Optimización combinatoria en el diseño computacional de vacunas

Iker Malaina

Tesis doctoral

Departamento de Matemáticas

Universidad del País Vasco (UPV/EHU)

Bilbao, 2022

Memoria para optar al grado de Doctor en Ciencias Exactas, en la rama de Matemática Aplicada, por la Universidad del País Vasco - Euskal Herriko Unibertsitatea. Realizada bajo la dirección de:

Dra. Virginia Muto, doctora por la Technical University of Denmark (Lyngby, Dinamarca), profesora titular del Departamento de Matemáticas de la UPV/EHU, del área de Matemática Aplicada.

You can't get to a time before the Big Bang because there was no time before the Big Bang. We have finally found something that doesn't have a cause, because there was no time for a cause to exist in. For me this means that there is no possibility of a creator, because there is no time for a creator to have existed in. It's my view that the simplest explanation is that there is no God. No one created the universe and no one directs our fate. This leads me to a profound realization: there is probably no heaven and afterlife either. I think belief in an afterlife is just wishful thinking. There is no reliable evidence for it, and it flies in the face of everything we know in science. I think that when we die we return to dust. But there's a sense in which we live on, in our influence, and in our genes that we pass on to our children. We have this one life to appreciate the grand design of the universe, and for that I am extremely grateful.

(Hawking, S., (2018), *Brief answers to the big questions*).

Agradecimientos

Dado que mi andadura como investigador comenzó hace ya varios años, las personas a las que agradecer haber llegado hasta aquí se han ido acumulando, y siendo consciente de que con mi memoria es imposible que no se me escape alguien, optaré por hacer esto lo más resumido posible.

En cuanto a lo profesional, quiero agradecer a todos los que en algún momento me habéis ayudado, tanto en mi trayectoria investigadora como en la docente. Gracias a los miembros del Departamento de Matemáticas de la UPV/EHU, por hacer que ir a trabajar sea algo agradable, a pesar de los contratiempos que puedan surgir.

Me siento tentado de mencionar a personas concretas a las que me gustaría dirigir un agradecimiento especial, pero espero que si leen estas líneas se den por aludidas, ya sea por haber sido mis mentores, mis referentes, o por haber querido compartir un café conmigo.

Sin embargo, debo hacer dos excepciones, la primera, con mi directora de tesis y hasta muy poco directora del departamento, Virginia Muto. Quiero agradecerte todo tu trabajo para ayudarme a llegar hasta aquí. De corazón te digo que ha sido un privilegio trabajar con alguien que se preocupa tanto por el bienestar de sus compañeros.

La segunda, con mi exdirector de la primera tesis y amigo, Luis Martínez. Gracias por tu apoyo durante todos estos años, y también por no rendirte con las líneas de investigación que compartimos, tan poco reconocidas y valoradas. Mis éxitos serán siempre en parte tuyos.

Para terminar, quiero agradecer a Vir, por quererme tal como soy; a mi familia, por apoyarme incondicionalmente en todo momento; y cómo no, a mis amigos, por aguantarme sin matices. De verdad que me siento muy afortunado de teneros a todos en mi vida.

Finalmente, como es habitual, cito a continuación algunas de las fuentes de financiación que durante esta tesis han hecho que nuestros trabajos fueran posibles:

- Proyecto de investigación "Groups, topology and applications" financiado por el Gobierno Vasco. Ref: IT974-16.

- Proyecto de investigación "Investigación multidisciplinar en nuevas estrategias para el diagnóstico temprano y tratamiento personalizado del cáncer" financiado por el Gobierno Vasco. Ref: Elkartek18/11.

- Proyecto de investigación "Aplicaciones de la matemática discreta al diseño computacional de vacunas y de ADN-códigos" financiado por la UPV/EHU y el BCAM. Ref: US18/21.

Índice

Introducción

"*Las matemáticas poseen no solo la verdad, sino cierta belleza suprema. Una belleza fría y austera, como la de una escultura*" (North & Russell, 1913). Gracias a la objetividad y el rigor que aportan las matemáticas, su importancia ha ido creciendo con el paso de los siglos, haciéndose cada vez más importantes en el resto de ciencias, donde cada vez es más habitual ver grupos multidisciplinares que integran a un especialista en ciencias exactas.

La biomedicina, campo sobre el cuál se profundizará en esta tesis, no es una excepción. Durante toda la historia, la relación entre las matemáticas y la biomedicina se ha ido estrechando. Uno de los primeros en utilizar las matemáticas para predecir sucesos biológicos fue Leonardo de Pisa (c. 1175-1250) (más conocido como Fibonacci), quién desarrolló el famoso modelo matemático para controlar el crecimiento de los conejos (publicado el 1202, y traducido al inglés moderno en (Fibonacci, 2002)). Sin embargo, no fue hasta la era Moderna cuando la sinergia entre las matemáticas y la biomedicina se hizo realmente evidente. Entre las figuras más relevantes cabe destacar a Karl Pearson (1857-1936), quien formuló matemáticamente una gran variedad de problemas biológicos y en particular relativos a la evolución, y cuyos estudios inspiraron a otro gran matemático y biólogo llamado Roland Fisher (1890-1962), que, entre sus logros, fue quien desarrolló el Análisis de la Varianza (ANOVA) (Fisher, 1950).

Sin embargo, si tuviera que destacar la contribución de una única persona a la matematización de la biomedicina, sería la de Francis Galton (1822-1911). Entre sus aportaciones, se encuentran el haber ayudado a desarrollar el carácter combinatorio de la genética Mendeliana (Galton, 1877), demostrar que varias características de la población, tanto psíquicas como físicas, siguen una distribución normal (Galton, 1889/1), ser el primero en definir el coeficiente de correlación (Galton, 1889/2), o cofundar *Biometrika*, la primera revista científica dedicada a promover el estudio de las matemáticas y la estadística en la biología.

Los esfuerzos de éste y otros muchos matemáticos, biólogos y médicos, han hecho que hoy en día sea más natural el abordar problemas biomédicos desde las matemáticas. En esta tesis, profundizaremos en el diseño computacional de vacunas, tanto universales como personalizadas, una disciplina enmarcada en el campo de la inmunología.

Este campo de las ciencias biomédicas se encarga del estudio del sistema inmunitario, que tiene como función el reconocer patógenos y elementos perjudiciales en nuestro organismo, para posteriormente generar una respuesta contra ellos (Murphy & Weaver, 2016). Gracias a los avances tecnológicos, el volumen de datos inmunológicos con los que trabajar se ha incrementado enormemente (Tong & Ren, 2009), lo que ha hecho necesaria la utilización de herramientas computacionales y técnicas matemáticas para poder procesar dicha información. Los trabajos que involucran las matemáticas en esta área son muy variados, pero cabe destacar los estudios sobre la diversidad del virus del dengue (Khan et al., 2006/1), o la variabilidad del virus causante de la gripe A (Heiny et al., 2007).

A continuación, se presenta la contextualización específica de las 4 investigaciones sobre el diseño de vacunas presentadas en esta tesis. Estas investigaciones pueden abordarse como dos bloques diferenciados dentro del mismo campo, uno sobre vacunas antivirales universales (investigaciones nº1 y nº2), y otro sobre vacunas antitumorales personalizadas (investigaciones nº3 y nº4). Esta diferenciación se debe a que, la virtud principal del método que presentamos en esta tesis para el diseño de vacunas, las λ-supercadenas, radica en su capacidad para proteger contra todas las variantes seleccionadas de un patógeno. Sin embargo, al obtener la información para desarrollar vacunas personalizadas, y en particular, antitumorales, no se dispone de una secuenciación del ADN de cada célula del tumor de manera individualizada (lo que podría situarnos en un escenario similar al de las vacunas antivirales y nos permitiría aplicar el principio de λ-supercadena), sino que se obtiene información de las mutaciones de manera "más global". Esto implica que el abordaje deba ser distinto, como se verá más adelante.

I.1. Antecedentes de la investigación nº1

Las enfermedades infecciosas son la causa diaria de la muerte de millones de personas, y la mejor forma de prevenirlas es la vacunación. Por lo tanto, los inmunólogos centran sus esfuerzos principalmente en mejorar la predicción de antígenos eficientes que nos protejan contra los patógenos (Nielsen et al., 2010), y mejorar así la selección de antígenos a incluir en la vacunación (Khan et al., 2006/2). Existen dos tipos de inmunidad, y su eficiencia varía según el patógeno. La inmunidad humoral implica la producción de anticuerpos a través de las células B que interactúan con la superficie de los patógenos, o con las toxinas que secretan. Cada anticuerpo se une a un antígeno, que es una estructura tridimensional de aminoácidos que puede ser contactada a través de la región variable del anticuerpo. Por otro lado, la inmunidad celular depende de los antígenos para las células T que son generados por otras células, como las células presentadoras de antígenos, que generan antígenos a partir de la degradación de patógenos o la síntesis de proteínas. Estas cortas cadenas de aminoácidos generadas a partir de la degradación intracelular o de la síntesis de proteínas, se unen a las moléculas del complejo mayor de histocompatibilidad (MHC), que se compone de dos tipos de moléculas: las moléculas de clase I (denotado HLA-I en humanos, por Human Leucocyte Antigen), que son reconocidas por los linfocitos T CD8+ y a las que se unen antígenos compuestos por 8-9 aminoácidos, y las de clase II (HLA-II), que son reconocidas por los linfocitos T CD4+ y que aceptan péptidos más largos, habitualmente entre 12-15 aminoácidos (Hemmer et al., 2000).

Con el objeto de optimizar la selección de estos antígenos, nuestro grupo introdujo en Martínez et al. 2015, el criterio de ser una λ-supercadena, junto con el problema de optimización combinatoria asociado. En resumen, el criterio dice que, dados dos conjuntos de cadenas, uno llamado *host strings* y representando en este caso las variantes de proteína de los virus expresadas en aminoácidos, y otro llamado *target strings* representando los antígenos, una λ-supercadena será una cadena candidata a vacuna que contendrá, como subcadenas, al menos λ antígenos de cada variante (esto es, de cada *host string*). Esto significa que cada vacuna recubre al menos λ antígenos de cada paciente (con el objeto de aclarar los elementos del diseño de vacunas a través de λ-supercadenas, en la Figura I.1 se ilustran, de manera esquemática, dichos componentes). El problema combinatorio asociado es el de encontrar λ-supercadenas de longitud mínima, lo que se traduce en hallar un candidato a vacuna lo más corto posible, para un λ dado. En Martínez et al. 2015, se mostró que este problema puede resolverse de manera exacta (en nuestro caso fue resuelto a través de programación entera, ver Material y Métodos M.3), o también de manera aproximada (utilizando por ejemplo algoritmos genéticos multiobjetivo (ver Material y Métodos M.4)).

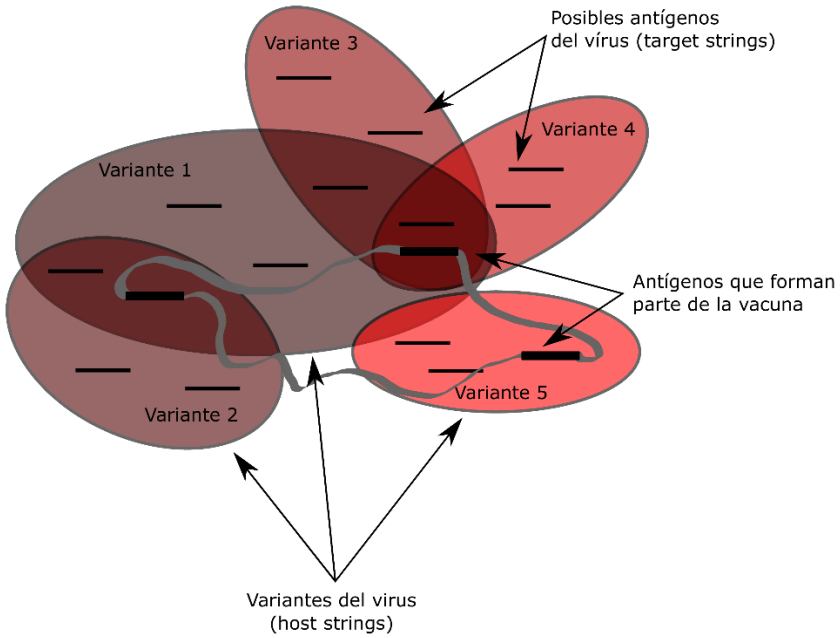

Figura I.1. Esquema ilustrativo de los componentes del diseño de vacunas. En este esquema se representan los elementos principales del diseño de vacunas utilizando λ-supercadenas. Los óvalos representan las diferentes variantes del virus (host strings); las rayas horizontales son los potenciales antígenos (target strings); las rayas más gruesas son los antígenos seleccionados para formar parte de la vacuna; finalmente, la banda gris representa la vacuna, compuesta por los antígenos seleccionados.

En este trabajo, hemos generalizado el concepto de λ-supercadena, considerando una función peso para cada antígeno representando su inmunogenicidad, acercando así el criterio a la realidad biológica, donde distintos antígenos generan respuestas del sistema inmunológico cuantitativamente distintas. En resumen, una λ-supercadena ponderada se define como una cadena tal que, para cada variante de proteína

considerada, el mínimo de las sumas de las inmunogenicidades de los antígenos que están tanto en la variante como en el candidato a vacuna, es al menos λ.

Este trabajo fue publicado en PloS one (Q1 en el campo Multidisciplinar de acuerdo con SCOPUS) como: Martínez, L., Milanič, M., Malaina, I., Álvarez, C., Pérez, M. B., & M. de la Fuente, I. (2019). Weighted lambda superstrings applied to vaccine design. *PloS one*, 14(2), e0211714.

I.2. Antecedentes de la investigación nº2

La epidemia del coronavirus disease 19 (COVID-19), ha representado la mayor amenaza a la salud global de la humanidad, con más de 505 millones de personas infectadas y más de 6.2 millones de muertes desde que se detectó hace 2 años (COVID-19 website, https://covid19.who.int). Para acabar con la pandemia, se han desarrollado varios tipos de vacunas de manera acelerada (Awadasseid et al., 2021; Flanagan et al., 2020; Krammer, 2020; Gaebler y Nussenzweig, 2020; Poland et al., 2020). A pesar de que se están aplicando nuevas tecnologías para la producción de vacunas, como las vacunas de ARN mensajero (Wang et al., 2020; Yi et al., 2020), éstas siguen habiéndose desarrollado con la concepción antigua de diseño de vacunas, donde se protege contra los virus más frecuentes, y no contra todas las variantes posibles hasta la fecha. Por otro lado, cada vez es más frecuente el uso de las herramientas bioinformáticas para abordar diferentes aspectos de la enfermedad, como modelizar el virus (Estrada, 2020), identificar mapas de antígenos (Sikora et al., 2021), diseñar proteínas inhibidoras (Jaiswal y Kumar, 2020), u optimizar anticuerpos (Chen et al., 2021).

En este trabajo, basándonos en estimaciones bioinformáticas de las principales cualidades de los antígenos, se ha diseñado una vacuna peptídica basada en las λ-supercadenas ponderadas (Martínez et al., 2019) que es capaz de proteger contra todas las variantes de virus consideradas, y de esta manera, puede minimizar la probabilidad de que una nueva variante de virus pueda evitar nuestra vacuna. La bondad de las vacunas peptídicas radica en tres criterios: primero, es fácil de manufacturar; segundo, el coste-efectividad de su producción es alto; y tercero, son más seguras que las vacunas con virus completos (Flanagan et al., 2020; Dong et al., 2020; Hodgson et al., 2021). Sin embargo, estas vacunas tienen la limitación de que hace falta elegir adecuadamente los antígenos en cuestión, y dado el número de posibilidades y de combinaciones entre ellos, es humanamente imposible hacerlo en el laboratorio.

Es aquí donde tiene cabida nuestra metodología, ya que, como se ha mencionado, a través de herramientas bioinformáticas somos capaces de ponderar los antígenos que, aplicando de un algoritmo de programación entera (Material y Métodos, M.3) y exigiendo el criterio de λ-supercadena, formarán candidatos a vacuna que sorteen el defecto de las vacunas peptídicas.

Este trabajo ha sido publicado en Scientific Reports (Q1 en el campo Multidisciplinar) como: Martínez, L., Malaina, I., Salcines-Cuevas, D., Terán-Navarro, H., Zeoli, A., Alonso, S., de la Fuente, I.M., González-López, E., Ocejo-Vinyals, J.G., Gonzalo-Margüello, M., Calvo-Montes, J., & Alvarez-Dominguez, C. (2022). First

computational design using lambda-superstrings and in vivo validation of SARS-CoV-2 vaccine. *Scientific Reports*, 12(1), 1-12.

I.3. Antecedentes de la investigación nº3

En los últimos años, la vacunación antitumoral personalizada ha surgido como una prometedora e innovadora alternativa para el tratamiento de varios tipos de cáncer (Vormehr et al., 2020; Vermaelen, 2019; Sahin et al., 2017; Kakimi et al., 2017). El elemento clave para elegir la vacunación personalizada contra las células cancerosas es que, por un lado, los tumores tienen un gran número de mutaciones, y, por otro lado, que aproximadamente el 95% de dichas mutaciones parecen únicas y particulares para dicho tumor (Stratton, 2011). Por lo tanto, estas mutaciones constituyen una diana ideal para el tratamiento individualizado de tumores (Kreiter et al., 2012), y en particular, para la vacunación personalizada.

Sin embargo, a pesar de que haya un gran número de mutaciones en los tumores, para desarrollar una vacuna efectiva, el primer paso es separar las mutaciones que sólo suceden en el tumor, de las mutaciones que se dan en células no cancerosas. Así, se presenta el concepto de neoantígeno como un tipo de péptido surgido a partir de mutaciones específicas del tumor, y se une al complejo mayor de histocompatibilidad (MHC) (Peng et al., 2019). Estos neoantígenos no han sido detectados previamente por el sistema inmunitario, y por lo tanto el organismo no aplica mecanismos de tolerancia contra ellos, siendo por lo tanto reconocibles como ajenos (Vormehr et al., 2019).

Dado que hay miles de mutaciones que podrían dar lugar a tantos posibles neoantígenos, y como nos interesarían las que tengan mayores probabilidades de generar una buena respuesta inmunitaria, el siguiente paso sería evaluar experimentalmente las características de cada neoantígeno de manera individual, lo que es demasiado costoso (a día de hoy irrealizable) tanto en tiempo como en dinero. Por suerte, en las últimas décadas se han desarrollado varias herramientas bioinformáticas que son capaces de predecir algunas de las propiedades biológicas de los péptidos, y en particular, de los potenciales neoantígenos, permitiendo así realizar una primera criba en la conocida como fase *in silico*.

A pesar de que el uso de estas técnicas va en aumento, una vez estimadas las características bioinformáticas, la manera de combinarlas para dar lugar a una potencial vacuna se ha basado en algoritmos de ordenación muy simples, que son incapaces de seleccionar los mejores neoantígenos considerando todas las variables simultáneamente (Sahin et al., 2017).

En este trabajo, a la hora de seleccionar la mejor combinación de neoantígenos, hemos utilizado funciones agregativas lineales, eligiendo soluciones del frente de Pareto, que dan lugar a soluciones más equilibradas que los órdenes lexicográficos habituales. Finalmente, hemos sintetizados dicha vacuna, encapsulada en partículas de ácido poliláctico-co-glicolítico recubiertas de polietilimina, y hemos testado *ex vivo* (esto es, reproduciendo en un ambiente externo, con toda la precisión posible, las condiciones naturales originales del organismo) su efectividad y especificidad.

Dicho trabajo está bajo revisión en BMC Bioinformatics (Q1 en el campo Mathematics: Applied Mathematics) bajo el nombre de: Computational and

experimental evaluation of the immune response of neoantigens for personalized vaccine design (2021), y ha sido elaborado por: Malaina, I., Martínez, L., González-Melero, L., Salvador, A., Sánchez-Díez, A., Asumendi, A., Margareto, J., Carrasco, J., Legarreta, L., García, M.A., Pérez, M.B., Izu, R., De la Fuente, I.M., Igartua, M., Alonso, S., Hernández, R.M., & Boyano, M.D.

I.4. Antecedentes de la investigación nº4

A pesar del creciente interés en el desarrollo de vacunas antitumorales personalizadas, y del hecho de que utilizar neoantígenos como diana ha obtenido resultados prometedores (Tanyi et al., 2018; Hu et al., 2018), cuando se considera todo el espectro de mutaciones del tumor (conocido como mutanoma), el número de posibles neoantígenos es demasiado grande como para evaluarlos uno a uno en el laboratorio. Por otro lado, si eligiéramos alguno de estos neoantígenos "a ciegas", no hay garantía alguna de que estos vayan a generar una respuesta inmunitaria alta.

Como alternativa, como se ha mencionado en el punto I.3, se han desarrollado varias herramientas bioinformáticas para la predicción de propiedades biológicas a partir de su secuencia de aminoácidos (Lundegaard et al., 2011; Zhang et al., 2008; Soria-Guerra et al., 2015).

Dado que generar una respuesta inmunitaria contra un péptido mutado depende directamente de la capacidad del HLA del paciente para anclarse a estos neoantígenos, y presentarlos a los linfocitos (Fritsch et al., 2014), se ha considerado un buen primer criterio para elegir los neoantígenos más prometedores el coger los que se unan más eficientemente a dichas moléculas. Para estimar esta propiedad, las herramientas más utilizadas son las desarrolladas por IEDB para la estimación de afinidad de unión a las moléculas HLA-I y HLA-II (Zhang et al., 2008).

Sin embargo, a pesar de haber sido ampliamente usadas para estimar la afinidad de unión de péptidos en general, surge la cuestión siguiente: dado que los neoantígenos son capaces a priori de generar una respuesta inmunitaria, y solo difieren de su versión no-mutada en un único aminoácido, y ésta no genera respuesta alguna, ¿son capaces las herramientas bioinformáticas de detectar cuantitativamente esta diferencia? o dicho de otro modo, ¿es cierto que las versiones mutadas de los péptidos (neoantígenos) son más inmunogénicos que sus versiones no mutadas, de acuerdo con estas herramientas bioinformáticas?

En esta investigación, con el objeto de responder a dicha pregunta, hemos secuenciado y analizado parte del mutanoma de seis pacientes que presentaban melanoma cutáneo, y hemos comparado la estimación de afinidad de unión a las moléculas HLA-I y II de los neoantígenos, y su respectiva versión no mutada.

Este trabajo ha sido publicado en LNCS (Q4 en el campo "Mathematics: Theoretical Computer Science", y Q2 en "Computer Science") como: Malaina, I., Legarreta, L., Boyano, M.D., Alonso, S., De la Fuente, I.M., Martinez, L. (2020). Analyzing the Immune Response of Neoepitopes for Personalized Vaccine Design. In: Rojas, I., Valenzuela, O., Rojas, F., Herrera, L., Ortuño, F. (eds) Bioinformatics and

Biomedical Engineering. IWBBIO 2020. *Lecture Notes in Computer Science*, vol 12108. Springer, Cham.

Como resultado, en esta tesis presentamos cuatro estudios (tres de ellos publicados) donde se han utilizado las matemáticas en el diseño de vacunas, que posteriormente se han testado, ofreciendo resultados prometedores. A través de dichos trabajos, se espera evidenciar la sinergia (cada vez más difícilmente discutible) entre las ciencias exactas y las ciencias de la salud.

Finalmente, a modo de anexo, además de las publicaciones íntegras, se incluye el registro y descripción de una patente para una vacuna contra el SARS-CoV-2 diseñada a partir de técnicas cuantitativas y algoritmos de optimización heurísticos, basada en la metodología descrita en esta tesis.

Material y Métodos

λ-supercadenas ponderadas

A continuación, presentamos una sección donde se profundizará en el concepto de λ-supercadena ponderada, un criterio combinatorio desarrollado por nuestro grupo de investigación, que ha sido posteriormente aplicado al diseño computacional de vacunas, y en particular, en las investigaciones nº1 y nº2 de esta tesis. Para llegar a este concepto, primero, vamos a presentar en qué consiste dicho criterio combinatorio. En lo que sigue, se expondrán y demostrarán algunos resultados teóricos, asociándolos a un problema de teoría de grafos. Finalmente, se presentarán dos maneras de obtener λ-supercadenas ponderadas óptimas y subóptimas, la primera, a través de programación entera, y la segunda, utilizando una modificación de un algoritmo genético multiobjetivo.

Nótese que toda la teoría sobre las λ-supercadenas (definiciones, teoremas y proposiciones) han sido desarrolladas íntegramente por nuestro grupo. Además, tanto el algoritmo de programación entera como el algoritmo genético han sido diseñados y desarrollados por nuestro grupo (a pesar de que este último tenga como punto de partida el conocido algoritmo genético de optimización multiobjetivo NSGA-II).

M.1. Introducción al concepto de λ-supercadena

Sea A un alfabeto finito (en nuestro caso formado por los 20 aminoácidos) y $A^* = \bigcup_{n=1}^{\infty} A^n \cup \{\theta\}$ el conjunto de todas las posibles cadenas formadas por elementos de A, donde θ representa la cadena vacía. El conjunto A^* constituye un semigrupo para la operación concatenación (que denotaremos como $+$), donde:

$$t + t' = (t_1, \ldots, t_n) + (t'_1, \ldots, t'_m) = (t_1, \ldots, t_n, t'_1, \ldots, t'_m).$$

Es fácil ver que para cualesquiera dos elementos de A^*, su concatenación, como hemos indicado arriba, será otra cadena, que a su vez será un elemento de A^*; por otro

lado, es evidente que se cumple la propiedad asociativa, ya que el orden en el que se aplique la concatenación no altera el resultado. Por lo tanto, se tiene que $(A^*, +)$ es un semigrupo. Además, θ es el elemento neutro para la operación, por lo que además de ser un semigrupo es un monoide.

Diremos que una cadena $t = (t_1, \dots, t_m)$ es una *subcadena* de otra cadena $h = (h_1, \dots, h_n)$ cuando podamos escribir $h = a + t + b$ para algunas cadenas $a, b \in A^*$. Definimos a continuación el grado de solapamiento entre dos cadenas t y t' como:

$$solapamiento(t, t') = max\{i \in \{0,1, \dots, min\{m, n\}\} \mid t_{n-i+j} = t_j{}', para\ j = 1, \dots, i\}.$$

Definimos por lo tanto la operación de suma con solapamiento $+'$ en un conjunto A^* como:

$$(t_1, \dots, t_n) +' (t_1{}', \dots, t_m{}') = \left(t_1, \dots, t_{n-solapamiento(t,t')}\right) + (t_1{}', \dots, t_m{}').$$

Nótese que el conjunto A^* con la operación $+'$ no es un semigrupo, porque la propiedad asociativa no se mantiene: sean $t_1 = t_3 = a$, y $t_2 = t_4 = b$, entonces $\left((t_1 +' t_2) +' t_3\right) +' t_4 = abab$, mientras que $(t_1 +' t_2) +' (t_3 +' t_4) = ab$.

Consideramos ahora un conjunto $T \subset A^*$ como el conjunto de cadenas objetivo (*Target Strings)*, que en nuestro caso se corresponderán con los posibles antígenos (sustancias que pueden ser reconocidas por el sistema inmunitario e inducir una respuesta que genere anticuerpos), y el valor $\lambda \in \mathbb{N}$, que será el valor que maximizaremos y que al generalizarlo más adelante podrá tomar valores reales. Con estos elementos, podemos escribir la definición de λ-supercadena:

Definición 1. Sean $H_1, \dots, H_k \subseteq A^*$ (las posibles variantes del virus, denominadas también *Host Strings*) y $T \subseteq A^*$ (los posibles antígenos), entonces llamando $C(h, v)$ al conjunto de todas las subcadenas comunes de h y v, definimos una **λ-supercadena** v (el candidato a vacuna) para el conjunto $(H_1, \dots, H_k, T)$ como una cadena $v \in A^*$ tal que $|C(H_i, v) \cap T| \geq \lambda,\ \forall\ i = 1, \dots, k$.

M.2. Resolución del *Shortest weighted λ-superstring problem*

Para llegar a presentar el *Shortest weighted λ-superstring problem*, vamos a comenzar primero exponiendo la versión no ponderada, esto es, el *Shortest λ-superstring problem*.

Shortest λ-superstring problem
Sean $H_1, \dots, H_n \subseteq A^*$ un conjunto finito de cadenas del alfabeto A, sea $T \subseteq A^*$ el conjunto de cadenas objetivo, y sea $\lambda \in \mathbb{N}$, encontrar una λ-supercadena $v \in A^*$ para $(H_1, \dots, H_k, T)$ de longitud mínima.

Éste generaliza tanto el *Shortest common superstring problem*, como el *Set cover problem*:

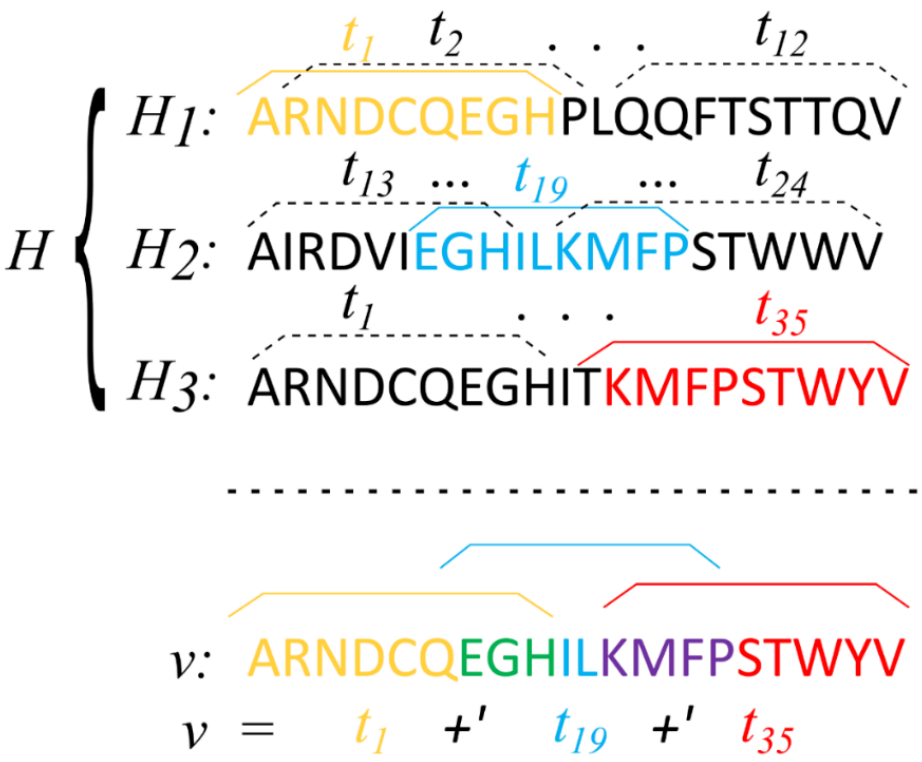

Figura M.1. Obtención de una λ-supercadena. En este ejemplo, hay 3 variantes de virus (H_1, H_2, H_3), 35 posibles antígenos distintos de longitud 9 ($t_1, t_2, \dots, t_{35} \in T$, nótese que t_1 aparece en H_1 y en H_3), y el candidato a vacuna resultante es la suma con solapamiento $v = t_1 +' t_{19} +' t_{35}$. Por lo tanto, se trata de una 1-supercadena.

Shortest common superstring problem
Dado un conjunto finito $T \subseteq A^*$ de cadenas de un alfabeto A, encontrar una cadena $v \in A^*$ tal que contiene a cada cadena $t \in T$ como subcadena, y es lo más corta posible.

Set cover problem
Dado un conjunto finito de elementos U y un grupo de subconjuntos F cuya unión sea U, encontrar el menor número de conjuntos de F tales que su unión siga siendo U.

Además, como veremos más adelante, es polinomialmente reducible (y polinómicamente equivalente) al *Shortest λ-cover superstring problem*:

Definición 2. Una **λ-cover superstring** es una cadena $v \in A^*$ tal que dados un conjunto $(H_1, \dots, H_k)$ y un valor $\lambda \in \mathbb{N}$, $\forall i = 1, \dots, k$ al menos λ elementos de H_i son subcadenas de v.

Shortest λ-cover superstring problem
Sean $H_1, \dots, H_n \subseteq A^*$ un conjunto finito de cadenas del alfabeto A, y sea $\lambda \in \mathbb{N}$, encontrar una λ-cover superstring $v \in A^*$ para $(H_1, \dots, H_k)$ de longitud mínima.

Dada esta equivalencia, varios de los resultados que expondremos a continuación se probarán para el *Shortest λ-cover superstring problem*, y posteriormente se extenderán al *Shortest λ-superstring problem.* Por lo tanto, inicialmente, daremos una serie de definiciones y demostraremos la equivalencia previamente mencionada, para posteriormente formular el *Shortest λ-cover superstring problem* a través de la teoría de grafos.

Definición 3. Decimos que el tiempo de ejecución de un algoritmo $T(n)$ es $O(f(n))$ si hay constantes c y n_0 tales que $T(n) \leq c \cdot f(n), \forall n \geq n_0$.

Por ejemplo, si el tiempo de ejecución del algoritmo es $T(n) = 2n^4 + n$, entonces el algoritmo es $O(n^4)$, ya que para $n_0 = 0$ y $c = 3$ se tiene que $2n^4 + n \leq 3n^4, \forall n \geq 0$.

Definición 4. Se dice que un algoritmo/función es de **tiempo polinomial** si su tiempo de ejecución $T_{algoritmo}$ (suma del tiempo de ejecución de todos los fragmentos del código) está limitado por una expresión polinómica, esto es, dado una dimensión de la entrada n, entonces $T_{algoritmo}(n) = O(n^k)$, para algún k positivo, donde O es la notación para la cota superior asintótica.

Veamos este concepto a través de un ejemplo. Si definimos un algoritmo como:

bucles(n):

$x = 0$;

for $i = 1{:}\,n$

for $j = 1{:}\,n$

$x = x + 1$;

end

end

Entonces este algoritmo será de tiempo polinomial, y en particular $T_{bucles}(n) = O(n^2)$, porque el total de iteraciones es $n \cdot n$ (n veces el primer bucle y dentro de cada uno, n veces la operación).

Definición 5. Dados dos problemas de minimización Π_1 y Π_2, diremos que Π_1 es **polinomialmente reducible** a Π_2 si cada elemento de entrada I_1 de Π_1 puede ser llevado a través de una función F en tiempo polinomial a los elementos de entrada I_2 de Π_2, de manera que las siguientes dos condiciones se mantengan:

1. $F_{\Pi_1}(I_1) = F_{\Pi_2}(I_2)$,
2. $f_1(x) = f_2(x), \forall x \in F_{\Pi_1}(I_1)$, donde f_i es la función objetivo del problema Π_i, para $i = 1{,}2$.

Definición 6. Diremos que dos problemas de minimización son **polinomialmente equivalentes**, si cada uno de ellos es polinomialmente reducible al otro.

A continuación, probaremos el teorema que dice que el *Shortest λ-superstring problem* y el *Shortest λ-cover superstring problem* son polinomialmente equivalentes:

Teorema 1. El *Shortest λ-superstring problem* y el *Shortest λ-cover superstring problem* son polinomialmente equivalentes.

Demostración. Para demostrar este teorema, primero probaremos que el *Shortest λ-superstring problem* es polinomialmente reducible al *Shortest λ-cover superstring problem*, y a continuación probaremos la afirmación recíproca, esto es, que el *Shortest λ-cover superstring problem* es polinomialmente reducible al *Shortest λ-superstring problem*.

Primera parte: Sean $I = (A, H_1, \dots, H_k, T, \lambda)$ los elementos de entrada del *Shortest λ-superstring problem*. Entonces describimos una transformación en tiempo polinomial para convertir I en una entrada equivalente del *Shortest λ-cover superstring problem* $I' = (A', X_1, \dots, X_n, \lambda')$:

1. Hacemos $n = k, \lambda' = \lambda$, y $A' = A$.
2. Para cada $i \in \{1, \dots, n\}$, definimos X_i como el conjunto de todas las cadenas $t \in T$ que son subcadenas de H_i.

Claramente, I' puede ejecutarse a partir de I en tiempo polinomial. Veamos ahora que el conjunto de soluciones factibles para ambos problemas es el mismo. Primero, supongamos que $v \in A^*$ es una solución factible (esto es, una solución que satisface todas las restricciones) del *Shortest λ-superstring problem* dada una entrada I. Entonces, para cada $i \in \{1, \dots, n\}, \exists$ un subconjunto $T_i \subseteq T$ con cardinal al menos λ, tal que todo elemento de T_i es una subcadena tanto de H_i como de v (por ser una solución). En particular, $T_i \subseteq X_i$ y cada elemento de T_i es una subcadena de v. Por lo tanto, v es una solución factible del *Shortest λ-cover superstring problem* para la entrada I'.

Del mismo modo, supongamos que $v \in A^*$ es una solución factible del *Shortest λ-cover superstring problem* para la entrada I'. Entonces, para cada $i \in \{1, \dots, n\}$, existe un subconjunto $T_i \subseteq X_i$ de cardinal al menos λ para el cual todos los elementos son subcadenas de v. Cada elemento T_i, por definición de X_i, es un elemento de T y una subcadena de H_i. Por lo tanto, al menos hay λ *target strings* distintas que son subcadenas tanto de H_i como de v, y, por lo tanto, v es una solución factible del *Shortest λ-superstring problem* para una entrada I. Finalmente, dado que la segunda condición de la definición 4 se obtiene directamente de la definición de ambos problemas, esta parte de la demostración está completa.

Segunda parte: Sean $I = (A, X_1, \dots, X_n, \lambda)$ los argumentos de entrada del *Shortest λ-cover superstring problem*. Describimos ahora una transformación en tiempo polinomial de I a una entrada equivalente $I' = (A', H_1, \dots, H_k, T, \lambda')$ del *Shortest λ-superstring problem*:

1. Fijamos $k = n, A' = A \cup \{*\}$, donde $* \notin A$ y $\lambda' = \lambda$.
2. Para cada $i \in \{1, \dots, n\}$, sea $X_i = \{x_1^i, \dots, x_{n_i}^i\}$. Construimos la cadena H_i como concatenación de todas las cadenas en X_i separadas por $*$:
$$H_i = x_1^i + * + x_2^i + * + \cdots + * + x_{n_i}^i.$$
3. Fijamos $T = \bigcup_{i=1}^{n} X_i$.

Claramente, I' puede obtenerse a partir de I en tiempo polinomial. Ahora, veamos que el conjunto de soluciones factibles de ambos problemas es el mismo. Supongamos que $v \in A^*$ es una solución factible del *Shortest λ-cover superstring problem* dada la entrada I. Para cualquier índice $i \in \{1, \dots, k\}$, como estamos suponiendo que v es solución, existe un subconjunto $T_i \subseteq X_i$ de cardinal al menos λ para el cual todos sus elementos son subcadenas de v. Sea $t \in T_i$. Como $T_i \subseteq X_i$, se tiene que $t = x_j^i$ para algún $j \in \{1, \dots, n_i\}$. Por lo tanto, t es una subcadena de H_i. Además, por construcción de T, también se tiene que $t \in T$. En particular, T_i es un conjunto de λ cadenas de T que son a su vez subcadenas comunes tanto de H_i como de v. Como $A \subseteq A'$, $v \in (A')^*$, y, por lo tanto, v es una solución factible del *Shortest λ-superstring problem* dada la entrada I'.

Asimismo, supongamos que $v \in (A')^*$ es una solución factible del *Shortest λ-superstring problem* para una entrada I'. Nótese que el símbolo $*$ no está presente en ninguna cadena de T. Para cualquier $i \in \{1, \dots, n\}$, existe un conjunto $T_i \subseteq T$ formado por al menos λ cadenas comunes tanto de H_i como de v. Como ningún T_i contiene el símbolo $*$, se tiene que $T_i \subseteq A^*$. En particular, dada la estructura de H_i, $\forall t \in T_i, \exists j \in \{1, \dots, n_i\} | t = x_j^i \in X_i$. Por lo tanto, el conjunto T_i es un subconjunto de X_i de cardinal al menos λ, donde todos los elementos son subcadenas de v, lo que implica que v es una solución factible del *Shortest λ-cover superstring problem* para una entrada I. Finalmente, como sucedía en la primera parte, dado que la segunda condición de la definición 4 se obtiene directamente de la definición de ambos problemas, la otra parte de la prueba queda demostrada. ■

Como se ha mencionado, utilizando esta equivalencia, en vez de resolver el *Shortest λ superstring problem*, se procederá a la resolución del *Shortest λ-cover superstring problem*, utilizando la teoría de grafos. Nuestra aproximación al problema se basa en modelarlo como una generalización del *generalized Traveling Salesman Problem* para grafos dirigidos presentado en (Henry-Labordere, 1969; Saksena, 1970; y Srivastava et al. 1969):

Generalized Traveling Salesman Problem

Sea $G = (V, E)$ un grafo completo dirigido, donde $V = \{v_1, \dots, v_n\}$ representan el conjunto de ciudades y un conjunto de clusters $W = \{W_1, \dots, W_m\}$, con $0 < m \leq n$, y donde cada ciudad $v_i \in V$ pertenece a un único cluster. Sea c_{ij} el "coste" de viajar entre las ciudades v_i y v_j. El objetivo es encontrar un ciclo dirigido que visite exactamente una única vez cada cluster, y que minimice la suma de los costes del viaje.

Sean $(A, H_1, \dots, H_n, \lambda)$ los elementos del *Shortest λ-cover superstring problem*, se construye un grafo ponderado dirigido completo $G = (V, E, c)$ (que denotaremos como *grafo de distancia*) donde V es el conjunto de vértices, E el de aristas, y c es la función de coste, de la siguiente manera:

- El conjunto de vértices V está compuesto por el conjunto de todas las cadenas de entrada, junto con un nuevo vértice s^*:

$$V = \bigcup_{i=1}^{n} H_i \cup \{s^*\},$$

- Para cada dos vértices distintos $s, t \in V \backslash \{s^*\}$, añadimos la arista (s, t) a E y le asignamos el coste $c_{s,t} = l(s) - solapamiento(s, t)$. Como puede verse, los costes están bien definidos y son no-negativos.
- Para cada vértice $s \in V \backslash \{s^*\}$, añadimos la arista (s, s^*) a E y le asignamos el coste $c_{s,s^*} = l(s)$.
- Para cada vértice $s \in V \backslash \{s^*\}$, añadimos la arista (s^*, s) a E y le asignamos el coste $c_{s^*,s} = 0$.

De acuerdo con la siguiente proposición, resolver el *Shortest λ-cover superstring problem* es equivalente a encontrar un ciclo dirigido C en G a través de s^* de longitud mínima, sujeto a la restricción de que para cada H_i, al menos λ cadenas de H_i aparecen como vértices de C.

Definición 7. Se dice que un subgrafo H de G **cubre** una cadena $s \in T$ si existe un vértice $t \in V(H) \cap T | s \subseteq t$.

Para cada $i \in \{1, \dots, n\}$, denotamos el conjunto de todas las cadenas en H_i cubiertas por H como $V_i(H)$.

Definición 8. diremos que un ciclo dirigido C es **factible** si satisface las siguientes condiciones:

- $s^* \in V(C)$;
- para cada dos vértices distintos $s, t \in V(C) \cap T$, s no es una subcadena de t;
- $\forall i \in \{i, \dots, n\}$, se tiene que $|V_i(C)| \geq \lambda$.

Definiendo la función coste de un ciclo dirigido C en G con un conjunto de aristas F como $\sum_{e \in F} c_e$, la siguiente proposición establece la conexión entre las *λ-cover superstring* y los ciclos dirigidos factibles derivados de grafos de distancia:

Proposición 1. Existe una *λ-cover superstring* para $(H_1, \dots, H_n)$ de longitud máxima $l \Leftrightarrow$ G contiene un ciclo dirigido factible C con coste máximo l.

Demostración: Primero, supongamos que v es una *λ-cover superstring* para $(H_1, \dots, H_n)$ de longitud máxima l. Entonces, para cada $i \in \{1, \dots, n\}$, tenemos $|Y_i| \geq \lambda$, donde Y_i denota el conjunto de elementos de H_i que son subcadena de v. Sea Z el conjunto maximal de elementos de $Y \coloneqq \bigcup_{i=1}^{n} V_i(C)$ parcialmente ordenados con respecto a la relación de subcadenas, esto es:

$$Z = \{s \in Y : (\forall y \in Y)(si\ s \subseteq y \Longrightarrow s = y)\}.$$

A continuación, ordenamos los elementos de Z como $(z_1, \dots, z_p)$ de acuerdo con el orden de aparición como subcadenas de v. Dado que ninguna cadena de Z es subcadena de otra cadena en Z, este ordenamiento está bien definido y es único. Nótese que $(z_1, \dots, z_p)$ define un camino dirigido en $G - s^*$. Extendemos este camino a través de s^* a un ciclo $C = (z_1, \dots, z_p, s^*)$ en G. Afirmamos que este C es un ciclo factible de coste como máximo $l(v)$, que es la longitud de v. Por la definición de grafo de distancia, el coste de C es igual a:

$$\sum_{i=1}^{p} c_{z_i, z_{i+1}} + c_{z_p, s^*} + c_{s^*, z_1} = \sum_{i=1}^{p} \big(l(z_i) - solapamiento(z_i, z_{i+1})\big) + l(z_p),$$

esto es, la longitud de la suma con solapamiento de las cadenas $z_1, \dots, z_p$, en dicho orden. Esto no es más que el número total de caracteres de v que aparecen en la primera aparición de algún z_i como subcadena de v. Queda verificar que C satisface las propiedades de ciclo factible. Primero, tenemos que $s^* \in V(C)$ por construcción. Segundo, cada dos vértices distintos s, t de $V(C) \cap T$, pertenecen a Z, y por lo tanto ninguna es subcadena de otra cadena. Finalmente, sea $i \in \{1, \dots, n\}$, y consideremos el conjunto $V_i(C) \subseteq H_i$ que son las cadenas de H_i cubiertas por C. Por definición de C, el conjunto $V_i(C)$ es igual al conjunto de todas las cadenas de H_i que son subcadenas de alguna cadena en Z, lo que, en consecuencia, es igual al conjunto de cadenas de H_i que son subcadenas de v. Por lo tanto, $V_i(C) = Y_i$, y el requerimiento de que $|V_i(C)| \geq \lambda$ es consecuencia del requerimiento correspondiente de v.

