

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **3 001 611**

51 Int. Cl.:

G16B 30/00 (2009.01)

G16B 40/20 (2009.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **15.10.2021** **PCT/EP2021/078639**

87 Fecha y número de publicación internacional: **21.04.2022** **WO22079255**

96 Fecha de presentación y número de la solicitud europea: **15.10.2021** **E 21791393 (8)**

97 Fecha y número de publicación de la concesión europea: **04.12.2024** **EP 4229640**

54 Título: **Método, sistema y producto de programa informático para determinar la inmunogenicidad de péptidos**

30 Prioridad:

15.10.2020 EP 20202140

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
05.03.2025

73 Titular/es:

MYNEO NV (100.00%)
Ottergemsesteenweg Zuid 808 bus B511
9000 Gent, BE

72 Inventor/es:

PFITZER, LENA;
FANT, BRUNO y
BOGAERT, CEDRIC

74 Agente/Representante:

ELZABURU, S.L.P

ES 3 001 611 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Método, sistema y producto de programa informático para determinar la inmunogenicidad de péptidos

Campo de la invención

- 5 La presente invención se refiere a un método, sistema y producto de programa informático para determinar la inmunogenicidad de péptidos, así como al uso de los mismos para el tratamiento del cáncer, vacunación de virus, etc.

Antecedentes

La predicción de epítomos de células T desempeña un papel importante en el diseño experimental de inmunización y la preparación de vacunas.

- 10 Actualmente, la mayor parte de la investigación de la predicción de epítomos se centra en el procesamiento y la presentación de péptidos, p. ej., escisión proteosomal, transportador asociado con la combinación de procesamiento de antígenos (TAP) y complejo mayor de histocompatibilidad (MHC).

- 15 Hasta la fecha, sin embargo, el mecanismo para la inmunogenicidad de los epítomos sigue sin estar claro. Generalmente se acepta que en la inmunogenicidad de las células T puede influir la ajenidad, accesibilidad, peso molecular y estructura, conformación molecular, propiedades químicas y propiedades físicas de los péptidos objetivo en diferentes grados.

- 20 Para el desarrollo de vacunas (es decir, una vacuna contra el cáncer, vacuna vírica), es esencial identificar péptidos que puedan producir una respuesta inmunitaria. La potencia de una vacuna depende en gran medida de la inmunogenicidad de los epítomos, es decir, la capacidad de un posible péptido objetivo para producir una respuesta inmunitaria. El estado actual de la técnica para el ensayo de inmunogenicidad implica ensayos que requieren mano de obra y tiempo y que requieren medidas de seguridad en un entorno de laboratorio.

- Las estrategias actuales para identificar péptidos capaces de producir una respuesta inmunitaria usan la afinidad de unión predicha (CI50) a moléculas del MHC. Ejemplos de dichas estrategias se discuten en los documentos US 2013 330 335, US 2016 132 631 y US 2019 346 442. Sin embargo, no todos los péptidos con una afinidad de unión alta (predicha) también producen una respuesta inmunitaria.

- 25 El uso de dichos modelos de predicción de la unión que predicen si es probable que un péptido sea presentado en el MHC y suponiendo que esos péptidos seleccionados son también neoantígenos inmunogénicos, da como resultado un gran número de predicciones de falsos positivos. Tan poco como el 5% de dichos neoantígenos identificados a través de dichos modelos para predecir la unión/presentación del péptido en el MHC también son capaces de producir una respuesta inmunitaria, destacando la necesidad de una mejor identificación de péptidos realmente inmunogénicos.

- 30 El método predictivo de inmunogenicidad de IEDB y Ineo-Epp son dos herramientas de predicción de inmunogenicidad disponibles públicamente. El método predictivo de inmunogenicidad de IEDB es un modelo simple que clasifica péptidos basándose en la posición y propiedades de sus aminoácidos. Se basa en el enriquecimiento por posición de aminoácidos en péptidos inmunogénicos frente a no inmunogénicos. La puntuación de inmunogenicidad resultante de un péptido es la suma de sus puntuaciones de aminoácidos individuales. Ineo-Epp es un clasificador de bosque aleatorio entrenado en péptidos inmunogénicos de HLA-I de los que se extrajeron varias características, como propiedades fisicoquímicas de aminoácidos y puntuación de rango de percentil de probabilidad de ligando eluido (rango EL(%)). Ineo-Epp usa una lista de péptidos, junto con un alelo de HLA como entrada. Para la predicción de antígenos, se pueden usar los nueve supertipos principales de HLA-I (A1, A2, A3, A24, B7, B27, B44, B58, B62).

- 40 Ambas herramientas de predicción, sin embargo, tienen un rendimiento sustancialmente inferior en comparación con el sistema predictor de la presente invención. La presente invención tiene como objetivo proporcionar herramientas de predicción de inmunogenicidad mejoradas frente a las técnicas del estado de la técnica.

Resumen de la invención

En un primer aspecto, la invención se refiere a un método implementado por ordenador para determinar la inmunogenicidad de un péptido, según la reivindicación 1.

- 45 En un segundo aspecto, la invención se refiere a un sistema informático para determinar la inmunogenicidad de un péptido, según la reivindicación 11.

En un tercer aspecto, la invención se refiere a un producto de programa informático para determinar la inmunogenicidad de un péptido, según la reivindicación 12.

- 50 En un cuarto aspecto, la invención se refiere a un uso para determinar un tratamiento de cáncer de un sujeto, particularmente determinando la inmunogenicidad de un neoantígeno presentado por una célula tumoral del sujeto, según la reivindicación 13.

En un quinto aspecto, la invención se refiere a un uso para el cribado de un virus o bacteria en un objetivo de vacunación, particularmente determinando la inmunogenicidad de epítopos de proteínas víricas o bacterianas, según la reivindicación 14.

5 En un sexto aspecto, la invención se refiere a un uso para caracterizar reacciones de autoinmunidad de un sujeto, particularmente determinando la inmunogenicidad de autoantígenos presentados por una célula del sujeto, según la reivindicación 15.

10 En un séptimo aspecto, la invención se refiere a un uso para evaluar cambios de inmunogenicidad de un péptido tras alterar uno o más aminoácidos de la secuencia de aminoácidos de dicho péptido, en particular simular cambios de aminoácidos que pueden aumentar o disminuir la inmunogenicidad de péptidos particulares, según la reivindicación 16.

15 La presente invención es ventajosa ya que el clasificador de aprendizaje automático se puede entrenar por medio de conjuntos de registros, aprendiendo de este modo a reconocer la inmunogenicidad de péptidos basándose en una secuencia de aminoácidos de los mismos. Esto, por ejemplo, simplifica significativamente el procedimiento de desarrollo de vacunas, ya que los péptidos pueden seleccionarse basándose en una indicación de inmunogenicidad bien informada. Además, la evaluación comparativa muestra una mejora significativa en el rendimiento del clasificador de aprendizaje automático de la presente invención en comparación con otros clasificadores de péptidos inmunogénicos conocidos en la técnica.

Las realizaciones preferidas de la presente invención se comentan en las reivindicaciones 2-10, así como a lo largo de la descripción, ejemplos y figuras.

20 Descripción de las figuras

La Figura 1 muestra un flujo de trabajo esquemático de la presente invención.

Las Figuras 2 a 5 muestran la evaluación comparativa de característica operativa del receptor (ROC) y precisión-sensibilidad de diferentes configuraciones y clasificadores que se implementan según la presente invención.

25 La Figura 6 muestra la comparativa de ROC y precisión-sensibilidad de la presente invención respecto a otros clasificadores de péptidos inmunogénicos conocidos en la técnica.

La Figura 7 muestra histogramas y curvas de densidad de probabilidad de inmunogenicidad para los péptidos derivados de la envoltura de SARS-CoV-2, SARS-CoV y MERS-CoV.

30 La Figura 8 muestra símbolos de secuencias peptídicas que muestran el uso diferencial de aminoácidos de péptidos de 9 aminoácidos de longitud en el conjunto de datos positivos inmunogénicos (arriba) frente a péptidos de 9 aminoácidos de longitud en el conjunto de datos negativos no inmunogénicos (abajo). Los aminoácidos estaban no agrupados (Fig. 8A), agrupados por propiedades químicas (Fig. 8B), carga eléctrica (Fig. 8C) o tamaño (Fig. 8D).

35 La Figura 9 muestra la correlación de diferentes parámetros usados para predecir una reacción inmunogénica con resultados de ELISpot de 33 péptidos, los cuales se ha predicho por netMHCpan4.0 que son todos presentados. La Fig. 9A muestra la probabilidad predicha de que los péptidos sean inmunogénicos ensayado usando el método implementado por ordenador de la invención, el método predictivo de inmunogenicidad de IEDB y netMHCpan4.0, basado este último en un alelo de HLA específico o en todos los alelos de HLA de un sujeto, de grupos de péptidos con un resultado en ELISpot positivo o negativo. La Fig. 9B muestra una comparación de curvas de características operativas del receptor (ROC) de múltiples métodos predictivos de inmunogenicidad diferentes, incluyendo el método implementado por ordenador de la invención, el método predictivo de inmunogenicidad de IEDB y netMHCpan4.0, basado este último en un alelo de HLA específico o en todos los alelos de HLA de un sujeto. La Fig. 9C muestra el gráfico de ganancias acumulativas de múltiples métodos predictivos de inmunogenicidad diferentes que incluyen el método implementado por ordenador de la invención, el método predictivo de inmunogenicidad de IEDB y netMHCpan4.0, basado este último en un alelo de HLA específico o en todos los alelos de HLA de un sujeto. La curva de ganancia acumulativa ilustra qué porcentaje de los positivos verdaderos se identifica con la porción más pequeña posible de muestras.

45 La Figura 10 muestra la tasa de respuesta frente a la duración de la respuesta de pacientes con melanoma al bloqueo de CTLA-4. Los pacientes se dividen en dos grupos, indicados como "Alto" (línea continua) o "Bajo" (línea discontinua), basado en el número mediano de mutaciones (carga mutacional tumoral TMB), el número de mutaciones presentadas o el número de mutaciones inmunogénicas.

50 Descripción detallada de la invención

La invención se refiere a un método, sistema y producto de programa informático para determinar la inmunogenicidad de péptidos. Además, la invención se refiere al uso del método, el sistema y/o el producto, para el tratamiento del cáncer y la vacunación de virus. En lo que sigue, se describirá la invención en detalle, se explican realizaciones preferidas, y la invención se ilustrará por medio de ejemplos no limitantes.

A menos que se defina lo contrario, todos los términos usados en la descripción de la invención, incluyendo los términos técnicos y científicos, tienen el significado que entiende comúnmente un experto en la técnica a la que pertenece esta invención. A modo de guía, se incluyen definiciones de términos para apreciar mejor la enseñanza de la presente invención.

5 Como se usa en el presente documento, los siguientes términos tienen los siguientes significados:

"Un", "una", "el" y "la", como se usa en el presente documento, se refieren a referentes tanto singulares como plurales a menos que el contexto indique claramente lo contrario. A modo de ejemplo, "un compartimento" se refiere a uno o más de un compartimento.

10 "Aproximadamente", como se usa en el presente documento, que se refiere a un valor medible tal como un parámetro, una cantidad, una duración temporal y similares, pretende abarcar variaciones de +/-20% o menos, preferiblemente +/-10% o menos, más preferiblemente +/-5% o menos, incluso más preferiblemente +/-1% o menos, y aún más preferiblemente +/-0,1% o menos del valor especificado, en la medida en que dichas variaciones son apropiadas para tener lugar en la invención descrita. Sin embargo, debe entenderse que el valor al que se refiere el modificador "aproximadamente" también se describe en sí específicamente.

15 "Comprender", "que comprende" y "comprende" y "compuesto de" como se usa en el presente documento son sinónimos de "incluir", "que incluye", "incluye" o "contener", "que contiene", "contiene" y son términos inclusivos o abiertos que especifican la presencia de lo que sigue, p. ej., componente, y no excluyen o impiden la presencia de elementos, miembros, etapas adicionales no mencionadas, etc., conocidos en la técnica o descritas en la misma.

20 Además, los términos primero, segundo, tercero y similares en la descripción y en las reivindicaciones, se usan para distinguir entre elementos similares y no necesariamente para describir un orden secuencial o cronológico, a menos que se especifique. Debe entenderse que los términos así usados son intercambiables en circunstancias apropiadas y que las realizaciones de la invención descritas en el presente documento pueden funcionar en secuencias distintas de las descritas o ilustradas en el presente documento.

25 Mientras que las expresiones "uno o más" o "al menos uno", tales como uno o más o al menos un miembro(s) de un grupo de miembros, son claras por sí mismas, a modo de ejemplo adicional, la expresión abarca entre otras una referencia a uno cualquiera de dichos miembros, o a cualesquiera dos o más de dichos miembros, tal como, p. ej., cualquiera ≥ 3 , ≥ 4 , ≥ 5 , ≥ 6 o ≥ 7 , etc. de dichos miembros, y hasta todos dichos miembros.

La mención de intervalos numéricos por puntos finales incluye todos los números y fracciones incluidos dentro de ese intervalo, así como los puntos finales mencionados.

30 A menos que se defina lo contrario, todos los términos usados en la descripción de la invención, incluyendo los términos técnicos y científicos, tienen el significado que entiende comúnmente un experto en la técnica a la que pertenece esta invención. Por medio de una guía adicional, se incluyen definiciones para los términos usados en la descripción para apreciar mejor la enseñanza de la presente invención. Los términos o definiciones usados en el presente documento se proporcionan únicamente para ayudar en la comprensión de la invención.

35 La referencia a lo largo de esta memoria descriptiva a "una (1) realización" o "una realización" significa que un rasgo, estructura o característica particular descrito en relación con la realización se incluye en al menos una realización de la presente invención. Por lo tanto, las apariciones de las frases "en una (1) realización" o "en una realización" en varios lugares a lo largo de esta memoria descriptiva no se refieren necesariamente todas a la misma realización, pero pueden. Además, los rasgos, estructuras o características particulares pueden combinarse de cualquier manera adecuada, como resultará evidente para un experto en la técnica a partir de esta descripción, en una o más realizaciones. Además, aunque algunas realizaciones descritas en el presente documento incluyen algunas, pero no otras características incluidas en otras realizaciones, se pretende que las combinaciones de características de diferentes realizaciones estén dentro del alcance de la invención, y formen diferentes realizaciones, como entenderán los expertos en la técnica. Por ejemplo, en las siguientes reivindicaciones, cualquiera de las realizaciones reivindicadas puede usarse en cualquier combinación.

40

45

En un primer aspecto, la presente invención se refiere a un método implementado por ordenador para determinar la inmunogenicidad de un péptido según la reivindicación 1.

