Exploración del espacio de secuencias de los receptores de células T

Trabajo Práctico Final - Redes Complejas

Heli Magalí García Álvarez, Juan Ignacio Gossn, Santiago Scheiner

Resumen

Los linfocitos T y B, componentes fundamentales del sistema inmune adaptativo, son los responsables de reconocer a las células infectadas por patógenos en el organismo. En particular, hace varias décadas que se conoce que los linfocitos T reconocen pequeños trozos de péptidos, también llamados epítopos o antígenos, provenientes de proteínas degradadas por las células y presentadas en los complejos mayores de histocompatibilidad (MHC) en la superficie de las mismas. Los linfocitos T interactúan mediante sus receptores (TCR) con las células presentadoras de antígenos y, por ende, las propiedades de los mismos determinarán su afinidad por distintos péptidos o epítopos presentados. En este trabajo, se estudiaron las características topológicas de diferentes redes de receptores de células T (TCR), cuyos nodos corresponden a las secuencias aminoacídicas de la porción CDR3β de los receptores. Este conjunto de secuencias TCR elegido está asociado, a su vez, a diferentes epítopos correspondientes a 6 virus que infectan a los seres humanos. Para determinar las existencia o no de enlaces entre distintas secuencias CDR3β, los nodos de la red, se consideraron diversas métricas ya descritas en trabajos anteriores y se establecieron umbrales fijos a partir de los cuales se consideró la existencia, o no, de enlace entre secuencias. Para la totalidad de las redes construidas mediante las distintas métricas de similitud o distancia para los TCR, se evaluó la existencia de grupos o *clusters* de secuencias correspondientes a un mismo "tipo" de epítopos, considerando "tipo" como especie de virus asociada, gen del epítopo o epítopo único. Se halló que todas las redes estudiadas, armadas considerando las distintas métricas entre receptores T, poseen grupos de nodos pequeños de alta pureza vinculadas a epítopos específicos; aunque también contienen grupos asociados a nivel de genes epitópicos y especies de virus, pero con menor pureza que en el primer caso. Dadas las medidas consideradas para evaluar los algoritmos de particiones utilizados (InfoMap y Louvain), no se dispone aún de un criterio concluyente que establezca cuál de las métricas, ni cuál de los algoritmos de *clustering* empleados, se desempeña mejor en diferenciar las secuencias aminoacídicas CDR3β por alguno de los atributos de epítopos mencionados.

1. Introducción

El cuerpo humano se encuentra expuesto constantemente a patógenos externos, tales como bacterias y virus, para los que el sistema inmune debe articular una compleja red de interacción entre células con funciones de diversa complejidad [2]. El sistema inmune emplea un amplio repertorio de moléculas, células y órganos que actúan en conjunto con el fin de mantener sano al organismo. Al igual que en una red social, las células que componen el sistema inmune ejercen sus funciones de manera efectiva gracias a una colaboración y comunicación adecuadas [11]. Un subconjunto esencial de células pertenecientes al sistema inmune son las células (o linfocitos) T, que son producidas en la médula ósea, maduran en el timo, cuyas funciones son parte importante del sistema inmunitario adaptativo. Esencialmente, son las responsables de la inmunidad celular destruyendo células infectadas o activando macrófagos, linfocitos B u otros linfocitos T mediante citocinas y otras proteínas coestimulatorias que se encuentran en su membrana celular, llamadas receptores de células T"(TCR) (figura 1).

Es importante mencionar que la región CDR3 de las cadenas α y β del receptor T (Fig 1) constituyen una región nucleotídica muy variable y, por ende, única y específica para un dado clonotipo de célula T y toda su progenie. Más aún, la región CDR3 de estos receptores es aquella porción principalmente involucrada en interacciones con antígenos procesados de manera intracelular y expuestos por células presentadoras en los complejos mayores de histocompatibilidad (MHC). La interacción entre el receptor de célula T y el complejo péptido-MHC se ilustra en la figura 2.

En este trabajo se propuso representar al espacio de secuencias de CDR3β mediante un grafo. En particular, analizamos secuencias de la tercera región determinante de complementariedad (CDR3) de

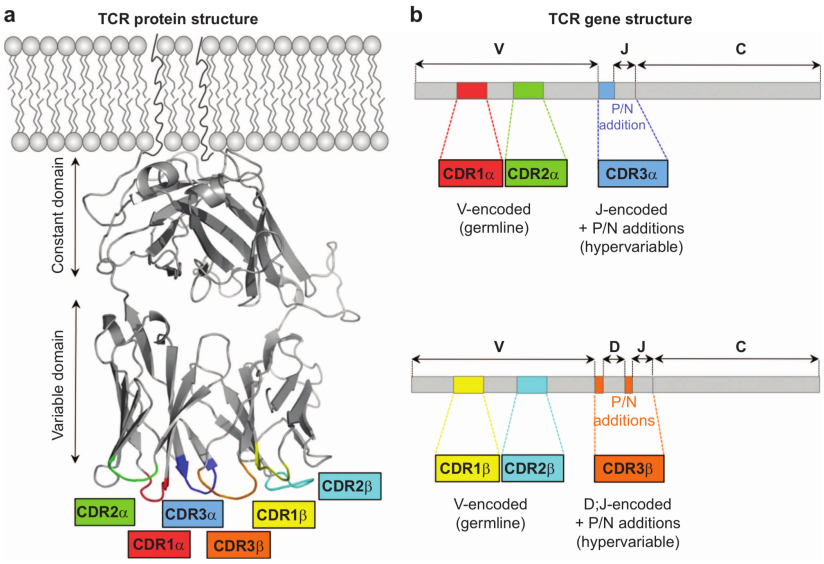


Figura 1: Representación del receptor de célula T (a) y su correspondiente estructura génica (b). La región de nuestro interés es la CDR3β. Figura extraída de Attaf *et al.*, 2015 [1].

la cadena β del receptor T (1). En nuestra representación, los nodos son secuencias únicas de CDR3β y los enlaces se establecen a partir de la similitud entre dichas secuencias. Para definir los enlaces se utilizan distintas métricas para establecer el grado de similitud entre las secuencias seleccionadas. Se seleccionó un subconjunto de secuencias correspondientes a receptores que interactúan con epítopos de 6 virus de humano conocidos: virus de Fiebre Amarilla, Citomegalovirus, virus de Epstein-Barr, VIH-1, virus de Influenza A y virus de la Hepatitis C. En términos generales, el objetivo del trabajo consistió en el estudio de la estructura de las redes construidas para las secuencias de los receptores T, así como también para las secuencias de sus epítopos asociados. Para alcanzar dicho objetivo, se realizaron diferentes análisis sobre las propiedades de las redes. En primer lugar, resultó razonable preguntarse si las secuencias CDR3β asociadas a los mismos antígenos se encuentran más cercanas en el grafo y cómo se agrupan las mismas en comunidades dentro de la red. En esta misma línea, cabe también observar si las comunidades definidas por un algoritmo de *clustering* están enriquecidas en CDR3β correspondientes a un mismo epítopo. Asimismo, se analizó si las secuencias de CDR3β que reconocen a péptidos de un mismo virus se aglomeran por esta característica más general o, en el mismo sentido, si se agrupan debido a que reconocen péptidos codificados por un mismo gen de un virus. En síntesis, es intentó responder con qué criterio se agrupan en comunidades las secuencias de CDR3β estudiadas y si esto tiene un correlato a lo observado en la red de secuencias de sus epítopos asociados.