Del mismo modo, supongamos ahora que $C = (z_1, \dots, z_p, s^*)$ es un ciclo dirigido factible en G con coste máximo l. Sea v una cadena definida como la suma con solapamiento de las cadenas $z_1, \dots, z_p$, en dicho orden. La definición de grafo de distancia implica que la longitud de v es igual al coste de C, y por lo tanto es como máximo l. Además, el hecho de que C sea factible implica que v es una λ-*cover superstring* para $(H_1, \dots, H_n)$. De hecho, $\forall i = \{1, .., n\}$, si denotamos como Y_i al conjunto de cadenas de H_i que son subcadenas de v, entonces $V_i(C) = Y_i$, y en consecuencia $|Y_i| \geq |V_i(C)| \geq \lambda$, lo que completa la demostración. ■

El *Shortest* λ*-superstring problem* fue generalizado en Martínez et al., 2019, donde se amplió el concepto de λ-supercadena introduciendo pesos para las cadenas objetivo, lo que dio origen a las λ-supercadenas ponderadas.

Definición 9. Sea $w: T \to \mathbb{R}$ la función que asigna a cada cadena objetivo un peso. Una **λ-supercadena ponderada** sobre los conjuntos $H, T \subseteq A^*$ es una cadena $v \in A^*$ tal que $\sum_{t \in C(h,v) \cap T} w(t) \geq \lambda, \forall\, h \in H$.

Esto a su vez dio origen a un nuevo problema combinatorio, el *Shortest weighted* λ*-superstring problem*:

Shortest weighted λ-superstring problem

Sean $H_1, \dots, H_n \subseteq A^*$ un conjunto finito de cadenas del alfabeto A, sea $T \subseteq A^*$ el conjunto de cadenas objetivo, sea $w: T \to \mathbb{R}$ la función peso, y sea $\lambda \in \mathbb{R}$, encontrar una λ-supercadena ponderada $v \in A^*$ para $(H_1, \dots, H_k, T, w)$ de longitud mínima.

Éste a su vez es polinómicamente equivalente al *Shortest weighted λ-cover superstring problem*, que es la extensión del previamente mencionado *Shortest λ-cover superstring problem*, y que enunciamos a continuación, ya que lo utilizaremos más adelante:

Shortest weighted λ-cover superstring problem

Sea $H \subseteq A^*$ un conjunto finito de cadenas del alfabeto A, y $w: T \to \mathbb{R}$, y sea $\lambda \in \mathbb{R}$, encontrar una λ-cover superstring ponderada $v \in A^*$ para (H, w) de longitud mínima, donde $T = \bigcup_{H_i \in H} H$.

Análogamente a lo hecho para el caso no ponderado, a continuación, se traduce el problema ponderado a lenguaje de teoría de grafos, y finalmente, se demuestra una proposición equivalente a la Proposición 1 para el caso ponderado.

Consideremos para ello primero los elementos de entrada (C, w, λ) del *Shortest weighted λ-cover superstring problem*, y sea $T = \bigcup_{H_i \in H} H_i$. Como en el caso no ponderado, construimos un grafo de distancia $G = (V, E, c)$ de la siguiente manera: (nótese que a los pesos de los vértices les hemos llamado coste, $(c: E \to \mathbb{Z}^+)$ para diferenciarlos de la función de ponderación w.)

- $V = T \cup \{s^*\}$.
- Para cada dos vértices distintos $s, t \in T$, añadimos la arista (s, t) a E y le asignamos el coste $c_{s,t} = l(s) - solapamiento(s, t)$.
- Para cada vértice $s \in T$, añadimos la arista (s, s^*) a E y le asignamos el coste $c_{s,s^*} = l(s)$.
- Para cada vértice $s \in T$, añadimos la arista (s^*, s) a E y le asignamos el coste $c_{s^*,s} = 0$.

En lo que sigue, identificaremos los vértices de $V \backslash \{s^*\}$ como T. Además, diremos que un subgrafo H de G se dice que *cubre* una cadena $s \in T$ si existe un vértice $t \in V(H) \cap T | s \subseteq t$. Para un elemento $X \in C$, denotaremos el conjunto de cadenas de X cubiertas por H como X_H. El coste del ciclo dirigido C se define como $\sum_{e \in E(C)} c(e)$, y un ciclo dirigido se dirá que es *w-factible* si satisface las siguientes condiciones:

- $s^* \in V(C)$;
- para cada dos vértices distintos s, t de $V(C) \cap T$, s no es una subcadena de t;
- $\forall X \in C$, se tiene que $\sum_{t \in X_C} w(t) \geq \lambda$.

Ahora estamos en condiciones de enunciar la Proposición 2:

Proposición 2. Sea (C, w, λ) la entrada del *Shortest weighted λ-cover superstring problem*, y sea G su grafo de distancia. Entonces, existe una *weighted λ-cover superstring* para (C, w) de longitud máxima $l \Leftrightarrow G$ contiene un ciclo dirigido *w-factible* C con coste máximo l.

Demostración. Primero, supongamos que v es una *weighted λ-cover superstring* para (C, w), de longitud máxima l. $\forall X \in C$, denotamos $X_v \subseteq X$ al conjunto de cadenas de X

que son subcadenas de v. Entonces, tenemos que $\sum_{t \in X_v} w(t) \geq \lambda$. Llamemos Z al conjunto de elementos maximales del conjunto $Y \coloneqq \bigcup_{X \in C} X_v$ parcialmente ordenados con respecto a la relación de ser subcadena, esto es:

$$Z = \{s \in Y : (\forall y \in Y)(si\ s \subseteq y \Longrightarrow s = y)\}.$$

Después, ordenamos los elementos de Z como $(z_1, \dots, z_p)$ de acuerdo con el orden de aparición como subcadenas de v. Dado que ninguna cadena de Z es subcadena de otra cadena en Z, este ordenamiento está bien definido y es único. Nótese que $(z_1, \dots, z_p)$ define un camino dirigido en $G - s^*$. Extendemos este camino a través de s^* a un ciclo $C = (z_1, \dots, z_p, s^*)$ en G. Afirmamos que este C es un ciclo w-factible de coste como máximo $l(v)$. Por la definición de grafo de distancia, el coste de C es igual a:

$$\sum_{i=1}^{p} c_{z_i, z_{i+1}} + c_{z_p, s^*} + c_{s^*, z_1} = \sum_{i=1}^{p} \big(l(z_i) - solapamiento(z_i, z_{i+1})\big) + l(z_p),$$

esto es, la longitud de la suma con solapamiento de las cadenas $z_1, \dots, z_p$, en dicho orden. Esto es el número total de caracteres de v que aparecen en la primera aparición de algún z_i como subcadena de v, que claramente no excede la longitud de v. Deben verificarse además las propiedades de ciclo w-factible:

Primero, por construcción se tiene que $s^* \in V(C)$. Segundo, cualesquiera dos vértices distintos s, t de $V(C) \cap T$, pertenecen a Z, por lo que ninguna es subcadena de otra cadena. Por último, sea $X \in C$, y consideremos el conjunto $X_C \subseteq X$, que son las cadenas de X cubiertas por C. Por definición de C, el conjunto X_C es igual al conjunto de todas las cadenas de X que son subcadenas de alguna cadena en Z, lo que, en consecuencia, es igual al conjunto de cadenas de X que son subcadenas de v. Por lo tanto, $X_C = X_v$, y el requerimiento de que $\sum_{t \in X_C} w(t) \geq \lambda$ es consecuencia del requerimiento correspondiente de v.

Para probar la otra implicación, supongamos ahora que $C = (z_1, \dots, z_p, s^*)$ es un ciclo dirigido w-factible en G con coste como máximo l. Sea v una cadena definida como la suma con solapamiento de las cadenas $z_1, \dots, z_p$, en ese orden. La definición de grafo de distancia implica que la longitud de v es igual al coste de C, y por lo tanto como máximo su valor es l. Por otro lado, como C es w-factible, tenemos que v es una *weighted λ-cover superstring* para (C, w). Asimismo, $\forall X \in C$, si llamamos X_v al conjunto de cadenas de X que son subcadenas de v, entonces $X_C \subseteq X_v$, y, por lo tanto, $\sum_{t \in X_v} w(t) \geq \sum_{t \in X_C} w(t) \geq \lambda$, completando así la demostración. ■

Para terminar con esta sección y de manera ilustrativa, en la Figura M.2. se representa gráficamente la relación entre el problema de optimización combinatoria y el problema en términos de teoría de grafos.

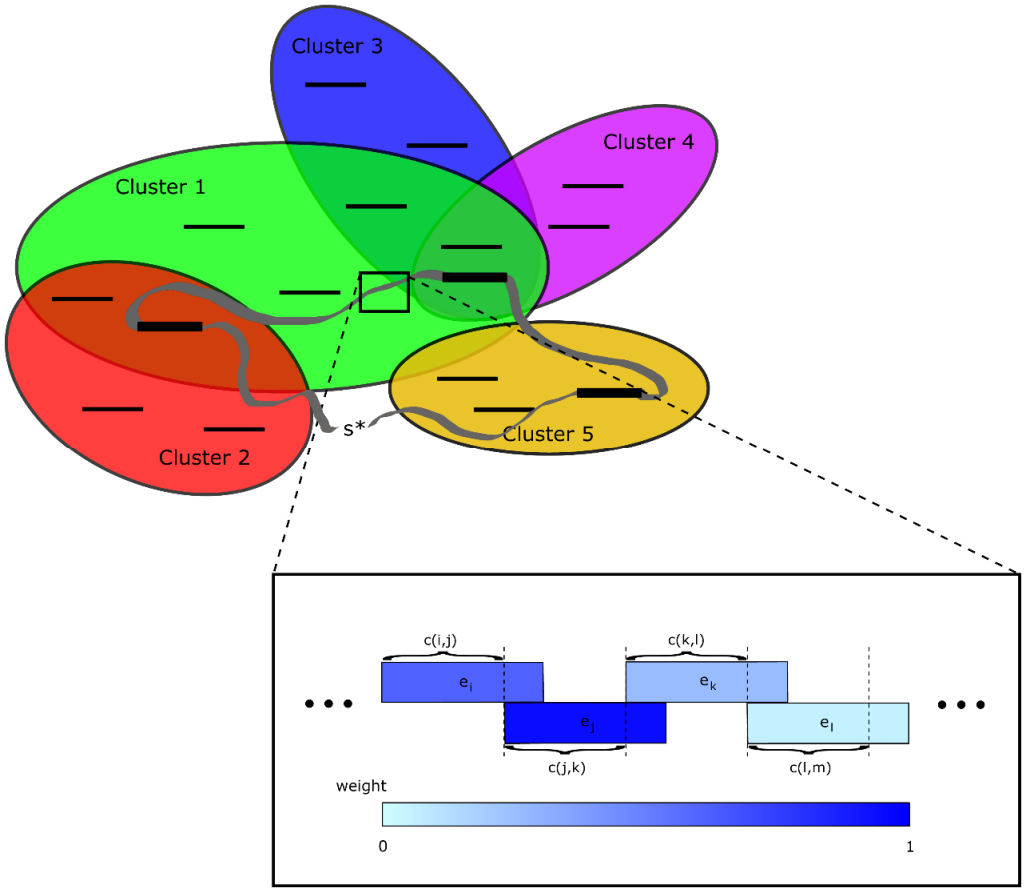

Figura M.2. Interpretación gráfica de la conexión entre el problema de optimización combinatoria y el problema de teoría de grafos. Los clusters asociados a las host strings (H) se muestran como óvalos con las correspondientes target strings (T) dentro de ellas. Cada target string tiene asociado un peso, que en el ejemplo hemos ilustrado con un código de colores que va de azul suave a fuerte, correspondiendo el 0 al azul más suave, y 1 al más fuerte. La λ-supercadena está representada a través de un lazo gris que atraviesa los clusters. Es cerrada porque una de las cadenas que lo conforma corresponde al vértice artificial s^*, el cual no es una host string, pero puede verse como una cadena vacía que "pega" los extremos de la λ-supercadena. La condición de que para cada cluster la suma de los pesos de las cadenas objetivo que están tanto en la λ-supercadena como en el cluster sea al menos λ se impone para las soluciones factibles. La longitud de la λ-supercadena se minimiza, y ésta puede obtenerse sumando los valores $c(i,j)$ de las cadenas que forman la λ-supercadena. Los valores $c(i,j)$ se representan en la figura como la longitud del fragmento del vértice i que no se solapa con el siguiente vértice j de la λ-supercadena, y los e_i representan las aristas del grafo.

M.3. Resolución del problema a través de programación entera

La Proposición 2 nos conduce a la siguiente formulación en términos de programación entera del *Shortest weighted λ-cover superstring problem*. El programa utiliza tres tipos de variables binarias: x_{ij}, donde (i,j) recorre todos los pares ordenadoes de distintos elementos de V; y_i, donde i recorre todos los elementos de V; y z_i, donde i recorre todos los elementos de T. Nótese que $c: E \to \mathbb{R}^+$ es la función de coste de los vértices en el grafo de distancia G. El problema es el siguiente:

$$Encontrar\ la\ \min \sum_{i,j} c(i,j)x_{ij}\ \ tal\ que:$$

$$y_{s^*} = 1$$

$$\sum_{i \in V: i \neq j} x_{ij} = y_j, \forall j \in V$$

$$\sum_{j \in V: j \neq i} x_{ij} = y_j, \forall i \in V$$

$$\sum_{i \in X} w(i) z_i \geq \lambda, \forall X \in C$$

$$\sum_{i \subseteq j} y_j \geq z_i, \forall i \in T$$

$$y_i + y_j \leq 1, \forall i, j \in T | i \subset j$$

$$0 \leq x_{ij} \leq 1, \qquad con\ x_{ij}\ entero$$

$$0 \leq y_i \leq 1, \qquad con\ y_i\ entero$$

$$0 \leq z_i \leq 1, \qquad con\ z_i\ entero.$$

Las soluciones factibles obtenidas por el algoritmo de programación entera descrito arriba se corresponden con los subgrafos H de G que contienen a s^* que consisten en uno o más *subtours* (ciclos dirigidos de vértices disjuntos) en los cuales los vértices que no son s^* corresponden a un conjunto de cadenas que son incompatibles dos a dos con respecto a ser subcadenas las unas de las otras, y por lo tanto la condición de cubrimiento

$$\sum_{t \in X_H} w(t) \geq \lambda$$

se satisface. Para aplicar la Proposición 2, nosotros sólo estamos interesados en las soluciones que estén constituidas por un único ciclo dirigido, y para obtener esta solución y no las formadas por subtours, una posibilidad es introducir en el algoritmo las siguientes restricciones y variables adicionales ($u_i, i \in V$), lo que es conocido como la formulación Miller-Tucker-Zemlin (Miller et al., 1960):

$$u_{s^*} = 1$$

$$2 \leq u_i \leq |V|, \qquad \forall i \neq s^*$$

$$u_i - u_j + 1 \leq (|V| - 1)\big(1 - x_{ij}\big), \qquad \forall i \neq s^*, \forall j \neq s^*, i \neq j.$$

Esto elimina los subtours, ya que la última restricción para los pares (i, j) fuerza a que $u_j \geq u_i + 1$ cuando $x_{ij} = 1$, y si una solución factible del problema tiene más de un subtour, entonces al menos uno de esos subtours no contendrá el nodo s^*, y a través de este subtour los valores de u_i tendrían que crecer hasta el infinito.

M.4. Resolución del problema a través de un Algoritmo Genético multiobjetivo (GA)

Además del método basado en programación entera para obtener las soluciones óptimas, dado que, en la práctica, la cantidad de cadenas y antígenos a manejar es muy grande, el IP es poco viable por su enorme coste computacional. Por lo tanto, para resolver este problema, también hemos desarrollado un método basado en Algoritmos Genéticos (GA) para encontrar soluciones sub-óptimas, donde se optimizan dos parámetros simultáneamente.

En particular, a través de este algoritmo se busca encontrar una λ-supercadena ponderada con λ lo más grande posible, para un conjunto de cadenas de longitud similar (que serán las variantes de la proteína objetivo del virus) y que, como hemos mencionado al principio del capítulo, denominaremos *Host Strings* (H), y un conjunto de cadenas formado por subcadenas de las *Host Strings* de longitud fija (en este caso 9 aminoácidos, que es la longitud por defecto para los antígenos de Clase I), que llamaremos *Target Strings* (T). Asimismo, se buscará mantener el ordenamiento con respecto a las variantes de la proteína, a través de la optimización del alineamiento. Dicho de otro modo, buscamos diseñar una proteína sintética que contenga el mayor número de los mejores antígenos por variante de proteína, y que además se asemeje a las secuencias de las proteínas originales, de manera que mantenga su estructura secundaria y terciaria, lo que aumenta las probabilidades de que sea reconocida por el sistema inmunitario y no sea rechazada por éste. Por lo tanto, nos encontramos ante un problema de optimización multiobjetivo.

Para abordar este tipo de problemas, consideramos funciones $f: P \to \mathbb{R}^n$, siendo P el conjunto de soluciones factibles, que asigna a cada elemento $v \in P$ una n-tupla de entradas reales $(f_1(v), \dots, f_n(v))$ donde cada entrada indica un objetivo parcial de la función. En general, no es posible encontrar una solución $v \in P$ para la cual todas las funciones parciales f_i alcancen el máximo valor, por lo que se evalúa la optimalidad a través del concepto de óptimo de Pareto: dadas dos soluciones factibles $v, w \in P$, diremos que la solución $v = (v_1, \dots, v_n)$ está dominada por la solución $w = (w_1, \dots, w_n)$ si se cumple que $\forall i, v_i \leq w_i$ y además, para algún $j, v_j < w_j$. Considerando los elementos no dominados de P formaremos un conjunto de soluciones conocido como el frente de Pareto de P.

Entre los algoritmos evolutivos propuestos para obtener Frentes de Pareto se encuentra el NSGA-II (Deb et al., 2002), que es uno de los más rápidos y utilizados. En resumen, este algoritmo se estructura así: dado un conjunto $P' \subseteq P$ de soluciones posibles, se asignan dos valores a cada elemento $v \in P'$, el rango de dominación v_{rango} y la distancia respecto al conjunto de soluciones v_{dist}. Estos valores se calculan de la siguiente manera: a los elementos no-dominados del Frente de Pareto de P' se les asigna

el rango 1, y forman el conjunto F_1. Después se asigna el rango 2 a las soluciones no-dominadas del conjunto $P' - F_1$, que formarán el conjunto F_2, y así sucesivamente. Por otro lado, la distancia respecto al conjunto de una solución se obtiene tomando el promedio (para los n valores de las funciones objetivo del vector solución) de las distancias de dos puntos que "rodean" a la solución v. Este proceso origina un ordenamiento en P' definido como: $v \prec w$ si $v_{rango} < w_{rango}$ o, en caso de que $v_{rango} = w_{rango}$, si $v_{dist} > w_{dist}$ (favoreciendo así la elección de elementos más alejados entre ellos en el frente de Pareto).

El algoritmo procede de la siguiente manera: fijado el tamaño de población m, se construye de manera aleatoria una población P_0 y se ordena utilizando la relación $\prec$. Después, se realiza una selección por torneo binario de dichos elementos (esto es, se obtienen primero dos soluciones aleatorias de la población; a continuación, si las soluciones son de diferente rango F_i, se escoge la que tenga menor rango, y en caso de ser ambas del mismo rango, se escoge la de mayor distancia respecto del conjunto de soluciones). Posteriormente, se realizan mutaciones y cruzamientos entre dichos elementos, dando pie a una población Q_0 también de tamaño m. Ahora, se forma una población combinada $R_0 = P_0 \cup Q_0$, y se ordena según su nivel de dominación. A continuación, se genera una población P_1 tomando los elementos de R_0 en función de su rango (los de rango 1 de F_1, rango 2 de F_2…), en orden descendente, hasta llegar a cierto F_i para el cual no podamos incluir todos sus elementos (porque excedan el tamaño m) de P_1. En tal caso, se ordenarán dichos valores según su distancia respecto al conjunto de soluciones, y se tomarán los elementos con mayor valor, formando así el conjunto P_1 de tamaño m. Este proceso se itera tantas veces como se haya indicado en el parámetro $niter$, hasta obtener el conjunto P_{niter}. Una descripción más detallada y el pseudocódigo del algoritmo NSGA-II pueden encontrarse en Deb et al., 2002. A modo ilustrativo, en la Figura M.3. se representa el proceso de elitismo del algoritmo.

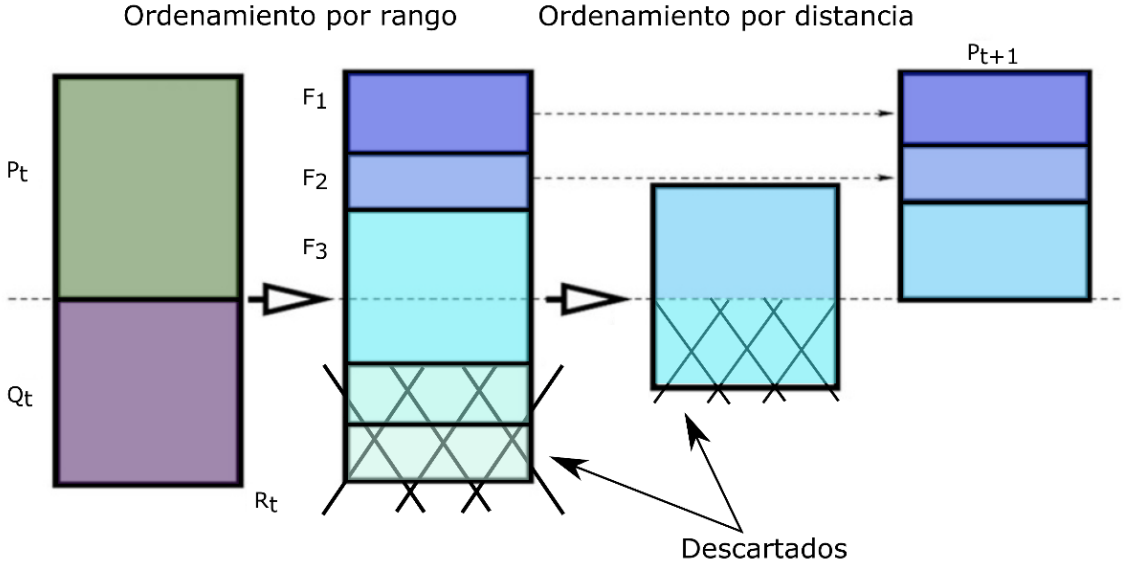

Figura M.3. Selección de las cadenas por elitismo del algoritmo NSGA-II.

En nuestro problema particular, buscamos obtener una λ-supercadena ponderada para un conjunto $H = \{H_1, \ldots, H_{spop}\}$ de *Host Strings* (siendo *spop* el número total de Host strings, correspondiente a las distintas variantes del virus), y un conjunto T de *Target*

Strings formado por todas las subsecuencias distintas de 9 aminoácidos obtenidas a partir de las *Host Strings*, que serán ponderadas a partir de una función $w: T \to \mathbb{R}^2$. Los cromosomas del algoritmo genético serán secuencias formadas por *Target Strings*, y el fenotipo de un cromosoma v será la suma con solapamiento $o(v)$ de las cadenas objetivo que lo forman (en función de la posición en la que se encuentran en v). Por lo tanto, la función fitness de cada cromosoma v de la población vendrá dada por $f(\lambda(v), al(v))$, donde:

- $\lambda(v)$, es una estimación del máximo valor para el cual $o(v)$ es una λ-supercadena ponderada para el conjunto (H, T), definida como:

$$\lambda(v) = \min\{ \sum_{t \in v, t\ subcadena\ de\ H_i} w(t) : i = 1, \dots, spop\}$$

  (es una estimación porque el verdadero valor de λ podría ser mayor, ya que al considerar la cadena completa como unión de subcadenas, se obtienen nuevos antígenos recubiertos).

- $al(v)$, es el promedio de los alineamientos globales dos a dos (Gusfield, 1997) obtenidos al comparar $o(v)$ con cada *Host String* h_i.

En nuestra versión modificada del NSGA-II, hemos aumentado el tamaño de la población Q_0 (que originalmente era $m = spop$) para que $|R_i| > 2|P_i| \forall i$, y no hemos escogido la población inicial P_0 de manera aleatoria, sino que está formada por secuencias de cadenas de T siguiendo el orden de aparición en H. Para el cruzamiento de dos cromosomas $(v_1, \dots, v_{l_1})$ y $(w_1, \dots, w_{l_2})$, se ha elegido de manera aleatoria el punto de corte c comprendido entre 1 y $\min\{l_1, l_2\} - 1$, y se ha tomado como primer descendiente $(v_1, \dots, v_c, w_{c+1}, \dots, w_{l_2})$, y como segundo, $(w_1, \dots, w_c, v_{c+1}, \dots, v_{l_1})$. Para realizar las mutaciones en un cromosoma $(v_1, \dots, v_{l_1})$, primero se seleccionó una posición de la cadena aleatoriamente, después se seleccionó una de las *Host Strings*, y se sustituyó el aminoácido de la posición seleccionada por el correspondiente a esa posición en la cadena original. De esta manera, las mutaciones que se realizaron no fueron al azar, si no que reflejaron posibles mutaciones que se hubieran dado en el virus.

En resumen, el proceso sería el siguiente:

Paso 1: Se genera una población inicial P_0.

⇓

Paso 2: Se ordena utilizando la relación $\prec$.

⇓

Paso 3: Se seleccionan m cadenas de P_0 por torneo binario y se realizan mutaciones y cruzamientos (dando pie a la población Q_0).

⇓

Paso 4: Se combinan P_0 y Q_0 para formar R_0.

⇓

Paso 5: Se ordenan siguiendo el nivel de dominación.

⇓

Paso 6: Se genera una población P_1 a partir de los "mejores" elementos de R_0.

⇓

Paso 7: Volvemos al Paso 2 con P_1, y así hasta repetir *niter* veces.

⇓

Paso 8: Obtenemos el frente de Pareto no dominado.

El código completo de este algoritmo ha sido publicado en el repositorio de libre acceso Zenodo: https://zenodo.org/record/1487837.

Objetivos

O.1. Objetivo principal de la investigación nº1

Mejorar la eficiencia de las vacunas contra virus con alta tasa de mutación introduciendo la condición de λ-supercadena ponderada.

Objetivos específicos:

- Extender los resultados de las λ-supercadenas al caso ponderado.
- Desarrollar algoritmos capaces de dar solución al problema combinatorio asociado.
- Probar dichos algoritmos para el diseño de una vacuna contra el VIH.

Este trabajo ha sido publicado en PloS one (Q1 en el área "Multidisciplinary" de SJR): Martínez, L., Milanič, M., Malaina, I., Álvarez, C., Pérez, M. B., & M. de la Fuente, I. (2019). Weighted lambda superstrings applied to vaccine design. *PloS one*, 14(2), e0211714.

O.2. Objetivo principal de la investigación nº2

Diseñar y testar una vacuna eficaz contra el SARS-CoV-2 utilizando la metodología de las λ-supercadenas.

Objetivos específicos:

- Adaptar el problema de las λ-supercadenas ponderadas a la particularidad del virus SARS-CoV-2.
- Elaborar un sistema de ponderación de cadenas acorde con el desarrollo de una vacuna universal.
- Elaborar y testar en células, ratones y humanos, una de las vacunas obtenidas, para probar su eficacia.

Este trabajo ha sido publicado en Scientific Reports (Q1 en el área "Multidisciplinary" de SJR) como: Martínez, L., Malaina, I., Salcines-Cuevas, D., Terán-Navarro, H., Zeoli, A., Alonso, S., de la Fuente, I.M., González-López, E.,

Ocejo-Vinyals, J.G., Gonzalo-Margüello, M., Calvo-Montes, J., & Alvarez-Dominguez, C. (2022). First computational design using lambda-superstrings and in vivo validation of SARS-CoV-2 vaccine. *Scientific Reports*, 12(1), 1-12.

O.3. Objetivo principal de la investigación nº3

Testar experimentalmente una vacuna personalizada contra el melanoma diseñada a partir de optimización combinatoria y técnicas bioinformáticas, validando su efectividad, y la de dichos métodos cuantitativos.

Objetivos específicos:

- Extraer los elementos necesarios para diseñar vacunas peptídicas a partir del genoma, identificando los potenciales neoantígenos.
- Estimar las principales propiedades de los potenciales neoantígenos.
- Llevar a cabo una selección objetiva y óptima que resulte en la elaboración de una vacuna personalizada.
- Testar dicha vacuna y probar su efectividad.

Este trabajo está bajo revisión en BMC Bioinformatics (Q1 en el área "Mathematics: Applied Mathematics" de SJR) como: Computational and experimental evaluation of the immune response of neoantigens for personalized vaccine design, y ha sido elaborado por: Malaina, I., Martínez, L., González-Melero, L., Salvador, A., Sánchez-Díez, A., Asumendi, A., Margareto, J., Carrasco, J., Legarreta, L., García, M.A., Pérez, M.B., Izu, R., De la Fuente, I.M., Igartua, M., Alonso, S., Hernández, R.M., & Boyano, M.D.

O.4. Objetivo principal de la investigación nº4

Evaluar la capacidad de las herramientas bioinformáticas de diferenciar la capacidad inmunogénica de neoantígenos y cadenas no-mutadas.

Objetivos específicos:

- Obtener los neoantígenos a partir del mutanoma de 6 pacientes, así como su versión no mutada.
- Estudiar el número de péptidos que potencialmente se unirían a moléculas del complejo mayor de histocompatibilidad de clases I y II.
- Analizar si el número de péptidos mutados es significativamente mayor que el de no-mutados de acuerdo a la afinidad de unión de las moléculas HLA-I y HLA-II.

Este trabajo ha sido publicado en LNCS (Q4 en el campo "Mathematics: Theoretical Computer Science", y Q2 en "Computer Science" de SJR) como: Malaina, I., Legarreta, L., Boyano, M.D., Alonso, S., De la Fuente, I.M., Martinez, L. (2020).

Analyzing the Immune Response of Neoepitopes for Personalized Vaccine Design. In: Rojas, I., Valenzuela, O., Rojas, F., Herrera, L., Ortuño, F. (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2020. *Lecture Notes in Computer Science*, vol 12108. Springer, Cham.

Resultados

Investigación nº1: λ-supercadenas ponderadas en el diseño computacional de vacunas

R.1.1. Relevancia

Los métodos de vacunación tradicionales han demostrado no ser eficientes a la hora de proteger contra virus con alta tasa de mutación. El criterio de seleccionar los antígenos más frecuentes conduce a una desprotección contra las variantes menos habituales, que acaban sorteando el sistema inmune a través de mutaciones de escape, y, en definitiva, perpetuando la infección.

En este trabajo, hemos generalizado el concepto de λ-supercadena, introduciendo una ponderación que tiene en cuenta la inmunogenicidad del antígeno y el alineamiento con la proteína a la que pretende asemejarse, acercando el criterio de λ-supercadenas a la realidad biológica. Se abre por lo tanto un nuevo marco para el diseño de vacunas contra virus como el HIV, que aún no dispone de vacuna eficiente.

R.1.2. Resolución del problema a través de programación entera

Tanto en esta sección (R.1.2) como en la siguiente (R.1.3), los algoritmos se aplicaron al conjunto de 169 cadenas correspondientes a la proteína Nef del HIV-1 subtipo B, que fueron obtenidas a partir de GenBank (GenBank website), y que fueron usadas por Nickle et al. (Nickle et al., 2007), y posteriormente por nuestro grupo en Martínez et al. 2015. La razón de usar el mismo conjunto de cadenas fue la de ser capaces de comparar el método aquí propuesto con los dos mencionados, comparativa que puede encontrarse al final de la sección R.1.3.

Para resolver el problema de manera exacta se utilizó la programación entera. El conjunto de antígenos (experimentalmente testados) se obtuvo de la base de datos HIV Molecular Immunology Database (HIV Molecular Immunology Database website), mientras que sus inmunogenicidades estimadas se obtuvieron de Immune Epitope Database and Analysis Resource (IEDB Class-I Immunogenicity). Los antígenos seleccionados cumplieron simultáneamente los siguientes criterios:

- Podían encontrarse en al menos una de las 169 cadenas analizadas.
- Eran antígenos para la proteína Nef según HIV Molecular Immunology Database.
- Tenían un valor asociado $p + n$ positivo, donde p y n fueron el número de ensayos positivos y negativos respectivamente, en la sección de MHC Ligand Assays de la base de datos IEDB.

El peso de los antígenos (su inmunogenicidad estimada) se obtuvo como el ratio $\frac{p}{(p+n)}$. La razón para considerar los ensayos para ponderar los antígenos es que la respuesta que genera un antígeno concreto solo puede verificarse a través de ensayos, por lo que consideramos que la manera más realista de aproximar dicha respuesta era a través de esta ratio. Más aún, utilizamos los ensayos de MHC Ligand Assays porque como indican varios estudios, existe una correlación entre la respuesta inmune generada y la estabilidad (Rasmussen et al., 2016) o la afinidad con el complejo MHC (Sette et al., 1994), y además ha sido utilizado anteriormente para predecir posibles antígenos para linfocitos T (Paul et al., 2013). Nótese que un re-escalamiento no lineal de los pesos (como normalizarlos entre 0 y 1) cambiaría el problema de optimización. Los antígenos seleccionados y los pesos asociados pueden encontrarse en la Tabla 1.1.

Tabla 1.1. Valores experimentales de la inmunogenicidad de los antígenos

Antígeno	Peso	Antígeno	Peso
AAVDLSHFL	0	LTFGWCFKL	1
AFHHVAREL	1	LTFGWCFKLV	1
AVDLSHFL	0	PLTFGWCYKL	0
AVDLSHFLK	0.57	QEILDLWVY	0.63
EWRFDSRL	1	QVPLRPMTYK	0.68
FPDWQNYT	0	RPMTYKAAL	0.41
FPVRPQVPL	0.94	RPQVPLRPM	1
FPVTPQVPL	0.31	RYPLTFGWCF	1
KAAVDLSHFL	1	TPGPGIRYPL	1
KEKGGLEGL	0.5	TPGPGVRYPL	1
KRQEILDLWVY	1	TQGYFPDWQNY	1
VLEWRFDSRL	0.2	VPLRPMTY	1
YPLTFGWCF	1	DLSHFLKEKGGLEGL	0.5
HHVARELHPEYFKNC	1	RLAFHHVARELHPE	1
EWRFDSRLAFHHVAREL	1	GVRYPLTFGWCYKLVP	1
PEKEVLVWKFDSRLAFHH	1	YKAAVDLSHFLKEKGGL	0.75

Realizamos dos análisis con el algoritmo basado en programación entera (IP). En el primero, variamos el valor de λ entre 1 y 3.3, con incrementos de 0.1, y para cada uno de los valores, se minimizó la longitud de la λ-supercadena correspondiente. Como se detalla en Martínez et al. 2019, el algoritmo se programó en lenguaje Java (Java website) y se resolvió de manera óptima utilizando IBM ILOG CPLEX Optimization Studio (IBM ILOG CPLEX Optimization Studio website). La razón para que se escogiera 3.3 como el valor máximo de λ fue que, para ese valor, o superiores, no se pudo encontrar solución. En la Tabla 1.2 se muestran los valores de las soluciones no dominadas obtenidas por el algoritmo.

Tabla 1.2. Soluciones óptimas de mínima longitud para un valor de λ dado

λ = 1.0	**λ = 2.5**
λ-supercadena óptima: TQGYFPDWQN YVPLRPMTYPLTFGWCF	λ-supercadena óptima: TQGYFPDWQ NYPLTFGWCFKLVFPVRPQVPLRPMTY KAAVDLSHFLK
Longitud de la λ-supercadena óptima: 27	Longitud de la λ-supercadena óptima: 47
Coverage de la solución: 1	Coverage de la solución: 2.51
λ = 1.5	**λ = 2.6**
λ-supercadena óptima: TFGWCFKLVFP VRPQVPLRPMTYKAAVDLSHFLK	λ-supercadena óptima: FPVRPQVPLR PMTYKAAVDLSHFLKEKGGLTQGYFP DWQNYTPGPGVRYPLTFGWCFKLV
Longitud de la λ-supercadena óptima: 35	Longitud de la λ-supercadena óptima: 60
Coverage de la solución: 1.51	Coverage de la solución: 2.68
λ = 1.9	**λ = 2.9**
λ-supercadena óptima: KAAVDLSHFLT FGWCFKLVFPVRPQVPLRPMTYTQGYFP DWQNY	λ-supercadena óptima: TPGPGVRYPLF PVRPQVPLRPMTYKAAVDLSHFLKTPG PGIRYPLTFGWCFKLVTQGYFPDWQNY
Longitud de la λ-supercadena óptima: 44	Longitud de la λ-supercadena óptima: 65
Coverage de la solución: 1.94	Coverage de la solución: 2.94
λ = 2	**λ = 3.2**
λ-supercadena óptima: KAAVDLSHFLKL TFGWCFKLVFPVRPQVPLRPMTYTQGYF PDWQNY	λ-supercadena óptima: TPGPGIRYPLT PGPGVRYPLTFGWCFKLVPEKEVLVWK FDSRLAFHHQEILDLWVYFPVRPQVPLR PMTYKAAVDLSHFLKEKGGLEGLTQGYF PDWQNY
Longitud de la λ-supercadena óptima: 46	Longitud de la λ-supercadena óptima: 100
Coverage de la solución: 2	Coverage de la solución: 3.25

Dado que nuestro objetivo en esta sección fue el de encontrar la λ-supercadena más corta, y a la vez con mayor valor del *coverage* (relación entre los pesos de los antígenos que se encuentran en el candidato a vacuna, divididos entre el total de los pesos de los antígenos considerados), consideramos interesante analizar los candidatos con mayores ratios. En este caso, el valor más alto fue alcanzado por la cadena de longitud 47, que obtuvo un *coverage* de 2.51. Los siguientes mejores resultados se obtuvieron con las cadenas con pares de valores (longitud, *coverage*): (44, 1.94), (60, 2.68) y (65, 2.94).

Por otro lado, también se realizó un segundo análisis donde para una longitud dada (variando desde 10 hasta 200), se maximizó el valor de λ. Los resultados de este análisis fueron muy similares a los del primero, por lo que no se ha incluido aquí, pero puede encontrarse en Martínez et al. 2019.

R.1.3. Resolución del problema a través del algoritmo genético

En esta segunda parte, en vez de considerar el conjunto de antígenos mencionados en la Tabla 3.1, consideramos todas las cadenas de 9 aminoácidos presentes en alguna de las 169 cadenas correspondientes a la proteína Nef del HIV-1 y mencionadas en el apartado R.1.2, las cuales también fueron utilizadas para esta parte del estudio. En este caso, el conjunto de datos era demasiado grande como para resolver el problema de manera exacta, por lo que se utilizó un algoritmo heurístico, concretamente el algoritmo genético multiobjetivo NSGA-II descrito en la sección M.4 de Material y Métodos.

A diferencia de la sección anterior, las inmunogenicidades (esto es, los pesos) de los antígenos no fueron obtenidas de manera experimental, principalmente porque no existen ensayos sobre muchos de los potenciales 9-meros considerados. Por el contrario, dado que cuantificar la respuesta inmunitaria de un antígeno es un aspecto muy importante del diseño de vacunas, existen muchos algoritmos especializados en ello (Bergmann-Leitner et al., 2013; Bryson et al., 2010; Khan et al., 2012; Moreau et al., 2008). Para nuestro análisis, elegimos la herramienta de IEDB denominada T-cell epitopes-Immunogenicity Prediction (Calis et al., 2013).

El algoritmo genético fue programado en Mathematica (Mathematica website), utilizando los siguientes parámetros: $niter = 500, spop = 169, prmut = 0.01, m = 1352, sd = 1$. Para estimar el alineamiento (esto es, la similitud en cuanto al ordenamiento de los aminoácidos) del candidato a vacuna con las 169 secuencias de la proteína Nef, se utilizó el comando "NeedlemanWunschSimilarity", que proporciona valores de acuerdo a la coincidencia de los aminoácidos en cada posición, comparando cadenas dos a dos (a mayor coincidencia, el número devuelto es más alto). Por lo tanto, se calcularon los valores de cada par "candidato/variante de Nef", y se promediaron los 169 valores obtenidos. El algoritmo fue ejecutado 20 veces, y las soluciones no-dominadas fueron almacenadas, para formar posteriormente parte del frente de Pareto (mostrado en la Figura 1.1).

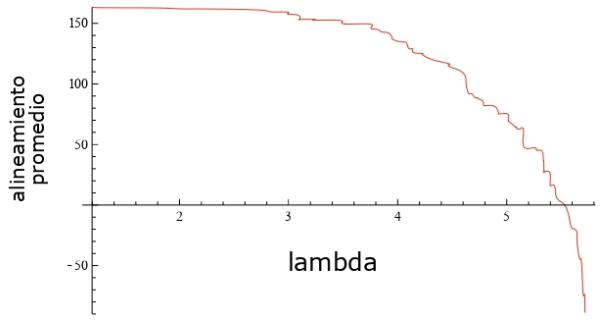

Figura 1.1. Estimación del frente de Pareto a partir del algoritmo genético. La línea roja representa las soluciones no-dominadas encontradas por el algoritmo genético, donde el eje X indica el valor de λ, mientras que el eje Y representa el alineamiento promedio.

La estimación resultante del frente de Pareto devolvió un valor máximo de λ de 5.71, y mínimo de 1.2 (4.32 ± 1.14, media ±SD). Los alineamientos estuvieron comprendidos entre -88.47 y 163.33 (87.73 ± 67.43, media ±SD). De entre estas soluciones, seleccionamos la que contaba con un λ de 2.18 y un alineamiento de 163.33, por diversas razones: la primera, que el λ de esta cadena es mayor que la de cada una de las cadenas de la población de 169 cadenas, que contaban con un valor de λ medio de

-1.70 y un máximo de 1.60; la segunda, que contaba con mejor alineamiento que las cadenas objetivo, ya que la media del valor de alineamiento fue de 143.34, con un máximo de 157.66; por último, este candidato cuenta con varias regiones altamente conservadas de la proteína, lo cual aumenta sus probabilidades de éxito como antígeno. La secuencia de amino-ácidos de esta solución es la siguiente:

MGGKWSKRSGVGWPTVRERMRRAEPAADGVGAVSRDLEKHGAITSSNTAATNADC
AWLEAQEEEEVGFPVRPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIYSQKRQDILD
LWIYHTQGYFPDWQNYTPGPGIRYPLTFGWCFKLVPVEPEKVEEANEGENNSLLHP
MSLHGMEDPEKEVLEWKFDSRLAFHHMARELHPEYYKDC

Con el objeto de evaluar la funcionalidad y la estructura de candidatos a vacuna obtenidos a través de nuestra técnica, se utilizaron varias herramientas bioinformáticas sobre la secuencia mencionada. La importancia de que el candidato sea similar a una proteína Nef existente en la naturaleza radica en que nuestro candidato se ha creado artificialmente, y se desconoce a priori si se plegará como es debido, o si será una proteína que mantenga las características de la original. A continuación, se explicarán los análisis más relevantes; para encontrar el estudio completo, acudir a Martínez et al. 2019.

La longitud de esta cadena obtenida fue 206, que coincide con la longitud de la cadena de referencia establecida para la proteína Nef (O'Neill et al., 2006), y que además está muy cerca de la media de las 169 longitudes consideradas como cadenas objetivo (esto es, variantes de las proteínas del virus), que fue de 207.11. A continuación, analizamos si nuestra solución mantenía las regiones más conservadas del virus, comparando nuestra solución con las secuencias de O'Neill et al., 2006. En particular, buscamos las regiones que se habían conservado en al menos un 90% de los casos, y comprobamos que nuestra solución contenía a dichas regiones.