50 En un segundo aspecto, la presente invención se refiere a un sistema informático para determinar la inmunogenicidad de un péptido. Estando configurado el sistema informático para realizar el método implementado por ordenador según el primer aspecto.

En un tercer aspecto, la presente invención se refiere a un producto de programa informático para determinar la inmunogenicidad de un péptido, en donde el producto de programa informático comprende instrucciones que, cuando el producto de programa informático es ejecutado por un ordenador, hacen que el ordenador lleve a cabo el método implementado por ordenador según el primer aspecto.

55

En un cuarto aspecto, la presente invención se refiere al uso del método implementado por ordenador del primer aspecto y/o el sistema informático del segundo aspecto y/o el producto de programa informático del tercer aspecto, para determinar un tratamiento de cáncer de un sujeto, particularmente determinando la inmunogenicidad de un neoantígeno presentado por una célula tumoral del sujeto.

- 5 En un quinto aspecto, la presente invención se refiere al uso del método implementado por ordenador del primer aspecto y/o el sistema del segundo aspecto y/o el producto del tercer aspecto, para el cribado de un virus en un objetivo de vacunación, particularmente determinando la inmunogenicidad de epítomos de proteínas víricas.

- 10 En un sexto aspecto, la invención se refiere a un uso del método implementado por ordenador del primer aspecto y/o el sistema del segundo aspecto y/o el producto del tercer aspecto, para caracterizar reacciones de autoinmunidad de un sujeto, particularmente determinando la inmunogenicidad de autoantígenos presentados por una célula del sujeto.

En un séptimo aspecto, la invención se refiere a un uso del método implementado por ordenador del primer aspecto y/o el sistema del segundo aspecto y/o el producto del tercer aspecto, para el cribado de una bacteria en un objetivo de vacunación, particularmente determinando la inmunogenicidad de epítomos de proteínas bacterianas.

- 15 En un octavo aspecto, la invención se refiere a un uso del método implementado por ordenador del primer aspecto y/o el sistema del segundo aspecto y/o el producto del tercer aspecto, para evaluar cambios de inmunogenicidad de un péptido tras alterar uno o más aminoácidos de la secuencia de aminoácidos de dicho péptido. En particular, un uso para simular cambios de aminoácidos que pueden aumentar o disminuir la inmunogenicidad de péptidos particulares.

- 20 La invención proporciona un método implementado por ordenador, un sistema informático y un producto de programa informático para determinar la inmunogenicidad de un péptido, así como un uso de cualquiera de cualquiera del método, sistema o producto para el tratamiento del cáncer, vacunación de virus, caracterización de reacciones de autoinmunidad de un sujeto, cribado de una bacteria u otros patógenos en un objetivo de vacunación y para evaluar cambios de inmunogenicidad de un péptido tras alterar la secuencia de aminoácidos de dicho péptido. Un experto en la técnica apreciará que el método se implementa en el producto de programa informático y se ejecuta usando el sistema informático. También está claro para un experto en la técnica que la predicción de inmunogenicidad puede usarse para el tratamiento del cáncer, vacunación de virus, caracterización de reacciones de autoinmunidad de un sujeto y cribado de una bacteria en un objetivo de vacunación. Por lo tanto, los ocho aspectos de la presente invención en lo sucesivo se tratan juntos.

- 30 Una realización sencilla de la invención proporciona preferiblemente la obtención de una secuencia de aminoácidos del péptido del que se va a determinar la inmunogenicidad. Preferiblemente, el péptido es un péptido presentado en la superficie celular. Más preferiblemente, un péptido unido al complejo mayor de histocompatibilidad (MHC) presentado en la superficie celular. Dichos péptidos preferiblemente se unen al MHC de clase I y desencadenan una respuesta de células T CD8, o se unen al MHC de clase II y desencadenan una respuesta de células T CD4. La invención es por lo tanto aplicable en algunas realizaciones a alelos de MHC I y en algunas realizaciones a alelos de MHC II. Preferiblemente, se debería saber que las secuencias peptídicas para las que se predecirá la inmunogenicidad usando la invención (secuencias problema) son presentadas en el MHC. La ventaja de esto es que dichos péptidos problema se parecerán más estrechamente a los péptidos de los que se deriva el conjunto de datos de entrenamiento (véase más adelante) y que se sabe que todos son presentados en el MHC. Por lo tanto, las propiedades relacionadas con la inmunogenicidad y no la presentación en el MHC son los factores diferenciadores más dominantes entre los péptidos positivos y negativos (véase más adelante).

- 40 El péptido puede obtenerse de cualquier sujeto. "Sujeto", como se usa en el presente documento, se refiere a un término conocido en el estado de la técnica, que debe entenderse preferiblemente como un cuerpo humano o animal, lo más preferiblemente un cuerpo humano. Como se usa en el presente documento, "animal" se refiere preferiblemente a vertebrados, preferiblemente a aves y mamíferos, incluso más preferiblemente a mamíferos. "Sujeto que lo necesita", como se usa en el presente documento, debe entenderse como un sujeto que se beneficiará del tratamiento (profiláctico), p. ej., cáncer o vacunación de virus. Además, el péptido para el que se va a determinar la inmunogenicidad comprende una secuencia de aminoácidos (aa) y preferiblemente tiene una longitud entre 9 y 11 aa, o una longitud entre 8 y 12, o una longitud entre 13 y 25.

- 50 Una realización sencilla de la invención proporciona preferiblemente la obtención para cada aminoácido proteínogénico de una pluralidad de índices numéricos, en donde cada uno de dichos índices numéricos está relacionado con una propiedad fisicoquímica de un aminoácido proteínogénico correspondiente. Los "aminoácidos proteínogénicos", como se usa en el presente documento, son aminoácidos que se incorporan biosintéticamente en proteínas durante la traducción. "Proteínogénico" significa "que crea proteína". En toda la vida conocida, hay 22 aminoácidos (proteínogénicos) codificados genéticamente, 20 en el código genético estándar y 2 adicionales que pueden incorporarse mediante mecanismos de traducción especiales. Por el contrario, los "aminoácidos no proteínogénicos" son aminoácidos que o bien no se incorporan en proteínas (como GABA, L-DOPA o triyodotironina), se desincorporan en lugar de un aminoácido codificado genéticamente, o no son producidos directamente y de manera aislada por maquinaria celular convencional (como hidroxiprolina). Esta última resulta frecuentemente de la modificación postraducciona de proteínas. Obteniendo información sobre ácidos proteínogénicos, se obtiene información sobre las "unidades estructurales" de cada péptido de un sujeto. Un péptido puede comprender por tanto aminoácidos proteínogénicos, aminoácidos no proteínogénicos, o una combinación de ambos.

Las propiedades fisicoquímicas se obtienen preferiblemente de una base de datos conocida en la técnica, tal como la base de datos AAindex en "<https://www.genome.jp/aaindex/>". AAindex es una base de datos de índices numéricos que representan diversas propiedades fisicoquímicas y bioquímicas de aminoácidos y pares de aminoácidos. AAindex ahora consiste en tres secciones. La sección de interés es la AAindex1 que comprende el índice de aminoácidos de 20 valores numéricos. Los ejemplos de propiedades fisicoquímicas incluyen, pero no se limitan de ninguna manera a, hidrofobicidad, energía libre, distribución de aa, etc.

Una realización sencilla de la invención preferiblemente proporciona el entrenamiento de un modelo de clasificación en el conjunto de datos de entrenamiento, y la determinación de la probabilidad de que dicho péptido produzca una respuesta inmunitaria por medio del modelo de clasificación entrenado. En el contexto del entrenamiento de modelos matemáticos tales como algoritmos de clasificación, se usan los siguientes términos, y por lo tanto se explican adicionalmente como guía.

El "conjunto de entrenamiento" es el conjunto de observaciones de datos (también denominadas "registros", "ejemplos" o "casos") que se usa para entrenar o aprender el modelo. Un modelo analítico tiene parámetros que es necesario estimar con el fin de hacer buenas predicciones. Esto se traduce en encontrar los valores de parámetros óptimos para el modelo analítico. Por esta razón, se usa el conjunto de entrenamiento para encontrar o estimar los valores de parámetros óptimos. Una vez que se tiene un modelo entrenado, se puede usar para hacer predicciones. En una tarea de clasificación supervisada, también se adjuntan etiquetas de clase (p. ej., "inmunogénicos", "no inmunogénicos") a cada observación para estimar los valores de parámetros óptimos. Esto permite entrenar al algoritmo en patrones que son útiles para identificar casos de fraude.

El "conjunto de validación" se refiere a modelos con parámetros que no pueden estimarse directamente a partir de los datos. Sin embargo, con el fin de encontrar también valores óptimos para esos parámetros (denominados hiperparámetros), se usa el denominado conjunto de validación. Típicamente, se puede identificar un conjunto de valores candidatos para los hiperparámetros. Se elige un valor candidato, se entrena el modelo en el conjunto de entrenamiento y se evalúa el rendimiento predictivo en el conjunto de validación. A continuación, se elige el siguiente valor candidato y se procede de una manera similar hasta que se han probado todos los valores candidatos. Al final, para cada valor candidato se obtiene una estimación correspondiente del rendimiento predictivo. Basándose en los rendimientos estimados en el conjunto de validación, se puede elegir el valor candidato que corresponde al rendimiento óptimo. El conjunto de entrenamiento y el conjunto de validación están preferiblemente separados estrictamente en todo el proceso con el fin de obtener estimaciones de rendimiento fiables. Es decir, las observaciones en el conjunto de validación no pueden estar en el conjunto de entrenamiento (o conjunto de prueba para ese asunto). Alternativamente, el conjunto de entrenamiento y el conjunto de validación no están separados estrictamente. Este es el caso, por ejemplo, de la validación cruzada.

El "conjunto de prueba", también "muestra retenida", es el conjunto de observaciones de datos que se usa para probar si el modelo entrenado hace buenas predicciones. Es decir, en la fase de evaluación del modelo, se conocen los valores verdaderos de las observaciones de prueba y se puede comprobar cuántos de los valores predichos son correctos comparándolos con los valores verdaderos. Es importante indicar que aquí las etiquetas de clase solo se usan para evaluar el rendimiento predictivo (p. ej., precisión) del modelo de clasificación. Es decir, las observaciones en el conjunto de prueba no pueden estar en el conjunto de entrenamiento o el conjunto de validación. La separación estricta es crucial, porque se desea el modelo para hacer predicción sobre observaciones que no se han usado en el proceso de entrenamiento. Sólo cuando esto está garantizado y el modelo muestra un buen rendimiento, se puede estar seguro de que el modelo también funcionará bien con nuevos datos previamente no vistos.

La "estrategia de retención" o "estrategia de división de entrenamiento y prueba única" se refiere a la división más simple porque los datos se dividen en dos subconjuntos: uno para entrenamiento y otro para prueba. Se puede entrenar el modelo con el primero y luego probarlo con el último. Obsérvese que el proceso de entrenamiento y prueba se hace solo una vez. Esta división de datos se hace aleatoriamente, es decir, las observaciones se asignan aleatoriamente para que pertenezcan al conjunto de entrenamiento o de prueba. El rendimiento se evalúa en el conjunto de prueba, normalmente para un conjunto de modelos candidatos, y se elige el mejor modelo. Además, a menudo es necesario comparar el modelo entrenado y validado, con el fin de tener en cuenta el sobreajuste. Algunos modelos tienen parámetros que no pueden estimarse a partir de los datos directamente. Se denominan hiperparámetros. Se puede confiar en un conjunto de validación para encontrar el mejor modelo. Aquí, se pueden dividir los datos en tres subconjuntos: uno para entrenamiento, uno para validación, y uno para prueba. La división también se realiza de manera aleatoria. Con la ayuda del conjunto de validación, se puede encontrar el modelo con los valores de hiperparámetros óptimos (es decir, una selección de modelo), y el mejor modelo se evalúa finalmente en el conjunto de prueba. Obsérvese que la elección para la selección del mejor modelo de predicción, entre un conjunto de diversos modelos candidatos, se hace basándose en el rendimiento medido en el conjunto de prueba, p. ej., puede ser necesario decidir si el modelo de regresión logística, el árbol de decisiones o el bosque aleatorio es el modelo de mejor rendimiento. Para tomar esta decisión, el rendimiento del conjunto de prueba es crucial. Cuando se encuentra el modelo de predicción final, se puede poner en práctica en el sistema operativo para hacer predicciones para datos nuevos, previamente no vistos.

La expresión "estrategia de validación cruzada de k particiones" se refiere a una alternativa a la división simple de entrenamiento y prueba. Corresponde a una división repetitiva de entrenamiento y prueba, de modo que el conjunto de prueba se desplaza sistemáticamente. A continuación, se promedian los rendimientos obtenidos en los conjuntos

de prueba. La ventaja de esta estrategia es que cada observación estará una vez en el conjunto de prueba. Aún, más importante, el rendimiento predictivo estimado se vuelve más fiable, lo que proporciona una mejor imagen del rendimiento de generalización del modelo.

El conjunto de datos de entrenamiento de la presente invención comprende al menos un conjunto de datos positivos que comprende una pluralidad de secuencias de aminoácidos de péptidos inmunogénicos. La "inmunogenicidad" o "inmunorreactividad", como se usa en el presente documento, es el resultado de un biomaterial que es detectado por el sistema inmunitario del cuerpo como un objeto extraño. Los biomateriales inmunorreactivos, tales como péptidos, se detectan mediante reacciones antigénicas en las células. Se produce entonces una cascada bioquímica, por la que las células T auxiliares migran hacia el biomaterial. Por consiguiente, el modelo de clasificación se entrena para reconocer inmunogenicidad.

Según una realización preferida, los péptidos inmunogénicos del conjunto de datos positivos se obtienen: obteniendo secuencias de aminoácidos de péptidos capaces de inducir una respuesta de células T; obteniendo secuencias de aminoácidos correspondientes a péptidos del proteoma; comparando las secuencias de aminoácidos que inducen respuesta de células T con las secuencias de aminoácidos del proteoma para determinar coincidencias entre las mismas; comprendiendo el conjunto de datos positivos las secuencias de aminoácidos que inducen respuesta de células T excepto la secuencia de aminoácidos de dichas coincidencias. Preferiblemente, las secuencias de aminoácidos que inducen respuesta de células T obtenidas son secuencias lineales.

