo el algoritmo de Needleman-Wunsch por el de Smith-Waterman. Aunque esta modificación del método fue propuesta por Fontana et al. (Fontana et al., 2005) no fue aplicada a un problema concreto y no hemos identificado otros artículos que la utilicen. El código para el cálculo de similitud entre proteínas se escribió en Python.

En el método de Przytycka et al. el score final corresponde al score del último casillero de la matriz de programación dinámica de un alineamiento global. Dado que trabajamos con un algoritmo de alineamiento local, nuestro score final corresponde al score máximo obtenido en la matriz de programación dinámica.

La longitud de las secuencias de los VFs es variable entre las dos clases y también dentro de cada clase. Lo mismo sucede para la longitud de las secuencias de sus VFPs, por lo que no se esperaría encontrar relación entre la longitud de los VFs y sus VFPs. Por esta razón resulta esperable que falle el método de normalización propuesto por Przytycka et al., donde el score final del alineamiento se divide entre el promedio los largos de las secuencias del par de proteínas. Proponemos una segunda modificación de la métrica, donde el score final se normaliza por el promedio de la longitud de las regiones alineadas para cada par de proteínas.

El trabajo de Przytycka et al. no realiza un análisis detallado del rol de la penalización del gap en el alineamiento (ver sección 2.1). Para proponer una tercera modificación de la métrica, analizamos las matrices de similitud obtenidas para valores de penalización del gap entre 0 y 10, eligiendo trabajar con un valor penalización igual a 5 (Fig. 21). Este valor maximiza el score al comparar secuencias de una misma clase, y minimiza el score al comparar secuencias de distinta clase. Mientras que para los pares de VFs la similitud es muy alta dentro de cada clase y baja entre pares de distinta clase, cuando se trata similitud entre VFs y VFPs esta diferencia no es tan clara. Para todos los casos, el valor de penalización de gap que maximiza la similitud dentro de una misma clase es 5.

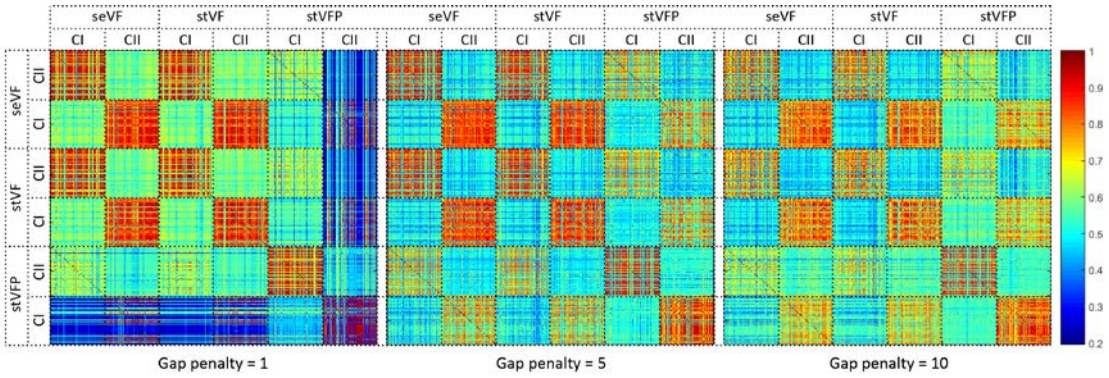


Figura 21. Matrices de similitud entre VFs y VFPs de fusógenos virales de clase I y II para valores de penalización de gap 1, 5 y 10. La escala de color corresponde a los valores de similitud obtenidos al aplicar el algoritmo de similitud a cada par de proteínas. VF: set de entrenamiento VF, stVF: set de test VF, stVFP: set de test VFP, CI: fusógenos virales de clase I, CII: fusógenos virales de clase II.

4.2 Evaluación de la métrica para clasificar fusógenos virales de clase I y II

Las similitudes calculadas en la sección 4.1.4 fueron transformadas en distancias, y con estas distancias se armó una matriz de distancias. Esta matriz se utilizó para evaluar el desempeño de distintos clasificadores utilizando la métrica trabajada en la sección anterior. Primero se clasificaron únicamente secuencias de VFs, entrenando con clase I y clase II, o únicamente con una de las clases como clase positiva, con el fin de tener una primera aproximación al desempeño de la métrica. Posteriormente se clasificaron secuencias de VFPs, entrenando con VFs de clase I y clase II, o únicamente una de las clases como positiva, con el fin de evaluar el funcionamiento de la modificación del algoritmo para alineamientos locales.

El método SVM ha sido ampliamente utilizado en el análisis de secuencias biológicas. En este trabajo transformamos las distancias a través de un kernel exponencial:

$$k(x) = \exp(\gamma d(X, Y))$$

Se trabajó con el paquete LIBSVM (Chang & Lin, 2011) para Python. LIBSVM permite generar un clasificador a partir de un kernel precalculado y estimar su desempeño.

4.2.1 Clasificación de VFs entrenando con dos clases

Entrenamos un clasificador de tipo SVM de dos clases con un conjunto de VFs de clase I y clase II. La clasificación se realizó con otro conjunto de VFs de clase I y clase II. El desempeño del clasificador depende de los parámetros C y γ. El parámetro C aporta cierta flexibilidad en la clasificación, permite algunos errores pero también los penaliza.

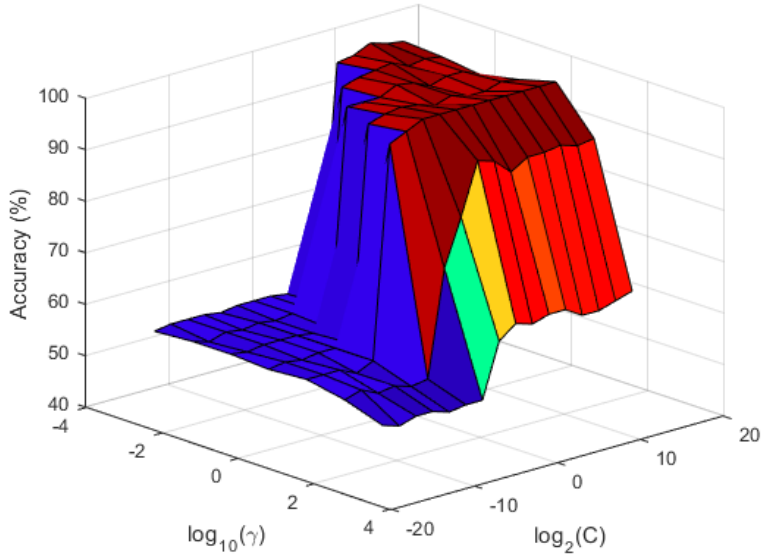


Figura 22. Grid search para la selección de parámetros de SVM para la clasificación de VFs entrenando con dos clases.

El parámetro γ define qué tan lejos llega la influencia de una muestra. La mejor combinación de C y γ se seleccionó a partir de una búsqueda "grid search" utilizando "10-fold cross-validation" con valores de C entre 2^{-15} y 2^{15} y valores de γ entre $5e^{-4}$ y $5e^{2}$ en intervalos uniformes (Fig. 22). Existe una región de combinación de los parámetros γ y C donde el desempeño no se ve afectado significativamente. Por lo tanto se seleccionaron como parámetros óptimos aquellos que se ubican en el centro de esta región: C = 2^{-15} y γ = $5{,}4e^{-1}$. Para estos parámetros la exactitud de la clasificación de VFs de clase I y clase II fue del 99.0 % (Tabla 1).

Para obtener una segunda evaluación de la métrica propuesta clasificamos el conjunto de VFs utilizando k-NN como método de clasificación, para 1, 3 y 7 NNs. La exactitud de la clasificación fue del 98.5 %, 98.5 % y 97.5 % respectivamente (Tabla 1). Estos resultados fueron satisfactorios para los cuatro clasificadores probados.

4.2.2 Clasificación de VFs entrenando con una clase positiva

Los clasificadores de tipo SVM clásicos están basados en el entrenamiento a partir de muestras de dos clases (por ejemplo positivas y negativas). Sin embargo, en diversas situaciones se dispone solamente de muestras positivas para el entrenamiento. Este es el caso del problema planteado originalmente por este trabajo, donde se dispone de un set de entrenamiento de proteínas que se conoce que son fusogénicas, y se pretende seleccionar proteínas candidatas a fusógenos de otro set de proteínas muy diverso. Dada la variabilidad del set de proteínas candidatas, resulta complejo identificar muestras negativas representativas. Schölkopf et al. describen el método "one-class classification" que permite entrenar el modelo con solamente muestras positivas.

Para evaluar el desempeño de un OC-SVM con este tipo de datos se entrenó un clasificador de tipo One-class SVM con VFs clase I como muestras positivas. Se evaluó la clasificación de un conjunto de VFs de clase I (distinto al conjunto de entrenamiento) y un conjunto de VFs de clase II.

Para esta etapa también se trabajó con el paquete LIBSVM que dispone de esta metodología. Se aplicó el mismo kernel a los datos y se seleccionó la mejor combinación de parámetros. En este caso el parámetro ν sustituye al parámetro C, el cual controla la influencia de cada vector de soporte. El significado del parámetro ν es análogo al significado de C, pero solamente puede tomar valores entre 0 y 1. Análogamente a la parte anterior la mejor combinación de ν y γ se seleccionó a partir de una búsqueda "grid search" utilizando "10-fold cross-validation" con valores de ν entre 0,05 y 1, y valores de γ entre $5e^{-4}$ y $5e^{2}$ en intervalos uniformes (Fig. 23). En este caso la búsqueda de parámetros no devuelve una región "plana" donde el desempeño no varía significativamente con los parámetros seleccionados. Se seleccionó la combinación de parámetros para el cual la exactitud es máxima.

Se seleccionaron como parámetros óptimos ν = 0,05 y γ = $5{,}4^{-1}$. Para estos parámetros la exactitud de la clasificación de VFs de clase I y clase II fue del 92.0 % (Tabla 1). Aunque el desempeño disminuyó respecto de la clasificación en dos clases esta se puede seguir considerando satisfactoria. Dado que este es un trabajo exploratorio, obtener proteínas candidatas con más de un 90% de exactitud es significativo.

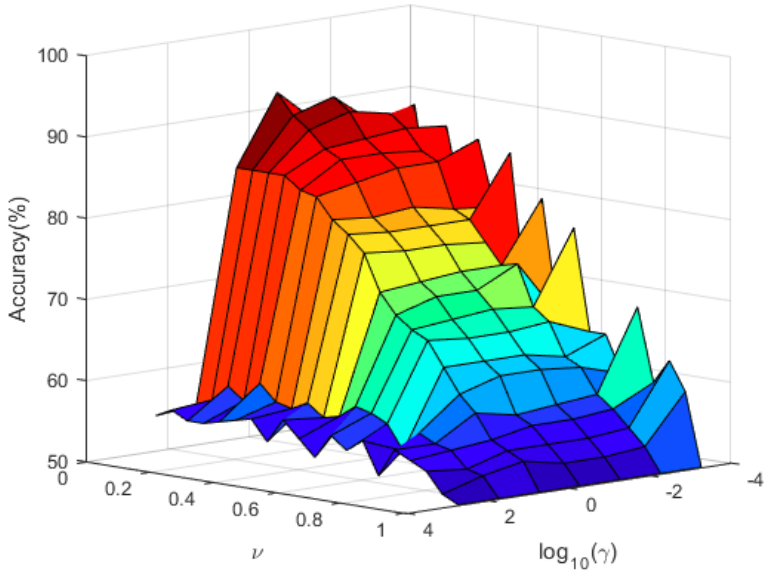


Figura 23. Grid search para la selección de parámetros de OCSVM para la clasificación de VFs entrenando con una sola clase.

4.2.3 Clasificación de VFPs entrenando con dos clases

A fin de evaluar el funcionamiento del algoritmo adaptado para alineamientos locales, evaluamos el clasificador SVM de dos clases entrenado anteriormente, esta vez clasificando un conjunto de VFPs. Se obtuvo una exactitud del 90.5 %. La clasificación utilizando k-NN resultó en una exactitud del 98.5%, 97.5% y 95.5% para 1, 3 y 7 k-NN (Tabla 1). Aunque la exactitud disminuyó sensiblemente respecto de la clasificación de VFs, los resultados siguen siendo satisfactorios.

4.2.4 Clasificación de VFPs entrenando con una clase positiva

La clasificación de VFPs de clase I y II entrenando solamente con VFs de clase I reduce su desempeño de forma considerable, obteniendo una exactitud del 69.5 %.

Tabla 1. Exactitud de clasificación para cada clasificador de dos clases al clasificar VFs o VFPs.

Clasificación	1-NN	3-NN	7-NN	SVM	OC-SVM
VFs	98,5	98,5	97,5	99	92
VFPs	98,5	97,5	95,5	90,5	69,5

4.2.5 Discusión

La métrica desarrollada a partir de la propuesta por Przytycka et al. permitió clasificar de forma satisfactoria VFs a partir de los tres métodos propuestos (k-NN, SVM y OC-SVM). La clasificación de VFPs utilizando k-NN presentó una exactitud similar a la obtenida para la clasificación de VFs. También obtuvimos una exactitud aceptable clasificando VFPs con un modelo basado en SVM. A pesar de esto, el desempeño se reduce considerablemente al clasificar los VFPs con un modelo basado en OC-SVM, por lo que sería de utilidad seguir trabajando sobre la métrica para mejorar la clasificación de proteínas para las cuales únicamente una región de su secuencia corresponde a la región fusogénica. El objetivo a futuro es poder obtener scores idénticos al enfrentar una VF y su correspondiente VFP al resto de las VFs. Este trabajo se basó en un conjunto reducido seleccionado al azar del conjunto original de secuencias de VFs y VFPs extraídas de UniProt dado que los cálculos requieren mucho tiempo. Además, este conjunto original había descartado un subconjunto importante de secuencias por no contener anotaciones de la posición del clivaje de la VFP.

Por otra parte proponemos trabajar en detalle sobre la influencia de la penalización del gap sobre la métrica. Evaluaciones no presentadas en este trabajo demostraron que el valor del gap no influye en el desempeño de los clasificadores de tipo k-NN, pero sí influye en los clasificadores de tipo SVM y OC-SVM. En este trabajo utilizamos un sistema de gaps constante, donde el valor de la penalización siempre es el mismo. Sería interesante evaluar cómo se comporta la métrica trabajando con un sistema de gaps lineal (dependiente de la longitud del gap) o de tipo afín, donde se penaliza de forma distinta la apertura y la elongación del gap.

A pesar de que se podría mejorar el desempeño de los clasificadores refinando la métrica, los resultados fueron satisfactorios. En particular, se podría evaluar el funcionamiento un sistema de gaps, sistema de scores o normalización alternativos. Dado el tiempo de cálculo y la necesidad de ajustarnos al calendario del proyecto no se trabajó en estas alternativas.

El método llevado a cabo en este punto 4 podría permitir la identificación de fusógenos virales hasta ahora desconocidos para algunos virus envueltos a partir de su genoma o proteoma, así como también la identificación de proteínas con capacidad fusogénica en diversos modelos biológicos. En este último punto se focalizó el segundo objetivo de este trabajo.

Este trabajo fue publicado en Megrian et al. 2017 (Megrian, Aguilar, & Lecumberry, 2017).

OBJETIVO 2

5 Búsqueda de proteínas con capacidad fusogénica en VEs.

Desde un punto de vista topológico-celular, la fusión entre VEs y la célula blanco es equivalente a la fusión entre células, o entre una célula y un virus con cubierta membranosa. Sin embargo los mecanismos por los cuales fusionan permanecen inciertos. Las proteínas de fusión de membranas extracelulares conocidas comparten similitud estructural con proteínas de fusión viral (sincitinas con fusógenos de retrovirus, EFF-1, AFF-1 y HAP2 con fusógenos virales de clase II). Por esta razón, el segundo objetivo de este trabajo se basó en la utilización de la métrica desarrollada en el objetivo anterior para identificar proteínas presentes en VEs con capacidad fusogénica utilizando como conjunto de entrenamiento proteínas de fusión viral de clase I, II y III. En la Fig. A2 se esquematiza todo el procedimiento así como los resultados de este objetivo.

Dado que únicamente se conoce la clase positiva (las proteínas de fusión viral) y la clase negativa abarca todo lo que no sea clase positiva, se utilizaron clasificadores de tipo OCC. Dado que las proteínas sincitinas, EFF-1, AFF-1 y HAP2 son las únicas conocidas con capacidad fusogénica además de los fusógenos virales se decidió utilizar estas como control positivo. El objetivo es identificar proteínas de VEs que hayan divergido a partir de algún fusógeno viral de la misma forma que lo hicieron los controles positivos.

5.1 Obtención de secuencias de aminoácidos

- Proteínas de VEs – set de evaluación

La base de datos de proteínas de VEs fue generada a partir de la base de datos pública de proteínas y ARNm de Vesiclepedia (Kalra et al., 2012) versión 3 correspondiente al 9 de enero de 2015. La misma contiene 17.922 proteínas identificadas por su GeneID correspondiente a la base de datos de genes del NCBI. Utilizando la tabla gene2accession del NCBI y búsqueda manual se mapearon estos GeneID a 22.200 identificadores de proteínas de la base de datos pública UniProt. Estas 22.200 proteínas se clusterizaron al 98% de identidad utilizando la herramienta CD-HIT, conservando solamente una proteína representante de cada clase con el objetivo de evitar redundancia de secuencias en la base de datos. Se obtuvieron 14528 proteínas de VEs, sobre las cuales se trabajó de aquí en adelante.

- Proteínas control positivo

Dado que las proteínas sincitinas, EFF/AFF y HAP2 son las únicas proteínas conocidas con capacidad fusogénica célula-célula y su similitud con fusógenos virales, estas se utilizaron como control positivo de la clasificación. Se descargaron de la base de datos pública InterPro todas aquellas proteínas identificadas como pertenecientes a las familias TLV/ENV coat polyprotein (id IPR018154) y Cell-cell fusogen EFF/AFF (id IPR029213). La primera familia agrupa proteínas que presentan el dominio ENV de retrovirus y comprende 4464 proteínas, de las cuales 899 no corresponden a virus sino a secuencias endógenas retrovirales en diversos organismos, principalmente vertebrados. Dentro de las secuencias endógenas retrovirales de humanos y primates aparece la proteína sincitina-1, codificada por el gen ERVW-1 (endogenous retrovirus group W envelope member 1). Esta proteína pertenece a la clase descripta en la sección 1.3.2.2.1. Estas 899 proteínas se clusterizaron al 98% de identidad con el programa CD-HIT,

obteniendo 139 secuencias representativas. La segunda familia comprende a las proteínas FF descriptas en la sección 1.3.2.2.2. Esta familia de InterPro agrupa 212 proteínas, de las cuales el 98% corresponden a nematodos. Éstas también se clusterizaron de acuerdo a su porcentaje de identidad, obteniendo 28 secuencias representativas. También se descargaron las 417 proteínas que contienen el dominio Generative cell specific-1, HAP2-GCS1 (id IPR018928), el cual fue descripto en la sección 1.3.2.2.3. Se obtuvieron 101 secuencias representativas según su porcentaje de identidad.

- Proteínas virales – set de entrenamiento

Se trabajó de forma similar que en la sección anterior, pero para aquellos fusógenos para los que no se conocía dónde se ubicaba el dominio transmembrana se realizaron predicciones con el software Philius (Reynolds, Kall, Riffle, Bilmes, & Noble, 2008) para recortar la región fusogénica. En todos los casos el fusógeno se encontró en la posición N terminal respecto del único (o primer) dominio transmembrana de la proteína. Se clusterizaron las proteínas al 95% de identidad, obteniéndose en total 528 proteínas. Éstas están repartidas en 384 fusógenos de clase I (26 de la familia Arenaviridae, 33 de Coronaviridae, 16 de Filoviridae, 170 de Orthomyxoviridae, 64 de Paramyxoviridae y de 75 Retroviridae), 99 fusógenos de clase II (9 de la familia Bunyaviridae, 51 de Flaviviridae y 39 de Togaviridae) y 45 fusógenos de clase III correspondientes a la familia Herpesviridae.

De aquí en más nos referiremos a todas las proteínas de acuerdo a su identificador de UniProt.