2. Métodos

2.1. Base de datos

Las distintas secuencias de aminoácidos de CDR3β y la información del epítopo al cual se vinculan fue extraída de la base de datos curada VDJdb [14]. Tal como se mencionó en la introducción se seleccionaron secuencias CDR3β con especificidad por antígenos de 6 virus de humano conocidos. En este trabajo resultaron relevantes la información sobre la especie del virus, el gen que codifica al epítopo del virus y la secuencias aminoacídicas tanto de CDR3β como de sus epítopos asociados. Cabe destacar que las secuencias CDR3β elegidas de VDJdb fueron filtradas por dos criterios más:

1. Se consideraron secuencias con *Confidence score* $\geq$ *0*, es decir que se seleccionaron aquellas secuencias para las cuales hay algún grado de confianza positivo en cuanto a la especificidad antigénica del CDR3β anotado, y

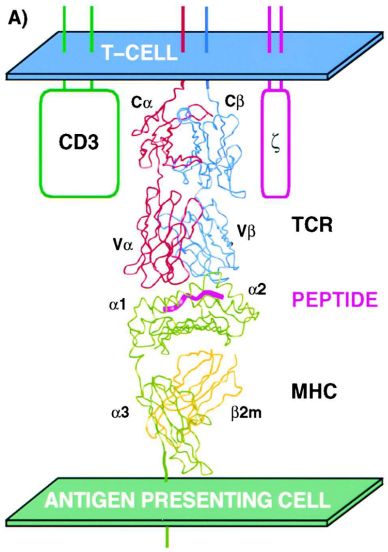


Figura 2: Representación de la interacción entre el receptor de célula T y el complejo de mayor histocompatibilidad (MHC) cargado con un epítopo (o péptido). Figura extraída de Hennecke *et al.*, 2001 [7].

2. Se consideraron únicamente secuencias cuyos epítopos se presentan en MHC de clase I.

La restricción 2) se debe a que las vías de procesamiento y presentación de péptidos difieren en gran medida entre MHC de clase I y MHC de clase II. Debido a que los virus inducen la expresión intracelular de proteínas virales, los fragmentos de proteínas resultantes del proceso de degradación de las mismas en el proteasoma tienen la potencialidad de unirse al MHC de clase I. Ya que se han seleccionado secuencias de CDR3β que se unen a péptidos de virus, los cuales en su mayoría se presentan por la vía del MHC de clase I, se ha aplicado el filtro 2) para generar el set de datos a analizar.

2.2. Métricas consideradas para establecer una distancia entre secuencias

Las métricas consideradas se hallan descritas en los trabajos de Meysman *et al.* (2018) [10], Shen *et al.* (2014) [12] y Kim *et al.* (2009) [9]. También se han empleado métodos tradicionales como el alineamiento local de secuencias por el algoritmo de Smith-Waterman [13], para evaluar la similitud entre secuencias de CDR3β o epítopos. En la Tab. 1 se encunetran listadas las sucesivas redes construidas. Las primeras 7, de T1 a T7, corresponden a redes en las cuales los nodos son CDR3β, mientras que las últimas dos corresponden a redes cuyos nodos son epítopos, E1 y E2. Para cada una de ellas se utilizó una métrica distinta para evaluar la distancia entre secuencias, lo cual luego se empleó para constituir los enlaces de la red.

ID	**Tipo de nodos**	**Métrica**
T1	CDR3β	Alignment Score [BLOSUM 62]
T2	CDR3β	GapAlign Score
T3	CDR3β	Profile Score
T4	CDR3β	Trimer Score
T5	CDR3β	Dimer Score
T6	CDR3β	Levenshtein Distance Score
T7	CDR3β	Kernel Distance Score
E1	Epítopo	Alignment Score [BLOSUM 62]
E2	Epítopo	Alignment Score [PMBEC o "pMHCMatrix"]

Cuadro 1: Listado de redes construidas con sus métricas asociadas.

Las métricas de T1, E1 y E2 consisten en el uso del algoritmo de Smith-Waterman para el alinea-

miento local de secuencias de a pares. Las matrices de sustitución utilizadas para asignar puntajes a los *matches* o *mis-matches* entre aminoácidos fueron la matriz BLOSUM 62 [6] y la matriz PMBEC [9]. Cabe destacar que se utilizaron valores de *Gap Opening Penalty* de -10 y *Gap Extension Penalty* de -1 para el caso de T1 y E1 (BLOSUM 62); y valores de *Gap Opening Penalty* de -1 y *Gap Extension Penalty* de -0.1 para el caso E2 (PMBEC). En ambos casos, más allá de las diferencias de escala entre las matrices, se puede observar que se penalizó fuertemente la apertura de un *gap*, siguiendo la línea de lo propuesto por Dash *et al.*, 2017 [3] en su definición de la *TCRdist*.

La métrica GapAlign empleada en T2 consiste en una adaptación propuesta por Meysman *et al.*, 2018, de la métrica *TCRdist* de Dash *et al.*, 2017 [3], restringida al esquema de puntajes para secuencias de CDR3β. En breve, usa un alineamiento de secuencias permitiendo *gaps* únicos, y una matriz de sustitución de aminoácidos derivada de la matriz BLOSUM 90.

La métrica propuesta para T3, Profile Score, se basa en las diferencias físico-químicas entre dos secuencias CDR3β, derivada del enfoque propuesto por De Neuter *et al.*, 2018 [4]. El perfil de una secuencia CDR3β se constituye a partir de la basicidad, hidrofobicidad y helicidad de dicha secuencia aminoacídica.

En cuanto a las métricas de T4 y T5, Trimer y Dimer Score, las mismas consisten en el cálculo del porcentaje de dímeros y trímeros compartidos entre dos secuencias de CDR3β, tal como se desarrolla en Glanville *et al.*, 2017 [5].