Según una realización para determinar la inmunogenicidad de péptidos autoinmunitarios, los péptidos inmunogénicos del conjunto de datos positivos se obtienen: obteniendo secuencias de aminoácidos de péptidos capaces de inducir una respuesta de células T; obteniendo secuencias de aminoácidos correspondientes a péptidos del proteoma; comparando las secuencias de aminoácidos que inducen respuesta de células T con las secuencias de aminoácidos del proteoma para determinar coincidencias entre las mismas; comprendiendo el conjunto de datos positivos las secuencias de aminoácidos que inducen respuesta de células T de dichas coincidencias. Preferiblemente, las secuencias de aminoácidos que inducen respuesta de células T obtenidas son secuencias lineales.

Según una realización para determinar la inmunogenicidad de péptidos resultantes de mutaciones, los péptidos inmunogénicos del conjunto de datos positivos se obtienen: obteniendo secuencias de aminoácidos de péptidos capaces de inducir una respuesta de células T; obteniendo secuencias de aminoácidos correspondientes a péptidos del proteoma; comparando las secuencias de aminoácidos que inducen respuesta de células T con las secuencias de aminoácidos del proteoma para determinar coincidencias entre las mismas; comprendiendo el conjunto de datos positivos las secuencias de aminoácidos que inducen respuesta de células T que están estrechamente relacionadas pero no son idénticas a péptidos del proteoma (p. ej., humano). Preferiblemente, "estrechamente relacionadas pero no idénticas a" significa que se permiten 1, 2 o 3 discordancias de aminoácidos entre dichos péptidos. Un experto en la materia conocerá métodos para comparar péptidos problema (péptidos a evaluar) con péptidos del proteoma humano y cómo identificar el número de discordancias. La ventaja de esto es que filtrando péptidos que son idénticos a péptidos del proteoma humano, los péptidos en el conjunto de datos positivos se parecen más estrechamente a los péptidos de interés que son péptidos que tienen una secuencia alterada debido a la mutación.

Según una realización, los péptidos inmunogénicos del conjunto de datos positivos se obtienen: obteniendo secuencias de aminoácidos de péptidos capaces de inducir una respuesta de células T; obteniendo secuencias de aminoácidos correspondientes a péptidos del proteoma; comparando las secuencias de aminoácidos que inducen respuesta de células T con las secuencias de aminoácidos del proteoma para determinar coincidencias entre las mismas; comprendiendo el conjunto de datos positivos las secuencias de aminoácidos que inducen respuesta de células T que tienen 1 o más, 2 o más, 3 o más, 4 o más, 5 o más,... discordancias con péptidos del proteoma (p. ej., humano). Un experto en la técnica conocerá métodos para comparar péptidos problema (péptidos a evaluar) con péptidos del proteoma (humano) y cómo identificar el número de discordancias. Este conjunto de datos positivos puede usarse cuando se requiere un modelo que sea específico para antígenos extraños (p. ej., si se compara con secuencias del proteoma humano, no humano o no mamífero, p. ej., bacteriano o vírico). Dicho conjunto de datos positivos obtenido en relación con péptidos que tienen múltiples discordancias respecto a péptidos del proteoma (humano) puede parecerse a un conjunto de datos positivos obtenido de dichos antígenos extraños (p. ej., bacterianos o víricos).

Los péptidos capaces de inducir respuesta de células T pueden caracterizarse experimentalmente mediante un ensayo de células T humanas. El ensayo de células T puede medir la liberación de citocinas, citotoxicidad o unión cualitativa de células T a una célula presentadora de antígeno (APC). La citocina medida para un ensayo de células T puede seleccionarse del grupo que consiste en interferón gamma (IFN γ), factor de necrosis tumoral alfa (TNF α), interleucina-2 (IL-2), interleucina-4 (IL-4), interleucina-5 (IL-5), interleucina-6 (IL-6), interleucina-8 (IL-8), interleucina-10 (IL-10), interleucina-17 (IL-17), interleucina-21 (IL-21), interleucina-22 (IL-22), granzima A y granzima B. En algunas variaciones, la unión cualitativa de células T puede determinarse mediante tinción de multímero de MHC. En alguna variación, el ensayo de células T puede realizarse ex vivo o in vitro. Los péptidos caracterizados experimentalmente reconocidos por las células T se pueden seleccionar de la Base de Datos de Epítomos Inmunitarios (IEDB).

Según una realización preferida, el conjunto de datos de entrenamiento comprende además un conjunto de datos negativos.

Dicho conjunto de datos negativos puede obtenerse: obteniendo secuencias de aminoácidos de péptidos capaces de unirse a y/o ser presentados en el complejo principal de histocompatibilidad (MHC), preferiblemente identificados por medios tales como ensayos de unión y/o datos de espectrometría de masas; obteniendo secuencias de aminoácidos correspondientes a péptidos del proteoma constitutivo; comparando dichas secuencias de aminoácidos de unión a MHC/presentadas en el MHC con las secuencias de aminoácidos del proteoma constitutivo para determinar coincidencias entre las mismas; comprendiendo el conjunto de datos negativos las coincidencias entre las mismas, es decir, las secuencias de aminoácidos presentadas en el MHC que corresponden a proteínas constitutivas. Dichos ensayos de unión y/o espectrometría de masas para identificar péptidos capaces de unirse a y/o ser presentados en el MHC pueden ser cualquier tipo de ensayos de unión y/o espectrometría de masas conocidos en la técnica. Los tipos de ensayos de unión y/o espectrometría de masas son conocidos por un experto en la técnica. Preferiblemente, las secuencias de aminoácidos presentadas en el MHC obtenidas son secuencias lineales.

En el presente contexto, "proteoma constitutivo" se refiere a un conjunto de proteínas implicadas en las funciones básicas de una célula o conjunto de células en un organismo.

Según una realización para determinar la inmunogenicidad de péptidos que resultan de mutaciones, el conjunto de datos negativos se obtiene: obteniendo secuencias de aminoácidos de péptidos capaces de unirse y/o ser presentados en el complejo principal de histocompatibilidad; obteniendo secuencias de aminoácidos correspondientes a péptidos del proteoma; comparando dichas secuencias de aminoácidos de unión al MHC/presentadas en el MHC con las secuencias de aminoácidos del proteoma para determinar coincidencias entre las mismas; comprendiendo el conjunto de datos negativos las secuencias de aminoácidos de unión al MHC/presentadas en el MHC que están estrechamente relacionadas pero no son idénticas a péptidos del proteoma humano. Preferiblemente, "estrechamente relacionadas pero no idénticas a" significa que se permiten 1, 2 o 3 discordancias de aminoácidos entre dichos péptidos. Un experto en la técnica conocerá métodos para comparar péptidos problema (péptidos a evaluar) con péptidos del proteoma (p. ej., humano) y cómo identificar el número de discordancias. La ventaja de esto es que filtrando péptidos que son idénticos a péptidos del proteoma (humano), los péptidos en el conjunto de datos negativos se parecen más estrechamente a péptidos que tienen una secuencia alterada debido a mutación. Eliendo este conjunto de datos negativos se evita que el modelo de clasificación sea entrenado para distinguir péptidos mutados de péptidos no mutados en lugar de péptidos inmunogénicos de no inmunogénicos. Otra ventaja o ventaja asociada es que el conjunto de datos negativos se parecerá estrechamente a los péptidos que serán probados por el modelo, aumentando de ese modo la especificidad del modelo.

Según una realización, el conjunto de datos negativos se obtiene: obteniendo secuencias de aminoácidos de péptidos capaces de ser presentados en el complejo principal de histocompatibilidad; obteniendo secuencias de aminoácidos correspondientes a péptidos del proteoma; comparando dichas secuencias de aminoácidos presentadas en el MHC con las secuencias de aminoácidos del proteoma para determinar coincidencias entre las mismas; comprendiendo el conjunto de datos negativos las secuencias de aminoácidos presentadas en el MHC que tienen 1 o más, 2 o más, 3 o más, 4 o más, 5 o más,... discordancias con péptidos del proteoma (p. ej., humano). Un experto en la técnica conocerá métodos para comparar péptidos problema (péptidos a evaluar) con péptidos del proteoma (humano) y cómo identificar el número de discordancias. Este conjunto de datos negativos puede usarse cuando se requiere un modelo que sea específico para antígenos extraños (no humanos o no mamíferos, p. ej., bacterianos o víricos). Dicho conjunto de datos negativos obtenido de péptidos que tienen múltiples discordancias respecto a péptidos del proteoma humano puede parecerse a un conjunto de datos negativos obtenido de dichos antígenos extraños (p. ej., - si se compara con secuencias del proteoma humano - de no humanos o no mamíferos, p. ej., bacterianos o víricos).

Según una realización preferida, el conjunto de datos de entrenamiento comprende un conjunto de datos positivos y negativos, en donde el conjunto de datos negativos comprende sustancialmente más registros en comparación con el conjunto de datos positivos. Preferiblemente, en una relación de al menos 3:1. Según otra realización, el conjunto de datos de entrenamiento comprende un conjunto de datos positivos y negativos, en donde el conjunto de datos negativos comprende una cantidad similar de registros en comparación con el conjunto de datos positivos, o menos registros en comparación con el conjunto de datos positivos.

Según una realización preferida, el modelo de clasificación es un algoritmo de aprendizaje automático de clasificación, preferiblemente un algoritmo de aprendizaje automático de clasificación supervisado, más preferiblemente en donde el algoritmo de aprendizaje automático es un algoritmo de clasificación que es uno o más de un clasificador de perceptrón multicapa, un clasificador de árbol de decisión, un clasificador de Bayes ingenuo gaussiano, un clasificador de proceso gaussiano, un clasificador de descenso de gradiente estocástico, una máquina de soporte vectorial lineal, una máquina de soporte vectorial de núcleo, un clasificador de K vecinos más cercanos o un clasificador de bosque aleatorio. Después de la evaluación comparativa, el clasificador de bosque aleatorio muestra los mejores resultados. Además, los autores de la invención observan que los algoritmos de los clasificadores de bosque aleatorio requieren la menor cantidad de trabajo en términos de ingeniería de variables y ajuste de parámetros. Por consiguiente, el modelo de clasificación más preferido según la presente invención es un clasificador de bosque aleatorio.

Una realización sencilla de la invención proporciona preferiblemente, antes del entrenamiento del modelo de clasificación, la realización de un análisis de componentes principales en los índices numéricos de cada uno de los aminoácidos proteinogénicos para obtener varios componentes principales para cada aminoácido proteinogénico analizado. Un "análisis de componentes principales" o abreviado "PCA", como se usa en el presente documento, se

emplea para reconstruir los subespacios de características más discriminativas, que se usan posteriormente como entrada en la clasificación basada en representación para la predicción. El PCA mejora la predicción, así como la procesabilidad de los datos, pero además elimina el ruido si se restringe a un número limitado de componentes principales. La transformación de las propiedades de los aminoácidos en componentes principales reduce la dimensionalidad y establece el foco en las propiedades más distintivas entre los aminoácidos.

Una realización sencilla de la invención proporciona preferiblemente, antes del entrenamiento del modelo de clasificación, la obtención de un vector de características para las secuencias de aminoácidos del conjunto de datos de entrenamiento y la secuencia de aminoácidos de dicho péptido, en donde se obtiene un vector de características para una secuencia de aminoácidos sustituyendo cada aminoácido de dicha secuencia de aminoácidos por uno o más primeros componentes principales correspondientes; en donde el modelo de clasificación se entrena en los vectores de características del conjunto de datos de entrenamiento; y en donde la probabilidad de que dicho péptido sea inmunogénico se determina para el vector de características de dicho péptido. La traducción de los péptidos en vectores de características conserva la información posicional de las propiedades fisicoquímicas.

Preferiblemente, uno o más primeros componentes principales, se debe entender como uno o más de los primeros componentes principales, más preferiblemente los 10 primeros componentes principales, incluso más preferiblemente los 9 primeros componentes principales, incluso más preferiblemente los 8 primeros componentes principales, incluso más preferiblemente los 7 primeros componentes principales, incluso más preferiblemente los 6 primeros componentes principales, incluso más preferiblemente los 5 primeros componentes principales, incluso más preferiblemente los 4 primeros componentes principales, y lo más preferiblemente los 3 primeros componentes principales.

Por consiguiente, un vector de características para una secuencia de aminoácidos se obtiene sustituyendo cada aminoácido de dicha secuencia de aminoácidos por entre 2 y 10 primeros componentes principales correspondientes, así como cualquier intervalo entre los mismos. Lo más preferiblemente, un vector de características para una secuencia de aminoácidos se obtiene sustituyendo cada aminoácido de dicha secuencia de aminoácidos por 3 primeros componentes principales correspondientes. Los autores de la invención observan que, en su experiencia, restringir el vector de características a solamente los tres primeros componentes principales impide requisitos de procesamiento excesivos a la vez que se retiene suficiente información para una predicción muy precisa.

Adicionalmente, antes del análisis de componentes principales, la pluralidad de índices numéricos se transforma preferiblemente en valores z, lo que caracteriza la probabilidad de que cada característica se ajuste a un valor normal.

Según una realización, la invención proporciona además el entrenamiento de un algoritmo de entrenamiento no supervisado para definir adicionalmente características de péptidos inmunogénicos y/o no inmunogénicos.

Según una realización, la invención proporciona además la obtención de propiedades a nivel de péptido, tales como solubilidad del péptido, peso molecular, estabilidad del péptido, estructura 3-dimensional del péptido, localización global del péptido en el genoma (p. ej., codificado como distancia al centro nuclear), similitud con péptido no mutado, etc., que se pueden usar para entrenar adicionalmente el modelo de clasificación.

Como será evidente para un experto en la técnica, el método implementado por ordenador, el sistema informático y el producto de programa informático de la presente invención pueden usarse para identificar péptidos inmunogénicos (epítopos inmunogénicos) dentro de un conjunto de péptidos presentados en el MHC (epítopos) y conducir a la identificación de dianas terapéuticas. Adicionalmente, se puede usar para evaluar los cambios de inmunogenicidad de un péptido tras alterar uno o más aminoácidos de la secuencia de aminoácidos de dicho péptido, en particular para simular cambios de aminoácidos que pueden aumentar o disminuir la inmunogenicidad de péptidos particulares.

Ejemplos y descripción de las figuras

La invención se describe adicionalmente mediante los siguientes ejemplos no limitantes que ilustran adicionalmente la invención, y no pretenden, ni deben interpretarse como, limitantes del alcance de la invención.

Ejemplo 1: Realización preferida

El presente ejemplo se refiere a una realización preferida para predecir la probabilidad de que un péptido presentado en la superficie celular produzca una respuesta inmunitaria.

Este presente ejemplo acepta péptidos de entrada de 9-11 aminoácidos de longitud que son presentados en una superficie celular (determinado por medio de un algoritmo de predicción o experimentalmente).

El resultado del presente ejemplo es una puntuación de probabilidad entre 0 y 1, que describe la probabilidad de que el péptido sea inmunogénico.

