(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 717 280**

(51) Int. CI.:

C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **18.06.2013 PCT/IB2013/055004**

(87) Fecha y número de publicación internacional: **27.12.2013 WO13190468**

(96) Fecha de presentación y número de la solicitud europea: **18.06.2013 E 13759573 (2)**

(97) Fecha y número de publicación de la concesión europea: **26.12.2018 EP 2861763**

(54) Título: **Factor predictivo basado en ordenador para cáncer de próstata**

(30) Prioridad:

19.06.2012 IT MI20121066

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
20.06.2019

(73) Titular/es:

EUROCLONE S.P.A. (100.0%)
Via Figino, 20/22
20016 Pero (MI), IT

(72) Inventor/es:

VENDRAMIN, ANNA;
SACCANI, ANDREA;
SONEGO, PAOLO y
CAPPUCCILLI, GUIDO

(74) Agente/Representante:

LINAGE GONZÁLEZ, Rafael

DESCRIPCIÓN

Factor predictivo basado en ordenador para cáncer de próstata

El objeto de la presente invención es un método para diagnosticar un cáncer de próstata, que comprende el análisis de datos relacionados con la expresión de determinados genes mediante algoritmos predictivos.

Hoy en día, el cáncer de próstata representa una de las enfermedades de cáncer más comunes en países occidentales, y se espera que la incidencia del mismo aumente en el futuro.

De la misma manera que para otros cánceres, ahora se cree en general que un diagnóstico temprano, es decir, en las primeras fases de desarrollo de la enfermedad, es crucial para garantizar un desenlace positivo del tratamiento. Hoy en día, los métodos de diagnóstico tradicionales comprenden el examen morfológico de muestras de tejido prostático. Sin embargo, tales métodos tienen limitaciones inherentes que no son desdeñables, por ejemplo, con respecto a la experiencia del operario que organiza y analiza la muestra. Por otro lado, las técnicas de diagnóstico molecular estudian la presencia y expresión de un patrón génico específico, es decir, el patrón de genes específicos implicados en una enfermedad. Estas técnicas evitan el problema de interpretar la preparación histológica; por tanto, desde este punto de vista, estos métodos son más fiables, aunque deben normalizarse e implementarse de modo que se completen en tiempos reducidos y sean accesibles a un coste mucho más reducido. Esto podría ser posible, por ejemplo, limitando el número de los denominados genes marcadores, es decir, genes cuyos niveles de expresión se ha observado que cambian de manera que pueden relacionarse directamente con el desarrollo de una enfermedad dada. Sin embargo, al hacerlo, podría obtenerse un resultado no muy fiable, sólo porque se basa en una conjunto reducido de evaluaciones. En realidad, se conocen métodos, que se basan en diferentes algoritmos posibles, para identificar un conjunto de genes significativos (los genes “marcadores” mencionados anteriormente) a partir de un análisis molecular sobre un gran número de genes. Sin embargo, tal como se indicó anteriormente, una vez que se ha identificado tal conjunto de genes significativos, el problema para garantizar que un diagnóstico, llevado a cabo sobre dicho conjunto, conduce a resultados suficientemente fiables, tan cercanos como sea posible a los resultados que pueden alcanzarse a través de un análisis a mayor escala, permanece sin resolver.

Por otro lado, será evidente que llevar a cabo un ensayo molecular sobre un gran número de genes implica necesariamente un coste mayor.

Por tanto, existe la necesidad de equilibrar dos requisitos diferentes: por un lado, tener en cuenta un número suficiente de genes marcadores, tal como para obtener un resultado altamente fiable, y, por otro lado, proporcionar una herramienta de diagnóstico a costes competitivos.

Además, una vez que se ha establecido un conjunto de genes significativos conocidos, surge la necesidad adicional, en aras de eficiencia de costes y simplicidad de procedimientos, pasar del “conjunto significativo de genes marcadores” mencionado anteriormente a un “subconjunto óptimo”, que comprende un número de genes reducido adicionalmente, siempre que sean capaces de proporcionar resultados suficientemente fiables. Una necesidad de este tipo no se cumple mayoritariamente cuando se usan las soluciones conocidas.