5.2 Tratamiento de datos de vesículas

5.2.1 MSA, predicción de estructura secundaria y cálculo de similitudes

Se utilizó el paquete HHsuite para realizar la predicción de la estructura secundaria de cada proteína. Para cada una de las proteínas de VEs, proteínas del control positivo y fusógenos virales se realizó un alineamiento múltiple con la herramienta HHblits del paquete HHsuite. Para esto se utilizó la base de datos uniprot20 provista por HHsuite, la cual es generada a partir del clustering de las bases de datos UniProt y nr del NCBI en grupos de secuencias que se alinean al menos en un 80% de su longitud y comparten un 20% de identidad. De esta forma se generaron MSA precisos y diversos. Luego se corrió el programa Psipred sobre cada MSA con el objetivo de obtener una predicción de la estructura secundaria más precisa. Se realizó un análisis de la distribución de la estructura secundaria de cada familia, género y subfamilia estudiada (Fig. 24). Los fusógenos virales de clase I presentan las mayores proporciones de hélice alfa, variando entre 34% para la subfamilia Pneumovirinae y 69% para la subfamilia Orthoretrovirinae. El porcentaje de hoja beta varía entre el 3% y el 26%. Las longitudes también varían significativamente, entre 137 y 544 aminoácidos. Los fusógenos virales de clase II presentan características más homogéneas, presentando un porcentaje de hélice alfa de entre 2% y 9%, en contraste con un 28% a 49% de hoja beta. Todos los fusógenos virales de clase III analizados en este trabajo corresponden a virus de la familia Herpesviridae. Estos son significativamente más extensos que los fusógenos de clase I y II, y no presentan porcentajes de elementos de estructura secundaria tan extremos.

Cada una de las predicciones de estructura secundaria fue colapsada según se describe en la sección 2.1.4.2.1. Finalmente se calculó la similitud de cada par de proteínas con la métrica desarrollada en la sección anterior.

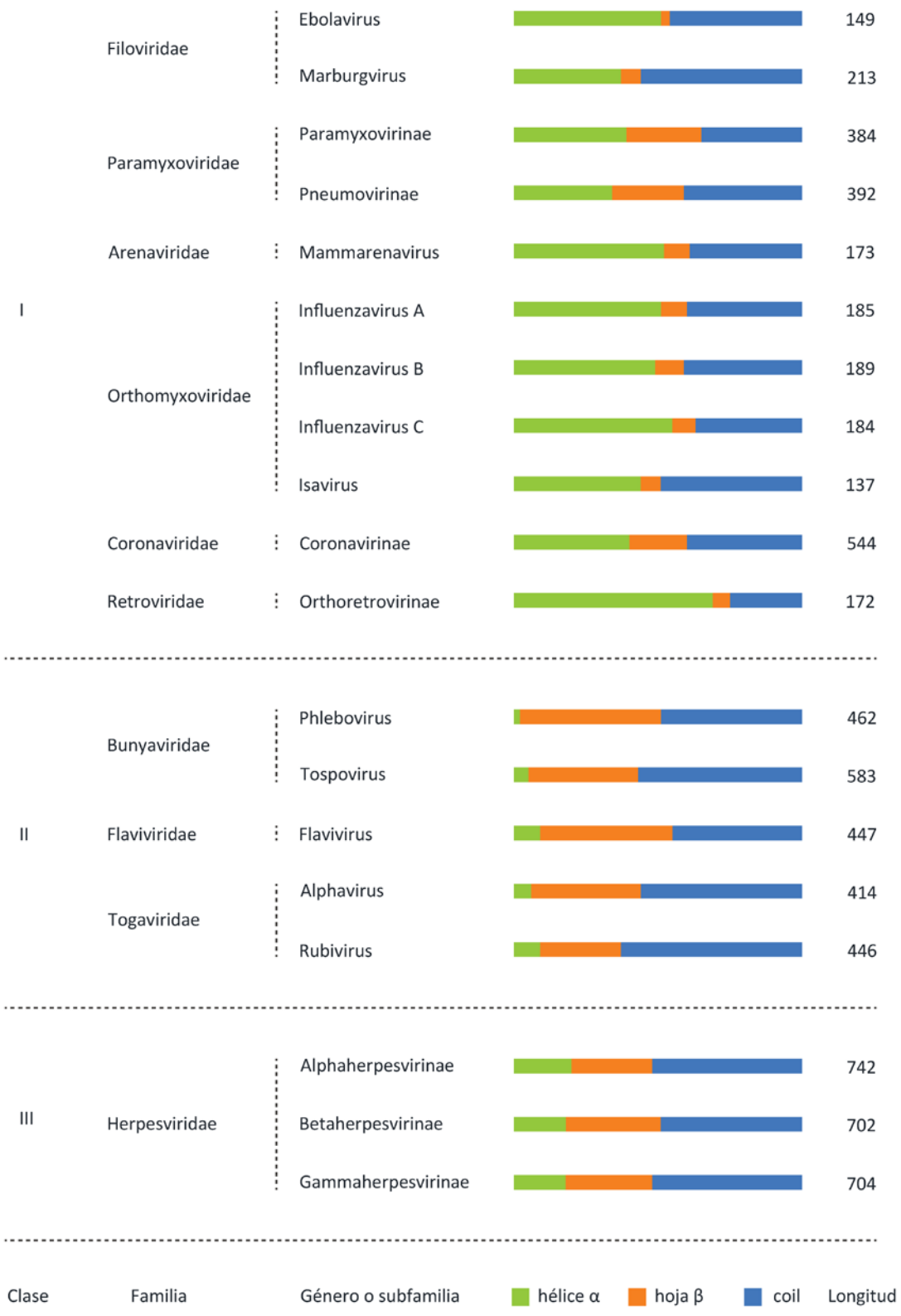


Figura 24. Composición de estructura secundaria y longitud del ectodominio de fusógenos virales de clase I, II y III representados como promedio por género o subfamilia viral. Las diferencias tanto en composición como longitud dentro de cada género o subfamilia son mínimas.

5.3 Armado de matriz de distancia

A partir de cada par de similitudes se generó una matriz de distancias, y con esta se trabajó de aquí en adelante. Primero obteniendo vectores de características para cada proteína, y luego utilizando estos vectores de características para entrenar modelos con fusógenos virales, evaluar su desempeño y finalmente clasificar proteínas de VEs como potenciales fusógenos.

Como primer paso, dado que la métrica devuelve un valor de similitud que varía entre 0 y 1 para cada par de proteínas, se convirtió este valor a una distancia restándoselo a 1.

Con los valores de distancia para cada par de proteínas se armó una matriz de distancias simétrica de 16732 por 16732, cuya diagonal corresponde a la identidad por lo que vale 0. En la sección anterior la clasificación de tipo OC se limitó a vecinos más cercanos y SVM ya que estos aceptan como input una matriz de distancias. En esta etapa se quiso evaluar una mayor cantidad de clasificadores, por lo cual se utilizó escalamiento multidimensional (MDS) (Kruskal, 1964) para transformar la matriz de distancias en vectores de características. El método MDS toma una matriz de disimilitud y devuelve un set de coordenadas de forma que las distancias entre estos son aproximadamente iguales a las disimilitudes. Un set de distancias de n puntos puede ser representado en un máximo de n-1 dimensiones. Se obtuvieron los 16731 valores propios y se graficaron (Fig. 25) de forma de seleccionar una cantidad de dimensiones manejable con las que trabajar que representen correctamente los puntos. Se seleccionaron los 20 primeros vectores, por lo que de aquí en más se trabajó con una matriz de 16732 puntos descriptos por 20 vectores.

En la Fig. 26 se grafican las tres primeras coordenadas para los fusógenos de las 10 familias virales y 500 proteínas de VEs seleccionadas al azar. Esto permitió verificar visualmente que es posible separar espacialmente los fusógenos virales de acuerdo a su clase. La densidad representada por los fusógenos virales de clase III (marrón) se encuentra más cercana a las densidades representadas por los fusógenos virales de clase II. Dada la necesidad de agrupar sus representantes a otra clase por la poca cantidad de elementos representativos, se decidió trabajar por un lado con clasificadores entrenados con fusógenos virales de clase I, y por otro con clase II y clase III.

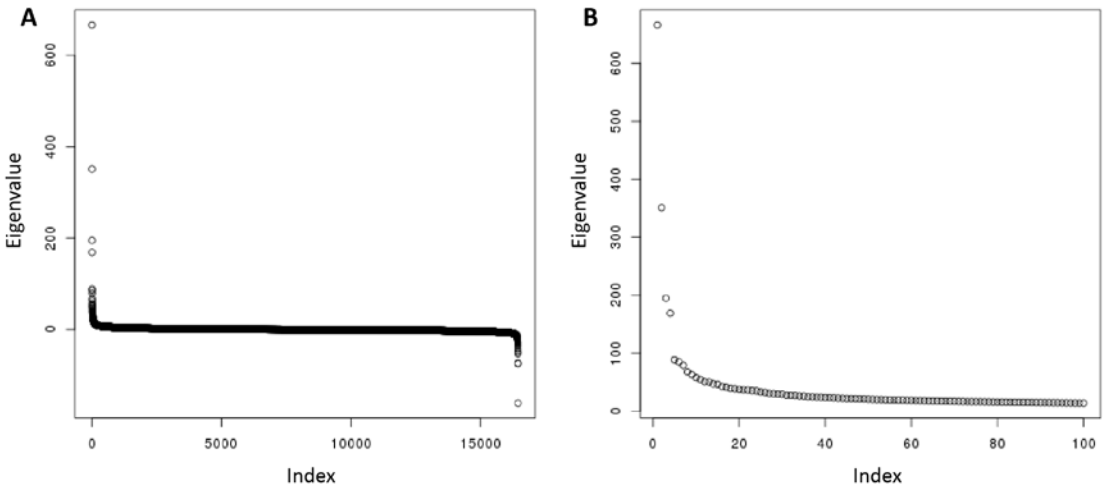


Figura 25. A. Eigenvalues para las n-1 dimensiones. B. Eigenvalues para las primeras 100 dimensiones.

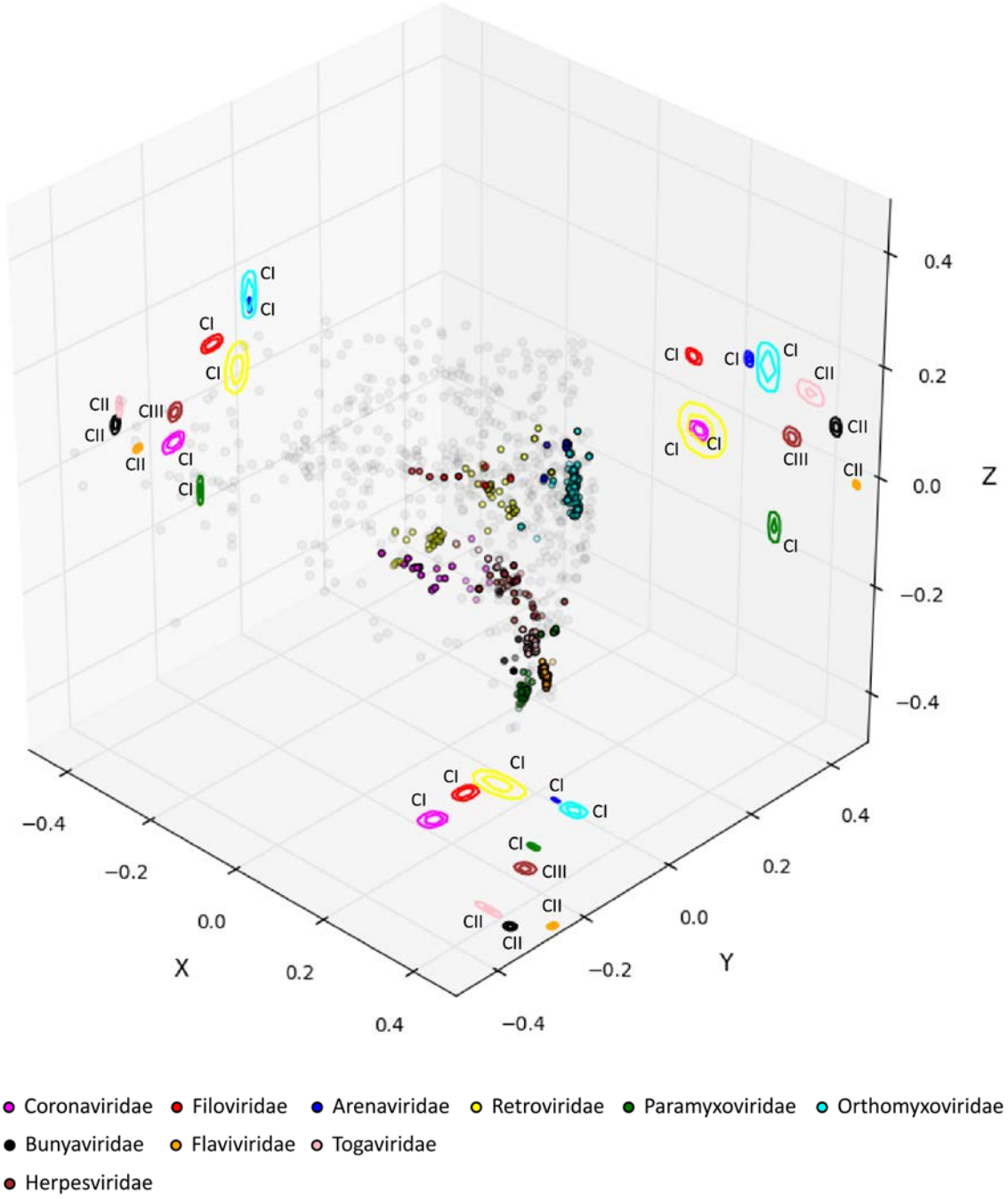


Figura 26. Primeras tres dimensiones del embedding obtenido con MDS. Los puntos de colores representan cada fusógeno viral. Cada color corresponde a una familia. Coronoviridae, Filoviridae, Arenaviridae, Retroviridae, Paramyxoviridae y Orthomyxoviridae corresponden a clase I, Bunyaviridae, Flaviviridae y Togaviridae a clase II, y Herpesviridae a clase III. Los puntos sin colorear corresponden a 500 proteínas de VEs seleccionadas al azar. En los planos XY, YZ y ZX se representaron las densidades de cada familia. Las densidades de clase II (negro, naranja y rosado) se encuentran cercanas entre sí y separadas de clase I (fucsia, rojo, azul, amarillo, verde y celeste). La densidad de clase III (marrón) se encuentra más cercana a las densidades de clase II.

5.4 Aplicación de métodos de clasificación

Para clasificar las proteínas de VEs como posibles fusógenos se evaluaron los nueve métodos incluidos en el paquete DDtools de Matlab descriptos en la sección 3.5. Estos fueron Gaussian model, Estimador de Parzen, Vecinos más cercanos, Support vector data description, k-means, Minimum spanning tree, Gaussian mixture model, Autoencoder neural networks y Linear programming data description.

Se evaluaron todos los métodos en paralelo para clasificar proteínas en por un lado fusógenos virales de clase I, y por otro fusógenos virales de clase II y clase III. Dado el número muy reducido de elementos para clase III y su cercanía a clase II se decidió agrupar estas dos clases en la clasificación. Para los clasificadores de proteínas similares a fusógenos de clase I se utilizó como set de entrenamiento el conjunto de fusógenos virales de clase I. En este caso se le denomina clase positiva a los fusógenos virales de clase I. Los fusógenos virales de clase II o clase III forman parte de la clase negativa. Para los clasificadores de proteínas similares a fusógenos de clase II o clase III se utilizó como set de entrenamiento el conjunto de fusógenos virales de clase II y clase III. Recíprocamente, en este caso se le denomina clase positiva a los fusógenos virales de clase II y clase III, y forman parte de la clase negativa a los fusógenos virales de clase I.

Para cada modelo se realizó una validación cruzada con 10 iteraciones.

Se realizó un grid-search para cada clasificador utilizando como fracción de error los valores $1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$ y 1. El segundo parámetro evaluado para cada modelo se detalla en la Tabla 2. Para elegir la mejor combinación de parámetros para cada modelo se tuvieron en cuenta cuatro condiciones:

1. Al realizar la validación cruzada, la exactitud de la clasificación para la clase positiva sea mayor al 95%.
2. El modelo generado no clasifica como positivo ningún fusógeno perteneciente a la clase negativa.
3. El modelo clasifica como clase positiva menos del 10% de las proteínas de VEs.
4. El modelo generado clasifica como positiva alguna proteína del control positivo (sincitinas para clase I, EFF/AFF y HAP2 para clase II).

En caso que varias combinaciones de parámetros cumplieran con las cuatro condiciones, se seleccionó aquella que maximice el porcentaje del punto cuatro.

Tabla 2. Parámetros explorados para cada clasificador.

Modelo	Parámetro	Definición de parámetro	Valores evaluados
Gaussian model	R	Parámetro de regularización	$1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$, 1
Estimador de Parzen	H	Ancho de ventana	$1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$, 1, $1e^{1}$, $1e^{2}$, $1e^{3}$, $1e^{4}$, $1e^{5}$
Vecinos más cercanos	K	Número de vecinos	2^{2}, 2^{3}, 2^{4}, 2^{5}, 2^{6}
Support vector data description	SIGMA	Ancho del kernel de función de base radial	$1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$, 1, $1e^{1}$, $1e^{2}$, $1e^{3}$, $1e^{4}$, $1e^{5}$
k-means	K	Número de clusters	2^{2}, 2^{3}, 2^{4}, 2^{5}, 2^{6}

Minimum spanning tree	N	Número de aristas de longitud máxima	0
Gaussian mixture model	N	Número de clusters	1, 2, 3, 4, 5, 6, 7
Autoencoder neural networks	N	Número de unidades ocultas	1, 5, 10
Linear programming data description	S	Parámetro de escala para sigmoide	$1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$, 1, $1e^{1}$, $1e^{2}$, $1e^{3}$, $1e^{4}$, $1e^{5}$

Dentro de los clasificadores de fusógenos virales de clase I se obtuvieron siete modelos que cumplen con las cuatro condiciones para al menos una combinación de parámetros, mientras que para los clasificadores de fusógenos virales de clase II y clase III fueron ocho. Estos se detallan en las Tablas 3 y 4. No se consiguió una combinación de parámetros del clasificador Linear programming data description de clase I para el cual se cumplan las cuatro condiciones anteriores.

Tabla 3. Clasificadores de fusógenos de clase I seleccionados.*

Método	FRACREJ	Parámetro 2	% Clase II clasificado como Clase I	% Sincitinas clasificado como Clase I	% VEs clasificado como Clase I	% Clase I clasificado como Clase I (CV)
autoenc	$1e^{-5}$	$1e^{-1}$	0,0	8,6	1,9	99,5
gauss	$1e^{-2}$	$1e^{-5}$	0,0	5,0	0,2	97,7
kmeans	$1e^{-2}$	$3e^{1}$	0,0	9,4	0,8	98,7
knn	$1e^{-2}$	4	0,0	5,0	0,3	99,7
mog	$1e^{-5}$	3	0,0	7,9	0,2	98,4
mst	$1e^{-2}$	0	0,0	21,6	2,1	99,7
parzen	$1e^{-4}$	$1e^{-1}$	0,0	9,4	1,4	99,2

Tabla 4. Clasificadores de fusógenos de clase II y III seleccionados.*

Método	FRACREJ	Parámetro 2	% Clase I clasificado como Clase II	% EFF/AFF/HAP2 clasificado como Clase II	% VEs clasificado como Clase II	% Clase II clasificado como Clase II (CV)
autoenc	$1e^{-4}$	5	0,0	26,4	1,9	98,6
gauss	$1e^{-4}$	$1e^{-1}$	0,0	14,0	1,1	99,3
kmeans	$1e^{-4}$	$2e^{1}$	0,0	23,3	2,6	97,2
knn	$1e^{-4}$	8	0,0	20,2	2,4	98,6
lp	$1e^{-4}$	$1e^{1}$	0,0	44,2	6,3	97,9
mog	$1e^{-4}$	1	0,0	5,4	0,3	97,2
mst	$1e^{-4}$	0	0,0	40,3	5,8	100,0
parzen	$1e^{-4}$	$1e^{-1}$	0,0	35,7	4,5	99,3

* autoenc: Autoencoder neural networks, gauss: Gaussian model, kmeans; k-means, knn: Vecinos más cercanos, lp: Linear programming data description, mog: Gaussian mixture model, mst: Minimum spanning tree, parzen: Estimador de Parzen.