Preprocesamiento

Las entradas se preprocesan de la siguiente manera:

Las propiedades fisicoquímicas para los 20 aminoácidos proteínogénicos se descargaron de la base de datos AAindex en "<https://www.genome.jp/aaindex/>" (con fecha 13.2.2017). AAindex es una base de datos de índices numéricos que representan diversas propiedades fisicoquímicas y bioquímicas de aminoácidos y pares de aminoácidos. AAindex consiste en tres secciones ahora: AAindex1 para el índice de aminoácidos de 20 valores numéricos, AAindex2 para la matriz de mutación de aminoácidos y AAindex3 para los potenciales contactos de proteínas estadísticos. Todos los datos se obtienen de la bibliografía publicada. Los valores numéricos de AAindex1 se transformaron en valores z (media establecida en 0 y desviación estándar de 1). Se realizó un análisis de componentes principales en las propiedades fisicoquímicas ajustadas a escala usando el módulo PCA sklearn para reducir las dimensiones de los datos. Cada aminoácido en un péptido se traduce entonces en los valores de los 3 primeros componentes principales del péptido respectivo, dando como resultado un vector de características de longitud $(9-11) \times 3$. Si la longitud del péptido es más corta que 11 aminoácidos, los aminoácidos ausentes se codifican como tres ceros cada uno en el extremo N (inicio del péptido). Por lo tanto, el vector de características es siempre de longitud 11×3 .

Clasificador supervisado:

El presente ejemplo utiliza un módulo clasificador de bosque aleatorio del paquete de python de sklearn, en particular:

```
ClasificadorBosqueAleatorio (n_estimadores=1000, criterio='entropía', profundidad_máx=Ninguna,
división_muestras_mín=3, hoja_muestras_mín=1, características_máx=2, remuestreo=Falso).
```

Clasificador de entrenamiento

El conjunto de datos de entrenamiento incluye péptidos de 9, 10 y 11 aminoácidos de longitud y 3 veces más negativos que los péptidos positivos. El conjunto de datos de entrenamiento se transforma en una matriz. Cada péptido en el conjunto de datos se traducía en un vector de características como se describió anteriormente. Cada fila en la matriz corresponde por lo tanto al vector de características de uno de los péptidos en el conjunto de datos. Esta matriz se usó para entrenar el modelo de bosque aleatorio descrito anteriormente y se guardó el modelo entrenado.

Datos de entrenamiento

Los datos de entrenamiento para entrenar al clasificador supervisado se construyen como sigue:

Datos negativos (péptidos no inmunogénicos):

Los péptidos de genes constitutivos endógenos se supone que son no inmunogénicos ya que son presentados constantemente en todas las células sanas. Por lo tanto, se construye un conjunto de datos negativos de entrenamiento a partir de péptidos que proceden de dichos genes constitutivos que se ha demostrado que son presentados en moléculas del MHC. La predicción de presentación en el MHC no es el objetivo del presente modelo. Además, no se desea que sea un factor en el modelo. Por lo tanto, los péptidos positivos, así como los péptidos negativos se unen al MHC. Los datos de IEDB de péptidos que se unen al MHC se importaron de "https://www.iedb.org/database_export_v3.php" (versión 19 de febrero de 2020). Los péptidos también se filtraron para ser una secuencia lineal, de homo sapiens, dar como resultado al menos un ensayo positivo y ser de 9, 10 u 11 aminoácidos de longitud. La lista de péptidos resultante contiene solamente péptidos únicos.

Además, se recuperaron genes constitutivos de "<https://www.proteinatlas.org/human-proteome/tissue/housekeeping>". Estos se filtraron para excluir genes anotados con "implicación en la enfermedad" o "especificidad de cáncer de ARN". Las secuencias Fasta de las proteínas de genes constitutivos se recuperaron de Uniprot basándose en las ID de genes de la etapa anterior y se construyó una base de datos Blast a partir de esas secuencias. Los péptidos que se unen al MHC de IEDB se consultaron entonces frente a la base de datos Blast de genes constitutivos. Todas las coincidencias perfectas de péptidos (péptidos resultantes de estos genes constitutivos) se retuvieron como conjunto de datos negativos.

Conjunto de datos positivos (péptidos inmunogénicos):

Los datos positivos (es decir, péptidos inmunogénicos) se obtienen de la base de datos IEDB (con fecha 19 de febrero de 2020): péptidos con datos de ensayo de células T asociados. Los péptidos se filtraron adicionalmente para ser una secuencia lineal, sometidos a ensayo para homo sapiens, dar como resultado al menos un ensayo positivo y ser de 9 aminoácidos de longitud. La lista de péptidos resultante contiene solamente péptidos únicos. Después de filtrar los péptidos, se asegura que no hay solapamiento con el proteoma humano filtrando las coincidencias perfectas con proteínas humanas. Esto se logra filtrando péptidos con coincidencias perfectas con la base de datos Blast hecha a partir del proteoma humano.

La Figura 1 ilustra el flujo de trabajo del presente ejemplo como se discutió anteriormente. El conjunto de datos de entrenamiento (1) comprende tanto un conjunto de datos positivos (102) como negativos (103). El conjunto de datos negativos se obtiene comparando las secuencias de aminoácidos presentadas en el MHC/secuencias de péptidos capaces de unirse a y/o ser presentadas en el MHC (104) con las secuencias de aminoácidos constitutivas (105) para determinar las coincidencias (106) entre las mismas, en donde el conjunto de datos negativos comprende las secuencias de aminoácidos de dichas coincidencias. El conjunto de datos positivos se obtiene comparando las secuencias de aminoácidos que inducen respuesta de células T (107) con las secuencias de aminoácidos del proteoma (108) para

determinar las coincidencias (109) entre las mismas, en donde el conjunto de datos positivos comprende las secuencias de aminoácidos que inducen respuesta de células T excepto la secuencia de aminoácidos de dichas coincidencias. Tanto el conjunto de datos negativos como positivos se compilan (110) 3:1 en el conjunto de datos de entrenamiento.

La figura 1 muestra además el preprocesamiento según el presente ejemplo. En primer lugar, se obtiene una pluralidad de índices numéricos (111), cada uno relacionado con una propiedad fisicoquímica, para cada aminoácido proteinogénico. En segundo lugar, para cada uno de los aminoácidos proteinogénicos, se realiza un análisis de componentes principales (112) en dichos índices numéricos. Por lo tanto, se determinan los componentes numéricos principales (113) para cada uno de los aminoácidos proteinogénicos. Estos pueden usarse posteriormente para un péptido para caracterizar la traducción tanto de los péptidos en el conjunto de datos de entrenamiento (114), así como para el péptido de interés (115, 116). La matriz de características resultante (117) se usa para entrenar (118) el clasificador de bosque aleatorio (119) para obtener un clasificador entrenado (120), que puede usarse para predecir (122) la inmunogenicidad para el péptido de interés por medio del vector de características (121) así determinado. La predicción obtenida según el presente ejemplo se muestra preferiblemente como una puntuación (123).

La Figura 2 muestra la característica operativa del receptor (ROC) del presente ejemplo. Por consiguiente, la tasa de verdaderos positivos (proporción de epítomos inmunogénicos reales que se denominaron como tales, es decir, sensibilidad) se muestra en función de la tasa de falsos positivos (proporción de epítomos no inmunogénicos denominados inmunogénicos, es decir, 1-especificidad). La curva de precisión-sensibilidad muestra el equilibrio entre precisión y sensibilidad para umbrales diferentes. La superficie del área bajo la curva (AUC) de la curva ROC es 0.88.

La Figura 4 muestra resultados de ROC de entrenamiento con diferentes relaciones de datos positivos a negativos, mientras se valida en una relación 1:3. Aumentando la relación de datos negativos a datos positivos, se mejoran los resultados.

La Figura 5 muestra métricas para entrenamiento y validación usando 1,2,3,5,10 o 20 de los componentes principales, respectivamente. El rendimiento aumenta cuando se usan 2 o más componentes principales. Sin embargo, cuando se usan más de 10 componentes principales, el procesamiento aumenta significativamente.

Ejemplo 2: Evaluación comparativa de clasificadores

El presente ejemplo se refiere al rendimiento de algoritmos de clasificación supervisada disponibles en scikit-learn.

La Figura 3 muestra la característica operativa del receptor (ROC) y la evaluación comparativa de precisión-sensibilidad de diferentes clasificadores conocidos en la técnica. Las características del flujo de trabajo del Ejemplo 1 se siguen para la implementación de cada uno de los clasificadores.

Los algoritmos de clasificación probados incluyen: un clasificador de perceptrón multicapa (MLP) (Figura 3; 1*1), un clasificador de bosque aleatorio (Figura 3; 1*2), un clasificador de K vecinos más cercanos (Figura 3; 1*3), un clasificador de Bayes ingenuo gaussiano (Figura 3; 2*1), un clasificador de proceso gaussiano (Figura 3; 2*2), un clasificador de soporte vectorial (Figura 3; 2*3). Los valores de AUC correspondientes son: 0.77; 0.88; 0.73; 0.74; 0.82; 0.82.

Aunque todos los clasificadores son eficaces con el flujo de trabajo actual, el clasificador de bosque aleatorio supera claramente a los otros clasificadores probados.

Ejemplo 3: Evaluación comparativa de la técnica anterior

El presente ejemplo se refiere a una evaluación comparativa del flujo de trabajo del Ejemplo 1 respecto a otros clasificadores de péptidos inmunogénicos conocidos en la técnica.

El flujo de trabajo según el Ejemplo 1 se comparó en un estudio comparativo con otras dos herramientas de predicción de inmunogenicidad disponibles públicamente, particularmente el método predictivo de inmunogenicidad de IEDB e Ineo-Epp.

El método predictivo de inmunogenicidad de IEDB es un modelo que clasifica péptidos basándose en la posición y propiedades de sus aminoácidos. Ineo-Epp es un clasificador de bosque aleatorio entrenado en péptidos inmunogénicos de HLA-I de los que se extrajeron varias características, como propiedades fisicoquímicas de aminoácidos y puntuación de rango de percentil de probabilidad de ligando eluido (rango EL(%)). Ineo-Epp usa una lista de péptidos, junto con un alelo de HLA como entrada. Para la predicción de antígenos, se pueden usar los nueve supertipos principales de HLA-I (A1, A2, A3, A24, B7, B27, B44, B58, B62).

Para este estudio comparativo, se obtuvieron péptidos inmunogénicos de la base de datos de epítomos inmunitarios (IEDB) (a la que se accede el 15.06.2020). Para evitar un sesgo en los datos comparativos para los epítomos presentes en los datos de entrenamiento de cualquiera de las herramientas, se obtuvieron datos de epítomos de antígenos de publicaciones recientes de 2020. Los datos se filtraron adicionalmente (ensayos de células T positivos, humanas, de epítomo lineal) y se limpiaron de epítomos que ya se habían depositado en versiones anteriores de la base de datos. Esto redujo el número de péptidos positivos de 2213 a 64 epítomos únicos, todos de 9 aminoácidos de longitud. Como datos negativos, se usaron los péptidos no inmunogénicos publicados en "Chowell, D. et al. TCR contact residue

hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1754-E1762 (2015)". Los péptidos no inmunogénicos en el conjunto de datos son autopéptidos presentados en el MHC-I eluidos con ligando que se han procesado antigénicamente y unidos a MHC. Se filtró para una longitud de 9-11 aminoácidos, dando como resultado 4254 no epítotos únicos.

- 5 Como todos los péptidos en el conjunto de datos comparativo son ligantes de MHC, se podrían usar todos como entrada para el flujo de trabajo del Ejemplo 1 y el método predictivo de inmunogenicidad de IEDB. Para Ineo-Epp, se realizaron predicciones para los nueve supertipos de HLA-I y solo se consideraron las puntuaciones para los ligantes predichos.

10 A partir de las puntuaciones de predicción de las tres herramientas, se construyeron curvas de la característica operativa del receptor (ROC) y se muestran en la Figura 6. Las curvas muestran la tasa de predicciones de verdaderos positivos y de falsos positivos para una puntuación de punto corte decreciente.

En la comparación de las tres herramientas, el flujo de trabajo según el Ejemplo 1 (41) superó al método predictivo de inmunogenicidad de IEDB (43) y al Ineo-Epp (42). El AUC del Ejemplo 1 era de 0.84, que está cerca del AUC de validación interna de 0.88. El método predictivo de inmunogenicidad de IEDB tenía un AUC de 0.57. Ineo-Epp también tenía un menor rendimiento (AUC de 0.65).

15 Ejemplo 4: Vacuna contra el cáncer

Para demostrar la validez de la puntuación de inmunogenicidad del esquema de trabajo descrito en el ejemplo 1, el esquema de trabajo se aplica a datos de un estudio publicado sobre una vacuna contra el cáncer de mama.

20 El estudio, es decir, "Dillon, P. M. et al. A pilot study of the immunogenicity of a 9-peptide breast cancer vaccine plus poly-ICLC in early stage breast cancer. *J. Immunother. Cancer* 5, 1-10 (2017)", evaluaba la respuesta inmunitaria de doce pacientes con cáncer de mama a una vacuna que consistía en nueve péptidos asociados al cáncer de mama restringidos por MHC de clase I. Las respuestas de células T CD8+ a la vacuna se evaluaron mediante ensayos ELISpot de interferón gamma tanto directos como estimulados. Dicho ensayo ELISpot evalúa cuantitativamente la frecuencia de secreción de citocinas para una célula en respuesta a un estimulante, y es por lo tanto un método adecuado para evaluar la inmunogenicidad de dicha vacuna.

25 En los ensayos de ELISpot estimulados, se detectaron respuestas de células T CD8+ específicas de péptido en cuatro de once pacientes evaluables. Se observaron dos respuestas ELISpot en respuesta al péptido HLA-A2 CEA571-579 modificado (antígeno YLS-D) y se observaron dos respuestas en respuesta a HLA-A3 CEA27-35 (antígeno HLF). Un paciente tuvo una respuesta límite a HLA-A3 MAGE-A196-104.

30 El esquema de trabajo del presente ejemplo se usó para puntuar la inmunogenicidad de cada uno de los nueve péptidos en la vacuna (Tabla 1). Como muestra la tabla 1, los dos antígenos que dieron como resultado ensayos ELISpot positivos (antígeno YLS-D y HLF) tenían las puntuaciones de esquema de trabajo/inmunogenicidad más altas.

