Universidad Europea MADRID

Predicción de epítopos B neutralizantes

Trabajo de Fin de Máster

Jarque Canalías, Isidro

Tutor: Reche Gallardo, Pedro Antonio
Departamento de Inmunología
Facultad de Medicina
Universidad Complutense de Madrid

Universidad Europea de Madrid
Facultad de Ciencias Biomédicas y de la Salud
Máster Universitario en Bioinformática
Curso 2021-2022

INDICE

Universidad Europea de Madrid
Facultad de Ciencias Biomédicas y de la Salud
Máster Universitario en Bioinformática
Curso 2021-2022

Universidad Europea de Madrid
Facultad de Ciencias Biomédicas y de la Salud
Máster Universitario en Bioinformática
Curso 2021-2022

Resumen

Un epítopo de célula B es la región de un antígeno donde se une un anticuerpo. Generar herramientas que faciliten la predicción de epítopos de células B tiene gran interés para el diseño de vacunas o anticuerpos específicos. En el presente estudio se procesaron secuencias de aminoácidos de epítopos lineales neutralizantes y no neutralizantes procedentes de Immune Epitope Data Base (IEDB). Se automatizó un algoritmo de selección y extracción de secuencias repetidas entre observaciones solapadas para reducir la redundancia y enriquecer la muestra con patrones subepitópicos potencialmente relevantes. Se calcularon diversas variables fisicoquímicas y estructurales para la aplicación de un modelo de *machine learning* y un modelo manual para la clasificación y predicción de la capacidad de neutralización. Aunque las predicciones no resultaron estadísticamente significativas, la aplicación del algoritmo de selección mejoró los parámetros de predicción.

Palabras clave: predicción de epítopos; inmunoinformática; neutralización antigénica

Abstract

A B cell epitope is the antigen region to which an antibody binds. Generating tools to facilitate B cell epitope predictions is very important for specific vaccines and antibodies design. In this study, amino acid sequences of neutralizing and non-neutralizing linear epitopes from Immune Epitope Data Base (IEDB) were processed. An algorithm for selection and extraction of repeated sequences between overlapping observations was automated to achieve lower redundancy and enrich the sample with potentially relevant subepitope patterns. Several physicochemical and structural variables were calculated to perform both a machine learning model and a manual model for epitope neutralization ability classification and prediction. Although the predictions were not statistically significant, the application of the selection algorithm improved the prediction parameters.

Keywords: epitope prediction; immunoinformatics; antigen neutralizing

1. Introducción

1.1 ¿Qué es un epítopo B?

Los antígenos son las moléculas frente a las que se dirige la respuesta inmune adaptativa mediada por linfocitos B y T; aunque su naturaleza puede ser diversa, la mayoría son proteínas. Las subregiones de un antígeno que participan activamente de esta respuesta se denominan epítopos. Los epítopos T, pequeños péptidos digeridos proteolíticamente y presentados en moléculas del sistema mayor de histocompatibilidad son diana del receptor de las células T (TCR)[1] y no son objeto de este trabajo; análogamente, los epítopos B son diana del receptor de células B (BCR) o de sus formas solubles, denominadas inmunoglobulinas o anticuerpos, y merecen una explicación más detallada.

Los epítopos B son el objeto del presente estudio por lo que, de aquí en adelante, serán referidos tan solo como epítopos. Éstos fueron, de hecho, los primeros en ser concebidos como denotan sus raíces etimológicas ἐπί (sobre) y τόπος (lugar), que hacen referencia a su posición superficial en un antígeno en estado nativo, donde son accesibles a los anticuerpos[2].
Sin embargo, epítopo es uno de esos conceptos caleidoscópicos y huidizos que solo parecen estables cuando se los contempla desde cierta distancia.

Partiendo de una clasificación elemental, los epítopos B se pueden diferenciar en lineales y discontinuos: los primeros consisten en residuos secuencialmente consecutivos de entre 5 y 22 aminoácidos[3], y los segundos incluyen residuos no correlativos y a veces distantes en la estructura primaria que, sin embargo, se ubican próximos en la estructura terciaria[4]. No obstante, es importante advertir que el término epítopo es completamente relacional y se define, en todos los sentidos, por interacción con anticuerpos, siendo principalmente una notación funcional sobre el comportamiento de éstos que, en general, resulta ambigua cuando se examina desde una perspectiva estructural.

Bajo este prisma, epítopo es un término vago que acoge datos procedentes de ensayos metodológica e instrumentalmente diversos, desde mediciones con sueros policlonales de respuesta fisiológica a la estimulación con antígeno hasta trabajos de mapeo epitópico con anticuerpos monoclonales que buscan delimitar y caracterizar su estructura[5].

Estudios de este último tipo han demostrado como la sustitución de ciertos residuos de un epítopo lineal no altera la capacidad de unión al anticuerpo, es decir, han probado su discontinuidad estructural[6]; aquellos aminoácidos irremplazables y, por otro lado, insuficientes por sí mismos para que se produzca la unión del anticuerpo se precisan como *epítopo funcional* en oposición a *epítopo estructural*, que alude al péptido completo[7]. Nociones cercanas a la de epítopo funcional populares en inmunología del trasplante son *triplet* y *eplet*, que señalan núcleos epitópicos lineales de tres aminoácidos y motivos espaciales concretos, respectivamente[8].

Por otra parte, algunos epítopos discontinuos pueden estar conformados por pequeños segmentos que son, a su vez, epítopos lineales para otros anticuerpos y, por supuesto, *viceversa*[9].

En resumidas cuentas, epítopo es un vocablo polisémico que empleado genéricamente puede referir tanto a estructuras completamente caracterizadas definidas por un anticuerpo monoclonal, como a sus núcleos funcionales o, en los casos menos precisos, a simples fragmentos proteicos que muestran capacidad de unir anticuerpos indeterminados generados policlonalmente y de los que, en cierto modo, cabría afirmar que no son epítopos, sino que los contienen.