5.5 Obtención y análisis de listas de proteínas candidatas

Se seleccionaron como potenciales proteínas candidatas a proteínas de fusión aquellas proteínas de VEs que hayan sido clasificadas como positivas para cualquiera de los 15 modelos. De esta forma de obtuvieron 1920 proteínas candidatas, 721 con potencial de similitud a fusógenos virales de clase I, y 1199 a clase II o clase III (Tabla 5). Algunas de estas proteínas se detallan en las Tablas 7 y 8 (ver punto 5.5.1).

Tabla 5. Clasificación de proteínas de VEs.

	Clasificados negativos	**Clasificados positivos**	**Clasificados positivos por cantidad de clasificadores**							
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Clase I	13797	721	488	128	56	31	11	5	2	-
Clase II	13319	1199	374	255	180	138	115	78	37	22

Con el objetivo de evaluar a grandes rasgos el desempeño de los clasificadores se exploró la distribución de las proporciones de hélice α, hoja β y coil en las proteínas clasificadas como positivas y comparar estos resultados con el set de proteínas de VEs inicial. A partir de la Fig. 27 se pudo verificar que los clasificadores de clase I seleccionaron un set de proteínas de VEs candidatas enriquecido en hélice α y una menor proporción de hoja β comparado con el set inicial de proteínas de VEs. Esto se condice con lo explorado en la Fig. 24. También se verificó que los clasificadores de clase II/III seleccionaron un set de proteínas de VEs candidatas enriquecido en hoja β y una menor proporción de hélice α, similar a lo conocido para fusógenos virales de clase II/III.

No hay relación entre la cantidad de proteínas clasificadas por cada clasificador para cada clase, de todas formas los clasificadores basados en Autoencoder neural networks, Minimum spanning tree y Estimador de Parzen son los que más proteínas de VEs clasificaron como positivas para ambas clases. No entra en comparación el clasificador Linear programming data description, el cual se utilizó solamente para clasificar clase II y III (Fig. 28).

En resumen, entre los clasificadores de clase I y clase II/III se clasificaron como proteínas candidatas a fusógenos un 13% del total de las proteínas de VEs. Aunque la mayoría de la proteínas de VEs candidatas a clase I fueron clasificadas por un único clasificador, casi la mitad de las clasificadas a clase II/III fueron clasificadas por al menos tres clasificadores. Se pudo verificar que las proteínas candidatas se encuentran enriquecidas en elementos de estructura secundaria de manera similar a los fusógenos con los que se entrenaron los modelos.

Dado que es necesario refinar y reducir el set de proteínas candidatas para considerarlos candidatos reales a una validación biológica, la etapa siguiente consistió en filtrar las proteínas candidatas de acuerdo a características conocidas para proteínas de fusión viral.

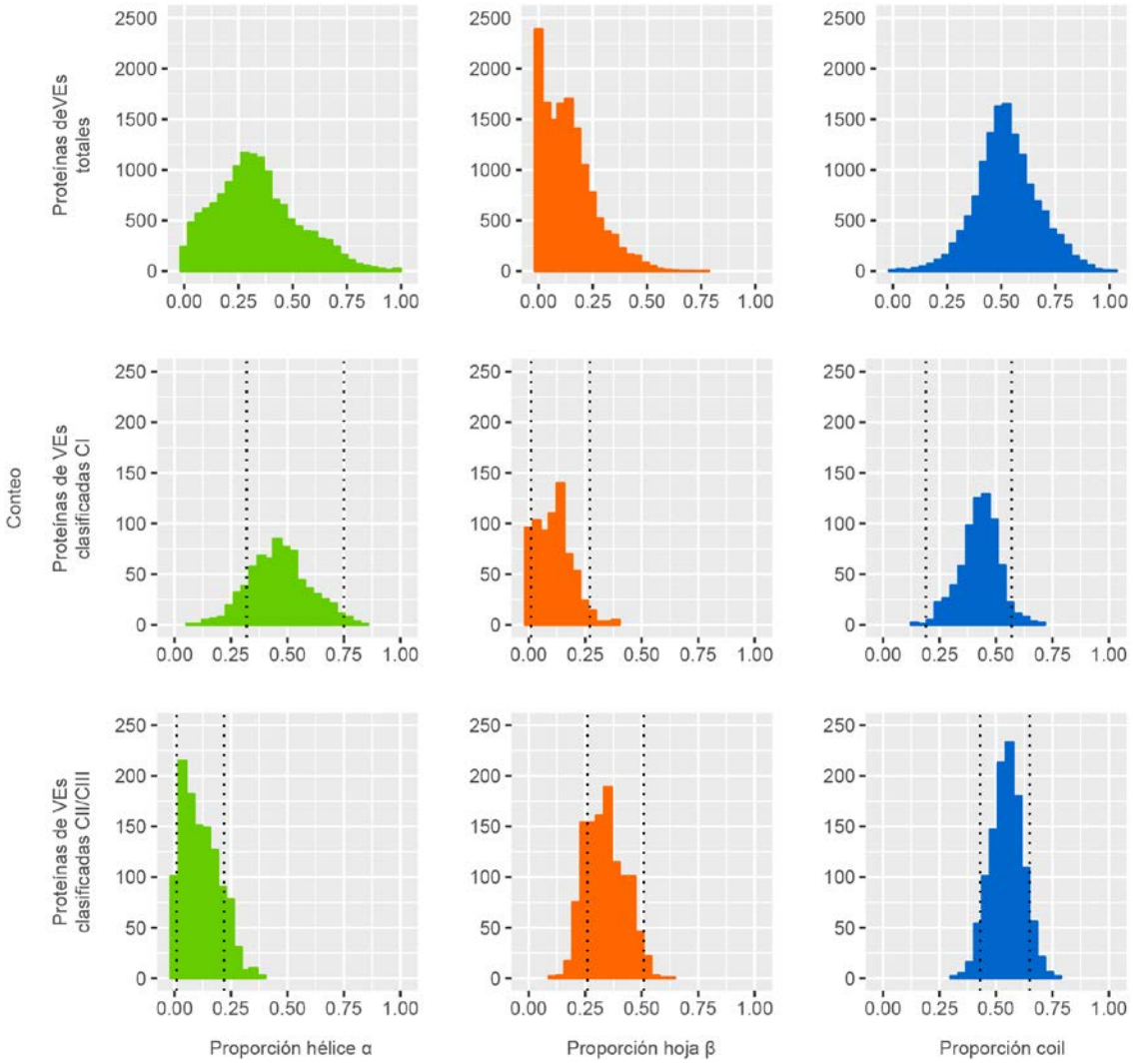


Figura 27. Proporción de elementos de estructura secundaria en distintos sets de proteínas de VEs. Fila superior: todas las proteínas de VEs. Fila intermedia: proteínas de VEs clasificadas como positivas para algún clasificador de fusógenos de clase I. Las líneas punteadas representan la proporción menor y mayor de cada estructura según el análisis de la Fig. 24 para los fusógenos virales de clase I. Fila inferior: proteínas de VEs clasificadas como positivas para algún clasificador de fusógenos de clase II/III. Las líneas punteadas representan la proporción menor y mayor de cada estructura según el análisis de la Fig. 24 para los fusógenos virales de clase II/III.

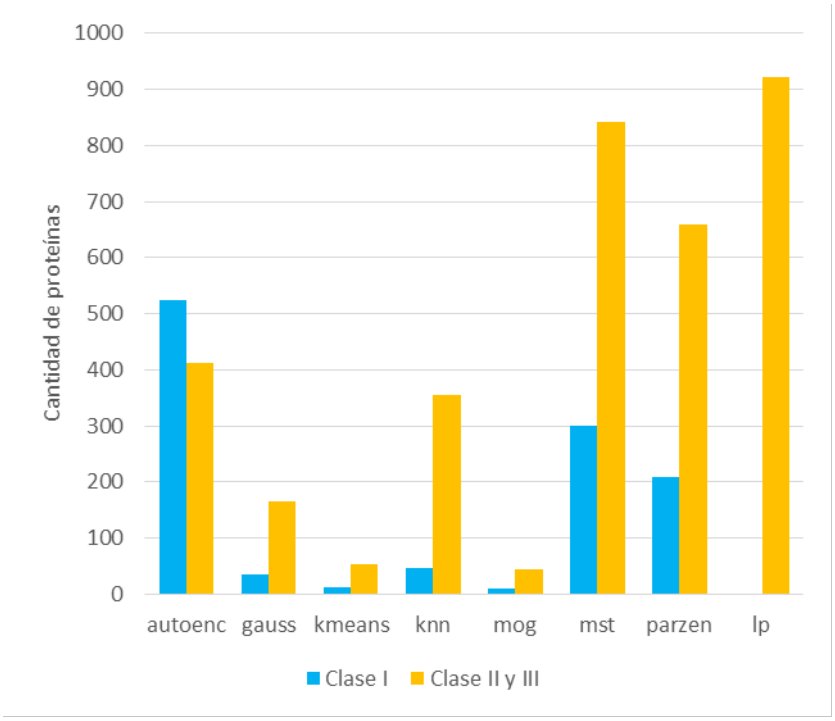


Figura 28. Cantidad de proteínas de VEs clasificadas por cada clasificador de clase I y clase II y III.

5.5.1 Predicción de topología

De acuerdo a la observación sobre el set de fusógenos virales de la sección 5.1 donde se verificó que la región extracelular fusogénica de todas las proteínas de fusión viral se encuentra en la región N terminal respecto del único o primer dominio transmembrana de la proteína, y las longitudes que presentan los fusógenos virales conocidos detallados en la Fig. 24, se filtraron las proteínas clasificadas como positivas de acuerdo a estas características a priori imprescindibles para una proteína de fusión viral.

5.5.1.1 Predicción de dominios transmembrana

Utilizando el servidor online TOPCONS (Bernsel, Viklund, Hennerdal, & Elofsson, 2009) se predijo la topología de cada una de las 1920 potenciales proteínas candidatas. A partir de la posición de los dominios transmembrana predichos se determinaron las longitudes de las regiones extracelulares o intracelulares de las proteínas. No se hizo distinción entre estas regiones ya que las predicciones pueden malinterpretar la posición de estas respecto a la membrana a la que están ancladas. Para clase I un 23% de las proteínas de VEs clasificadas positivas presentan dominios transmembrana, por lo que potencialmente podrían cumplir las características de un fusógeno viral. Lo mismo sucede para un 31% de las proteínas de VEs clasificadas como clase II/III. Si tenemos en cuenta la composición total del set de proteínas de VEs, un 20% de ellas son proteínas transmembrana, por lo que los clasificadores de clase II y III seleccionaron de forma preferencial proteínas transmembrana. Se estudiaron también las proteínas para las que no se identificó un dominio transmembrana, pero sí un péptido señal ya que a veces los predictores no pueden distinguir los mismos. También se especificaron las proteínas con al menos un

dominio transmembrana que presentan un péptido señal, ya que el péptido señal indica la posición de la proteína en una membrana (Tabla 6).

Tabla 6. Predicción de dominios transmembrana (DTM) y péptido señal (SP) en las proteínas clasificadas como positivas para los clasificadores de clase I y clase II y III, y para el total de las proteínas de VEs.

Predicción TOPCONS	Clasificadas		Total
	CI	CII + CIII	
No DTM	557 (77%)	822 (69%)	11322 (80%)
No DTM + SP	82	295	1755
1 DTM	51 (7%)	346 (29%)	1503 (11%)
2 DTM	18 (3%)	19 (1%)	274 (2%)
3 DTM	95 (13%)	12 (1%)	1074 (7%)
DTM + SP	34	284	996

5.5.1.2 Evaluación de longitud y posición de las proteínas candidatas

Teniendo en cuenta las longitudes de los fusógenos virales de cada clase como se muestra en la Fig. 24 se determinó cuál es la longitud mínima que tiene un fusógeno de cada clase. Para darle flexibilidad a las búsquedas se redujo el valor de longitud mínima un 20%. Entonces, para clase I se estableció una longitud mínima de 120 aminoácidos y para clase II y III de 300. De esta forma se seleccionaron aquellas potenciales proteínas candidatas de VEs que contengan al menos una región extracelular N-terminal o intracelular (pero en posición N-terminal de un dominio transmembrana) de longitud mínima según se explicó. No se seleccionaron aquellas proteínas cuyo única región extra o intracelular de longitud mínima se ubicara en el extremo C terminal, ya que no se conocen fusógenos que presenten esa característica. De esta forma se obtuvieron 48 proteínas similares a fusógenos virales de clase I y 214 similares a clase II o clase III, que se representan en las Fig. 29 y 30. Estas se recortaron, obteniendo únicamente la región candidata, sin péptido señal ni dominios transmembrana. En prácticamente todos los casos estas regiones coincidieron con el dominio extracelular de la proteína. Con estas regiones se trabajó de aquí en adelante.

Hay dos proteínas que a pesar de no pasar este filtro son clasificadas como positivas para los siete clasificadores de clase I por lo que resulta interesante caracterizarlas. Estas son la Q5U305 y la Q96FX8. La primera es la proteína ER lumen protein-retaining receptor 2 de rata, codificada por el gen Kdelr2. Esta proteína es requerida para la retención de proteínas luminales del retículo endoplasmático. También es requerida para el tráfico normal a través del aparato de Golgi (Lewis & Pelham, 1992). Tiene 212 aminoácidos de longitud y siete dominios transmembrana. Los dominios citoplasmáticos o luminales varían entre seis y 15 aminoácidos. La segunda proteína es la denominada p53 apoptosis effector related to PMP-22 de humano, codificada por el gen PERP. Es un componente intercelular de las uniones de tipo desmosoma que cumple una función en la integridad de epitelios estratificados y en la adhesión celular promoviendo el ensamblaje de desmosomas (Ihrie et al., 2005). También participa como efectora en la vía apoptótica dependiente de TP53. Tiene 193 aminoácidos y cuatro dominios transmembrana. Los dominios citoplasmáticos o luminales varían entre 11 y 47 aminoácidos.

5.5.2 Análisis del conjunto de proteínas resultantes

De acuerdo a la información depositada en la base de datos UniProt sobre su función, se clasificaron las 262 proteínas resultantes en tres grupos:

1. Aquellas para las que existe evidencia experimental de su función molecular.
2. Aquellas para las que existe una predicción directa de su función molecular, o que la evidencia experimental no determine la función molecular.
3. Aquellas para las que la predicción de su función no es determinante o directamente no existe.

Así para las proteínas similares a fusógenos virales de clase I se obtuvieron 16 para la primera categoría, ocho para la segunda y 24 para la tercera. Mientras que para las similares a clase II/III se obtuvieron 78 para la primera categoría, 45 para la segunda, y 91 para la tercera.

Para aquellas proteínas para las que se conoce una función molecular se realizó un análisis de Gene Ontology (H. Mi et al., 2010) (Fig. 29). Dado que la estructura y la función de una proteína están relacionadas, el objetivo de este análisis fue explorar las características funcionales de las proteínas seleccionadas como positivas de acuerdo a su estructura secundaria y su topología.

Las proteínas clasificadas como positivas por los clasificadores de clase I son principalmente proteínas de unión a ácidos nucleicos, helicasas, ligasas y oxidorreductasas. Las proteínas de unión a ácidos nucleicos presentan varios dominios conservados, como dedos de zinc, hélice-giro-hélice, hélice-loop-hélice o cierre de leucina. En general una misma proteína puede presentar múltiples repetidos del mismo dominio o puede contener distintos dominios de unión a lo largo de su secuencia. Las proteínas con actividad helicasa o ligasa también presentan este tipo de dominios ya que se unen al ADN. En resumen, las proteínas clasificadas como positivas por los clasificadores de clase I fueron proteínas con gran contenido de estructuras de tipo hélice.

Dentro de las proteínas clasificadas como positivas por los clasificadores de clase II/III hay mayor diversidad de funciones asociadas, aunque la mayor proporción corresponde a proteínas de unión a proteínas y proteínas de unión a iones calcio. Para este último grupo no se reportó particular presencia de hojas β en su estructura. Sin embargo para el primer grupo, uno de los dominios más reportado es el inmunoglobulina. Este dominio ha sido identificado en diversas proteínas con distintas funciones, que tienen en común la interacción proteína-proteína. Este dominio consiste en un sándwich de hojas β.

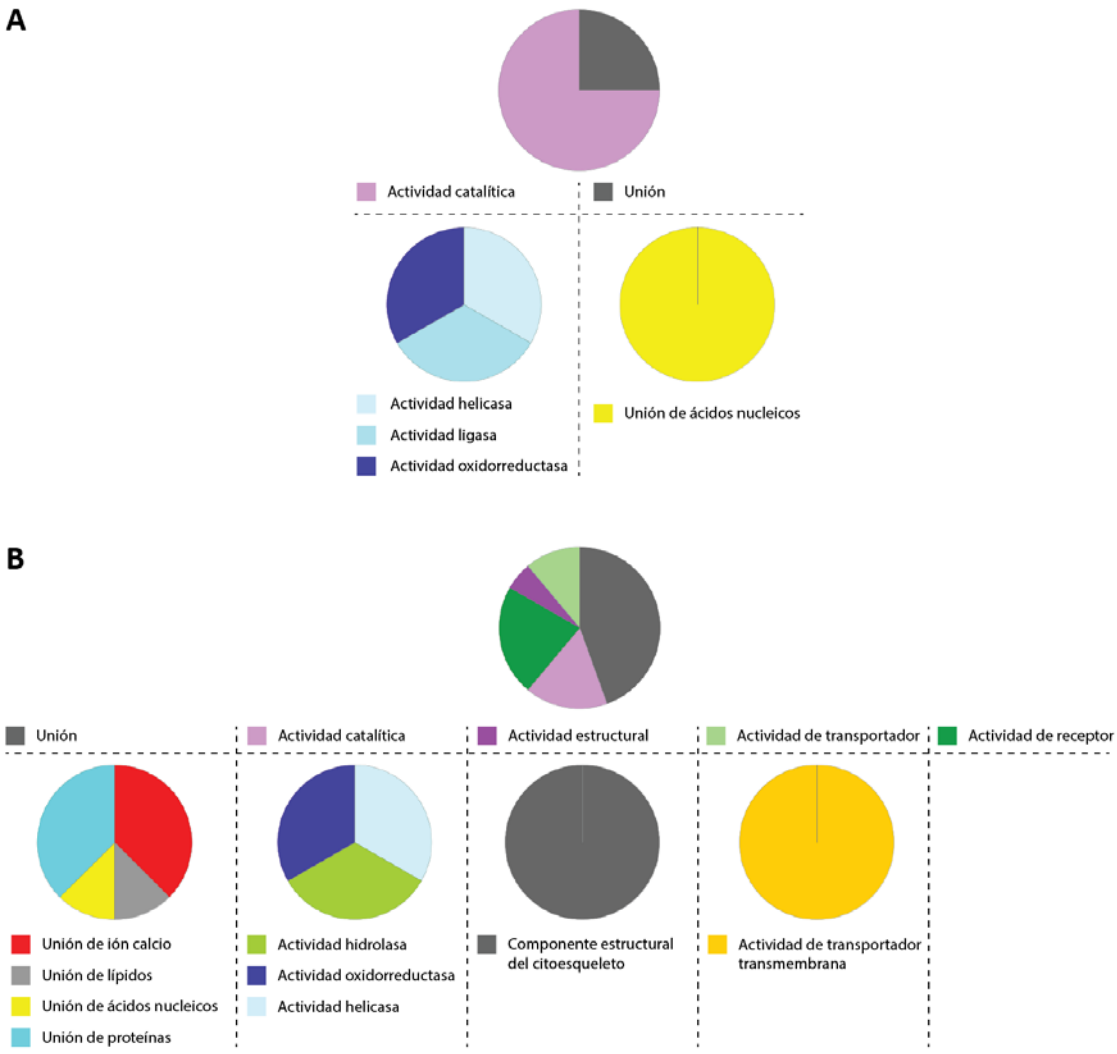


Figura 29. Análisis de Gene Ontology para las proteínas candidatas de función conocida.