Tabla 1 Péptidos de vacuna con respuestas en ELISpot y puntuación del esquema de trabajo

Alelo	Secuencia	Epítoto	ELISpot	Puntuación
HLA-A1	EADPTGHSY	MAGE-A1 ₁₆₁₋₁₆₉	-	0.11
	EVDPIGHLY	MAGE-A3 ₁₆₈₋₁₇₆	-	0.13
HLA-A2	KIFGSLAFL	Her-2/neu ₃₆₉₋₃₇₇	-	0.39
	YLSGADLNL	CEA ₅₇₁₋₅₇₉	2 respuestas	0.49
	GLYDGMHL	MAGE-A1 ₁₀₂₅₄₋₂₆₂	-	0.30
HLA-A3	HLFGYSWYK	CEA ₂₇₋₃₅	2 respuestas	0.39
	VLRENTSPK	Her-2/neu ₇₅₄₋₇₆₂	-	0.25
	SLFRAVITK	MAGE-A1 ₉₆₋₁₀₄	1 respuesta límite	0.28
HLA-A3/HLA-A31	ASGPGGGAPR	NY-ESO-1 ₅₃₋₆₂	-	0.22

Ejemplo 5: Inmunogenicidad vírica

35 Las predicciones de inmunogenicidad de epítotos derivados de proteínas víricas permiten el cribado rápido de nuevos virus en objetivos de vacunación.

El clasificador de inmunogenicidad descrito en el ejemplo 1 se entrena en péptidos inmunogénicos de cepas de virus

ya probadas y se puede usar para precubar la lista completa de epítomos potenciales de una nueva cepa de virus. Reduce el número de ensayos in vitro necesarios para esta cepa particular, minimizando de manera efectiva los plazos y costes asociados. Adicionalmente, ayuda a optimizar la vacuna para péptidos objetivo que es probable que causen una respuesta inmunitaria en una parte más grande de la población.

5 Para mostrar la aplicabilidad del presente algoritmo para péptidos víricos, se utilizó para comparar epítomos inmunogénicos de SARS-CoV-2 con epítomos de SARS-CoV y MERS-CoV. Estos tres virus están estrechamente relacionados, pero provocan síntomas de órdenes de magnitud diferentes. De hecho, los síntomas de SARS-CoV son más graves (aproximadamente 10% de mortalidad) que para el coronavirus actual (estimado en 1,3%), mientras que MERS-CoV fue incluso más mortal que ambos (20,4-69,2% de mortalidad).

10 Después de reunir las secuencias de las proteínas de la envuelta de SARS-CoV-2 así como para SARS-CoV y MERS-CoV, se extrajeron todos los 9-meros posibles. Estos se analizaron a través del algoritmo neoMS para determinar cuáles era probable que fueran presentados en la superficie de células infectadas. Posteriormente, el algoritmo de la presente invención clasificó los péptidos presentados según su inmunogenicidad predicha. Se examinó la distribución de probabilidad de inmunogenicidad resultante de este conjunto reducido de péptidos presentados y, como puede verse en la Figura 7, SARS-CoV-2 presenta en promedio menos péptidos potencialmente inmunogénicos que el SARS-CoV, que a su vez tiene menos péptidos inmunogénicos probables que MERS-CoV. Esto implica que la carga inmunogénica de estos tres virus parece correlacionarse con su patogenicidad corroborando de este modo la hipótesis por "Mcmanus, L. M. y Mitchell, R. N. Pathobiology of human disease: a dynamic encyclopedia of disease mechanisms. en Pathobiology of human disease: a dynamic encyclopedia of disease mechanisms. 1036 (Elsevier, 2014)". La Figura 7 muestra el histograma y curvas de densidad asociadas de probabilidad de inmunogenicidad para los péptidos presentados en la superficie de células derivados de la envoltura de SARS-CoV-2 (rojo), SARS-CoV (azul) y MERS-CoV (verde).

Ejemplo 6: Uso diferencial de aminoácidos en péptidos en los que se basa el conjunto de datos positivos y negativos de la presente invención

25 Como se ha comentado anteriormente, los conjuntos de datos positivos y negativos se basan en péptidos que se supone que producen una respuesta inmunitaria (conjunto de datos positivos) o que se supone que no producen una respuesta inmunitaria (conjunto de datos negativos).

El análisis de uso diferencial de aminoácidos (dagLogo: <https://doi.org/10.1371/journal.pone.0242030>) evaluaba si existen patrones de secuencia conservados dentro de los péptidos del conjunto positivo y negativo y si estos patrones difieren entre los dos grupos. Para esto, el conjunto de datos positivos se basó en péptidos inmunogénicos sin coincidencias con el proteoma humano, mientras que el conjunto de datos negativos se basó en péptidos presentados en el MHC sin coincidencias con el proteoma humano.

30 La Figura 8 muestra logotipos de secuencias peptídicas que muestran el uso diferencial de aminoácidos de péptidos de 9 aminoácidos de longitud en el conjunto de datos positivos inmunogénico (arriba) frente a péptidos de 9 aminoácidos de longitud en el conjunto de datos negativos no inmunogénico (abajo). Los aminoácidos estaban no agrupados (Fig. 8A), agrupados por propiedades químicas (Fig. 8B, ácidos, básicos, amfipáticos, hidroxilicos, que contienen azufre, aromáticos o alifáticos), carga eléctrica (Fig. 8C; positiva, negativa o neutra) o tamaño (Fig. 8D; muy pequeños, pequeños, medios, grandes o grandísimos).

A partir de la figura, está claro que hay diferencias fisicoquímicas entre ambos conjuntos de péptidos, en los que se construye y entrena el modelo. Además, también muestra que ciertas propiedades de aminoácidos o aminoácidos en ciertas posiciones están correlacionadas con la inmunogenicidad.

Ejemplo 7: Ensayo de la eficiencia del método implementado por ordenador de la presente invención en un conjunto de datos recién generado y comparación con otros métodos disponibles

45 Se realizó un ensayo de inmunogenicidad experimental de 33 péptidos mediante el ensayo ELISpot en células T de donantes sanos. Los ensayos se realizaron con dos rondas de estimulación a través de células dendríticas autólogas cargadas con péptidos. Todos los ensayos se realizaron por duplicado o triplicado.

Se predijo que todos los péptidos eran presentados en al menos dos alelos de HLA de donante usando netMHCpan4.0, una herramienta para la predicción de la unión de péptidos a una molécula de MHC usando redes neuronales artificiales. Esta herramienta se entrena en ligandos eluidos de forma natural y en datos de afinidad de unión. Devuelve dos propiedades: la probabilidad de que un péptido se convierta en un ligando natural, o la afinidad de unión predicha.

50 Un resultado positivo en ELISpot se definió como un mínimo de 25 puntos por 5×10^4 células y un aumento de al menos 2x con respecto al control. De los 33 péptidos que se predijo que eran todos presentados, ocho péptidos (23.5%) dieron positivo mediante el ensayo ELISpot.

55 Se analizó la inmunogenicidad de los mismos péptidos usando el método implementado por ordenador de la presente invención. El conjunto de datos positivos se basó en péptidos inmunogénicos sin coincidencias con el proteoma humano, mientras que el conjunto de datos negativos se basó en péptidos presentados en el MHC sin coincidencias con el proteoma humano.

Los diferentes sistemas de puntuación:

- Puntuación del método de la presente invención: usando el método implementado por ordenador de la invención, se calculó una puntuación de inmunogenicidad para cada péptido.

5 - Puntuación de inmunogenicidad de IEDB: usando el método predictivo de inmunogenicidad de IEDB, se calculó una puntuación para cada péptido.

- Puntuación de rango de netMHCpan (HLA-A*02:01): usando netMHCpan4.0, se calculó la puntuación de rango de netMHCpan para cada péptido para el alelo HLA-A*02:01.

10 - Puntuación de rango de netMHCpan (todos los HLA de donante): usando netMHCpan4.0, se calculó la puntuación de rango de netMHCpan para cada péptido para todos los alelos de HLA de donante y se consideró la puntuación más baja.

15 La Figura 9A muestra la probabilidad predicha de que los péptidos sean inmunogénicos como se ensayó usando el método implementado por ordenador de la invención de grupos de péptidos con un resultado en ELISpot positivo o negativo ($p = 0.019$). A partir de esos resultados, está claro que los péptidos con un resultado en ELISpot positivo ($n = 8$) tenían una puntuación de inmunogenicidad significativamente mayor usando el método de la presente invención en comparación con los péptidos con un resultado negativo ($n = 25$). El método implementado por ordenador de la presente invención es por lo tanto capaz de diferenciar entre péptidos inmunogénicos y no inmunogénicos ($p=0.019$). La diferencia no fue significativa para las herramientas IEDB y netMHCpan.

20 La Figura 9B muestra una comparación de curvas de característica operativa del receptor (ROC) de múltiples métodos predictivos de inmunogenicidad diferentes, incluyendo el método implementado por ordenador de la invención, el método predictivo de inmunogenicidad de IEDB y netMHCpan4.0, basado este último en un alelo de HLA específico o en todos los alelos de HLA de un sujeto. Los valores del área bajo la curva (valores de AUC) para las curvas ROC eran:

Herramienta predictiva	Valor de AUC
Método implementado por ordenador de la invención	0.80
Método predictivo de inmunogenicidad de IEDB	0.47
netMHCpan4.0 HLA-A02:01	0.52
netMHCpan4.0 todos los HLA donantes	0.42

25 Los resultados muestran que la precisión de diferenciación entre péptidos inmunogénicos y no inmunogénicos del método implementado por ordenador de la presente invención es mejor que la precisión de los otros métodos, tales como el método predictivo de inmunogenicidad de IEDB y netMHCpan4.0, prediciendo este último únicamente la afinidad de unión y no la inmunogenicidad.

30 En una etapa posterior, los péptidos se ordenaron de puntuaciones de inmunogenicidad más altas a más bajas tal como se obtuvieron mediante el método de la presente invención y se presentaron en una curva de elevación. Todos los péptidos positivos ensayados por ELISpot estaban presentes en el 60% superior de los péptidos, haciendo posible descartar el 40% de todos los péptidos sin perder ningún verdadero positivo, como se muestra en la Figura 9C, que muestra el porcentaje de verdaderos positivos en la población (como ganancia acumulativa). La curva de ganancia acumulativa ilustra qué porcentaje de los verdaderos positivos se identifica con la porción más pequeña posible de muestras. En comparación con el método predictivo de inmunogenicidad de IEDB y netMHCpan4.0, el uso del método de la presente invención da como resultado un aumento significativo de la especificidad de 23% a 40% sin perder sensibilidad, reduciendo sustancialmente el número de neoantígenos candidatos.

35 En conclusión, el método implementado por ordenador de la presente invención permite clasificar los neoepítomos presentados mejor por su potencial inmunogénico para mejorar la predicción de neoepítomos y reducir los falsos positivos significativamente. Priorizar neoepítomos con inmunogenicidad predicha alta mediante el método implementado por ordenador de la presente invención aumenta la probabilidad de seleccionar neoantígenos susceptibles de actuación. El método implementado por ordenador de la presente invención es, por lo tanto, una herramienta poderosa tanto en la mejora de la selección de objetivos para inmunoterapias personalizadas como en la oferta de carga de péptidos inmunogénicos como biomarcador mejorado para la supervivencia del paciente y respuesta a la terapia.

45 Ejemplo 8: Identificación de pacientes con baja carga mutacional tumoral que no obstante se beneficiarían de la terapia con inhibidores de puntos de control inmunitario

Un biomarcador comúnmente usado para seleccionar pacientes para recibir terapia con inhibidores de puntos de control inmunitarios (ICI), una forma de terapia contra el cáncer que tiene como objetivo reactivar el sistema inmunitario de un paciente, es una carga mutacional tumoral (TMB) de >100 mutaciones (ns = no sinónimas). Se ha demostrado que una TMB alta está correlacionada con beneficio clínico de los ICI. Sin embargo, hay algunos pacientes con baja TMB, que no obstante se beneficiarían de ICI y es crucial identificar esos pacientes.

Se investigaron tres biomarcadores potenciales para la respuesta a la terapia en una cohorte de pacientes con melanoma de baja TMB que recibían ICI.

El primer biomarcador era la TMB (el número total de mutaciones no sinónimas). La cohorte se dividió en dos grupos de igual tamaño basados en la TMB, "Alto" que contenía el 50% de los pacientes con el número más alto de mutaciones y "Bajo" que contenía el 50% de los pacientes con el número más bajo de mutaciones.

La Figura 10 muestra una curva de Kaplan Meier, que ilustra la probabilidad de un suceso en un intervalo de tiempo respectivo. Se puede observar que la TMB (Fig. 10 panel izquierdo) no es un buen biomarcador para identificar pacientes que muestran una respuesta a la terapia.

Posteriormente, el experimento examinó el número de mutaciones presentadas por paciente. El número de mutaciones presentadas se obtuvo ejecutando netMHCpan4.1 para todos los péptidos posibles de 9-11 aminoácidos de longitud que resultaban de cada mutación. Se consideró que una mutación estaba presente si se identificaba al menos un péptido con una puntuación de rango <2. La cohorte se dividió en dos grupos de igual tamaño basándose en el número de mutaciones presentadas, "Alto" que contenía el 50% de los pacientes con el número más alto de mutaciones presentadas y "Bajo" que contenía el 50% de los pacientes con el número más bajo de mutaciones presentadas.

El panel central de la Fig. 10 muestra que el número de mutaciones presentadas, similar a TMB, no es un buen biomarcador para identificar pacientes que muestran una respuesta a la terapia.

Finalmente, el experimento examinó el número de mutaciones inmunogénicas por paciente. El número de mutaciones inmunogénicas se obtuvo ejecutando netMHCpan4.1 para todos los péptidos posibles de 9-11 aminoácidos de longitud que resultaban de cada mutación. Para todos los péptidos considerados que eran presentados (puntuación de rango <2) se calculó una puntuación de inmunogenicidad usando el método implementado por ordenador de la presente invención. Una mutación se consideró inmunogénica si se identificaba al menos un péptido con una puntuación de inmunogenicidad > 0.75. La cohorte se dividió en dos grupos de igual tamaño basándose en el número de mutaciones inmunogénicas, "Alto" que contenía el 50% de los pacientes con el número más alto de mutaciones inmunogénicas y "Bajo" que contenía el 50% de los pacientes con el número más bajo de mutaciones inmunogénicas.

El panel derecho de la Fig. 10 muestra que el número de mutaciones inmunogénicas, es un mejor biomarcador para identificar pacientes que muestran una respuesta a la terapia que la TMB o el número de mutaciones presentadas.