Después, estudiamos la estructura terciaria (esto es, la disposición tridimensional) de nuestro candidato a través de la herramienta bioinformática I-Tasser (Zhang, 2008). En resumen, este método primero compara la secuencia en cuestión con las proteínas disponibles en las bases de datos, con el objeto de identificar estructuras similares, alineando dichas secuencias de amino-ácidos. Después estima la estructura de las secuencias no alineadas *ab initio* (desde cero, sin utilizar plantillas), y finalmente realiza una simulación de posibles ensamblamientos para las secuencias alineadas y no alineadas, creando un conjunto de posibles estructuras. Entonces, selecciona las de menor energía libre de Gibbs (esto es, las secuencias a las que les "cueste menos" mantenerse unidas), y realiza una segunda ronda de los pasos mencionados, con el objeto de refinar el resultado. Una vez obtenidas las estructuras más probables, I-Tasser nos ofrece un parámetro de bondad de predicción denominado C-score, que está comprendido entre -5 y 2, donde un C-score mayor de -1.5 puede considerarse una predicción fiable. En nuestro caso, el candidato mencionado obtuvo un C-score de 1.42, muy cerca del máximo, lo que indica que la estructura que obtendríamos al sintetizar el candidato sería muy probablemente la descrita por I-Tasser, ilustrada en el panel a de la Figura 1.2. Por otro lado, para cotejar la predicción de I-Tasser, utilizamos otro programa para estimar la estructura terciaria, Phyre-2 (Singh et al., 2011). En la Figura 1.2b puede verse el resultado obtenido a través de este software, donde puede apreciarse que el plegamiento es muy similar.

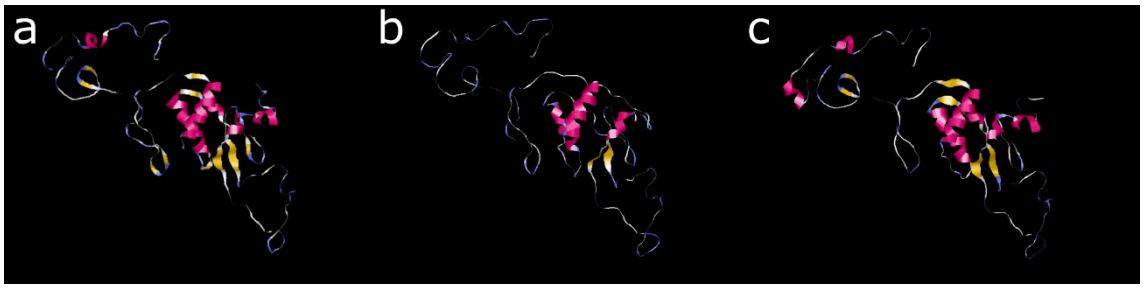

Figura 1.2. Estructuras terciarias obtenidas con I-Tasser y Phyre-2. En las figuras se muestran las diferentes conformaciones predichas de las proteínas. En rosa se representan las alpha hélices, y en amarillo las beta láminas. Como puede observarse, la estructura de las tres es muy similar.

Finalmente, predijimos la estructura terciaria de una secuencia de Nef de referencia (la secuencia 2XI1 de Protein Data Bank) a través de I-Tasser, ilustrada en la Figura 1.2c. Como puede verse, su estructura es muy similar a las obtenidas anteriormente para nuestro candidato, y en particular, a la que aparece en el panel a, obtenida con I-Tasser. Esta comparativa refleja que nuestro método es capaz de generar cadenas que se asemejan mucho a las originales (consecuencia de la optimización del alineamiento).

A continuación, con el objeto de probar que el método aquí propuesto ofrece mejores candidatos a vacuna que los métodos descritos en la literatura, realizamos una comparativa analizando tanto la inmunogenicidad de clase I (IEDB Class-I Immunogenicity) de cada candidato, como su desalineamiento promedio (esto es, cuánto difieren las posiciones de los aminoácidos respecto de la cadena de referencia de la proteína Nef). Para ello, obtuvimos tres nuevos candidatos: el primero se extrajo de nuestro trabajo anterior (Martínez et al., 2015) de λ-supercadenas no ponderadas; el segundo se obtuvo a través de la técnica Epigraph (LANL's Epigraph website); y el tercero a través de la técnica Consensus (LANL's Consensus website).

En la Tabla 1.3 pueden encontrarse los resultados para los 4 candidatos (el obtenido a través de las λ-supercadenas ponderadas, y los 3 mencionados en el párrafo anterior). Como era de esperar, el mayor valor de inumogenicidad lo consigue nuestro candidato ponderado, sugiriendo que generaría una mayor respuesta inmunitaria. Por otro lado, la proporción de aminoácidos desalineados fue similar en todos los candidatos, con la excepción de las λ-supercadenas no ponderadas, lo cual era presumible, ya que, a diferencia de los demás métodos, éste último no tiene en cuenta el alineamiento.

Además, con el objeto de comparar el algoritmo genético que diseñamos para este trabajo con otras técnicas heurísticas, desarrollamos un algoritmo hill-climbing mutiobjetivo, como el descrito en (Díaz & Suárez, 2001). Los resultados obtenidos indicaron que la aproximación al frente de Pareto utilizando este método fue mucho peor que la obtenida utilizando nuestro algoritmo (los valores obtenidos y un gráfico representado el frente de Pareto pueden encontrarse en Martínez et al., 2019).

Tabla 1.3. Comparación entre los candidatos obtenidos a través de distintos métodos

	Inmunogenicidad	Promedio de AA desalineados
Lambda-supercadena ponderada	1.8685	0.5115
Lambda-supercadena no ponderada	1.8409	1
Epigraph	1.2307	0.5114
Consensus	1.4103	0.5109

Investigación nº2: Primer diseño de una vacuna contra el SARS-CoV-2 usando λ-supercadenas

.

R.2.1. Relevancia

El COVID-19 es la mayor amenaza global de nuestro tiempo, y para combatirlo se han utilizado enormes esfuerzos tanto desde el ámbito público como desde el privado. Sin embargo, y a pesar de los avances en vacunas contra esta enfermedad, la eficacia a largo plazo de éstas no está clara. Por lo tanto, es aconsejable continuar desarrollando vacunas contra este virus, que a su vez sean lo más eficaces posible contra sus variantes venideras.

En este trabajo, hemos diseñado una serie de vacunas contra el SARS-CoV-2 aplicando nuestra técnica de λ-supercadenas ponderadas, y, además, hemos probado experimentalmente su eficacia, lo cual refuerza como prueba de concepto, nuestra metodología.

R.2.2. Diseño de la vacuna

Con el objeto de diseñar una vacuna contra el SARS-CoV-2, nuestra proteína objetivo (a partir de la cual obtener nuestros antígenos y diseñar nuestra vacuna) ha sido la proteína Spike, ya que, además de hacer de intermediaria entre el virus y las células del anfitrión a través del receptor ACE2, es una proteína que se expone en la superficie (Samrat et al., 2020; Zhang et al., 2020), lo cual es una virtud para que el sistema inmunitario pueda detectarla cuanto antes.

Para obtener las cadenas objetivo, hemos utilizado todas las secuencias de dicha proteína (22 secuencias) que estaban disponibles en GenBank (GenBank website) y GISAID (GISAID website) hasta el 4 de marzo del 2020. A continuación, se seleccionaron como potenciales antígenos (*Target strings, T*) los 9-meros contenidos en cualquiera de esas 22 cadenas, y estimamos su inmunogenicidad y afinidad al complejo

mayor de histocompatibilidad de clase I (HLA-I en humanos). Dichas estimaciones fueron combinadas para dar lugar a una función peso $w(s)$ que será la que maximicemos, y que se describe a continuación:

1. Se calcula la inmunogenicidad estimada $i(s)$ del antígeno s utilizando el "T cell class I pMHC immunogenicity predictor" de IEDB (como se ha hecho en la investigación nº1).
2. Después, nos restringimos a las cadenas que cuentan afinidad para con alelos que superan la barrera del 1% en la estimación de afinidad de MHC-I "Peptide binding to MHC class I molecule" (esta característica indica el grado de afinidad con las moléculas de histocompatibilidad de clase I. A mayor afinidad, más probable es que el sistema reconozca el antígeno y genere una respuesta inmunitaria). A este grupo lo denotaremos como $AI(s)$.
3. A continuación, con el objeto de hacer una vacuna lo más universal posible y dar más importancia a los alelos más frecuentes en la población, se ponderan los alelos por su frecuencia global estimada (obtenida del "The Allele Frequency Net Database"), y se obtiene un valor de la afinidad a través de $bI(s) = \sum_{i \in AI(s)} f(a)$, donde $f(a)$ indica la frecuencia del alelo a.
4. Ahora, normalizamos dichos pesos de la manera siguiente:

   - $i_N(s) = \frac{i(s) - m_i}{M_i - m_i}$, donde $min_{s \in T}\ i(s)$ y $M_i = max_{s \in T}\ i(s)$;
   - $bI_N(s) = \frac{bI(s) - m_{bI}}{M_{bI} - m_{bI}}$, donde $m_{bI} = min_{s \in T}\ bI(s)$ y $M_{bI} = max_{s \in T}\ bI(s)$;

5. Finalmente, obtenemos el peso del antígeno s a través de:

$$w(s) = \frac{3\ i_N(s) + bI_N(s)}{4}.$$

La sobreponderación de la inmunogenicidad se debió a que ésta es una estimación determinista, mientras que la de afinidad de unión es una estimación probabilística, en la que no todos los alelos posibles se han considerado, si no que solo se han tenido en cuenta los más frecuentes.

Una vez obtenidos los pesos, utilizamos el programa CPLEX Optimizer (IBM ILOG CPLEX Optimization Studio website) para ejecutar el algoritmo de programación entera descrito en la sección de métodos (ver Material y Métodos, M.4), obteniendo los resultados óptimos (esto es, con máximo valor de λ posible) para longitudes desde 9 hasta 280. En la Fig 2.1, representamos el ajuste de un modelo de regresión lineal para ajustar el λ en función de la longitud del candidato: λ = -0.579005 + 0.446982 · l (los resultados de la regression se encuentran en la Tabla 2.1). El R^2 obtenido por el modelo fue 0.999668, lo que indica un ajuste muy bueno, y sugiere que la pérdida del valor de λ no sufre grandes cambios a medida que la longitud de la vacuna aumenta.

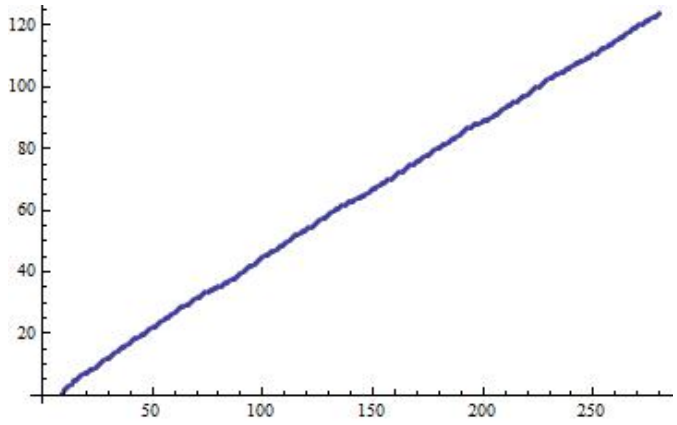

Fig 2.1. Scatterplot para λ. El eje de abcisas indica la longitud del candidato, mientras que el eje de ordenadas muestra el valor del λ correspondiente.

Tabla 2.1. Resultados de la regresión lineal

	Estimación del parámetro	**Error standard**	**P-valor**
Constante	-0.579005	0.0815211	1.08742×10^{-11}
Coef. de l	0.446982	0.000495704	1.07761×10^{-471}

A continuación, para llevar a cabo las pruebas experimentales, se seleccionó una de las soluciones del algoritmo (esto es, uno de los potenciales candidatos a vacuna). Para ello, se utilizó el algoritmo de VaxiJen (VaxiJen database), que sirve para estimar la potencialidad de los candidatos a generar respuesta. En particular, aquellos que obtienen una puntuación superior a 0.4 se consideran potencialmente inmunogénicos. Ése fue el caso de los candidatos de longitudes 22, 24, 67, 68, 69, 70, 175 y los de una longitud mayor de 184. Entre ellos, el que mayor puntuación obtuvo (un valor de 0.5545) fue el de longitud 22, a saber, el péptido STQDLFLPFFSNVTWFHAIHVS. Por lo tanto, dicha cadena fue la seleccionada como nuestro candidato a testar experimentalmente.

R.2.3. Ensayos experimentales de la eficiencia de la vacuna

Una vez seleccionado el candidato (denotado como CoVPSA), lo primero fue sintetizar el péptido, para realizar las pruebas de inmunogenicidad y eficiencia *in vivo*.

Las primeras pruebas se llevaron a cabo midiendo el retardo en la respuesta a la vacuna. Por un lado, se cargaron las células dendríticas (DC) con el candidato a vacuna, y por otro, los ratones fueron expuestos durante 7 días intraperitonealmente con péptidos del COVID-19. A continuación, se inoculó la vacuna en la pata izquierda del ratón, y se usó la pata derecha como control. Finalmente, se midió, cuarenta y ocho horas después, la respuesta generada en la pata izquierda, en comparación con su pata

derecha. Además, se utilizaron como control adicional células dendríticas cargadas con péptidos de bacterias no relacionadas con el COVID-19. Los resultados (barras azules de la Figura 2.2) indicaron respuestas inmunes mucho mayores cuando las DC se cargaron con la vacuna, que en los otros dos casos.

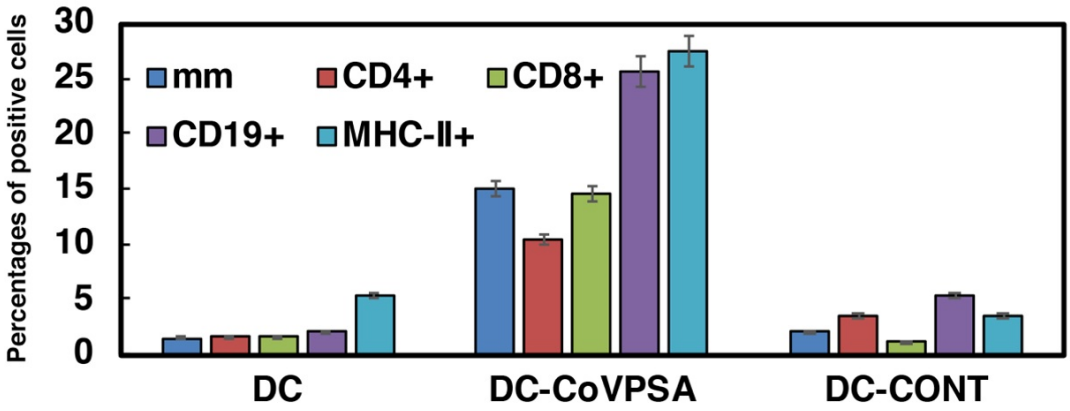

Fig.2.2. Inmunogenicidad del péptido CoVPSA. Las patas de los ratones (cepa C57BL/6, n = 5), fueron inoculadas con las células dendríticas (10^6 células por ratón) cargadas con diferentes péptidos (DC: sin péptido; DC-CoVPSA: con el candidato a vacuna; DC-CONT: con el péptido de control no relacionado). Se compararon los grosores de las patas medidos en mm (media$\pm SD$) con un test apareado. Después, se extrajeron los nódulos linfáticos de los ratones y se analizaron las poblaciones de células de interés, a través de citometría de flujo.

Los resultados de las diferentes poblaciones de células estudiadas indicaron que cuando se utilizó la vacuna, hubo una clara inducción del sistema inmune, involucrando formación de anticuerpos y la estimulación de células B, DC, y $CD4^+$ T.

Después, se analizó la producción de citoquinas antivirales (Figura 2.3). En particular, se observaron altos niveles de IFN-γ e IL-12, involucrados en la eficiencia de las vacunas y la respuesta antiviral, respectivamente.

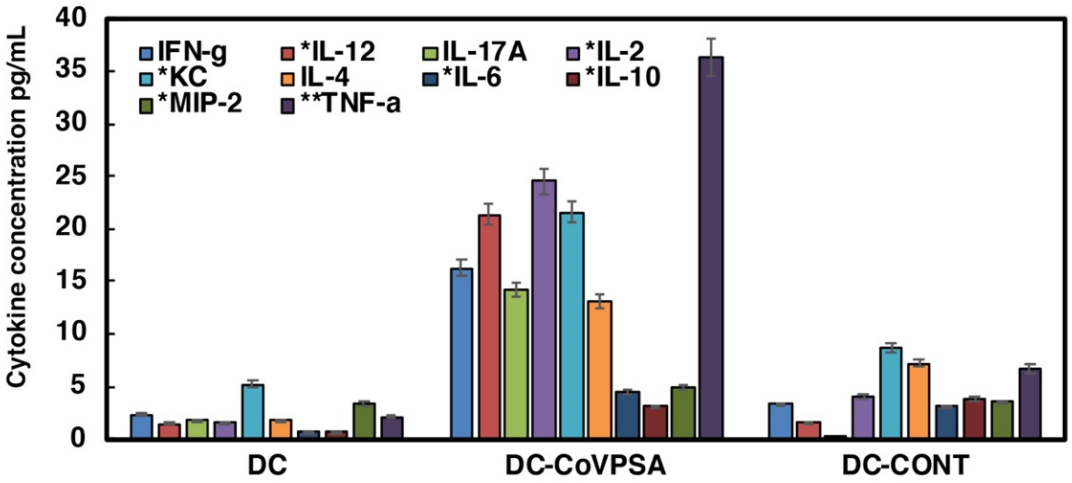

Fig 2.3. Niveles de citoquinas en ratones inoculados con células dendríticas. Los niveles de citoquinas se cuantificaron en suero del ratón, y fueron medidos utilizando un quit multiparamétrico Luminex desarrollado por Merck. Los resultados están expresados en pg/mL. * El nivel de citoquinas debe multiplicarse por 10. ** el nivel de citoquinas debe dividirse por 2.

Investigación nº3: Evaluación experimental de una vacuna personalizada contra el melanoma diseñada a través de optimización combinatoria

R.3.1. Relevancia

A pesar del creciente uso de las herramientas bioinformáticas, la comunidad biomédica, a día de hoy, aún se muestra reacia a confiar en los algoritmos y la computación para diseñar vacunas. Prueba de esto puede ser el desarrollo de las principales vacunas contra el virus del Covid-19, donde a pesar de los avances en computación, se han utilizado en el diseño y elaboración, una vez más, técnicas biológicas (ya sean vacunas de ARN mensajero o basadas en adenovirus). Es por eso que consideramos que son especialmente importantes los trabajos donde vacunas diseñadas computacionalmente muestran su potencial sobre datos experimentales.

Aquí, se presenta un estudio donde se han llevado a cabo todos los pasos hasta la comprobación *ex vivo* de la efectividad de una vacuna personalizada contra el melanoma, basada en parámetros bioinformáticos y optimización.

R.3.2. Selección de los neoantígenos

R.3.2.1. Obtención de muestras para los ensayos *ex vivo*

Inicialmente, el análisis partió de la información de seis pacientes con melanoma cutáneo en diferentes estadios (que se corresponde con en análisis de la investigación nº4), pero de cara a la prueba experimental de la vacuna, dadas las limitaciones para obtener sangre reciente de los pacientes (ya que varios habían fallecido para cuando se iban a comenzar los ensayos) y teniendo en cuenta las limitaciones económicas, se procedió con uno de los pacientes (al que nos referiremos como B057). Dicho paciente, fue diagnosticado con melanoma cutáneo en 1995, cuando tenía 32 años, y no contaba con antecedentes familiares de dicha enfermedad. El tumor primario fue clasificado como un melanoma nodular localizado en el tronco, en estadio IIB (de acuerdo con la

clasificación del American Joint Cancer Committee). El tumor tenía una profundidad de 4.4mm de acuerdo con el Índice de Breslow, sin ulceración, y presentaba la mutación BRAF-V600E. Se procedió a extirpar el tumor primario, y 54 meses después, se detectó metástasis linfática y se procedió a la operación.

R.3.2.2. Obtención de los posibles neoantígenos

Una vez obtenido el genoma de los pacientes, el siguiente paso fue el de determinar la estructura de los posibles neoantígenos partiendo de la mutación de DNA y su mutación correspondiente en el péptido asociado. Para ello, consideramos las mutaciones localizadas en el centro de un péptido de 15 aminoácidos, dejando a cada lado 7 aminoácidos no mutados. Nuestro análisis se limitó a las mutaciones en las cuales el péptido quedaba unívocamente determinado por su mutación en el DNA, independientemente de las posibles transcripciones que puedan asociarse a los péptidos a partir de las cadenas de DNA. Así, a pesar de que en la mayoría de los casos exista un único frame de lectura, esto es, una única posibilidad de "pasar" de DNA a aminoácidos, esto no fue siempre el caso, y, por lo tanto, se consideraron a priori tres frames de lectura (recuérdese que tres bases se corresponden con un aminoácido, por lo que las cadenas, leídas como nucleótidos, tendrían una longitud de 45 bases). Para cada mutación situada en el i-ésimo nucleótido (donde el codón de lectura se ha coloreado en rojo) se consideraron los tres posibles frames de lectura como sigue:

$$\dots i-21, i-20, i-19, \dots, i, i+1, i+2, \dots, i+21, i+22, i+23 \dots$$

$$\dots i-22, i-21, i-20, \dots, i-1, i, i+1, \dots, i+20, i+21, i+22 \dots$$

$$\dots i-23, i-22, i-21, \dots, i-2, i-1, i, \dots, i+19, i+20, i+21 \dots$$

Cada una de ellas corresponde a una transcripción posible, que daría un péptido de longitud 15 con la mutación en el aminoácido central. Para obtener los transcritos finales se utilizó el comando "genomeToTranscript" del paquete ensembldb del programa R.

R.3.2.3. Estimación de las características de los neoantígenos utilizando herramientas bioinformáticas

Una vez determinado cada neoantígeno n de 15 aminoácidos, se calcularon, para cada uno de ellos los siguientes siete valores:

- Estimación de la inmunogenicidad de Clase I. Para cada uno de los 7 péptidos de longitud 9 contenidos en n, calculamos la estimación de su inmunogenicidad utilizando la herramienta "T cell Class-I immunogenicity predictor" de IEDB Analysis Resource (IEDB Class-I Immunogenicity), y sumamos los 7 valores, obteniendo para cada i-ésimo neoantígeno, un valor $\{imm_i\}$ que fue almacenado en forma de vector, de longitud igual al número de neoantígenos total. Después, se normalizaron los valores del vector $\{imm_i\}$ a partir del valor mínimo y máximo de las inmunogenicidades estimadas, esto es, se obtuvo un nuevo vector $\{Nimm_i\}$, donde $Nimm_i = \frac{imm_i - m}{M - m}$, siendo m y M el mínimo y máximo de $\{imm_i\}$, respectivamente. Un valor alto en este estadístico indica que el péptido,

una vez reconocido, generaría previsiblemente una respuesta alta del sistema inmune.

- Estimación de la afinidad de unión al HLA-I. En este caso, consideramos los péptidos de longitud 9 contenidos en n del percentil 1 de la herramienta "Peptide binding to MHC class I molecules" (IEDB MHC-I binding) de IEDB, junto con sus correspondientes alelos. Para cada una de dichas combinaciones alelo-péptido, se calculó $-0.5\,p + 1$, donde p indica el percentil, y a continuación se sumaron, para cada péptido, todos los valores de las combinaciones alelo-péptido asociadas, obteniendo el vector de valores $\{HLAI_i\}$. De esta manera, se dio más peso a los neoantígenos más probables. Finalmente, se normalizó el vector $\{HLAI_i\}$ como se ha explicado antes, obteniendo el nuevo vector normalizado $\{NHLAI_i\}$. Un valor alto en este indicador se traduciría en que las células con moléculas de unión de clase I reconocerían con mayor probabilidad este neoantígeno.

- Estimación de la afinidad de unión al HLA-II. Para esta unión, se tuvieron en cuenta los neoantígenos de longitud 15 directamente, y nos quedamos con los alelos del percentil 10 de la herramienta "Peptide binding to MHC class II molecules" (IEDB MHC-II binding). Para cada una de las combinaciones alelo-péptido, se calculó $-0.05\,p + 1$, donde p indica el percentil, y a continuación se sumaron, para cada péptido, todos los valores de las combinaciones alelo-péptido asociadas, obteniendo el vector de valores $\{HLAII_i\}$. Por último, se normalizó el vector $\{HLAII_i\}$, obteniendo el nuevo vector normalizado $\{NHLAII_i\}$. Como en el caso anterior, un valor alto indicaría en que las células con moléculas de unión de clase II reconocerían con mayor probabilidad este neoantígeno.

- Frecuencia de la variante mutada. En este caso, calculamos el vector $\{vf_i\}$, a partir de los valores de la frecuencia de la mutación con respecto a su variante no mutada, y a continuación lo normalizamos como en los casos anteriores, obteniendo el vector $\{Nvf_i\}$. De esta manera, se buscó dar más peso a las mutaciones que más aparecieran en el tumor, facilitando así que el sistema inmune los reconozca.

- Probabilidad de actuar como antígeno. Utilizando la herramienta VaxiJen (VaxiJen database), estimamos la probabilidad de una secuencia de ser considerada por el sistema inmunitario como antígeno, y en particular, como antígeno tumoral. Finalmente, normalizamos el vector entre 0 y 1 como en caso anteriores, obteniendo $\{Nap_i\}$. Así, se fomentó la selección de las secuencias con mayores posibilidades de actuar como neoantígeno y de generar una respuesta inmune.

- Hidrofobicidad. Primero, se calculó el vector $\{gr_i\}$ a partir de los valores correspondientes al índice GRAVY para cada neoantígeno, utilizando la herramienta ProtParam (ExPASy database). Después, normalizamos el vector entre 0 y 1, y por último, dado que nuestro estadístico se corresponde con la hidrofobicidad, y se busca maximizar la hidrofilicidad (ya que interesa un péptido lo más expuesto posible para que las células puedan reconocerlo), se

calculó la resta $Nhpl_i = 1 - Ngr_i$, donde $\{Nhpl_i\}$ es nuestro vector final, y Ngr_i el obtenido después de normalizar el índice GRAVY.

- TAP (transporte para la presentación de antígenos) y proteasoma. Para el cálculo de esta variable, se utilizó la herramienta "Proteasomal cleavage/TAP transport/MHC class I combined predictor" (IEDB TAP/transport), y a continuación se normalizó entre 0 y 1, obteniendo el vector $\{NTAP_i\}$. Al maximizar esta característica, aumentan las probabilidades de que el neoantígeno sea presentado y degradado con éxito por las células presentadoras de antígenos.

R.3.2.4. Optimización y diseño de la vacuna

Una vez estimadas las características de cada neoantígeno, una 7-tupla

$$v(n) = \left(Nimm_i(n), NHLAI_i(n), NHLAII_i(n), Nvf_i(n), Nap_i(n), Nhpl_i(n), NTAP_i(n)\right)$$

fue asociada a cada cadena n. A continuación, se consideró la función:

$$f(n) = 0.2\, Nimm_i(n) + 0.2\, NHLAI_i(n) + 0.2\, NHLAII_i(n) + 0.1\, Nvf_i(n) + 0.1\, Nap_i(n) + 0.1\, Nhpl_i(n) + 0.1\, NTAP_i(n),$$

y la función objetivo $F(S) = \sum_{n \in S} f(n)$, donde S es un subconunto de cardinal 6 del conjunto de los N neoantígenos. Esto es, se obtuvieron 6 péptidos $\{n_1, n_2, n_3, n_{4,} n_5, n_6\}$ tales que $F\left(\{n_1, n_2, n_3, n_{4,} n_5, n_6\}\right) = max_{S \subset N, |S|=6} F(S)$ (lo que a su vez es equivalente a seleccionar lo seis neoantígenos con mayor valor para $f(n)$). Como puede observarse, las características principales (inmunogenicidad de Clase I y afinidad de unión al HLA-I y HLA-II) fueron sobreponderadas con respecto al resto, dado que son las variables más comúnmente utilizadas en la literatura (Sahin et al., 2017). Finalmente, para la prueba experimental, esta solución fue agrupada en dos péptidos: el primero fue la concatenación de n_1, n_2 y n_3, y el segundo, la de $n_{4,} n_5$ y n_6.

Los resultados obtenidos para los 6 neoantígenos con mayor valor $f(n)$ para el paciente B057 están representados en la Tabla 3.1.

Tabla 3.1. Valores de las características bioinformáticas para los 6 neoantígenos seleccionados para el paciente B057

Neoantígeno	*Nimm*	*NHLA1*	*NHLAII*	*Nhpl*	*NTAP*	*Nap*	*Nvf*
DWLEWLRQLSLELLK	0.56	0.88	1	0.37	0.72	0.52	0.20
FRDQSLSYHHTMVVQ	0.34	1	0	0.50	0.56	0.45	0.30
IGRFANYFRNLLPSN	0.85	0.67	0.53	0.41	0.57	0.56	0.33
MRHSFFSEVNWQDVY	0.88	0	0.45	0.54	0.72	0.43	0.34
RLFMHHVFLEPITCV	1	0.37	0.43	0	0.60	0	0.76
CSRRFYQFTKLLDSV	0.52	0.56	0.79	0.40	0.61	0.38	0.47

R.3.3. Evaluación *ex vivo* de la respuesta inmune

Una vez seleccionados los neoantígenos, estos fueron agrupados en dos péptidos no solapados de longitud 45 (compuestos por tres péptidos consecutivos de longitud 15 cada uno):

Péptido 1: DWLEWLRQLSLELLKFRDQSLSYHHTMVVQIGRFANYFRNLLPSN

Péptido 2: RHSFFSEVNWQDVYRLFMHHVFLEPITCVCSRRFYQFTKLLDSV

Después, estos péptidos fueron sintetizados y encapsulados en dos nanopartículas (NPs), y posteriormente fueron testados *ex vivo*.

El objetivo final es el de activar las células T contra el antígeno seleccionado. Una vez los neoantígenos encapsulados son captados por las células dendríticas (DCs), éstas se transforman en células dendríticas maduras, lo que les permite presentar el antígeno a otras células para su posterior reconocimiento. En este proceso, las DCs sobre-expresan una serie de marcadores de superficie (HLA-DR, CD80, CD83 y CD86) (Aerts-Toegaert et al., 2007; Lu et al., 1997). Para determinar el efecto de la maduración de las células dendríticas, se estudió la expresión de dichos marcadores en cinco condiciones distintas: en células tratadas directamente con los neoantígenos libres (Ag_{B057}), tratadas directamente con péptidos de control (Ag_{ctrl}), tratadas con los neoantígenos encapsulados (NP_{B057}), tratadas con péptidos de control encapsulados (NP_{ctrl}), y, por último, tratadas con la nanopartícula sin péptido (NP_{blank}). Los resultados de este estudio pueden verse en la Figura 3.1, a partir de lo cual se deriva que los neoantógenos, y en particular, los neoantígenos encapsulados, indujeron la maduración de las DCs.

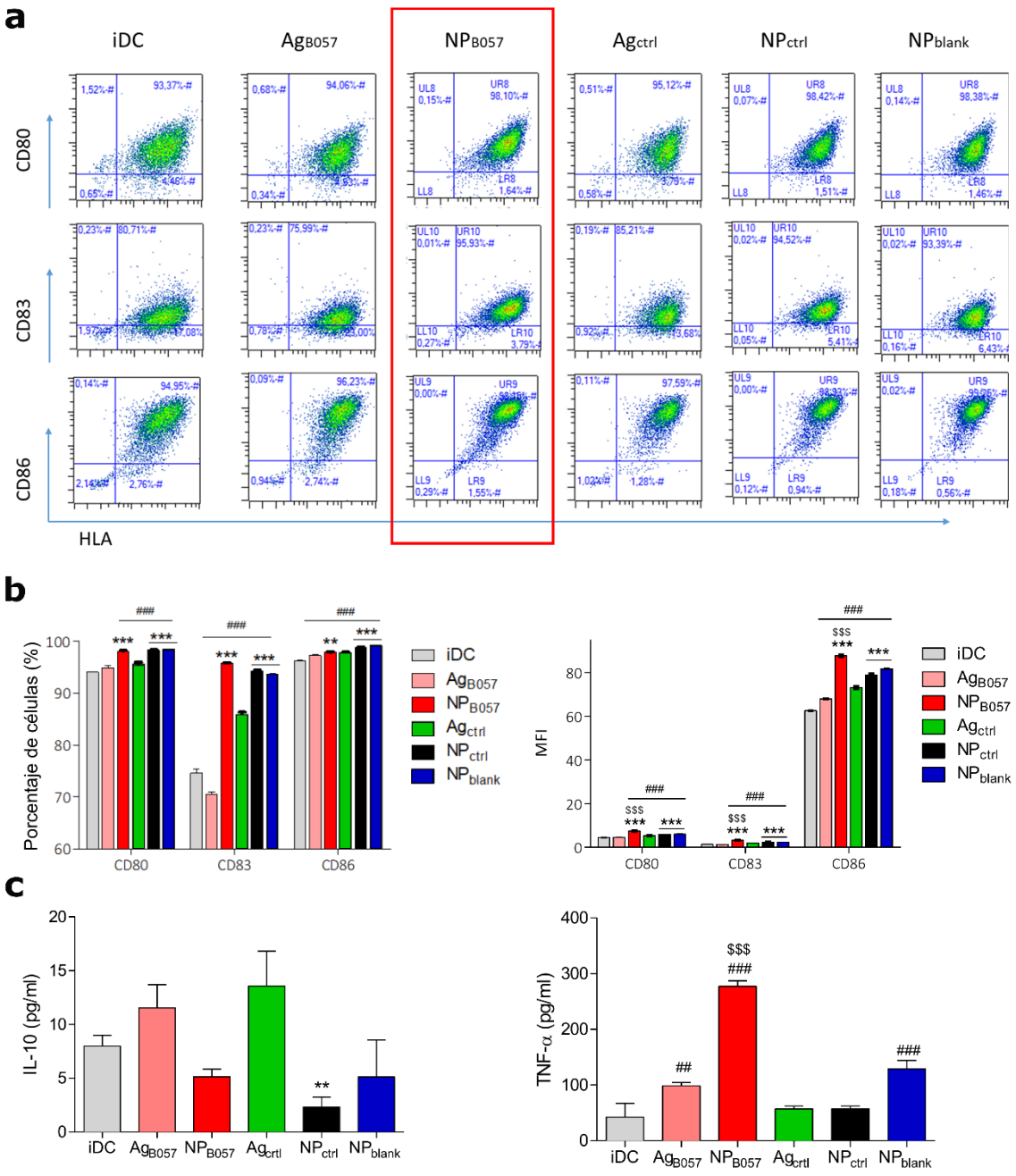

Fig. 3.1. Estudio de la expresión de los marcadores bajo las cinco condiciones. Cinco días después de aislar la sangre del paciente, las DCs inmaduras (iDC) fueron incubadas junto con los neantígenos y las nanopartículas, y su maduración fue evaluada el sexto día (estudiando el número de $CD80^+$-HLA-DR^+ DCs, $CD83^+$-HLA-DR^+ DCs y $CD86^+$-HLA-DR^+ DCs). **a** Citometría de flujo de las DCs después de la maduración. **b** Porcentaje de marcadores de maduración celular y media de la intensidad de la fluorescencia (MFI). **c** Resultados del ELISA para medir la secreción de TNF-α y IL-10. Todas las muestras fueron analizadas por triplicado, y para representar la signiticatividad estadística, se utilizó la notación: ## $p<0.01$ y ### $p<0.001$ con respecto a las iDCs; **$p<0.01$ y ***$p<0.001$ con respecto a los antígenos libres; \$\$\$ $p<0.001$ con respecto al resto de grupos).

Aparte del número de células que expresaron cada marcador, se estudió la media de la intensidad de la fluorescencia (MFI) de los marcadores de maduración, que son un indicativo de la cantidad de marcador expresado en cada célula. Los resultados

mostraron un patrón similar, pero se observó una mayor expresión del marcador de maduración para la nanopartícula con neoantígeno dirigida al paciente (NP_{B057}).

Para analizar a la secreción de citoquinas de las DCs, se estudiaron el TNF-α (factor de necrosis alpha) y la IL-10 (interleucina 10). El TNF-α es un pirógeno endógeno que regula la respuesta inmune y cuyos efectos incluyen la inhibición de la tumorgénesis (Wajant, 2009). Además, el TNF-α induce la migración de las DCs a los nodos linfáticos, donde se encuentran las células T (Tough, 2008). Por lo tanto, la nanopartícula que transporta los neoantígenos del paciente produce efectos beneficiales en la maduración y activación de las DCs, potenciando la respuesta inmune innata y facilitando también el inicio de una respuesta específica al fomentar la migración de las DCs y su interacción con las células T. Por otro lado, la IL-10 es una citocina con propiedades antiinflamatorias relacionada con mecanismos de inmunosupresión (y en particular con la supresión de citocinas de células T) (Schülke, 2018; Tucci et al., 2019). En este caso, no se observó un incremento significativo de este marcador en ninguno de los grupos, lo que se consideró como un factor positivo.

Después, se pasó a estudiar el efecto de la vacuna en la proliferación y secreción de citocinas de las células T a través del estudio de los linfocitos T $CD4^+$ y $CD8^+$ (necesarias para una correcta supresión del tumor, siendo las células $CD8^+$ las células citotóxicas, y las $CD4^+$ las necesarias para su correcta activación (Ostroumov et al., 2018)).

La proliferación de las células T (tanto $CD4^+$ como $CD8^+$) se vio incrementada cuando se utilizaron los neoantígenos del paciente. En el caso de las $CD4^+$, los neoantígenos libres parecieron inducir mayor proliferación que los encapsulados, y, por otro lado, los neoantígenos propios generaron mayor proliferación que los péptidos de control (Figura 3.2) (Flores et al., 2019). Por otro lado, tanto los neoantígenos libres (Ag_{B057}) como los encapsulados (NP_{B057}) fueron capaces de inducir proliferación de las células $CD8^+$, y de una manera significativamente mayor que los antígenos de control (Figura 3.2).

Como se muestra en la Figura 3.2, la secreción de IL-2, inductor del crecimiento, proliferación y supervivencia de los linfocitos (Boyman et al., 2010) fue significativamente mayor para NP_{B057}. Además, el IFN- γ (interferón gamma, activador de la respuesta inmunitaria) también fue significativamente mayor para NP_{B057} que para en el resto de grupos.

Estos resultados prueban que los neoantígenos encapsulados en nanopartículas (NP_{B057}) son capaces de producir la proliferación de células T $CD4^+$ y $CD8^+$, esenciales para obtener inmunidad antitumoral (Ostroumov et al., 2018), ya que ambas son necesarias para obtener una respuesta citotóxica óptima (Borst et al., 2018).

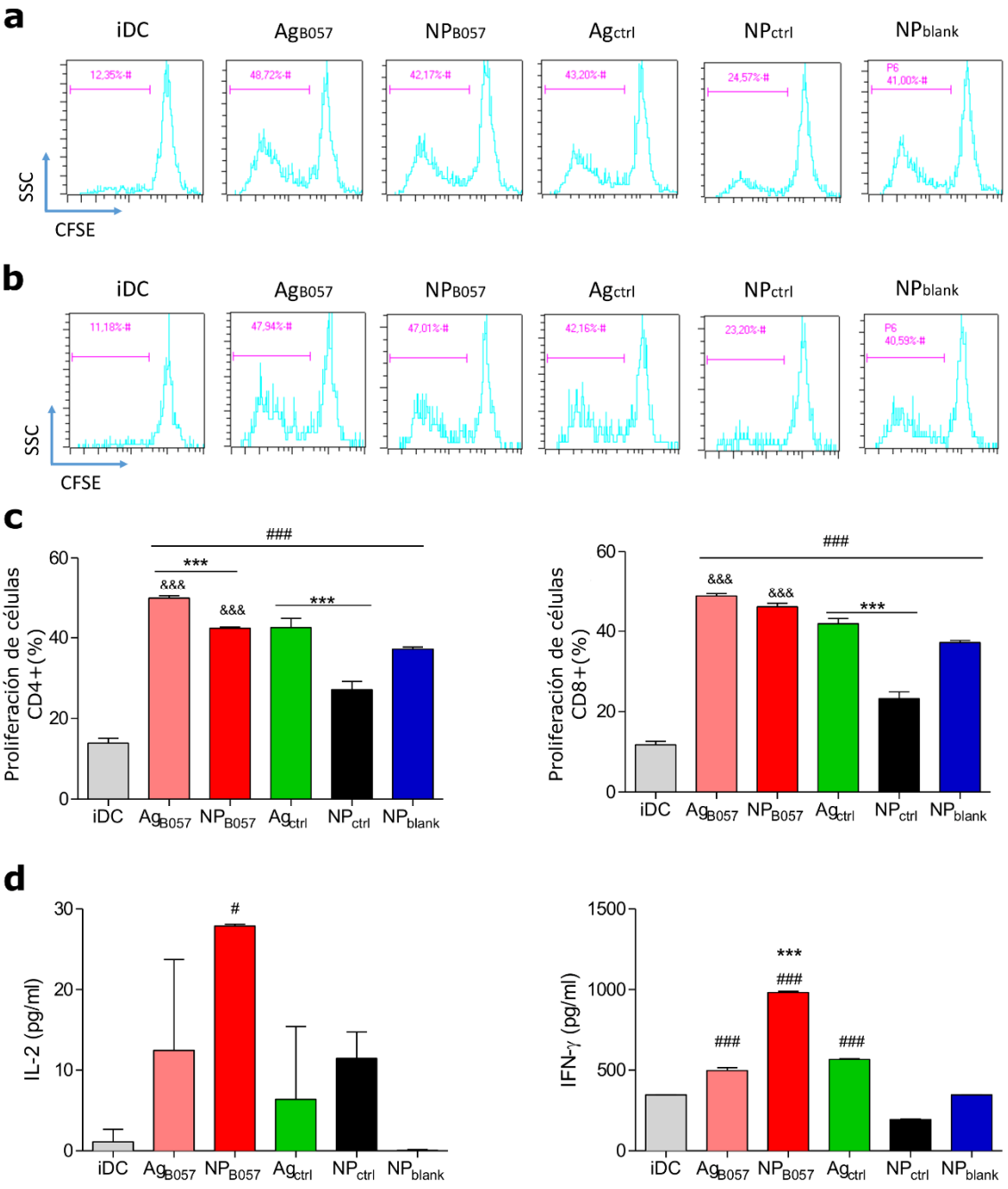

Fig. 3.2. Estudio del efecto de la vacuna en la proliferación y secreción de citocinas. a y **b** citometría de flujo de las células T $CD4^+$ y $CD8^+$ T, respectivamente. **c** porcentaje de proliferación de la respuesta de linfocitos $CD4^+$ y $CD8^+$ inducida por las DCs. **d** Liberación de IL-2 e IFN-γ. Las proliferaciones se analizaron por triplicado, y las citoquinas por duplicado (# $p<0.05$ y ### $p<0.001$ para las iDCs; ***$p<0.001$ para los antígenos libres; &&& $p<0.001$ para el control).

Investigación nº4: Análisis de la capacidad de detección de neoantígenos a través de las herramientas bioinformáticas

R.4.1. Relevancia

En las últimas décadas, la importancia de los neoantígenos a la hora de desarrollar vacunas antitumorales ha crecido considerablemente. Este tipo de antígenos, en teoría, generan una fuerte respuesta inmunitaria, mientras que, su versión no mutada, que habitualmente difiere solamente en un aminoácido, no genera respuesta alguna. Por otro lado, cada vez se utilizan más herramientas bioinformáticas para estimar la respuesta inmunitaria que generará un potencial antígeno. Inevitablemente, surge una pregunta, ¿son las herramientas bioinformáticas capaces de detectar esta diferencia de un aminoácido, de manera que la respuesta estimada de la versión mutada sea significativamente distinta?

Para responderlo, en este trabajo se han obtenido experimentalmente los neoantígenos de seis pacientes diagnosticados con melanoma cutáneo, y se han comparado las estimaciones de afinidad de unión de las moléculas HLA I y II de la versión mutada y no mutada.

R.4.2. Obtención de la muestra

Como se ha mencionado en la investigación nº3, este estudio partió de la muestra de las biopsias de seis pacientes diagnosticados con melanoma cutáneo en los hospitales de Basurto y Cruces. Con el objeto de obtener una diversidad mutacional suficiente, pero también una cantidad de mutaciones alta, se escogieron dichos pacientes con diferentes estadios del cáncer, pero a su vez, se seleccionaron fases avanzadas de éste. En particular, se escogió un paciente en estadio IB (hasta 2 centímetros de profundidad del tumor y sin ulceración), dos en estadio IIB (profundidad entre 2 y 4 centímetros con ulceración, o mayor de 4 cm sin ulceración; uno de estos pacientes fue para el cual se

diseñó la vacuna personalizada en la investigación nº3), y por último dos pacientes en estadio IIC (profundidad mayor de 4 cm y ulceración) (Thompson, 2002).

A continuación, dado que el análisis de todo el mutanoma de los pacientes era demasiado costoso, nos centramos en las regiones con mayor variabilidad en este tipo de cánceres, entre las que se encuentran las relacionadas con las regiones que codifican las proteínas BRAF, NRAS, MAP2K1 o MAP2K2 (Edlundh-Rose et al., 2006; Nikolaev et al., 2012). Para seleccionar las mutaciones propias del tumor y descartar las que suceden como consecuencia de la división celular (presentes tanto en células cancerosas como no cancerosas), secuenciamos también las regiones de células regulares de sangre, y nos quedamos con aquellas que solo estaban presentes en el tumor.