Los documentos de la técnica anterior de Federica Rizzi *et al.* (“A novel gene signature for molecular diagnosis of human prostate cancer by RT-PCR”, PLOS ONE, vol. 3, n.º 10, 1 de enero de 2008 página e3617) y Bettuzzi S *et al.* (“Molecular Diagnosis of human prostate cancer (CAP) by RRT-QPCR determination of gene expression signature”, European Urology Supplements, XX,XX, vol 5, n.º 14, 1 de septiembre de 2006) dan a conocer una firma de ocho genes marcadores más dos genes de mantenimiento y su uso en el diagnóstico o en la determinación del pronóstico de cáncer de próstata, pero no dan a conocer el uso del algoritmo de máquina de soporte vectorial (SVM) para el análisis de la expresión de dichos genes.

El documento de la técnica anterior US 2009/215058 da a conocer el uso del algoritmo de máquina de soporte vectorial (SVM) en el contexto de una selección de un subconjunto de genes de factores predictivos.

Por tanto, la presente invención aspira a proporcionar un método diagnóstico que, en las diversas realizaciones del mismo, satisface las necesidades anteriormente mencionadas.

Objeto de la invención

Por tanto, en un primer aspecto, la invención describe un método para el diagnóstico del tumor de próstata (o cáncer) según la reivindicación 1 y las reivindicaciones dependientes 2 a 4.

Descripción de los dibujos

Las figuras 1 y 2 representan resultados de predicción (PREDICHO), en comparación con los resultados observados realmente (OBSERVADO), cuantificados en cuanto a tasa de error (o “tasa de error global” OE), tasa de falsos positivos (FPR), tasa de falsos negativos (FNR), obtenidos aplicando varios ejemplos del método según la invención,

y particularmente un procesamiento basado en tres algoritmos predictivos diferentes (k-NN, bosque aleatorio, SVM).

Descripción detallada de la invención

Según un primer objeto de la invención, un método para el diagnóstico de cánceres de próstata, que comprende las etapas de:

a. determinar una pluralidad de valores de expresión de cada uno de un grupo de genes de factores predictivos, en una muestra de tejido prostático aislada,

b. proporcionar a un algoritmo predictivo entrenado, que funciona en espacios vectoriales, un vector de entrada que contiene en elementos vectoriales respectivos cada valor de expresión de dicha pluralidad de valores de expresión en dicha muestra de tejido prostático que va a analizarse;

c. procesar el vector de entrada, mediante el algoritmo predictivo entrenado;

d. generar un resultado representativo de dicho diagnóstico para dicha muestra de tejido prostático;

en el que:

el algoritmo predictivo entrenado es una máquina de soporte vectorial, y el grupo de genes de factores predictivos se compone de (i) genes que codifican para HMBS, CLU, GAS I y ODC;

comprendiendo además dicho método, antes de la etapa a., la etapa de:

- entrenar el algoritmo predictivo, usando, como entrada, valores de expresión conocidos en muestras de tejido prostático que corresponden al diagnóstico de cáncer de próstata.

Según una terminología que los expertos en el campo de algoritmos conocen, los términos “algoritmo predictivo entrenado” (es decir, “algoritmo de factor predictivo entrenado” es decir, “algoritmo de predicción de aprendizaje automático”) se usan para indicar, por ejemplo, un algoritmo que pertenece a la clase de algoritmos que pueden estimar un desenlace (es decir, habitualmente, que estiman un resultado dentro de un conjunto discreto de resultados posibles), basándose en datos de entrada conocidos, en aquellos casos en que tal desenlace no puede determinarse de manera analítica basándose en los datos de entrada. La clase de algoritmos predictivos entrenados que pueden aprovecharse en la presente invención (tal como se muestra con más detalle a continuación en el presente documento) comprende, por ejemplo, algoritmos paramétricos, cuyos resultados dependen de parámetros ajustables. En tales algoritmos, el ajuste de los parámetros se lleva a cabo basándose en datos de entrada, cuyo resultado ya se conoce, y se lleva a cabo comparando los resultados generados por el algoritmo, a medida que los parámetros ajustables cambian, en un conjunto de datos de entrada dado, con los respectivos resultados que ya se conocen. La etapa de ajustar los parámetros, es decir, el “entrenamiento”, se realiza habitualmente como etapa de entrenamiento independiente, antes del uso del algoritmo con una función predictiva, pero también puede realizarse de manera continua durante la propia operación del algoritmo.