Esto muestra que no solo el simple número de mutaciones en un tumor dirige la respuesta a la terapia, especialmente en pacientes con un número bajo de mutaciones, sino también el potencial inmunogénico de estas mutaciones. Por lo tanto, el método implementado por ordenador de la presente invención puede usarse para identificar pacientes con baja carga mutacional tumoral que no obstante se beneficiarían de la terapia con inhibidores de puntos de control inmunitarios, basándose en el potencial inmunogénico de las mutaciones.

La presente invención no se limita en modo alguno a las realizaciones descritas en los ejemplos y/o mostradas en las figuras. Por el contrario, los métodos según la presente invención pueden realizarse de muchas maneras diferentes sin apartarse del alcance de la invención.

REIVINDICACIONES

1. Método implementado por ordenador para determinar la inmunogenicidad de un péptido, preferiblemente un péptido presentado en la superficie celular, que comprende las etapas de:

- 5 - obtener una secuencia de aminoácidos de dicho péptido, en donde dicho péptido comprende aminoácidos proteinogénicos y/o no proteinogénicos;
- obtener para cada aminoácido proteinogénico una pluralidad de índices numéricos, cada uno relacionado con una propiedad fisicoquímica;
- 10 - obtener un conjunto de datos de entrenamiento que comprende un conjunto de datos positivos y un conjunto de datos negativos, en donde el conjunto de positivos comprende datos (numéricos) relacionados con una pluralidad de secuencias de aminoácidos de péptidos inmunogénicos, en donde el conjunto de negativos comprende datos (numéricos) relacionados con una pluralidad de secuencias de aminoácidos de péptidos no inmunogénicos;
- entrenar un modelo de clasificación matemático en el conjunto de datos de entrenamiento;
- determinar una probabilidad de que dicho péptido produzca una respuesta inmunitaria por medio del modelo de clasificación entrenado;
- 15 en donde el método comprende las etapas de:
- realizar un análisis de componentes principales en los índices numéricos de cada uno de los aminoácidos proteinogénicos para obtener componentes principales para cada aminoácido proteinogénico analizado;
- obtener un vector de características para las secuencias de aminoácidos del conjunto de datos de entrenamiento y la secuencia de aminoácidos de dicho péptido, en donde se obtiene un vector de características para una secuencia de aminoácidos sustituyendo cada aminoácido de dicha secuencia de aminoácidos por uno o más componentes principales correspondientes;
- 20 en donde el modelo de clasificación se entrena en los vectores de características del conjunto de datos de entrenamiento;
- en donde la probabilidad de que dicho péptido sea inmunogénico se determina para el vector de características de dicho péptido;
- 25 caracterizado por que el conjunto de datos negativos que comprende datos (numéricos) relacionados con una pluralidad de secuencias de aminoácidos de péptidos no inmunogénicos, se obtiene mediante
- obtención de secuencias de aminoácidos de péptidos capaces de unirse a y/o ser presentados en un MHC;
- obtención de secuencias de aminoácidos correspondientes a péptidos del proteoma;
- 30 - comparación de las secuencias de aminoácidos de péptidos capaces de unirse a y/o ser presentados en el MHC con una secuencia de aminoácidos del proteoma para determinar coincidencias entre las mismas;
- comprendiendo el conjunto de datos negativos la secuencia de aminoácidos de péptidos capaces de unirse a y/o ser presentados en el MHC que están estrechamente relacionados pero no son idénticos a péptidos del proteoma, teniendo dichas secuencias de aminoácidos estrechamente relacionadas 1, 2 o 3 discordancias de aminoácidos en comparación con péptidos del proteoma.
- 35
2. Método según la reivindicación precedente 1, en donde el conjunto de datos negativos que comprende datos (numéricos) relacionados con una pluralidad de secuencias de aminoácidos de péptidos no inmunogénicos, se obtiene mediante:
- obtención de secuencias de aminoácidos de péptidos capaces de unirse a y/o ser presentados en un complejo mayor de histocompatibilidad (MHC), preferiblemente identificados mediante ensayos de unión y/o espectrometría de masas;
- obtención de secuencias de aminoácidos correspondientes a proteínas constitutivas;
- comparación de las secuencias de aminoácidos de péptidos capaces de unirse a y/o ser presentados en el MHC con las secuencias de aminoácidos correspondientes a proteínas constitutivas para determinar coincidencias entre las mismas;
- 40
- comprendiendo el conjunto de datos negativos las secuencias de aminoácidos de dichas coincidencias.
- 45
3. Método según la reivindicación precedente 1 o 2, en donde las secuencias de aminoácidos presentadas en el MHC obtenidas son secuencias lineales.

4. Método según cualquiera de las reivindicaciones precedentes 1 a 3, en donde los péptidos inmunogénicos del conjunto de datos positivos se obtienen mediante:

- obtención de secuencias de aminoácidos de péptidos capaces de inducir una respuesta de células T;
- obtención de secuencias de aminoácidos que corresponden a péptidos del proteoma;

5 - comparación de las secuencias de aminoácidos que inducen respuesta de células T con las secuencias de aminoácidos del proteoma para determinar coincidencias entre las mismas;

comprendiendo el conjunto de datos positivos las secuencias de aminoácidos que inducen respuesta de células T, excepto la secuencia de aminoácidos de dichas coincidencias.

10 5. Método según cualquiera de las reivindicaciones precedentes 1 a 3, en donde los péptidos inmunogénicos del conjunto de datos positivos se obtienen mediante:

- obtención de secuencias de aminoácidos de péptidos capaces de inducir una respuesta de células T;
- obtención de secuencias de aminoácidos correspondientes a péptidos del proteoma;
- comparación de las secuencias de aminoácidos que inducen respuesta de células T con secuencias de aminoácidos del proteoma para determinar coincidencias entre las mismas;

15 comprendiendo el conjunto de datos positivos las secuencias de aminoácidos que inducen respuesta de células T que están estrechamente relacionadas, pero no son idénticas a los péptidos del proteoma, teniendo dichas secuencias de aminoácidos estrechamente relacionadas 1, 2 o 3 discordancias de aminoácidos en comparación con los péptidos del proteoma.

20 6. Método según cualquiera de las reivindicación precedente 4 o 5, en donde las secuencias de aminoácidos que inducen respuesta de células T son secuencias lineales.

7. Método según cualquiera de las reivindicaciones precedentes, en donde el conjunto de datos de entrenamiento comprende un conjunto de datos positivos y negativos, en donde el conjunto de datos negativos comprende sustancialmente más registros en comparación con el conjunto de datos positivos.

25 8. Método según cualquiera de las reivindicaciones precedentes, en donde se obtiene un vector de características para una secuencia de aminoácidos sustituyendo cada aminoácido de dicha secuencia de aminoácidos por al menos 2 primeros componentes principales correspondientes, preferiblemente entre 2 y 10 primeros componentes principales correspondientes, lo más preferiblemente 3 primeros componentes principales correspondientes.

9. Método según cualquiera de las reivindicaciones precedentes, en donde antes del análisis de componentes principales la pluralidad de índices numéricos se transforma en valores z.

30 10. Método según cualquiera de las reivindicaciones precedentes, en donde el modelo de clasificación es un algoritmo de aprendizaje automático de clasificación supervisada, más preferiblemente en donde el algoritmo de aprendizaje automático es uno o más de un clasificador de perceptrón multicapa, un clasificador de Bayes ingenuo gaussiano, una máquina de soporte vectorial lineal, una máquina de soporte vectorial de núcleo, un clasificador de K vecinos más cercanos o un clasificador de bosque aleatorio, lo más preferiblemente en donde el algoritmo de aprendizaje automático es un clasificador de bosque aleatorio.

35 11. Sistema informático para determinar la inmunogenicidad de un péptido, estando configurado el sistema informático para realizar el método implementado por ordenador según una cualquiera de las reivindicaciones precedentes.

40 12. Producto de programa informático para determinar la inmunogenicidad de un péptido, comprendiendo el producto de programa informático instrucciones que, cuando el producto de programa informático es ejecutado por un ordenador, hacen que el ordenador lleve a cabo el método implementado por ordenador según una cualquiera de las reivindicaciones precedentes.

45 13. Uso del método implementado por ordenador según cualquiera de las reivindicaciones precedentes 1 a 10 y/o el sistema informático según la reivindicación precedente 11 y/o el producto de programa informático según la reivindicación precedente 12, para determinar un tratamiento del cáncer de un sujeto, preferiblemente determinando la inmunogenicidad de un neoantígeno presentado por una célula tumoral del sujeto.

14. Uso del método implementado por ordenador según cualquiera de las reivindicaciones precedentes 1 a 10 y/o del sistema informático según la reivindicación precedente 11 y/o del producto de programa informático según la reivindicación precedente 12, para el cribado de un virus o bacteria en un objetivo de vacunación, preferiblemente determinando la inmunogenicidad de epítomos de proteínas víricas o bacterianas.

15. Uso del método implementado por ordenador según cualquiera de las reivindicaciones precedentes 1 a 10 y/o el sistema informático según la reivindicación precedente 11 y/o el producto de programa informático según la reivindicación precedente 12, para caracterizar reacciones de autoinmunidad de un sujeto, preferiblemente determinando la inmunogenicidad de autoantígenos presentados por una célula del sujeto.

- 5 16. Uso del método implementado por ordenador según cualquiera de las reivindicaciones precedentes 1 a 10 y/o el sistema informático según la reivindicación precedente 11 y/o el producto de programa informático según la reivindicación precedente 12, para evaluar los cambios de inmunogenicidad de un péptido tras alterar uno o más aminoácidos de la secuencia de aminoácidos de dicho péptido.