1.2 Neutralización antigénica

La neutralización antigénica se define como la inhibición de la actividad biológica característica de un antígeno acontecida tras la unión de anticuerpos[10]. Formalmente, la capacidad de neutralizar debiera atribuirse a los anticuerpos, que son los agentes neutralizantes y no a los epítopos, que son parte del objeto neutralizado, sin embargo, el tándem epítopo-anticuerpo constituye una unidad funcional interdependiente y los epítopos involucrados en este fenómeno reciben también el adjetivo de neutralizante[11].

Pueden concebirse diversas características relacionadas con epítopos que modulen o determinen la neutralización, algunas extrínsecas como su número y posición espacial respecto al antígeno o el grado de proximidad a otros epítopos, pero también intrínsecas, como patrones estructurales[12].

2. Hipótesis y objetivos

La hipótesis de partida es que deben existir motivos estructurales diferenciables entre epítopos neutralizantes y no neutralizantes que puedan deducirse a partir de su secuencia de aminoácidos.

El principal objetivo de este trabajo es encontrar diferencias entre un subconjunto muestral de péptidos clasificados como epítopos lineales neutralizantes y otro subconjunto de péptidos clasificados como epítopos lineales no neutralizantes que sean extrapolables a la población y aplicables como elementos predictivos de neutralización.

3. Material y métodos

3.1 *Hardware* y entorno de programación

Todo el proceso computacional de acondicionamiento, cálculo y análisis de datos se diseñó en lenguaje R v.4.2.1[13] empleando RStudio en entorno Windows 10 sobre un PC portátil con procesador Intel ® Core™ i5 2.49 GHz y 8 GB de RAM (ver anexo).

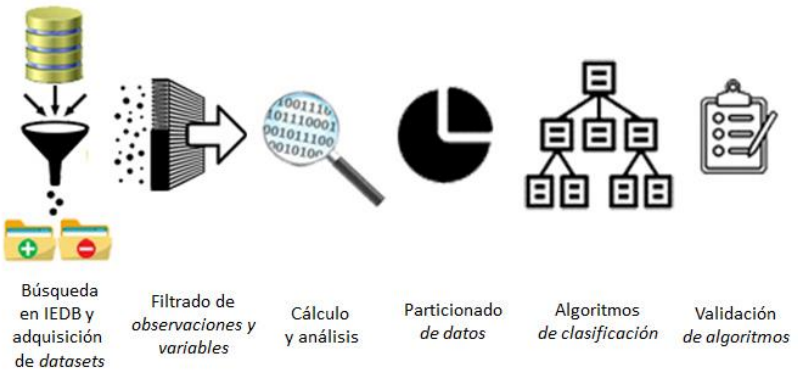

Figura 1. Diagrama del flujo de trabajo

3.2 Obtención de datos

IEDB es una base de datos curada[14] financiada por el *National Institute of Allergy and Infectious Diseases*, de acceso libre y gratuito a través de www.iedb.org, que cosecha y cataloga información sobre epítopos procedente de experimentación con células B y T mediante la explotación sistemática de literatura especializada.

Aprovechando el motor de búsqueda y clasificación de IEDB integrado en la interfaz gráfica de su plataforma *web*, que permite el empaquetado de información en base a diversas características como la conformación de la molécula estudiada, el tipo celular ensayado o la actividad biológica observada, entre otras[15], se compusieron y descargaron dos subconjuntos de datos en formato *CSV* explorando las siguientes características: fragmentos proteicos lineales que fueron diana de inmunoglobulinas procedentes de ensayos en los que se observó actividad neutralizante, en adelante *dataset* positivo o LNP; y segmentos proteicos con estructura lineal que fueron diana de inmunoglobulinas procedentes de ensayos en los que se no observó actividad neutralizante, en adelante *dataset* negativo o LNN.

Las secuencias de aminoácidos completas de las proteínas parentales se obtuvieron de la base de datos UniProt[16].

4. Resultados

4.1 Descripción, exploración y acondicionamiento de datos

Los paquetes de datos obtenidos de IEDB en formato CSV constan de una serie de observaciones dispuestas en filas definidas parcialmente por hasta 28 columnas entre las que destacan el identificador único IEDB del epítopo, su secuencia de aminoácidos, el identificador único UniProt de la proteína parental y las posiciones de inicio y final del epítopo dentro de la proteína.

Como primera medida, fueron suprimidas las observaciones con aminoácidos no canónicos o químicamente modificados, así como aquellas no referenciadas al identificador UniProt. Un análisis exploratorio de los datos permite detectar 209 observaciones posicionadas ambivalentemente en sendos *dataset* positivo y negativo. Para indagar en esta incoherencia se revisó la información de la literatura originaria de una pequeña muestra aleatoria de estos casos a través del servidor IEDB; en algunos de ellos la ambigüedad quedó justificada por provenir de ensayos con diferentes organismos huésped que produjeron resultados distintos, en otros esta imprecisión no logró explicarse. De todos modos, las 209 observaciones fueron desechadas para evitar contrariedades.

Continuando con el análisis, puede apreciarse un amplio rango de longitudes peptídicas que abarca de 2 a 829 aminoácidos. Esta heterogeneidad es menos sorprendente teniendo en cuenta las consideraciones mencionadas en la introducción sobre la imprecisión del término epítopo y la enorme variabilidad metodológica e instrumental con que estos resultados se producen.

Observando el tamaño de los péptidos en los datos sin procesar en los gráficos de distribución de frecuencias de la *Figura 2* parece razonable fijar un límite superior en torno a 25-26 aminoácidos asimilando así, groseramente, la muestra a una función gaussiana y al abanico de longitudes esperables en un epítopo ya citado en la introducción[17]. Sin embargo, un examen detallado de las secuencias revela que existe redundancia por solapamiento, total o parcial, entre epítopos procedentes de una misma proteína. Este hallazgo motivó la idea de programar un algoritmo automatizado de comparación posicional y selección de secuencias entre epítopos de una misma proteína antes de acotar las longitudes de la muestra. El algoritmo procede de la siguiente manera: cuando detecta solapamiento total entre las posiciones de dos péptidos del *dataset* positivo, es decir, que uno contiene al otro, compara la identidad de las regiones superpuestas y, si coincide, procede a eliminar la observación de mayor tamaño; cuando el solapamiento es parcial y superior al 50%, tras verificar la identidad de las regiones solapadas, construye una observación artificial cuya secuencia corresponde con la región compartida; la verificación de identidad entre secuencias evita desechar de la muestra mutaciones puntuales potencialmente informativas. Si el solapamiento total ocurre entre secuencias del *dataset* negativo la selección procede a la inversa, conservando la de mayor longitud; el solapamiento parcial del *dataset* negativo no ha sido tenido en cuenta. El algoritmo es iterado hasta eliminar la superposición redundante.