En una aplicación para estimaciones de diagnóstico, los resultados del algoritmo pueden ser de un tipo binario, es decir, “verdadero” o “falso” (o, en otros términos, “sano” o “enfermo”), o pueden articularse según una escala discreta, indicativa, por ejemplo, de diferentes niveles de gravedad de la enfermedad.

Según un aspecto preferido, los genes de factores predictivos (marcadores) usados para los fines de la presente invención pertenecen a un conjunto que comprende:

- gen que codifica para ornitina descarboxilasa (ODC);
- gen específico de detención del crecimiento (GAS 1, también denominado GAS I);
- gen que codifica para clusterina (CLU);
- gen que codifica para hidroximetilbilano sintasa (HMBS);

Según un aspecto de la invención, la expresión mencionada anteriormente de un gen de factor predictivo es un valor representativo de la presencia, particularmente del aspecto funcional, del gen de factor predictivo correspondiente en la muestra de tejido prostático.

Según un aspecto de la invención, la muestra de tejido prostático se aísla previamente por biopsia de un paciente y a continuación se trata según las técnicas conocidas en la técnica. En un aspecto preferido de la invención, después de la extracción, la muestra se incrusta y conserva en parafina; por tanto, es una muestra “de repertorio” o “de

archivo". Por supuesto, antes del uso para los fines de la presente invención, la muestra incrustada en parafina se trata adecuadamente, por ejemplo "se desparafina", según técnicas conocidas en la técnica.

Con respecto al análisis de expresión génica, este se implementa mediante PCR en tiempo real, según métodos conocidos por los expertos en la técnica.

Particularmente, se usan preferiblemente los siguientes cebadores:

Gas1 fwd 5'- CGCACCGTCATTGAGGAC

Gas1 rev 5'- CACGCAGTCGTTGAGCAG

Clu fwd 5'- CCTCACTTCTTCTTTCCCAAG

Clu rev 5'- GTACGGAGAGAAGGGCATC

Odc fwd 5'- CAGTCTGTCGTCTCAGTGTG

Odc rev 5'- TTCGCCCGTTCCAAAAGGAG

Hmbs fwd 5'- CGCTGCATCGCTGAAAGG

Hmbs rev 5'- ACGGCTACTGGCACACTG

Las secuencias descritas anteriormente representan en sí mismas un objeto adicional de la presente invención.

Según un aspecto de la divulgación el algoritmo predictivo entrenado es un algoritmo que opera en espacios vectoriales, y la etapa b. mencionada anteriormente (procesamiento) comprende: procesar un vector de entrada que contiene, en elementos vectoriales respectivos, cada expresión de la pluralidad de expresiones; y generar, dependiendo de tal procesamiento, un resultado representativo del diagnóstico y/o pronóstico.

Según la invención, el algoritmo predictivo es: máquina de soporte vectorial (SVM).

Según un aspecto adicional de la invención, el método comprende además, antes de la etapa a., la etapa de entrenar el algoritmo predictivo, usando como información de entrada los valores de expresión conocidos en las muestras de tejido prostático, de las que se conoce el diagnóstico y/o pronóstico, para calibrar parámetros algorítmicos ajustables y para obtener parámetros algorítmicos entrenados.

Según un aspecto adicional de la divulgación, el método comprende además la etapa de determinar el grupo de genes de factores predictivos, del conjunto de genes de factores predictivos mencionado anteriormente, dependiendo de los valores de expresión conocidos de tales genes en las muestras de tejido prostático, de las que se conoce el diagnóstico y/o pronóstico.

Según un aspecto adicional de la divulgación, la etapa mencionada anteriormente de determinar el grupo de genes de factores predictivos se implementa mediante un algoritmo para la reducción de la dimensión de espacios vectoriales.

Debe indicarse que el "conjunto de genes de factores predictivos" mencionado anteriormente puede referirse a un conjunto de genes marcadores significativos, particularmente, a un conjunto de genes marcadores que se conoce en el campo médico/de diagnóstico considerado en el presente documento. Preferiblemente, tal conjunto es el conjunto ya mencionado anteriormente.