Por otro lado, la distancia de Levenshtein para T6 consiste en la distancia de edición entre dos secuencias CDR3β, lo que significa que cuantifica el número mínimo de cambios (mutaciones, supresiones e inserciones) que deben realizarse sobre una de las secuencias para obtener la otra.

Por último la métrica propuesta para T7 es la llamada Distance Kernel, desarrollada por Shen *et al.*, 2014 [12]. Para medir la similitud entre secuencias CDR3β se define un kernel sobre *strings* o k-meros (de 1 a 30 aminoácidos de longitud en este caso) de las secuencias utilizando la estructura de las mismas y una matriz de sustitución de aminoácidos (BLOSUM 62-2).

2.3. Criterio de definición del enlace

A partir de los puntajes computados para cada uno de todos los posibles pares de nodos, se consideró la existencia de enlace (no dirigido) en un par si el puntaje asociado al mismo excedía un umbral (de puntaje) determinado a partir de un percentil fijo (de la distribución de puntajes), igual para todas las métricas estudiadas. En el caso de la red de CDR3β el percentil utilizado fue 99.95 y en caso de la red de epítopos fue 95 para el conjunto de datos 1 (DS1). Se utilizaron dichos valores elevados con el propósito de acotar el análisis en primera instancia y asimismo para comparar los resultados con lo obtenido en la literatura, por ejemplo con lo reportado en la Tab. 1 de Meysman *et al.*, 2018 [10]. A continuación, en la figura 3a, se puede observar la distribución de puntajes y el umbral elegido, para una red de CDR3β (red T7, DS1). En la figura 3b, se muestra lo mismo para una red de epítopos (red E2, DS1).

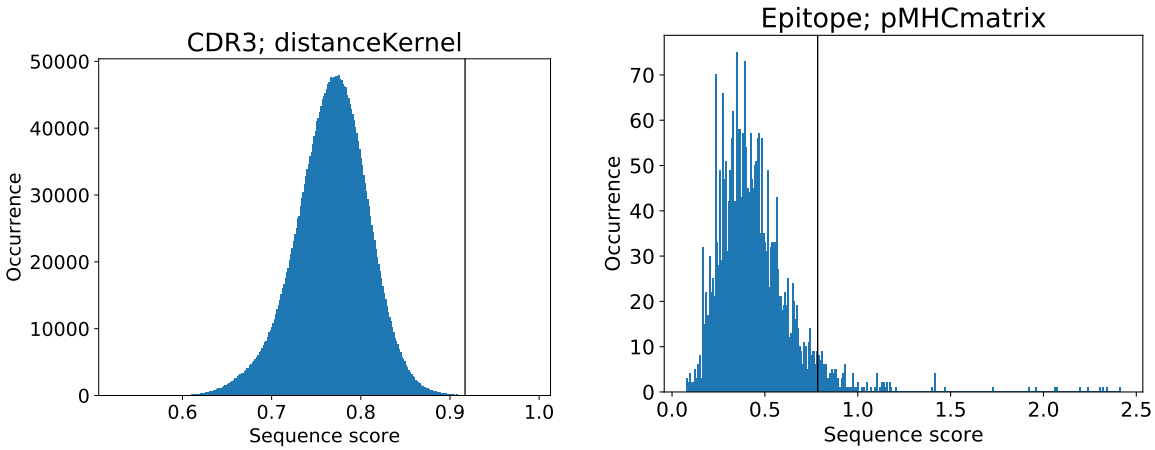
Consideramos que en un futuro trabajo sería interesante estudiar las características topológicas de cada red al variar los valores del umbral, en vez de realizar el análisis para un umbral fijo. De esta forma se podría determinar la existencia de un umbral óptimo, para el cual, luego de aplicar los algoritmos de *clustering* planteados, se obtuviesen comunidades constituidas por CDR3β asociados, en gran medida, a un único epítopo. Cabe destacar que, para el conjunto de datos 2 (DS2), los percentiles (de las distribuciones de puntajes) que determinaron el puntaje umbral para la construcción de las redes fueron 99.25, para las secuencias CDR3β, y 85, para las secuencias de epítopos.

2.4. Conjuntos o sets de datos utilizados

Se utilizaron dos conjuntos para el análisis realizado los cuales se detallan a continuación:

1. Set global: secuencias de CDR3β que interaccionan con epítopos de 6 virus de humano conocidos, seleccionados por las características detalladas en la sección 2.1.

2. Set reducido: secuencias de CDR3β del set global que interaccionan con epítopos del gen p24 del virus HIV-1.



(a) Red T7 [DS1]. Puntaje umbral: 0.917.

(b) Red E2 [DS1]. Puntaje umbral: 0.784

Figura 3: Ejemplos de histogramas de puntajes de similitud de secuencias computados para cada posible par de CDR3β (a) o epítopos (b). Los umbrales se indican con una línea vertical.

Cuadro 2: Características de sendos conjuntos de CDR3β utilizados en el presente trabajo.

ID	# **Secuencias TCR**β	# **Especies**	# **Genes**	# **Epítopos**
DS1 (Global)	2462	6	31	75
DS2 (Reducido)	515	1(HIV-1)	1(p24)	19

2.5. Detección de comunidades

Con el objetivo de evaluar cómo se agrupan las secuencias de CDR3β en comunidades se aplicaron dos algoritmos de *clustering*, InfoMap y Louvain, sobre las redes T1 a T7 estudiadas. Es importante mencionar que se aplicaron dichos algoritmos sobre las componentes gigantes de las redes T1 a T7, debido a que, para todas las redes estudiadas, muchos nodos quedaron sin enlazar y otros cuantos resultaron unidos a unos pocos vecinos (ver sección 3.1) . Se consideró que dadas las redes resultantes, con los criterios de definición de enlace mencionados en la sección 2.3, lo más oportuno para este caso sería analizar la componentes gigantes remanentes en cada red.

2.6. Medidas para evaluar la bondad de las particiones

Aparte de la modularidad y la silueta (promedio y nodo a nodo), consideramos las siguientes medidas de desempeño del proceso de *clustering* (descritas en Meysman *et al.*, 2018 [10]) para evaluar la bondad de las comunidades detectadas.

Estos nuevos observables se introducen particularmente para evaluar en qué medida las comunidades detectadas constituyen grupos de secuencias CDR3β con una misma especificidad epitópica.