La lógica detrás de este proceso adopta la premisa de que estas observaciones surgen de ensayos donde el epítopo lineal es definido funcional y parcialmente, es decir, donde se ha confirmado que un péptido concreto capta los anticuerpos estudiados, pero no se ha restringido sucesivamente su tamaño hasta ubicar el núcleo que conserva esta propiedad. Aceptando este supuesto, se hipotetiza que la aparición recurrente de regiones concordantes entre diversas secuencias de péptidos procedentes de un mismo estudio o de ensayos independientes puede indicar la presencia de un motivo nuclear estructuralmente relevante. No obstante, es de justicia apuntar otras hipótesis que pueden explicar el suceso, como la posible multiplicidad de epítopos insertos en una cadena peptídica. Según este último supuesto los epítopos no compartidos se perderían tras aplicar el algoritmo, aun así, puede presumirse cierta tendencia o predilección inmunológica hacia aquellos que han sido rescatados por repetición, por lo que aportarían mayor peso al modelo computacional que aquellos excluidos.
El distinto criterio selectivo introducido en el algoritmo para sendos *datasets* pretende acentuar el contraste de los datos, enriqueciendo la muestra con presuntos motivos neutralizantes mientras se expande todo aquello que no lo es.

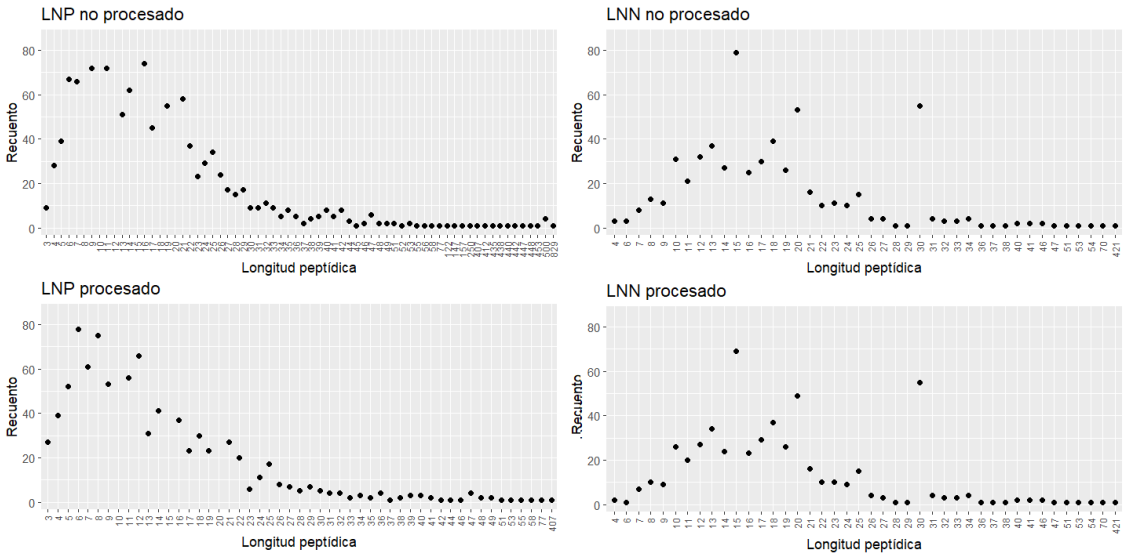

Figura 2. Distribución de longitudes peptídicas antes y después de aplicar el algoritmo de selección

La aplicación del algoritmo moldea las distribuciones reduciendo las observaciones y recuentos extremos siguiendo un criterio intrínseco a la naturaleza de los datos, como puede apreciarse en la *Figura 2*, especialmente en las gráficas izquierdas que corresponden a LNP.

4.2 Cálculo y análisis

4.2.1. Parametrización de los datos

Tras acondicionar los datos se ejecutó un análisis composicional de las observaciones y se calcularon múltiples índices explicativos de propiedades fisicoquímicas y estructurales derivados de sus secuencias de aminoácidos, los cuales se detallan a continuación. Al finalizar esta etapa cada observación se cuenta con 8502 variables.

1. Composición de aminoácidos: Cálculo de la proporción dentro de un péptido de cada uno de los veinte aminoácidos canónicos relativizado con respecto a las frecuencias observadas en el universo proteico UniProt.

2. Composición de aminoácidos por categorías: Cálculo de la proporción dentro de un péptido de aminoácidos agrupados según tamaño, aromaticidad y polaridades de la cadena lateral.

3. Composición de dipéptidos y tripéptidos: Cálculo de la proporción dentro de un péptido de cada una de las 400 combinaciones de dipéptidos posibles (AA, AC, […], YW, YY)* y de las 8000 combinaciones de tripéptidos posibles (AAA, AAC, […], YYW, YYY).
**Nomenclatura de aminoácidos de una sola letra.*

4. Composición de dipéptidos y tripéptidos agrupados: Cálculo de la proporción dentro de un péptido de cada una de las 16 combinaciones de dipéptidos posibles y de las 64 combinaciones

de tripéptidos posibles agrupados según la polaridad de sus cadenas laterales en ácidos, básicos, polares e hidrofóbicos.

5. Carga neta: Determinación de la carga neta de un péptido calculada a *pH* 7 mediante la ecuación de Henderson-Hasselbalch [18].