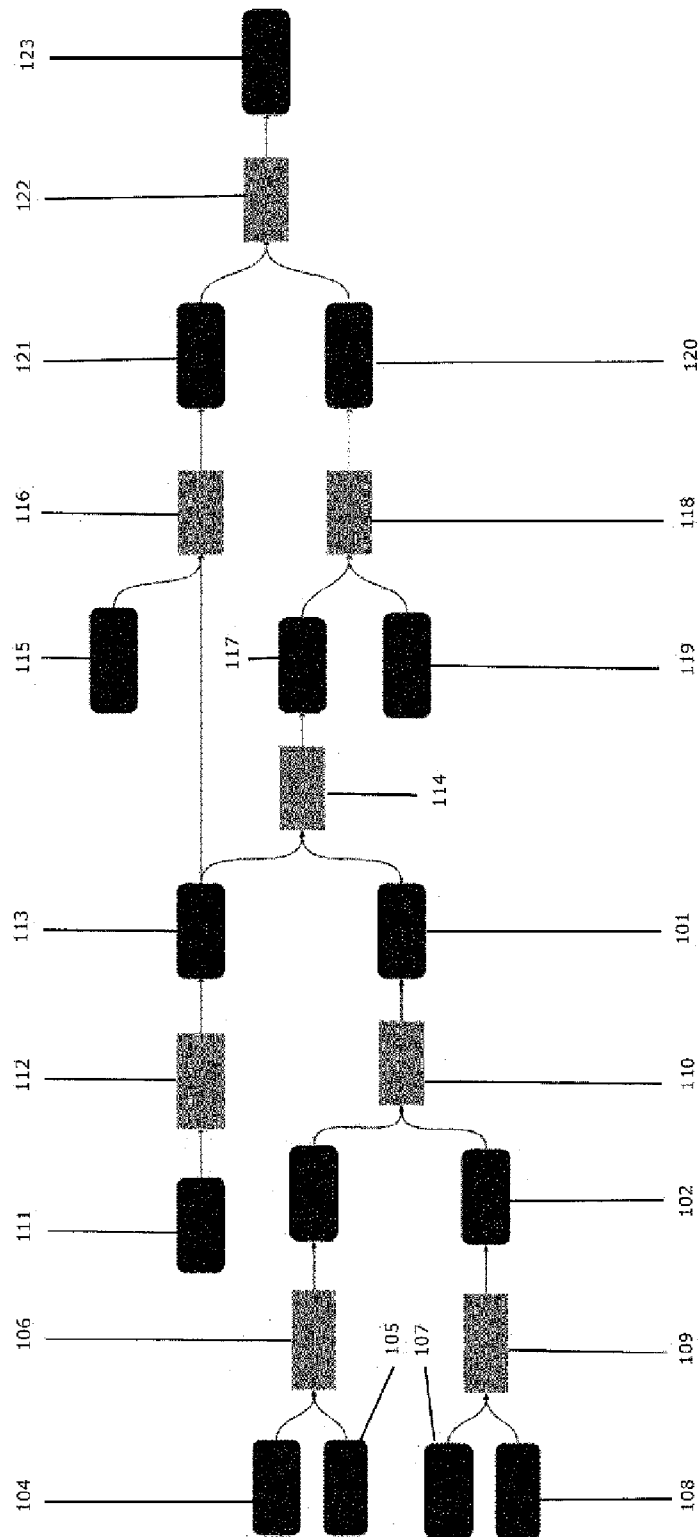


Fig. 1

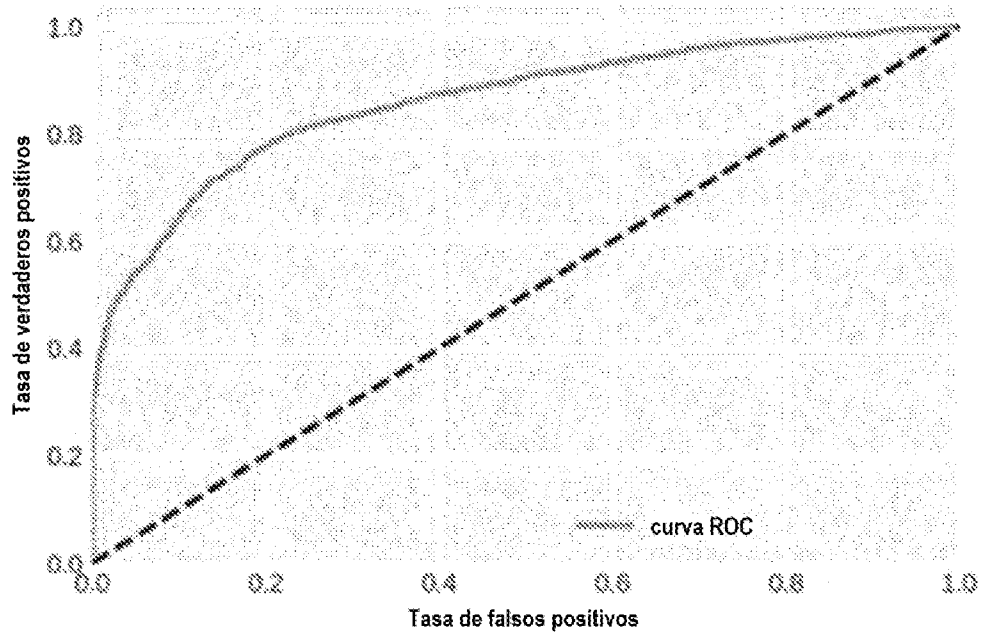


Fig. 2

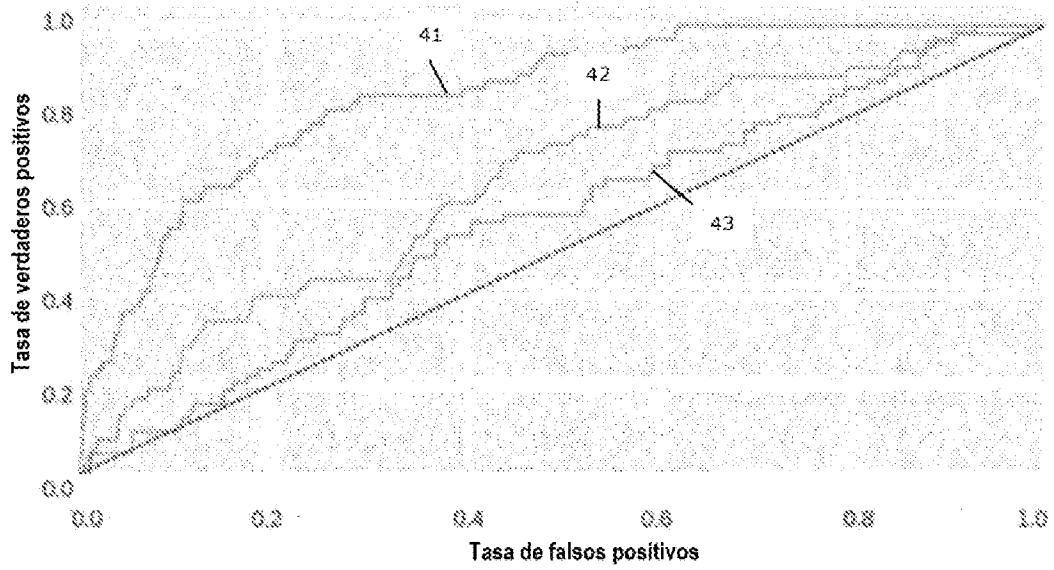


Fig. 6

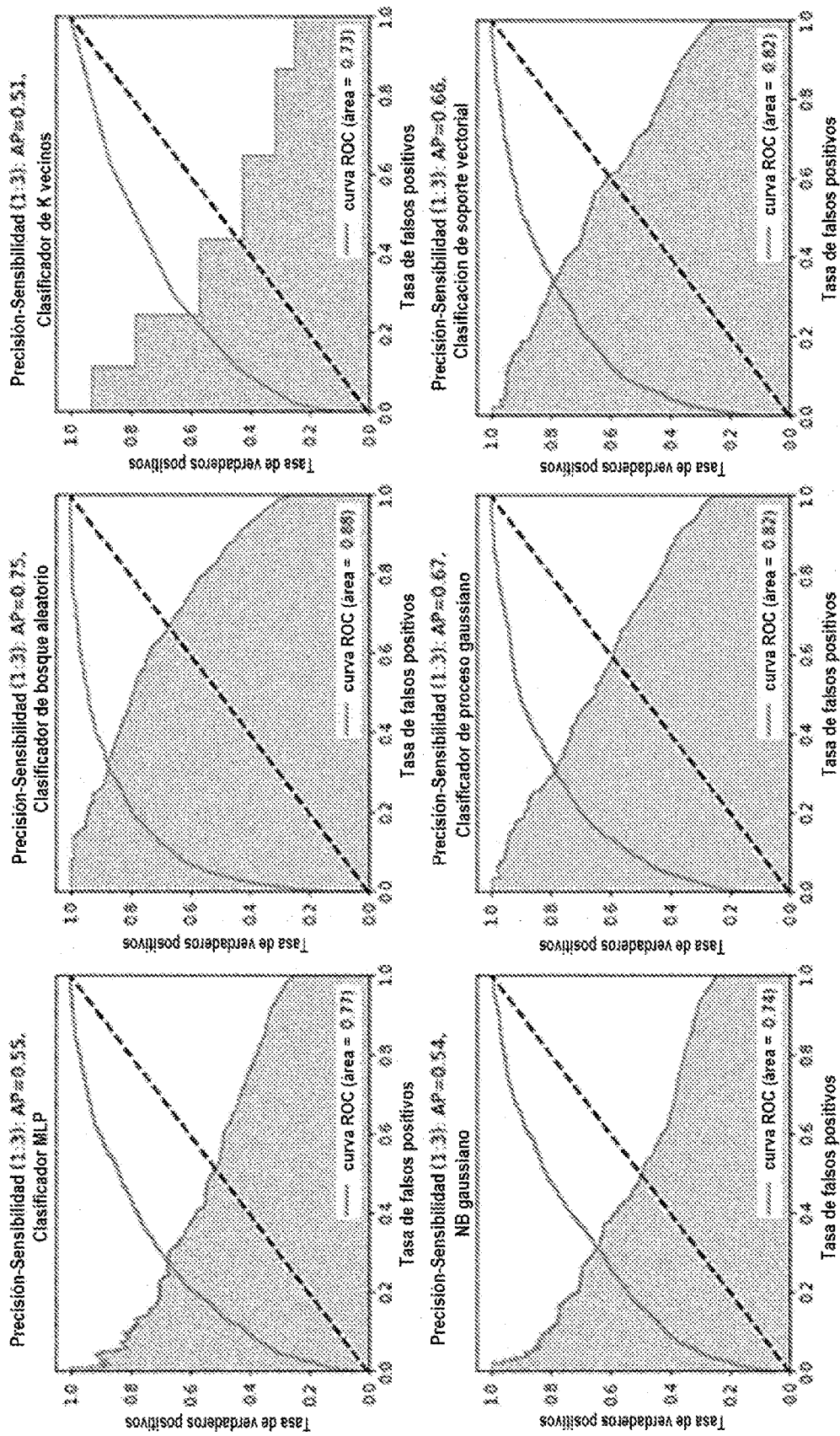


Fig. 3

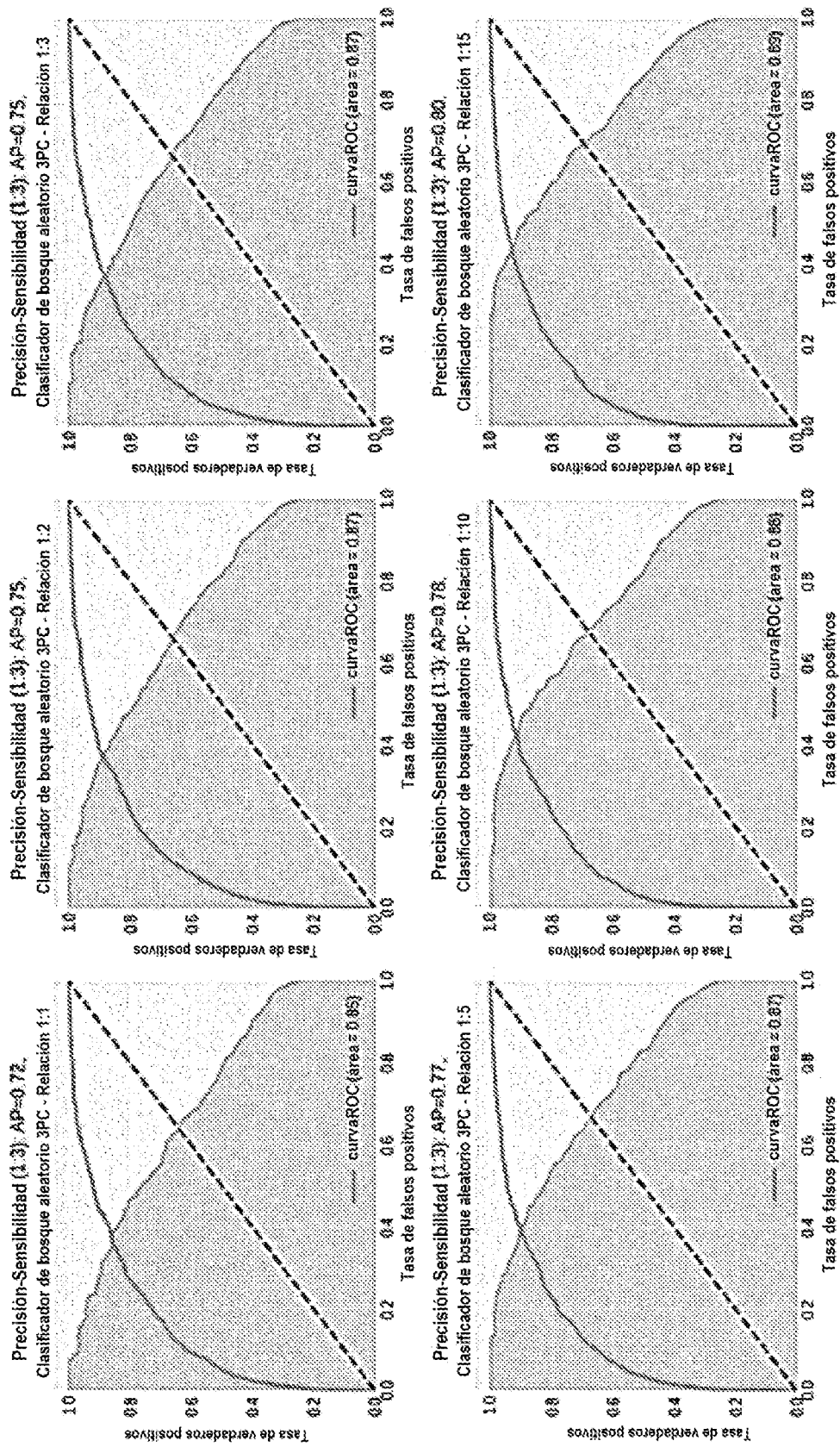


Fig. 4

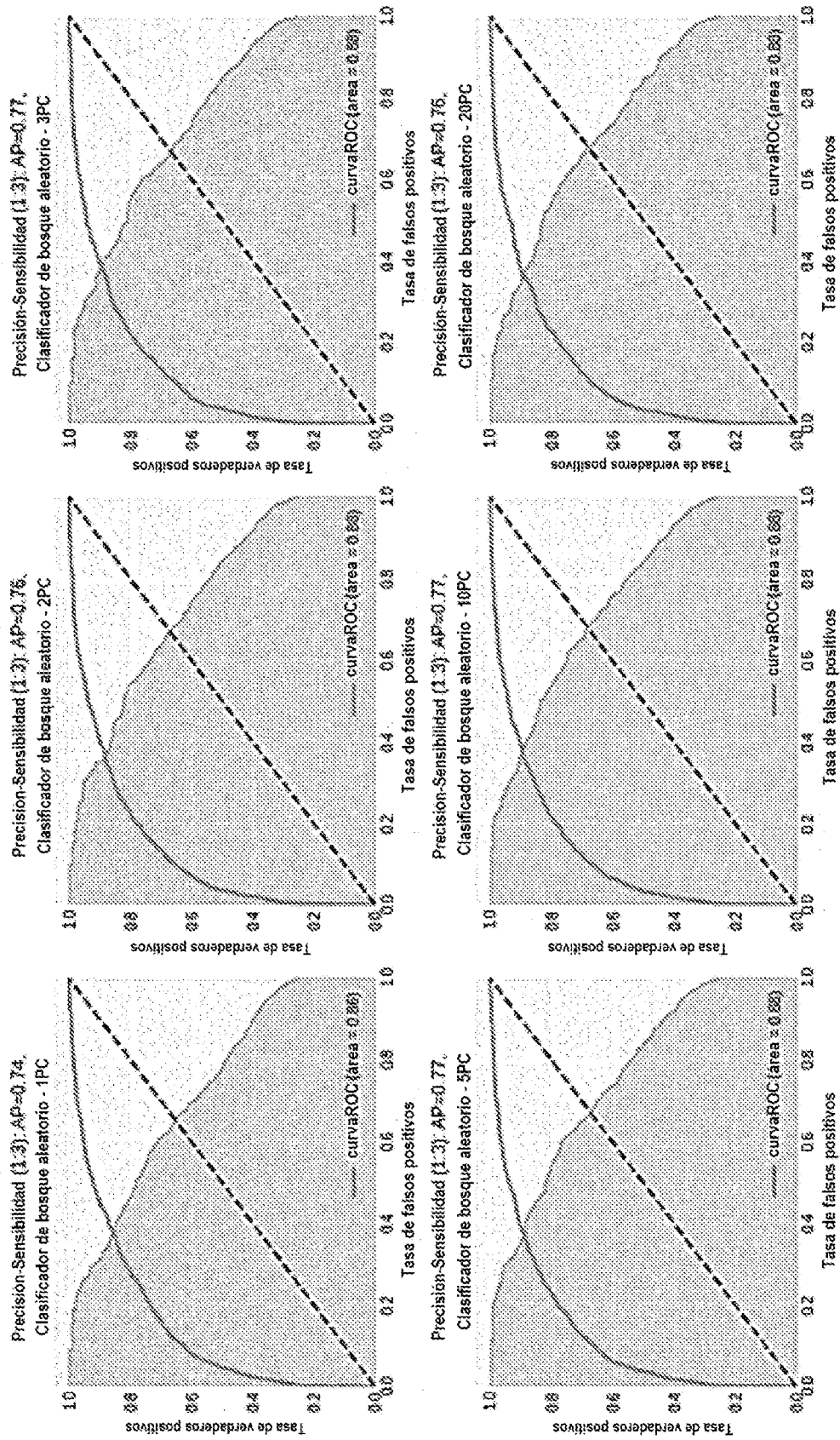


Fig. 5

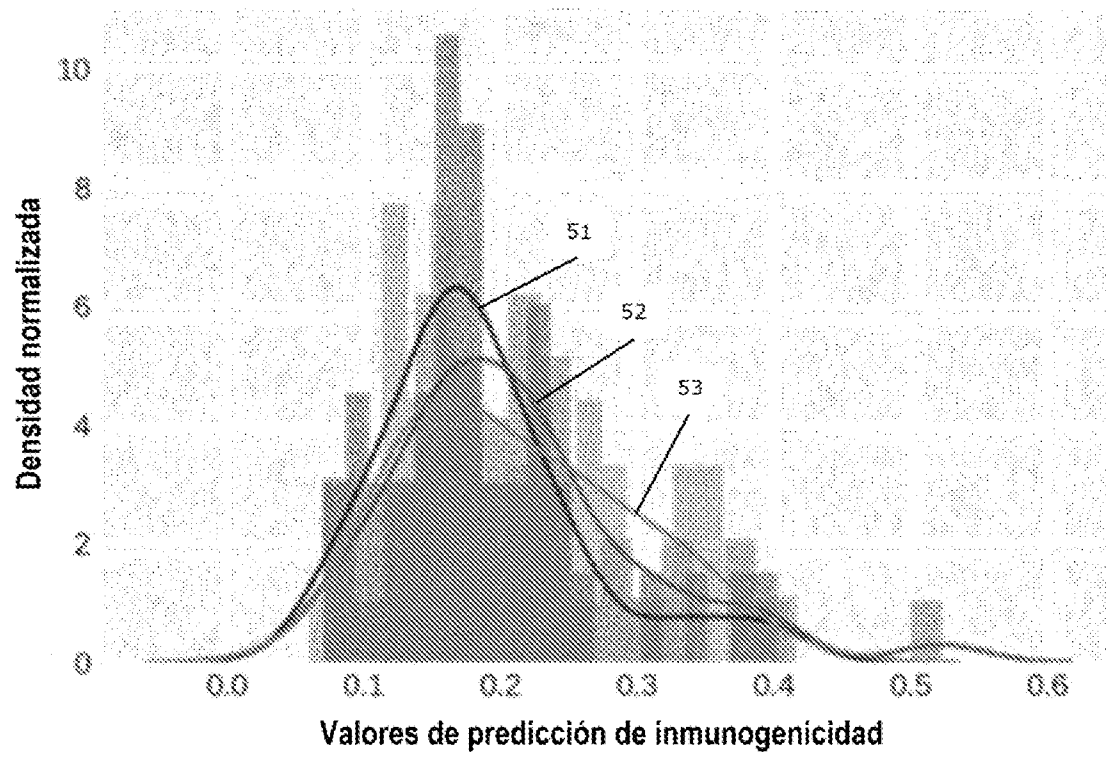