A continuación se definen dichas medidas mencionadas:

Retención. La retención, ecuación (2.1), indica qué fracción, del total de las secuencias CDR3β, se agrupan en algún cluster c, sumando sobre todas las particiones definidas C.

$$Retencion = \frac{\sum_{c \in C} \#\{TCR \in c\}}{\#\{TCR\}} \tag{2.1}$$

Pureza. La pureza, para un dado cluster c (ecuación 2.2), considera al epítopo modal o asociado en mayor número a las secuencias CDR3β del cluster c en cuestión y computa qué fracción de las secuencias del cluster reconocen al epítopo modal.

$$Pureza[c] = \frac{\#\{EPI \in c / EPI = mod(EPI \in c)\}}{\#\{EPI \in c\}} \tag{2.2}$$

Asimismo, la pureza media, Ec. (2.3), es equivalente al promedio de las purezas de cada cluster c detectado.

$$< Pureza >= \frac{1}{\#\{C\}} \sum_{c \in C} \frac{\#\{EPI \in c / EPI = mod(EPI \in c)\}}{\#\{EPI \in c\}} \tag{2.3}$$

Consistencia. La consistencia, para un dado epítopo EPI, ecuación (2.4), se define como la fracción de secuencias CDR3β asociadas al epítopo que pertecen al cluster c_{EPI}, es decir al cluster con más secuencias asociadas a dicho epítopo. Cabe destacar que, a diferencia de lo que se propone en Meysman *et al.*, 2018, no se impuso la restricción de clusters c_{EPI} únicos para cada epítopo en el conjunto de datos.

$$Consistencia[EPI] = \frac{\#\{EPI \in c_{EPI}\}}{\#\{EPI\}} \tag{2.4}$$

Asimismo, la consistencia media, ecuación (2.5), es equivalente al promedio de las consistencias de cada epítopo EPI detectado.

$$< Consistencia >= \frac{1}{n} \sum_{EPI_{i=1}}^{EPI_n} \frac{\#\{EPI_i \in c_{EPI_i}\}}{\#\{EPI_i\}} \tag{2.5}$$

Por otro lado también se considero definir una consistencia media pesada por la cantidad de secuencias asociadas a cada epítope, con el objetivo de darle más peso a la consistencia calculada para aquellos epítopes con más secuencias CDR3β asociadas.

Por último se definió la **cobertura**, ecuación (2.6), que indica la cantidad de nodos, secuencias de CDR3β o epítopos, de la red total definida (luego de computar los puntajes para cada par de secuencias posibles y definir los enlaces tal como se menciona en 2.3) que quedan remanentes en la componente gigante de la red.

$$Cobertura = \frac{\#n \in GC}{N} \tag{2.6}$$

Más allá de los criterios ya mencionados para evaluar la bondad de las particiones a nivel de epítopes únicos, se computó, para el caso del conjunto global (DS1, sección 3.1), el test exacto de Fisher para cada cluster (matriz de confusión de 2x2), teniendo en cuenta como criterio de clasificación de las secuencias en determinado cluster su asociación o no con un gen de un virus o un virus dado.

3. Resultados y discusión

3.1. Conjunto global (DS1)

A modo de ilustrar el tipo de redes obtenidas a partir de la metodología descrita, las figuras 4 y 5 muestran los grafos correspondientes a las redes T6 y T7 respectivamente, considerando el conjunto global de secuencias (DS1). Se observa que existen varios nodos solitarios y grupos disjuntos pequeños de nodos donde hay prevalencia de un único color (i.e., una única especie de virus), aparte de una componente gigante, donde, si bien hay colores entremezclados, se observan potenciales particiones de pureza alta.

Dados los umbrales elevados utilizados para determinar la existencia de enlace entre dos secuencias cualesquiera, la densidad de enlaces de todas las redes generadas es baja. Consecuentemente, lo son las coberturas (es decir, la fracción de nodos pertenecientes a la componente gigante de cada red), cuyos valores variaron para las redes de CDR3β entre 1.87 % (T6) y 13.61 % (T7). Contrario a lo que esperábamos, si bien los grupos de nodos formados suelen estar aglomerados en conjuntos de alta pureza, los mismos lo hacen en varios grupos pequeños, en vez de grupos grandes.

En el caso de la red de epítopos E2 (figura 7), así como el de la red de CDR3β T7 (figura 6), se observa que si bien existen aglomeraciones de misma especie, las agrupaciones están más bien dadas por gen. Es importante destacar que esto es evidente del grafo, dado que, en estos dos casos, la disposición de los nodos no es arbitraria; sino que corresponde al método dirigido de fuerzas de Kamada & Kawai, 1989 [8]. Dicho método asume que cada nodo es una masa puntual m y que cada enlace representa un resorte de cierta longitud natural l_0, y posiciona a los nodos en un equilibrio estático

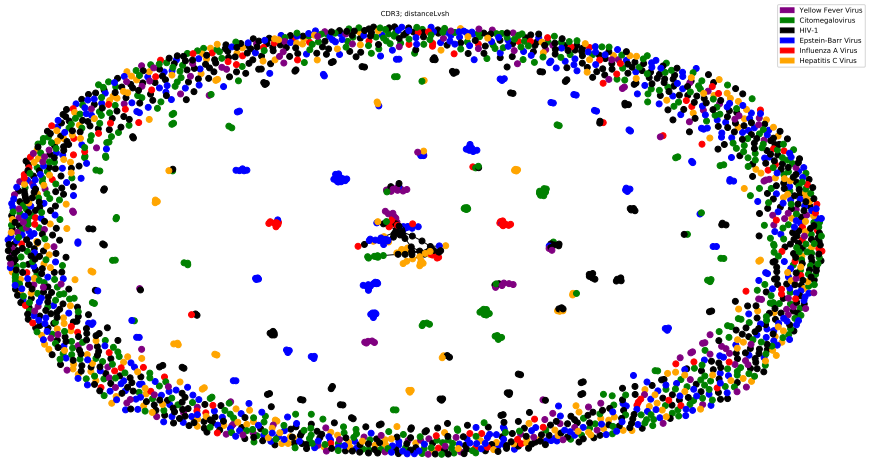


Figura 4: Red T6 (DS1) [Levenshtein Distance Score]. Los colores de los nodos representan las especies de virus a los cuales corresponden los epítopos.