6. Estimación de estructura secundaria: Aproximación a la estructura secundaria basada en las puntuaciones calculadas para tres posibles estados, que son hélice alfa, conformación extendida o superenrollamiento (*coiled coil*)[19].

7. Índice de interacción potencial de Boman: Este índice resulta de la suma, normalizada por longitud de un péptido, de los coeficientes de solubilidad de todos sus residuos como estimación de su capacidad para establecer interacciones hidrofílicas [20].

8. Índice alifático de Ikai: Volumen relativo de un péptido ocupado por cadenas laterales alifáticas [21].

9. Índice de inestabilidad de Guruprasad: Basado en diferencias de composición halladas entre proteínas estables e inestables[22]

10. Factores, escalas, propiedades y vectores: Bajo estos rótulos se agrupan diversos conjuntos de factores numéricos en escala continua que describen y simplifican los espacios multidimensionales que dimanan de la combinación de múltiples propiedades estudiadas en los aminoácidos y sus interacciones, mediante la aplicación de análisis multivariante y la descripción de componentes principales sobre tales espacios. Se detallan a continuación.

a) Los **factores de Kidera** concentran la información de 188 propiedades físicas [23]. Son diez, y su ámbito informativo puede resumirse del siguiente modo:

- KF_1: Tendencia a formar α-hélices o a formar curvas.
- KF_2: Volumen global de las cadenas laterales.
- KF_3: Tendencia a formar estructuras β.
- KF_4: Hidrofobicidad global.
- KF_5: Frecuencia normalizada de formación de dobles curvas.
- KF_6: Volumen específico parcial (volumen/masa).
- KF_7: Tendencia a adoptar conformaciones planas.
- KF_8: Frecuencia normalizada de formación de regiones α.
- KF_9: Constante de disociación carboxílica.
- KF_{10}: Hidrofobicidad circundante a las estructuras β.

b) Los **vectores FASGAI** (Factor Analysis Scales of Generalized Amino Acid Information) resultan del procesamiento de 335 variables de carácter diverso [24] que convergen en los siguientes seis elementos:

- F_1: Hidrofobicidad.

- F_2: Tendencia a formar estructuras α o β.
- F_3: Volumen.
- F_4: Características composicionales.
- F_5: Flexibilidad local.
- F_6: Propiedades electrónicas.

c) Los **vectores HSE** (*Vectors of Hydrophobic, Steric and Electronic properties*) se originan del cómputo de 50 variables fisicoquímicas categorizadas como propiedades hidrofóbicas, estéricas y electrónicas [25] analizadas por separado y destiladas en $VHSE_1$-$VHSE_2$, $VHSE_3$-$VHSE_4$ y $VHSE_5$-$VHSE_8$, respectivamente.

d) Las **escalas Z** parten de propiedades fisicoquímicas obtenidas por resonancia magnética nuclear (RMN) y cromatografía de capa fina (TLC) para calcular cinco factores [26] que compilan la siguiente información:

- Z_1: Lipofilicidad
- Z_2: Volumen estérico y polarizabilidad
- Z_3: Propiedades electrónicas
- Z_4, Z_5: Relacionan electrogenicidad, entalpía de formación y dureza química

e) Las **escalas T** (*Topological*) sintetizan en cinco únicos factores (T_1-T_5) 67 variables de carácter topológico [27].

f) Las **escalas ST (*Structural-Topological*)** codifican en ocho factores numéricos (ST_1-ST_8) 827 variables de corte topológico-estructural [28].

g) Las **propiedades de Cruciani** resumen el comportamiento observado al confrontar aminoácidos con sondas químicas de diversa índole en tres factores que describen polaridad, hidrofobicidad y formación de enlaces de hidrógeno [29].

h) Las **Escalas MS-WHIM** (*Molecular Surfaces - Weighted Holistic Invariant Molecular*) son tres factores ($MS\text{-}WHIM_1$- $MS\text{-}WHIM_3$) derivados de 36 propiedades geométricas y electrostáticas [30].

4.2.2. Clasificación mediante *Random Forest*

Para escrutar el poder informativo de las variables se implementó un algoritmo estándar de selección de predictores basado en el grado de variabilidad entre observaciones. Este algoritmo penaliza la baja proporción de valores únicos en la muestra por lo que se excluyeron, entre otras observaciones, los recuentos de dipéptidos y tripéptidos no agrupados en categorías, dada la gran proporción de valores nulos naturalmente existente para cada una de las 8400 combinaciones. Estas variables, que no participaron del presente modelo, se recuperaron y procesaron en paralelo como se explica en apartados ulteriores.
El *Random Forest* es una técnica de clasificación ampliamente utilizada en inteligencia artificial y aprendizaje automático que busca definir las relaciones entre variables de una muestra que

mejor logren explicar un determinado atributo, para tratar de definir estas relaciones entre variables como algoritmos predictores.
Como su nombre sugiere, se basa en el ensamblaje de múltiples árboles decisionales generados individual y aleatoriamente[31]. Un árbol de decisión individual prueba distintos modos de jerarquizar las variables que computa hasta consolidar aquel que minimiza la impureza de sus nodos terminales, entendiendo impureza como ambigüedad en la clasificación de los atributos[32].

Para efectuar un *Random Forest* la muestra es dividida en subconjuntos de observaciones aleatorios y, para cada uno de ellos, se crean distintos árboles que procesan, aleatoriamente, diferentes muestras.

Cada árbol consolidado aporta su estimación particular al aparato predictivo global que pondera por mayoría simple la contribución de todos sus árboles para emitir su veredicto.
Por supuesto, el procedimiento debe ser validado efectuando predicciones sobre datos de características conocidas que no han participado del mismo[33].
Una aproximación para robustecer el modelo y no sobreestimar su capacidad es someterlo a validación cruzada que, brevemente, consiste en fragmentar los datos de entrenamiento en *k* porciones e iterar el proceso *k* veces utilizando alternativamente una de las fracciones como datos de pre-validación. La robustez se incrementa aún más si se repite varias veces el proceso completo[34].