5.5.3 Búsqueda de similitud con dominios conocidos

Para aquellas proteínas candidatas que cumplieron con la condición de topología y según UniProt no se conoce una función molecular se hizo una búsqueda contra la base de datos Pfam de dominios de proteínas utilizando la herramienta HHsearch del paquete HHsuite. Para cada proteína se realizó un alineamiento múltiple HHblits, se realizó un modelo de Markov oculto con HHmake y finalmente se realizó la búsqueda con HHsearch. El objetivo fue determinar si para estas proteínas HHsearch era capaz de identificar similitud con alta probabilidad a lo largo de una fracción significativa de la región extracelular de la proteína contra algún dominio depositado en la base de datos. Esta información se representa en las Fig. 30 y 31 y las Tabla 7 y 8. Aunque no parece haber un patrón claro en las características de las proteínas candidatas clasificadas como clase I, para las de clase II se observan principalmente proteínas de adhesión celular, con motivos ricos en hoja β.

Tabla 7. Proteínas de VEs que cumplen la topología de fusógeno viral de clase I y no tienen función molecular asignada según UniProt

UniProt ID	Gen	Proteína	Función UniProt	Predicción dominios	Autoencoder neural networks	Gaussian model	k-means	Vecinos más cercanos	Gaussian mixture model	Minimum spanning tree	Estimador de Parzen	Suma de clasificadores	DTMs	SP
A0A024R7I2	CYP4F3	Cytochrome P450		Cytochrome P450	1	0	0	0	0	0	0	1	3	0
A0A0D9SEG8	Adgre5	Adhesion G protei coupled receptor E5		GPCR-Autoproteolysis Inducing; GPCR proteolysis site	0	0	0	0	0	1	0	1	7	1
A0A0G2JYA3	Sppl2a	Signal peptide peptidase-like 2/		PA	1	1	0	0	0	1	1	4	9	1
A8K3Q8		cDNA FLJ75069			1	0	0	0	0	0	0	1	1	0
B5RIT8	Atpalpha-RD	Sodium/potassium transporting ATPase subunit alpha		Cation transport ATPase; Pyrimidine 5'-nucleotidase	1	0	0	0	0	0	1	2	6	0
D3DQ48	TMEM59	Transmembrane protein 59		Brain specific membrane anchored protein	1	0	0	0	0	1	0	2	1	1
D3ZG81	March8	Membrane-associated ring finger		RING-variant	0	0	0	0	0	1	0	1	2	0
D3ZNG3	Sppl2a	Signal peptide peptidase-like 2/		PA	0	1	0	0	0	1	0	2	9	1
E1BDT5	CYP26C1	Uncharacterized protein		Cytochrome P450	1	0	0	0	0	0	0	1	1	1
E9QMJ5	Adgre5	Adhesion G protei coupled receptor E5		GPCR-Autoproteolysis Inducing; GPCR proteolysis site	0	0	0	0	0	1	0	1	7	1
F2Z4F2	CYP19A1	Aromatase		Cytochrome P450	1	0	0	0	0	0	0	1	3	0
P43570	GYP8	GTPase-activating protein GYP8			1	0	0	0	0	0	0	1	4	0
Q0P6H9	TMEM62	Transmembrane protein 62		Metallophosphoesterase Glucodextranase	1	0	0	0	0	0	0	1	5	1
Q1XID4	Atp6ap2	Renin/prorenin receptor		Renin receptor-like protein	0	0	0	0	0	0	1	1	1	1
Q4DZJ9	Tc00.1047053508799.160	Cytochrome b5-like		Cytochrome b5-like Heme/Steroid binding domain	1	0	0	0	0	0	0	1	3	0
Q59FU8		Tumor necrosis factor receptor superfamily		TNFR/NGFR cystein-rich region	1	0	0	0	0	0	0	1	1	0
Q5FVF4	Tm7sf3	Transmembrane 7 superfamily member 3			0	0	0	0	0	1	0	1	7	1
Q642D2	Acp2	Acid phosphatase 2		Histidine phosphatase superfamily	1	0	0	0	0	0	0	1	1	1
Q96K49	TMEM87B	Transmembrane protein 87B		Pombe specific 5TM protein	0	0	0	0	0	1	0	1	7	1
Q96QC4	MICA	MHC class I polypeptide-related sequence A		Class I Histocompatibility antigen; Domain of unknown function	1	0	0	0	0	0	0	1	1	1
Q9HV31	pmrB	PmrB: two-component regulator system signal sensor kinas PmrB		Two-component sensor kinase	1	0	0	0	0	1	0	2	1	1
Q9HVY6	PA4429	Probable cytochrome c1		Cytochrome C1 family	1	0	0	0	0	0	0	1	1	1
Q9I0Y8	mexF	Resistance-Nodulation-Cell Division		AcrB/ActD/AcrF family	0	0	0	0	0	1	0	1	12	0
Q9I4F8	phoQ	Two-component sensor PhoC		PhoQ Sensor	1	1	0	0	1	1	1	5	1	1

Tabla 8. Proteínas de VEs que cumplen la topología de fusógeno viral de clase II y III y no tienen función molecular asignada según UniProt

UniProt ID	Gen	Proteína	Función UniProt	Predicción dominios	Autoencoder neural networks	Gaussian model	k-means	Vecinos más cercanos	Linear programming data descriction	Gaussian mixture model	Minimum spanning tree	Estimador de Parzen	Suma de clasificadores	DTMs	SP
A0A024R0K5	CEACAM5	Carcinoembryonic antigen-related cell adhesion molecule 5		Adhesion molecule	1	0	0	1	0	0	1	1	4	1	1
A0A024R7C1	hCG_2033729	HCG2033729		Adhesion molecule	1	0	0	1	1	0	1	1	5	1	1
A0A024R8Y9	SELP	Selectin P		Lectin C-type; Sushi repeat	0	0	0	0	1	0	0	1	2	2	0
A0A024RB01	ITGA5	Integrin		Integrin alpha	0	0	0	0	1	0	1	0	2	1	0
A0A024RDY3	LAMP1	Lysosomal-associated membrane protein 1		Lysosome-associated membrane glycoprotein	1	1	0	1	1	0	1	1	6	1	1
A0A087WVM2	CD177	CD177 antigen		Phospholipase A2 inhibitor; Ly6/PLAUR	0	0	0	0	1	0	1	0	2	1	1
A0A087WWD4	NCAM1	Neural cell adhesion molecule 1		Intercellular adhesion molecule; Domain of unknown function	0	0	0	0	1	0	0	0	1	1	1
A0A087X0M8	CHL1	Neural cell adhesion molecule L1-like protein		Intercellular adhesion molecule; Domain of unknown function	0	0	0	0	1	0	0	0	1	1	1
A0A0A0MSV9	TAPBP	Tapasin		Intercellular adhesion molecule	1	1	1	1	1	0	1	1	7	1	1
A0A0A0MT98	TAPBP	Tapasin		Intercellular adhesion molecule	1	0	0	1	1	0	1	1	5	1	1
A0A0A6YX40	Igsf3	Immunoglobulin superfamily member 3		Intercellular adhesion molecule; ICOS V-set	1	1	0	1	1	0	1	1	6	1	1
A0A0C4DG49	PVR	Poliovirus receptor		ICOS V-set; Intercellular adhesion molecule	1	0	0	1	1	0	1	1	5	1	1
A0A0C4DG76	HEPH	Hephaestin		Multicopper oxidase	0	0	0	1	1	0	1	0	3	1	1
A0A0G2JNR3	LILRB1	Leukocyte immunoglobulin-like receptor subfamily B member 1		Intercellular adhesion molecule	0	0	0	0	1	0	0	0	1	1	1
A0A0G2JSI0	Fmo3	Dimethylaniline monooxygenase		Flavin-binding monooxygenase-like	0	0	0	0	1	0	1	0	2	2	1
A0A0G2JSP1	Umod	Uromodulin		Zona pellucida-like	0	0	0	1	1	0	1	1	4	1	1
A0A0G2JSW2	Ceacam1	Carcinoembryonic antigen-related cell adhesion molecule 1		Intercellular adhesion molecule	1	0	0	1	1	0	1	1	5	1	1
A0A4Z1	CADM1	CADM1 protein		Intercellular adhesion molecule	1	0	0	1	0	0	1	1	4	1	0
A1A4T2	Ganab	Alpha glucosidase 2 alpha neutral subunit		Maltase-glucoamylase; Galactose mutarotase-like; Glycosyl hydrolases family 31	1	0	0	0	1	0	0	0	2	1	1
A2DCZ1	TVAG_237910	Uncharacterized protein			1	1	0	1	1	0	1	1	6	1	0
A2EFA8	TVAG_239650	Uncharacterized protein			0	0	0	0	1	0	1	1	3	1	0
A5D7Q6	PCDHGA2	PCDHGA2 protein		Domain of unknown function	0	0	0	0	1	0	0	0	1	1	1
A5PK12	CD177	CD177 protein		Phospholipase A2 inhibitor	0	0	0	1	1	0	1	0	3	1	1
A6QLR1	PVRL4	Nectin-4		Adhesion molecule	0	0	0	0	1	0	0	0	1	1	1
A6QQ97	ITGAD	ITGAD protein		Integrin beta chain VWA; Integrin alpha	0	0	0	0	1	0	1	0	2	1	1
A8K6K4		cDNA FLJ77565		Adhesion molecule	0	0	0	0	1	0	1	1	3	1	1
B0BNG3	Lman2	Lectin		Adhesion Legume-like lectin family	0	0	0	0	1	0	0	0	1	1	0
B2RAL6		cDNA FLJ94991		Integrin beta chain VWA; Integrin alpha	0	0	0	0	1	0	1	1	3	1	1
B4DHC3		cDNA FLJ61254		Intercellular adhesion molecule; Domain of unknown function	1	1	0	1	1	0	1	1	6	1	1
B4DLZ2		cDNA FLJ55893		Putative ephrine-receptor like; Autophagy-related protein 27; Mannose-6-phosphate receptor	0	0	0	0	1	0	0	0	1	1	0
B4E282		cDNA FLJ52401		Integrin beta chain VWA; Integrin alpha	1	0	0	0	1	0	1	1	4	1	0
B7ZL91	MEP1A	Meprin A subunit		Astacin; MAM; MATH	0	0	0	0	1	0	1	0	2	1	1

E1BDY7	PLXNC1	Uncharacterized protein		Sema; Plexin repeat; IPT/TIG	0	0	0	0	1	0	0	0	1	2	1
E1BFQ6	ITGA6	Uncharacterized protein		FG-GAP repeat; Integrin-alpha FG-GAP repeat; Integrin alpha	0	0	0	0	1	0	1	0	2	1	1
E1BRR6	SCARB2	Uncharacterized protein		CD36 family	0	0	0	0	1	0	1	0	2	2	0
E7ES21	HEPH	Hephaestin		Multicopper oxidase	0	0	0	0	1	0	1	0	2	1	1
E9PDN6	CNTNAP4	Contactin-associated protein-like 4		Laminin G domain; Fibrillar collagen	1	1	0	0	1	0	1	1	5	1	0
E9PZD8	Cp	Ceruloplasmin		Multicopper oxidase	1	0	0	1	1	0	1	1	5	1	1
E9Q5M7	Itgal	Integrin alpha-L		Integrin beta chain VWA; Integrin alpha	1	0	0	1	1	0	1	1	5	1	1
F1LP44	Itgal	Integrin subunit alpha L		Integrin beta chain VWA; Integrin alpha	0	0	0	0	1	0	1	0	2	1	1
F1MFH3	FREM2	Uncharacterized protein		Domain of unknown function; Cadherin-like; Domain of unknown function	0	0	0	0	1	0	0	0	1	1	1
F1MHN8	ALCAM	CD166 antigen		SLAM protein; Intercellular adhesion molecule	0	0	0	0	0	0	0	1	1	1	1
F1MLJ1	PCDHB1	Uncharacterized protein		Domain of unknown function	0	0	0	0	1	0	0	0	1	1	1
F1MSE7	SUSD2	Uncharacterized protein		Anthrax receptor; AMOP; von Willebrand factor type D; Sushi	1	1	0	0	1	0	1	1	5	1	1
F1MVI0	CNTN1	Contactin-1		Intercellular adhesion molecule; Domain of unknown function	1	1	0	1	1	0	1	1	6	1	1
F1N1A9	ITGAV	Integrin alpha-V		FG-GAP repeat; Integrin-alpha FG-GAP repeat; Integrin alpha	0	0	0	0	1	0	1	1	3	1	1
F1N720	TLR2	Toll-like receptor		Leucine-rich repeat	0	0	0	0	1	0	0	0	1	1	1
G3V824	Igf2r	Insulin-like growth factor 2 receptor		Mannose-6-phosphate receptor; Cation independent mannose-6-phosphate receptor repeat	0	0	0	0	1	0	0	1	2	1	1
G3X9Q1	Itga7	Integrin alpha 7		FG-GAP repeat; Integrin-alpha FG-GAP repeat; Integrin alpha	0	0	0	0	1	0	0	0	1	1	1
H3BN02	ITGAX	Integrin alpha-X		Integrin beta chain VWA; Integrin alpha	1	0	0	0	1	0	1	1	4	1	1
O14498	ISLR	Immunoglobulin superfamily containing leucine-rich repeat protein		Leucine-rich repeat; Intercellular adhesion molecule	1	0	0	0	1	1	1	0	4	1	1
O75054	IGSF3	Immunoglobulin superfamily member 3		Intercellular adhesion molecule; ICOS V-set	1	0	0	1	1	0	1	1	5	1	1
O94856	NFASC	Neurofascin	Cell adhesion, ankyrin-binding protein which may be involved in neurite extension, axonal guidance, synaptogenesis, myelination and neuron-glial cell interactions. {ECO:0000250}.	Intercellular adhesion molecule; Domain of unknown function	0	0	0	0	1	0	0	0	1	1	1
P46992	YJL171C	Cell wall protein YJL171C		Glycine-rich protein; Putative TOS1-like glycosyl hydrolase	0	0	0	0	1	0	1	1	3	1	1
Q004B2	CEACAM1	Carcinoembryonic antigen-related cell adhesion molecule 1		Intercellular adhesion molecule	1	1	1	1	1	0	1	1	7	1	1
Q05204	LAMP1	Lysosome-associated membrane glycoprotein 1		Lysosome-associated membrane glycoprotein	1	1	1	1	1	1	1	1	8	1	1
Q0P6H9	TMEM62	Transmembrane protein 62		Metallophosphoesterase ; Glucodextranase	1	0	0	0	1	0	1	1	4	5	1
Q1RMV1	ENG	Endoglin		Zona pellucida-like	0	1	0	0	1	0	1	1	4	1	1
Q2HZ94	Mrc1	Macrophage mannose receptor 1		Lectin C-type	0	0	0	0	1	0	1	1	3	1	1
Q2TBL2	CADM1	Cell adhesion molecule 1		Intercellular adhesion molecule	1	0	0	1	0	0	1	1	4	1	1
Q3TZS3	Itga7	Putative uncharacterized protein		FG-GAP repeat; Integrin-alpha FG-GAP repeat; Integrin alpha	0	0	0	0	1	0	1	0	2	1	1
Q3U1U4	Itgam	Integrin alpha-M		Integrin beta chain VWA; Integrin alpha	1	0	0	0	1	0	1	1	4	1	1

Q4DDF0	Tc00.1047053510517.40	Uncharacterized protein			0	0	0	0	1	0	0	0	1	1	1
Q4DDU4	Tc00.1047053506751.50	Trans-sialidase		BNR repeat-like; Pentaxin family	1	1	0	1	1	0	1	1	6	3	0
Q4LE35	ITGA7	ITGA7 variant protein		FG-GAP repeat; Integrin-alpha FG-GAP repeat; Integrin alpha	0	0	0	0	1	0	0	0	1	2	0
Q59EZ1		Protein tyrosine phosphatase		Intercellular adhesion molecule; Domain of unknown function	0	0	0	1	1	0	1	1	4	1	0
Q5E9G7	CDH16	Cadherin 16		Cadherin domain; Domain of unknown function	0	0	0	0	1	0	0	1	2	1	1
Q5R341	SELP	P-selectin		Lectin C-type; Sushi	0	0	0	0	1	0	1	1	3	2	0
Q5U355	Itfg1	Itfg1 protein		Integrin-alpha FG-GAP repeat-containing protein 2; ASPIC and UnbV	1	1	0	1	1	0	1	1	6	1	1
Q6P6W1		Lysosome-associated membrane glycoprotein 2	Plays an important role in chaperone-mediated autophagy, a process that mediates lysosomal degradation of proteins in response to various stresses and as part of the normal turnover of proteins with a long biological half-live (PubMed:8662539, PubMed:11082038). Binds target proteins, such as GAPDH, and targets them for lysosomal degradation (PubMed:8662539, PubMed:11082038, PubMed:18644871). Plays a role in lysosomal protein degradation in response to starvation (PubMed:11082038). Required for the fusion of autophagosomes with lysosomes during autophagy. Cells that lack LAMP2 express normal levels of VAMP8, but fail to accumulate STX17 on autophagosomes, which is the most likely explanation for the lack of fusion between autophagosomes and lysosomes. Required for normal degradation of the contents of autophagosomes. Required for efficient MHCII-mediated	Lysosome-associated membrane glycoprotein	1	1	1	1	1	1	1	1	8	1	1
Q6UVY6	MOXD1	DBH-like monooxygenase protein 1		Cytochrome; Copper type II ascorbate-dependent monooxygenase	1	1	0	1	1	0	1	1	6	1	1
Q6ZQA6	Igsf3	Immunoglobulin superfamily member 3		Intercellular adhesion molecule; ICOS V-set	1	0	0	1	1	0	1	1	5	1	1
Q7Z407	CSMD3	CUB and sushi domain-containing protein 3		CUB domain; Corticotropin-releasing factor binding protein	0	0	0	1	1	0	1	1	4	1	0
Q7Z408	CSMD2	CUB and sushi domain-containing protein 2		CUB domain; Corticotropin-releasing factor binding protein	1	1	0	1	1	0	1	1	6	1	0
Q86WI1	PKHD1L1	Fibrocystin-L		PA14; IPT/TIG; G8	1	0	0	0	1	0	1	1	4	1	1
Q8C5K0	Lamp2	Lysosomal membrane glycoprotein 2		Lysosome-associated membrane glycoprotein	1	1	0	1	1	0	1	1	6	1	1
Q8CC06	Itga6	Putative uncharacterized protein		FG-GAP repeat; Integrin-alpha FG-GAP repeat; Integrin alpha	0	0	0	0	1	0	1	0	2	1	1
Q8IYS2	KIAA2013	Uncharacterized protein KIAA2013		Uncharacterized conserved protein	1	0	0	1	1	0	1	1	5	2	0
Q8K094	Pvr	Poliovirus receptor		ICOS V-set; Intercellular adhesion molecule	1	0	0	1	1	0	1	1	5	1	1
Q8K4C0	Fmo5	Dimethylaniline monooxygenase	In contrast with other forms of FMO it does not seem to be a drug-metabolizing enzyme. {ECO:0000250}.	Flavin-binding monooxygenase-like	0	0	0	0	0	0	1	0	1	1	1
Q8N766	EMC1	ER membrane protein complex subunit 1		PQQ-like; Protein of unknown function	1	0	0	0	1	0	1	1	4	1	1
Q96PQ0	SORCS2	VPS10 domain-containing receptor SorCS2		Sortilin	0	0	0	0	1	0	1	0	2	1	0

Q98TT7		Integrin alpha 3A subunit cytoplasmic domain variant	FG-GAP repeat; Integrin-alpha FG-GAP repeat; Integrin alpha	1	0	0	1	1	0	1	1	5	1	1
Q99JC6	Tapbp	TAP binding protein	Intercellular adhesion molecule	1	0	0	0	1	0	1	1	4	1	1
Q9DC13	Lamp1	Lysosomal membrane glycoprotein 1	Lysosome-associated membrane glycoprotein	1	1	0	1	1	1	1	1	7	1	1
Q9HU71	PA5113	Uncharacterized protein	Protein of unknown function	0	0	0	1	1	0	1	1	4	1	1
Q9JI30	Itgam	Integrin beta 2 alpha subunit	Integrin beta chain VWA; Integrin alpha	1	0	0	1	1	0	1	1	5	1	1
X5D7A8	CADM1	Cell adhesion molecule 1	Intercellular adhesion molecule	0	0	0	0	1	0	0	0	1	1	0
X5DQR8	CADM1	Cell adhesion molecule 1	Intercellular adhesion molecule	1	0	0	1	1	0	1	1	5	1	0
X5DQS5	CADM1	Cell adhesion molecule 1	Intercellular adhesion molecule	0	0	0	0	1	0	0	0	1	1	0