Por otro lado, el "grupo de genes de factores predictivos" mencionado anteriormente, que forma la base para el diagnóstico a través de un algoritmo predictivo entrenado, puede definirse también como un "subconjunto optimizado significativo de genes de factores predictivos", es decir, "subconjunto reducido de marcadores" en los que puede implementarse la predicción, donde se lleva a cabo la transición del conjunto de genes marcadores al subconjunto optimizado (es decir, el "grupo") de genes marcadores, por ejemplo, por medio del algoritmo mencionado anteriormente para la reducción de la dimensión de espacios vectoriales.

En casos peculiares, también comprendidos en la invención, "grupo" y "conjunto" de genes marcadores pueden coincidir.

Según un aspecto adicional de la divulgación, la etapa mencionada anteriormente de determinar el grupo de genes de factores predictivos está comprendida en la etapa de entrenamiento inicial.

Según un ejemplo de implementación preferido del método, el algoritmo predictivo entrenado es una máquina de

soporte vectorial, y el grupo de genes de factores predictivos se compone de genes que codifican para HMBS, CLU, GAS I, ODC.

Debe indicarse que los dos ejemplos expuestos anteriormente, con referencia al grupo de genes de factores predictivos, no pretender ser limitativo. De hecho, tal como ya se notificó, en una realización del método de la presente divulgación el grupo de genes de factores predictivos se determina mediante un algoritmo para la reducción de la dimensión de espacios vectoriales, y por tanto tal grupo también puede ser diferente de los mencionados.

Por tanto, en el método de la presente divulgación puede obtenerse un resultado de diagnóstico (por ejemplo, “enfermo” - “sano”) en una muestra de tejido, a través de un algoritmo predictivo entrenado, usando como conjunto de genes marcadores todo el conjunto de partida, o bien de los grupos (es decir, subconjuntos) de genes marcadores expuestos anteriormente, con referencia a los dos ejemplos preferidos mencionados anteriormente, o bien otro grupo de genes marcador, siempre que se identifique del conjunto inicial mediante un algoritmo adecuado para la reducción de la dimensión de espacios vectoriales.

También se da a conocer un kit para el diagnóstico y/o para el pronóstico de cánceres de próstata, configurado para implementar el método descrito anteriormente. Particularmente, el kit comprende: medios de determinación para determinar una pluralidad de información representativa, siendo representativo cada elemento de información de una expresión de cada uno de un grupo de genes de factores predictivos, en una muestra de tejido prostático; y medios de procesamiento que pueden conectarse operativamente a los medios de determinación para recibir la pluralidad de información representativa de la pluralidad de expresiones, estando configurados los medios de procesamiento para procesar la información representativa de la pluralidad de expresiones por medio de un algoritmo predictivo entrenado, para obtener un resultado representativo de dicho diagnóstico y/o pronóstico.

Los genes de factores predictivos dados a conocer pertenecen al conjunto ya mencionado que comprende:

- gen que codifica para ornitina descarboxilasa (ODC);
- gen que codifica para antizima ornitina descarboxilasa (OAZ);
- gen que codifica para adenosilmetionina descarboxilasa (AdoMetDC);
- gen que codifica para espermidina/espermina N(1)-acetiltransferasa (SSAT);
- gen que codifica para histona H3 (H3);
- gen específico de detención del crecimiento (GAS 1, también denominado GAS I);
- gen que codifica para clusterina (CLU);
- gen que codifica para hidroximetilbilano sintasa (HMBS);
- gen que codifica para gliceraldehído 3-fosfato deshidrogenasa (GAPDH); y
- gen que codifica para fosfoglicerato cinasa (PGK1).

Genes de factores predictivos (marcadores) dados a conocer son:

- gen que codifica para ornitina descarboxilasa 1 (ODC1);
- gen que codifica para antizima ornitina descarboxilasa 1 (OAZ1);
- gen que codifica para adenosilmetionina descarboxilasa 1 (AMD1);
- gen que codifica para espermidina/espermina N(1)-acetiltransferasa 1 (SAT1);
- gen que codifica para agrupación de histonas 1, H3c (HIST1H3C);
- gen específico de detención del crecimiento 1 (GAS 1);
- gen que codifica para clusterina (CLU);
- gen que codifica para hidroximetilbilano sintasa (HMBS);

- gen que codifica para gliceraldehído 3-fosfato deshidrogenasa (GAPDH); y
- gen que codifica para fosfoglicerato cinasa 1 (PGK1).