Para tratar de encontrar predictores de neutralización en la muestra de epítopos se ejecutaron diversos *Random Forest* con validación cruzada *k*=10 y 5 repeticiones sobre distintas configuraciones de los datos para comparar sus niveles de ajuste y predicción. Para todas ellas, la muestra seleccionada fue particionada destinando el 85% de sus observaciones al entrenamiento del *Random Forest* y el 15%, ajeno al modelo, se reservó para su validación. Por motivos de síntesis y nitidez solo se presentan las matrices de confusión y estadísticos de validación de los cuatro más relevantes.

1) *Random Forest* sobre datos NO procesados por el algoritmo de selección. Restricción de observaciones según longitud entre 3 y 25 aminoácidos.

		Referencia	
		neutralizante	no neutralizante
Predicción	neutralizante	0	0
	no neutralizante	75	229

Accuracy : 0.7533	Sensitivity : 1.0000	F1 : 0.8593
95% CI : (0.7009, 0.8007)	Specificity : 0.0000	Prevalence : 0.7533
No Information Rate : 0.7533	Pos Pred Value : 0.7533	Detection Rate : 0.7533
P-Value [Acc > NIR] : 0.531	Neg Pred Value : NaN	Detection Prevalence : 1.0000
Kappa : 0	Precision : 0.7533	Balanced Accuracy : 0.5000
Mcnemar's Test P-Value : <2e-16	Recall : 1.0000	

Tabla 1. Estadísticos de Random Forest

2) *Random Forest* sobre datos procesados por el algoritmo de selección. Restricción de observaciones según longitud entre 3 y 25 aminoácidos.

		Referencia	
		neutralizante	no neutralizante
Predicción	neutralizante	28	35
	no neutralizante	39	120

Accuracy : 0.6667	Sensitivity : 0.7742	F1 : 0.7643
95% CI : (0.6005, 0.7283)	Specificity : 0.4179	Prevalence : 0.6982
No Information Rate : 0.6982	Pos Pred Value : 0.7547	Detection Rate : 0.5405
P-Value [Acc > NIR] : 0.8632	Neg Pred Value : 0.4444	Detection Prevalence : 0.7162
Kappa : 0.1954	Precision : 0.7547	Balanced Accuracy : 0.5961
Mcnemar's Test P-Value : 0.7273	Recall : 0.7742	

Tabla 2. Estadísticos de Random Forest

10.000.000	Helix conformation trend
8.478.069	Extended conformation trend
4.016.550	Aliphatic index
3.968.134	Coiled- conformation trend
3.733.293	Hydrophobic moment
3.059.039	Charge
3.034.138	Aromatic residues proportion
2.876.908	KF7
2.792.821	F4
2.748.881	MSWHIM1

Tabla 3. Contribución al modelo de las diez variables más explicativas

3) *Random Forest* sobre datos procesados por el algoritmo de selección. Restricción de observaciones según longitud entre 3 y 30 aminoácidos.

		Referencia	
		neutralizante	no neutralizante
Predicción	neutralizante	73	131
	no neutralizante	4	24

Accuracy : 0.4181	Sensitivity : 0.1548	F1 : 0.2623
95% CI : (0.3539, 0.4844)	Specificity : 0.9481	Prevalence : 0.6681
No Information Rate : 0.6681	Pos Pred Value : 0.8571	Detection Rate : 0.1094
P-Value [Acc > NIR] : 1	Neg Pred Value : 0.3578	Detection Prevalence : 0.1207
Kappa : 0.0727	Precision : 0.8571	Balanced Accuracy : 0.5514
Mcnemar's Test P-Value : <2e-16	Recall : 0.1548	

Tabla 4. Estadísticos de Random Forest

4) *Random Forest* sobre datos procesados por el algoritmo de selección. Restricción de observaciones según longitud entre 8 y 25 aminoácidos.

		Referencia	
		neutralizante	no neutralizante
Predicción	neutralizante	9	16
	no neutralizante	58	122

Accuracy : 0.639	Sensitivity : 0.8841	F1 : 0.7673
95% CI : (0.5692, 0.7048)	Specificity : 0.1343	Prevalence : 0.6732
No Information Rate : 0.6732	Pos Pred Value : 0.6778	Detection Rate : 0.5951
P-Value [Acc > NIR] : 0.8676	Neg Pred Value : 0.3600	Detection Prevalence : 0.8780
Kappa : 0.0219	Precision : 0.8571	Balanced Accuracy : 0.5092
Mcnemar's Test P-Value : 1.878e-16	Recall : 0.8841	

Tabla 5. Estadísticos de Random Forest

Como se observa en las *tablas 1, 2, 4 y 5* los resultados no son estadísticamente significativos, sin embargo, puede apreciarse como la aplicación del algoritmo de selección mejora la sensibilidad y especificidad de la predicción. Así mismo, el intervalo de longitudes peptídicas escogido influye en estos parámetros, alcanzando un punto equilibrado entre 3 y 25 aminoácidos.

4.2.3. Clasificador basado en dipéptidos y tripéptidos

Tras una exploración visual de las variables se puede advertir cierto desequilibrio en la frecuencia con que se presentan algunos dipéptidos y tripéptidos (en adelante DPTP) entre los *datasets* positivo y negativo. Para examinar mejor estas diferencias se computaron los datos, previamente procesados por el algoritmo de selección, del siguiente modo:

1. Se consideró cada *dataset* como un conglomerado de DPTP a partir del recuento total de DPTP de sus observaciones.
2. El cálculo de la frecuencia de cada *DPTP* en cada población se relativizó en base a dicho conglomerado.
3. Se calculó la tasa de cambio (*fold change*) en logaritmo binario para cada DPTP.

En el gráfico de la *Figura 3* puede apreciarse como distintos DPTP aparecen sobredimensionados en sendas muestras de epítopos neutralizantes y no neutralizantes. Se han omitido puntos para aumentar la resolución gráfica.