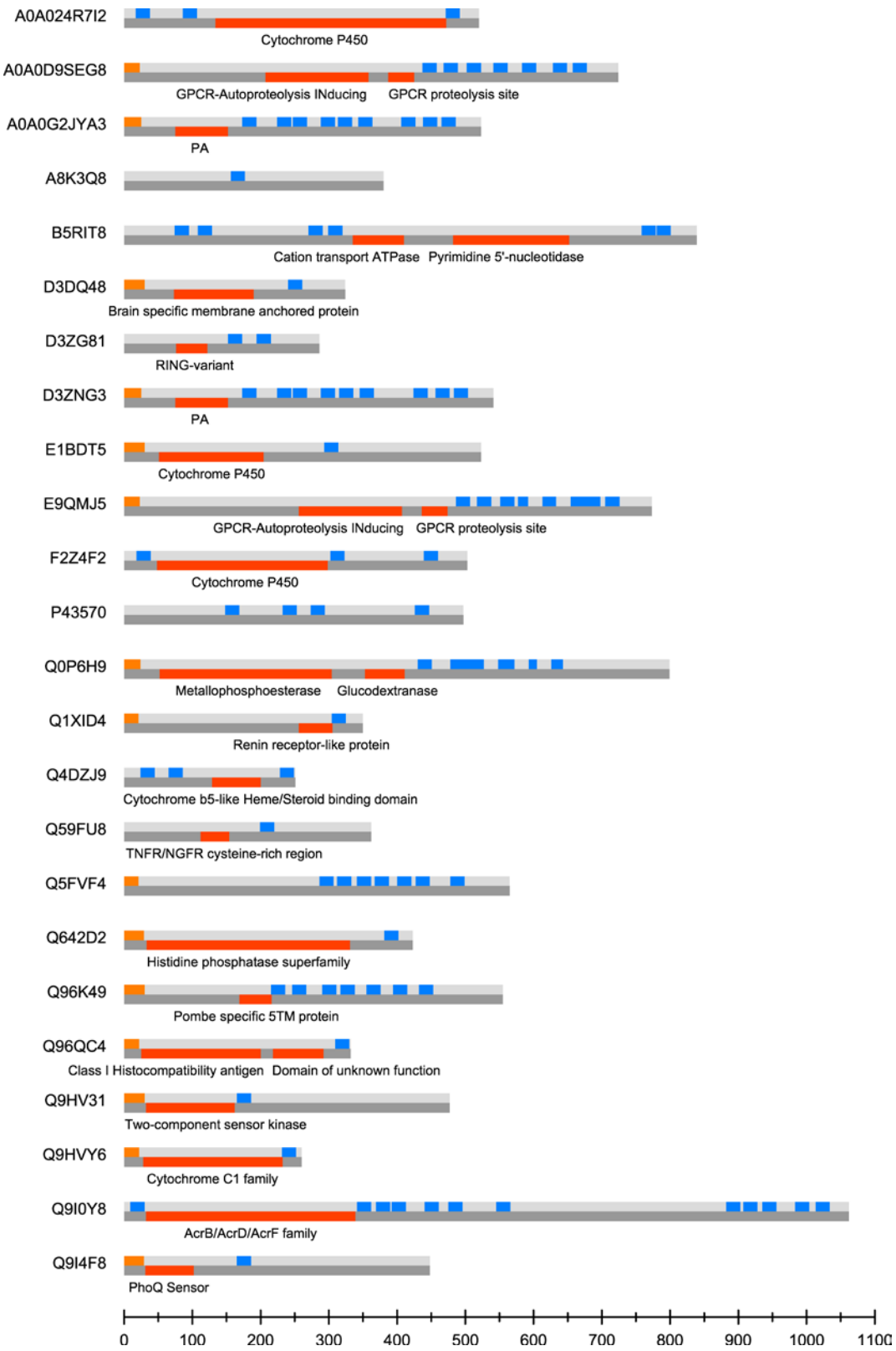
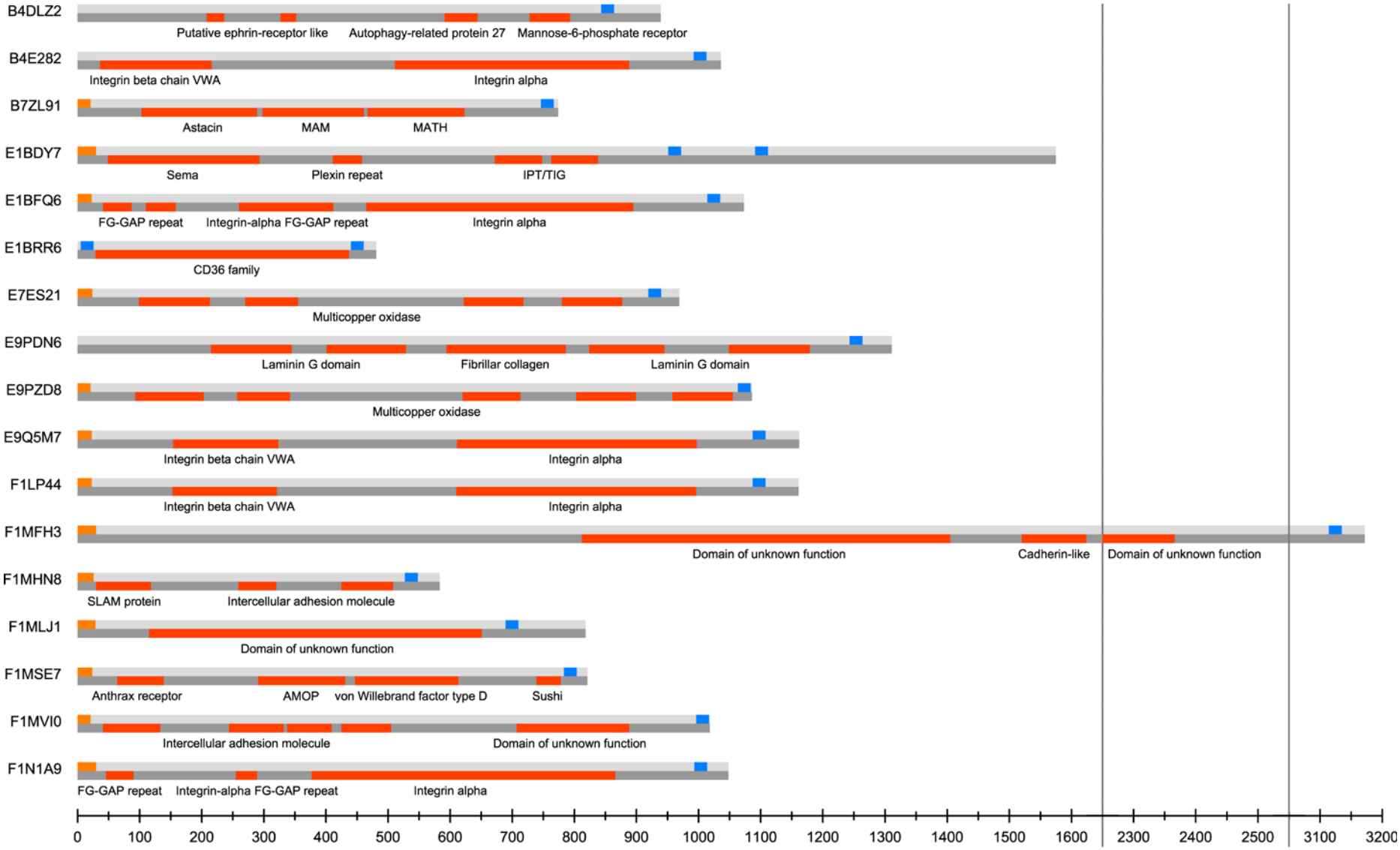
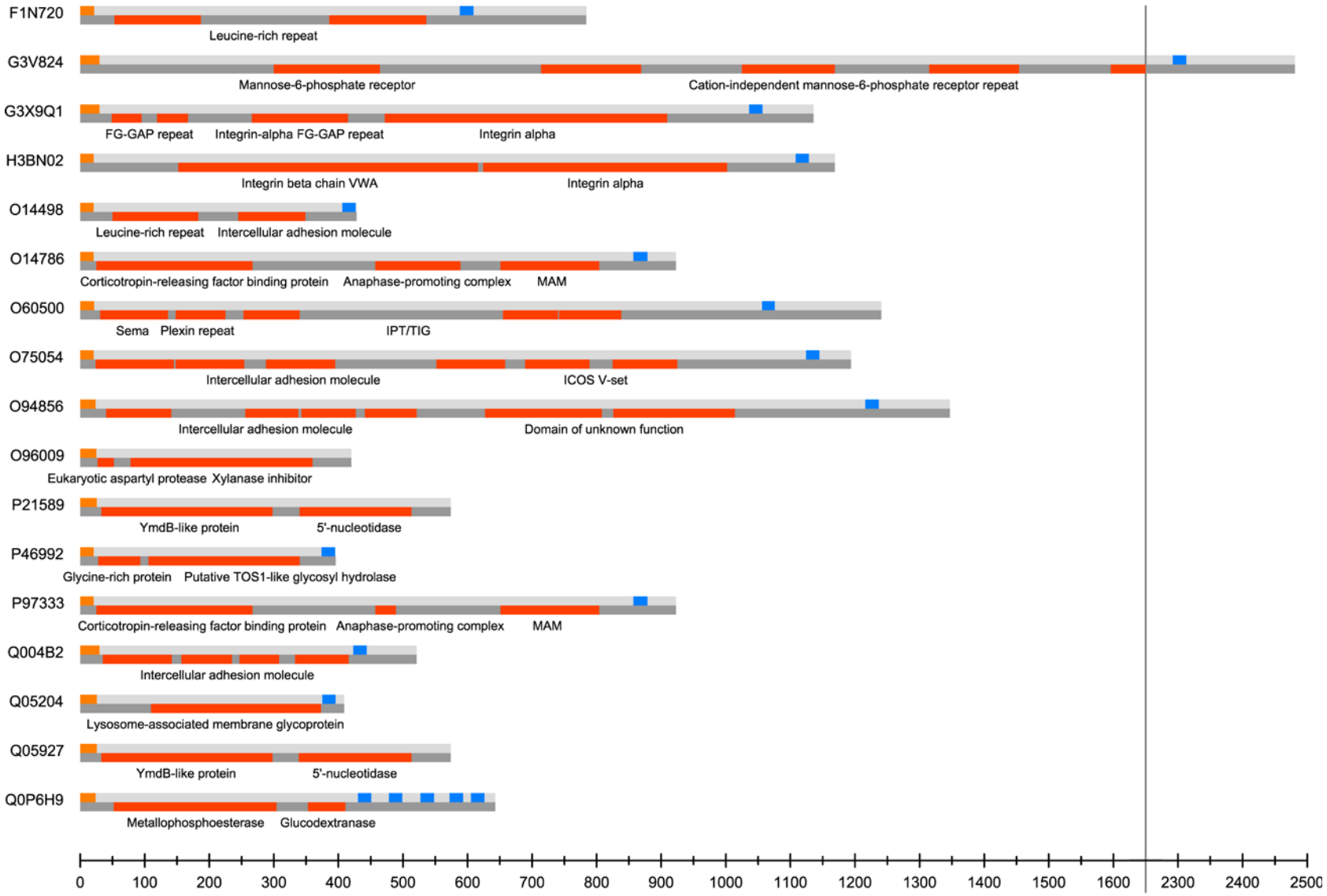
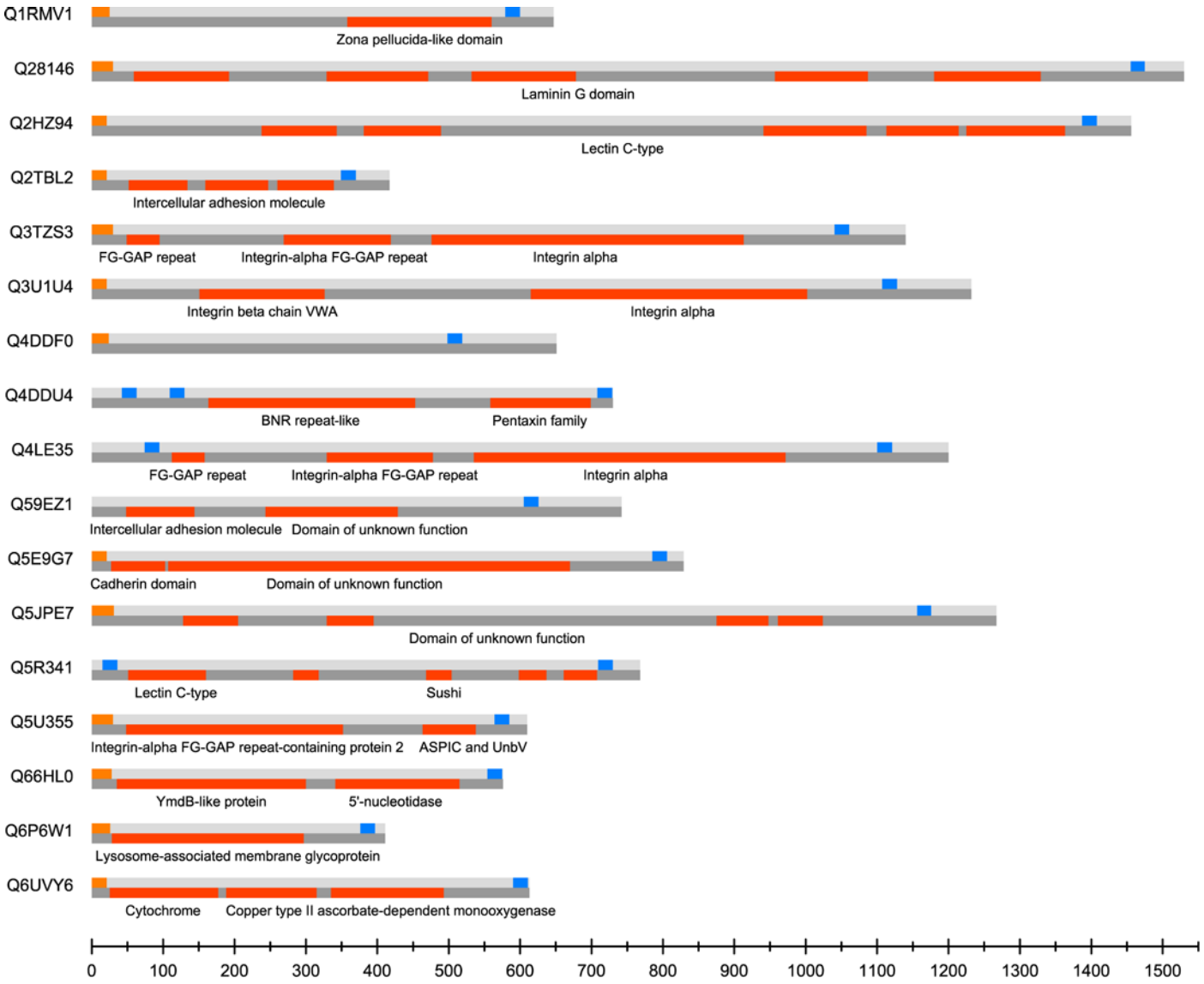


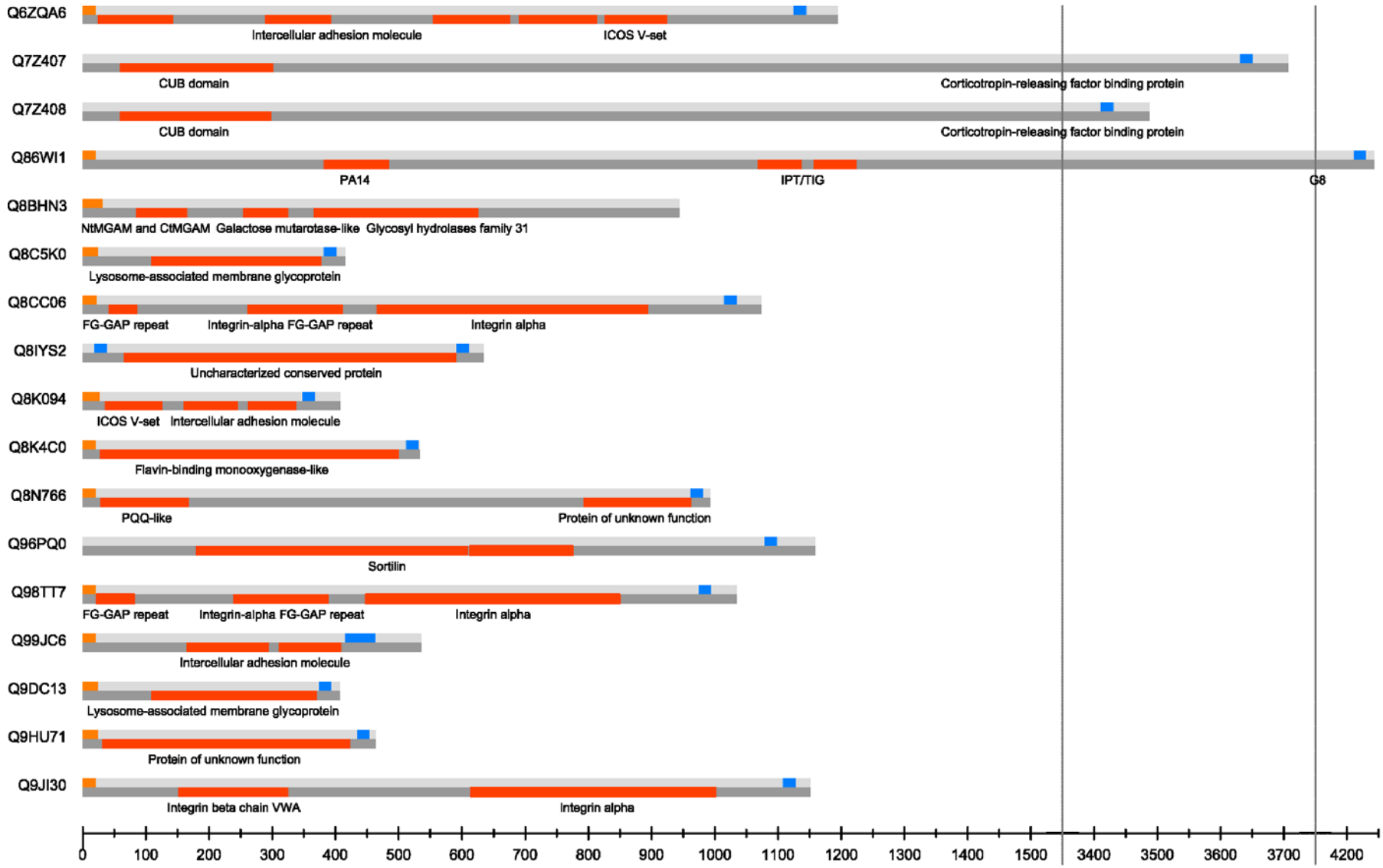
Figura 30. Representación de proteínas candidatas de clase I de función desconocida. Se representan las posiciones de los péptidos señal, dominios transmembrana y dominios predichos para las proteínas clasificadas como clase I para las que no se conoce su función molecular. La primera línea representa los péptidos señal (naranja) y dominios transmembrana (azul). La segunda línea ubica los dominios predichos con alta probabilidad según HHsearch (rojo).