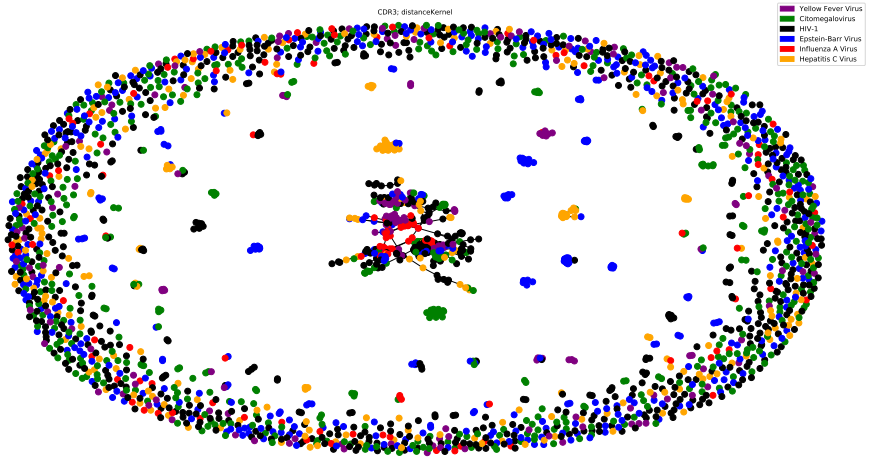


Figura 5: Ídem figura 4, pero para la red T7 (DS1) [Kernel Distance Score].

del sistema. Retomando, por ejemplo, la red E2, se observa que la métrica distingue en general mejor entre genes que entre especie. Asimismo, un conjunto de nodos "aglomerados" de un gen (por ejemplo, gen NS3, nodos amarillos, virus de la Hepatitis C), puede hallarse más cercano a un conjunto de un gen perteneciente a otra especie (por ej. gen p24, nodos negros, virus HIV-1) que otro conjunto de la misma especie. Esto mismo se puede ver en la red T7. Nuevamente esto ilustra que las métricas distinguen más detalladamente de lo que esperábamos, es decir, a nivel de genes epitópicos más que a nivel de especies de virus. El análisis del conjunto reducido DS2 nos muestra que más aún, las métricas distinguen grupos a nivel de epítopos específicos más que a nivel de genes epitópicos (véase §3.2).

Las figuras 8a a 9b muestran los valores medios de los indicadores: modularidad, silueta, consistencia y pureza para las redes T1-T7 (DS1), dados los métodos de partición InfoMap y Louvain. En general, las *performances* obtenidas para sendos métodos de partición son similares; aunque Louvain logra valores más altos de modularidad y consistencia medias; mientras que InfoMap logra valores más altos de silueta y pureza medias. A esta altura esperábamos encontrar un método de partición que se desempeñara destacablemente mejor que los demás. Pero, dado que por el momento carecemos de un criterio para considerar alguna de las medidas estudiadas como más relevante respecto de las restantes, no tenemos un argumento concluyente que establezca cuál de los dos métodos de partición resulta mejor que el otro.

Por otro lado, si bien existen diferencias en las medidas obtenidas para las distintas redes, las mismas tampoco son tan significativas como para aseverar que el desempeño de una prima por sobre las restantes. Si bien al considerar la pureza y la consistencia, la red T6 (Levenshtein Distance Score)

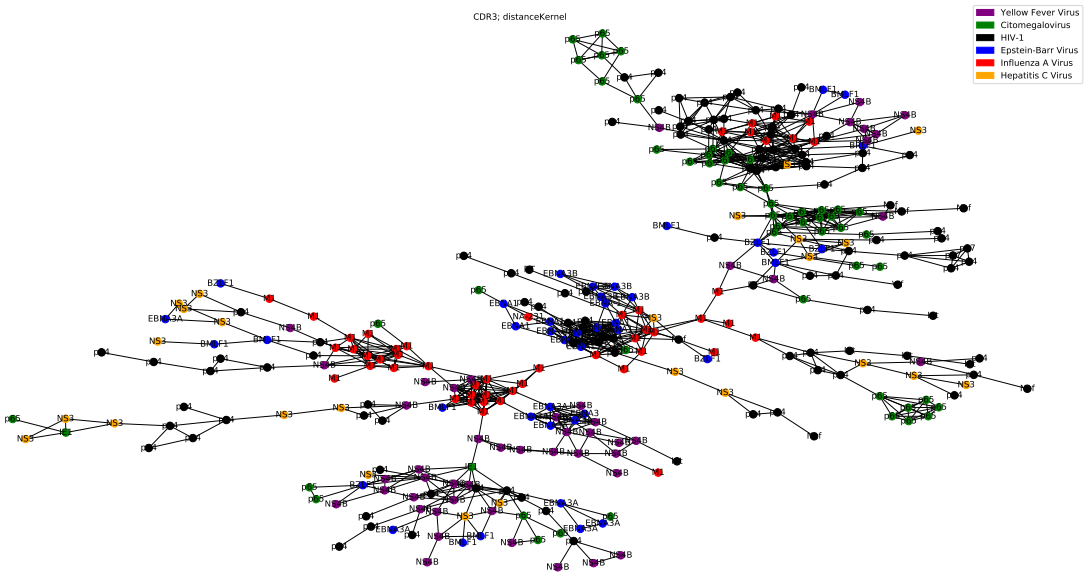


Figura 6: Ídem figura 4, pero para la componente gigante de la red T7 (DS1) [Kernel Distance Score]. Las etiquetas sobre los nodos corresponden a genes de virus.

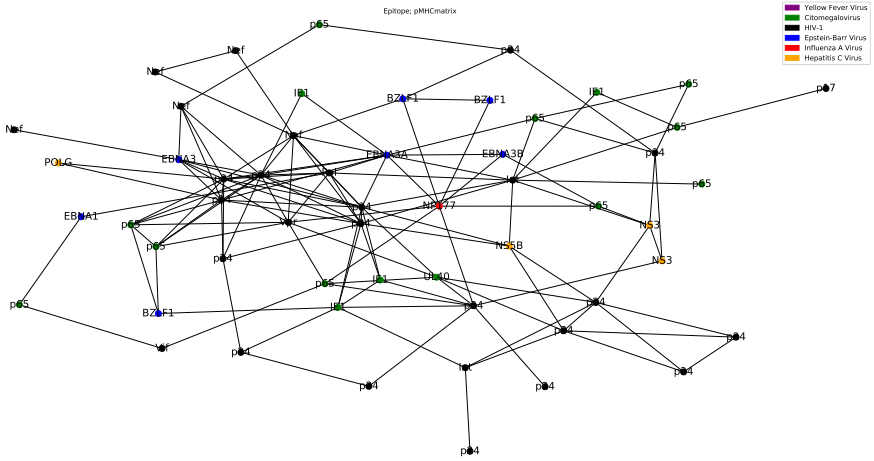
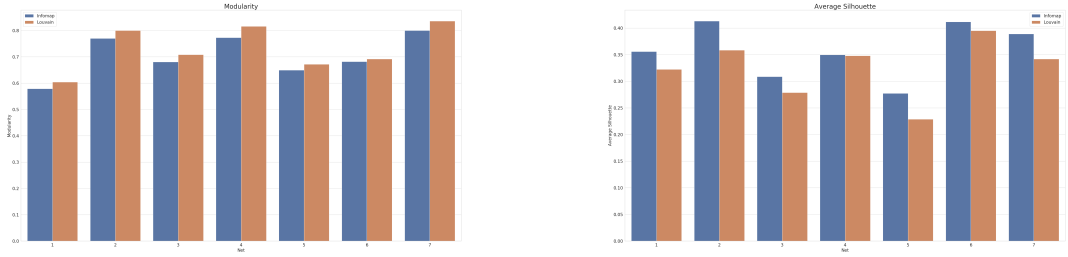