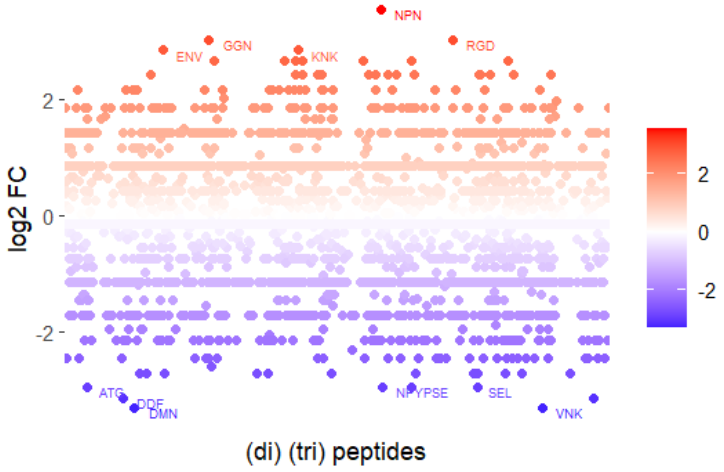

Figura 3. Fold Change de la proporción de DPTP. Rojo LNP, azul LNN.

Encontrar estos resultados inspiró la construcción de un modelo de clasificación cimentado en el cálculo de puntuaciones probabilísticas para cada DPTP.

Los datos fueron fraccionados, dedicando el 85% para configurar el clasificador y retirando el 15% restante para su validación. Se verificó la presencia en la partición de validación identificadores UniProt compartidos entre sendos LNP y LNN.

La atribución de puntuaciones a los DPTP se diseñó del siguiente modo:

1. Para aquellos con *fold change* mayor de 2 se calculó un *puntaje positivo* como el cociente entre la probabilidad de aparición de un DPTP en un epítopo neutralizante y su probabilidad de aparición en un epítopo no neutralizante. Cada probabilidad se calculó como la proporción del DPTP entre el total de DPTPs positivos y negativos, respectivamente.
2. Para aquellos con *fold change* menor de 2 se calculó un *puntaje negativo* como el cociente entre la probabilidad de aparición de un DPTP en un epítopo no neutralizante y su probabilidad de aparición en un epítopo neutralizante. Cada probabilidad se calculó como la proporción del DPTP entre el total de DPTPs negativos y positivos, respectivamente.

Cuando alguna de estas operaciones dio como resultado infinito, debido a la completa ausencia de un DPTP en alguno de los *dataset*, se le asignó un puntaje de 0.5 ajustado empíricamente tras tantear diversos valores.

Las secuencias de las proteínas parentales fueron digeridas en péptidos 30-mer solapados 10 residuos entre sí. Se automatizó un algoritmo de búsqueda de DPTPs y asignación de sus puntos asociados para cada uno de los fragmentos. Como componente predictivo, los fragmentos fueron clasificados como negativo, nulo o positivo según la posición del valor de su puntaje total respecto a dos umbrales, negativo y positivo, fijados empíricamente en 0.5 y 15, respectivamente. Como componente observacional, se verificó la presencia de las secuencias de los epítopos de validación en cada uno de los fragmentos. Cada fragmento fue entonces etiquetado como positivo, negativo, ambivalente o nulo dependiendo de si fueron encontrados en su secuencia epítopos positivos, negativos, ambos o ninguno,

respectivamente. Sendas etiquetas, predicción y observación, fueron cotejadas entre sí contabilizando el número de verdaderos positivos (VP), verdaderos negativos (VP), falsos positivos (FP) y falsos negativos (FN) en cada proteína. Las observaciones nulas no se tuvieron en cuenta independientemente de su predicción asociada.

Por último, se calcularon los parámetros de validación estándar sensibilidad, especificidad, valor predictivo positivo (VPP) y valor predictivo negativo (VPN). La *tabla 5* muestra la media de estos parámetros y su desviación típica asociada calculados en conjunto para todas las proteínas probadas. Como puede apreciarse, la elevada variabilidad entre predicciones individuales que se refleja en los altos valores de las desviaciones típicas indica que acaecen predicciones desastrosas, pero también exitosas, entre distintas proteínas.

Sería necesario emprender un análisis pormenorizado del comportamiento del algoritmo ante las variaciones particulares de cada proteína para continuar refinando su funcionamiento, además, la combinación de los DPTP con alguno de los factores que fueron relevantes en *Random Forest* mostrados en la *Tabla 3* podría incrementar su capacidad predictiva. Todo ello sobrepasa los confines del presente trabajo.

	Media	Desviación Típica
Sensibilidad	54.82	44.67
Especificidad	15.76	33.25
VPP	48.38	40.32
VPN	17.03	35.40

Tabla 5. Parámetros de validación de clasificador basado en DPTP

5. Discusión

Existe un interés creciente en el ámbito de la inmunoterapia por las herramientas bioinformáticas que facilitan la predicción de epítopos, pues la definición de regiones altamente inmunogénicas dentro de una proteína repercute directamente en el diseño de vacunas[35], receptores celulares quiméricos[36], anticuerpos y otros constructos[37].

Actualmente, existen diversos servidores en línea que ofrecen predicciones sobre epítopos de células T como EPISOPT[38] y de células B como Epitopia[39], BCEPS[40] o BepiPred[41], que implementan distintos algoritmos de *machine learning* para asignar puntuaciones a las regiones de la secuencia de una proteína relacionadas con su probabilidad de devenir en epítopo. La principal dificultad en este tipo análisis radica en disponer de datos negativos de contraste: abundan en la literatura descripciones de elementos que han sido calificados de epítopos, pero no es tan común encontrarlas de aquello que no es un epítopo. Es práctica habitual, pues, tomar secuencias creadas a partir de la digestión aleatoria de proteínas alojadas en bases de datos para tratar de paliar la carencia de datos negativos[42]. En el presente trabajo este problema es sorteado al virar el objetivo hacia la búsqueda de diferencias entre epítopos neutralizantes y

epítopos no neutralizantes, de este modo, los datos utilizados como negativos fueron minados de publicaciones donde se han declarado manifiestamente como epítopos que no provocaron neutralización. Sin embargo, existen otros inconvenientes a los que se ha aludido en apartados anteriores, como la presencia de observaciones clasificadas ambivalentemente en ambas categorías, la redundancia de las observaciones o la heterogeneidad de tamaños peptídicos que lastran y embrutecen los resultados.