Después, extrajimos los neoantígenos siguiendo el procedimiento descrito en la investigación nº3 (punto R.3.2.2.), con la salvedad de que se consideraron cadenas de longitud 17, con la mutación en el noveno aminoácido. Nótese que en este estudio se utilizaron cadenas con 2 aminoácidos más que en la investigación anterior. Esto se debe a que, para la investigación nº3, se buscó evitar que la mutación pudiera quedar en el extremo de la 9-tupla al dividir el péptido en ventanas de 9 aminoácidos, ya que, al buscar el reconocimiento por el sistema inmunitario, los aminoácidos de los extremos suelen servir de "anclaje", y a priori podría ser más difícil su reconocimiento. Sin embargo, para este estudio no es relevante, ya que estos tecnicismos no son considerados por los algoritmos analizados. Al tomar cadenas de longitud 17, como hemos dicho, nos aseguramos de que cogiendo ventanas de 9 aminoácidos (que es el número de aminoácidos para el cual los programas bioinformáticos han sido optimizados), la mutación quede dentro. Por otro lado, el hecho de aumentar la longitud sirvió para aumentar la muestra de "fragmentos mutados" de longitud 9 que serían posteriormente considerados.

Antes de poder evaluar la afinidad de unión de los neoantígenos obtenidos, dado que los complejos HLA son altamente polimórficos y varían dependiendo del individuo, se secuenciaron los genes responsables de codificar el complejo mayor de histocompatibilidad (que se encuentra en el sexto cromosoma (Moutaftsi et al., 2006)), y se identificaron los alelos HLA de clases I y II correspondientes a cada paciente.

Finalmente, se utilizaron las herramientas de IEDB (Zhang et al., 2008) para estimar la afinidad de unión de las moléculas HLA-I y II, escogiendo para la primera las longitudes de 9 a 14, y para la segunda, longitud 15.

R.4.3. Comparación de las cadenas

Una vez obtenidas las estimaciones de afinidad para cada cadena, siguiendo las indicaciones de IEDB, utilizamos la variable "percentile rank" para separar las cadenas con potencial de unión, de las que no lo tenían. Como se ha mencionado en la investigación nº3, se consideró que tenían potencial de unión aquellas cadenas que obtuvieron un percentile rank $\leq 1\%$ para las estimaciones de clase I, y $\leq 10\%$ para las de clase II. En la tabla 4.1 se muestran el número de cadenas que potencialmente se unirían a las moléculas de los complejos de histocompatibilidad I y II, dividido por

pacientes. Además, en la Figura 4.1, se muestra la distribución de dichos valores a través de un diagrama de caja.

Tabla 4.1. Número de cadenas con potencial de unión a moléculas HLA-I y II, divididos por número de paciente. Además, M indica análisis sobre cadenas mutadas (esto es, el neoantígenos), y NM sobre su versión no mutada

	1º paciente	2º paciente	3º paciente	4º paciente	5º paciente	6º paciente
HLA-I M	57	88	8	121	20	26
HLA-I NM	49	70	4	114	20	30
HLA-II M	51	229	26	144	47	21
HLA-II NM	45	193	22	118	33	22

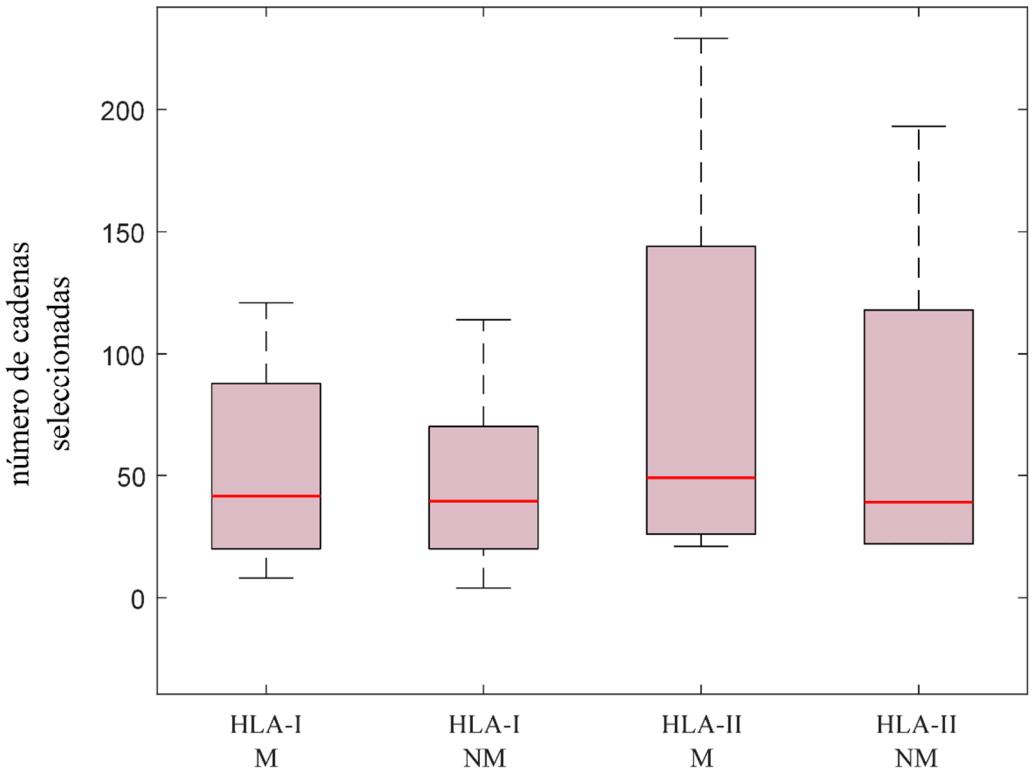

Fig. 4.1. Diagrama de caja representando la distribución de péptidos que superaron el umbral del "percentile rank". La línea roja indica la mediana, M indica versión mutada, y NM versión no-mutada.

Además, se analizaron las medias $\pm$ desviación estándar de los porcentajes de péptidos que pasaron el corte. El porcentaje de péptidos mutados que pasaron el corte de la herramienta para las moléculas de clase I fue $0.69\% \pm 0.31\%$, variando de 0.27% a 0.98%; para los no-mutados, fue $0.65\% \pm 0.36\%$, variando entre 0.14% y 1.13%; en el caso del HLA-II, para los mutados se obtuvo que pasaron el corte el $15.24\% \pm 2.96\%$, variando de 10.63% a 19.88%; y finalmente, para los no-mutados, los resultados fueron $13.02\% \pm 2.93\%$, variando de 9.38% a 16.75%.

Después, para estudiar la hipótesis de este trabajo, esto es, para analizar si el número de péptidos mutados es significativamente mayor que el de no-mutados de acuerdo con las herramientas bioinformáticas (comparando los resultados de la Tabla 4.1), se realizaron dos contrastes, uno comparando los resultados del HLA-I, y otro comparando los del HLA-II.

Primero, para utilizar los test estadísticos apropiados, se analizó la normalidad de la distribución de las variables, obteniendo unos p-valores de 0.837, 0.978, 0.43 y 0.476, para HLA-I mutado, HLA-I no-mutado, HLA-II mutado y HLA-II no mutado, respectivamente. Por lo tanto, no se rechazó la hipótesis de normalidad, y se aplicaron t-test apareados para cada comparativa.

Para el primer test estadístico, se utilizó la hipótesis nula de que la diferencia entre las medias del número de potenciales antígenos mutados por paciente era menor o igual que la de los no-mutados, el p-valor para la comparativa del HLA-I fue 0.068, con un t-estadístico de 1.78, y un intervalo de confianza de $CI^{0.95}_{\mu_1 \leq \mu_2} = (-0.74, \infty)$. Por lo tanto, no se rechazó la hipótesis nula, y no pudo concluirse que para el HLA-I, el número de antígenos no-mutados detectado fuera menor que los mutados.

Finalmente, se hizo la misma prueba para el número de péptidos que pasaron el corte en el caso del HLA-II, obteniendo un p-valor de 0.03, un t-estadístico de 2.43, y un intervalo de confianza de $CI^{0.95}_{\mu_1 \leq \mu_2} = (2.44, \infty)$. En consecuencia, se rechazó la hipótesis nula, aceptando que, de acuerdo a las herramientas bioinformáticas utilizadas, los neoantígenos se unirían con mayor afinidad a las moléculas HLA-II que su versión no mutada.

Referencias

- Aerts-Toegaert, C., Heirman, C., Tuyaerts, S., Corthals, J., Aerts, J. L., Bonehill, A., ... & Breckpot, K. (2007). CD83 expression on dendritic cells and T cells: correlation with effective immune responses. *European journal of immunology*, *37*(3), 686-695.

- Awadasseid, A., Wu, Y., Tanaka, Y., & Zhang, W. (2021). Current advances in the development of SARS-CoV-2 vaccines. *International journal of biological sciences*, *17*(1), 8.

- Bergmann-Leitner, E. S., Chaudhury, S., Steers, N. J., Sabato, M., Delvecchio, V., Wallqvist, A. S., ... & Angov, E. (2013). Computational and experimental validation of B and T-cell epitopes of the in vivo immune response to a novel malarial antigen. *PloS one*, *8*(8).

- Borst, J., Ahrends, T., Bąbała, N., Melief, C. J., & Kastenmüller, W. (2018). CD4+ T cell help in cancer immunology and immunotherapy. *Nature Reviews Immunology*, *18*(10), 635-647.

- Boyman, O., Cho, J. H., & Sprent, J. (2010). The role of interleukin-2 in memory CD8 cell differentiation. In *Memory T Cells* (pp. 28-41). Springer, New York, NY.

- Bryson, C. J., Jones, T. D., & Baker, M. P. (2010). Prediction of immunogenicity of therapeutic proteins. *BioDrugs*, *24*(1), 1-8.

- Calis, J. J., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., ... & Peters, B. (2013). Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS computational biology*, *9*(10).

- Chen, J., Wu, F., Lin, D., Kong, W., Cai, X., Yang, J., ... & Cao, P. (2021). Rational optimization of a human neutralizing antibody of SARS-CoV-2. *Computers in biology and medicine*, *135*, 104550.

- COVID-19 website. https://covid 19.who.int.

- Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, *6*(2), 182-197.

- Díaz, R., & Suárez, A. R. (2001). A study of the capacity of the stochastic Hill

Climbing to solve multi-objective problems. In *Proceedings of the Third International Symposium on Adaptive Systems-Evolutionary Computation and Probabilistic Graphical Models, La Habana: Institute of Cybernetics, Mathematics and Physics* (pp. 37-40).

- Dong, Y., Dai, T., Wei, Y., Zhang, L., Zheng, M., & Zhou, F. (2020). A systematic review of SARS-CoV-2 vaccine candidates. *Signal transduction and targeted therapy*, *5*(1), 1-14.
- Edlundh-Rose, E., Egyha, S., Omholt, K., Mansson-Brahme, E., Platz, A., Hansson, J., & Lundeberg, J. (2006). NRAS and BRAF mutations in melanoma tumours in relation to clinical characteristics: a study based on mutation screening by pyrosequencing. *Melanoma research*, *16*(6), 471-478.
- Estrada, E. (2020). COVID-19 and SARS-CoV-2. Modeling the present, looking at the future. *Physics Reports*, *869*, 1-51.
- ExPASy database. https://web.expasy.org/protparam/protparam-doc.html
- Fibonacci, L. (2002). Liber abaci. *SIGLER, LE Fibonacci's Liber Abaci A Translation into Modern English of Leonardo Pisano's Book of Calculation. New York: Springer*.
- Fisher, R. A. (1950). Statistical methods for research workers., (llth ed. revised).
- Flanagan, K. L., Best, E., Crawford, N. W., Giles, M., Koirala, A., Macartney, K., ... & Wen, S. C. (2020). Progress and pitfalls in the quest for effective SARS-CoV-2 (COVID-19) vaccines. *Frontiers in immunology*, 2410.
- Flores, I., Hevia, D., Tittarelli, A., Soto, D., Rojas-Sepúlveda, D., Pereda, C., ... & López, M. N. (2019). Dendritic cells loaded with heat shock-conditioned ovarian epithelial carcinoma cell lysates elicit T cell-dependent antitumor immune responses in vitro. *Journal of immunology research*, *2019*.
- Fritsch, E. F., Rajasagi, M., Ott, P. A., Brusic, V., Hacohen, N., & Wu, C. J. (2014). HLA-binding properties of tumor neoepitopes in humans. *Cancer immunology research*, *2*(6), 522-529.
- Gaebler, C., & Nussenzweig, M. C. (2020). All eyes on a hurdle race for a SARS-CoV-2 vaccine. *Nature*, *586*, 501-502.
- Galton, F. (1877). Typical laws of heredity. III. *Nature*, *15*(389), 512-514.
- Galton, F. (1889/1). Natural Inheritance. Macmillan & Co.
- Galton, F. (1889/2). I. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, *45*(273-279), 135-145.
- GenBank website. www.ncbi.nlm.nih.gov/genbank/.
- GISAID website. https://www.gisaid.org/

- Gusfield, D. (1997). Algorithms on stings, trees, and sequences: Computer science and computational biology. *Acm Sigact News*, *28*(4), 41-60.

- Heiny, A. T., Miotto, O., Srinivasan, K. N., Khan, A. M., Zhang, G. L., Brusic, V., ... & August, J. T. (2007). Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. *PloS one*, *2*(11).

- Hemmer, B., Kondo, T., Gran, B., Pinilla, C., Cortese, I., Pascal, J., ... & Martin, R. (2000). Minimal peptide length requirements for CD4+ T cell clones—implications for molecular mimicry and T cell survival. *International immunology*, *12*(3), 375-383.

- Henry-Labordere, A. L. (1969). The record balancing problem: A dynamic programming solution of a generalized traveling salesman problem. *Revue Francaise D Informatique DeRecherche Operationnelle*, *3*(2), 43-49.

- HIV Molecular Immunology Database website. www.hiv.lanl.gov/content/immunology.

- Hodgson, S. H., Mansatta, K., Mallett, G., Harris, V., Emary, K. R., & Pollard, A. J. (2021). What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. *The lancet infectious diseases*, *21*(2), e26-e35.

- Hu, Z., Ott, P. A., & Wu, C. J. (2018). Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nature Reviews Immunology*, *18*(3), 168-182.

- IBM ILOG CPLEX Optimization Studio website. www-03.ibm.com/software/products/us/en/ibmilogcpleoptistud/.

- IEDB Class-I Immunogenicity. http://tools.iedb.org/immunogenicity/

- IEDB MHC-I binding. http://tools.iedb.org/mhci/

- IEDB MHC-II binding. http://tools.iedb.org/mhcii/

- IEDB TAP/transport. http://tools.iedb.org/processing/

- Jaiswal, G., & Kumar, V. (2020). In-silico design of a potential inhibitor of SARS-CoV-2 S protein. *PLoS One*, *15*(10), e0240004.

- Java website. www.java.com

- Kakimi, K., Karasaki, T., Matsushita, H., & Sugie, T. (2017). Advances in personalized cancer immunotherapy. *Breast Cancer*, *24*(1), 16-24.

- Khan, A. M., Heiny, A. T., Lee, K. X., Srinivasan, K. N., Tan, T. W., August, J. T., & Brusic, V. (2006/1). Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus. In *BMC bioinformatics* (Vol. 7, No. 5, p. S4). BioMed Central.

- Khan, A. M., Miotto, O., Heiny, A. T., Salmon, J., Srinivasan, K. N., Nascimento, E. J., ... & August, J. T. (2006/2). A systematic bioinformatics approach for selection of

epitope-based vaccine targets. *Cellular immunology*, *244*(2), 141-147.

- Khan, J. M., Kumar, G., & Ranganathan, S. (2012). In silico prediction of immunogenic T cell epitopes for HLA-DQ8. *Immunome Research*, *8*(1), 1-9.

- Krammer, F. (2020). SARS-CoV-2 vaccines in development. *Nature*, *586*(7830), 516-527.

- Kreiter, S., Castle, J. C., Türeci, Ö., & Sahin, U. (2012). Targeting the tumor mutanome for personalized vaccination therapy. *Oncoimmunology*, *1*(5), 768-769.

- LANL's Consensus website. https://www.hiv.lanl.gov/content/sequence/CONSENSUS/SimpCon.html.

- LANL's Epigraph website. https://www.hiv.lanl.gov/content/sequence/EPIGRAPH/epigraph.html.

- Lu, P., Wang, Y. L., & Linsley, P. S. (1997). Regulation of self-tolerance by CD80/CD86 interactions. *Current opinion in immunology*, *9*(6), 858-862.

- Lundegaard, C., Lund, O., & Nielsen, M. (2011). Prediction of epitopes using neural network-based methods. *Journal of immunological methods*, *374*(1-2), 26-34.

- Martínez, L., Milanič, M., Legarreta, L., Medvedev, P., Malaina, I., & Ildefonso, M. (2015). A combinatorial approach to the design of vaccines. *Journal of Mathematical biology*, *70*(6), 1327-1358.

- Martínez, L., Milanič, M., Malaina, I., Álvarez, C., Perez, M. B., & de la Fuente, I. M. (2019). Weighted lambda superstrings applied to vaccine design. *PloS one*, *14*(2).

- Mathematica website. https://www.wolfram.com/mathematica/.

- Miller, C. E., Tucker, A. W., & Zemlin, R. A. (1960). Integer programming formulation of traveling salesman problems. *Journal of the ACM (JACM)*, *7*(4), 326-329.

- Moreau, V., Fleury, C., Piquer, D., Nguyen, C., Novali, N., Villard, S., ... & Molina, F. (2008). PEPOP: computational design of immunogenic peptides. *Bmc Bioinformatics*, *9*(1), 71.

- Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H. H., ... & Sette, A. (2006). A consensus epitope prediction approach identifies the breadth of murine TCD8+-cell responses to vaccinia virus. *Nature biotechnology*, *24*(7), 817-819.

- Murphy, K., & Weaver, C. (2016). *Janeway's immunobiology*. Garland science.

- Nickle, D. C., Rolland, M., Jensen, M. A., Pond, S. L. K., Deng, W., Seligman, M., ... & Jojic, N. (2007). Coping with viral diversity in HIV vaccine design. *PLoS computational biology*, *3*(4).

- Nielsen, M., Lund, O., Buus, S., & Lundegaard, C. (2010). MHC class II epitope

predictive algorithms. *Immunology*, *130*(3), 319-328.

- Nikolaev, S. I., Rimoldi, D., Iseli, C., Valsesia, A., Robyr, D., Gehrig, C., ... & Antonarakis, S. E. (2012). Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nature genetics*, *44*(2), 133-139.

- North, A., & Russell, B. (1913). Principia Mathematica. *Cambridge MA: Cambridge University*.

- O'Neill, E., Kuo, L. S., Krisko, J. F., Tomchick, D. R., Garcia, J. V., & Foster, J. L. (2006). Dynamic evolution of the human immunodeficiency virus type 1 pathogenic factor, Nef. *Journal of virology*, *80*(3), 1311-1320.

- Ostroumov, D., Fekete-Drimusz, N., Saborowski, M., Kühnel, F., & Woller, N. (2018). CD4 and CD8 T lymphocyte interplay in controlling tumor growth. *Cellular and molecular life sciences*, *75*(4), 689-713.

- Paul, S., Weiskopf, D., Angelo, M. A., Sidney, J., Peters, B., & Sette, A. (2013). HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *The Journal of Immunology*, *191*(12), 5831-5839.

- Peng, M., Mo, Y., Wang, Y., Wu, P., Zhang, Y., Xiong, F., ... & Zeng, Z. (2019). Neoantigen vaccine: an emerging tumor immunotherapy. *Molecular cancer*, *18*(1), 1-14.

- Poland, G. A., Ovsyannikova, I. G., Crooke, S. N., & Kennedy, R. B. (2020, October). SARS-CoV-2 vaccine development: current status. In *Mayo Clinic Proceedings* (Vol. 95, No. 10, pp. 2172-2188). Elsevier.

- Rasmussen, M., Fenoy, E., Harndahl, M., Kristensen, A. B., Nielsen, I. K., Nielsen, M., & Buus, S. (2016). Pan-specific prediction of peptide–MHC class I complex stability, a correlate of T cell immunogenicity. *The Journal of Immunology*, *197*(4), 1517-1524.

- Sahin, U., Derhovanessian, E., Miller, M., Kloke, B. P., Simon, P., Löwer, M., ... & Türeci, Ö. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, *547*(7662), 222-226.

- Samrat, S. K., Tharappel, A. M., Li, Z., & Li, H. (2020). Prospect of SARS-CoV-2 spike protein: Potential role in vaccine and therapeutic development. *Virus research*, *288*, 198141.

- Saskena, J. P. (1970). Mathematical model of scheduling clients through welfare agencies. *Journal of the Canadian Operational Research Society*, *8*, 185-200.

- Schülke, S. (2018). Induction of interleukin-10 producing dendritic cells as a tool to suppress allergen-specific T helper 2 responses. *Frontiers in immunology*, *9*, 455.

- Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., ... & Sidney, J. (1994). The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *The Journal of Immunology*, *153*(12), 5586-5592.

- Sikora, M., von Bülow, S., Blanc, F. E., Gecht, M., Covino, R., & Hummer, G. (2021). Computational epitope map of SARS-CoV-2 spike protein. *PLoS computational biology*, *17*(4), e1008790.

- Singh, P., Yadav, G. P., Gupta, S., Tripathi, A. K., Ramachandran, R., & Tripathi, R. K. (2011). A novel dimer-tetramer transition captured by the crystal structure of the HIV-1 Nef. *PloS one*, *6*(11).

- Soria-Guerra, R. E., Nieto-Gomez, R., Govea-Alonso, D. O., & Rosales-Mendoza, S. (2015). An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *Journal of biomedical informatics*, *53*, 405-414.

- Srivastava, S. S., Kumar, S., Garg, R. C., & Sen, P. (1969). Generalized traveling salesman problem through n sets of nodes. *CORS journal*, *7*(2), 97.

- Stratton, M. R. (2011). Exploring the genomes of cancer cells: progress and promise. *Science*, *331*(6024), 1553-1558.

- Tanyi, J. L., Bobisse, S., Ophir, E., Tuyaerts, S., Roberti, A., Genolet, R., ... & Kandalaft, L. E. (2018). Personalized cancer vaccine effectively mobilizes antitumor T cell immunity in ovarian cancer. *Science translational medicine*, *10*(436), eaao5931.

- Thompson, J. A. (2002). The revised American Joint Committee on Cancer staging system for melanoma. In *Seminars in oncology* (Vol. 29, No. 4, pp. 361-369). WB Saunders.

- Tong, J. C., & Ren, E. C. (2009). Immunoinformatics: current trends and future directions. *Drug discovery today*, *14*(13-14), 684-689.

- Tough, DF. (2008). Cytokines produced by dendritic cells. *Handbook of dendritic cells*. 355-383

- Tucci, M., Passarelli, A., Mannavola, F., Felici, C., Stucci, L. S., Cives, M., & Silvestris, F. (2019). Immune system evasion as hallmark of melanoma progression: the role of dendritic cells. *Frontiers in oncology*, *9*, 1148.

- VaxiJen database. http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html

- Vermaelen, K. (2019). Vaccine strategies to improve anti-cancer cellular immune responses. *Frontiers in immunology*, *10*, 8.

- Vormehr, M., Türeci, Ö., & Sahin, U. (2019). Harnessing tumor mutations for truly individualized cancer vaccines. *Annual Review of Medicine*, *70*, 395-407.

- Vormehr, M., Diken, M., Türeci, Ö., Sahin, U., & Kreiter, S. (2020). Personalized neo-epitope vaccines for cancer treatment. In *Current Immunotherapeutic Strategies in Cancer* (pp. 153-167). Springer, Cham.

- Wajant, H. (2009). The role of TNF in cancer. *Death Receptors and Cognate Ligands in Cancer*, 1-15.

- Wang, F., Kream, R. M., & Stefano, G. B. (2020). An evidence-based perspective on mRNA-SARS-CoV-2 vaccine development. *Medical science monitor: international medical journal of experimental and clinical research*, *26*, e924700-1.

- Yi, C., Yi, Y., & Li, J. (2020). mRNA vaccines: possible tools to combat SARS-CoV-2. *Virologica Sinica*, *35*(3), 259-262.

- Zhang, J., Zeng, H., Gu, J., Li, H., Zheng, L., & Zou, Q. (2020). Progress and prospects on vaccine development against SARS-CoV-2. *Vaccines*, *8*(2), 153.

- Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P. E., ... & Peters, B. (2008). Immune epitope database analysis resource (IEDB-AR). *Nucleic acids research*, *36*(suppl_2), W513-W518.

- Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, *9*(1), 40.

Conclusiones

A pesar de que la sinergia entre las matemáticas y las ciencias biomédicas ha crecido durante los siglos, hoy en día sigue habiendo barreras que las separan y las mantienen como dos áreas independientes. En esta tesis, se ha buscado avanzar en el camino de la unificación de estos campos, profundizando en un campo particular, a saber, el diseño de vacunas a través de técnicas de optimización. Las conclusiones derivadas de las investigaciones previamente presentadas se recogen a continuación. Además, al final de este documento, se expone una sección con posibles líneas futuras de investigación.

C.1. λ-supercadenas ponderadas en el diseño computacional de vacunas

1. El criterio de λ-supercadenas ponderadas mejora el criterio de λ-supercadena, permitiendo maximizar no solo el número de antígenos, si no eligiendo los mejores en función del peso definido (inmunogenicidad, afinidad de unión, alineamiento, etc.).

2. El método propuesto acerca las técnicas de optimización combinatoria a la realidad biológica, al tener en cuenta la inmunogenicidad generada por los antígenos y el alineamiento con las proteínas originales.

3. Nuestra propuesta es capaz de conseguir prácticamente el mismo alineamiento que los métodos más utilizados para el diseño computacional de vacunas, pero puede ofrecer candidatos con mayor inmunogenicidad, y, además, añade la propiedad de ser λ-supercadenas, y, por lo tanto, de proteger contra todas las variantes consideradas.

4. La programación entera utilizada dio mejores resultados que el algoritmo genético, que a su vez mejoró los del Hill-Climbing. Sin embargo, consideramos que, para desarrollar vacunas contra virus con alta tasa de mutación, el más apropiado sería el algoritmo genético, ya que la programación entera depende mucho de la capacidad del equipo, y aumenta considerablemente el coste computacional.

C.2. Primer diseño de una vacuna contra el SARS-CoV-2 usando λ-supercadenas

1. Las λ-supercadenas son un método eficaz y prometedor de diseño de vacunas.

2. La vacuna generó una respuesta inmune proinflamatoria con alta estimulación de células involucradas en la generación de anticuerpos.

3. La vacuna es capaz de incrementar las citoquinas beneficiosas para una respuesta eficaz.

4. Las pruebas tanto en ratones como en humanos (para más información ver el trabajo completo en el anexo) fueron exitosas.

C.3. Evaluación experimental de una vacuna personalizada contra el melanoma diseñada a través de optimización combinatoria

1. La fase *in silico* llevada a cabo a través de herramientas bioinformáticas permite una selección eficiente de neoantígenos.

2. La proliferación de células T se ve aumentada al utilizar neoantígenos específicos del paciente.

3. La liberación de citocinas es mayor al utilizar neoantígenos encapsulados en nanopartículas que al utilizar neoantígenos libres, lo que conlleva una mejor activación de los linfocitos T.

4. La respuesta inmune es antígeno y paciente específica, ya que las células T no reconocieron los antígenos de control ni las nanopartículas sin péptido.

5. Las herramientas bioinformáticas y la optimización combinatoria son eficaces a la hora de diseñar vacunas antitumorales personalizadas.

C.4. Análisis de la capacidad de detección de neoantígenos a través de las herramientas bioinformáticas

1. El porcentaje de neoantígenos que superó el umbral para ser considerados con potencial de unión a las moléculas HLA-I y II fue mayor que el porcentaje de péptidos no-mutados que superaron dicho umbral.

2. De acuerdo con los resultados obtenidos por la herramienta que estima la afinidad de unión de las moléculas HLA-I, no puede concluirse que el número de antígenos no-mutados detectados sea menor que el número de los mutados.

3. De acuerdo con los resultados obtenidos por la herramienta que estima la afinidad de unión de las moléculas HLA-II, los neoantígenos se unen con mayor afinidad a las moléculas HLA-II que su versión no mutada.

Líneas futuras sobre vacunas universales contra enfermedades infecciosas

Actualmente, como continuación de estos trabajos, el concepto de λ-supercadena ponderada ha sido ampliado para poder aplicarlo diseñando combinaciones de cadenas cortas, en vez de obtener una cadena larga como candidata a vacuna, lo cual también ha dado buenos resultados experimentales, y tiene la ventaja de ser menos susceptible a que futuras variantes puedan escapar de la vacuna. Como resultados, también debemos mencionar la patente que figura en el Anexo, obtenida a partir de los conceptos mencionados. A futuro, se continuará trabajando con este concepto, acercándolo más a la realidad biológica y a las necesidades sociosanitarias.

Líneas futuras sobre vacunas personalizadas antitumorales

Los trabajos de vacunas personalizadas surgieron como una colaboración multidisciplinar en un proyecto Elkartek. Como trabajo futuro, se nos ha concedido un nuevo proyecto donde probar, con un número mayor de pacientes, la eficiencia de las vacunas personalizadas utilizando nuestro método. Además, se contempla la posibilidad de extender nuestra metodología para trabajar con pacientes de cáncer renal.

Finalmente, me gustaría terminar este trabajo con una conclusión: las matemáticas no podrán ayudar a la humanidad si no las extendemos y aplicamos a otras áreas, mientras que la biomedicina no será capaz de dar respuestas eficientes sin incluir criterios objetivos, obtenidos a través de las matemáticas. Gracias a la convergencia entre estos campos, en un futuro no tan lejano seremos capaces de mejorar los diagnósticos médicos, entender mejor las enfermedades, u optimizar los tratamientos, lo que en definitiva se traducirá en un aumento en la calidad de vida de toda la sociedad.

Anexo: trabajos publicados

Para concluir, además de las publicaciones de las investigaciones desarrolladas en esta tesis en su versión íntegra, se ha incluido, el registro y descripción de una patente para una vacuna contra el SARS-CoV-2 diseñada a partir de las técnicas cuantitativas y algoritmos basados en la metodología de las λ-supercadenas. Finalmente, como se indica en la normativa, especificamos los indicios de calidad de las revistas donde se han publicado los tres trabajos:

- *Weighted lambda superstrings applied to vaccine design* fue publicado en PloS one en 2019. De acuerdo con SCOPUS, dicho año ocupaba la posición 10 de 111 revistas del área Multidisciplinary, con un factor de impacto de 1.023 de acuerdo con el SJR.

- *First computational design using lambda-superstrings and in vivo validation of SARS-CoV-2 vaccine* fue publicado en Scientific Reports en 2022. De acuerdo con SCOPUS, en 2020, el último año para el cual hay datos, ocupaba la posición 8 de 110 revistas del área Multidisciplinary, con un factor de impacto de 1.240 de acuerdo con el SJR.

- *Analyzing the Immune Response of Neoepitopes for Personalized Vaccine Design* fue publicado en Lecture Notes in Computer Science en 2020. De acuerdo con SCOPUS, dicho año ocupaba la posición 91 de 120 revistas del área Mathematics: Theoretical Computer Science, con un factor de impacto de 0.249 de acuerdo con el SJR.

MINISTERIO
DE INDUSTRIA, ENERGÍA
Y TURISMO

Oficina Española
de Patentes y Marcas

Justificante de presentación electrónica de solicitud de patente

Este documento es un justificante de que se ha recibido una solicitud española de patente por vía electrónica utilizando la conexión segura de la O.E.P.M. De acuerdo con lo dispuesto en el art. 16.1 del Reglamento de ejecución de la Ley 24/2015 de Patentes, se han asignado a su solicitud un número de expediente y una fecha de recepción de forma automática. La fecha de presentación de la solicitud a la que se refiere el art. 24 de la Ley le será comunicada posteriormente.

Número de solicitud:	P202030467	
Fecha de recepción:	20 mayo 2020, 12:41 (CEST)	
Oficina receptora:	OEPM Madrid	
Su referencia:	P5483ES00	
Solicitante:	Universidad del País Vasco / Euskal Herriko Unibertsitatea	
Número de solicitantes:	3	
País:	ES	
Título:	COMPOSICIÓN INMUNOGÉNICA	
Documentos enviados:	Descripcion.pdf (4590 p.) Reivindicaciones.pdf (2 p.) Resumen.pdf (1 p.) Dibujos.pdf (2 p.) OLF-ARCHIVE.zip POWATT.pdf (1 p.) OTRO-1.pdf (1 p.) SEQLPDF.pdf (4 p.) SEQLTXT.txt	package-data.xml es-request.xml application-body.xml es-fee-sheet.xml feesheet.pdf request.pdf
Enviados por:	CN=CONTRERAS PEREZ YAHEL TERESA - 24414521K,SN=CONTRERAS PEREZ,givenName=YAHEL TERESA,serialNumber=IDCES-24414521K,C=ES	
Fecha y hora de recepción:	20 mayo 2020, 12:41 (CEST)	
Codificación del envío:	B8:94:64:55:33:CB:ED:89:87:B6:0E:DA:7F:0E:B4:08:C6:34:6D:94	

AVISO IMPORTANTE

Las tasas pagaderas al solicitar y durante la tramitación de una patente o un modelo de utilidad son las que se recogen en el Apartado "Tasas y precios públicos" de la página web de la OEPM (http://www.oepm.es/es/propiedad_industrial/tasas/). Consecuentemente, si recibe una comunicación informándole de la necesidad de hacer un pago por la inscripción de su patente o su modelo de utilidad en un "registro central" o en un "registro de internet" posiblemente se trate de un fraude.

La anotación en este tipo de autodenominados "registros" no despliega ningún tipo de eficacia jurídica ni tiene carácter oficial.

En estos casos le aconsejamos que se ponga en contacto con la Oficina Española de Patentes y Marcas en el correo electrónico informacion@oepm.es.

ADVERTENCIA: POR DISPOSICIÓN LEGAL LOS DATOS CONTENIDOS EN ESTA SOLICITUD PODRÁN SER PUBLICADOS EN EL BOLETÍN OFICIAL DE LA PROPIEDAD INDUSTRIAL E INSCRITOS EN EL REGISTRO DE PATENTES DE LA OEPM, SIENDO AMBAS BASES DE DATOS DE CARÁCTER PÚBLICO Y ACCESIBLES VÍA REDES MUNDIALES DE INFORMÁTICA.

Para cualquier aclaración puede contactar con la O.E.P.M.

/Madrid, Oficina Receptora/

MINISTERIO
DE INDUSTRIA, ENERGÍA
Y TURISMO

Oficina Española
de Patentes y Marcas

(1) MODALIDAD: **PATENTE DE INVENCIÓN** **MODELO DE UTILIDAD**	 [✓] [ ]
(2) FORMULARIO 5101. TIPO DE SOLICITUD: PRIMERA PRESENTACIÓN SOLICITUD DIVISIONAL CAMBIO DE MODALIDAD TRANSFORMACIÓN SOLICITUD PATENTE EUROPEA PCT: ENTRADA FASE NACIONAL	 [✓] [ ] [ ] [ ] [ ]
(3) EXP. PRINCIPAL O DE ORIGEN: MODALIDAD: N.ºSOLICITUD: FECHA SOLICITUD:	
4) LUGAR DE PRESENTACIÓN:	OEPM, Presentación Electrónica
(5-1) SOLICITANTE 1: DENOMINACIÓN SOCIAL: UNIVERSIDAD PÚBLICA NACIONALIDAD: CÓDIGO PAÍS: NIF/NIE/PASAPORTE: CNAE: PYME: DOMICILIO: LOCALIDAD: PROVINCIA: CÓDIGO POSTAL: PAÍS RESIDENCIA: CÓDIGO PAÍS: TELÉFONO: FAX: CORREO ELECTRÓNICO: EMPRENDEDOR: PERSONA DE CONTACTO: MODO DE OBTENCIÓN DEL DERECHO: INVENCIÓN LABORAL: CONTRATO: SUCESIÓN: OTROS: PORCENTAJE DE TITULARIDAD:	 Universidad del País Vasco / Euskal Herriko Unibertsitatea [✓] España ES Q4818001B C. Barrio Sarriena s/n LEIOA 48_Vizcaya 48940 España ES [ ] [✓] [ ] [ ] [ ] 050,50 %
(5-2) SOLICITANTE 2: DENOMINACIÓN SOCIAL: UNIVERSIDAD PÚBLICA NACIONALIDAD: CÓDIGO PAÍS: NIF/NIE/PASAPORTE: CNAE: PYME: DOMICILIO: LOCALIDAD: PROVINCIA:	 CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (CSIC) [ ] España ES Q2818002D C. Serrano, 117 MADRID 28_Madrid

CÓDIGO POSTAL:	28006
PAÍS RESIDENCIA:	España
CÓDIGO PAÍS:	ES
TELÉFONO:	
FAX:	
CORREO ELECTRÓNICO:	
EMPRENDEDOR:	[ ]
PERSONA DE CONTACTO:	
MODO DE OBTENCIÓN DEL DERECHO:	
INVENCIÓN LABORAL:	[✓]
CONTRATO:	[ ]
SUCESIÓN:	[ ]
OTROS:	[ ]
PORCENTAJE DE TITULARIDAD:	047,50 %
(5-3) SOLICITANTE 3:	
DENOMINACIÓN SOCIAL:	FUNDACIÓN BIOFÍSICA BIZKAIA/BIOFISIKA BIZKAIA FUNDAZIOAFUNDACIÓN BIOFÍSICA BIZKAIA/BIOFISIKA BIZKAIA FUNDAZIOA
UNIVERSIDAD PÚBLICA	[ ]
NACIONALIDAD:	España
CÓDIGO PAÍS:	ES
NIF/NIE/PASAPORTE:	G95453775
CNAE:	
PYME:	
DOMICILIO:	Barrio Sarriena
LOCALIDAD:	LEIOA
PROVINCIA:	48_Vizcaya
CÓDIGO POSTAL:	48940
PAÍS RESIDENCIA:	España
CÓDIGO PAÍS:	ES
TELÉFONO:	
FAX:	
CORREO ELECTRÓNICO:	
EMPRENDEDOR:	[ ]
PERSONA DE CONTACTO:	
MODO DE OBTENCIÓN DEL DERECHO:	
INVENCIÓN LABORAL:	[ ]
CONTRATO:	[✓]
SUCESIÓN:	[ ]
OTROS:	[ ]
PORCENTAJE DE TITULARIDAD:	002,00 %
(6-1) INVENTOR 1:	
APELLIDOS:	KNAFO FARHI
NOMBRE:	Dina Shira
NACIONALIDAD:	
CÓDIGO PAÍS:	
DOMICILIO:	C. Ibaibide 20
LOCALIDAD:	GETXO
PROVINCIA:	48_Vizcaya
CÓDIGO POSTAL:	48930
PAÍS RESIDENCIA:	España
CÓDIGO PAÍS:	ES
TELÉFONO:	
FAX:	
CORREO ELECTRÓNICO:	
EL INVENTOR RENUNCIA A SER MENCIONADO:	[ ]

(6-2) INVENTOR 2:	
APELLIDOS:	MALAINA CELADA
NOMBRE:	Iker Andoni
NACIONALIDAD:	
CÓDIGO PAÍS:	
DOMICILIO:	C. Zubilleta nº13 1ºB
LOCALIDAD:	GETXO
PROVINCIA:	48_Vizcaya
CÓDIGO POSTAL:	48991
PAÍS RESIDENCIA:	España
CÓDIGO PAÍS:	ES
TELÉFONO:	
FAX:	
CORREO ELECTRÓNICO:	
EL INVENTOR RENUNCIA A SER MENCIONADO:	[ ]
(6-3) INVENTOR 3:	
APELLIDOS:	MARTÍNEZ DE LA FUENTE MARTÍNEZ
NOMBRE:	Ildefonso
NACIONALIDAD:	
CÓDIGO PAÍS:	
DOMICILIO:	CEBAS-CSIC Calle Campus Universitario, 3ª
LOCALIDAD:	MURCIA
PROVINCIA:	30_Murcia
CÓDIGO POSTAL:	30100
PAÍS RESIDENCIA:	España
CÓDIGO PAÍS:	ES
TELÉFONO:	
FAX:	
CORREO ELECTRÓNICO:	
EL INVENTOR RENUNCIA A SER MENCIONADO:	[ ]
(7) TÍTULO DE LA INVENCIÓN:	COMPOSICIÓN INMUNOGÉNICA
(8) NÚMERO DE INFORME TECNOLÓGICO DE PATENTES (ITP):	
(9) SOLICITA LA INCLUSIÓN EN EL PROCEDIMIENTO ACELERADO DE CONCESIÓN	
SI	[ ]
NO	[✓]
(10) EFECTUADO DEPÓSITO DE MATERÍA BIOLÓGICA:	
SI	[ ]
NO	[✓]
(11) DEPÓSITO:	
REFERENCIA DE IDENTIFICACIÓN:	
INSTITUCIÓN DE DEPÓSITO:	
NÚMERO DE DEPÓSITO:	
ORÍGEN BIOLÓGICO:	
(12) RECURSO GENÉTICO:	
NÚMERO DE REGISTRO:	
NÚMERO DE CERTIFICADO DE ACCESO AL RECURSO:	
UTILIZACIÓN DEL RECURSO GENÉTICO:	
CONOCIMIENTO TRADICIONAL ASOCIADO A UN RECURSO GENÉTICO:	
(13) DECLARACIONES RELATIVAS A LA LISTA DE SECUENCIAS:	
LA LISTA DE SECUENCIAS NO VA MÁS ALLÁ DEL CONTENIDO DE LA SOLICITUD	[✓]
LA LISTA DE SECUENCIAS EN FORMATO PDF Y ASCII SON IDENTICOS	[ ]
(14) EXPOSICIONES OFICIALES:	
NOMBRE:	
LUGAR:	
FECHA:	
(15) DECLARACIONES DE PRIORIDAD:	
PAÍS DE ORIGEN:	
CÓDIGO PAÍS:	

NÚMERO: FECHA:	
(16) REMISIÓN A UNA SOLICITUD ANTERIOR: PAÍS DE ORIGEN: CÓDIGO PAÍS: NÚMERO: FECHA:	
(17) AGENTE DE PROPIEDAD INDUSTRIAL: APELLIDOS: NOMBRE: CÓDIGO DE AGENTE: NÚMERO DE PODER:	 CONTRERAS PÉREZ Yahel 1062/7
(18) DIRECCIÓN A EFECTOS DE COMUNICACIONES: DIRECCIÓN ASOCIADA AL PRIMER SOLICITANTE DOMICILIO: LOCALIDAD: CÓDIGO POSTAL: PAÍS RESIDENCIA: CÓDIGO PAÍS: TELÉFONO: FAX: CORREO ELECTRÓNICO: MEDIO PREFERENTE DE COMUNICACIÓN	
(19) RELACIÓN DE DOCUMENTOS QUE SE ACOMPAÑAN: DESCRIPCIÓN: REIVINDICACIONES: DIBUJOS: RESUMEN: FIGURA(S) A PUBLICAR CON EL RESUMEN: ARCHIVO DE PRECONVERSION: DOCUMENTO DE REPRESENTACIÓN: LISTA DE SECUENCIAS PDF: ARCHIVO PARA LA BUSQUEDA DE LS: OTROS (Aparecerán detallados): -OTRO1.pdf Poder especifico solicitante	 [✓] N.º de páginas: 4590 [✓] N.º reivindicaciones: 15 [✓] N.º de dibujos: 2 [✓] N.º de páginas: 1 [✓] N.º de figura(s): [✓] [✓] N.º de páginas: 1 [✓] N.º de páginas: 4 [✓] [✓] N.º de páginas: 1
(20) EL SOLICITANTE SE ACOGE A LA REDUCCIÓN DE TASAS PARA EMPRENDEDORES PREVISTA EN EL ART. 186 DE LA LEY 24/2015 DE PATENTES Y, A TAL EFECTO, APORTA LA SIGUIENTE DOCUMENTACIÓN ADJUNTA:	[ ]
(21) NOTAS:	
(22) FIRMA: FIRMA DEL SOLICITANTE O REPRESENTANTE: LUGAR DE FIRMA: FECHA DE FIRMA:	 CONTRERAS PEREZ YAHEL TERESA - 24414521K Barcellona 20 Mayo 2020

Composición inmunogénica

La presente invención se refiere a una composición inmunogénica que comprende un péptido seleccionado de los péptidos de SEC ID NO: 1 a 17 o cualquier combinación de los mismos, así como a vacunas que comprenden dicha composición inmunogénica y su uso para la prevención total o parcial de la infección causada por el virus SARS-CoV-2 y la enfermedad que provoca.