A0A024R0K5
Adhesion molecule
A0A024R7C1
Adhesion molecule
A0A024R8Y9
Lectin C-type
Sushi repeat
A0A024RB01
Integrin alpha
A0A024RDY3
Lysosome-associated membrane glycoprotein
A0A087WVM2
Phospholipase A2 inhibitor
Ly6/PLAUR
A0A087WWD4
Intercellular adhesion molecule
Domain of unknown function
A0A087X0M8
Intercellular adhesion molecule
Domain of unknown function
A0A0A0MSV9
Intercellular adhesion molecule
A0A0A0MT98
Intercellular adhesion molecule
A0A0A6YX40
Intercellular adhesion molecule
ICOS V-set
A0A0C4DG49
ICOS V-set
Intercellular adhesion molecule
A0A0C4DG76
Multicopper oxidase
A0A0G2JNR3
Intercellular adhesion molecule
A0A0G2JSI0
Flavin-binding monooxygenase-like
A0A0G2JSP1
Zona pellucida-like
A0A0G2JSW2
Intercellular adhesion molecule
A0A4Z1
Intercellular adhesion molecule
A1A4T2
Maltase-glucoamylase
Galactose mutarotase-like
Glycosyl hydrolases family 31
A2DCZ1
A2EFA8
A5D7Q6
Domain of unknown function
A5PK12
Phospholipase A2 inhibitor
A6QLR1
Adhesion molecule
A6QQ97
Integrin beta chain VWA
Integrin alpha
A8K6K4
Adhesion molecule
B0BNG3
Adhesion Legume-like lectin family
B2RAL6
Integrin beta chain VWA
Integrin alpha
B4DHC3
Intercellular adhesion molecule
Domain of unknown function
0
100
200
300
400
500
600
700
800
900
1000
1100
1200


B4DLZ2
Putative ephrin-receptor like
Autophagy-related protein 27
Mannose-6-phosphate receptor
B4E282
Integrin beta chain VWA
Integrin alpha
B7ZL91
Astacin
MAM
MATH
E1BDY7
Sema
Plexin repeat
IPT/TIG
E1BFQ6
FG-GAP repeat
Integrin-alpha FG-GAP repeat
Integrin alpha
E1BRR6
CD36 family
E7ES21
Multicopper oxidase
E9PDN6
Laminin G domain
Fibrillar collagen
Laminin G domain
E9PZD8
Multicopper oxidase
E9Q5M7
Integrin beta chain VWA
Integrin alpha
F1LP44
Integrin beta chain VWA
Integrin alpha
F1MFH3
Domain of unknown function
Cadherin-like
Domain of unknown function
F1MHN8
SLAM protein
Intercellular adhesion molecule
F1MLJ1
Domain of unknown function
F1MSE7
Anthrax receptor
AMOP
von Willebrand factor type D
Sushi
F1MVI0
Intercellular adhesion molecule
Domain of unknown function
F1N1A9
FG-GAP repeat
Integrin-alpha FG-GAP repeat
Integrin alpha
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
2300
2400
2500
3100
3200


F1N720
Leucine-rich repeat
G3V824
Mannose-6-phosphate receptor
Cation-independent mannose-6-phosphate receptor repeat
G3X9Q1
FG-GAP repeat
Integrin-alpha FG-GAP repeat
Integrin alpha
H3BN02
Integrin beta chain VWA
Integrin alpha
O14498
Leucine-rich repeat
Intercellular adhesion molecule
O14786
Corticotropin-releasing factor binding protein
Anaphase-promoting complex
MAM
O60500
Sema
Plexin repeat
IPT/TIG
O75054
Intercellular adhesion molecule
ICOS V-set
O94856
Intercellular adhesion molecule
Domain of unknown function
O96009
Eukaryotic aspartyl protease
Xylanase inhibitor
P21589
YmdB-like protein
5'-nucleotidase
P46992
Glycine-rich protein
Putative TOS1-like glycosyl hydrolase
P97333
Corticotropin-releasing factor binding protein
Anaphase-promoting complex
MAM
Q004B2
Intercellular adhesion molecule
Q05204
Lysosome-associated membrane glycoprotein
Q05927
YmdB-like protein
5'-nucleotidase
Q0P6H9
Metallophosphoesterase
Glucodextranase
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
2300
2400
2500


Q1RMV1
Zona pellucida-like domain
Q28146
Laminin G domain
Q2HZ94
Lectin C-type
Q2TBL2
Intercellular adhesion molecule
Q3TZS3
FG-GAP repeat
Integrin-alpha FG-GAP repeat
Integrin alpha
Q3U1U4
Integrin beta chain VWA
Integrin alpha
Q4DDF0
Q4DDU4
BNR repeat-like
Pentaxin family
Q4LE35
FG-GAP repeat
Integrin-alpha FG-GAP repeat
Integrin alpha
Q59EZ1
Intercellular adhesion molecule
Domain of unknown function
Q5E9G7
Cadherin domain
Domain of unknown function
Q5JPE7
Domain of unknown function
Q5R341
Lectin C-type
Sushi
Q5U355
Integrin-alpha FG-GAP repeat-containing protein 2
ASPIC and UnbV
Q66HL0
YmdB-like protein
5'-nucleotidase
Q6P6W1
Lysosome-associated membrane glycoprotein
Q6UVY6
Cytochrome
Copper type II ascorbate-dependent monooxygenase
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500


Q6ZQA6
Intercellular adhesion molecule
ICOS V-set
Q7Z407
CUB domain
Corticotropin-releasing factor binding protein
Q7Z408
CUB domain
Corticotropin-releasing factor binding protein
Q86WI1
PA14
IPT/TIG
G8
Q8BHN3
NtMGAM and CtMGAM Galactose mutarotase-like Glycosyl hydrolases family 31
Q8C5K0
Lysosome-associated membrane glycoprotein
Q8CC06
FG-GAP repeat
Integrin-alpha FG-GAP repeat
Integrin alpha
Q8IYS2
Uncharacterized conserved protein
Q8K094
ICOS V-set
Intercellular adhesion molecule
Q8K4C0
Flavin-binding monooxygenase-like
Q8N766
PQQ-like
Protein of unknown function
Q96PQ0
Sortilin
Q98TT7
FG-GAP repeat
Integrin-alpha FG-GAP repeat
Integrin alpha
Q99JC6
Intercellular adhesion molecule
Q9DC13
Lysosome-associated membrane glycoprotein
Q9HU71
Protein of unknown function
Q9JI30
Integrin beta chain VWA
Integrin alpha
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
3400
3500
3600
3700
4200

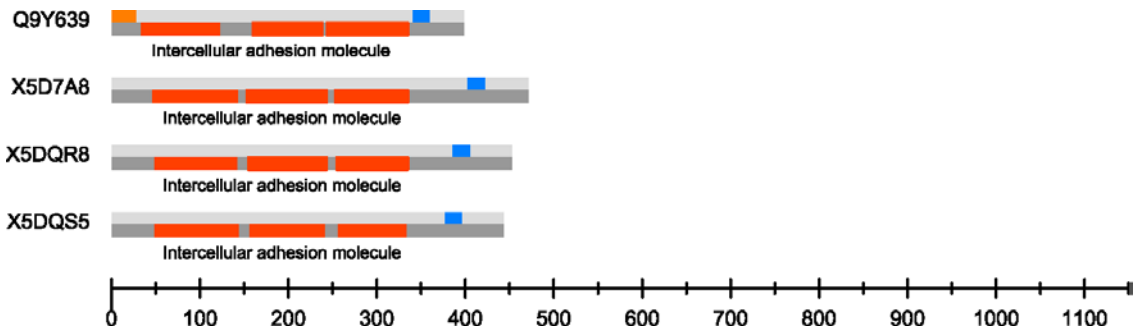


Figura 31. Representación de proteínas candidatas clase II y III de función desconocida. Se representan las posiciones de los péptidos señal, dominios transmembrana y dominios predichos para las proteínas clasificadas como clase II y III para las que no se conoce su función molecular. La primera línea representa los péptidos señal (naranja) y dominios transmembrana (azul). La segunda línea ubica los dominios predichos con alta probabilidad según HHsearch (rojo).

5.5.4 Armado de perfiles y evaluación contra PDB

Se descargaron todas las proteínas correspondientes a fusógenos virales de clase I, II y II de la base de datos Protein Data Bank (PDB) (Berman et al., 2000). Esta base de datos pública contiene información sobre estructuras tridimensionales de proteínas obtenidas en su mayoría por cristalografía de rayos X o resonancia magnética nuclear (RMN). Se generó una base de datos para cada clase. Cada base de datos de modelos de Markov ocultos fue generada con las herramientas HHblits, addss.pl y HHmake del paquete HHsearch.

Se exploraron las proteínas candidatas del punto anterior nuevamente con la herramienta HHsearch, pero utilizando dichas bases de datos de estructuras tridimensionales de fusógenos virales. El objetivo de la búsqueda contra la base de datos de fusógenos virales fue identificar regiones extensas (más de 50 aminoácidos) de similitud con fusógenos virales aunque la probabilidad del hit sea baja, ya que análisis previos para las proteínas EFF-1 y HAP2 contra esta base de datos tuvieron resultados con esas características. En caso de identificar similitud que cumpla con las características mencionadas, asumiríamos que la función de la proteína en cuestión es la misma que la de su hit de PDB, dada la estrecha relación entre estructura tridimensional y función (Fig. 32).

Se identificó únicamente una proteína con similitud con un fusógeno viral de clase I según HHsearch. De todas formas, esa región de similitud no se corresponde con la porción N-terminal de la proteína respecto de dominio extracelular de acuerdo a la topología conocida para fusógenos virales.

Se identificaron varias proteínas con similitud con fusógenos virales de clase III. Sin embargo, la mayoría corresponde a regiones de la proteína para las que ya se había predicho un dominio conocido con alta probabilidad. Las regiones de similitud que no caen sobre dominios conocidos se ubican en proteínas muy extensas, y no se conoce ningún fusógeno viral con estas características. Las regiones de similitud correspondientes sobre los fusógenos de clase III corresponden a distintos dominios.

Las regiones de similitud con proteínas de clase II se corresponden en su mayoría con el dominio inmunoglobulina (dIII) del fusógeno viral. Dado que el dominio inmunoglobulina es un dominio muy distribuido en la naturaleza estas proteínas candidatas no resultan particularmente interesante. Sin embargo se observó algo particular para un conjunto de proteínas. Las proteínas A0A0G2JSP1 y P07911 (Uromodulina), O88393 (TGFBR3), y Q05996 y Q9BH10 (Zona pellucida sperm-binding protein 2 (ZP2)) presentan un dominio Zona pellucida-like. En las proteínas Uromodulina y TGFBR3 se identificó una región de similitud con el dominio inmunoglobulina del fusógeno viral de clase I E1 del virus Chikungunya. Este dominio se localiza justo antes del dominio transmembrana de igual forma que en el fusógeno viral. Esta región de similitud coincide con la ubicación del dominio Zona pellucida-like, posiblemente indicando que parte de este dominio adopta una conformación tipo inmunoglobulina. La proteína Uromodulina es la proteína más abundante en orina, sin embargo no se conoce su función. Se cree que participa en la biogénesis y organización de la membrana apical de las células epiteliales de la rama ascendente gruesa del asa de Henle, donde promueve la formación de la estructura filamentosa que tendría un rol en la permeabilidad de la barrera de agua (Schaeffer, Santambrogio, Perucca, Casari, & Rampoldi, 2009). El fragmento que corresponde al dominio Zona pellucida-like, correspondiente a la región de polimerización, fue cristalizado recientemente. Por su parte, la proteína TGFBR3 se cree que está involucrada en la captura y retención de TGFB para la presentación a receptores de transducción de señales (Wang et al., 1991). Finalmente, la proteína ZP2 presenta una región de similitud con parte de los dominios dI y dII del fusógeno viral de clase II E del virus del Dengue. Esta región de similitud abarca casi la totalidad de los subdominios 1 y 2 del dominio dII e incluye también el péptido de fusión. La región de similitud no está superpuesta a ningún dominio conocido en la proteína. Resulta interesante que en la proteína de fusión EFF-1 de Caenorhabditis elegans se identificó en una posición equivalente la misma región de similitud con el virus del Dengue. La topología de la proteína ZP2 sigue los mismos patrones que la topología de EFF-1 (y por consiguiente de los fusógenos virales de clase II), presentando una longitud similar, un péptido señal en el extremo N-terminal, y un dominio transmembrana en la región C-terminal. Además, considerando que la predicción del dominio inmunoglobulina de la proteína E1 del virus Chikungunya sobre la región C-terminal del dominio Zona-pellucida-like es correcta, la proteína ZP2 presentaría también un dominio inmunoglobulina en la posición C-terminal, tal como los fusógenos virales de clase II. Para la proteína ZP2 bajo el identificador Q9BH10 se identificó una región similar al dominio inmunoglobulina de la proteína E1 del virus Chikungunya aunque es menor a 50 aminoácidos. Esta región de similitud se corresponde con la posición característica del dominio inmunoglobulina en los fusógenos virales de clase II.

Una de las proteínas ZP2 identificadas es de humano y aparece en Vesiclepedia por su identificación en el trabajo de Peterson et al. (Peterson et al., 2008), mientras que la otra es de bovino y se identificó en el trabajo de (Reinhardt, Lippolis, Nonnecke, & Sacco, 2012). Ambos corresponden a experimentos de espectrometría de masas, pero el primero corresponde a un análisis de células epiteliales y el segundo a leche. Sin embargo, en la bibliografía se ha reportado su expresión únicamente en ovocitos.

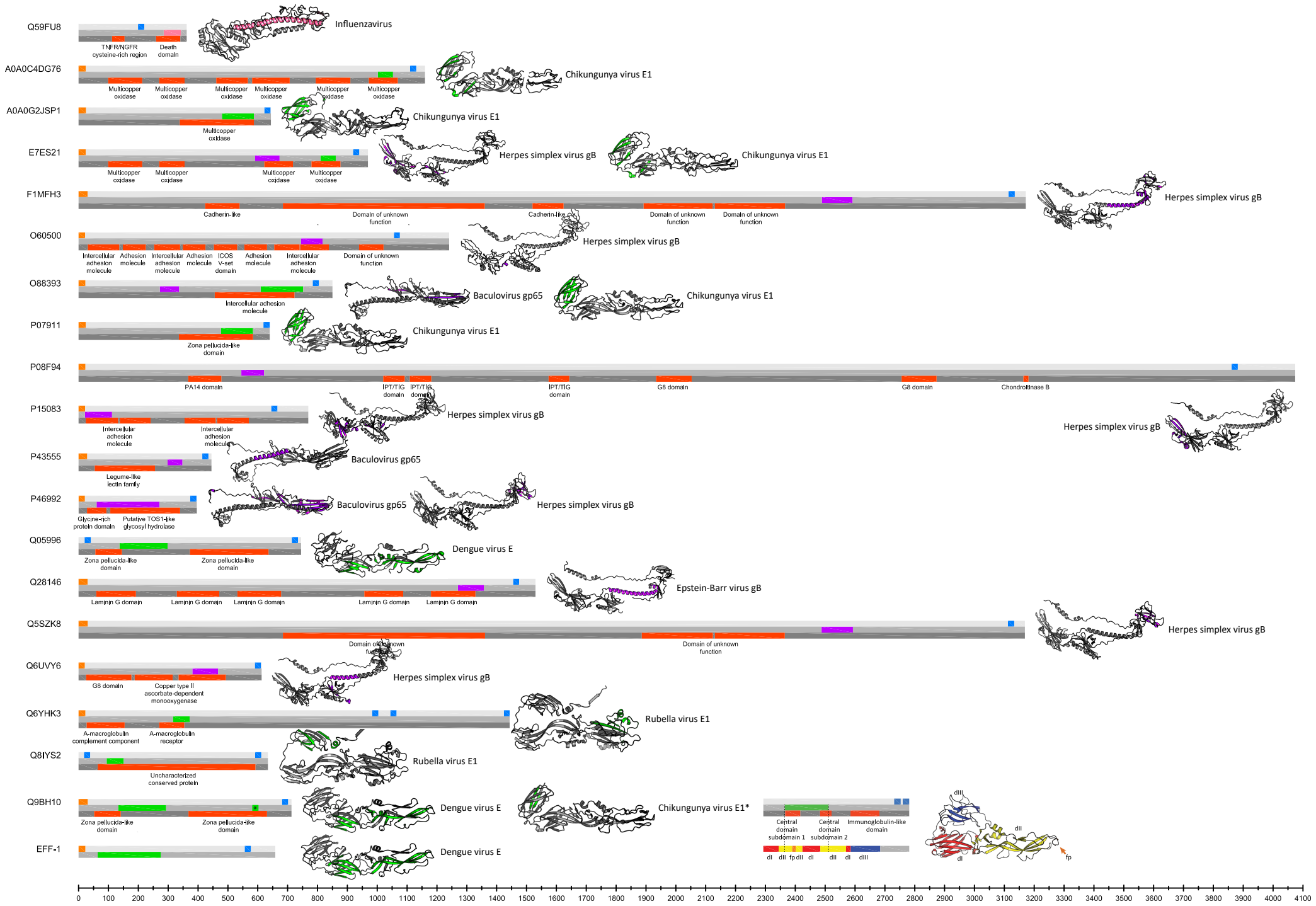

Q59FU8
TNFR/NGFR cysteine-rich region
Death domain
Influenzavirus
A0A0C4DG76
Multicopper oxidase
Chikungunya virus E1
A0A0G2JSP1
Chikungunya virus E1
E7ES21
Herpes simplex virus gB
Chikungunya virus E1
F1MFH3
Cadherin-like
Domain of unknown function
Herpes simplex virus gB
O60500
Intercellular adhesion molecule
Adhesion molecule
ICOS V-set domain
Herpes simplex virus gB
O88393
Intercellular adhesion molecule
Baculovirus gp65
Chikungunya virus E1
P07911
Zona pellucida-like domain
Chikungunya virus E1
P08F94
PA14 domain
IPT/TIG domain
G8 domain
Chondroitinase B
P15083
Herpes simplex virus gB
Herpes simplex virus gB
P43555
Legume-like lectin family
Baculovirus gp65
P46992
Glycine-rich protein domain
Putative TOS1-like glycosyl hydrolase
Baculovirus gp65
Herpes simplex virus gB
Q05996
Zona pellucida-like domain
Dengue virus E
Q28146
Laminin G domain
Epstein-Barr virus gB
Q5SZK8
Domain of unknown function
Herpes simplex virus gB
Q6UVY6
G8 domain
Copper type II ascorbate-dependent monooxygenase
Herpes simplex virus gB
Q6YHK3
A-macroglobulin complement component
A-macroglobulin receptor
Rubella virus E1
Q8IYS2
Uncharacterized conserved protein
Rubella virus E1
Q9BH10
Dengue virus E
Chikungunya virus E1*
Central domain subdomain 1
Central domain subdomain 2
Immunoglobulin-like domain
dI dII fp dII dI dII dI dIII
dIII
dII
dI
fp
EFF-1
Dengue virus E
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000 4100

Figura 32. Proteínas candidatas con similitud con fusógenos virales según HHsearch. Para cada proteína candidata se presentan tres líneas. La primera línea representa los péptidos señal (naranja) y dominios transmembrana (azul). La segunda línea ubica los dominios predichos con alta probabilidad según HHsearch (rojo). La última línea representa la región de similitud con un fusógeno viral mayor a 50 aminoácidos según HHsearch. Rosado representa un hit con un fusógeno viral de clase I, verde con un fusógeno viral de clase II y violeta con un fusógeno viral de clase III. A la derecha de cada proteína candidata se representa la estructura del fusógeno viral contra el que se encontró similitud y la región de similitud. La última proteína representada es EFF-1, donde se señala también la región de similitud con un fusógeno viral. A la derecha de esta, se representa la proteína E del virus del Dengue.

5.6 Discusión

La métrica desarrollada tomando como punto de partida la métrica propuesta por Przytycka et al. permitió cumplir con los objetivos de este trabajo. Sin embargo, teniendo en cuenta los resultados obtenidos para la clasificación de VFPs entrenando con una clase positiva es posible que el ajuste de la métrica hubiese permitido una mejor exactitud en la selección de proteínas de VEs con potencial fusogénico. La clasificación de VFPs entrenando con una clase positiva tenía como objetivo verificar si la adaptación del algoritmo a realizar alineamientos locales tenía un desempeño satisfactorio. Sin embargo los resultados no fueron satisfactorios. Es posible entonces, que en la etapa de clasificación de proteínas de VEs como potenciales proteínas fusogénicas, en caso de existir proteínas cuya actividad fusogénica esté reducida a una región de la misma, no se hayan clasificado positivamente.