Fig. 7

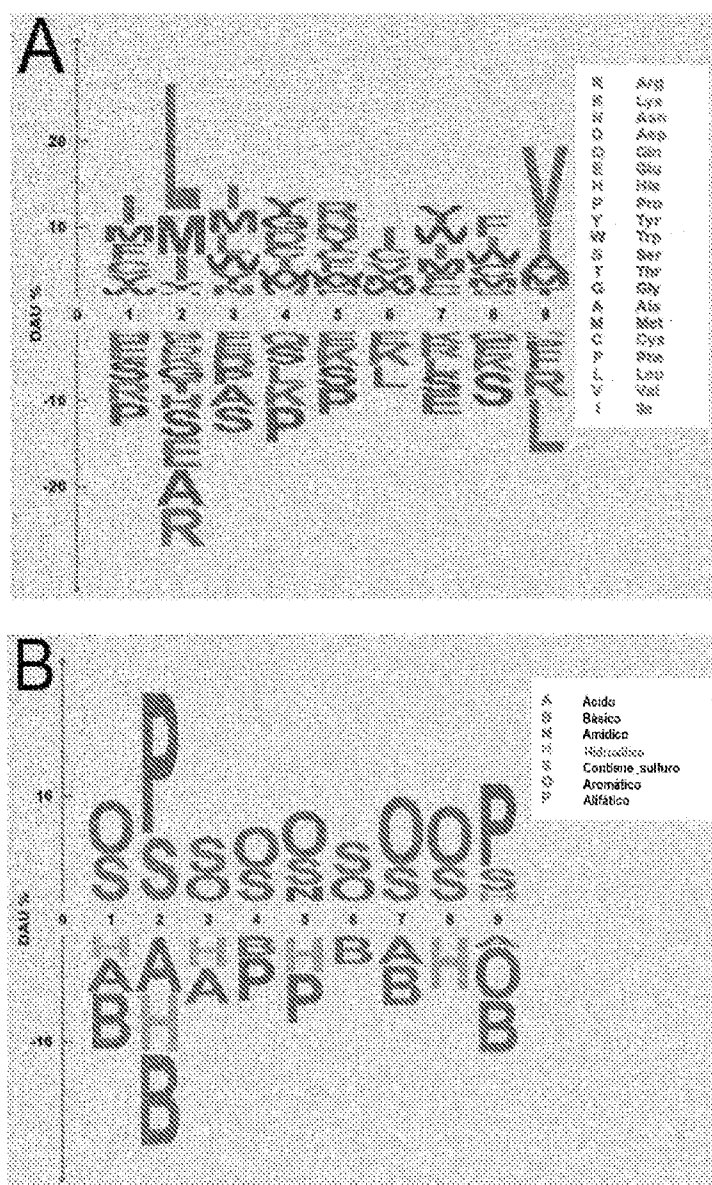


Fig. 8

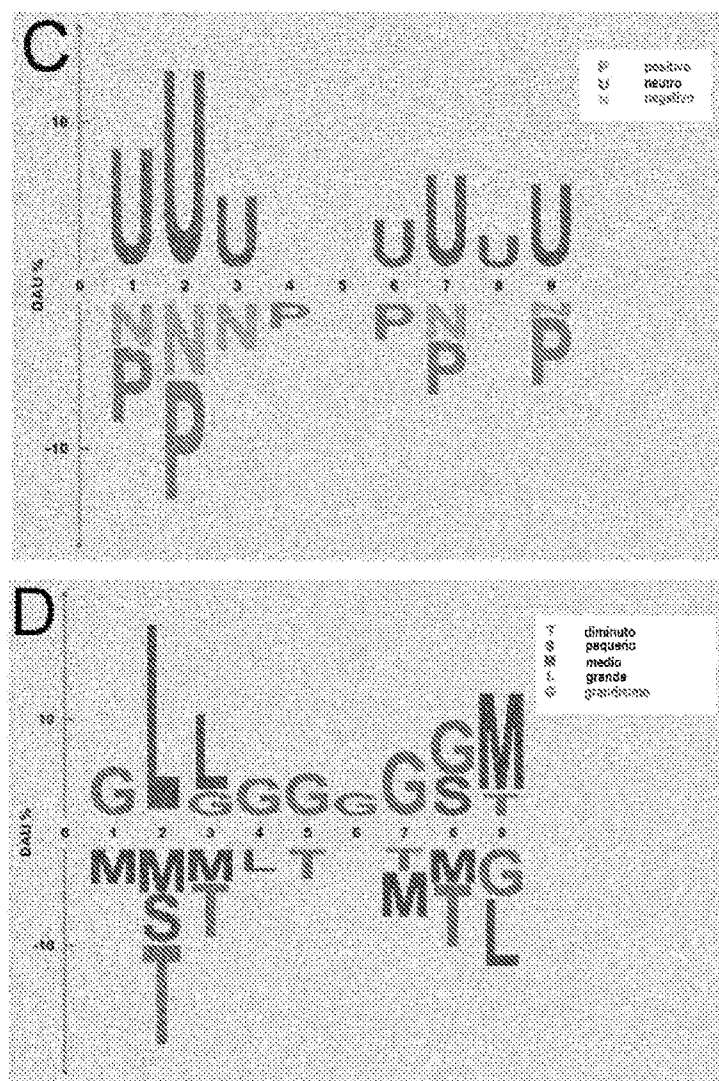


Fig. 8 (continuación)

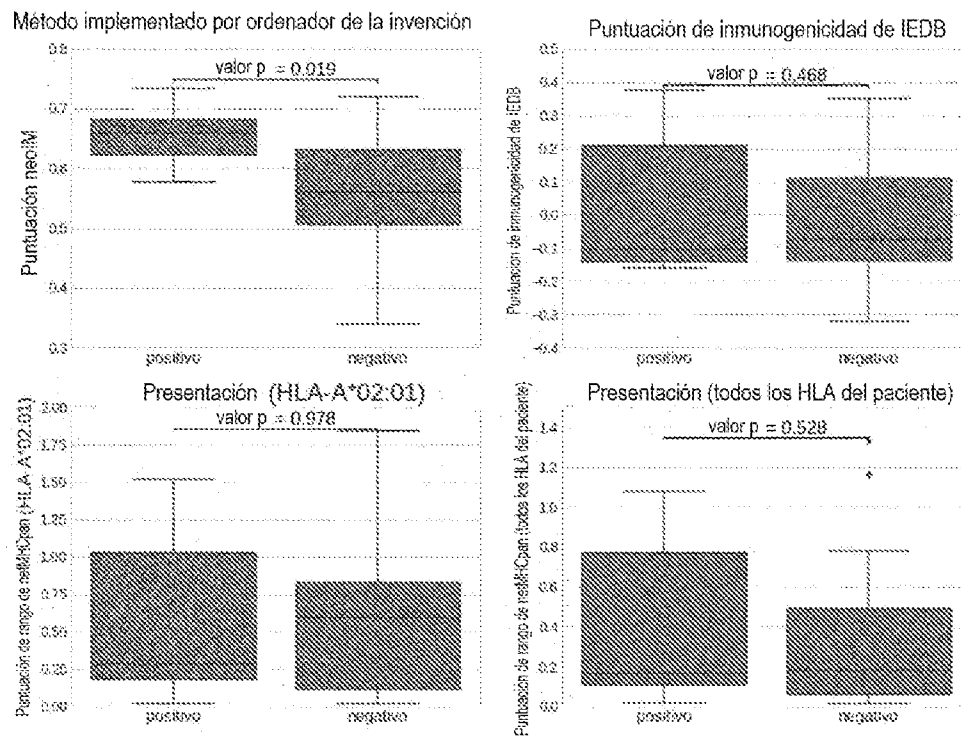


Fig. 9A

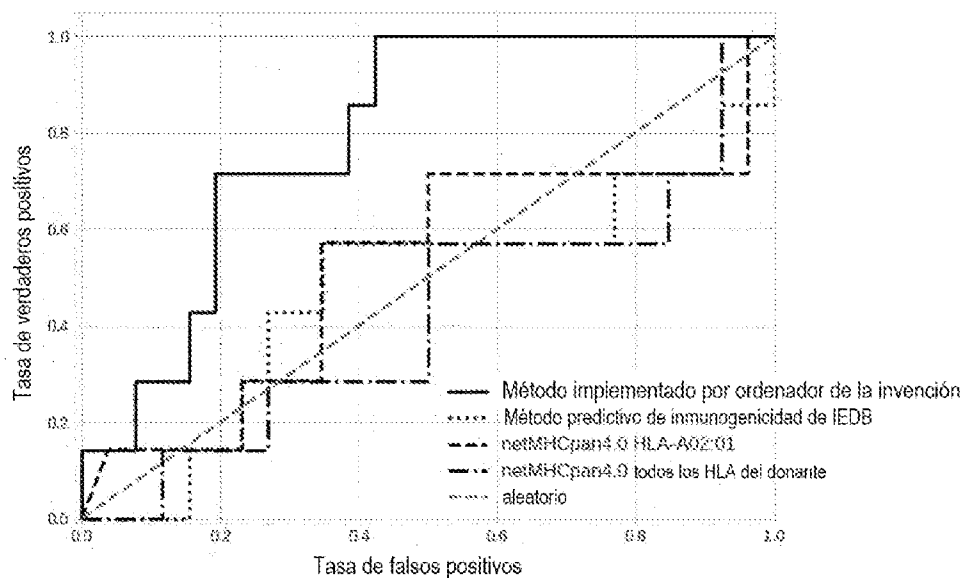


Fig. 9B

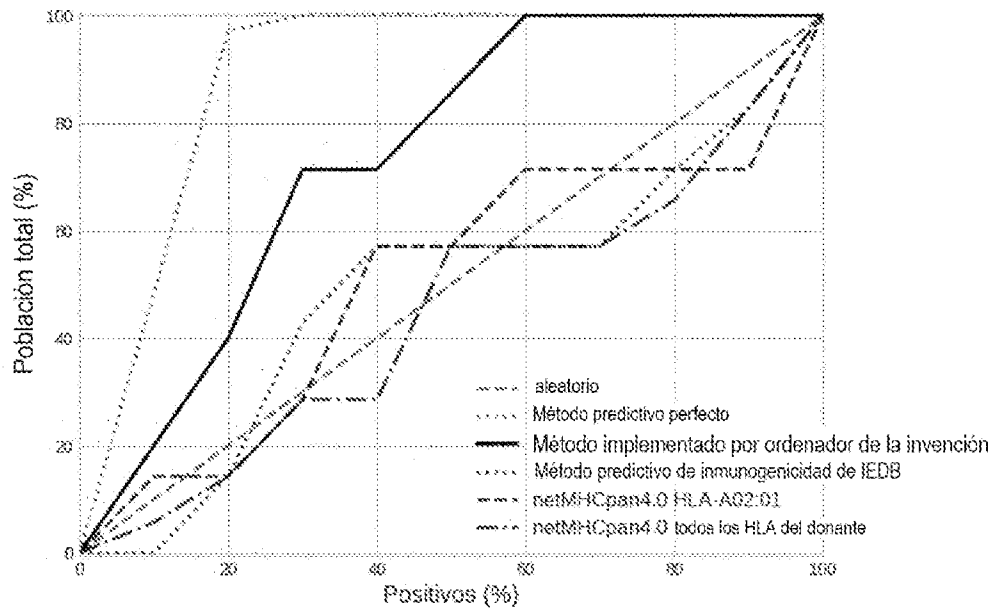


Fig. 9C

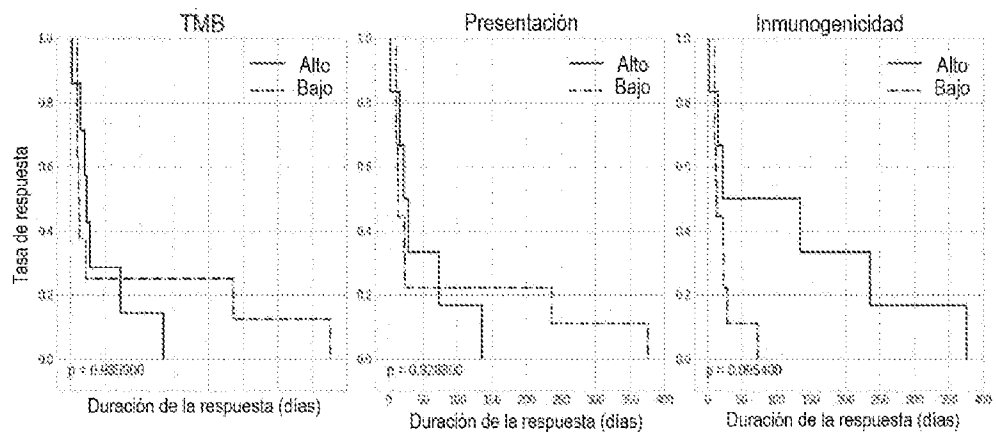


Fig. 10