PLOS ONE

RESEARCH ARTICLE

Weighted lambda superstrings applied to vaccine design

Luis Martínez[1,2,3]*, **Martin Milanič**[4], **Iker Malaina**[1,2], **Carmen Álvarez**[5], **Martín-Blas Pérez**[1], **Ildefonso M. de la Fuente**[1,6]

1 Department of Mathematics, University of the Basque Country UPV/EHU, Bilbao, Spain, **2** Biocruces Bizkaia Health Research Institute, Barakaldo, Spain, **3** Basque Center for Applied Mathematics BCAM, Bilbao, Spain, **4** University of Primorska, UP IAM and UP FAMNIT, Koper, Slovenia, **5** IDIVAL Valdecilla Biomedical Research Institute, Santander, Spain, **6** Department of Nutrition, CEBAS-CSIC Institute, Murcia, Spain

* luis.martinez@ehu.eus

Citation: Martínez L, Milanič M, Malaina I, Álvarez C, Pérez M-B, M. de la Fuente I (2019) Weighted lambda superstrings applied to vaccine design. PLoS ONE 14(2): e0211714. https://doi.org/10.1371/journal.pone.0211714

Editor: Claude Loverdo, UPMC, FRANCE

Received: August 16, 2018

Accepted: January 19, 2019

Published: February 8, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files. The code for the genetic algorithm can be found in: https://zenodo.org/record/1487837.

Funding: This research was supported in part by the Basque Government, grants IT753-13 and IT974-16 and by the UPV/EHU and Basque Center of Applied Mathematics, grant US18/21. This research was also in part by the Slovenian Research Agency (I0-0035, research program P1-0285, and research projects N1-0032, J1-7051, and J1-9110). The funders had no role in study

Abstract

We generalize the notion of λ-superstrings, presented in a previous paper, to the notion of weighted λ-superstrings. This generalization entails an important improvement in the applications to vaccine designs, as it allows epitopes to be weighted by their immunogenicities. Motivated by these potential applications of constructing short weighted λ-superstrings to vaccine design, we approach this problem in two ways. First, we formalize the problem as a combinatorial optimization problem (in fact, as two polynomially equivalent problems) and develop an integer programming (IP) formulation for solving it optimally. Second, we describe a model that also takes into account good pairwise alignments of the obtained superstring with the input strings, and present a genetic algorithm that solves the problem approximately. We apply both algorithms to a set of 169 strings corresponding to the Nef protein taken from patiens infected with HIV-1. In the IP-based algorithm, we take the epitopes and the estimation of the immunogenicities from databases of experimental epitopes. In the genetic algorithm we take as candidate epitopes all 9-mers present in the 169 strings and estimate their immunogenicities using a public bioinformatics tool. Finally, we used several bioinformatic tools to evaluate the properties of the candidates generated by our method, which indicated that we can score high immunogenic λ-superstrings that at the same time present similar conformations to the Nef virus proteins.

Introduction

Infectious and transmissible diseases cause deaths of millions of people every year. The best immunological measures to prevent such diseases are vaccines. Therefore, the main efforts of immunologists are focused towards improving our predictions of effective epitopes that would confer protection against pathogens [1] and towards enhancing our ability to select appropriate epitopes for inclusion in an efficient vaccine [2]. Protective immunity requires humoral or cellular immunity depending on the pathogen.

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Humoral immunity implies the production of antibodies by B cells that interact with surface or secreted toxins of pathogens. Each antibody binds to an epitope, defined as the three-dimensional structure of amino acids that can be contacted by the variable region of an antibody. There are two types of B-cell epitopes: (i) linear or continuous epitopes, which are short peptides that correspond to a fragment of a protein, and (ii) conformational epitopes, composed of amino acids not contiguous in primary sequence of the protein but brought in close proximity within the folded 3D structure. The length of these epitopes is variable, ranging from 8 to 20 amino acids [3].

Cellular immunity depends on T-cell epitopes generated in other cell types, the antigen presenting cells (or APC) that generate linear epitopes from pathogen degradation or protein synthesis. These short linear amino acids generated from intracellular degraded or synthesized proteins from the microorganisms bind to two types of major histocompatibility complexes (MHC), class I MHC that attach epitopes of 8-9-mer lengths and class II MHC that fit epitopes of 12-15-mer lengths [4]. CD4+ T cells recognize class II MHC epitopes and CD8+ T cells recognize class I MHC epitopes in APC.

Bionformatics methods that predict B-cell epitopes are based on certain correlations between some physicochemical properties of amino acids and the locations of linear B-cell epitopes with protein sequences [5]. Therefore, hydrophilicity, flexibility, turns, and solvent accessibility generated propensity scales for B-cell epitope prediction. However, propensity scale predictions have failed to predict B-cell epitopes since they are mainly based on fixed lengths and require flexibility [6].

Mapping of T-cell epitopes has been based on using complete sets of overlapping peptides or biochemical elution methods from MHC molecules. Both methods, when applied to a classical T cell-mediated pathogen as *Listeria monocytogenes* were costly, time consuming and, more importantly, failed to generate predictive rules [7], [8]. More recently, bioinformatics methods have also been applied to T-cell epitopes via their ability to bind MHC molecules [9]. However, they have not been able to predict efficient epitopes for vaccine design. Therefore, a mathematical method of epitope prediction able to be applied either to B or T-cell epitopes is important in the immunology field of vaccination. This has been highlighted in the last outbreaks of world wide infectious diseases, such as flu every year or Ebola in the most recent years.

Martínez et al. [10] introduced the notion of a λ-*superstring* along with an optimization problem associated to it, and gave an application to the computational design of vaccines. Given two sets of strings, a set of *host strings*, which models a set of instances of a protein (which in our case will be amino acid sequences of the protein for a given pathogen), and a set of *target strings*, which models a set of epitopes, a λ-superstring was defined to be a string that models a candidate vaccine containing, as substrings, at least λ target strings from each host string. This means that the vaccine covers at least λ epitopes in each patient. The associated optimization problem was to find a λ-superstring of minimum length, which means to find a candidate vaccine as short as possible. The aforementioned problem in [10] was shown to generalize both the shortest common superstring problem and the set cover problem, and in order to solve it they gave two approaches, one to find exact solutions and the other one to obtain approximate solutions. The approach giving exact solutions was based on an integer programming formulation of the problem, under the assumption that no two target strings are comparable with respect to the substring relation.

Motivated by the necessity of selecting the most effective epitopes mentioned at the beginning of this section, we give in this paper a generalization of the notion of λ-superstring and of the corresponding optimization problem, which is more biologically meaningful. We consider a weight function for the target strings, which represents the immunogenicity of each epitope.

A *weighted λ-superstring* is then defined as a string such that for every host string, the sum of the weights of all target strings covered simultaneously by the string and the host string is at least λ (i.e., the minimum of the sums of predicted immunogenicities of the epitopes in each protein variant considered is at least λ). Note that, in principle, the model allows for negative weights. On one hand, the more negative the immunogenicity of an epitope is, the less we prefer the corresponding target string to be a substring of a weighted λ-superstring. However, it could happen that a short weighted λ-superstring necessarily contains target strings representing epitopes of large positive immunogenicity (indicating that its epitopes will likely induce an immune response), which together cover a target string representing an epitope of negative immunogenicity (meaning that it is very unlikely for those covered epitopes to generate an immune response). Furthermore, in the Materials and Methods section we will present a model that also takes into account good pairwise alignments of the obtained superstring with the host strings, in which case target strings with negative weights could be essential. Therefore, we cannot simply disregard target strings with negative weights from the model.

We give two methods for obtaining short weighted λ-superstrings in the Materials and Methods Section. In the first subsection, a mathematical formulation of the problem is presented. In the second subsection, following the approach of [10], a graph theoretic formulation of the problem is given, from which an integer program is derived leading to optimal solutions to the problem of finding shortest weighted λ-superstrings. Next, in the third subsection, a genetic algorithm is introduced to obtain suboptimal solutions in the case when the integer programming approach cannot be used due to the large number of variables in the IP formulation. This algorithm, besides getting the λ-superstring criterion closer to biological reality, considers an additional objective to be optimized simultaneously, the alignment of the protein. By optimizing the alignment, we can obtain vaccine candidates that resemble the virus proteins that are recognized by the immune system, and therefore, build a pseudo-protein that will have a stable structure, recognizable by the MHC-complex. Our genetic algorithm is based on the NSGA-II algorithm [11], which is one of the most used heuristic techniques for solving multi-objective problems, which stands out due to its high speed, elitism, and non-necessity of specifying a sharing parameter for the optimization. In the Results section we give an application to the design of a weighted λ-superstring for a set of target strings corresponding to the Nef protein of HIV-1. We chose Nef because it is highly immunogenic [12] and plays an important role in HIV pathogenesis [13]. In order to evaluate the goodness of our candidate in silico, we have used several bioinformatic tools such as Blast, VaxiJen, I-Tasser and Phyre-2. In addition, we have studied the mismatch proportion, and compared our candidate to a candidate obtained by LANL's Epigraph, a consensus sequence and to one of the solutions using the unweighted algorithm from [10]. Finally, in the Discussion section, the main conclusions are presented and some future lines of research are outlined.

Materials and methods

The shortest weighted λ-superstring problem

In this subsection, we give a mathematical formulation of the problem. We first recall some notation and terminology for finite strings (that is, finite sequences) over a finite alphabet A. We denote by ϵ the empty string, and by A^* the set $A^* = \bigcup_{n=1}^{\infty} A^n \cup \{\epsilon\}$ of all finite strings over A. It is well known (and can be easily seen) that the set A^* forms a semigroup with respect to the operation + of concatenation $(s_1, \ldots, s_n) + (t_1, \ldots, t_m) = (s_1, \ldots, s_n, t_1, \ldots, t_m)$. Given a string $\mathbf{s} = (s_1, \ldots, s_n) \in A^*$, we denote by $\ell(\mathbf{s})$ the *length* of $\mathbf{s}$, that is, n. We say that a string $\mathbf{s}$ is a *substring* of another string $\mathbf{t}$, and denote this relation by $\mathbf{s} \subseteq \mathbf{t}$, if $\mathbf{t}$ can be written as $\mathbf{t} = \mathbf{u} + \mathbf{s} + \mathbf{v}$ for some strings $\mathbf{u}$ and $\mathbf{v}$ over A. We also use $\subset$ to denote the proper substring relation, that

is, $\mathbf{s} \subset \mathbf{t}$ if and only if $\mathbf{s} \subseteq \mathbf{t}$ and $\mathbf{s} \neq \mathbf{t}$. Given two strings $\mathbf{s} = (s_1, \ldots, s_n)$, $\mathbf{t} = (t_1, \ldots, t_m)$ in A^*, the *degree of overlapping* of $\mathbf{s}$ and $\mathbf{t}$ is defined as

$$ov(\mathbf{s}, \mathbf{t}) = \max\{i \in \{0, 1, \ldots, \min\{m, n\}\} \mid s_{n-i+j} = t_j \text{ for } j = 1, \ldots, i\}.$$

The operation of the *overlapping sum*$+'$ in A^* is defined by

$$(s_1, \ldots, s_n) +' (t_1, \ldots, t_m) = (s_1, \ldots, s_{n-ov(\mathbf{s},\mathbf{t})}) + (t_1, \ldots, t_m).$$

We remark that this operation is not associative.

The combinatorial approach to the design of vaccines described in [10] is based on the notions of λ-superstrings and λ-cover superstrings, which we now recall. Given two finite sets $H, T \subseteq A^*$ of *host* and *target* strings (modeling the set of instances of the chosen pathogen protein and the set of epitopes), respectively, and a positive integer λ, a λ-*superstring* for (H, T) is a string $\mathbf{v} \in A^*$ such that for every host string $\mathbf{h} \in H$, there exist at least λ strings in T that are common substrings of both $\mathbf{h}$ and $\mathbf{v}$. Similarly, given a collection $\mathcal{C}$ of finitely many finite sets of strings over A (that is, $\mathcal{C} = \{X_1, \ldots, X_n\}$ where $X_i \subseteq A^*$ for all $i \in \{1, \ldots, n\}$) and a positive integer λ, a λ-*cover superstring* for $\mathcal{C}$ is a string $\mathbf{v} \in A^*$ such that for every $X \in \mathcal{C}$, at least λ strings in X are substrings of $\mathbf{v}$.

We now generalize these notions and the corresponding optimization problems to the weighted case.

Definition 1 *Let $H, T \subseteq A^*$ be two finite sets of host and target strings, respectively, let each target string $\mathbf{t} \in T$ be equipped with a weight $w(\mathbf{t}) \in \mathbb{R}$, and let $\lambda \in \mathbb{R}$. A weighted λ-superstring for (H, T, w) is a string $\mathbf{v} \in A^*$ such that for every $\mathbf{h} \in H$, the sum of the weights of the target strings that are common substrings of both $\mathbf{h}$ and $\mathbf{v}$ is at least λ.*

More formally, denoting by $CS(\mathbf{s}, \mathbf{t})$ the set of all common substrings of two strings $\mathbf{s}$ and $\mathbf{t}$, a weighted λ-superstring for (H, T, w) is a string $\mathbf{v} \in A^*$ such that

$$\sum_{\mathbf{t} \in CS(\mathbf{h},\mathbf{v}) \cap T} w(\mathbf{t}) \geq \lambda \text{ for all } \mathbf{h} \in H.$$

Clearly, if $w(\mathbf{t}) = 1$ for all $\mathbf{t} \in T$, then a string $\mathbf{v}$ is a weighted λ-superstring for (H, T, w) if and only if $\mathbf{v}$ is a λ-superstring for (H, T).

The corresponding optimization problem (Box 1) is the following:

The restriction of the Shortest Weighted λ-Superstring problem to instances such that $w(\mathbf{t}) = 1$ for all $\mathbf{t} \in T$ is equivalent to the Shortest λ-Superstring problem defined in [10].

Definition 2 *Let $\mathcal{C}$ be a collection of finitely many finite sets of strings over A, let $T = \cup_{X \in \mathcal{C}} X$, let $w : T \to \mathbb{R}$, and let $\lambda \in \mathbb{R}$. A* weighted λ-cover superstring *for $(\mathcal{C}, w)$ is a string $\mathbf{v} \in A^*$ such that for every $X \in \mathcal{C}$, the sum of the weights $w(\mathbf{t})$ of the strings $\mathbf{t} \in X$ that are substrings of $\mathbf{v}$ is at least λ. Formally, for every $X \in \mathcal{C}$, we have $\sum_{\mathbf{t} \in X, \mathbf{t} \subseteq \mathbf{v}} w(\mathbf{t}) \geq \lambda$.*

Box 1

Shortest Weighted λ-Superstring

Instance: A finite set of $H \subseteq A^*$ of *host* strings, a finite set of $T \subseteq A^*$ of *target* strings, a weight function $w : T \to \mathbb{R}$, a *covering requirement* $\lambda \in R$.

Task: Find a weighted λ-superstring for (H, T, w) of minimum length.

Box 2

Shortest Weighted λ-Cover Superstring

Instance: A collection $\mathcal{C}$ of finitely many finite sets of finite strings over alphabet A, a weight function $w : \cup_{X\in\mathcal{C}} X \to \mathbb{R}$, a *covering requirement* $\lambda \in \mathbb{R}$.

Task: Find a weighted λ-cover superstring for $(\mathcal{C}, w)$ of minimum length.

Clearly, the case of unit weights corresponds to the notion of a λ-cover superstring. The corresponding optimization problem (Box 2) is the following:

The restriction of the Shortest Weighted λ-Cover Superstring problem to instances such that $w(\mathbf{t}) = 1$ for all $\mathbf{t} \in \cup_{X\in\mathcal{C}} X$ is equivalent to the Shortest λ-Cover Superstring problem defined in [10]. In that paper, it was proved that the Shortest λ-Superstring problem is polynomially equivalent to the Shortest λ-Cover Superstring problem. This equivalence extends straightforwardly to the weighted versions of the problems. Moreover, since the weighted versions of the problem generalize the unweighted ones, hardness results from [10] immediately carry over to the weighted ones. In particular:

Theorem 3 *1. For every* $\epsilon > 0$, *there is no polynomial time algorithm approximating the* Shortest Weighted λ-Superstring *problem within a factor of* $(1-\epsilon)\ln|H|$, *unless* P = NP, *even for the case of the binary alphabet* $A = \{0, 1\}$, *a constant weight function* $w \equiv 1$, *and* $\lambda = 1$.

2. For every $\epsilon > 0$, *there is no polynomial time algorithm approximating the* Shortest Weighted λ-Cover Sperstring *problem within a factor of* $(1-\epsilon)\ln|\mathcal{C}|$ *unless* P = NP, *even for the case of the binary alphabet, a constant weight function* $w \equiv 1$, *and* $\lambda = 1$.

The corresponding hardness results from [10] are stated with a multiplicative constant of $c > 0.2267$ instead of $1-\epsilon$. However, exactly the same approach as the one used to prove Theorem 3.9 and Corollary 3.10 in [10] can be used to derive Theorem 3; one only needs to use the more recent, stronger inapproximability result on the set cover problem due to Dinur and Steurer [14] instead of the one due to Alon et al. [15].

Theorem 3 suggests that most likely the two problems cannot be solved optimally or approximately by efficient algorithms, and motivate the development of exact exponential time algorithms and of suboptimal heuristic approaches. This is what we do in the next two subsections.

Graph theoretic and integer programming formulations of the shortest weighted λ-cover superstring problem

In this section, we extend the graph theoretic and integer programming (IP) formulations of the Shortest λ-Cover Superstring problem from [10] to the weighted case. (For background on integer programming, see, e.g., [16]). Following [10], we model the problem as a generalization of the *generalized Traveling Salesman Problem*. In this problem, the set of vertices of a given complete directed edge-weighted graph is divided into clusters and the objective is to find a minimum-cost tour passing through at least one node from each cluster.

The graph theoretic model for the Shortest λ-Cover Superstring problem from [10] is based on a derived complete edge-weighted directed graph G with vertex set $T = \cup_{X\in\mathcal{C}} X$ plus one special vertex. Roughly speaking, the main idea is the following. Given a λ-cover superstring $\mathbf{v}$ for $\mathcal{C}$, one can identify a set of substrings of $\mathbf{v}$ that are pairwise incomparable with

respect to the substring relation and contain, as substrings, at least λ strings from each cluster $X \in \mathcal{C}$. Sorting these strings in order of their first appearance in **v** yields a directed path in G that can be extended to a directed cycle in G through the special vertex. By construction, the vertices of this cycle "cover" (in the sense of substring relation, when viewed as strings) at least λ vertices from each cluster $X \in \mathcal{C}$. The weights of the edges are defined so that the length of the resulting cycle does not exceed the length of **v**. And conversely, every directed cycle in G through the special vertex satisfying the above covering property and such that no two strings corresponding to (non-special) vertices of the cycle are comparable with respect to the substring relation can be transformed into a λ-cover superstring **v**, by taking the overlapping sum of the strings corresponding to the non-special vertices of the cycle. The weights of the edges are defined so that the length of the cycle equals the length of the obtained superstring.

We now formalize these notions and explain the extension to the weighted case. Consider an instance $(\mathcal{C}, w, \lambda)$ of the Shortest Weighted λ-Cover Superstring problem, and let $T = \cup_{X \in \mathcal{C}} X$. Following [10], we construct a complete directed edge-weighted graph $G = (V, E, c)$, called the *distance graph*. To distinguish the edge weights from the weights from the input weight function w, the weights on edges will also be referred to as *costs* and will be specified with a function $c : E \to \mathbb{Z}_+$. The construction is the same as in [10]:

- $V = T \cup \{s^*\}$.
- For every two distinct vertices $s, t \in T$, add the arc (s, t) to E and assign to it the cost $c(s, t) = \ell(s) - ov(s, t)$. Clearly, the costs are well defined and non-negative.
- For every vertex $s \in T$, add the arc (s, s^*) to E and assign to it cost $c(s, s^*) = \ell(s)$.
- For every vertex $s \in T$, add the arc (s^*, s) to E and assign to it zero cost, $c(s^*, s) = 0$.

We emphasize that in what follows, we identify the vertices of G other than s^* with the corresponding strings from T. In particular, for $i, j \in V(G) \setminus \{s^*\}$, notation $i \subseteq j$ means that i is a substring of j and $i \subset j$ that i is a proper substring of j. One more definition is needed to express the problem as a graph problem. A subgraph H of G is said to *cover* a string $\mathbf{s} \in T$ if there exists a vertex $\mathbf{t} \in V(H) \cap T$ such that $\mathbf{s} \subseteq \mathbf{t}$. For $X \in \mathcal{C}$, we will denote the set of all strings in X covered by H by X_H. The *cost* of a directed cycle C in G is defined as $\sum_{e \in E(C)} c(e)$.

Definition 4 *A directed cycle C in the distance graph G is said to be* w-feasible *if it satisfies the following conditions*:

1. $s^* \in V(C)$.
2. *For every two distinct vertices* **s**, **t** *from* $V(C) \cap T$, **s** *is not a substring of* **t**.
3. *For every* $X \in \mathcal{C}$, *we have* $\sum_{\mathbf{t} \in X_C} w(\mathbf{t}) \geq \lambda$.

Proposition 5 *Let $(\mathcal{C}, w, \lambda)$ be an instance to the* Shortest Weighted λ-Cover Superstring *problem, and let G be its derived distance graph. Then, there exists a weighted λ-cover superstring for $(\mathcal{C}, w)$ of length at most ℓ if and only if G contains a w-feasible directed cycle C of cost at most ℓ.*

We give a proof of Proposition 5 in S1 Appendix.

Proposition 5 leads to the following IP formulation for the Shortest Weighted λ-Cover Superstring problem. The program has three types of binary variables: x_{ij}, where (i, j) ranges over all ordered pairs of distinct elements of V, y_i, where i ranges over all elements of V, and z_i, where i ranges over all elements of T. Recall that $c : E \to \mathbb{R}_+$ is the cost function on the edges

of the distance graph G.

$$
\begin{aligned}
\min \quad & \sum_{i,j} c(i,j)x_{ij} & \\
s.t. \quad & y_{s^*} = 1 & \\
& \sum_{i \in V\,:\, i \neq j} x_{ij} = y_j & \forall j \in V \\
& \sum_{j \in V\,:\, j \neq i} x_{ij} = y_i & \forall i \in V \\
& \sum_{i \in X} w(i) z_i \geq \lambda & \forall X \in \mathcal{C} \\
& \sum_{i \subseteq j} y_j \geq z_i & \forall i \in T \\
& y_i + y_j \leq 1 & \forall i, j \in T \text{such that } i \subset j \\
& 0 \leq x_{ij} \leq 1, & x_{ij} \text{ integer} \\
& 0 \leq y_i \leq 1, & y_i \text{ integer} \\
& 0 \leq z_i \leq 1, & z_i \text{ integer}
\end{aligned}
$$

The feasible solutions of the IP described above are in correspondence with subgraphs H of G containing s^* that consist of one or more *subtours* (vertex-disjoint directed cycles) in which the vertices other than s^* correspond to a set of strings that are pairwise incomparable with respect to the substring relation and such that the covering requirement

$$\sum_{\mathbf{t} \in X_H} w(\mathbf{t}) \geq \lambda\,.$$

is satisfied.

To be able to apply Proposition 5, we are only interested in solutions that consist of a single directed cycle. As discussed in [10], this can be achieved in several ways (see, e.g., [17]), for instance using the Miller-Tucker-Zemlin (MTZ) formulation [18], the subtour formulations, or with a combined approach resulting in a cutting-plane algorithm.

In Fig 1, we represent an ilustrative sketch linking the combinatorial optimization problem to the graph problem.

A genetic algorithm

In this section we will present a genetic algorithm well suited to find solutions to a problem with potential applications to vaccine design posed, for unweighted λ-superstrings, in the concluding section of [10]. The problem is the following: Given a set of host strings of approximately similar lengths corresponding to the same protein with different mutations in a set of patients, find a λ-superstring of about one-gene length with λ as big as possible when the set T of target strings is formed by all the substrings of a given length ℓ of the set of host strings, while keeping, as much as possible, the relative order of the elements in T. In other words, the goal is to design a synthetic protein enriched in the sense that it covers many epitopes in each host string. In our more general setting of weighted λ-superstrings we require these epitopes to be very immunogenic. As the second objective of our multi-objective optimization program, we have chosen to optimize the amino acid resemblance with the virus peptides. By using the alignment as target to be optimized, we will be able to choose candidates that have a structure

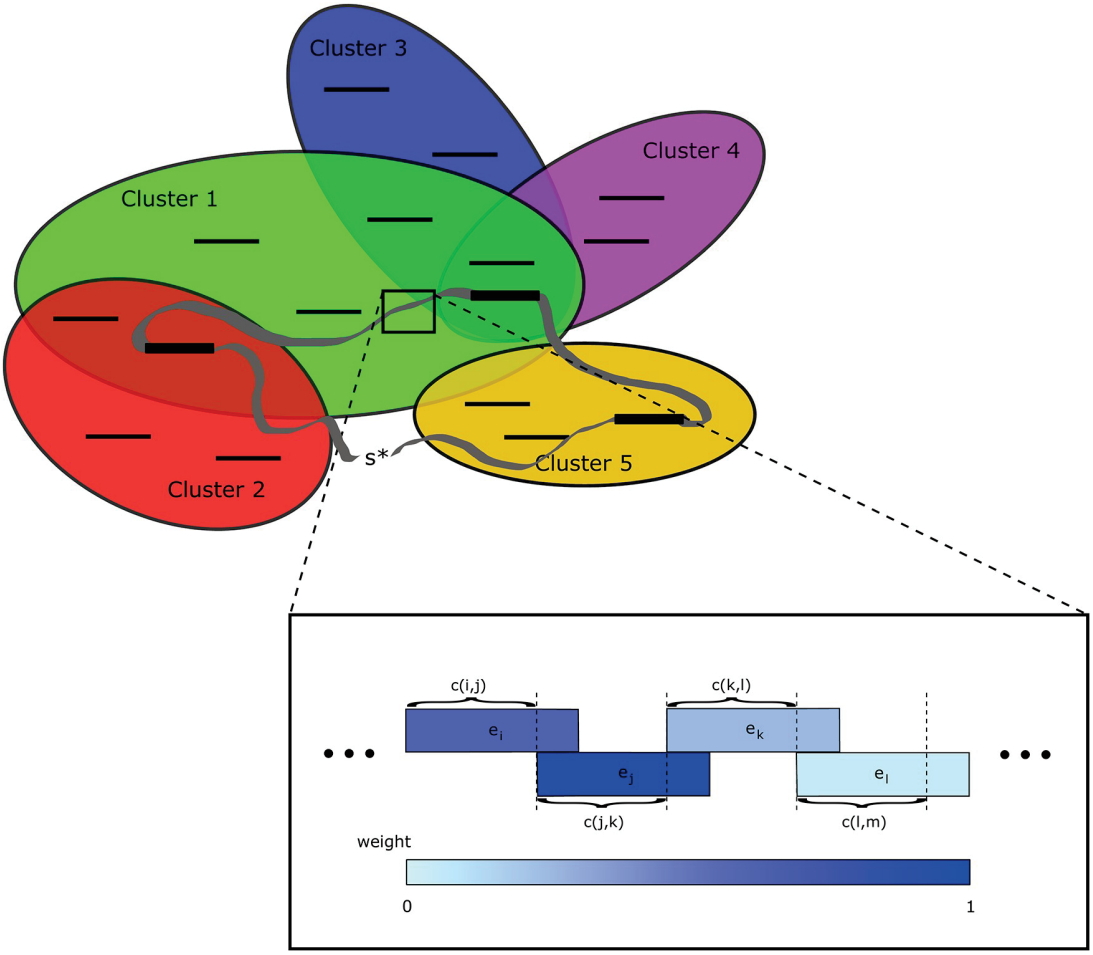

Fig 1. Graphical interpretation of the connection of the combinatorial optimization problem to the graph problem. The clusters associated to the host strings are shown in ovals with the corresponding target strings inside them. Each target string has an associated weight, which is shown in this example using a color code from light blue to strong blue, with extreme values corresponding to 0 and 1, respectively. The λ-superstring is represented with a closed ribbon which travels among the clusters. It is closed because one of the strings forming it corresponds to the artificial vertex s^*, which is not a host string, but can be viewed as an empty string gluing the extremes of the λ-superstring. The condition that for each one the clusters, the sum of the weights of the target strings that are both in the λ-superstring and in the cluster is at least λ is imposed in the feasible solutions. The length of the λ-superstring is minimized, and this length can be obtained by summing up the $c(i, j)$ values of the strings forming the λ-superstring. The $c(i, j)$ values are shown in the figure as the length of the part of the vertex labelled by i not overlapping with the next vertex in the λ-superstring, which is labelled by j.

https://doi.org/10.1371/journal.pone.0211714.g001

similar to those which already interacted with HIV patients, and therefore will likely be recognized by the immune system.

We opt for a genetic algorithm in this case because the high number of target strings makes the use of integer programming impractical; employing heuristic methods of optimization is thus a good alternative. We do sacrifice on optimality; nevertheless, suboptimal solutions can be satisfactory in practice.

We are faced with a multi-objective optimization problem. To solve such problems, multi-objective functions f: $P \rightarrow R^n$ are considered, where P is the set of feasible solutions, that assign to each element $x \in P$ an n-tuple $(f_1(x), \ldots, f_n(x))$ with real entries, each of which indicates a partial objective function. Without loss of generality, we can assume that we want to *maximize*

each partial objective function, because minimizing $f_i(x)$ is equivalent to maximizing the opposite function $-f_i(x)$. Obviously, it is not possible in general to get a solution $x \in P$ in which all partial objective functions f_i attain maximum value. Instead, optimality of a solution is established in terms of *Pareto domination*: given two feasible solutions $x, y \in P$, we say that $x = (x_1, \ldots, x_n)$ is dominated by $y = (y_1, \ldots, y_n)$ if $x_i \leq y_i$ for every i and $x_j < y_j$ for some j. The *Pareto front* is formed by the elements in P which are not dominated by any element of P.

Very often evolutionary algorithms are used to evolve an initial population $P_0 \subseteq P$ to obtain a sequence P_i of populations which get closer to the Pareto front, and it is desirable to obtain wide-spread sets of solutions. In particular, several genetic algorithm approaches have been proposed for these kinds of problems. One of the most reliable and quick ones among them is NSGA-II [11], and we have used it for our optimization problem. For definitions and results on genetic algorithms we refer the reader to [19].

We outline here the structure of the NSGA-II algorithm. We refer to [11] for details.

Given a set $P' \subseteq P$ of feasible solutions, two key values are assigned to each $x \in P'$: the non-domination rank x_{rank} and the crowding distance x_{distance}. The process of assignment of non-domination ranks is as follows. The non-dominated elements, that is, the elements in the Pareto front of P' are assigned rank 1, and they form the set F_1. If we take $P' - F_1$, the non-dominated elements in this set are assigned rank 2, and they form the set F_2, and so on. This ordering is done using the fast non-dominated sorting described in [11]. The crowded distance x_{distance} is calculated by taking the average distance of two points on either side of x along each of the n objectives. This leads to a strict partial order on P' defined by

$$x \prec y \text{ if } x_{\text{rank}} < y_{\text{rank}} \text{ or if } x_{\text{rank}} = y_{\text{rank}} \text{ and } x_{\text{distance}} > y_{\text{distance}}.$$

The general process in NSGA-II is as follows:

First, given a parameter m, a random population P_0 of size m is constructed, and it is sorted according to the relation $\prec$ defined above. Then, a binary tournament selection is done considering the relation $\prec$. In the tournament selection it is theoretically possible, although it is unlikely, that two different elements are not comparable with respect to the relation, because they have the same rank and the same crowded distance. In this case, one of them is chosen uniformly at random. After the tournament selection is completed, mutation and crossing is done on the selected elements, to create an offspring population Q_0 of size m. Now a combined population $R_0 = P_0 \cup Q_0$ is formed, and the elements in R_0 are sorted according to their domination level. Then, a new population P_1 is formed by collecting the elements in R_0 in ascending order of ranks, that is, we take the elements in the set F_1 formed by the elements of rank 1, then the elements in F_2, and so on, until all the elements of a certain set F_{i-1} have been allocated but there is no place to allocate all the elements of F_i, that is, until $|F_1 \cup \cdots \cup F_{i-1}| \leq m$ but $|F_1 \cup \cdots \cup F_i| > m$. Then, we rank the elements in F_i according to its crowding distance and we select elements in non-increasing order of crowding distance until we have m elements in P_1. Now, given a parameter *niter*, the process is iterated *niter* times to obtain a population P_{i+1} from a population P_i for any i in the same way that we obtained P_1 from P_0.

Next we will describe how we use NSGA-II for our particular problem.

We want to find a weighted λ-superstring for a set $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_{spop}\}$ of host strings, a set T of target strings formed by all the subsequences of a given length ℓ of the strings of H, and a weight mapping w assigning real values to elements of T. The chromosomes in the genetic algorithm will be sequences of target strings. The phenotype of a chromosome u will be the overlapping sum $o(u)$ of the target strings which constitute it (according to the sequence in which they appear in u). The fitness function that we consider for each chromosome u in the population is taken to be $f(u) = (\lambda(u), al(u))$, where:

- $\lambda(u)$ is an estimate of the maximum value for which $o(u)$ is a weighted $\lambda(u)$-superstring for (H, T), defined by

$$\lambda(u) = \min\left\{\sum_{t\in u, t \text{ substring of } \mathbf{h}_i} w(t) : i = 1, \ldots, spop\right\}$$

  (it is an estimate because the true value of the maximum λ could, in principle, be different from $\lambda(u)$ if there are elements of T covered by $o(u)$ which are not in u), and

- $al(u)$ is the average value of the scorings for the pairwise global alignments of $o(u)$ and each of the strings $\mathbf{h}_i$.

The specific scoring scheme may depend on the application; in the Results section we specify it for our particular biological application. (For background on string alignment, see [20]).

We have used a modified version of NSGA-II in which we take the Q_i sets of a cardinality m greater than *spop*, so that $|R_i| > 2|P_i|$ for every i. Also, instead of taking the initial population P_0 randomly, we have taken it to be formed by the sequences of target strings corresponding to the set $\{\mathbf{h}_1, \ldots, \mathbf{h}_{spop}\}$ of host strings, in the order of appearance in each host string.

For the crossing of two chromosomes $(u_1, \ldots, u_{\ell_1})$ and $(v_1, \ldots, v_{\ell_2})$, we have used a one-point crossing in which we select randomly a crossing point c between 1 and $\min\{\ell_1, \ell_2\} - 1$ and take $(u_1, \ldots, u_c, v_{c+1}, \ldots, v_{\ell_2})$ as the first child and $(v_1, \ldots, v_c, u_{c+1}, \ldots, u_{\ell_1})$ as the second child.

Once the crossing has been done, we have assigned a probability of mutation *prmut* in each gene of each child chromosome. A mutation in the i-th position of a chromosome $u = (u_1, \ldots, u_{\ell_1})$ is done by selecting first a random integer j obtained by rounding a real number sampled according to the normal distribution with mean i and standard deviation defined by a parameter *sd*, choosing then uniformly at random a sequence $(v_1, \ldots, v_{\ell_2})$ associated to a host string from the initial population and substituting u_i with v_j in the chromosome u if $1 \leq j \leq \ell_2$; in any other case, the mutation is not done. The idea of this mutation that we have just described is to substitute the u_i with an element 'not far from the i-th position', in the sense that it is close to an element in the i-th position in a chromosome of the initial population formed by the host strings.

Results

In this section, an application of the IP-based algorithm and of the genetic algorithm is given to find weighted λ-superstrings for a set of 169 host strings whose GenBank [21] access numbers appear in S1 Table, corresponding to the Nef protein, and two sets of target strings (epitopes) chosen in a way that will be made clear soon. The 169 sequences were from HIV-1 subtype B independently infected individuals, and this specific set was first considered by Nickle et al. in [22], and later by our group in [10]. Thus, we used this same set in order to be able to compare the method here proposed, to our previous work [10]. This comparison can be found at the end of this section.

Applying the integer programming formulation

We begin with the IP-based algorithm described in the Materials and methods section. We consider the set of epitopes shown in S2 Table.

The weights corresponding to the immunogenicities of epitopes were experimentally obtained from the data appearing in the Immune Epitope Database and Analysis Resource

(IEDB) [23]. We selected the epitopes for the Nef protein satisfying simultaneously the following three conditions:

1. they are covered by at least one of the 169 host strings analyzed;
2. they appear in the HIV Molecular Immunology Database [24];
3. they appear in IEDB with a positive value of $p + n$, where p and n are the number of positive and negative results, respectively, in the MHC Ligand Assays section.

We took the ratio $p/(p + n)$ as the weighting of the epitopes. Note that a non-linear rescaling of the weights (i.e., normalizing them) would change the optimization problem. However, we consider that to justify a rescaling we would require empirical evidence pointing that the candidates give better results, and that is out of the scope of this work. The main reason for considering this weighting is that the empirical response of an epitope can only be verified through assays, so we estimated it numerically by the aforementioned ratio. Moreover, we used the MHC Ligand Assays, because there are several works stating that there exists a correlation between the generated immune response and MHC complex stability [25] or MHC affinity [26], and it has been used to predict T Cell epitopes [27]. The values are also shown in S2 Table.

The solutions found with the IP-based algorithm and the values of the corresponding parameters are shown in Tables 1 and 2, which we now explain.

In the analysis whose results are shown in Table 1, the value of λ was varied from 1.0 up to 3.3 in increments of 0.1, and for each value of λ, the total length of the λ-superstring was minimized. Solutions were obtained by implementing the integer program descrived in Materials and Methods (extended with the MTZ formulation) in Java [28] and solving it to optimality using IBM ILOG CPLEX Optimization Studio [29]. The integer program corresponding to the case λ = 3.3 turned out to be infeasible; all the others were feasible. In the table we also show the *covering value* of the obtained solution, that is, the value of $\min_{X \in \mathcal{C}} \sum_{i \in X} w(i) z_i$ (using notation from the Materials and methods section). Only the results not dominated by others are shown, in the sense that in cases when for different values of λ the same optimal solution strings were found, only the highest value of λ is shown.

Table 2 shows the results of a "dual" analysis in which we were maximizing the value of λ subject to imposing an upper bound on the length of a λ-superstring for the given sets of host and target strings. The results were obtained by solving a straightforward modification of integer program (and its extension with the MTZ formulation), again using Java and CPLEX. The modification of the IP consists in treating λ as a variable, replacing the objective function $\Sigma_{i,j}$ $c(i, j)x_{ij}$ with λ and min with max, and adding the constraint $\Sigma_{i,j}\, c(i, j)x_{ij} \leq \ell$, where ℓ is a given upper bound on the string length. Clearly, since we are maximizing λ, in any optimal solution the value of λ will be equal to the covering value, that is, $\lambda = \min_{X \in \mathcal{C}} \sum_{i \in X} w(i) z_i$ (again, using notation from the Materials and methods section).

The upper bound ℓ on the length of the λ-superstring was varied from 10 to 200 in increments of 10. Increasing the upper bound on the string length from 100 to anywhere up to 200 did not result in any increase in the covering value λ. We therefore only display in Table 2 the results for the values of the upper bounds up to 100. Since in this second model the length of the obtained solution was only constrained by an upper bound and not taken into account in the objective function, it should not be surprising that the corresponding solutions found for upper bounds between 100 and 200 were of different lengths, despite the fact of being equally good in terms of their covering values. A similar phenomenon occurred also for values of the upper bound ℓ displayed in the table: the optimal covering values of the solutions corresponding to the upper bounds in each of the ranges 10–20 and 70–90 were the same.

Table 1. Optimal solutions of minimum length for a given value of λ.

λ = 1.0
Optimal λ superstring: TQGYFPDWQNYVPLRPMTYPLTFGWCF
Optimal λ superstring length: 27
Covering value of the solution: 1
λ = 1.5
Optimal λ superstring: LTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLK
Optimal λ superstring length: 35
Covering value of the solution: 1.51
λ = 1.9
Optimal λ superstring: KAAVDLSHFLTFGWCFKLVFPVRPQVPLRPMTYTQGYFPDWQNY
Optimal λ superstring length: 44
Covering value of the solution: 1.94
λ = 2
Optimal λ superstring: KAAVDLSHFLKLTFGWCFKLVFPVRPQVPLRPMTYTQGYFPDWQNY
Optimal λ superstring length: 46
Covering value of the solution: 2
λ = 2.5
Optimal λ superstring: TQGYFPDWQNYPLTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLK
Optimal λ superstring length: 47
Covering value of the solution: 2.51
λ = 2.6
Optimal λ superstring: FPVRPQVPLRPMTYKAAVDLSHFLKEKGGLTQGYFPDWQNYTPGPGVRYPLTFGWCFKLV
Optimal λ superstring length: 60
Covering value of the solution: 2.68
λ = 2.9
Optimal λ superstring: TPGPGVRYPLFPVRPQVPLRPMTYKAAVDLSHFLKTPGPGIRYPLTFGWCFKLVTQGYFPDWQNY
Optimal λ superstring length: 65
Covering value of the solution: 2.94
λ = 3.2
Optimal λ superstring: TPGPGIRYPLTPGPGVRYPLTFGWCFKLVPEKEVLVWKFDSRLAFHHQEILDLWVYFPVRPQVPLRPMTYKAAVDLSHFLKEKGGLEGLTQGYFPDWQNY
Optimal λ superstring length: 100
Covering value of the solution: 3.25

https://doi.org/10.1371/journal.pone.0211714.t001

We are interested in high covering values while keeping the length of the λ-superstring small. It is therefore interesting to analyze which of the solutions found by the above analysis have the best (that is, highest) ratio between the covering value and the length. In this respect, the best solution found by the above analysis is the λ-superstring of length 47 achieving a covering value of 2.51 (see Table 1). The same covering value is also achieved by the string of length 47 shown in Table 2. Only slightly worse ratios were achieved by the solutions from the above tables corresponing to the following (length, covering value) pairs: (44, 1.94), (60, 2.68), (65, 2.94) (all from Table 1).

Another aspect of such analysis that might be potentially interesting for vaccine design applications would be to identify the maximum possible covering value that can be achieved for a given set of host and target strings (without any restriction on the length of the λ-superstring), and then find a shortest substring realizing this covering value. In the instance analyzed above, this maximum covering value is equal to 3.25, and the shortest length of aλ-superstring achieving this covering value is 100.

Table 2. Optimal solutions with maximum λ for a given upper bound on the length of the string.

Upper bound on string length = 10
Optimal value of λ = 0.0
Optimal λ superstring: AVDLSHFL
Optimal λ superstring length: 8
Upper bound on string length = 20
Optimal value of λ = 0.0
Optimal λ superstring: AVDLSHFL
Optimal λ superstring length: 8
Upper bound on string length = 30
Optimal value of λ = 1.0
Optimal λ superstring: TQGYFPDWQNYPLTFGWCFQVPLRPMTYK
Optimal λ superstring length: 29
Upper bound on string length = 40
Optimal value of λ = 1.51
Optimal λ superstring: LTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLKEKGGL
Optimal λ superstring length: 40
Upper bound on string length = 50
Optimal value of λ = 2.51
Optimal λ superstring: TQGYFPDWQNYPLTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLK
Optimal λ superstring length: 47
Upper bound on string length = 60
Optimal value of λ = 2.68
Optimal λ superstring: TQGYFPDWQNYTPGPGVRYPLTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLKEKGGL
Optimal λ superstring length: 60
Upper bound on string length = 70
Optimal value of λ = 2.94
Optimal λ superstring: TPGPGIRYPLTQGYFPDWQNYTPGPGVRYPLTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLKEKGGL
Optimal λ superstring length: 70
Upper bound on string length = 80
Optimal value of λ = 2.94
Optimal λ superstring: TPGPGIRYPLTQGYFPDWQNYTPGPGVRYPLTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLKEKGGL
Optimal λ superstring length: 70
Upper bound on string length = 90
Optimal value of λ = 2.94
Optimal λ superstring: TPGPGIRYPLTQGYFPDWQNYTPGPGVRYPLTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLKEKGGL
Optimal λ superstring length: 70
Upper bound on string length = 100
Optimal value of λ = 3.25
Optimal λ superstring: QEILDLWVYTQGYFPDWQNYTPGPGIRYPLPEKEVLVWKFDSRLAFHHTPGPGVRYPLTFGWCFKLVFPVRPQVPLRPMTYKAAVDLSHFLKEKGGLEGL
Optimal λ superstring length: 100

https://doi.org/10.1371/journal.pone.0211714.t002

Applying the multiobjective genetic algorithm

We used the NSGA-II multiobjective genetic algorithm described in the Materials and methods section for the same set of 169 host strings used in the previous subsection whose GenBank IDs appear in S1 Table. The set of target strings was taken to be the set of all 9-mers present in the host strings. Unlike in the previous subsection, immunogenicities were not obtained experimentally, because of the technical difficulty and the high cost of estimating empirically the

immunogenicity of a large number of sequences. In this case, the immunogenicity associated to each of the target strings (that is, the value of the weight function $w(\mathbf{t})$) was computationally assessed.

Several algorithms to estimate numerically the immunogenicity of epitopes have been proposed in the literature, see, for instance, [30–39]. We selected in our analysis the algorithm proposed in [34], where a tool was also given in the "T-cell" epitopes-Immunogenicity Prediction" of the "IEDB Analysis Resource" [40].

We ran the genetic algorithm by using the program Mathematica [41] with the following set of parameters:

$$niter = 500, spop = 169, prmut = 0.01, m = 1352 \text{ and } sd = 1.$$

We used the Mathematica command `NeedlemanWunschSimilarity`, which gives the number of one-element matches in the alignment, for calculating the scorings of the global alignments that are averaged to obtain the values of $al(u)$ described in the Materials and methods section.