Según la presente divulgación, los medios determinantes mencionados anteriormente son kits de diagnóstico, que tienen una estructura que se conoce *per se*, configurados para operar sobre el grupo seleccionado de genes de factores predictivos.

Según la presente divulgación, los medios de determinación comprenden un interfaz de comunicación de datos a los medios de procesamiento.

Según la presente divulgación, la provisión de los datos que proceden de los medios de determinación a los medios de procesamiento se lleva a cabo de otras maneras conocidas, por ejemplo, introduciéndolos manualmente.

Según la presente divulgación, los medios de procesamiento mencionados anteriormente son un procesador, tal como un ordenador, o un ordenador personal, o un ordenador portátil, o una estación de trabajo. Tal procesador está dotado de interfaz y medios de presentación visual de resultados, conocidos *per se*; con una memoria *per se*, en la que se almacenan uno o más algoritmos predictivos entrenados y programas de ejecución respectivos; y con un procesador, conocido *per se*, para ejecutar tales programas y algoritmos.

Según la presente divulgación, el kit según la invención está configurado para implementar una realización correspondiente del método según la invención, entre las descritas anteriormente.

A continuación en el presente documento, se proporcionarán detalles adicionales sobre el método descrito anteriormente, según diferentes ejemplos de implementación de la invención, y particularmente sobre los algoritmos predictivos usados.

Particularmente, se proporciona el uso de algoritmos que están configurados para operar en espacios vectoriales. Por ejemplo, se sometieron a prueba diferentes algoritmos de este tipo (conocidos *per se* en el campo de las matemáticas), y demostraron ser eficientes:

- Máquina de soporte vectorial (SVM);
- Bosque aleatorio;
- k vecino más cercano (k-NN);

para usarse alternativamente o en combinación entre sí.

Tal como conocen los expertos en la técnica de tales algoritmos, tales algoritmos definen una “clasificación” de muestras de entrada con respecto a posibles resultados de salida. El algoritmo de k-NN permite una clasificación automática de las muestras asignando una muestra a una clase basándose en la mayoría de las indicaciones (“votos”) de sus k vecinos, donde k es un número entero positivo (normalmente no muy grande).

El algoritmo de bosque aleatorio es en principio una extensión de métodos basados en árboles de clasificación en los que se genera un “bosque” de árboles clasificadores, cada uno de los que propone una clasificación para un único elemento en una clase dada. Comparando las propuestas de clasificación proporcionadas por cada árbol en el bosque, se identifica la clase a la que el elemento puede asignarse, es decir, la clase que recibió la mayoría de “votos”.

Las máquinas de soporte vectorial (SVM) pertenece al grupo de los separadores lineales (tales como, por ejemplo, las “redes neuronales artificiales”) y permiten resolver problemas de clasificación que no pueden resolverse mediante un separador lineal en un espacio con pocas dimensiones, mediante un mapeo, definido por una función “núcleo”, de cada dato en un espacio con una dimensionalidad mayor, en la que los grupos de clasificación pueden estar separados linealmente.

Cada uno de los algoritmos mencionados anteriormente opera en un espacio vectorial; por tanto, puede considerarse como una “caja negra”, que recibe en la entrada un vector y que produce, como función del vector de entrada, un resultado de entre un conjunto predefinido discreto de posibles resultados.

En la presente divulgación el vector de entrada es representativo de las expresiones de genes de factores predictivos. Particularmente, el vector de entrada comprende diferentes elementos vectoriales, estando asociado cada uno a un gen de factor predictivo correspondiente. Cada elemento de vector contiene un valor representativo de la presencia del gen correspondiente en la muestra de tejido prostático analizada, tal como se detecta en la misma muestra.

Según un ejemplo de implementación preferido, tal valor representativo puede tomar valores numéricos, dentro de un intervalo discreto o continuo, correspondiente a una medición de los niveles de expresión del gen de factor predictivo correspondiente en la muestra de tejido prostático analizada. Ventajosamente, esto permite caracterizar mejor el tejido prostático, en comparación con una mera indicación de la presencia o la ausencia del gen de factor predictivo, y por consiguiente mejorar la exactitud de predicción.