Como ya se discutió, existen tres aspectos que podrían ser explorados en mayor detalle con el objetivo de mejorar la exactitud de la clasificación. El primero, el sistema de score. Se trabajó con el mismo sistema de score propuesto en la métrica original. En la bibliografía de identificaron modificaciones al sistema de score, permitiendo tomar como una única hélice (u hoja) dos hélices (o dos hojas) separadas por un loop. Sin embargo en ningún caso se dan detalles de dicha modificación del algoritmo. El segundo aspecto es la normalización. Dado que se modificó el algoritmo para realizar alineamientos locales se tuvo que proponer un sistema de normalización adecuado para que los scores sean comparables entre pares de proteínas. Para que el score tomara valores entre 0 y 1 se normalizó por el promedio de las longitudes de las regiones alineadas. Sería interesante proponer y evaluar los resultados de otro tipo de normalización, por ejemplo normalizar por la longitud de región alineada más extensa o más corta. Dada la complejidad del algoritmo no resulta intuitivo el resultado final. Finalmente, el tercer aspecto a explorar es la penalidad de gap. Se trabajó con un sistema de gap lineal, dada la simplicidad de su implementación. Sin embargo, sería interesante implementar un sistema de gap afín, donde se penalice más el inicio de un gap, que la extensión de uno existente. Resulta menos probable la existencia de varios gaps en un alineamiento que la existencia de uno sólo de mayor longitud.

Por otra parte, para la evaluación de la métrica se trabajó únicamente con dos tipos de clasificadores SVM y vecinos más cercanos. Dado que al momento de llevarse a cabo las actividades de esa sección no se había considerado la utilización de MDS, se tuvo que trabajar únicamente con aquellos clasificadores que permitieran una matriz de distancias como input.

Podrían obtenerse mejores resultados utilizando alguno de los clasificadores utilizados en el segundo objetivo.

Analizando en detalle los resultados de la clasificación de proteínas de VEs como potenciales fusógenos, se obtuvo una cantidad importante de proteínas candidatas. A pesar de que en conjunto se pudo comprobar que los clasificadores seleccionaron proteínas que tuvieran características similares en la composición de estructura secundaria con la que fueron entrenados, la mayoría de las candidatas no cumplen con características esenciales conocidas para las proteínas de fusión viral. Estas características están asociadas a su topología. Todos los fusógenos virales, independientemente de su clase, son proteínas de membrana de tipo I. Estas tienen una secuencia señal en su extremo N terminal que las dirige a su membrana blanco, ubicándose la porción N terminal en la matriz extracelular, y el dominio C terminal en el interior celular. Además, esta región extracelular tiene una longitud mínima de acuerdo a la clase a la que pertenece el fusógeno. Estos filtros redujeron el set inicial de proteínas clasificadas como positivas en un 70%, permaneciendo únicamente 262 proteínas candidatas.

Como último filtro se pretendió mantener únicamente aquellas proteínas para las cuales no se conoce una función molecular. No resulta trivial determinar con seguridad si una función asignada a una proteína corresponde a una función molecular definida, pero se trabajó con el curado bibliográfico y predictivo de UniProt para establecer la existencia o falta de funcionalidad de cada proteína candidata. Así, se seleccionaron 115 proteínas candidatas a las que se les hizo una predicción de dominios Pfam con el objetivo de caracterizarlas y proponerles una función molecular. Por otra parte, se caracterizó qué proteínas fueron clasificadas positivas. En este caso, estas proteínas ya no corren como candidatas al conocer su función, pero nos permitieron verificar que los clasificadores seleccionaron preferentemente proteínas cuyas características de estructura secundaria son similares. Los clasificadores de clase I seleccionaron preferentemente proteínas con dominios de unión a ADN, con patrones característicos de hélice, mientras que los de clase II seleccionaron preferentemente proteínas con dominios de unión a proteína, abundantes en hoja beta. En definitiva, los clasificadores seleccionaron como candidatas proteínas que siguen sus patrones de abundancia de estructura secundaria, aunque en su gran mayoría se pudo verificar que no son candidatos válidos, ya sea por sus características topológicas o por su función conocida.

Finalmente, a partir del análisis con HHsuite se pudo tener otra opinión respecto a las proteínas candidatas. La proteína más interesante resultó ser la proteína ZP2 de la zona pellucida. La zona pellucida está compuesta de tres o cuatro glicoproteínas (ZP1, ZP2 y ZP3) que forman una matriz filamentosa. De acuerdo a la bibliografía se cree que ZP2 podría actuar como un receptor secundario (Bleil, Greve, & Wassarman, 1988).

Se identificaron patrones en la proteína que sugieren que podría actuar como fusógeno. Presenta una longitud similar a los fusógenos virales de clase II y la homología se identificó en una región poco conservada (respecto al dominio inmunoglobulina) de la proteína. Otro punto en común presenta un dominio inmunoglobulina en la porción C-terminal del ectodominio. Análisis preliminares no presentados en este trabajo con la herramienta de predicción de la estructura terciaria I-TASSER (Yang et al., 2015) también encontraron similitud con fusógenos virales, aunque la predicción final de la estructura no fue la esperada. Sin embargo, la predicción de la estructura terciaria del fusógeno EFF-1 tampoco es predicha correctamente por I-TASSER.

Durante la fecundación, el espermatozoide inicialmente se une a la zona pellucida (ZP). La zona pellucida es una cubierta glicoproteica extracelular que rodea el ovocito de mamífero. Esta

matriz tiene un rol importante en la unión especie-específica y en la inducción de la reacción acrosómica (Gupta, Bansal, Ganguly, Bhandari, & Chakrabarti, 2009). También es importante en la prevención de la polispermia, que da lugar a la formación de embriones poliploides no viables, y en la protección del embrión hasta su implantación. Las proteínas de la zona pellucida son proteínas transmembrana que se clivan y se ensamblan para formar la ZP. De existir únicamente como proteína de la matriz no podría actuar como fusógeno de membranas. De todas formas, la proteína ZP2 es anclada a la membrana plasmática de la célula para luego clivarse su ectodominio y ser liberada a la matriz. Es posible que exista proteína intacta en la membrana en el momento de la fusión, aunque no ha sido reportado.

La proteína ZP2 se cliva nuevamente al unirse el espermatozoide al ovocito, previniendo la polispermia (Burkart, Xiong, Baibakov, Jimenez-Movilla, & Dean, 2012). Otros resultados preliminares no presentados en este trabajo muestran que la pérdida de este fragmento N-terminal da lugar a un ectodominio con características topológicas aún más similares a los fusógenos virales de clase II. Por su parte, el mismo procedimiento para prevenir la polispermia ha sido sugerido para la proteína HAP2 al unirse los gametos (Liu, Misamore, & Snell, 2010).

En resumen, este trabajo de tesis tuvo resultados satisfactorios tanto en el primer objetivo donde se desarrolló una métrica para alinear estructuras secundarias de proteínas, como en el segundo objetivo, donde se aplicó esta métrica para identificar proteínas candidatas a fusógenos de membranas. Se identificó una proteína candidata, ZP2, que cumple con todas las características topológicas buscadas, que se encuentra en el momento y lugar indicados para llevar a cabo la fusión de gametos, pero que presenta algunas evidencias experimentales en contra. Será necesario hacer una búsqueda bibliográfica más a fondo sobre ZP2 con el objetivo de diseñar experimentos que permitan validar este resultado. Además se proponen otras 17 proteínas candidatas, y se caracterizaron proteínas que no pasaron todos los filtros, pero para las cuales no se conocía su función molecular.

ANEXO

6. Especificaciones de funciones de Dd_tools

- Gauss_dd

La clase positiva se modela como una distribución Gaussiana. Se evita la estimación de la densidad y se utiliza solamente la distancia de Mahalanobis:

$$f(x) = (x-\mu)^T \Sigma^{-1}(x-\mu)$$

El clasificador se define como:

$$h(x) = \begin{cases} target\ if\ f(x) \leq \Theta \\ outlier\ if\ f(x) > \Theta \end{cases}$$

La media μ y la matriz de covarianza Σ son estimados de la muestra. El umbral Θ es determinado de acuerdo al error de la clase positiva dado por el usuario.

En el algoritmo de k-means implementado por dd_tools, para clasificar un nuevo elemento, se mide sus distancias a todos los prototipos y se promedian. Este valor se usa para determinar hasta qué punto es un outlier.

- Mog_dd

La clase positiva se modela utilizando una mezcla de k Gaussianas para crear una descripción más flexible. El modelo se define como:

$$f(x) = \sum_{i=1}^{K} P_i \exp(-(x-\mu_i)^T \sum_i^{-1} (x-\mu_i))$$

El clasificador se define como:

$$h(x) = \begin{cases} target\ if\ f(x) \geq \Theta \\ outlier\ if\ f(x) < \Theta \end{cases}$$

Los parámetros P_i, μ_i y Σ_i son optimizados usando el algoritmo de esperanza-maximización. Dado que en un espacio de alta dimensión y un número elevado de clusters K el número de parámetros libres puede ser enorme, las matrices de covarianza pueden ser restringidas.

- Parzen_dd

Es un modelo de densidad flexible con un kernel Gaussiano centrado en cada objeto de entrenamiento:

$$f(x) = \sum_{i=1}^{N} \exp(-(x-x_i)^T h^{-2}(x-x_i))$$

El parámetro libre h es optimizado maximizando la probabilidad sobre los datos de entrenamiento utilizando validación cruzada dejando uno fuera.

El clasificador se define como:

$$h(x) = \begin{cases} target\ if\ f(x) \geq \Theta \\ outlier\ if\ f(x) < \Theta \end{cases}$$

- Autoenc_dd

Una red neuronal es entrenada para reconstruir el patrón x del input en el output NeurN(x) de la red. La diferencia entre los patrones del input y del output es utilizada como la caracterización de la clase positiva. Esto resulta en:

$$f(x) = (x - NeurN(x))^2$$

El clasificador se define como:

$$h(x) = \begin{cases} target\ if\ f(x) \geq \Theta \\ outlier\ if\ f(x) < \Theta \end{cases}$$

- Kmeans_dd

Los datos son descriptos por k clusters ubicados de forma que la distancia promedio a un centro de cluster es minimizada. Los centros de cluster c_i son ubicados utilizando el procedimiento estándar de k-means (REF). La clase positiva es caracterizada por:

$$f(x) = \min_i (x - c_i)^2$$

El clasificador se define como:

$$h(x) = \begin{cases} target\ if\ f(x) \geq \Theta \\ outlier\ if\ f(x) < \Theta \end{cases}$$

- Mst_dd

Se ajusta un mínimum spanning tree a los datos de entrenamiento. La distancia a las aristas es utilizada como la similitud con la clase positiva.

- Knndd

En método del vecino más cercano, cada objeto es evaluado calculando la distancia a su vecino más cercano NN(x). Esta distancia es normalizada por la distancia del vecino más cercano de este objeto (es decir, la distancia entre el objeto NN(x) y NN(NN(x)). Este algoritmo se extiende a la distancia a sus k vecinos más cercanos.

- Svdd

Ajusta una hiperesfera alrededor de la clase positiva. Al introducir kernels, este modelo inflexible se vuelve mucho más poderoso, y puede dar excelentes resultados cuando se utiliza un kernel adecuado. Es posible optimizar el método para rechazar una fracción predefinida de datos. Esto significa que para distintas tasas de rechazo la forma de la barrera cambia. Esta implementación utiliza el kernel RBF.

- Lpdd

Está construido para describir objetos positivos representados en términos de distancias a otros objetos. El clasificador tiene la forma:

$$f(x) = \sum_i w_i d(x, x_i)$$

Los pesos w son optimizados de forma que solamente algunos pocos pesos son distintos a cero, y la barrera está lo más ajustada posible alrededor de los datos.

7. Figuras

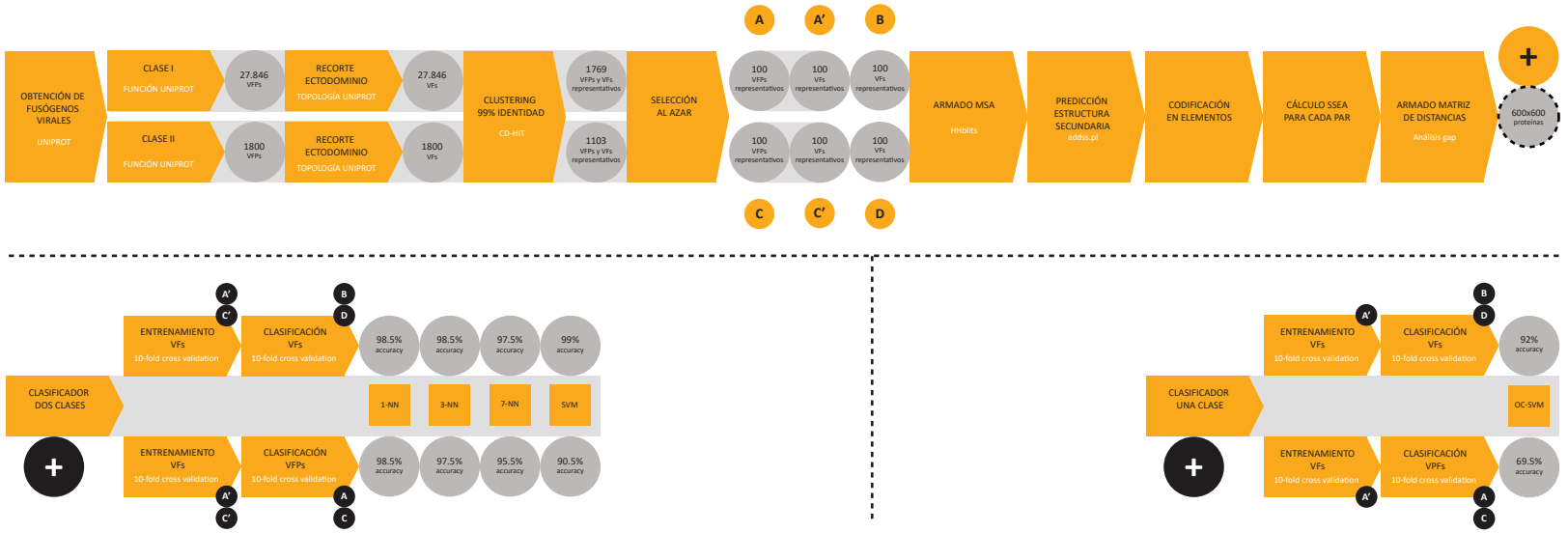


Figura A1. Diagrama de Objetivo 1

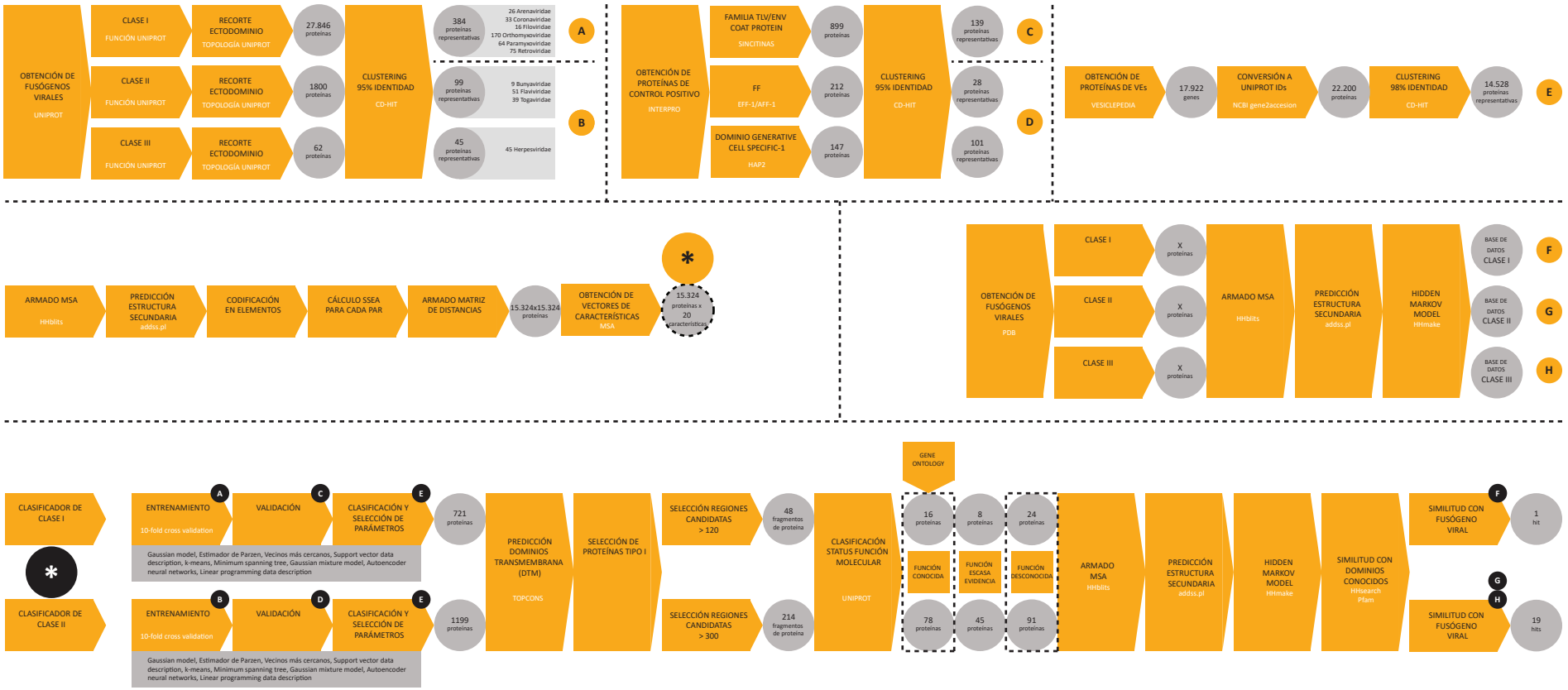


Figura A2. Diagrama de Objetivo 2

BIBLIOGRAFÍA

Aguilar, P. S., Baylies, M. K., Fleissner, A., Helming, L., Inoue, N., Podbilewicz, B., . . . Wong, M. (2013). Genetic basis of cell-cell fusion mechanisms. *Trends Genet, 29*(7), 427-437. doi:10.1016/j.tig.2013.01.011

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

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

Alvarez-Erviti, L., Seow, Y., Yin, H., Betts, C., Lakhal, S., & Wood, M. J. (2011). Delivery of siRNA to the mouse brain by systemic injection of targeted exosomes. *Nat Biotechnol, 29*(4), 341-345. doi:10.1038/nbt.1807

Andre, F., Schartz, N. E., Movassagh, M., Flament, C., Pautier, P., Morice, P., . . . Zitvogel, L. (2002). Malignant effusions and immunogenic tumour-derived exosomes. *Lancet, 360*(9329), 295-305. doi:10.1016/S0140-6736(02)09552-1

Avinoam, O., Fridman, K., Valansi, C., Abutbul, I., Zeev-Ben-Mordehai, T., Maurer, U. E., . . . Podbilewicz, B. (2011). Conserved eukaryotic fusogens can fuse viral envelopes to cells. *Science, 332*(6029), 589-592. doi:10.1126/science.1202333

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res, 28*(1), 235-242.

Bernsel, A., Viklund, H., Hennerdal, A., & Elofsson, A. (2009). TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res, 37*(Web Server issue), W465-468. doi:10.1093/nar/gkp363

Bi, P., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J. R., Shelton, J. M., . . . Olson, E. N. (2017). Control of muscle formation by the fusogenic micropeptide myomixer. *Science, 356*(6335), 323-327. doi:10.1126/science.aam9361

Bleil, J. D., Greve, J. M., & Wassarman, P. M. (1988). Identification of a secondary sperm receptor in the mouse egg zona pellucida: role in maintenance of binding of acrosome-reacted sperm to eggs. *Dev Biol, 128*(2), 376-385.

Booth, A. M., Fang, Y., Fallon, J. K., Yang, J. M., Hildreth, J. E., & Gould, S. J. (2006). Exosomes and HIV Gag bud from endosome-like domains of the T cell plasma membrane. *J Cell Biol, 172*(6), 923-935. doi:10.1083/jcb.200508014

Bouvier, N. M., & Palese, P. (2008). The biology of influenza viruses. *Vaccine, 26 Suppl 4*, D49-53.