We run 20 times the NSGA-II algorithm and collected the non-dominated solutions obtained in each of the runs. We eliminated the dominated solutions to obtain a final estimation of the Pareto front. The values are shown in Fig 2 and in Table 3. The resultant estimation of the Pareto front gave a set of non-dominated sequences with a maximum λ of 5.71 and a minimum value of 1.2 (average ± SD of 4.32±1.14). The alignments ranged between -88.47 and 163.33 (average± SD of 87.73±67.43). The distributions of λ and the alignment values are represented in S1 Fig panel (a) and (b), respectively.

We selected and analyzed in the estimation of the Pareto front the solution with scoring value 161.93 and λ value 2.1794. We have chosen this sequence due to several reasons. First, theλ and the scoring are greater than the ones of all the members in the initial population of 169 strings, for which the mean of the λ values was -1.70395, the maximum λ value was 1.59422, the mean of the scores was 143.34 and the maximum score was 157.66. Second, there

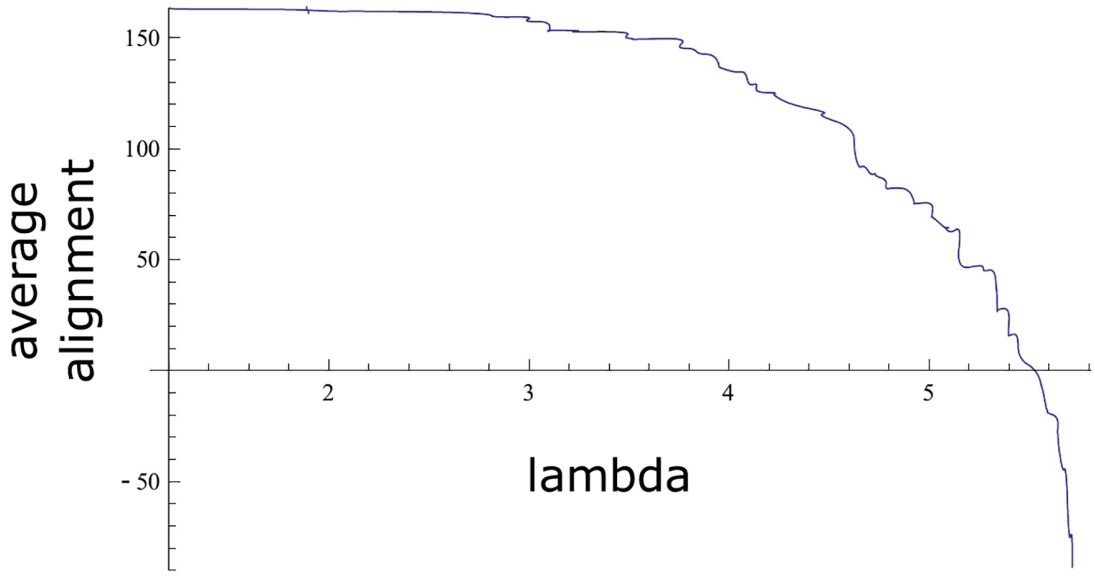

Fig 2. Estimation of the Pareto front in the genetic algorithm. The line represents the non-dominated solutions found with the genetic algorithm. The X axis indicates the λ value, while the Y axis indicates the alignment score.

https://doi.org/10.1371/journal.pone.0211714.g002

Table 3. Numerical values for the estimation of the Pareto front in the genetic algorithm.

(1.2017,163.33)	(3.7513,149.3)	(4.6507,92.16)	(5.3242,44.29)
(1.293,162.95)	(3.7547,145.87)	(4.6771,91.94)	(5.3374,36.46)
(1.7287,162.79)	(3.8153,144.98)	(4.7089,88.77)	(5.3391,36.19)
(1.8909,162.49)	(3.855,142.96)	(4.7308,88.71)	(5.3413,27.89)
(1.8977,162.44)	(3.9156,142.07)	(4.7392,87.72)	(5.3492,27.59)
(2.0255,161.96)	(3.9461,138.79)	(4.7815,85.76)	(5.3959,26.6)
(2.1794,161.93)	(3.9526,136.87)	(4.787,82.31)	(5.3974,17.13)
(2.5507,161.58)	(3.9797,135.85)	(4.8195,82.26)	(5.4096,16.14)
(2.7806,160.63)	(4.0195,134.82)	(4.8925,81.32)	(5.4404,15.24)
(2.8411,159.48)	(4.0502,134.62)	(4.9253,76.17)	(5.4582,6.15)
(2.9918,159.25)	(4.08,133.92)	(4.9355,75.31)	(5.5322,-0.18)
(2.9923,157.56)	(4.0977,129.8)	(5.0094,75.04)	(5.57,-9.33)
(3.0863,156.8)	(4.1161,128.93)	(5.0147,69.88)	(5.585,-16.12)
(3.1024,153.34)	(4.1377,128.84)	(5.0249,68.67)	(5.5929,-18.14)
(3.1083,153.31)	(4.1458,125.92)	(5.078,64.63)	(5.5988,-19.13)
(3.2322,153.22)	(4.2259,125.13)	(5.0834,64.36)	(5.6403,-21.67)
(3.2357,152.79)	(4.2286,124.14)	(5.1092,62.62)	(5.6454,-30.74)
(3.2449,152.69)	(4.3204,120)	(5.1528,62.41)	(5.6671,-43.62)
(3.4688,152.4)	(4.4694,116.65)	(5.1562,48.33)	(5.6859,-47.61)
(3.4855,150.26)	(4.47,114.67)	(5.2502,47.24)	(5.6998,-72.72)
(3.5152,149.62)	(4.602,108.9)	(5.2719,46.02)	(5.7141,-74.6)
(3.546,149.37)	(4.6296,98.75)	(5.2798,45.08)	(5.717,-88.47)

https://doi.org/10.1371/journal.pone.0211714.t003

is a remarkable level of maintenance of the highly conserved regions of the protein for this solution. Nonetheless, other solutions in the estimation of the Pareto front with greater values of λ and lower scorings could, of course, be useful in practice.

The sequence of amino acids of the selected solution is

MGGKWSKRSGVGWPTVRERMRRAEPAADGVGAV

SRDLEKHGAITSSNTAATNADCAWLEAQEEEEVGF

PVRPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIYS

QKRQDILDLWIYHTQGYFPDWQNYTPGPGIRYPLT

FGWCFKLVPVEPEKVEEANEGENNSLLHPMSLHG

MEDPEKEVLEWKFDSRLAFHHMARELHPEYYKDC.

Since the main goal in this section is to study the structure and functionality of a protein modelled by a sequence with given fixed values of λ and of the average score, we performed several bioinformatics analyses to the string showed in the previous paragraph.

The average value of the lengths of the 169 sequences whose GenBank IDs appear in S1 Table is 207.11, and the length of our sequence is 206, which is very close to that average. In fact, 206 is the length established for Nef in [42], where the distribution of the amino acids of 1643 Nef sequences was analyzed. This does not imply, of course, that the protein has a well-defined length (there are deletions and insertions in certain positions for some of the sequences) and there is not the same amino acid residue for each position in all sequences. Given the high variability of the protein, it is more appropriate to see the protein as a non-

deterministic distribution of residues conserving to some extent the secondary and tertiary structures and the functionality.

In order to study to what extent our solution captures the well conserved regions of the protein, we considered the sequences of residues conserved at 90% and their starting positions searching in the table of O'Neill et al. [42, Fig 1]. The sequences and positions are shown in S3 Table.

In our solution, all the sequences appear at the same positions as in the table, so all the oligopeptides conserved at 90% are kept.

In order to analyze the structure of the candidate sequence, we have used the bioinformatics tool I-TASSER [43], [44], which is an open source software implemented by Zhang Lab—University of Michigan.

Among the available software, we chose I-TASSER because it has ranked several years as the top method in Critical Assessment of protein Structure Prediction (CASP) experiment, a worldwide test which every two years evaluates the protein structure prediction methods proposed by research groups. More precisely, I-TASSER ranked n°1 in CASP7 (2006), CASP8 (2008), CASP9 (2010), CASP10 (2012), CASP11 (2014), and CASP12 (2016).

In short, the method first compares the proposed sequence with the ones in protein databases to identify similar structural templates and align its amino acid sequences. Next, the unaligned sequences are built by *ab initio* folding and a simulation of different assemblies with the aligned and unaligned sequences is made by Monte Carlo simulations, creating a set of possible candidates. Then, a selection of the lowest free-energy conformations is made and, starting from this model, a second round of assembly simulation is performed in order to refine the global topology [45].

To evaluate the goodness of the predictions, in addition to the TM-Score and the residual RMSD present in the literature, Zhang Lab—University of Michigan has defined a parameter called C-Score. When it is used to evaluate the structural properties, C is typically in the range of [-5,2], where a higher value implies higher confidence in the structure prediction, and models with a C-Score greater than -1.5 are considered reliable predictions.

We analyzed the secondary structure of the candidate sequence. In Fig 3, the secondary structure predicted with I-TASSER for the candidate sequence is displayed. To show the plausibility of the predicted secondary structure, we emphasize that in sequence 2XI1 of Protein Data Bank, which is based in the work by Singh et al. [46], a secondary structure for most part of the C-terminal highly conserved domain of HIV1-Nef is showed, in which there is a high level of agreement with our prediction. Residues 149-178 are disordered in the crystal structure obtained in [46], and hence in that region the sequence is recorded but no coordinates are determined. In 2XI1 the following substructures appear:

alpha helix: 83–95

alpha helix: 106–120

beta strand: 145–149

beta strand: 183–187

3/10 helix: 189–192

alpha helix: 196–200,

which are in good agreement with the structure predicted by I-TASSER.

We also analyzed the tertiary structure of the candidate sequence, which is represented graphically in Fig 4, panel (a). The prediction obtained for our candidate is highly reliable,

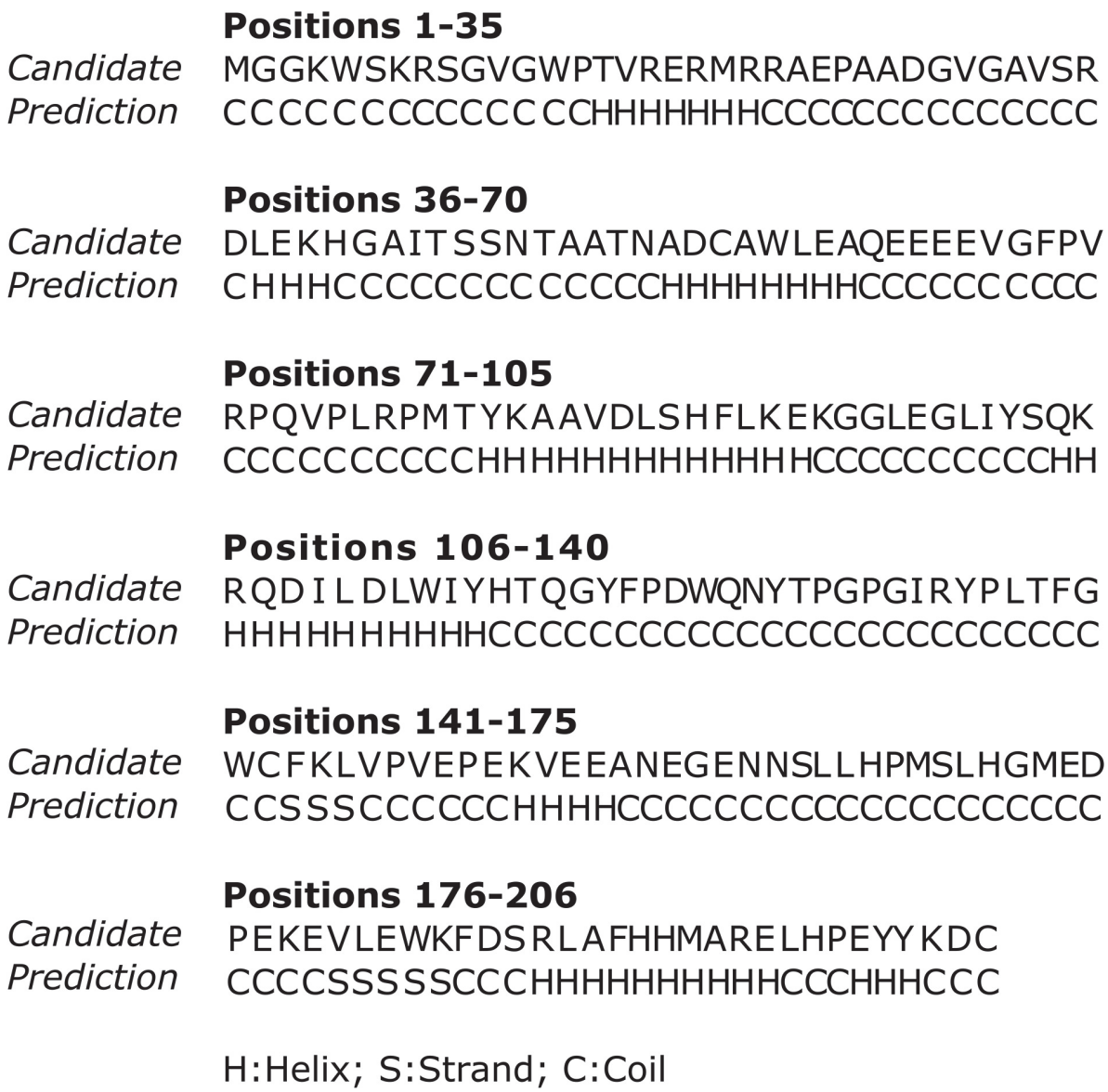

Fig 3. Prediction of the secondary structure of the candidate sequence by I-TASSER. In this graph we represent the amino acid sequence of our candidate, and below, the secondary structure associated to each AA predicted with I-TASSER. Here, H indicates Helix, S Strand, and C Coil.

https://doi.org/10.1371/journal.pone.0211714.g003

since the C-Score of the model is 1.42, and the cutoff value to consider a good prediction is -1.5. Besides, the predicted structure is very similar to the one observed in the Nef protein 3TB8 of Protein Data Bank. This similarity achieved a TM-score of 0.896 in I-TASSER. The TM-score scales the structural similarity between two protein structures. The TM-score ranges on a scale from 0 to 1, with 1 denoting a perfect match and where a scoring greater than 0.5 means that it assumes generally the same fold [47].

In addition to the analysis done with I-TASSER, we have used Phyre2 [48] web portal for protein folding to estimate the structure of the candidate. In Fig 4, panel (b) we illustrate the tertiary structure obtained by Phyre2. It can be observed that the folding is very similar to the

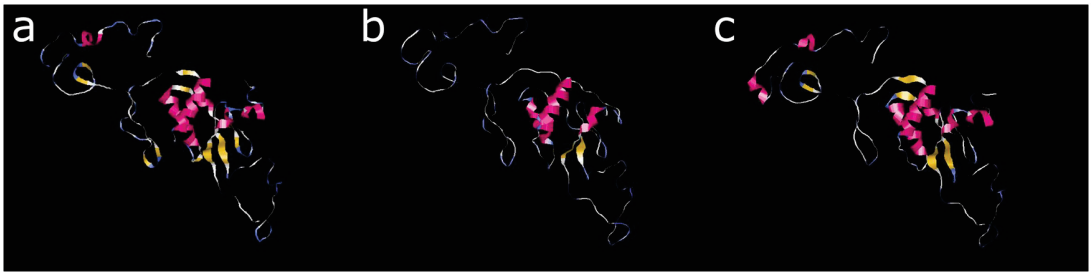

Fig 4. Tertiary structures by I-tasser and Phyre-2. In this molecular graph we illustrate the resemblance between the tertiary structure of (a) the candidate with I-Tasser, (b) the candidate with Phyre-2, and (c) the sequence 2XI1 with Phyre-2.

https://doi.org/10.1371/journal.pone.0211714.g004

one obtained by I-TASSER, depicted in Fig 4, panel (a). Moreover, results of Phyre2 indicate that 98% of the residues were modeled with a confidence >90%, using as model the 3TB8 protein, which is the same that I-TASSER used as model for its predictions. Therefore, in this case, both predictions coincide, which reinforces the likelihood that the candidate will fold as predicted.

Additionally, we have studied the sequence 2XI1 with Phyre2. As in the prediction of the candidate with Phyre2, the main model to estimate the tertiary structure of 2XI1 is the protein 3TB8, with 94% of the residues modeled with a confidence >90% using this template. In Fig 4, panel (c) we illustrate the predicted folding of the sequence 2XI1 by Phyre2, which is very similar to the one obtained with the candidate by using Phyre2, and even more similar to the folding obtained by I-TASSER.

Finally, we can see that the folding predictions done with I-TASSER and Phyre2 were based in the same protein (3TB8) and were very similar (see Fig 4).

We did also a BLAST [49] search of the candidate sequence, and we obtained that the five most similar sequences to the candidate sequence were the following ones:

1. **AAX86040.1**, with a total score of 420 and an identity of 97%. It corresponds to a synthetic construct of a HIV-1 Clade B consensus Nef protein presented in [50], where Kavanagh et al. transfected antigen-presenting cells (APCs) with mRNA encoding Gag-p24 and cytoplasmic, lysosomal, and secreted forms of Nef. They found that transfection of APCs with a Nef construct bearing lysosomal targeting sygnals produced rapid and prolongued antigen presentation to $CD4^+$ and $CD8^+$ T cells [50].
2. **AAX39503.1**, with a total score of 418 and an identity of 97%. It corresponds to a synthetic construct of a consensus Nef protein, which was used in [51], along with other sequences, to validate the FATT-CTL assay, which is a novel method for the measurement of CTL-mediated cytotoxicity.
3. **AAA87523.1**, with a total score of 416 and an identity of 95% and **AAA87527.1**, with a total score of 415 and an identity of 94%. They corresponds to 2 of the 88 sequences of Nef protein of HIV-I, analyzed by Michael et al. in [52].
4. **AAA63871.1**, with a total score of 414 and an identity of 94%. It corresponds to 1 of the 90 sequences of a Nef protein of HIV-I, analyzed by Huang, Zhang, and Ho in [53].

In Fig 5, we depict the alignments of the candidate with the five sequences for the BLAST analysis. When the residues were identical, they were shaded in black; if they were not identical but at least similar, they were colored in grey; finally, when there were no similarities among residuals, they were shaded in white.

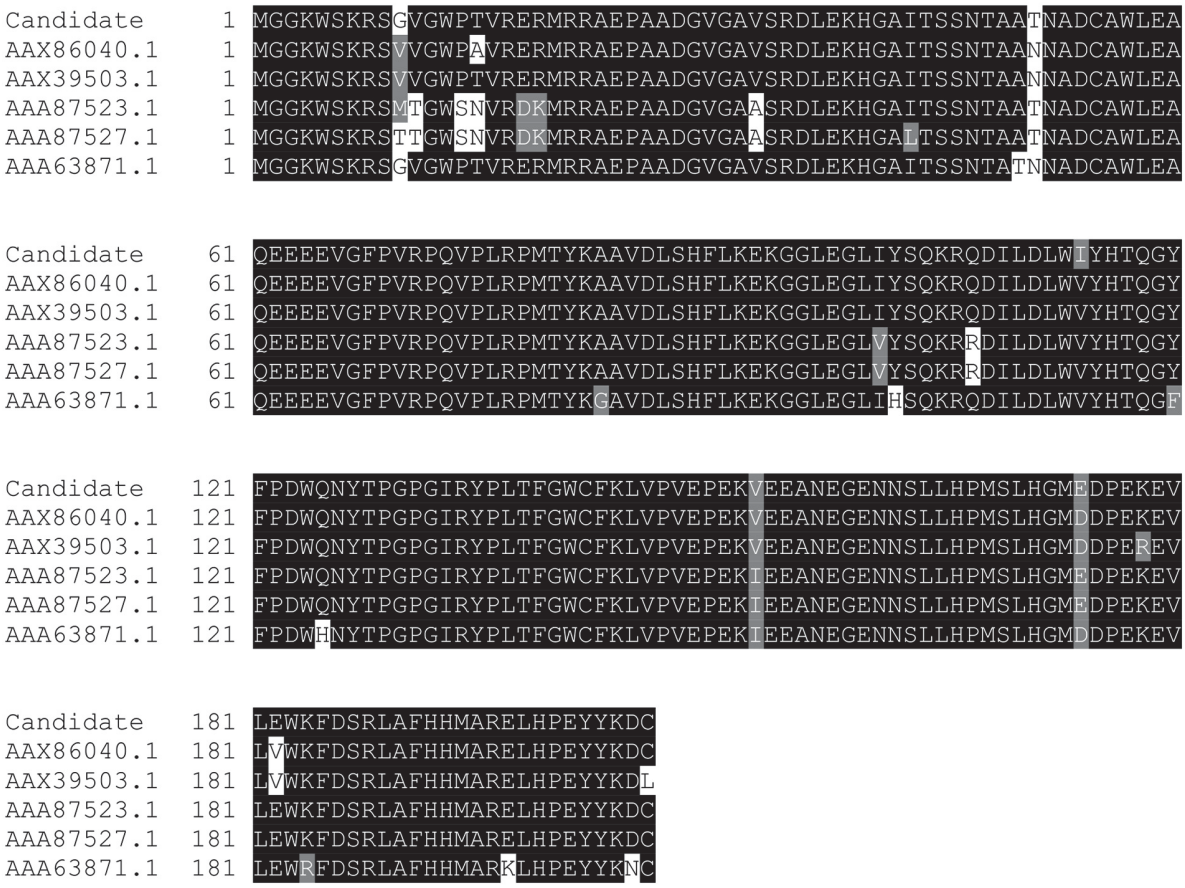

Fig 5. BLAST alignment of the candidate with AAX86040.1, AAX39503.1, AAA87523.1, AAA87527.1, and AAA63871.1. In this graph we depict the alignments of the candidate with the five sequences for the BLAST analysis (AAX86040.1, AAX39503.1, AAA87523.1, AAA87527.1, and AAA63871.1). When the residues were identical, they were shaded in black; if they were not identical but at least similar, they were colored in grey; finally, when there were no similarities among residuals, they were shaded in white.

https://doi.org/10.1371/journal.pone.0211714.g005

In addition, we used VaxiJen [54], which is a server for alignment-independent prediction of protective antigens. It uses bacterial, viral and tumour protein datasets to derive models for prediction of whole protein antigenicity. With our candidate sequence the overall prediction for the antigen obtained with VaxiJen selecting "Virus" as target organism was 0.6895 (Probable antigen). The threshold value to be considered probable antigen was 0.4. For more information about VaxiJen, we refer the reader to [55].

The overall predictions obtained in VaxiJen for the 5 strings closer to our candidate sequence in the BLAST search were:

0.6380 for AAX86040.1

0.6409 for AAX39503.1

0.6688 for AAA87523.1

0.6747 for AAA87527.1

0.6599 for AAA63871.1

Table 4. Comparison between the weighted, unweighted, epigraph, and consensus candidates.

	Class I immunogenicity	mismatch average
Weighted	1.8685	0.5115
Unweighted	1.8409	1
Epigraph	1.2307	0.5114
Consensus	1.4103	0.5109

https://doi.org/10.1371/journal.pone.0211714.t004

Next, we have compared our candidate with other sequences obtained by three different algorithms. The first is one of the candidates obtained with the unweighted algorithm in our previous work [10] (we selected among our solutions the candidate with the number of amino acids closest to 206, i.e., closest to the length established for Nef [42], but without exceeding this number); the second was obtained by using LANL's Epigraph [56]; and the third was a consensus sequence obtained by LANL's Consensus [57].

In Table 4, the resultant estimated Class I immunogenicity [40] and mismatch proportion for the four strings can be found. As expected, the estimated immunogenicity value of our weighted candidate was better than the ones of the other three, suggesting that it would generate a more immunogenic response. The mismatch proportion was very similar (near to 0.511) between the weighted, epigraph, and consensus candidates. This result was expected, since we chose a candidate with high alignment, which implies a smaller number of mismatches, and both epigraph and consensus methods are expected to resemble the natural proteins [56, 57]. Finally, since the unweighted candidate did not take into account the alignment, it scored a very high mismatch ratio (equal to 1).

For the purpose of comparison, we have used also a hill-climbing algorithm, as we did in [10]. In this case we used a multi-objective hill climbing algorithm analogous to the one described in [58]. As we did in the Materials and methods section, we considered sequences u of target strings and the corresponding phenotypes $o(u)$ obtained by taking the overlapping sum of the strings in u. We selected randomly 10 sequences $h_{i_1}, \ldots, h_{i_{10}}$ from the set of host strings and the corresponding sequences $u_{i_1}, \ldots, u_{i_{10}}$ of target strings, and for each sequence u_{i_j} we performed the following procedure:

First, we initialized to $\{u_{i_j}\}$ the set ND_i of non-dominated solutions. Then, we tried to simulate mutations sequentially in positions of the sequences in ND_i, by replacing a target string by another target string at the same position in some of the host strings $h_1, \ldots, h_{169}$. If at some point we get a sequence u' non-dominated for some sequence in ND_i, then we add the sequence u' to the set ND_i and we repeat the process from the beginning. Instead of repeating this process until no new non-dominated sequence is found, due to the excessive time to required to attain this, we simulated a total of 10^6 mutations.

We took the union of the non-dominated sets $ND_1, \ldots, ND_{10}$ and selected the phenotypes of the non-dominated elements in this union as an approximation to the true Pareto front, which is shown in Fig 6 and in Table 5. The approximation to the Pareto front is worse than the one shown in Fig 2, obtained by using the genetic algorithm, in the sense that every solution shown in Fig 6 is dominated by at least one solution in Fig 2.

Discussion

In this paper, we generalized the notion of λ-superstrings from [10] to the weighted case. We developed an exact algorithm for a corresponding combinatorial optimization problem based on integer programming, extending the model from the previous paper by introducing a weight function on the target strings (which can take both positive and negative values). We

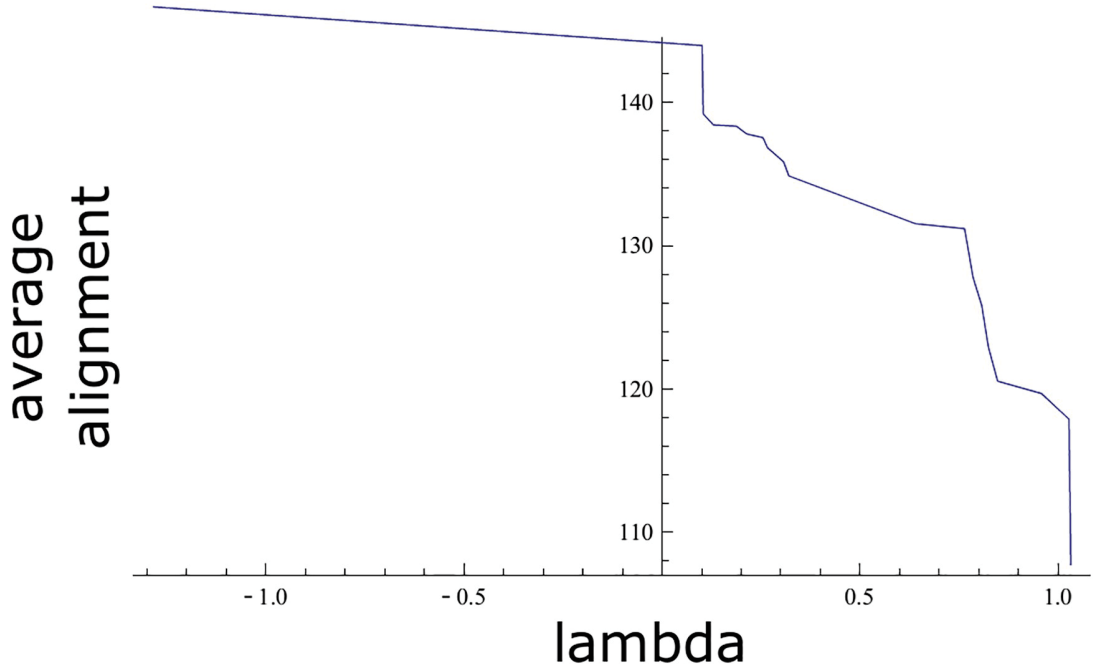

Fig 6. Estimation of the Pareto front in the hill-climbing algorithm. The line represents the non-dominated solutions found with the multi-objective hill climbing algorithm. The X axis indicates the λ value, while the Y axis indicates the alignment score.

https://doi.org/10.1371/journal.pone.0211714.g006

consider that weighted λ-superstring criterion could be useful to fight the high mutability and escape mutations of viruses like HIV, HCV, or Influenza, since it gives a balanced protection against all the variants considered, by ensuring that at the overall the immunogenicity of the epitopes in each variant is at least λ. We also described a model taking into account good

Table 5. Numerical values for the estimation of the Pareto front in the hill-climbing algorithm.

(-1.2852,146.68)
(0.10136,143.982)
(0.10365,139.213)
(0.12965,138.432)
(0.18779,138.349)
(0.21379,137.811)
(0.25488,137.55)
(0.26566,136.87)
(0.30675,135.87)
(0.32028,134.876)
(0.63875,131.544)
(0.76386,131.183)
(0.78531,127.781)
(0.80699,125.817)
(0.82376,122.917)
(0.84713,120.538)
(0.959,119.663)
(1.02748,117.893)
(1.03195,107.686)

https://doi.org/10.1371/journal.pone.0211714.t005

pairwise alignments of the obtained superstring with the host strings, and presented a heuristic approach based on a multi-objective genetic algorithm. By considering the alignment as a target to optimize by our algorithm, the weighted λ-superstrings obtained by using the genetic algorithm correspond to pseudoproteins structurally similar to the original ones taken from the patients, instead of being just epitope aggregates, opening the doors to possible improvements in the current methodology of epitope vaccine design.

In order to evaluate the performance of our algorithm, first, we analyzed the estimation of the Pareto front obtained with a multi-objective hill-climbing algorithm, which gave worse solutions than the one obtained by the genetic algorithm. Then, we selected a vaccine candidate from the Pareto front and studied its effectiveness *in silico*. Due to the weighted λ-superstring condition, and the positive λ value, this pseudo-protein would likely protect against all virus variants considered. Besides, VaxiJen analysis corroborated that the vaccine would be a probable antigen. Next, the structure and resemblance to the native protein were evaluated by several bioinformatic tools (such as Blast, Phyre 2 or I-Tasser), which indicated that our candidate was very similar to HIV-1 2XI1 and 3TB8 sequences. Then, we performed a comparison among our weighted candidate, one of the candidates obtained with the unweighted algorithm in our previous work [10], a candidate obtained by using LANL's Epigraph, and a consensus sequence. In this analysis, we observed that the mismatch proportion was worse in the unweighted candidate, which was expected, since the algorithm in [10] did not optimize the alignment. Besides, the estimated Class I immunogenicity [40] of the weighted candidate was bigger than the estimated immunogenicity for the candidates found with other methods, suggesting that it would generate a more immunogenic response.

Additionally, in order to study the sensitivity of the method, we have also analyzed D and G HIV subtypes, and they yielded similar results, indicating that the method is robust. These analyses can be found in S2 Appendix.

An important point of future work on weighted λ-superstrings is to determine the extent of practical applicability of the presented models and algorithms to vaccine design, in particular to assess the immunological value of the resulting candidate vaccines. In this regard, we have recently described a functional method to decipher T-cell epitopes of the bacterial and human pathogen *Listeria monocytogenes (Listeria)* based on combination of bioinformatics predictions of epitopes binding to MHC molecules and functional assays [59]. Our hypothesis was based in the use of two *Listeria* antigens, the listeriolysin O (LLO) and the glyceraldehyde-3-phosphate-dehydrogenase (GAPDH) that elicits strong CD4+ and CD8+ T cell responses [60], [61]. Our method to test vaccine candidates was based in the use of predicted peptides from the bioinformatics analysis to activate dendritic cells *in vitro* and elicit high delayed T hypersensitivity (DTH) responses *in vivo*, combined to measurements of IL-12 production as the cytokine that best correlates with immune protection.

In order to adapt the methodology just described to the framework of vaccine design using weighted λ-superstrings, we will use in future work the full-length sequence of LLO for the thirteen recognized serotypes of *Listeria Monocytogenes* to design B and T-cell epitope vaccines applying weigthed λ-superstrings that gather the genetic diversity of the pathogen by means of the consideration of the different serotypes, and we will compare the epitopes obtained with those of previous studies. Next, we will use the weighted λ-superstrings obtained with the selected epitopes in our functional method of vaccine candidates testing. Finally, our success in predicting efficient LLO epitopes for vaccination and the construction of the subsequent λ-superstrings will be relevant for other intracellular bacteria for which we currently lack available vaccines, such as *Mycobacterium tuberculosis, Salmonella enteritidis*, or *Chlamydia trachomatis*, among others.

One of the lines considered as future work is to evaluate if the algorithm gives better results when we consider near-matches of the epitopes instead of exact matches, by changing the fitness function. By this approach, we would obtain vaccine candidates that induce cross-reactive T-Cells, which could be activated during the infection of an unrelated heterologous virus. Cross-reaction and its benefits have been widely observed in several infections [62, 63], and since their positive effects in vaccination is promising [64, 65], we consider that this approach might enhance the effectiveness of our method.

Moreover, we would like to consider, besides the weights corresponding to immunogenicities, other kinds of weights at the same time, addressing different biologically motivated goals with different weights. For example, one could consider weights associated to the relative frequencies of the epitopes.

In summary, here we have presented two algorithms for computational vaccine design. Our results indicate that with this methodology, we can obtain weighted λ-superstrings that resemble native protein structures, and protect well-balancedly against the whole group of considered virus variants, minimizing the chances of perpetuating the infection due to escape mutations.

Supporting information

S1 Appendix. Mathematical proof.
(PDF)

S2 Appendix. Additional sub-analyses.
(PDF)

S1 Fig. Distribution of the values in the Pareto front. Histograms representing the frequencies of the λ (a) and alignment (b) values of the estimation of the Pareto front obtained with the genetic algorithm. The Y axis represents the frequency, while the X axis indicates the λ value (a) and the alignment score (b).
(TIF)

S1 Table. GenBank IDs of the sequences for the Nef protein.
(PDF)

S2 Table. Experimental values of the immunogenicities of the epitopes.
(PDF)

S3 Table. Positions and sequences of the conserved regions for the Nef protein.
(PDF)

Acknowledgments

Supported in part by the Basque Government, grants IT753-13 and IT974-16 and by the UPV/EHU and Basque Center of Applied Mathematics, grant US18/21. Supported in part by the Slovenian Research Agency (I0-0035, research program P1-0285, and research projects N1-0032, J1-7051, and J1-9110). Technical and human support provided by IZO-SGI, SGIker (UPV/EHU, MICINN, GV/EJ, ERDF and ESF) is gratefully acknowledged. We would like to thank the referees for their helpful suggestions.

Author Contributions

Conceptualization: Luis Martínez, Martin Milanič, Carmen Álvarez, Ildefonso M. de la Fuente.

Data curation: Iker Malaina.

Investigation: Luis Martínez, Martin Milanič, Iker Malaina.

Methodology: Luis Martínez.

Software: Luis Martínez, Martin Milanič, Iker Malaina, Martín-Blas Pérez.

Supervision: Luis Martínez.

Writing – original draft: Luis Martínez, Martin Milanič, Iker Malaina, Carmen Álvarez, Martín-Blas Pérez, Ildefonso M. de la Fuente.

References

1. Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. Immunology. 2010; 130: 319–328. https://doi.org/10.1111/j.1365-2567.2010.03268.x PMID: 20408898
2. Khan AM, Miotto O, Heiny AT, Salmon J, Srinivasan KN, Nascimento EJ, et al. A systematic bioinformatics approach for selection of epitope-based vaccine targets. Cell Immunol. 2006; 244: 141–147. https://doi.org/10.1016/j.cellimm.2007.02.005 PMID: 17434154
3. Wang HW, Lin YC, Pai TW, Chang HT. Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification. Biomed Res Int. 2011;
4. Hemmer B, Kondo T, Gran B, Pinilla C, Cortese I, Pascal J, et al. Minimal peptide length requirements for CD4+ T cell clones—implications for molecular mimicry and T cell survival. Int Immunol. 2000; 12: 375–383. https://doi.org/10.1093/intimm/12.3.375 PMID: 10700472
5. Sharon J, Rynkiewicz MJ, Lu Z, Yang CY. Discovery of protective B-cell epitopes for development of antimicrobial vaccines and antibody therapeutics. Immunology. 2014; 142: 1–23. https://doi.org/10.1111/imm.12213 PMID: 24219801
6. El-manzalawy Y, Dobbs D, Honavar V. Predicting flexible length linear B-cell epitopes. Computational Systems Bioinformatics. 2008; 7: 121–132. https://doi.org/10.1142/9781848162648_0011 PMID: 19642274
7. Geginat G, Schenk S, Skoberne M, Goebel W, Hof H. A novel approach of direct ex vivo epitope mapping identifies dominant and subdominant CD4 and CD8 T cell epitopes from Listeria monocytogenes. The Journal of Immunology. 2001; 166: 1877–1884. https://doi.org/10.4049/jimmunol.166.3.1877 PMID: 11160235
8. Skoberne M, Geginat G. Efficient in vivo presentation of Listeria monocytogenes-derived CD4 and CD8 T cell epitopes in the absence of IFN-γ. The Journal of Immunology. 2002; 168: 1854–1860. https://doi.org/10.4049/jimmunol.168.4.1854 PMID: 11823519
9. Kim Y, Ponomarenko J, Zhu Z, Tamang D, Wang P, Greenbaum J, et al. Immune epitope database analysis resource. Nucleic Acids Res. 2012; 40: 525–530. https://doi.org/10.1093/nar/gks438
10. Martinez L, Milanic M, Legarreta L, Medvedev P, Malaina I, De la Fuente IM. A combinatorial approach to the design of vaccines. J Math Biol. 2015; 70: 1327–1358. https://doi.org/10.1007/s00285-014-0797-4 PMID: 24859149
11. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE T Evol Comput. 2002; 6: 182–197. https://doi.org/10.1109/4235.996017
12. Manzourolajdad A, Gonzalez M, Spouge JL. Changes in the Plasticity of HIV-1 Nef RNA during the Evolution of the North American Epidemic. PloS One. 2016; 11: e0163688. https://doi.org/10.1371/journal.pone.0163688 PMID: 27685447
13. Sharma D, Bhattacharya J. Cellular & molecular basis of HIV-associated neuropathogenesis. Indian J Med Res. 2009; 129: 637. PMID: 19692743
14. Dinur I, Steurer D. Analytical approach to parallel repetition. In Proceedings of the forty-sixth annual ACM symposium on Theory of computing; 2014; 624-633.
15. Alon N, Moshkovitz D, Safra S. Algorithmic construction of sets for k-restrictions. ACM Transactions on Algorithms (TALG). 2006; 2: 153–177. https://doi.org/10.1145/1150334.1150336
16. Schrijver A. Theory of Linear and Integer Programming. Amsterdam: John Wiley & Sons; 1998.
17. Pataki G. Teaching integer programming formulations using the traveling salesman problem. SIAM Rev. 2003; 45: 116–123. https://doi.org/10.1137/S00361445023685

18. Miller CE, Tucker AW, Zemlin RA. Integer programming formulation of traveling salesman problems. J ACM. 1960; 7: 326–329. https://doi.org/10.1145/321043.321046
19. Haupt RL, Haupt SE. Practical genetic algorithms. New Jersey: John Wiley & Sons; 2004.
20. Gusfield D. Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge: Cambridge university press; 1997.
21. GenBank website. [cited 06 August 2018]. Available from: www.ncbi.nlm.nih.gov/genbank/.
22. Nickle DC, Rolland M, Jensen MA, Pond SLK, Deng W, Seligman M, et al. Coping with viral diversity in HIV vaccine design. PLoS Comput Biol. 2007; 3:e75. https://doi.org/10.1371/journal.pcbi.0030075 PMID: 17465674
23. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, et al. The immune epitope database 2.0. Nucleic Acids Res. 2009; 38: 854–862. https://doi.org/10.1093/nar/gkp1004
24. HIV Molecular Immunology Database website. [cited 06 August 2018]. Available from: www.hiv.lanl.gov/content/immunology.
25. Rasmussen M, Fenoy E, Harndahl M, Kristensen AB, Nielsen IK, Nielsen M, et al. Pan-Specific Prediction of Peptide–MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. The Journal of Immunology. 2016; 1600582.
26. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. The Journal of Immunology. 1994; 153:5586–5592. PMID: 7527444
27. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. The Journal of Immunology. 2013; 1302101.
28. Java website. [cited 06 August 2018]. Available from: www.java.com
29. IBM ILOG CPLEX Optimization Studio website. [cited 06 August 2018]. Available from: www-03.ibm.com/software/products/us/en/ibmilogcpleoptistud/.
30. Bergmann-Leitner ES, Chaudhury S, Steers NJ, Sabato M, Delvecchio V, Wallqvist AS, et al. Computational and experimental validation of B and T-cell epitopes of the in vivo immune response to a novel malarial antigen. PloS One. 2013; 8: e71610. https://doi.org/10.1371/journal.pone.0071610 PMID: 23977087
31. Bryson CJ, Jones TD, Baker MP. Prediction of immunogenicity of therapeutic proteins. BioDrugs. 2010; 24: 1–8. https://doi.org/10.2165/11318560-000000000-00000 PMID: 20055528
32. Khan JM, Kumar G, Ranganathan S. In silico prediction of immunogenic T cell epitopes for HLA-DQ8. Immunome Research. 2012; 8: 1–9.
33. Moreau V, Fleury C, Piquer D, Nguyen C, Novali N, Villard S, et al. PEPOP: computational design of immunogenic peptides. BMC Bioinformatics. 2008; 9: 71. https://doi.org/10.1186/1471-2105-9-71 PMID: 18234071
34. Calis JJ, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC class I presented peptides that enhance immunogenicity. PLoS Comput Biol. 2013; 9: e1003266. https://doi.org/10.1371/journal.pcbi.1003266 PMID: 24204222
35. Paul S, Kolla RV, Sidney J, Weiskopf D, Fleri W, Kim Y, et al. Evaluating the immunogenicity of protein drugs by applying in vitro MHC binding data and the immune epitope database and analysis resource. Clinical and Developmental Immunology. 2013; 2013. https://doi.org/10.1155/2013/467852 PMID: 24222776
36. Ponomarenko JV, Van Regenmortel MH. B cell epitope prediction. Structural bioinformatics. 2009; 849–879.
37. Shmelkov E, Krachmarov C, Grigoryan AV, Pinter A, Statnikov A, Cardozo T. Computational prediction of neutralization epitopes targeted by human anti-V3 HIV monoclonal antibodies. PloS One. 2014; 9: e89987. https://doi.org/10.1371/journal.pone.0089987 PMID: 24587168
38. Tong JC, Tan TW, Ranganathan S. Methods and protocols for prediction of immunogenic epitopes. Brief Bioinform. 2006; 8: 96–108. https://doi.org/10.1093/bib/bbl038 PMID: 17077136
39. Tung CW, Ho SY. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. Bioinformatics. 2007; 23: 942–949. https://doi.org/10.1093/bioinformatics/btm061 PMID: 17384427
40. IEDB, T cell class I pMHC immunogenicity predictor website. [cited 06 August 2018]. Available from: http://tools.immuneepitope.org/immunogenicity/.
41. Mathematica website. [cited 06 August 2018]. Available from: https://www.wolfram.com/mathematica/.

42. O'Neill E, Kuo LS, Krisko JF, Tomchick DR, Garcia JV, Foster JL. Dynamic evolution of the human immunodeficiency virus type 1 pathogenic factor, Nef. J Virol. 2006; 80: 1311–1320. https://doi.org/10.1128/JVI.80.3.1311-1320.2006 PMID: 16415008

43. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008; 9: 40. https://doi.org/10.1186/1471-2105-9-40 PMID: 18215316

44. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010; 5: 725. https://doi.org/10.1038/nprot.2010.5 PMID: 20360767

45. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2015; 12: 7. https://doi.org/10.1038/nmeth.3213 PMID: 25549265

46. Singh P, Yadav GP, Gupta S, Tripathi AK, Ramachandran R, Tripathi RK. A novel dimer-tetramer transition captured by the crystal structure of the HIV-1 Nef. PloS One. 2011; 6: e26629. https://doi.org/10.1371/journal.pone.0026629 PMID: 22073177

47. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005; 33: 2302–2309. https://doi.org/10.1093/nar/gki524 PMID: 15849316

48. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015; 10: 845. https://doi.org/10.1038/nprot.2015.053 PMID: 25950237

49. BLAST website. [cited 06 August 2018]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi.

50. Kavanagh DG, Kaufmann DE, Sunderji S, Frahm N, Le Gall S, Boczkowski D, et al. Expansion of HIV-specific CD4+ and CD8+ T cells by dendritic cells transfected with mRNA encoding cytoplasm-or lysosome-targeted Nef. Blood. 2006; 107: 1963–1969. https://doi.org/10.1182/blood-2005-04-1513 PMID: 16249391

51. van Baalen CA, Kwa D, Verschuren EJ, Reedijk ML, Boon AC, de Mutsert G, et al. Fluorescent Antigen–Transfected Target Cell Cytotoxic T Lymphocyte Assay for Ex Vivo Detection of Antigen-Specific Cell-Mediated Cytotoxicity. The Journal of infectious diseases. 2005; 192: 1183–1190. https://doi.org/10.1086/444546 PMID: 16136460

52. Michael NL, Chang G, d'Arcy LA, Tseng CJ, Birx DL, Sheppard HW. Functional characterization of human immunodeficiency virus type 1 nef genes in patients with divergent rates of disease progression. J Virol. 1995; 69: 6758–6769. PMID: 7474087

53. Huang Y, Zhang L, Ho DD. Characterization of nef sequences in long-term survivors of human immunodeficiency virus type 1 infection. J Virol. 1995; 69: 93–100. PMID: 7983771

54. VaxiJen website. [cited 06 August 2018]. Available from: www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html.

55. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics. 2007; 8: 4. https://doi.org/10.1186/1471-2105-8-4 PMID: 17207271

56. LANL's Epigraph website [cited 02 November 2018]. Available from: https://www.hiv.lanl.gov/content/sequence/EPIGRAPH/epigraph.html.

57. LANL's Consensus website [cited 02 November 2018]. Available from: https://www.hiv.lanl.gov/content/sequence/CONSENSUS/SimpCon.html.

58. Diaz R, Suarez AR. A study of the capacity of the stochastic Hill Climbing to solve multi-objective problems. In Proceedings of the Third International Symposium on Adaptive Systems-Evolutionary Computation and Probabilistic Graphical Models, La Habana: Institute of Cybernetics, Mathematics and Physics. 2001; 37-40.

59. Calderon-Gonzalez R, Tobes R, Pareja E, Frande-Cabanes E, Petrovsky N, Alvarez-Dominguez C. Identification and characterisation of T-cell epitopes for incorporation into dendritic cell-delivered Listeria vaccines. J Immunol Methods. 2015; 424: 111–119. https://doi.org/10.1016/j.jim.2015.05.009 PMID: 26031451

60. Alvarez-Dominguez C, Madrazo-Toca F, Fernandez-Prieto L, Vandekerckhove J, Pareja E, Tobes R, et al. Characterization of a Listeria monocytogenes protein interfering with Rab5a. Traffic. 2008; 9: 325–337. https://doi.org/10.1111/j.1600-0854.2007.00683.x PMID: 18088303

61. Calderón-González R, Frande-Cabanes E, Bronchalo-Vicente L, Lecea-Cuello MJ, Pareja E, Bosch-Martínez A, et al. Cellular vaccines in listeriosis: role of the Listeria antigen GAPDH. Front Cell Infect Mi. 2014; 4: 22–33.

62. Welsh RM, Selin LK. No one is naive: the significance of heterologous T-cell immunity. Nature Reviews Immunology. 2002; 2:417. https://doi.org/10.1038/nri820 PMID: 12093008

63. Rehermann B, Shin EC. Private aspects of heterologous immunity. Journal of Experimental Medicine. 2005; 201:667–670. https://doi.org/10.1084/jem.20050220 PMID: 15753200

64. de Boer RJ, Perelson AS. How germinal centers evolve broadly neutralizing antibodies: the breadth of follicular helper T cell response. Journal of Virology. 2017; JVI-00983. https://doi.org/10.1128/JVI.00983-17 PMID: 28878083

65. Wang S. Optimal sequential immunization can focus antibody responses against diversity loss and distraction. PLoS Computational Biology. 2017; 13:e1005336. https://doi.org/10.1371/journal.pcbi.1005336 PMID: 28135270

scientific reports

Check for updates

OPEN

First computational design using lambda-superstrings and in vivo validation of SARS-CoV-2 vaccine

Luis Martínez[1,2,12]✉, Iker Malaina[1,3], David Salcines-Cuevas[4], Héctor Terán-Navarro[4], Andrea Zeoli[4], Santos Alonso[5,6], Ildefonso M. De la Fuente[1,11], Elena Gonzalez-Lopez[7], J. Gonzalo Ocejo-Vinyals[7], Mónica Gozalo-Margüello[8], Jorge Calvo-Montes[4,8,10] & Carmen Alvarez-Dominguez[4,9,12]✉

Coronavirus disease 2019 (COVID-19) is the greatest threat to global health at the present time, and considerable public and private effort is being devoted to fighting this recently emerged disease. Despite the undoubted advances in the development of vaccines against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19, uncertainty remains about their future efficacy and the duration of the immunity induced. It is therefore prudent to continue designing and testing vaccines against this pathogen. In this article we computationally designed two candidate vaccines, one monopeptide and one multipeptide, using a technique involving optimizing lambda-superstrings, which was introduced and developed by our research group. We tested the monopeptide vaccine, thus establishing a proof of concept for the validity of the technique. We synthesized a peptide of 22 amino acids in length, corresponding to one of the candidate vaccines, and prepared a dendritic cell (DC) vaccine vector loaded with the 22 amino acids SARS-CoV-2 peptide (positions 50-71) contained in the NTD domain (DC-CoVPSA) of the Spike protein. Next, we tested the immunogenicity, the type of immune response elicited, and the cytokine profile induced by the vaccine, using a non-related bacterial peptide as negative control. Our results indicated that the CoVPSA peptide of the Spike protein elicits noticeable immunogenicity in vivo using a DC vaccine vector and remarkable cellular and humoral immune responses. This DC vaccine vector loaded with the NTD peptide of the Spike protein elicited a predominant Th1-Th17 cytokine profile, indicative of an effective anti-viral response. Finally, we performed a proof of concept experiment in humans that included the following groups: asymptomatic non-active COVID-19 patients, vaccinated volunteers, and control donors that tested negative for SARS-CoV-2. The positive control was the current receptor binding domain epitope of COVID-19 RNA-vaccines. We successfully developed a vaccine candidate technique involving optimizing lambda-superstrings and provided proof of concept in human subjects. We conclude that it is a valid method to decipher the best epitopes of the Spike protein of SARS-CoV-2 to prepare peptide-based vaccines for different vector platforms, including DC vaccines.

Abbreviations

ACE2	Angiotensin-converting enzyme 2 receptor
APC	Allophycocyanin

[1]Department of Mathematics, Faculty of Science and Technology, University of the Basque Country, UPV/EHU, 48940 Leioa, Spain. [2]BCAM, Basque Center for Applied Mathematics, 48009 Bilbao, Spain. [3]BioCruces Health Research Institute, Cruces University Hospital, 48903 Barakaldo, Spain. [4]Instituto de Investigación Marqués de Valdecilla (IDIVAL), 39011 Santander, Spain. [5]Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country, UPV/EHU, 48940 Leioa, Spain. [6]María Goyri Building. Animal Biotechnology Center, University of the Basque Country, UPV/EHU, 48940 Leioa, Spain. [7]Servicio de Inmunología, Hospital Universitario Marqués de Valdecilla, 39008 Santander, Spain. [8]Servicio de Microbiología, Hospital Universitario Marqués de Valdecilla, 39008 Santander, Spain. [9]Universidad Internacional de La Rioja, 26006 Logroño, Spain. [10]CIBER Enfermedades Infecciosas, ISCIII, Madrid, Spain. [11]Department of Nutrition, CEBAS-CSIC Institute, Espinardo University Campus, 30100 Murcia, Spain. [12]These authors contributed equally: Luis Martínez and Carmen Alvarez-Dominguez. ✉email: luis.martinez@ehu.eus; carmen.alvarezd@scsalud.es

CV	Candidate vaccine
DC	Dendritic cell
DMEM	Dulbecco's Modified Eagle's Medium
DTH	Delayed type hypersensitivity
GM-CSF	Granulocyte-macrophage colony-stimulating factor
RBD	Receptor binding domain
7-AAD	7-Aminoactinomycin D

The coronavirus disease 2019 (COVID-19) epidemic represents the greatest global threat to human health at the current juncture, with more than 281 million people infected and more than 5.4 million mortalities worldwide since the disease was detected two years ago[1]. To end this epidemic, different types of vaccines are being developed in an accelerated manner[2–6].

Although new cutting-edge technologies are being used in the production of vaccines, such as the development of mRNA vaccines[7,8], which speeds up the manufacturing process and reduces the cost of fabrication, they do not take into account the mutations that arise as the pandemic progresses.

Several effective vaccines against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19, are currently available, such as the Pfizer, Moderna, Oxford Vaccine Group/AstraZeneca, Janssen, BIOCAD, and CanSino Biologics vaccines. Clinical trials, as well as data from ~50%–70% of vaccinated individuals in Europe and EEUU, show that the highest protection corresponds to the Pfizer (93%) and Moderna (90%) vaccines; the duration of this protection still requires evaluation. Thus, despite the important and promising advances that have been made in the design and development of vaccines against SARS-CoV-2, uncertainties still remain[9,10], making it imperative to continue the search for new candidate vaccines (CVs).

Several tools from computational biology are being used for different tasks in the fight against COVID-19, such as modeling[11], identifying epitope maps[12], designing protein inhibitors[13], identifying inhibitors of the interaction of the Spike protein with angiotensin-converting enzyme 2 (ACE2) receptor[14], optimizing antibodies[15], and designing CVs[16–19].

We present here two peptide-based CVs for SARS-CoV-2 designed entirely using computational methods and advanced tools of artificial intelligence. For one of these peptides, we provide proof of concept for immunogenicity.

Our technique is based in the concept of λ-superstring, introduced by our research group in a previous publication[20]. In that paper, we presented a new criterion for the selection of epitopes in the design of vaccines that was well suited to consider all the mutations, providing a balance with respect to the number of epitopes covered by the CV in the mutated versions of the target protein. Using this method, we consider the mutations in the pathogen's genome and develop a CV that performs well against all those mutations.

Specifically, we considered a set of target strings, formed by the epitopes that can be selected for the CV, and a set of host strings, constituted by the different variants of the target protein, in which the known mutations are considered. In that context, given the value of a parameter λ, a λ-superstring is a sequence of amino acids with properties that ensure that the string covers at least λ epitopes in each of the host strings.

The concept of λ-superstring was generalized in our subsequent publication[21] to that of a weighted λ-superstring by allowing the epitopes to be weighted by estimations of their immunogenicities. This generalization entails an important improvement in the applications to vaccine design[22], as this consideration of an epitope's immunogenicity more closely models biological and medical practice and increases the likelihood that the obtained CVs are effective. In fact, the use of weighted λ-superstrings could be useful in response to the high mutability of viruses such as human immunodeficiency virus (HIV), hepatitis C virus (HCV), and influenza, and the generation of escape mutations.

Peptide-based vaccines against SARS-CoV-2 are developed using recombinant technology, which is the most widely used vaccine strategy. In fact, 50 protein or peptide constructs are in pre-clinical trials and 8 out of these 50 are in clinical trials, being the receptor binding domain (RBD) region of the Spike protein a common epitope of COVID-19 vaccines[3], as well as an epitope detected in antibodies and T cells from COVID-19 patients[23,24]. These studies also indicated that the RBD region was not the only B or T cell epitope detected in patients, with other parts of the Spike protein also being detected. Therefore, peptide-based vaccines for SARS-CoV-2 should include other epitopes of the Spike protein to maximize the broad spectrum of cellular immune responses they might elicit.

The success of peptide-based vaccines relies on three criteria: (1) easy to manufacture, (2) cost-effective production, and (3) high safety profile compared with whole virus vaccines[3,23,25]. However, such vaccines also have limitations, including the fact that they require specific methodologies to design epitopes with high immunogenicity and to test their efficacy as vaccines. To address this issue, we used two combined methods in the development of COVID-19 peptide-based vaccines. First, a computational technique that uses advanced tools of artificial intelligence to design the best immunogenic epitopes[20,21]. Second, subcutaneous inoculation of DC vaccine platforms loaded with these peptides to test the delayed type hypersensitivity (DTH) reactions in mice, as an in vivo measurement of cellular immune responses that provides easy preparation for the clinic[22].

To establish an experimental proof of concept of the method of using λ-superstrings, in this work we used Integer Programming to obtain the best solution for different lengths of the CV, to achieve the maximum value of λ, and therefore optimal immunological protection.

We used the Spike protein as the target, which is the best suited antigen for SARS-CoV-2 vaccine research, because in addition to being an intermediate in the interaction with host cell binding to the ACE2 receptor, it is also a surface-exposed protein[26,27], which makes it a suitable target protein for vaccine development.

In this way, we obtained CVs that offer potential protection against all the virus variants considered for the study (in this case, all the sequences appearing in the GenBank[28] and GISAID[29] websites until March 4, 2020). Our objective was not only to develop an effective vaccine for the current pandemic, but also to confer protection against potential coronavirus mutations by considering sufficiently high values of λ. Thus, although future mutations might diminish the value of λ when new host strings associated to mutations appear, λ will nevertheless remain sufficiently high to be medically effective and inhibit the expansion and possible resurgence of the virus.

After obtaining the sequences, to select the most promising epitopes for the vaccine, we first associated a weight to each potential 9-mer epitope. More precisely, we used an aggregate function that combines estimations of the immunogenicity and the HLA-binding affinity of class I of these potential epitopes. The procedure is outlined briefly below, but is described in more detail in the "Methods" section. We used the "T cell class I pMHC immunogenicity predictor"[30] of IEDB for the estimation of immunogenicity, and the "Peptide binding to MHC class I molecules"[31] tool of IEDB for the estimation of HLA-I binding affinity. Since we had no specific information about the alleles of the different patients (whose number is much greater than the number of host strings) and our vaccine is not a personalized vaccine, we weighted each allele with its frequency appearing in "The Allele Frequency Net Database"[32] that corresponds to the alleles of the HLA-I allele reference set previously published [33].

We chose these two variables, immunogenicity and HLA-I binding affinity, because of the key need to engage T cells in the development of an effective vaccine response against SARS-CoV-2 to ensure long-lasting immunity[34].

We did not consider estimations of HLA-II binding affinity for three reasons. First, antigen presentation of viral proteins is mainly restricted by HLA-I molecules. Second, the computational complexity of the Integer Programming problem increases considerably, and it becomes impractical to obtain solutions in the range of lengths that we have considered. Moreover, it takes a prohibitive time for the algorithm to finish and requires a large amount of computer RAM. Third, prediction algorithms for MHC class II presentation are less accurate compared with class I algorithms[35].

Besides, although HLA-II binding affinity was not considered in the optimization, the experimental results have shown a strong humoral immunity elicited by the CV, as is indicated in the "Methods" section.

To assure the antigenicity of our CVs we used the VaxiJen[36] tool, which is a server for alignment-independent prediction of protective antigens that uses bacterial, viral and tumour protein datasets to derive models for prediction of whole protein antigenicity. We selected from among the 272 candidates those that exceeded the threshold to be considered probable antigens for the VaxiJen viral model.

As a result of our computations, we present in this work a map of optimal CVs with lengths varying from 9 to 280 amino acids, which on one hand optimize both immunogenicity and HLA-binding affinity, and on the other hand, confer balanced protection against all of the sequenced variants of COVID-19 surface protein obtained up to the moment at which the data were collected.

Then, to test the efficacy of our method, we first selected from this map a peptide of 22 amino acids contained within the NTD domain of the Spike protein (50–71 amino acids in the surface glycoprotein in the NCBI Reference Sequence YP_009724390.1). We selected a 22-mer of the NTD domain since it has a length feasible to be synthesized at high purity, it is an optimal length for loading onto DCs that admit epitopes lower than 30-mer[22], and because a 22-mer of *Listeria monocytogenes* has previously shown high immunogenicity as CV, as well as effective protection[22]. Next, we incorporated the peptide into DC-based vaccine vectors to explore the epitope safety and immunogenicity, and to determine the type of peptide-induced immune response.

Finally, we performed a small proof of concept experiment in COVID-19-vaccinated volunteers and detected high levels of neutralizing IgG COVID-19 antibodies against the NTD domain that indicated a protective response.

Methods

Setting of the problem. Given two sets H and T of strings, called host strings and target strings, respectively, and given a mapping $w : T \to \mathbb{R}$, we say that a string s is a weighted λ–superstring[21] if, for every $h \in H$, the inequality $\sum_{t \in CS(h,s) \cap T} w(t) \geq \lambda$ holds, where $CS(h, s)$ is the set of common substrings of h and s.

We solved in this work an instance of one of the combinatorial optimization problems settled in[21]: given a fixed length, find a λ-superstring with the maximum value of λ.

Extraction of the sequences. The sequences were taken from two sources, the GenBank database[28] and GISAID[29] (using the sequences available up to March 4, 2020).

To search the GenBank database, the search term "Severe acute respiratory sindrome coronavuirus 2" AND "Homo sapiens" was applied for a nucleotide search. The results of the search were saved as Coding Sequences in the FASTA protein format to obtain information about the corresponding amino acids. Then, in the generated file, sequences corresponding to "surface" or to "spike" protein (which both refer to the same protein) were selected.

To search GISAID, the term "Human" was selected in the host window and "complete" (> 29000 bp) was also selected to obtain only complete genomes in human hosts. Information about the surface protein was extracted from the genomes using the GeneWise[37] tool, by inserting the sequence of aminoacids corresponding to the surface glycoprotein (YP_009724390.1) product in the reference genome (NC_045512.2) into the protein window.

Duplicated sequences were removed, as well as sequences containing ambiguous characters such as "X", "B", "Z", "J", "O", "U", ".", "*". An anomalous short sequence of 35 amino acids was also discarded.

The resulting 22 sequences were taken as host strings, constituting the set H. The multiple alignment of the sequences, obtained using BioEdit, can be found in Additional file 1.

Weighting of the epitopes. The set T of target strings was taken to be the set of 9-tuples of elements of A (where A is the set of 20 amino acids) that are contained in at least one host string and that correspond to residues 1 to 1208, located before the transmembrane domain[38].

The weight $w(s)$ associated to a target string (epitope) s was calculated is as follows:

(1) The estimation $i(s)$ of the immunogenicity of s was calculated with the "T cell class I pMHC immunogenicity predictor" of IEDB.
(2) The set $AI(s)$ of alleles of the HLA-I allele reference set with the "Peptide binding to MHC class I molecules" tool of IEDB which pass the threshold was computed, and the number $bI(s) = \sum_{i \in AI(s)} f(a)$ was calculated, where $f(a)$ is the estimated global frequency of allele a in "The Allele Frequency Net Database".

Next, the normalized families were computed as follows:

$$i_N(s) = \frac{i(s) - m_i}{M_i - m_i}, \text{ where } m_i = min_{s \in T} i(s) \text{ and } M_i = max_{s \in T} i(s),$$

$$bI_N(s) = \frac{bI(s) - m_{bI}}{M_{bI} - m_{bI}}, \text{ where } m_{bI} = min_{s \in T} bI(s) \text{ and } M_{bI} = max_{s \in T} bI(s).$$

Finally, the weight of epitope s was taken as

$$w(s) = \frac{3i_N(s) + bI_N(s)}{4}.$$

The ponderation of the immunogenicity was taken to be larger than that of the binding affinity to favor the former, taking into account that it is a deterministic estimation, while the latter is a probabilistic estimation based on the frequencies of the most frequent alleles, and it does not cover exhaustively all possible alleles.

The target strings and weights can be found in Additional file 2.

Optimization with CPLEX. CPLEX Optimizer[39] was used with an Intel(R) Xeon(R) CPU E5-4620 v2 @ 2.60 GHz Processor with 512 GB of RAM to solve the Integer Programming algorithm described previously [21] to maximize the value of λ for a fixed length, with the set of host strings described in "Extraction of the sequences" and the set of target strings and weights described in "Weighting of the epitopes".

The Integer Programming is founded in a graph theoretic formulation of the optimization problem based in a generalization of the Traveling Salesman Problem.

Ranking the candidates with Vaxijen. The bioinformatics tool Vaxijen[36] was used with each one of the candidates obtained in the optimization with CPLEX. "Virus" was selected as the target organism. The overall prediction for the protective antigen was calculated for each sequence and those over the threshold of 0.4 established for this model were selected, which is the threshold of the highest accuracy value beyond which the sequence is considered to be a probable antigen.

Peptides. The sequence of different peptides from the Spike protein of SARS-CoV-2 were confirmed according to the published reference cryo-EM structure of the protein[38]. The peptide comprising amino acids 50–71 of the Spike protein with the sequence STQDLFLPFFSNVTWFHAIHVS, which is contained in the NTD domain (here designated as CoVPSA), as well as a non-related peptide of the same length selected from another pathogen unrelated to SARS-CoV-2, namely the Lmo 2459 virulence factor of *L. monocytogenes*, here referred as control peptide (CONT) with the sequence MTVKVGINGFGRIGRLAFRRIQ, or the RBD peptide of SARS-CoV2 Spike protein S1 from amino acids 330–530, were all synthesized by Genescript with a purity ≥ 99% by HPLC.

Preparation and safety of DC vaccines loaded with COVID-19 peptides. Bone-marrow-derived DCs were obtained from femurs of 8–12-week-old female mice and were cultured at 2×10^6 cells/mL in six-well plates in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 20% fetal calf serum, 1 mM glutamine, 1 mM nonessential amino acids, 50 µg/mL gentamicin, 30 µg/mL vancomycin (DMEM complete medium), and 20 ng/mL granulocyte–macrophage colony-stimulating factor (GM-CSF). On Day 7, cells were harvested and analyzed by fluorescence-activated cell sorting (FACS) to evaluate cell surface markers. Differentiated DCs showed a phenotype of 98% $CD11c^+MHC\text{-}II^+CD11b^{-/+}CD40^-CD86^-$ cells and were used in vivo for T cell responses.

The safety of DC vaccines loaded with peptides was explored with two assays to assess cell viability and apoptosis. Safety was considered optimal if the percentages of cell viability were higher than 95% and apoptosis induction was no higher than 6%–7%. Cell viability was explored after DC incubation with synthesized peptides at a concentration of 50 µg/mL for 16 hours, cells were then washed and stained with trypan blue. Results were expressed as the percentage viability compared with non-treated DCs of triplicate experiments ± SD ($P < 0.05$). Apoptosis was measured after labeling with the DNA fluorescent intercalating probe 7-aminoactinomycin D (7-AAD, BD Biosciences, San Jose, CA, USA) and cell surface analysis of the apoptotic marker, Annexin V conjugated with allophycocyanin (APC) fluorochrome, followed by incubation of DCs with peptides for 16 hours. Staining of DCs with 7-AAD corresponded to normal cell death and staining of DCs with annexin-V alone indicated the percentage of apoptotic cell death. The results were expressed as the mean ± SD ($P < 0.05$).

T cell responses elicited by DC-vaccines loaded with COVID-19 peptides. For DTH analysis, C57BL/6 mice that had been primed for 7 days intraperitoneally with COVID-19 peptides (50 µg/ml) were inoculated into the left hind footpad with DC vaccines (10^6 cells/mouse) pre-loaded with COVID-19 peptide or a control peptide protective against another unrelated infection (a 22-mer of the glyceraldehyde-3-phosphate-dehydrogenase of *L. monocytogenes*)[22]. DC vaccines were formulated in the presence of DIO-1 (2 µg/mL)[22]. Negative controls were the right hind footpads, since they were not inoculated. After 48 hours, the footpad thickness was measured with a caliper and the results were expressed in millimeters as the mean of three different experiments. Next, the popliteal lymph nodes were collected and homogenized, and cell homogenates were passed through cell strainers to analyze $CD4^+$ and $CD8^+$ T cells by flow cytometry. The results are expressed as the percentage of positive cells ± SD.

Cytokine measurements. Cytokines in mice sera, DCs, or mouse DC supernatants were quantified using multiparametric Luminex kits. In brief, interferon gamma (IFN-γ), IL-2, IL-4, IL-6, IL-10, IL-12 (p70), IL-17A, KC/CXCL1, MIP-2, and TNF-α levels in mice serum samples were quantified using the Luminex 200 platform with a magnetic system (Milliplex MAP Mouse High Sensitivity T Cell Magnetic Bead Panel, EMD Millipore Corporation, Billerica, MA, USA) following the manufacturer's instructions. Cytokine concentrations are expressed as the average of three replicates in pg/mL ± SD. Similarly, cytokines in the human sera of COVID-19 patients, and vaccinated or control volunteers were quantified using the multiparametric Luminex kit (Milliplex human 1xl HSTCMAG-28SK including the following cytokines: IFN-γ, IL-10, IL-17A, IL-2, IL-4, IL-6, IL-8, TNF-α, EMD Millipore Corporation) following the manufacturer´s instructions.

ELISA measurements of antibodies. Neutralization COVID-19 IgG antibodies were measured with the cPass™ kit from GenScript supplied by IES Medical (Leioa, Bizkaia, https://www.iesmedical.es) that measures the percentage of antibodies able to neutralize the SARS-CoV-2 virus in human sera. Antibodies of IgG isotype against the peptides RBD, CoVPSA, or CONT were assessed according to previously reported methods[40]. In brief, ninety-six well plates were coated with the different peptides (RBD and CoVPSA from the Spike protein and CONT from the Lmo2459 virulence factor) at a concentration of 50 µg/mL in carbonate buffer (pH 8.0) at 4ºC overnight, and were then washed and incubated with 1 mg/mL of BSA (fraction V) for blocking non-specific sites. The sera of COVID-19 patients with non-active infection (four patients: COV-1: NAT, COV-2: MAR, COV-3: AMA, COV-4: MABR), vaccinated volunteers with different COVID-19 vaccines (three with Pfizer and one with Moderna) (four volunteers: VAC-1: HEC, VAC-2: DAV, VAC-3: CAR, VAC-4: EDU), and healthy donors that were non-infected, non-vaccinated, and tested negative in a COVID-19 antigen test (four volunteers: CONT-1: CAD, CONT-2: AND, CONT-3: VIC, CONT-4: EFG) were 1/10 diluted and peptide-coated plates were incubated with the diluted sera for two hours at room temperature. Reactions were developed with goat anti-human IgG and the absorbances analyzed at 450 nm and expressed as optical units (OD) from the mean values ± SD of triplicate experiments ($P < 0.05$). Only results with OD ≥ 0.2 ± 0.01 were considered positive in Fig. 4.

Statistical analysis. For statistical analysis, the Student´s *t*-test was applied to all mouse assays and ELISA assays. All samples were evaluated in triplicate and experiments were performed at least three times. GraphPad software was used for the generation of all graphs presented.

Ethics statement. This study was carried out in accordance with the Guide for the Care and Use of Laboratory Animals of the Spanish Ministry of Science, Research and Innovation. The Committee on the Ethics of Animal Experiments of the University of Cantabria approved the protocol (Permit number: PI-10-17) that follows the Spanish legislation (RD 1201/2005). All surgeries were performed by cervical dislocation, and all efforts were made to minimize animal suffering. For the human data from COVID-19 patients and vaccinated non-infected donors, the study was approved by the Ethical Committee of Clinical Research of Cantabria at Instituto de Investigación Marqués de Valdecilla (Santander, Spain), reference number Acta 13/2021. All participants signed the Informed Consent documents and received and Information Document about the project. These documents are in the custody of physicians in accordance with Spanish Law (Ministry of Health). Similarly, for the use of human data on sera from COVID-19 patients, vaccinated volunteers, and healthy donors, all participants signed an Informed Consent form and received a General Project Information document, following approval from the Committee of Clinical Ethics of Cantabria (CEm) entitled: Trained immunity in the design of COVID-19 nanovaccines (reference INNVAL20-01).

Results

As described in the "Methods" section, the problem of finding optimal weighted λ–superstrings with a maximum value of λ for a given length that can serve as CVs against SARS-CoV-2 was solved using an Integer Programming algorithm (Methods, "Optimization with CPLEX"). This algorithm was fed with three elements, the host string set H, the target string set T and the weighting function w, obtained as follows:

First, the set H of host strings was taken as the 22 distinct sequences corresponding to the Surface protein of SARS-COV-2 that appear in the Genbank[28] and GISAID[29] databases (Methods, "Extraction of the sequences").

Next, we took as the set T of target strings (i.e., as potential epitopes), the 9-mers contained in some of the 22 host strings in positions corresponding to residues before the transmembrane domain.

Then, we assessed the weights of the epitopes using a function w in which the estimation of their immunogenicities and the estimation of the binding affinity to HLA-I was taken into account (Methods, "Weighting of the epitopes").

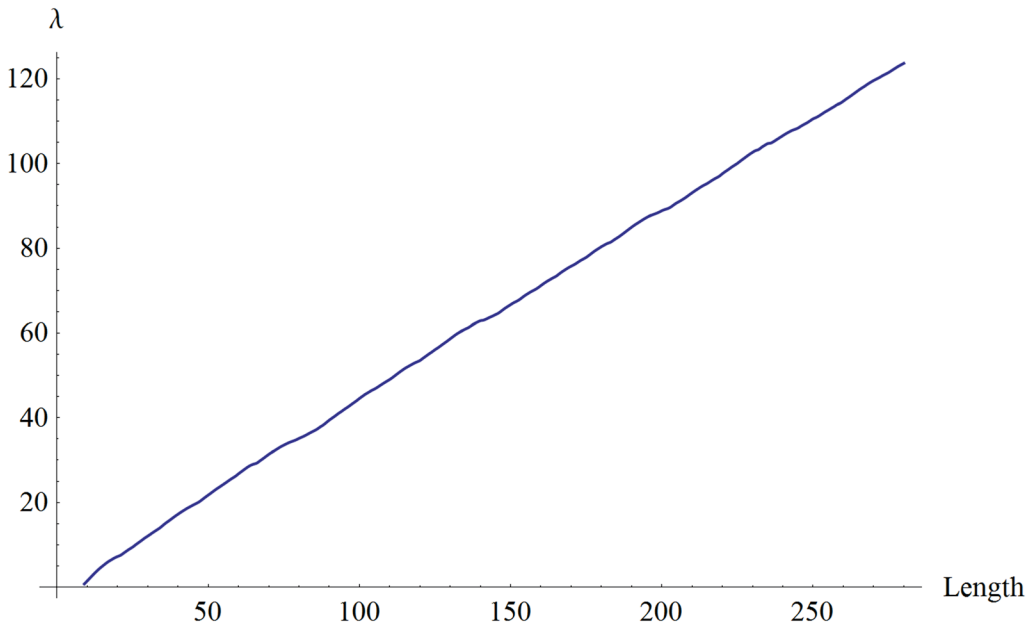

Figure 1. Scatterplot for λ. The abscissa axis shows the length of the candidate peptides, the ordinate axis shows the value of λ.

	Estimate	Standard error	P-value
1	−0.579005	0.0815211	1.08742×10^{-11}
x	0.446982	0.000495704	1.07761×10^{-471}

Table 1. Inference for the intercept and slope constants.

Then, we used the algorithm to calculate a weighted λ–superstring with maximum λ for each length between 9 and 280. A scatterplot for the value of λ as a function of the length of the CV is shown in Fig. 1. It can be well fitted by a least square line with the regression line $\lambda = -0.579005 + 0.446982 \cdot l$, where l is the length of the candidate. The intercept and slope of the line were accurately determined, with a low standard error and a low P-value, as shown in Table 1. The R-squared value of the fit was 0.999668, and the closeness of this value to 1 indicates a good fit. Thus, each one-unit increase in length is associated approximately with an increase of 0.4 in λ all along the range from 9 to 280. Therefore, the map is robust and there is no significant loss in the λ increase per unit length in the considered interval of lengths.

Furthermore, we calculated the VaxiJen overall prediction for each CV (Methods, "Ranking the candidates with Vaxijen"). These optimal weighted λ-superstring, as well as the corresponding λ values and VaxiJen predictions, are shown in the table in Additional file 3. The threshold of 0.4 indicated in VaxiJen for the viral model was surpassed by the candidates with lengths of 22,24,67,68,69,70, and 175, as well as those with a length of at least 184 amino acids (candidates shown in green in the above-mentioned table).

Each λ-superstring can be naturally divided as it constitutes a union of a small number of peptides located in different regions of the protein. These peptides are enumerated, for each λ-superstring, in the fourth column of the table. When a peptide has some intersection with a domain of the protein, the domain is annotated next to the peptide. For the λ-superstrings with lengths from 176 to 183, the third peptide intersects two domains, namely NTD and RBD, and for those with lengths from 237 to 247, the fourth peptide also intersects the same two domains.

The two CVs with the maximum value in the VaxiJen overall predictions are shown in Table 2. The first CV (22 amino acids in length) is contained in the NTD domain, and the second CV (277 amino acids in length) can be divided, as previously described, as it originates a multipeptide of five peptides. In particular, the third and fifth peptides intersect the NTD and RBD domains, respectively, making them appropriate targets for vaccine development against SARS-CoV-2[27].

We selected the first CV in Table 2 for further biological assays because it showed the maximum value in the overall prediction in VaxiJen. This peptide STQDLFLPFFSNVTWFHAIHVS is 22 amino acids in length, and is contained in the NTD domain, therefore being a valid candidate antigen for vaccine development[27,41,42], with an overall prediction of 0.5545 in VaxiJen.

After analyzing our results computationally, we synthesized the 22-amino acid SARS-CoV-2-NTD peptide (designated here as CoVPSA) and performed in vivo experiments to test its immunogenicity and putative efficiency.

Length	Lambda	Prediction	Sequence
22	8.03	0.5545	1:STQDLFLPFFSNVTWFHAIHVS(NTD)
277	122.44	0.5190	1:QSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRN-FYEPQIITTDNTFVSGNC(CD) 2:TEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQ 3:DLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGT-TLDSK(NTD) 4:FLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSN 5:FRVQPTESIVRFPNITNLCPFGEVFNATRFASVYA(RBD)

Table 2. Optimal weighted λ–superstring, λ values, and VaxiJen overall prediction for the two candidate vaccines with the maximum λ value. Column 1, the number of amino acids in the CVs; column 2, the value of λ; column 3, the VaxiJen overall prediction for antigenicity; column 4, the peptides whose union forms the λ-superstring.

Condition[a]	Cell viability[b]	Apoptosis[c]
NT	99 ± 0.3%	3 ± 0.3%
CoVPSA	98 ± 0.2%	4 ± 0.2%
DC-peptide CONT	97 ± 0.5%	4.5 ± 0.4%

Table 3. Safety of DC vaccines loaded with COVID-19 peptides. [a]DC cells were incubated with 50 μg/mL of peptides for 16 hours. [b]Cell viability was explored after trypan blue staining and microscopy was used to count viable (non-stained) and non-viable (blue stained) cells. Results are expressed as the percentage of viable versus total cells (viable and non-viable cells). [c]Apoptosis is detected after DC staining with the DNA marker 7-AAD-PE and the apoptotic marker Annexin-V-APC. Results show the percentage of apoptotic cells ± SD of triplicate samples. All experiments were performed at least three times.

A first proof of concept was related with the immunogenicity of CoVPSA, and it was determined using a previously described procedure[22] that evaluates the best immunogenic epitopes for preparing vaccines.

Safety was also examined by a cell viability assay after Trypan blue staining and apoptosis induction. Safety for DC vaccine vectors is considered as a percentage of cell viability higher than 95% and apoptosis induction lower than 7%–8%. Table 3 shows that DC vaccines loaded with CoVPSA peptides, or the unrelated bacterial peptide used as negative control, presented 98%–99% cell viability and lower than 4%–5% apoptosis. Therefore, we concluded that the DC vaccines loaded with peptides presented good safety profiles.

Next, we performed immunogenicity assays to measuring the DTH response of the vaccine vector. DCs were loaded with the peptide, then, mice were primed for 7 days intraperitoneally with COVID-19 peptides and then inoculated with the vaccine formulation (DC-CoVPSA) into the left hind footpads, with the non-inoculated right hind footpads acting as basal controls. Forty-eight hours later, we measured the DTH response as the swelling of the left hind footpads compared with the right hind footpads.

We also included empty DCs in these experiments and DCs loaded with a bacterial peptide unrelated to SARS-CoV-2 but with high CV efficiency against the bacterial pathogen[22]. Analysis of DTH responses (blue bars in Fig. 2) indicated that DCs loaded with COVID-19 designed peptide (DC-CoVPSA bars) elicited significantly stronger immune responses than DCs loaded with the control bacterial peptide (DC-CONT bars) or empty DCs (DC labelled bars). This may be explained by the fact that mice primed and DC-vaccinated with the same COVID-19 peptide produced high DTH responses, while mice primed and DC-vaccinated with different peptides were not able to elicit significant DTH responses (DC-control in Fig. 3). Next, we collected the popliteal lymph nodes and cultured them in vitro with 1 μg/mL of each peptide, CoVPSA, control peptide, or saline for 72 hours to examine the main immune cell populations by flow cytometry.

We observed that the highest percentages of immune cells corresponded to $CD19^+$ cells (25,63%) that usually correspond to B cells, followed by MHC-II^+ cells (27, 45%) that usually label DCs and macrophages, next $CD4^+$ T cells (10,39%) and $CD8^+$ T cells (14,61%).

The control peptide (CONT) produced no significant immune responses, as we observed only a small percentage of $CD19^+$ cells (5,3%) and moderate numbers of MHC-II^+ cells (13,5%) (DC-CONT bars in Fig. 2). Empty DCs (DC bars in Fig. 2) induced no significant numbers of immune cells.

These results indicated the clear induction of immune cells by DC vaccines loaded with CoVPSA peptide, with immune cells involved in antibody formation, such as B cells, DCs, and $CD4^+$ T cells, being stimulated. While not predominant, cytotoxic immune responses caused by $CD8^+$ T cells were also induced by DC vaccines loaded with CoVPSA peptide. These results were not surprising since $CD4^+$ and $CD8^+$ T cell epitopes are recovered from patients with mild and severe COVID-19 that are specific for the Spike protein[24].

A second proof of concept was related with the production of cytokines, either anti-viral cytokines, such as TNF-α, IFN-γ, IL-2, KC, and IL-12, or acute Th2 cytokines, such as IL-4, IL-6, MIP-2, or IL-10. The COVID-19 cytokine storm observed in patients with severe disease correlates with high levels of TNF-α, IL-6, IL-4, and IL-10, as well as with a clear deficiency in the production of IFN-related cytokines (i.e., IFN-α, IFN-γ, or IL-12)[43].

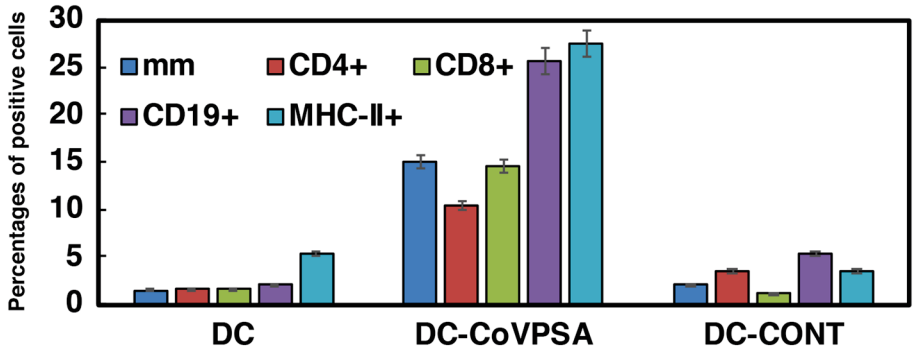

Figure 2. Immunogenicity of the CoVPSA peptide in vaccine platforms. The hind footpads of mice (C57BL/6, n = 5) were inoculated with the DC vaccine (10^6 cells/mouse) loaded with different peptides (COVID-CoVPSA or peptide control, CONT) or remained as empty DCs in formulations with the adjuvant DIO-1 (40 ng/mL). After 48 hours, footpad swelling was measured with a caliper (dark blue bars) and expressed as the difference in mm between left and right hind footpads. Results are the mean ± SD of three different experiments (*$P < 0.05$). Popliteal lymph nodes were then isolated from the legs of the mice and after homogenization, immune cell populations were analyzed by flow cytometry. The percentages of $CD4^+$ (red bars), $CD8^+$ T cells (green bars), CD19 + (B cells, purple bars), and MHC-II + positive cells, mainly DCs or macrophages (light blue bars), are shown. Results are expressed as the percentage of positive cells ± SD of three different experiments ($P < 0.05$). Experiments were performed six times.

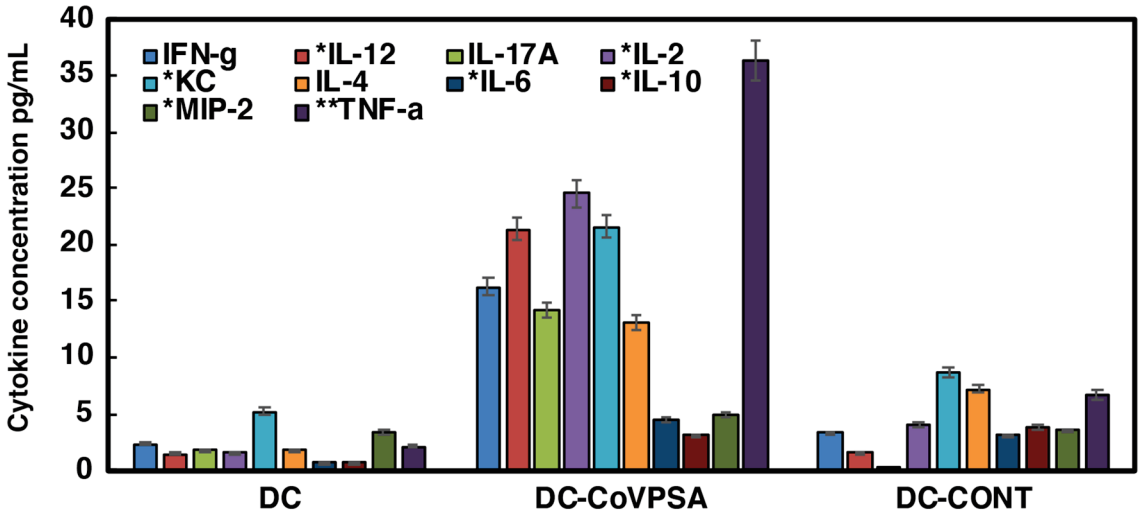

Figure 3. Cytokine levels of mice inoculated with DC vaccine platforms. Cytokine levels were detected in the sera of mice, as described in Fig. 2 , and were measured using a multiparametric Luminex kit from Merck. Results are expressed as pg/mL of each cytokine ± SD of triplicate samples ($P \leq 0.05$). Asterisk: Levels of cytokines should be multiplied tenfold. Double asterisk: Levels of cytokines should be divided twofold. Cytokine experiments were performed five times.

Our results in Fig. 3 show that DCs loaded with CoVPSA peptide produced mainly Th1-Th7 cytokines, IL-12, IL-17A, and IL-2. However, this DC-CoVPSA vaccine platform did not induce cytokines participating in the COVID-19 cytokine storm, such as IL-6, IL-10, or TNF-α (bars labelled with DC-CoVPSA in Fig. 3).

Interestingly, DC-CoVPSA vaccines induced high levels of IFN-γ (blue bars in Fig. 3) but barely detectable levels of MIP-2, an inflammatory cytokine that recruits inflammatory macrophages (grey bars in Fig. 3). The lack of significant levels of IL-4 (orange bars in Fig. 3) but high levels of IL-2 (red bars) strongly suggested the induction of Th1-Th17-type immune responses, but with no exacerbation of other cytokines, such as TFN-α or MIP-2.