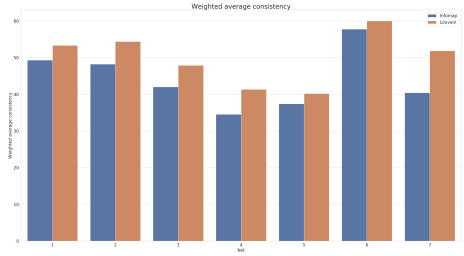
Figura 7: Ídem figura 4, pero para la componente gigante de la red E2 (DS1) [Alignment, matriz PMBEC o 'pMHCMatrix']. Las etiquetas sobre los nodos corresponden a genes de virus.



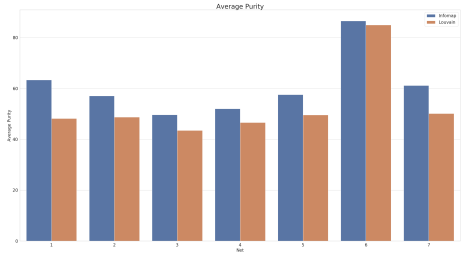
(a) Modularidad promedio para las redes T1 a T7. (b) Silueta promedio para las redes T1 a T7.

Figura 8: Modularidad y silueta para las particiones InfoMap (azul) y Louvain (rojo) en el conjunto DS1.

presenta mejores resultados, la misma posee una cobertura marcadamente más baja en comparación con otras redes de desempeño similar (cf. T6 y T7 Tabla 3). El hecho de que los métodos de partición



(a) Consistencia media para las redes T1 a T7.



(b) Pureza media para las redes T1 a T7.

Figura 9: Consistencia media y pureza media para las particiones InfoMap (azul) y Louvain (rojo) en el conjunto DS1.

hayan sido aplicados sobre las componentes gigantes (GC) sumado a la baja cobertura de T6 explica los valores elevados de pureza y consistencia medias para dicha red. Consideramos que, en una versión más avanzada de dicho trabajo, de podrían aplicar los métodos de partición sobre las redes completas, sin importar si las mismas son completamente conexas o no.

Cuadro 3: Cobertura, retención, nodos en la componente gigante y total para las redes T1-T7 (DS1).

ID	**Cobertura**	**Retención**	**# Nodos en la GC**	**# Nodos total**
T1	3.37	35.74	83	2462
T2	5.32	35.46	131	2462
T3	5.36	40.98	132	2462
T4	8.37	44.88	206	2462
T5	7.68	47.03	189	2462
T6	1.87	24.41	46	2462
T7	13.61	35.13	335	2462

Así como fueron obtenidas la consistencia y pureza medias, dichas magnitudes pueden ser calculadas epítopo a epítopo y partición a partición, respectivamente (véanse ecuaciones 2.4 y 2.2). La figura 10 muestra un histograma de pureza partición a partición para las redes T1-T7 (DS1) para el método de InfoMap.

Se observa en esta figura que los valores se concentran mayormente entre el 20 % y el 60 %. A partir de este valor, la densidad disminuye considerablemente. La excepción es claramente la red T6, que es la que más acumula para el valor de pureza del 100 %.

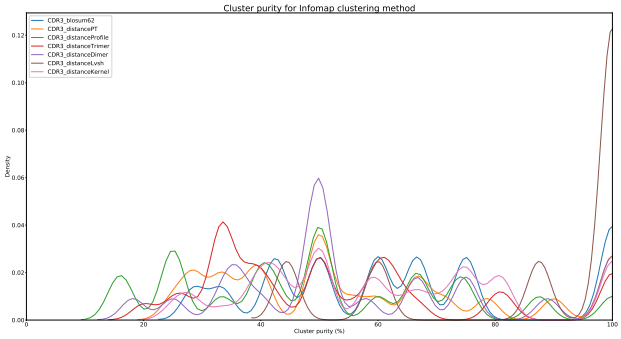
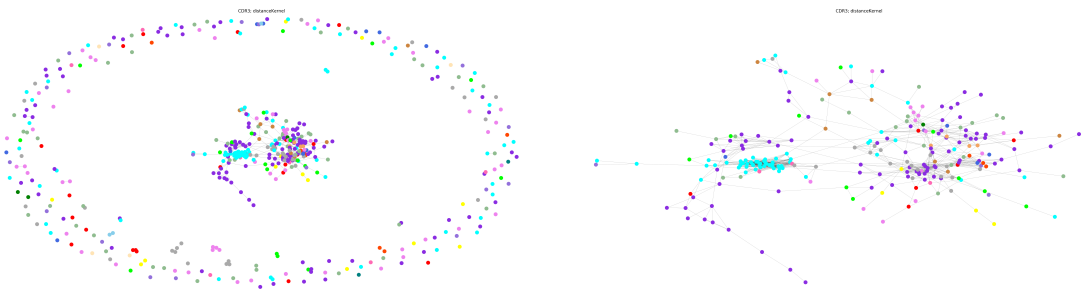


Figura 10: Histograma de la medida de "pureza" (partición a partición, ecuación (2.2)) para cada una de las redes T1-T7 mediante el método de partición InfoMap para las redes del DS1.



(a) Red completa (b) Componente gigante

Figura 11: Red T7 (DS2) [Kernel Distance Score] (a) y detalle de su componente gigante (b). Cada color representa un epítopo específico.

3.2. Conjunto reducido (DS2)

Como se ha detallado hasta aquí, se observó cualitativamente para la totalidad de las métricas estudiadas que los nodos se aglomeraban, en primera aproximación, más bien por genes epitópicos que por especies únicamente. Por este motivo, se estudió el caso de la red asociada a epítopos de un único gen (gen p24 del virus HIV-1) para entender si, en dicho caso, los nodos se agrupaban aún con más resolución, es decir a nivel de epítopos únicos (figuras 11a y 11b).