El hallazgo más revelador en este estudio es que la aplicación de un algoritmo de depuración y rescate de observaciones relativamente sencillo previo al notable cómputo de variables y al complejo análisis de *machine learning* mejora sustancialmente la calidad del proceso. El acceso a un volumen considerable de datos depurados y fiables es un punto limitante para este tipo de empresas. A mi entender, es menester prestar, al menos, la misma atención y dedicación al diseño de herramientas de refinado de datos que a la etapa de cálculo y clasificación pues la calidad que puede ofrecer un sistema analítico, por sofisticado que sea, depende estrechamente de los datos con que es alimentado.

En efecto, la predicción de epítopos B aún puede considerarse una quimera. Estudios independientes de evaluación de los métodos vigentes indican que su poder predictivo es cercano a la aleatoriedad[43]. Amén de las imprecisiones terminológicas y otras contrariedades metodológicas, el amplísimo repertorio combinacional de inmunoglobulinas que abarca virtualmente toda superficie antigénica[44], la interdependencia funcional epítopo-anticuerpo y los factores extrínsecos a la estructura epitópica que influyen en la neutralización hacen de la predicción de epítopos y sus efectos una tarea ardua.

6. Referencias bibliográficas

1. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. T Cells and MHC Proteins.

2. Jerne NK. Immunological speculations. Annu Rev Microbiol. 1960;14:341-58. doi: 10.1146/annurev.mi.14.100160.002013. PMID: 13789973.

3. Singh H, Ansari HR, Raghava GP. Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One. 2013 May 7;8(5):e62216. doi: 10.1371/journal.pone.0062216. PMID: 23667458; PMCID: PMC3646881.

4. Ferdous S, Kelm S, Baker TS, Shi J, Martin ACR. B-cell epitopes: Discontinuity and conformational analysis. Mol Immunol. 2019 Oct;114:643-650. doi: 10.1016/j.molimm.2019.09.014. Epub 2019 Sep 20. PMID: 31546099.

5. Johne B. Epitope mapping by surface plasmon resonance in the BIAcore. Mol Biotechnol. 1998 Feb;9(1):65-71. doi: 10.1007/BF02752698. PMID: 9592769.

6. Klasse PJ. Neutralization of Virus Infectivity by Antibodies: Old Problems in New Perspectives. Adv Biol. 2014;2014:157895. doi: 10.1155/2014/157895. Epub 2014 Sep 9. PMID: 27099867; PMCID: PMC4835181.

7. Cunningham BC, Wells JA. Comparison of a structural and a functional epitope. J Mol Biol. 1993 Dec 5;234(3):554-63. doi: 10.1006/jmbi.1993.1611. Erratum in: J Mol Biol 1994 Apr 8;237(4):513. PMID: 7504735.

8. Duquesnoy RJ. A structurally based approach to determine HLA compatibility at the humoral immune level. Hum Immunol. 2006 Nov;67(11):847-62. doi: 10.1016/j.humimm.2006.08.001. Epub 2006 Sep 1. PMID: 17145365; PMCID: PMC2169290.

9. Van Regenmortel MH. What is a B-cell epitope? Methods Mol Biol. 2009;524:3-20. doi: 10.1007/978-1-59745-450-6_1. PMID: 19377933.

10. Poignard P, Klasse PJ, Sattentau QJ. Antibody neutralization of HIV-1. Immunol Today. 1996 May;17(5):239-46. doi: 10.1016/0167-5699(96)10007-4. PMID: 8991386.

11. Ekiert DC, Friesen RH, Bhabha G, Kwaks T, Jongeneelen M, Yu W, Ophorst C, Cox F, Korse HJ, Brandenburg B, Vogels R, Brakenhoff JP, Kompier R, Koldijk MH, Cornelissen LA, Poon LL, Peiris M, Koudstaal W, Wilson IA, Goudsmit J. A highly conserved neutralizing epitope on group 2 influenza A viruses. Science. 2011 Aug 12;333(6044):843-50. doi: 10.1126/science.1204839. Epub 2011 Jul 7. PMID: 21737702; PMCID: PMC3210727.

12. Parker Miller E, Finkelstein MT, Erdman MC, Seth PC, Fera D. A Structural Update of Neutralizing Epitopes on the HIV Envelope, a Moving Target. Viruses. 2021 Sep 5;13(9):1774. doi: 10.3390/v13091774. PMID: 34578355; PMCID: PMC8472920.

[13] *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.*
URL https://www.R-project.org/.rgh 51, n.º 1 (marzo de 2021): 11-12, https://doi.org/10.4997/JRCPE.2021.104.

14. Vita R, Peters B, Sette A. The curation guidelines of the immune epitope database and analysis resource. Cytometry A. 2008 Nov;73(11):1066-70. doi: 10.1002/cyto.a.20585. PMID: 18688821; PMCID: PMC2597159.

15. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019 Jan 8;47(D1):D339-D343. doi: 10.1093/nar/gky1006. PMID: 30357391; PMCID: PMC6324067.

16. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D480-D489. doi: 10.1093/nar/gkaa1100. PMID: 33237286; PMCID: PMC7778908.

17. Singh H, Ansari HR, Raghava GP. Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One. 2013 May 7;8(5):e62216. doi: 10.1371/journal.pone.0062216. PMID: 23667458; PMCID: PMC3646881.

18. Lawrence J. Henderson, «Concerning the relationship between the strength of acids and their capacity to preserve neutrality», *American Journal of Physiology-Legacy Content* 21, n.º 2; 1908 Mar 02; 173-79, https://doi.org/10.1152/ajplegacy.1908.21.2.173.

19. J. Garnier, J. F. Gibrat, y B. Robson, «GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence», *Methods in Enzymology* 266 (1996): 540-53, https://doi.org/10.1016/s0076-6879(96)66034-0.