In summary, the high levels of IFN-γ and especially IL-12, involved in vaccine efficiency and anti-viral responses, respectively, prompted us to suggest that CoVPSA peptide might function as an immunogenic epitope. CoVPSA peptide might be a good candidate to prepare vaccine platforms that induce not only antibody production but strong anti-viral T cell responses. We also confirmed these results in samples of human sera that served as a third proof of concept of our vaccine design. We recruited four asymptomatic patients with non-active COVID-19, four vaccinated volunteers (three with the Pfizer vaccine and one with the Moderna vaccine prepared against the RBD region of Spike protein), and four healthy donors that were non-vaccinated and tested negative in a COVID-19 antigen test. We collected blood from these 12 volunteers and compared the titers of different

Human Samples	[a]Percentages of neutral-RBD	[b]Anti-CoVPSA	[b]Anti-CONT	[b]Anti-RBD	[c]IFN-γ	[c]IL-17A	[c]IL-2	[c]IL-4	[c]IL-6	[c]IL-8
COV-1	54% ± 0.4	0.38 ± 0.05	0.1 ± 0.01	0.14 ± 0.01	#7,97 ± 0.1	7.61 ± 0.2	1.66 ± 0.1	*191.7 ± 0.4	*25.29 ± 0.2	*36.68 ± 0.2
COV-2	70% ± 0.6	0.31 ± 0.04	0.04 ± 0.02	0.59 ± 0.02	#8.25 ± 0.1	5.56 ± 0.2	1.32 ± 0.1	*76.7 ± 0.2	8.15 ± 0.1	*15.89 ± 0.2
COV-3	95% ± 0.8	0.28 ± 0.02	0.08 ± 0.03	0.44 ± 0.03	#9.89 ± 0.1	5.06 ± 0.1	2.40 ± 0.1	*311.7 ± 0.2	*105.5 ± 0.3	*196.1 ± 0.3
COV-4	26% ± 0.4	1.01 ± 0.04	0.10 ± 0.02	0.74 ± 0.02	#6.40 ± 0.1	3.52 ± 0.1	1.29 ± 0.1	*137.1 ± 0.4	*49.4 ± 0.3	*182.25 ± 0.3
VAC1	94% ± 0.7	0.28 ± 0.02	0.11 ± 0.03	0.44 ± 0.02	22.0 ± 0.2	15.01 ± 0.2	4.42 ± 0.1	17.65 ± 0.2	3.10 ± 0.1	6.30 ± 0.1
VAC2	95% ± 0.6	0.37 ± 0.02	0.11± 0.01	0.45 ± 0.03	17.8 ± 0.2	10.72 ± 0.2	4.13 ± 0.1	10.99 ± 0.2	1.11 ± 0.1	4.26 ± 0.1
VAC3	95% ± 0.5	0.28 ± 0.03	0.07 ± 0.01	0.37 ± 0.02	13.8 ± 0.1	8.5 ± 0.2	3.64 ± 0.1	23.21 ± 0.2	2.48 ± 0.1	1.37 ± 0.1
VAC4	95% ± 0.4	0.53 ± 0.01	0.07 ± 0.01	0.34 ± 0.2	11.21 ± 0.2	5.33 ± 0.2	2.55 ± 0.1	10.90 ± 0.2	0.69 ± 0.1	1.31 ± 0.1
CONT1	6% ± 0.5	0.12 ± 0.01	0.01 ± 0.02	0.08 ± 0.01	4.23 ± 0.1	2.16 ± 0.1	0.2 ± 0.1	5.51 ± 0.1	0.4 ± 0.1	2.1 ± 0.1
CONT2	20% ± 0.7	0.16 ± 0.02	0.1 ± 0.01	0.07 ± 0.02	8.50 ± 0.2	2.55 ± 0.1	0.27 ± 0.1	7.11 ± 0.2	1.17 ± 0.1	3.48 ± 0.1
CONT3	25% ± 0.8	0.18 ± 0.01	0.07 ± 0.03	0.04 ± 0.01	8.22 ± 0.1	4.43 ± 0.1	0.12 ± 0.1	6.23 ± 0.1	2.99 ± 0.1	7.43 ± 0.2
CONT4	24% ± 0.5	0.20 ± 0.02	0.02 ± 0.01	0.02 ± 0.01	7.61 ± 0.1	3.54 ± 0.1	0.67 ± 0.1	5.97 ± 0.1	2.31 ± 0.1	4.45 ± 0.2

[a]Neutralization kit for RBD antibodies (IES Medical) expressed as percentages
[b]ELISA peptide expressed as $OD_{450} \geq 0.2 \pm 0.01$ (anti-CoVPSA, anti-CONT and anti-RBD)
[c]Citokine concentrations expressed as pg/mL
*Cytokine storm concentrations
#IFN-g low concentration

Figure 4. Correlation between neutralization IgG antibodies, anti-CoVPSA antibodies, and Th1-Th17 cytokines in COVID-19-vaccinated volunteers. The sera of 12 volunteers (four asymptomatic non-active COVID-19 patients: COV-1, COV-2, COV-3, and COV-4; four vaccinated volunteers who tested negative for SARS-CoV-2: VAC-1, VAC-2, and VAC-3 who received the Pfizer vaccine, and VAC-4 who received the Moderna vaccine; four control donors who were unvaccinated and tested negative for SARS-CoV-2: CONT1, CONT2, CONT3, and CONT4) were tested for IgG neutralization antibodies (IES Medical kit, first column), IgG anti-CoVPSA antibodies (second column), and IgG anti-RBD antibodies (third column). Reactions were developed with goat anti-human IgG and the absorbances (OD) were analyzed at 450 nm. Results are expressed as percentages of neutralization ± SD ([a]column), or as the mean ± SD of triplicate data.

IgG COVID-19 antibodies as follows: (i) IgG antibodies able to neutralize the virus (neutral-RBD column in Fig. 4) were assessed by a neutralization antibody assay, (ii) anti-RBD antibodies that correspond to IgG antibodies recognizing the whole Spike protein including the RBD region binding to the ACE2 receptor (anti-RBD column), and (iii) anti-CoVPSA antibodies that reflect the IgG antibodies against our designed peptide in the NTD region (anti-CoVPSA column). As expected, COVID-19 asymptomatic patients presented medium and varied titers of IgG viral neutralization antibodies and low but significant levels of whole Spike protein anti-RBD antibodies, as previously reported[44–46]. These COVID-19 patients also presented medium anti-CoVPSA IgG antibody titers (compare columns 1, 2 and 4 of Fig. 4). Interestingly, volunteers vaccinated with mRNA vaccines prepared against the RBD region of the Spike protein presented not only the highest titers of IgG viral neutralization antibodies that correlated with significant antibody titers against the RBD region[44–46], but also significant responses against the CoVPSA peptide. Analyses of the cytokine concentrations in the sera of these volunteers indicated that COVID-19 patients presented a storm cytokine pattern with high levels of IL-4, IL-6 and IL-8 and low or insignificant levels of IFN-γ, as previously published[44–50]. Interestingly, vaccinated volunteers presented high levels of IFN-γ as well as significant levels of IL-2 and IL-4. This indicates that mRNA vaccines induce good antibody responses, as well as significant anti-viral cellular responses, as measured with neutralizing anti-RBD antibodies, antibodies against other Spike protein regions, such as the NTD region, and high levels of anti-viral cytokines, such as IFN-γ, IL-17A and IL-2.

Discussion

In this work, we established a proof of concept for our computational vaccine design method using λ-superstrings, and we demonstrated the feasibility of this method to obtain effective CVs.

Unlike previous studies in the medical literature, we did not start from a single genome in our analysis, instead we considered several genomes corresponding to different mutated versions. Furthermore, unlike most of the vaccines currently developed, our candidate is a peptide vaccine that does not consider the entire Spike protein, but rather a set of computationally selected overlapping epitopes.

In our study, we first found a map of CVs against SARS-CoV-2 targeted to the Spike protein, in which the length of the candidates ranged from 9 to 280 amino acids. Then, we filtered the candidates of this map to those considered protective antigens according to the VaxiJen bioinformatics tool, and finally synthesized the candidate peptide with the highest value from the overall prediction. This peptide was 22 amino acids in length and comprised the sequence STQDLFLPFFSNVTWFHAIHVS, which is contained in the NTD domain of the protein.

To experimentally validate the viability of the candidate, we performed several in vivo assays. The result of these experiments was positive for the following reasons:

(1) The selected CV elicited a robust immune response in mice with Th1-Th17 pro-inflammatory features and strong stimulation of cells involved in antibody production.
(2) The selected CV elicited the production of anti-viral cytokines. The cytokine profile obtained was adequate because it mainly involved anti-viral cytokines IL-12, IL-17A, and IL-2, not cytokines participating in the COVID-19 cytokine storm, such as IL-6 or IL-10.
(3) The DC vaccines loaded with peptides showed good safety profiles.
(4) mRNA vaccines recognizing the RBD region of Spike protein induce a wide anti-Spike protein immune response, since we also observed significant levels of antibodies against other parts of the Spike protein, such as the NTD region (the CoVPSA peptide).
(5) Antibodies against the selected CV peptide, CoVPSA, were detected at good titers in COVID-19 patients, although the profiles varied (COV-4 patient was a good responder, whereas COV-1, COV-2 and COV-3 patients were medium responders, column 2 in Fig. 4). Interestingly, the vaccinated group that was inoculated with the RBD antigen (RNA against the SARS-CoV-2 RBD region) and not with the CoVPSA antigen, also presented significant levels of antibodies to anti-CoVPSA antigen (VAC-4 volunteer was a good responder, whereas VAC-1, VAC-2 and VAC-3 volunteers were medium responders, column 2 in Fig. 4). This supported the conclusion that CoVPSA might be a relevant antigen to incorporate in COVID-19 vaccines to induce high levels of anti-viral cytokines, such as IFN-γ, IL-17A, and IL-2.
(6) COVID-19-vaccinated volunteers presented high titers of anti-COVID-19 neutralization antibodies, varied responses to all regions of the Spike protein (RBD or NTD (CoVPSA peptide) regions), and anti-viral Th1-Th17 cytokines. These three features are characteristic of efficient vaccines.

Taken together, these results confirm the use of λ-superstrings as an effective means of detecting feasible vaccines against SARS-CoV-2.

It is worth noting that we also showed that peptides targeting the NTD region of the Spike protein, which were already known to be a good target for antibodies[48–50], can also induce potent cellular immunity. However, our techniques selected specific parts of the NTD region based on our combinatorial optimization method that considers all of the generated variants at once, and not the complete NTD region. Moreover, the selection of this area was not premeditated, but the respective weights of the peptides and the consideration of the whole virus variant directed the algorithm to choose epitopes in that region.

In future studies, it would be worthwhile testing the second CV in Table 2, which is a multipeptide, and also the other CVs indicated in green in Additional file 3, which are those that pass the threshold established in VaxiJen to be considered as probable antigens. This would increase the range of potentially effective vaccines against SARS-CoV-2 that would cover a high percentage of the population. The uncertainty surrounding the future effectiveness of currently available vaccines means that such endeavors may prove valuable.

In summary, we have proven that our methodology for designing CVs that utilizes λ-superstrings represents an efficient alternative approach to peptide-based vaccine design for SARS-CoV-2. This may aid the design of further safe and efficient COVID-19 vaccines in the near future.

Data availability

The datasets supporting the conclusions of this article are included within the article and its additional files.

Received: 20 October 2021; Accepted: 7 March 2022
Published online: 19 April 2022

References

1. https://covid19.who.int. Accessed 01 Apr 2022.
2. Awadasseid, A., Wu, Y., Tanaka, Y. & Zhang, W. Current advances in the development of SARS-CoV-2 vaccines. *Int. J. Biol. Sci.* **17**(1), 8–19. https://doi.org/10.7150/ijbs.52569 (2021).
3. Flanagan, K. L. *et al.* Progress and pitfalls in the quest for effective SARS-CoV-2 (COVID-19) vaccines. *Front. Immunol.* **11**, 579250. https://doi.org/10.3389/fimmu.2020.579250 (2020).
4. Gaebler, C. & Nussenzweig, M. C. All eyes on a hurdle race for a SARS-CoV-2 vaccine. *Nature* **586**, 501–502. https://doi.org/10.1038/d41586-020-02926-w (2020).
5. Krammer, F. SARS-CoV-2 vaccines in development. *Nature* **586**, 516–527. https://doi.org/10.1038/s41586-020-2798-3 (2020).
6. Poland, G. A., Ovsyannikova, I. G., Crooke, S. N. & Kennedy, R. B. SARS-CoV-2 vaccine development: Current status. *Mayo Clin. Proc.* **95**(10), 2172–2188. https://doi.org/10.1016/j.mayocp.2020.07.021 (2020).
7. Wang, F., Kream, R. M. & Stefano, G. B. An evidence based perspective on mRNA-SARS-CoV-2 vaccine development. *Med Sci. Monit.* **26**, e924700-1-e924700-8. https://doi.org/10.12659/MSM.924700 (2020).
8. Yi, C., Yi, Y. & Li, J. mRNA vaccines : Possible tools to combat SARS-CoV-2. *Virol. Sin.* **35**(3), 259–262. https://doi.org/10.1007/s12250-020-00243-0 (2020).
9. Subbarao, K. COVID-19 vaccines: Time to talk about the uncertainties. *Nature* https://doi.org/10.1038/d41586-020-02944-8 (2020).
10. Chakraborty, S., Mallajosyula, V., Tato, C. M., Tan, G. S. & Wang, T. T. SARS-CoV-2 vaccines in advanced clinical trials: Where do we stand?. *Adv. Drug Deliv. Rev.* **172**, 314–338. https://doi.org/10.1016/j.addr.2021.01.014 (2021).
11. Estrada, E. COVID-19 and SARS-CoV-2, modeling the present, looking at the future. *Phys. Rep.* **869**, 1–51. https://doi.org/10.1016/j.physrep.2020.07.005 (2020).
12. Sikora, M. *et al.* Computational epitope map of SARS-CoV-2 spike protein. *PLoS Comput. Biol.* **17**(4), e1008790. https://doi.org/10.1371/journal.pcbi.1008790 (2021).
13. Jaiswal, G. & Kumar, V. In-silico design of a potential inhibitor of SARS-CoV-2 S protein. *PLoS ONE* **15**(10), e0240004. https://doi.org/10.1371/journal.pone.0240004 (2020).

14. Baig, M. S., Alagumuthu, M., Rajpoot, S. & Saqib, U. Identification of a potential peptide inhibitor of SARS-CoV-2 targeting its entry into the host cells. *Drugs R D.* **20**, 161–169. https://doi.org/10.1007/s40268-020-00312-5 (2020).
15. Chen, J. *et al.* Rational optimization of a human neutralizing antibody of SARS-CoV-2. *Comput. Biol. Med.* **135**, 104550. https://doi.org/10.1016/j.compbiomed.2021.104550 (2021).
16. Kar, T. *et al.* A candidate multi-epitope vaccine against SARS-CoV-2. *Sci. Rep.* **10**(1), 10895. https://doi.org/10.1038/s41598-020-67749-1 (2020).
17. Saha, R., Ghosh, P. & Prasad Burra, V. L. S. Designing a next generation multi-epitope based peptide vaccine candidate against SARS-CoV-2 using computational approaches. *3 Biotech.* **11**(2), 47. https://doi.org/10.1007/s13205-020-02574-x (2021).
18. Ong, E., Huang, X., Pearce, R., Zhang, Y. & He, Y. Computational design of SARS-CoV-2 spike glycoproteins to increase immunogenicity by T cell epitope engineering. *Comput. Struct. Biotechnol. J.* **19**, 518–529. https://doi.org/10.1016/j.csbj.2020.12.039 (2021).
19. Chukwudozie, O. S. *et al.* Immuno-informatics design of a multimeric epitope peptide based vaccine targeting SARS-CoV-2 spike glycoprotein. *PLoS ONE* **16**(3), e0248061. https://doi.org/10.1371/journal.pone.0248061 (2021).
20. Martínez, L. *et al.* A combinatorial approach to the design of vaccines. *J. Math. Biol.* **70**(6), 1327–1358. https://doi.org/10.1007/s00285-014-0797-4 (2015).
21. Martínez, L. *et al.* Weighted lambda superstrings applied to vaccine design. *PLoS ONE* **14**(2), e0211714. https://doi.org/10.1371/journal.pone.0211714 (2019).
22. Calderón-González, R. *et al.* Identification and characterization of T-cell epitopes for incorporation into dendritic cell-delivered Listeria vaccines. *J. Immunol. Methods.* **424**, 111–119. https://doi.org/10.1016/j.jim.2015.05.009 (2015).
23. Dong, Y. *et al.* A systematic review of SARS-CoV-2 vaccine candidates. *Signal Transduct. Target Ther.* **5**(1), 237. https://doi.org/10.1038/s41392-020-00352-y (2020).
24. Peng, Y. *et al.* Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. *Nat. Immunol.* **21**(11), 1336–1345. https://doi.org/10.1038/s41590-020-0782-6 (2020).
25. Hodgson, S. H. *et al.* What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. *Lancet Infect. Dis.* **S1473–3099**(20), 30773–30778. https://doi.org/10.1016/S1473-3099(20)30773-8 (2020).
26. Samrat, S. K., Tharappel, A. M., Li, Z. & Li, H. Prospect of SARS-CoV-2 spike protein: Potential role in vaccine and therapeutic development. *Virus Res.* **288**, 198141. https://doi.org/10.1016/j.virusres.2020.198141 (2020).
27. Zhang, J. *et al.* Progress and prospects on vaccine development against SARS-CoV-2. *Vaccines (Basel).* **8**(2), 153. https://doi.org/10.3390/vaccines8020153 (2020).
28. https://www.ncbi.nlm.nih.gov/genbank/. Accessed 03 Apr 2020.
29. https://platform.gisaid.org/epi3/frontend#3898c7. Accessed 03 Apr 2020.
30. http://tools.iedb.org/immunogenicity/. Accessed 03 Apr 2020.
31. http://tools.iedb.org/mhci/. Accessed 03 Apr 2020.
32. http://www.allelefrequencies.net/. Accessed 03 Apr 2020.
33. https://help.iedb.org/hc/en-us/articles/114094151851. Accessed 03 Apr 2020.
34. Sauer, K. & Harris, T. An effective COVID-19 vaccine needs to engage T cells. *Front. Immunol.* **11**, 581807. https://doi.org/10.3389/fimmu.2020.581807 (2020).
35. Platten, M. & Offringa, R. Cancer immunotherapy: Exploiting neoepitopes. *Cell Res.* **25**(8), 887–888. https://doi.org/10.1038/cr.2015.66 (2015).
36. http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html. Accessed 03 Apr 2020.
37. https://www.ebi.ac.uk/Tools/psa/genewise/. Accessed 03 Apr 2020.
38. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**(6483), 1260–1263. https://doi.org/10.1126/science.abb2507 (2020).
39. https://www.ibm.com/analytics/cplex-optimizer. Accessed 03 Apr 2020.
40. Teran-Navarro, H. *et al.* A comparison between recombinant *Listeria* GAPDH proteins and m-RNA encoded GAPDH conjugated lipids as cross-reactive vaccines protecting against *Listeria, Mycobacterium* and *Streptococcus*. *Front. Immunol.* **12**, 632304. https://doi.org/10.3389/fimmu.2021.632304 (2021).
41. Belete, T. M. A review on promising vaccine development progress for COVID-19 disease. *Vacunas* **21**(2), 121–128. https://doi.org/10.1016/j.vacun.2020.05.002 (2020).
42. Chaudhry, S. N. *et al.* New insights on possible vaccine development against SARS-CoV-2. *Life Sci.* **260**, 118421. https://doi.org/10.1016/j.lfs.2020.118421 (2020).
43. Zhang, Y., Chen, Y. & Meng, Z. Immunomodulation for severe COVID-19 pneumonia: The state of the art. *Front. Immunol.* **11**, 577442. https://doi.org/10.3389/fimmu.2020.577442 (2020).
44. Trougakos, I. P. *et al.* Comparative kinetics of SARS-CoV-2 anti-spike protein RBD IgGs and neutralizing antibodies in convalescent and naive recipients of the BNT162b2 mRNA vaccine versus COVID-19 patient. *BMC Med.* **19**(1), 208. https://doi.org/10.1186/s12916-021-02090-6 (2021).
45. Favresse, J. *et al.* Neutralizing antibodies in COVID-19 patients and vaccine recipients after two doses of BNT162b2. *Viruses* **13**(7), 1364. https://doi.org/10.3390/v13071364 (2021).
46. Alvarez, C., & Soriano, V. COVID-19 prevention and vaccines. in *Challenges in the Pandemic: A Multidisciplinary Approach.* (eds. Varon, J.V., Marik, P., Iglesias, J., de Souza, C.). ISBN: 978-93-90553-42-6. (Thieme Medical and Sciences Publishers Private Limited, 2021)
47. Ward, J. D., Cornaby, C. & Schmitz, J. L. Indeterminate QuantiFERON gold plus results reveal deficient interferon gamma responses in severely Ill COVID-19 patients. *J. Clin. Microbiol.* **59**(10), e0081121. https://doi.org/10.1128/JCM.00811-21 (2021).
48. Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the spike protein of SARS-CoV-2. *Science* **369**(6504), 650–655. https://doi.org/10.1126/science.abc6952 (2020).
49. Jiang, S., Zhang, X., Yang, Y., Hotez, P. J. & Du, L. Neutralizing antibodies for the treatment of COVID-19. *Nat. Biomed. Eng.* **4**(12), 1134–1139. https://doi.org/10.1038/s41551-020-00660-2 (2020).
50. Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**(7821), 450–456. https://doi.org/10.1038/s41586-020-2571-7 (2020).

Acknowledgements

We acknowledge the participation of four COVID-19 patients (MAT, MAR, AMA, MABR), four vaccinated volunteers (HEC, DAV, CAR, EDU), and four healthy donors (CAD, AND, VIC, EFG) in this study and their provision of blood and COVID-19 antigen test samples. The authors thank for technical and human support provided by SGIker (UPV/EHU/ERDF, EU).

Author contributions

L.M. conceived, planned, and directed the design of the candidate vaccines. I.M. participated in the planning of the algorithms and in the use of the bioinformatics tools. I.M.D. participated in the interpretation of the results. D.S.C. participated in the inoculation of mice with the candidate vaccines and helped with the experimental analysis. H.T.N. participated in the execution of the experimental analysis of the candidate vaccines. A.Z. participated in the execution of the experimental analysis of the candidate vaccines. S.A. participated in the acquisition and interpretation of genetic data. E.G.L. participated in measurement of cytokines in mice sera. J.G.O.V. designed and analyzed the results of cytokines in mice sera. M.G.M. participated in the execution of the experimental analysis of the candidate vaccines. J.C.M. participated in the execution of the experimental analysis of the candidate vaccines. C.A.D. conceived, planned, and directed the experimental analysis of the candidate vaccines. All authors participated in discussions about the article and in the writing of the manuscript.

Funding

Luis Martínez and Iker Malaina were supported by the Basque Government, grants IT974-16 and KK-2018/00090 and by the UPV/EHU and Basque Center of Applied Mathematics, grants US18/21 and US21/27. Carmen Alvarez-Dominguez was funded by the Instituto de Salud Carlos III, grants DTS18-00022 and PI19-01580, co-funded in part with European FEDER funds "*A new way of making Europe*", the Instituto de Investigación Marqués de Valdecilla, grant INNVAL20/01, and the COST European action ENOVA CA-16231. David Salcines-Cuevas was supported by a predoctoral contract for the BioHealth research program of the Cantabria government. Hector Teran-Navarro salary was supported by the Instituto de Investigación Marqués de Valdecilla, grant INNVAL19/26. Andrea Zeoli was an Erasmus student from the University of Milan "La Statale" (Milan, Italy) performing a stay at IDIVAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-09615-w.

Correspondence and requests for materials should be addressed to L.M. or C.A.-D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Analyzing the Immune Response of Neoepitopes for Personalized Vaccine Design

Iker Malaina[1(✉)], Leire Legarreta[1], Mª Dolores Boyano[2], Santos Alonso[3], Ildefonso M. De la Fuente[1,4], and Luis Martinez[1]

[1] Department of Mathematics, University of the Basque Country UPV/EHU, 48080 Bilbao, Spain
{iker.malaina,leire.legarreta, mtpmadei,luis.martinez}@ehu.eus
[2] Department of Cell Biology and Histology, University of the Basque Country UPV/EHU, Bilbao, Spain
lola.boyano@ehu.eus
[3] Department of Genetics, Anthropology and Animal Physiology, University of the Basque Country UPV/EHU, Bilbao, Spain
santos.alonso@ehu.eus
[4] Department of Nutrition, CEBAS-CSIC Institute, Espinardo University Campus, Murcia, Spain

Abstract. In the last few years, the importance of neoepitopes for the development of personalized antitumor vaccines has increased remarkably. This kind of epitopes are considered to generate a strong immune reaction, while their non-mutated version, which sometimes differs only in a single amino-acid, does not generate a response at all. In order to study if, regardless the immune tolerance, neoepitopes are quantitatively more immunogenic than the original strings, we have obtained samples of mutated and non-mutated epitopes of six patients with cutaneous melanoma in different stages, and then we have compared them. More precisely, we have used several bioinformatic tools to study certain properties of the epitopes such as the HLA binding affinity of classes I and II, and found that some of them are in fact increased in their mutated versions, which supports the hypothesis, and also reinforces the use of neoepitopes for cancer vaccine design.

Keywords: Neoepitope · Vaccine design · Bioinformatics · HLA immunogenicity

1 Introduction

In the last few years, personalized antitumoral vaccination has been increasingly proposed as a novel and encouraging approach to treat several types of cancers [1–4]. The main reasons for choosing personalized vaccination against cancer cells is that, on one hand, tumors contain a large amount of mutations, and on the other hand, it is that in a patient's tumor, approximately 95% of the mutations seem to be unique to that tumor [5]. Therefore, these mutations make ideal oncological targets for efficiently targeting individual tumors [6], and more precisely, for personalized vaccination.

I. Rojas et al. (Eds.): IWBBIO 2020, LNBI 12108, pp. 40–48, 2020.
https://doi.org/10.1007/978-3-030-45385-5_4

Nevertheless, even if the amount of mutations in tumors is considerable, in order to make an effective vaccine, first we need to distinguish between the mutations that only appear in the tumor, and the ones that happen in the rest of our non-oncogenic cells. Here is where we introduce the concept of neoepitope, which is a class of peptide which bounds to the Major Histocompatibility Complex (MHC, that in humans was denoted as HLA, referring to Human Leucocytic Antigen) and emerges from tumor-specific mutations [7]. This kind of epitopes have not been encountered before by the immune system, and as a consequence, the system will not apply the tolerance mechanisms against them [8].

However, even if targeting neoepitopes has resulted in clinical benefits [9, 10], when we consider the whole mutation spectrum of a tumor (known as mutanome), the amount of potential neoepitopes is too vast, and if we pick them blindly when we develop our vaccine, there is no guarantee of obtaining a highly immunogenic response. Generating an immune response against a mutated peptide depends directly on the capacity of the patient's HLA to bind the neoepitope and present it to lymphocytes [11], and therefore, selecting those epitopes which will bind more effectively the cells of the immune system seems like a reasonable first criterion.

In order to do this, and since the individual evaluation of every neoepitope in a tumor is too expensive in both time and cost, there have been developed several bioinformatic tools for *in silico* prediction of immunogenicity, HLA-1 and HLA-2 binding affinity, TAP transport, etc. [12–14]. These tools have been widely used to identify potential epitopes [15], but as a consequence an inevitable question arises: since a mutated peptide is capable of generating a strong immune response while their non-mutated version, which sometimes differs only in a single amino-acid (aa), does not generate a response at all, are these tools capable of quantitatively detect these differences? Or in other words, is it true that mutated versions of peptides are more immunogenic than the non-mutated ones, according to bioinformatic tools?

With the aim of giving response to this question, first, we have experimentally sequenced part of the mutanome of six patients affected with cutaneous melanoma. Cutaneous melanoma is a type of skin cancer that is located in the epidermis, and it arises from the pigment-containing cells known as melanocytes [16]. In Spain, the percentage of cases yearly increases by 7%, and happens more often in women (corresponding to the 2.7% of the cancers) than in men (where it represents the 1.5% of the male cancers). More importantly, besides being considered very invasive, and metastatic [17], this kind of cancer presents a lot of mutations, which makes it a good candidate for addressing our problem.

Secondly, after identifying the amino-acid sequence of the peptide corresponding to the detected DNA mutations, we have studied the predicted HLA binding affinity of classes I and II with the IEDB predicting tool [13] of the total of 152 potential neoepitopes, and their respective non-mutated versions.

Finally, we have compared both groups and our results have indicated that bioinformatic tools support the hypothesis indicating that neoepitopes are more immunogenic than the original strings, which as a consequence reinforces the use of neoepitopes for cancer vaccine design.

2 Methodology

In order to analyze *in silico* characteristics of potential neoepitopes, the first step was to obtain tumor samples of patients with cutaneous melanoma. Tumor biopsies of six patients from Cruces University Hospital and Basurto Hospital (Spain) were obtained, and sequenced. With the objective of obtaining sufficient mutational diversity (and therefore an advanced stage of tumor development), but also a heterogeneous staging sample (where different severities of tumor could be analyzed), we selected cases with several cancer stages, but without metastasis. Particularly, the stage of the studied cases was: one IB (up to 2 mm thick without ulceration), one IIA (from 1 to 2 mm thick with ulceration or from 2.01 to 4 mm thick without ulceration), two IIB (from 2.01 to 4 mm thick with ulceration or greater than 4 mm thick without ulceration), and two IIC (greater than 4 mm thick with ulceration) [18].

Next, since performing an analysis of the whole mutanome of patients was out of our scope, we targeted the regions with most variability in this kind of cancer, which is related to the regions which codify proteins such as BRAF, NRAS, MAP2K1 or MAP2K2 [19, 20]. In order to select only the mutations appearing in the tumor and discard the ones derived as mistakes of normal cellular division (and present in non-oncogenic cells), we also sequenced regular blood cells and studied their mutations, and afterwards, we only kept as mutation candidates the ones appearing solely in the tumor.

Once mutations that potentially could originate neoepitopes were detected, the next step was to define the length of the neoepitopes that we were going to consider, or in other words, the amount of amino-acids around the mutated base (and its respective amino-acid). There are several approaches regarding this issue, for example, Sharin et al. [3] used 27-mer peptides with the mutation in the center (i.e., in the 14^{th} position), while Ott et al. [21] used variable lengths, ranging from 15 to 30, but maintaining the mutation in the center. One reason to maintain the mutation in the center and these range of lengths, is that class I MHC usually fit epitopes of 8-9-mer lengths while class II MHC attach epitopes of 12-15-mer lengths, and regardless the processing of the

antigens presented by the Antigen Presenting Cells (APC) such as dendritic cells or macrophages [22], the mutation will likely be included in one of the presented peptides. In our case, we considered epitopes of length 17, with the mutation in the 8th amino-acid. This way, on one hand, if we perform a sliding window for peptides with length 9 (which has been used traditionally as standard length for HLA-I restricted T-cell epitopes [23]), the mutation will be included in all of them. On the other hand, this length is long enough to allow us to estimate computationally the HLA-II binding affinity.

After fixing the potential neoepitopes, and in order to evaluate the binding affinity of both classes, since HLA complexes are highly polymorphic and vary from each patient, we also sequenced the genes responsible for coding the MHC, which are located in the 6th chromosome [24], and identified the allele variants corresponding to each patient. In Tables 1 and 2, we depict, divided by patient, the obtained HLA alleles of class I and II respectively.

Table 1. In the first row, the patient number; in the second, the first or second allele (one obtained from the father and the other one from the mother); in the rest of the rows, the respective class I HLA alleles: HLA-A, HLA-B and HLA-C.

HLA-I	1st Patient		2nd Patient		3rd Patient		4th Patient		5th Patient		6th Patient	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
A	01:01:01	02:01:01	02:01:01	32:01:01	24:02:01	32:01:01	02:01:01	11:01:01	24:02:01	24:02:01	03:01:01	26:01:01
B	08:01:01	44:27:01	35:11:01	51:01:01	07:02:01	51:01:01	27:05:02	40:01:02	35:01:01	40:01:03	07:02:01	18:01:01
C	07:01:01	07:04:01	02:02:02	04:01:01	07:02:01	15:02:01	01:02:01	03:04:01	03:04:01	04:01:01	02:02:02	07:02:01

Finally, we estimated the MHC class I and II binding affinity predictions with the respective tools of the Immune Epitope Database Analysis Resource [13]. The prediction method used was “IEDB recommended 2.22”, the selected species “human”, and in length, for class I prediction we selected “all lengths” (which ranged from 8 to 14), while for class II we selected “default”, (which fixed the epitope length to 15).

Table 2. In the first row, the patient number; in the second, the first or second allele (one obtained from the father and the other one from the mother); in the rest of the rows, the respective class II HLA alleles: HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4 and HLA-DRB5.

HLA-II	1st Patient		2nd Patient		3rd Patient		4th Patient		5th Patient		6th Patient	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
DPA1	01:03:01	02:01:01	01:03:01	01:03:01	01:03:01	01:03:01	01:03:01	01:03:01	01:03:01	01:03:01	01:03:01	02:01:01
DPB1	02:01:02	14:01:01	02:01:02	04:01:01	--:01:--	--:01:--	04:01:01	06:01:--	04:01:01	04:01:01	02:01:02	11:01:01
DQA1	01:02:02	01:04:01	01:02:01	05:05:01	01:02:01	01:03:01	03:01:01	04:01:01	01:01:01	04:01:01	01:03:01	05:05:01
DQB1	05:02:01	05:03:01	03:01:01	06:02:01	06:02:01	06:03:01	03:02:01	04:02:01	04:02:01	05:01:01	03:01:01	06:03:01
DRB1	16:01:01	14:54:01	15:01:01	11:04:01	15:01:01	13:01:01	08:01:--	04:04:01	01:01:01	08:02:01	--:--:--	--:--:--
DRB3	02:02:01	--:--:--	02:02:01	--:--:--	01:01:02	--:--:--	--:--:--	--:--:--	--:--:--	--:--:--	01:01:02	01:01:02
DRB4	--:--:--	--:--:--	--:--:--	--:--:--	--:--:--	--:--:--	01:03:01	01:03:01	--:--:--	--:--:--	--:--:--	--:--:--
DRB5	02:02:--	--:--:--	01:01:01	--:--:--	01:01:01	--:--:--	--:--:--	--:--:--	--:--:--	--:--:--	--:--:--	--:--:--

3 Results

In order to compare the MHC binding affinity (correlated to the generated immune response [25]) of potential neoepitopes and their non-mutated version, we estimated the MHC class I and II binding affinity using the bioinformatic tools offered by IEDB [13] described in Methods section. After obtaining the estimations, following the recommendations from IEDB, we used the "percentile rank" variable to filter the potential binders from the ones that predictably would not be good binders. To cover most of the

immune responses, IEDB recommends to select the strings with "percentile rank" of $\leq 1\%$ for MHC class I [26, 27], while for MHC class II the recommended "percentile rank" would be $\leq 10\%$. In Table 3, we depict the number of potential HLA-I and HLA-II binders, divided by patients, while in Fig. 1, we illustrate the distribution of these values in a box plot.

Table 3. Number of possible epitopes according to their predicted binding affinity, with "percentile rank" $\leq 1\%$ for MHC-I and $\leq 10\%$ for MHC-II, for both mutated (M) and non-mutated (NM) versions.

	1st patient	2nd patient	3rd patient	4th patient	5th patient	6th patient
HLA-I M	57	88	8	121	20	26
HLA-I NM	49	70	4	114	20	30
HLA-II M	51	229	26	144	47	21
HLA-II NM	45	193	22	118	33	22

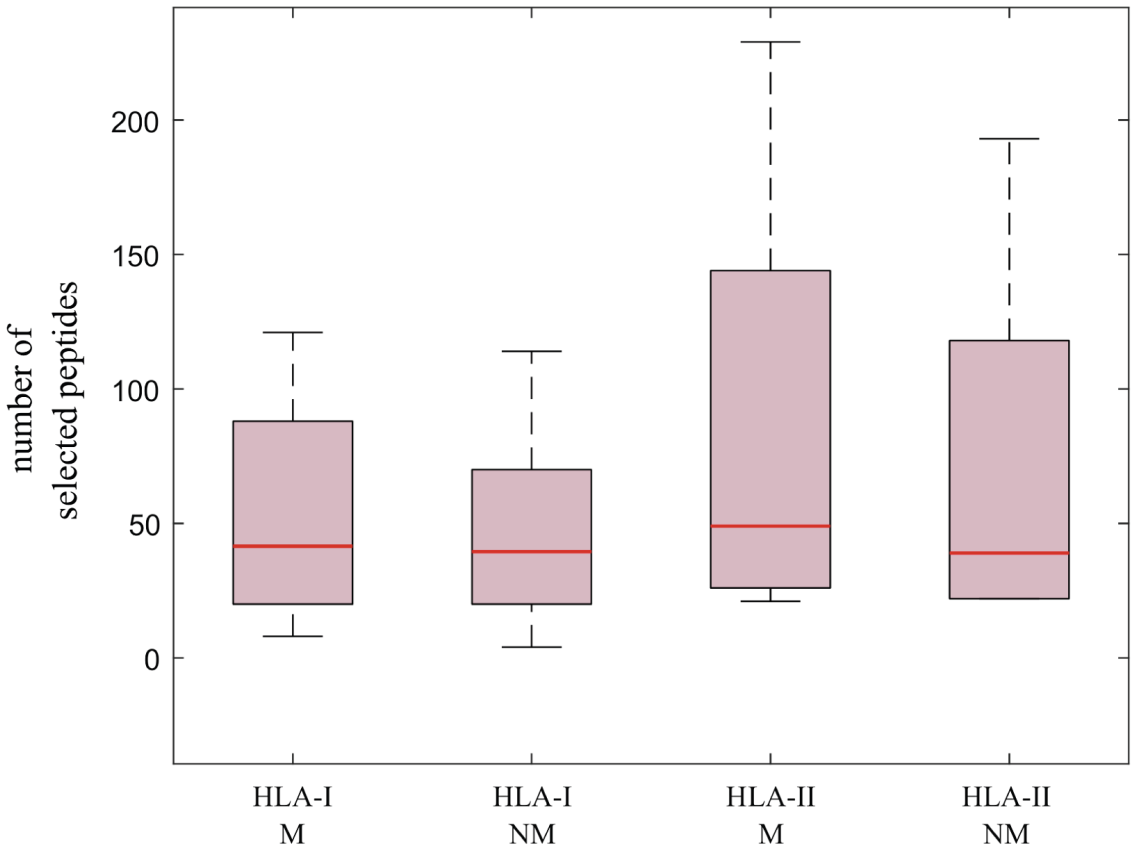

Fig. 1. Box plot of the number of peptides that pass the threshold. Box plot illustrating the distributions of the number of peptides that passed the respective cutoff points ($\leq 1\%$ for HLA-I and $\leq 10\%$ for HLA-II binding affinity). M indicates mutated peptides, while NM references the non-mutated ones. The maroon boxes represent the distribution of the central 50% of the values and the red lines represent the medians. The rest of the values are represented by the arms. (Color figure online)

Before studying the main hypothesis, we analyzed the relative percentages of each group. In the case of mutated peptides which would potentially bind HLA-I molecules, the average ± standard deviation was 0.69% ± 0.31%, ranging from 0.27% to 0.98%; for their non-mutated version, the results were 0.65% ± 0.36%, ranging from 0.14% to 1.13%; when HLA-II was analyzed, the mutated peptides were the 15.24% ± 2.96%, ranging from 10.63% to 19.88%; and finally the non-mutated ones were the 13.02% ± 2.93%, ranging from 9.38% to 16.75%.

Next, to verify the work hypothesis, i.e., to answer if is it true that the mutated versions of peptides are more immunogenic than the non-mutated ones, we performed two comparisons:

1. The first, studying if the number of predicted HLA-I mutated peptides is greater than the non-mutated amount.
2. The second, studying if the number of predicted HLA-II mutated peptides is greater than the non-mutated amount.

To make these comparisons, first we need to study if our variables are normally distributed. The p-values of the normality tests were 0.837, 0.978, 0.43 and 0.476, for HLA-I mutated, HLA-I non-mutated, HLA-II mutated and HLA-II non-mutated, respectively. Thus, normality was accepted in all cases, and we performed a paired T-test for each HLA couple.

Being the first null hypothesis that the difference between the average of mutated potential epitopes minus the average of non-mutated was greater than or equal to 0, with a significance of 5%, the p-value was 0.068, the confidence interval $CI^{0.95}_{\mu_1 \geq \mu_2} = (-0.74, \infty)$, and the t-statistic 1.78. Therefore, according to this data, no significant differences were found between the amount of mutated and non-mutated potential HLA-I peptides.

Finally, we performed the same comparison for the number of HLA-II peptides. In this case, the p-value was 0.03, the confidence interval $CI^{0.95}_{\mu_1 \geq \mu_2} = (2.44, \infty)$, and the t-statistic 2.43. Thus, we can conclude that mutated versions of peptides will bind significantly better to HLA-II molecules than non-mutated ones.

4 Discussion

In this work, for the first time, we have performed the comparison between potential neoepitopes and the corresponding peptides without the mutation, in experimental data obtained from six patients suffering from cutaneous melanoma in diverse stages (IB, IIA, IIB and IIC). To perform such study, we started sequencing both tumor and blood cells, selecting only mutations that happened in the tumor, and next, we identified the amino-acid sequence that surrounds the mutations, which gave us possible neoepitopes, for which we set the maximum length to 17aa. Then, we estimated the binding affinity for HLA classes I and II with the bioinformatic tools provided by IEDB, and for each case, enumerated the ones below the significance threshold.

Our results indicated that, even if the number of mutated strings (i.e., potential neoepitopes) presented higher binding affinity in almost every case, the difference was

not significant when we compared HLA-I binding affinity (p-value: 0.068), while in the case of HLA-II, the number of mutated vs non-mutated was significantly bigger (p-value: 0.03). Thus, we consider that this study answers, at least partially, the question raised by the medical community, which considered that mutated versions of peptides are more immunogenic than the non-mutated ones.

However, we want to acknowledge that our sample size is relatively small (n = 6), and since the differences between the number of possible HLA binders are also small, we hypothesize that increasing the sample size would lead to more significance in the results, and even also to difference between HLA-I binding mutated and non-mutated candidates.

In summary, here, we have performed a comparison between potential neoepitopes and their respective original peptides, and found that the mutated ones have significantly more HLA-II binding affinity, which supports the hypothesis indicating that mutated strings are indeed more immunogenic than their more common versions.

Acknowledgements. This work was supported by Basque Government funding (IT1974-16, KK-2018/00090), and by the UPV/EHU and Basque Center of Applied Mathematics (US18/21)

References

1. Vormehr, M., Diken, M., Türeci, Ö., Sahin, U., Kreiter, S.: Personalized neo-epitope vaccines for cancer treatment. In: Theobald, M. (ed.) Current Immunotherapeutic Strategies in Cancer. RRCR, vol. 214, pp. 153–167. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-23765-3_5
2. Vermaelen, K.: Vaccine strategies to improve anti-cancer cellular immune responses. Front. Immunol. **10**, 8 (2019). https://doi.org/10.3389/fimmu.2019.00008
3. Sahin, U., et al.: Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature **547**, 7662 (2017)
4. Kakimi, K., Karasaki, T., Matsushita, H., Sugie, T.: Advances in personalized cancer immunotherapy. Breast Cancer **24**, 16–24 (2017)
5. Stratton, M.R.: Exploring the genomes of cancer cells: progress and promise. Science **331**, 1553–1558 (2011)
6. Kreiter, S., Castle, J.C., Türeci, Ö., Sahin, U.: Targeting the tumor mutanome for personalized vaccination therapy. Oncoimmunology **1**, 768–769 (2012)
7. Leclerc, M., et al.: Recent advances in lung cancer immunotherapy: input of T-cell epitopes associated with impaired peptide processing. Front. Immunol. **10**, 1505 (2019)
8. Vormehr, M., Türeci, Ö., Sahin, U.: Harnessing tumor mutations for truly individualized cancer vaccines. Annu. Rev. Med. **70**, 395–407 (2019)
9. Tanyi, J.L., et al.: Personalized cancer vaccine effectively mobilizes antitumor T cell immunity in ovarian cancer. Sci. Transl. Med. **10**, eaao5931 (2018)
10. Hu, Z., Ott, P.A., Wu, C.J.: Towards personalized, tumour-specific, therapeutic vaccines for cancer. Nat. Rev. Immunol. **18**, 168 (2018)
11. Fritsch, E.F., Rajasagi, M., Ott, P.A., Brusic, V., Hacohen, N., Wu, C.J.: HLA-binding properties of tumor neoepitopes in humans. Cancer Immunol. Res. **2**, 522–529 (2014)
12. Lundegaard, C., Lund, O., Nielsen, M.: Prediction of epitopes using neural network-based methods. J. Immunol. Methods **374**, 26–34 (2011)

13. Zhang, Q., et al.: Immune epitope database analysis resource (IEDB-AR). Nucl. Acids Res. **36**, 513–518 (2008)
14. Soria-Guerra, R.E., Nieto-Gomez, R., Govea-Alonso, D.O., Rosales-Mendoza, S.: An overview of bioinformatics tools for epitope prediction: implications on vaccine development. J. Biomed. Inform. **53**, 405–414 (2015)
15. Martínez, L., Milanič, M., Malaina, I., Álvarez, C., Pérez, M.B., Ildefonso, M.: Weighted lambda superstrings applied to vaccine design. PLoS ONE **14**, e0211714 (2019)
16. Malaina, I., et al.: Metastasis of cutaneous melanoma: risk factors, detection and forecasting. In: Rojas, I., Ortuño, F. (eds.) IWBBIO 2018. LNCS, vol. 10813, pp. 511–519. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78723-7_44
17. Miller, A.J., Mihm, M.C.: Melanoma. N. Engl. J. Med. **355**, 51–65 (2006)
18. Thompson, J.A.: The revised american joint committee on cancer staging system for melanoma. In: Seminars in Oncology, vol. 29, pp. 361–369. WB Saunders (2002)
19. Edlundh-Rose, E., et al.: NRAS and BRAF mutations in melanoma tumours in relation to clinical characteristics: a study based on mutation screening by pyrosequencing. Melanoma Res. **16**, 471–478 (2006)
20. Nikolaev, S.I., et al.: Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat. Genet. **44**, 133 (2012)
21. Ott, P.A., et al.: An immunogenic personal neoantigen vaccine for patients with melanoma. Nature **547**, 217 (2017)
22. Mann, E.R., Li, X.: Intestinal antigen-presenting cells in mucosal immune homeostasis: crosstalk between dendritic cells, macrophages and B-cells. World J. Gastroenterol. WJG **20**, 9653 (2014)
23. Trolle, T., et al.: The length distribution of class I–restricted T cell epitopes is determined by both peptide supply and MHC allele–specific binding preference. J. Immunol. **196**, 1480–1487 (2016)
24. López-Martínez, A., Chávez-Muñoz, C., Granados, J.: Función biológica del complejo principal de histocompatibilidad. Revista de investigación clínica **57**, 132–141 (2005)
25. Sette, A., et al.: The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. J. Immunol. **153**, 5586–5592 (1994)
26. Moutaftsi, M., et al.: A consensus epitope prediction approach identifies the breadth of murine T CD8 + -cell responses to vaccinia virus. Nat. Biotechnol. **24**, 817 (2006)
27. Kotturi, M.F., et al.: The CD8+ T-cell response to lymphocytic choriomeningitis virus involves the L antigen: uncovering new tricks for an old virus. J. Virol. **81**, 4928–4940 (2007)