Sin embargo, según otros ejemplos de implementación comprendidos en la divulgación, el valor representativo también puede ser simplemente un valor binario (SÍ/NO, es decir, PRESENTE/AUSENTE), estando asociado al hecho de que el nivel de expresión detectado del gen correspondiente está por encima o por debajo de un umbral preestablecido.

La dimensión de los espacios vectoriales en los que los algoritmos predictivos operan corresponde a la dimensión de los vectores de entrada, por tanto, al número de genes de factores predictivos que pertenecen al grupo seleccionado. Según diferentes ejemplos de implementación de la invención, tal dimensión puede estar comprendida, por ejemplo, entre 4 y 10.

El algoritmo predictivo está configurado de modo que predice en primer lugar la presencia o la ausencia del cáncer. Además, en un ejemplo de implementación adicional, también puede estar configurado para proporcionar una predicción del grado de progresión/gravedad del cáncer. Según un ejemplo específico, en el caso de que se pronostique la presencia de un cáncer, también se clasifica en tres clases de Gleason (Gleason6, Gleason7, Gleason8) que corresponden a la misma cantidad de grados de pronóstico.

Por tanto, la información de salida del algoritmo (que proporciona, tal como ya se indicó, un resultado de entre un conjunto predefinido discreto de posibles resultados) puede ser simplemente un resultado binario, o puede ser no binario, mientras sea discreto. Por tanto, se observará que la salida del algoritmo predictivo entrenado, usado tal como se describió anteriormente, proporciona directamente un resultado representativo de un diagnóstico (por ejemplo, valores de salida “sano” o “enfermo”). Además, en un ejemplo de implementación particular, tal salida también puede comprender un resultado representativo del grado de gravedad/progresión de la enfermedad.

Ahora, teniendo en cuenta los modos de operación de los algoritmos, debe indicarse que cada uno de los algoritmos predictivos del tipo mencionado anteriormente se caracteriza por algunos “parámetros algorítmicos ajustables”, que afectan a los resultados. Particularmente, los “parámetros algorítmicos ajustables” son γ y *cost* (en SVM), *ntree* y *mtry* (en bosque aleatorio), *k* (en k-NN). Tales “parámetros algorítmicos ajustables” se calibran durante una etapa inicial, la denominada etapa de entrenamiento (entrenamiento), para obtener “parámetros algorítmicos entrenados”, que entonces se usan en la etapa de uso real del método con fines de diagnóstico/pronóstico.

La etapa de entrenamiento, que se proporciona independientemente para cada uno de los algoritmos, proporciona generalmente comenzar desde un valor nominal inicial de los “parámetros algorítmicos ajustables”, introducir los vectores de algoritmo que corresponden a situaciones para las que ya se conocen los resultados; luego, verificar el resultado obtenido; finalmente cambiar de manera iterativa los “parámetros algorítmicos ajustables”, hasta alcanzar una exactitud deseada en los resultados obtenidos. Particularmente, la etapa de entrenamiento proporciona el uso, como vectores de entrada, de vectores que contienen los resultados detectados en un conjunto de muestras de tejido prostático, del que se conoce el diagnóstico y/o pronóstico (en parte sano, en parte enfermo, y, cuando es enfermo, en diferentes fases de progresión de la enfermedad).

Debe indicarse que el algoritmo se denomina “entrenado” cuando los “parámetros algorítmicos ajustables” se establecieron en los valores que produjeron los resultados más fiables, durante la etapa de entrenamiento, obteniendo por tanto los “parámetros algorítmicos entrenados”.

Una vez que se ha entrenado el algoritmo, puede recibir en la entrada vectores específicos que contienen datos medidos en muestras de tejido prostático, referentes a los casos respectivos para los que se desea una predicción, y proporciona como salida el resultado de la predicción.