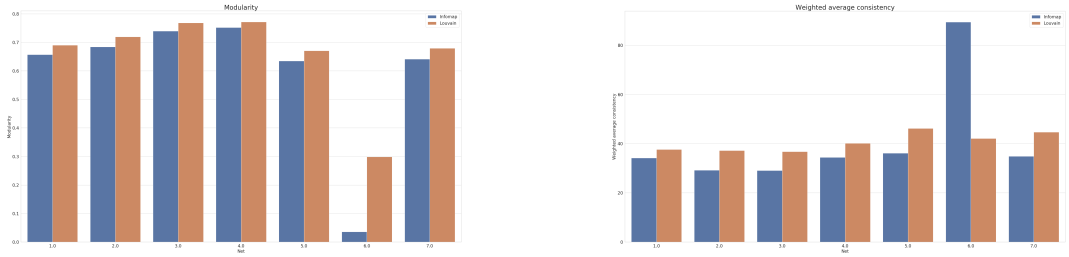
En la figura 11b se observa claramente que existe un criterio de aglomeración por especificidad epitópica. A continuación, en la Tabla 4 se muestran los valores característicos de las componentes gigantes analizadas para las redes del DS2.

Cuadro 4: Cobertura, retención, nodos en la componente gigante y total para las redes T1-T7 (DS2).

ID	**Cobertura**	**Retención**	**# Nodos en la GC**	**# Nodos total**
T1	45.47	71.07	234	515
T2	27.96	57.48	144	515
T3	53.20	68.16	274	515
T4	47.96	70.49	247	515
T5	48.93	73.59	252	515
T6	7.57	42.52	39	515
T7	47.77	60.19	246	515

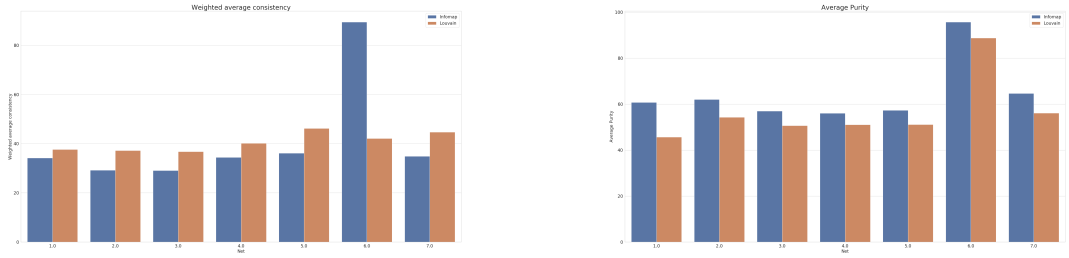
En este caso se puede ver que los valores de cobertura y retención son más elevados que los reportados para las redes del DS1 (ver Tabla 3) debido a que se utilizaron puntajes umbrales para la definición de los enlaces de las redes más laxos que en el caso anterior (ver Sección 2.3). Esto se modificó con el propósito de que quedara remanente un número apreciable de nodos en las componentes gigantes de las redes, comparables con aquellas de las redes del DS1. Una vez más evaluaremos la *performance* de los algoritmos de *clustering* y de las métricas de las redes construídas a partir de la modularidad, la silueta y los observables definidos en la Sección 2.6.

Se puede observar que, al igual que lo obtenido para el DS1 (ver Figs. 8a, 8b, 9a, 9b), el algoritmo de detección de comunidades Louvain alcanza valores más altos de modularidad y consistencia medias; mientras que InfoMap logra valores más altos de pureza media. En el caso de la silueta promedio, la misma da valores más altos para Louvain en el caso del DS2 (figura 12b) y valores más elevados para InfoMap en el caso del DS1 (figura 8b). Lo dicho vale si no tomamos en cuenta el caso atípico de la red T6, la cual está conformada por tan sólo 39 nodos y, en consecuencia, presenta un valor de modularidad más bajo que el resto de las redes (figura 12a). Asimismo, la red T6 es la que tiene una silueta, consistencia y pureza medias más alta, al menos para el algoritmo InfoMap. Nuevamente, este resultado no es muy confiable debido a que la red posee muchos menos nodos que las restantes, es decir que tiene una cobertura marcadamente más baja (Tabla 4). Este comportamiento registrado para la red T6 se repite para el DS1, lo cual nos hace pensar que esta métrica es la que se distingue más, en su comportamiento, del resto de las métricas estudiadas.



(a) Modularidad promedio para las redes T1 a T7. (b) Silueta promedio para las redes T1 a T7.

Figura 12: Modularidad y silueta para las particiones InfoMap (azul) y Louvain (rojo) en el conjunto DS2.



(a) Consistencia media para las redes T1 a T7. (b) Pureza media para las redes T1 a T7.

Figura 13: Consistencia media y pureza media para las particiones InfoMap (azul) y Louvain (rojo) en el conjunto DS2.

En líneas generales, también cabe mencionar que los valores de consistencia medios dieron, considerando los dos métodos de *clustering* y todas las redes en cuestión, levemente más bajos para el DS2 que para el DS1.

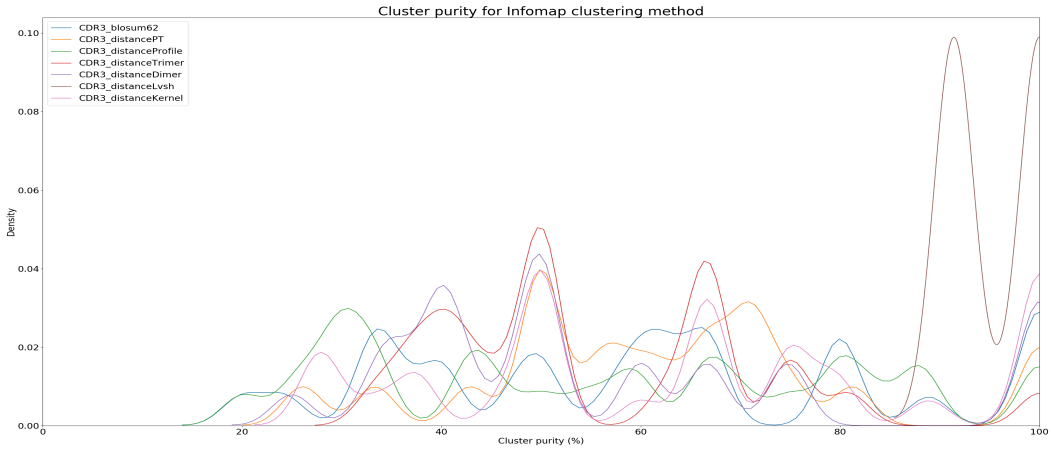
Por último, las figuras 14a y 14b muestran histogramas de pureza partición a partición para las redes T1-T7 (DS2) para los métodos de InfoMap (a) y Louvain (b), respectivamente Se observa en ambos casos que no se obtuvieron particiones con purezas inferiores al 20 %. A su vez, si bien los desempeños en purezas son similares, InfoMap presenta mayor cantidad de particiones con purezas cercanas al 100 % en todas las redes.