Burkart, A. D., Xiong, B., Baibakov, B., Jimenez-Movilla, M., & Dean, J. (2012). Ovastacin, a cortical granule protease, cleaves ZP2 in the zona pellucida to prevent polyspermy. *J Cell Biol, 197*(1), 37-44. doi:10.1083/jcb.201112094

Burri, L., Varlamov, O., Doege, C. A., Hofmann, K., Beilharz, T., Rothman, J. E., . . . Lithgow, T. (2003). A SNARE required for retrograde transport to the endoplasmic reticulum. *Proc Natl Acad Sci U S A, 100*(17), 9873-9877. doi:10.1073/pnas.1734000100

Burton, G. J., & Fowden, A. L. (2015). The placenta: a multifaceted, transient organ. *Philos Trans R Soc Lond B Biol Sci, 370*(1663), 20140066. doi:10.1098/rstb.2014.0066

Caby, M. P., Lankar, D., Vincendeau-Scherrer, C., Raposo, G., & Bonnerot, C. (2005). Exosomal-like vesicles are present in human blood plasma. *Int Immunol, 17*(7), 879-887. doi:10.1093/intimm/dxh267

Campbell, C. (2001). A Linear Programming Approach to Novelty Detection. *NIPS*.

Chalmin, F., Ladoire, S., Mignot, G., Vincent, J., Bruchard, M., Remy-Martin, J. P., . . . Ghiringhelli, F. (2010). Membrane-associated Hsp72 from tumor-derived exosomes mediates STAT3-

dependent immunosuppressive function of mouse and human myeloid-derived suppressor cells. *J Clin Invest, 120*(2), 457-471. doi:10.1172/JCI40483

Chan, D. C. (2006). Mitochondrial fusion and fission in mammals. *Annu Rev Cell Dev Biol, 22*, 79-99. doi:10.1146/annurev.cellbio.22.010305.104638

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology, 2*.

Chen, Y. A., & Scheller, R. H. (2001). SNARE-mediated membrane fusion. *Nat Rev Mol Cell Biol, 2*(2), 98-106. doi:10.1038/35052017

Chernomordik, L. V., & Kozlov, M. M. (2008). Mechanics of membrane fusion. *Nat Struct Mol Biol, 15*(7), 675-683. doi:10.1038/nsmb.1455

Ciechonska, M., Key, T., & Duncan, R. (2014). Efficient reovirus- and measles virus-mediated pore expansion during syncytium formation is dependent on annexin A1 and intracellular calcium. *J Virol, 88*(11), 6137-6147. doi:10.1128/JVI.00121-14

Cornelis, G., Vernochet, C., Carradec, Q., Souquere, S., Mulot, B., Catzeflis, F., . . . Heidmann, T. (2015). Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc Natl Acad Sci U S A, 112*(5), E487-496. doi:10.1073/pnas.1417000112

Dilcher, M., Veith, B., Chidambaram, S., Hartmann, E., Schmitt, H. D., & Fischer von Mollard, G. (2003). Use1p is a yeast SNARE protein required for retrograde traffic to the ER. *EMBO J, 22*(14), 3664-3674. doi:10.1093/emboj/cdg339

Doms, R. W. (2017). What Came First - the Virus or the Egg? *Cell, 168*(5), 755-757. doi:10.1016/j.cell.2017.02.012

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Duin, R. P. W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D. M. J., & Verzakov, S. (2007). PRTools4, A Matlab Toolbox for Pattern Recognition.

Fang, Y., Wu, N., Gan, X., Yan, W., Morrell, J. C., & Gould, S. J. (2007). Higher-order oligomerization targets plasma membrane proteins and HIV gag to exosomes. *PLoS Biol, 5*(6), e158. doi:10.1371/journal.pbio.0050158

Fasshauer, D. (2003). Structural insights into the SNARE mechanism. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1641*(2-3), 87-97. doi:10.1016/s0167-4889(03)00090-9

Fedry, J., Liu, Y., Pehau-Arnaudet, G., Pei, J., Li, W., Tortorici, M. A., . . . Krey, T. (2017). The Ancient Gamete Fusogen HAP2 Is a Eukaryotic Class II Fusion Protein. *Cell, 168*(5), 904-915 e910. doi:10.1016/j.cell.2017.01.024

Figueiredo, M. A. T. (2002). Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(3), 381-396.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., . . . Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res, 45*(D1), D190-D199. doi:10.1093/nar/gkw1107

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., . . . Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res, 44*(D1), D279-285. doi:10.1093/nar/gkv1344

Fontana, P., Bindewald, E., Toppo, S., Velasco, R., Valle, G., & Tosatto, S. C. (2005). The SSEA server for protein secondary structure alignment. *Bioinformatics, 21*(3), 393-395. doi:10.1093/bioinformatics/bti013

Gong, R., Peng, X., Kang, S., Feng, H., Huang, J., Zhang, W., . . . Xiao, G. (2005). Structural characterization of the fusion core in syncytin, envelope protein of human endogenous retrovirus family W. *Biochem Biophys Res Commun, 331*(4), 1193-1200. doi:10.1016/j.bbrc.2005.04.032

Graham, R. L., & Hell, P. (1985). On the History of the Minimum Spanning Tree Problem. *Annals of the History of Computing, 7*(1), 43-57.

Griparic, L., van der Wel, N. N., Orozco, I. J., Peters, P. J., & van der Bliek, A. M. (2004). Loss of the intermembrane space protein Mgm1/OPA1 induces swelling and localized

constrictions along the lengths of mitochondria. *J Biol Chem, 279*(18), 18792-18798. doi:10.1074/jbc.M400920200

Gupta, S. K., Bansal, P., Ganguly, A., Bhandari, B., & Chakrabarti, K. (2009). Human zona pellucida glycoproteins: functional relevance during fertilization. *J Reprod Immunol, 83*(1-2), 50-55. doi:10.1016/j.jri.2009.07.008

Harrison, S. C. (2008). Viral membrane fusion. *Nat Struct Mol Biol, 15*(7), 690-698. doi:10.1038/nsmb.1456

Harrison, S. C. (2015). Viral membrane fusion. *Virology, 479-480*, 498-507. doi:10.1016/j.virol.2015.03.043

Hristov, M., Erl, W., Linder, S., & Weber, P. C. (2004). Apoptotic bodies from endothelial cells enhance the number and initiate the differentiation of human endothelial progenitor cells in vitro. *Blood, 104*(9), 2761-2766. doi:10.1182/blood-2003-10-3614

Ihrie, R. A., Marques, M. R., Nguyen, B. T., Horner, J. S., Papazoglu, C., Bronson, R. T., . . . Attardi, L. D. (2005). Perp is a p63-regulated gene essential for epithelial integrity. *Cell, 120*(6), 843-856. doi:10.1016/j.cell.2005.01.008

Ishihara, N., Eura, Y., & Mihara, K. (2004). Mitofusin 1 and 2 play distinct roles in mitochondrial fusion reactions via GTPase activity. *J Cell Sci, 117*(Pt 26), 6535-6546. doi:10.1242/jcs.01565

Ishii, M., & Saeki, Y. (2008). Osteoclast cell fusion: mechanisms and molecules. *Mod Rheumatol, 18*(3), 220-227. doi:10.1007/s10165-008-0051-2

Itakura, E., Kishi-Itakura, C., & Mizushima, N. (2012). The hairpin-type tail-anchored SNARE syntaxin 17 targets to autophagosomes for fusion with endosomes/lysosomes. *Cell, 151*(6), 1256-1269. doi:10.1016/j.cell.2012.11.001

Jahn, R., & Scheller, R. H. (2006). SNAREs--engines for membrane fusion. *Nat Rev Mol Cell Biol, 7*(9), 631-643. doi:10.1038/nrm2002

Jahn, R., & Sudhof, T. C. (1999). Membrane fusion and exocytosis. *Annu Rev Biochem, 68*, 863-911. doi:10.1146/annurev.biochem.68.1.863

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol, 292*(2), 195-202. doi:10.1006/jmbi.1999.3091

Kalra, H., Simpson, R. J., Ji, H., Aikawa, E., Altevogt, P., Askenase, P., . . . Mathivanan, S. (2012). Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation. *PLoS Biol, 10*(12), e1001450. doi:10.1371/journal.pbio.1001450

Keller, S., Sanderson, M. P., Stoeck, A., & Altevogt, P. (2006). Exosomes: from biogenesis and secretion to biological function. *Immunol Lett, 107*(2), 102-108. doi:10.1016/j.imlet.2006.09.005

Kerr, J. F., Wyllie, A. H., & Currie, A. R. (1972). Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br J Cancer, 26*(4), 239-257.

Kielian, M. (2006). Class II virus membrane fusion proteins. *Virology, 344*(1), 38-47. doi:10.1016/j.virol.2005.09.036

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1-27.

Lenassi, M., Cagney, G., Liao, M., Vaupotic, T., Bartholomeeusen, K., Cheng, Y., . . . Peterlin, B. M. (2010). HIV Nef is secreted in exosomes and triggers apoptosis in bystander CD4+ T cells. *Traffic, 11*(1), 110-122. doi:10.1111/j.1600-0854.2009.01006.x

Lewis, M. J., & Pelham, H. R. (1992). Sequence of a second human KDEL receptor. *J Mol Biol, 226*(4), 913-916.

Liu, Y., Misamore, M. J., & Snell, W. J. (2010). Membrane fusion triggers rapid degradation of two gamete-specific, fusion-essential proteins in a membrane block to polygamy in Chlamydomonas. *Development, 137*(9), 1473-1481. doi:10.1242/dev.044743

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory, IT-28*(2), 129-137.

Marzesco, A. M., Janich, P., Wilsch-Brauninger, M., Dubreuil, V., Langenfeld, K., Corbeil, D., & Huttner, W. B. (2005). Release of extracellular membrane particles carrying the stem cell marker prominin-1 (CD133) from neural progenitors and other epithelial cells. *J Cell Sci, 118*(Pt 13), 2849-2858. doi:10.1242/jcs.02439

Mathivanan, S., Ji, H., & Simpson, R. J. (2010). Exosomes: extracellular organelles important in intercellular communication. *J Proteomics, 73*(10), 1907-1920. doi:10.1016/j.jprot.2010.06.006

McGuffin, L. J., & Jones, D. T. (2002). Targeting novel folds for structural genomics. *Proteins, 48*(1), 44-52. doi:10.1002/prot.10129

Megrian, D., Aguilar, P. S., & Lecumberry, F. (2017). Similarity measure for cell membrane fusion proteins identification. *Lecture Notes in Computer Science, 10125*, 257-265.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., & Thomas, P. D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res, 38*(Database issue), D204-210. doi:10.1093/nar/gkp1019

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., . . . McCoy, J. M. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature, 403*(6771), 785-789. doi:10.1038/35001608

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol, 48*(3), 443-453.

Ni, Q., & Zou, L. (2014). Accurate discrimination of outer membrane proteins using secondary structure element alignment and support vector machine. *J Bioinform Comput Biol, 12*(1), 1450003. doi:10.1142/S0219720014500036

Orso, G., Pendin, D., Liu, S., Tosetto, J., Moss, T. J., Faust, J. E., . . . Daga, A. (2009). Homotypic fusion of ER membranes requires the dynamin-like GTPase atlastin. *Nature, 460*(7258), 978-983. doi:10.1038/nature08280

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A, 37*(4), 205-211.

Pawelek, J. M., & Chakraborty, A. K. (2008). Chapter 10 The Cancer Cell–Leukocyte Fusion Theory of Metastasis. *101*, 397-444. doi:10.1016/s0065-230x(08)00410-7

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A, 85*(8), 2444-2448.

Perez-Vargas, J., Krey, T., Valansi, C., Avinoam, O., Haouz, A., Jamin, M., . . . Rey, F. A. (2014). Structural basis of eukaryotic cell-cell fusion. *Cell, 157*(2), 407-419. doi:10.1016/j.cell.2014.02.020

Peterson, D. B., Sander, T., Kaul, S., Wakim, B. T., Halligan, B., Twigger, S., . . . Ou, J. S. (2008). Comparative proteomic analysis of PAI-1 and TNF-alpha-derived endothelial microparticles. *Proteomics, 8*(12), 2430-2446. doi:10.1002/pmic.200701029

Pevsner, J. (2015). *Bioinformatics and functional genomics* (Third edition. ed.). Chichester, West Sussex, UK ; Hoboken, NJ, USA: John Wiley and Sons, Inc.

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing, 99*, 215-249. doi:10.1016/j.sigpro.2013.12.026

Pisitkun, T., Shen, R. F., & Knepper, M. A. (2004). Identification and proteomic profiling of exosomes in human urine. *Proc Natl Acad Sci U S A, 101*(36), 13368-13373. doi:10.1073/pnas.0403453101

Prada, I., & Meldolesi, J. (2016). Binding and Fusion of Extracellular Vesicles to the Plasma Membrane of Their Cell Targets. *Int J Mol Sci, 17*(8). doi:10.3390/ijms17081296

Primakoff, P., & Myles, D. G. (2007). Cell-cell membrane fusion during mammalian fertilization. *FEBS Lett, 581*(11), 2174-2180. doi:10.1016/j.febslet.2007.02.021

Przytycka, T., Aurora, R., & Rose, G. D. (1999). A protein taxonomy based on secondary structure. *Nat Struct Biol, 6*(7), 672-682. doi:10.1038/10728

Rand, R. P., & Parsegian, V. A. (1984). Physical force considerations in model and biological membranes. *Can J Biochem Cell Biol, 62*(8), 752-759.

Reinhardt, T. A., Lippolis, J. D., Nonnecke, B. J., & Sacco, R. E. (2012). Bovine milk exosome proteome. *J Proteomics, 75*(5), 1486-1492. doi:10.1016/j.jprot.2011.11.017

Reynolds, S. M., Kall, L., Riffle, M. E., Bilmes, J. A., & Noble, W. S. (2008). Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol, 4*(11), e1000213. doi:10.1371/journal.pcbi.1000213

Rochlin, K., Yu, S., Roy, S., & Baylies, M. K. (2010). Myoblast fusion: when it takes more to make one. *Dev Biol, 341*(1), 66-83. doi:10.1016/j.ydbio.2009.10.024

Santel, A., & Fuller, M. T. (2001). Control of mitochondrial morphology by a human mitofusin. *J Cell Sci, 114*(Pt 5), 867-874.

Schaeffer, C., Santambrogio, S., Perucca, S., Casari, G., & Rampoldi, L. (2009). Analysis of uromodulin polymerization provides new insights into the mechanisms regulating ZP domain-mediated protein assembly. *Mol Biol Cell, 20*(2), 589-599. doi:10.1091/mbc.E08-08-0876

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., & Platt, J. (2000). Support Vector Method for Novelty Detection. *MIT Press*, 582-588.

Sedgewick, R. (1988). *Algorithms* (2nd ed.). Reading, Mass.: Addison-Wesley.

Si, J. N., Yan, R. X., Wang, C., Zhang, Z., & Su, X. D. (2009). TIM-Finder: a new method for identifying TIM-barrel proteins. *BMC Struct Biol, 9*, 73. doi:10.1186/1472-6807-9-73

Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., . . . Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform, 3*(3), 265-274.

Simons, M., & Raposo, G. (2009). Exosomes--vesicular carriers for intercellular communication. *Curr Opin Cell Biol, 21*(4), 575-581. doi:10.1016/j.ceb.2009.03.007

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol, 147*(1), 195-197.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics, 21*(7), 951-960. doi:10.1093/bioinformatics/bti125

Sollner, T., Bennett, M. K., Whiteheart, S. W., Scheller, R. H., & Rothman, J. E. (1993). A protein assembly-disassembly pathway in vitro that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion. *Cell, 75*(3), 409-418.

Tax, D. M. J. (2015). DDtools, the Data Description Toolbox for Matlab.

Tax, D. M. J., & Duin, R. P. W. (2004). Support Vector Data Description. *Machine Learning, 54*, 45-66.

The UniProt, C. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res, 45*(D1), D158-D169. doi:10.1093/nar/gkw1099

Thery, C., Boussac, M., Veron, P., Ricciardi-Castagnoli, P., Raposo, G., Garin, J., & Amigorena, S. (2001). Proteomic analysis of dendritic cell-derived exosomes: a secreted subcellular compartment distinct from apoptotic vesicles. *J Immunol, 166*(12), 7309-7318.

Thery, C., Ostrowski, M., & Segura, E. (2009). Membrane vesicles as conveyors of immune responses. *Nat Rev Immunol, 9*(8), 581-593. doi:10.1038/nri2567

Turiak, L., Misjak, P., Szabo, T. G., Aradi, B., Paloczi, K., Ozohanics, O., . . . Vekey, K. (2011). Proteomic characterization of thymocyte-derived microvesicles and apoptotic bodies in BALB/c mice. *J Proteomics, 74*(10), 2025-2033. doi:10.1016/j.jprot.2011.05.023

Valansi, C., Moi, D., Leikina, E., Matveev, E., Grana, M., Chernomordik, L. V., . . . Podbilewicz, B. (2017). Arabidopsis HAP2/GCS1 is a gamete fusion protein homologous to somatic and viral fusogens. *J Cell Biol, 216*(3), 571-581. doi:10.1083/jcb.201610093

van den Boorn, J. G., Schlee, M., Coch, C., & Hartmann, G. (2011). SiRNA delivery with exosome nanoparticles. *Nat Biotechnol, 29*(4), 325-326. doi:10.1038/nbt.1830

van der Pol, E., Boing, A. N., Harrison, P., Sturk, A., & Nieuwland, R. (2012). Classification, functions, and clinical relevance of extracellular vesicles. *Pharmacol Rev, 64*(3), 676-705. doi:10.1124/pr.112.005983

Vella, L. J., Greenwood, D. L., Cappai, R., Scheerlinck, J. P., & Hill, A. F. (2008). Enrichment of prion protein in exosomes derived from ovine cerebral spinal fluid. *Vet Immunol Immunopathol, 124*(3-4), 385-393. doi:10.1016/j.vetimm.2008.04.002

Wang, X. F., Lin, H. Y., Ng-Eaton, E., Downward, J., Lodish, H. F., & Weinberg, R. A. (1991). Expression cloning and characterization of the TGF-beta type III receptor. *Cell, 67*(4), 797-805.

Willkomm, L., & Bloch, W. (2015). State of the art in cell-cell fusion. *Methods Mol Biol, 1313*, 1-19. doi:10.1007/978-1-4939-2703-6_1

Wubbolts, R., Leckie, R. S., Veenhuizen, P. T., Schwarzmann, G., Mobius, W., Hoernschemeyer, J., . . . Stoorvogel, W. (2003). Proteomic and biochemical analyses of human B cell-derived exosomes. Potential implications for their function and multivesicular body formation. *J Biol Chem, 278*(13), 10963-10972. doi:10.1074/jbc.M207550200

Xu, H., Aurora, R., Rose, G. D., & White, R. H. (1999). Identifying two ancient enzymes in Archaea using predicted secondary structure alignment. *Nat Struct Biol, 6*(8), 750-754. doi:10.1038/11525

Xu, Y., Rahman, N. A., Othman, R., Hu, P., & Huang, M. (2012). Computational identification of self-inhibitory peptides from envelope proteins. *Proteins, 80*(9), 2154-2168. doi:10.1002/prot.24105

Yan, R. X., Chen, Z., & Zhang, Z. (2011). Outer membrane proteins can be simply identified using secondary structure element alignment. *BMC Bioinformatics, 12*, 76. doi:10.1186/1471-2105-12-76

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat Methods, 12*(1), 7-8. doi:10.1038/nmeth.3213

Zhang, Z., Kochhar, S., & Grigorov, M. G. (2005). Descriptor-based protein remote homology identification. *Protein Sci, 14*(2), 431-444. doi:10.1110/ps.041035505

Zitvogel, L., Regnault, A., Lozier, A., Wolfers, J., Flament, C., Tenza, D., . . . Amigorena, S. (1998). Eradication of established murine tumors using a novel cell-free vaccine: dendritic cell-derived exosomes. *Nat Med, 4*(5), 594-600.