20. Boman HG. Antibacterial peptides: basic facts and emerging concepts. J Intern Med. 2003 Sep;254(3):197-215. doi: 10.1046/j.1365-2796.2003.01228.x. PMID: 12930229.

21. Ikai A. Thermostability and aliphatic index of globular proteins. J Biochem. 1980 Dec;88(6):1895-8. PMID: 7462208.

22. Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. 1990 Dec;4(2):155-61. doi: 10.1093/protein/4.2.155. PMID: 2075190.

23. Akinori Kidera et al., «Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids», *Journal of Protein Chemistry* 4, n.º 1 1985 Feb 2; 23-55, https://doi.org/10.1007/BF01025492.

24. Liang G, Chen G, Niu W, Li Z. Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human amphiphysin-1 SH3 domains and their peptide ligands. Chem Biol Drug Des. 2008 Apr;71(4):345-51. doi: 10.1111/j.1747-0285.2008.00641.x. Epub 2008 Mar 1. PMID: 18318694.

25. Mei H, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers. 2005;80(6):775-86. doi: 10.1002/bip.20296. PMID: 15895431.

26. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem. 1998 Jul 2;41(14):2481-91. doi: 10.1021/jm9700575. PMID: 9651153.

27. Feifei Tian, Peng Zhou, y Zhiliang Li, «T-Scale as a Novel Vector of Topological Descriptors for Amino Acids and Its Application in QSARs of Peptides», *Journal of Molecular Structure* 830, n.º 1; 2007 Mar 30; 106-15, https://doi.org/10.1016/j.molstruc.2006.07.004.

28. Yang L, Shu M, Ma K, Mei H, Jiang Y, Li Z. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. Amino Acids. 2010 Mar;38(3):805-16. doi: 10.1007/s00726-009-0287-y. Epub 2009 Apr 17. PMID: 19373543.

29. Gabriele Cruciani et al., «Peptide studies by means of principal properties of amino acids derived from MIF descriptors», *Journal of Chemometrics* 18; 2004 Mar 01; 146-55, https://doi.org/10.1002/cem.856.

30. Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. J Comput Aided Mol Des. 1997 Jan;11(1):79-92. doi: 10.1023/a:1008079512289. PMID: 9139115.

31. Tin Kam Ho, «Random decision forests», en *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1 (3rd International Conference on Document Analysis and Recognition, Montreal, Que., Canada: IEEE Comput. Soc. Press, 1995), 278-82, https://doi.org/10.1109/ICDAR.1995.598994.

32. Nunn ME, Fan J, Su X, Levine RA, Lee HJ, McGuire MK. Development of prognostic indicators using classification and regression trees for survival. Periodontol 2000. 2012 Feb;58(1):134-42. doi: 10.1111/j.1600-0757.2011.00421.x. PMID: 22133372; PMCID: PMC4305365.

33. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003 Nov-Dec;43(6):1947-58. doi: 10.1021/ci034160g. PMID: 14632445.

34. Christopher A. Ramezan, Timothy A. Warner, y Aaron E. Maxwell, «Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification», *Remote Sensing* 11, n.º 2; 2019 Jan 17; 185, https://doi.org/10.3390/rs11020185.

35. Chen J, Liu H, Yang J, Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 2007 Sep;33(3):423-8. doi: 10.1007/s00726-006-0485-9. Epub 2007 Jan 26. PMID: 17252308.

36. Guedan S, Calderon H, Posey AD Jr, Maus MV. Engineering and Design of Chimeric Antigen Receptors. Mol Ther Methods Clin Dev. 2018 Dec 31;12:145-156. doi: 10.1016/j.omtm.2018.12.009. PMID: 30666307; PMCID: PMC6330382.

37. Wang Q, Chen Y, Park J, Liu X, Hu Y, Wang T, McFarland K, Betenbaugh MJ. Design and Production of Bispecific Antibodies. Antibodies (Basel). 2019 Aug 2;8(3):43. doi: 10.3390/antib8030043. PMID: 31544849; PMCID: PMC6783844.

38. Molero-Abraham M, Lafuente EM, Flower DR, Reche PA. Selection of conserved epitopes from hepatitis C virus for pan-populational stimulation of T-cell responses. Clin Dev Immunol. 2013;2013:601943. doi: 10.1155/2013/601943. Epub 2013 Nov 21. PMID: 24348677; PMCID: PMC3856138.

39. Rubinstein ND, Mayrose I, Martz E, Pupko T. Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinformatics. 2009 Sep 14;10:287. doi: 10.1186/1471-2105-10-287. PMID: 19751513; PMCID: PMC2751785.

40. Ras-Carmona A, Pelaez-Prestel HF, Lafuente EM, Reche PA. BCEPS: A Web Server to Predict Linear B Cell Epitopes with Enhanced Immunogenicity and Cross-Reactivity. Cells. 2021 Oct 14;10(10):2744. doi: 10.3390/cells10102744. PMID: 34685724; PMCID: PMC8534968.

41. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017 Jul 3;45(W1):W24-W29. doi: 10.1093/nar/gkx346. PMID: 28472356; PMCID: PMC5570230.

42. Blythe MJ, Flower DR. Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci. 2005 Jan;14(1):246-8. doi: 10.1110/ps.041059505. Epub 2004 Dec 2. PMID: 15576553; PMCID: PMC2253337.

43. Zuo T, Gautam A, Wesemann DR. Affinity war: forging immunoglobulin repertoires. Curr Opin Immunol. 2019 Apr;57:32-39. doi: 10.1016/j.coi.2018.12.002. Epub 2019 Jan 25. PMID: 30690255; PMCID: PMC6511487.

7. Apéndice

7.1 Script R disponible en:

https://drive.google.com/file/d/1tWgOjCuUVwk7U0FzPx1IxbEP5Lnfbi2X/view?usp=share_link

7.2 Librerías R empleadas

- dplyr
- stringi
- stringr
- ggplot2
- seqinr
- Peptides
- caret