Si bien se observan algunas diferencias en el desempeño de los algoritmos de detección de comunidades y, a su vez, entre las redes construidas a partir de distintas métricas, tal como se enunció para el DS1, los resultados obtenidos hasta el momento no son los suficientemente robustos como para realizar una evaluación certera de estas dos variables.

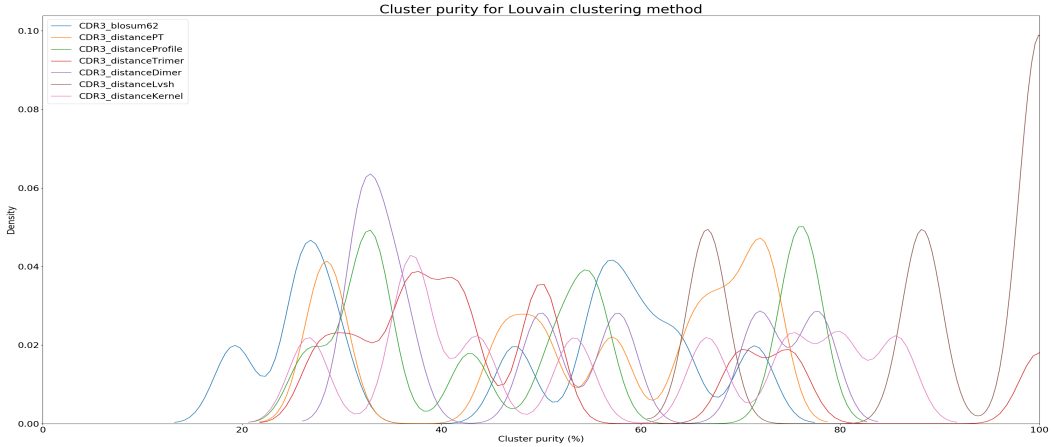
4. Conclusiones y perspectivas

Se estudiaron las propiedades de 7 redes construidas para las secuencias de los receptores T, así como también de 2 redes correspondientes a las secuencias de sus epítopos asociados. Las redes de secuencias de receptores T fueron analizadas cuantitativamente, mientras que las de epítopos fueron analizadas cualitativamente. En este sentido, la construcción de las redes de epítopos, como de las de los TCR, resultó ser una herramienta interesante para explorar ambos espacios de secuencias mencionados.

En primer lugar, se observó que el agrupamiento en comunidades de las redes no se da en grupos grandes en los que se reúnen todos los nodos que corresponden a una especie de virus (como se esperaba antes de iniciar el trabajo), sino en pequeñas comunidades en las que generalmente sí hay un enriquecimiento de alguna de las especies consideradas, es decir pequeños conjuntos de alta pureza. Más aún, lo que se encontró al analizar cada comunidad en particular, es que las secuencias se agrupaban a



(a) InfoMap



(b) Louvain

Figura 14: Histogramas de la medida de "pureza"(partición a partición, ecuación (2.2)) para cada una de las redes T1-T7, mediante los métodos de partición InfoMap y Louvain, para las redes del DS2.

nivel de epítopos únicos. En otras palabras, las métricas utilizadas distinguen las secuencias analizadas con un nivel de detalle mayor al que esperábamos en un principio. Por otra parte, la comparación de las redes construídas por distintas métricas dió como resultado un similar funcionamiento de todas las métricas consideradas. Las diferencias obtenidas para la red T6 (Levenshtein Distance Score) pueden ser explicadas por la baja cobertura de esta red, con lo que no es posible concluir que se tenga un mejor resultado mediante este método en particular.

En cuanto a las perspectivas a futuro, los autores vislumbramos dos posibilidades amplias de continuación del trabajo:

1. En primer lugar, surge la idea de estudiar los efectos de la variación del umbral para selección de enlaces entre secuencias en el momento del armado de las redes. Como se vio en el trabajo, pequeños cambios en el umbral de selección provocan grandes cambios en las estructuras de las redes y en las coberturas alcanzadas por las mismas, por lo que puede resultar de interés comparar los comportamientos de distintas estructuras para un mismo conjunto de secuencias.

2. Por otro lado, sería interesante comparar el espacio de TCRs y el de epítopos y examinar el grado de correspondencia entre las topologías de ambas redes. Asimismo proponemos agregar al grafo de TCRs la información del espacio de epítopos, y así condensar toda la información de

los grafos en uno solo. Dicho de otra manera, resultaría interesante observar las consecuencias de ensayar un *clustering* "supervisado".

Referencias

[1] ATTAF, Meriem, *et al.* (2015) $\alpha\beta$ T cell receptors as predictors of health and disease. *Cellular & Molecular Immunology*, 12, 391–399.

[2] BERGTHALER, Andreas; MENCHE, Jörg. (2017) The immune system as a social network. *Nature Immunology*, vol. 18, no 5, p. 481.

[3] DASH, Pradyot, *et al.* (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, vol. 547, no 7661, p. 89.

[4] DE NEUTER, Nicolas, *et al.* (2018) On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics*, 70, 159–168.

[5] GLANVILLE, Jacob, *et al.* (2017) Identifying specificity groups in the T cell receptor repertoire. *Nature*, vol. 547, no 7661, p. 94.

[6] HENIKOFF, Steven; HENIKOFF, Jorja G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, vol. 89, no 22, p. 10915-10919.

[7] HENNECKE, Jens; WILEY, Don C. (2001) T Cell Receptor–MHC Interactions up Close. *Cell*, Vol. 104, 1–4.

[8] KAMADA, Tomihisa; KAWAI, Satoru. (1989) An algorithm for drawing general undirected graphs, *Information Processing Letters, Elsevier*, 31 (1): 7-15.

[9] KIM, Yohan *et al.* (2009) Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. *BMC Bioinformatics*, 10:394.

[10] MEYSMAN, Pieter, *et al.* (2018) On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. *Bioinformatics-Oxford*, p. 1-7.

[11] RIECKMANN, Jan C., *et al.* (2017) Social network architecture of human immune cells unveiled by quantitative proteomics. *Nature Immunology*, vol. 18, no 5, p. 583.

[12] SHEN, Wen-Jun *et al.* (2014) Introduction to the Peptide Binding Problem of Computational Immunology: New Results. *Found Comput Math*, 14:951–984.

[13] SMITH, Temple F.; WATERMAN, Michael S. Comparison of biosequences. Advances in applied mathematics, 1981, vol. 2, no 4, p. 482-489.

[14] VDJdb; https://vdjdb.cdr3.net/