Según una realización del método, se realiza la etapa de entrenamiento usando la denominada técnica de “validación cruzada”. En un caso de este tipo, el conjunto de datos iniciales (del que se conoce el resultado) se divide en un subconjunto de entrenamiento (o “conjunto de entrenamiento”) y un subconjunto de prueba (o “conjunto de prueba”). El modelo para el factor predictivo dado, es decir, la definición de los parámetros ajustables, se construye usando cada uno de los “conjuntos de entrenamiento” y se evalúa en cada uno de los “conjuntos de prueba”. Por ejemplo, en la “validación cruzada 10 veces”, se dividen los datos en diez grupos que contienen aproximadamente el mismo número de pacientes, se entrena el algoritmo predictivo en nueve grupos y se valida en el décimo; entonces, se calcula la exactitud obtenida del modelo, y se repite el procedimiento hasta que se usa cada grupo como “conjunto de prueba” y los restantes como “conjunto de entrenamiento”. Al final del procedimiento mencionado anteriormente, se calcula el promedio de los errores a lo largo de la serie de repeticiones entera que se ha llevado a cabo.

A continuación en el presente documento se proporcionan datos relativos a pruebas experimentales que se han realizado a modo de ejemplo ilustrativo, no limitativo. Algunos de los resultados obtenidos se notifican en las figuras 1 y 2.

Se llevó a cabo una primera serie de pruebas basándose en un conjunto de 8 genes de factores predictivos (expuestos de manera convencional en el presente documento en letras minúsculas: gas I, h3, ssat, clu, odc, ado, oaz, hmbs). Tal primera serie de pruebas (cuyos resultados se ilustran en la primera fila de la figura 1) comprendió pruebas llevadas a cabo en un conjunto que comprendía 82 muestras, de las que 60 eran de enfermos y 22 eran de sanos; y se llevaron a cabo pruebas adicionales en un conjunto que comprendía 59 muestras de enfermos, en tres grados diferentes de gravedad (G6, G7, G8).

Se llevó a cabo una segunda serie de pruebas basándose en un conjunto de 10 genes de factores predictivos (indicados de manera convencional en el presente documento en letras minúsculas: gas I, h3, ssat, clu, odc, ado, oaz, hmbs, gapdh, pgK I). Tal segunda serie de pruebas (cuyos resultados se ilustran en la primera fila de la figura 2) comprendió pruebas llevadas a cabo en un conjunto que comprendía 90 muestras, de las que 64 eran de enfermos y 26 eran de sanos; y se llevaron a cabo pruebas adicionales en un conjunto que comprendía 63 muestras de enfermos, en tres grados diferentes de gravedad (G6, G7, G8).

Debe indicarse que los conjuntos de genes de factores predictivos considerados anteriormente se derivan de la práctica. También es oportuno indicar que, según un aspecto preferido de la presente invención, y a diferencia de la mayoría de los enfoques tradicionales, los genes en el conjunto se tratan y se consideran de manera equivalente, y no hay gen(es) que actúe(n) como “referencia”.

Tanto la primera como la segunda serie de pruebas se llevaron a cabo entrenando independientemente y luego usando cada de uno de los tres algoritmos predictivos mencionados anteriormente.

Se realizó el entrenamiento, (tal como ya se describió) a través de la calibración de los “parámetros algorítmicos ajustables” de los algoritmos, basándose en una comparación entre los resultados de predicción, tal como los proporcionó el algoritmo, y los resultados conocidos.

Particularmente, se recopilaron las denominadas “matrices de confusión”, notificadas en las figuras 1 y 2, cada una de las cuales contiene, en el elemento de filaN-columnaN, los casos negativos reconocidos por la predicción como negativos; en el elemento de filaP-columnaP, los casos positivos reconocidos por la predicción como positivos; en el elemento de filaN-columnaP, los casos negativos identificados por la predicción como positivos (FALSO POSITIVO); en el elemento de filaP-columnaN, los casos positivos identificados por la predicción como negativos (FALSO NEGATIVO).

Luego se han observado los resultados dividiéndolos en EXACTO, FALSO POSITIVO, FALSO NEGATIVO, y sintetizándolos a través de los parámetros de OE (tasa de error global), FPR (tasa de falsos positivos) y FNR (tasa de falsos negativos).

Los resultados cuantitativos expuestos en las figuras 1 y 2 son evidentes en muchos aspectos. Sin embargo, a continuación en el presente documento se notifican algunas observaciones.

Teniendo en cuenta las “matrices de confusión” de la primera fila de la figura 1, referente a la primera serie de pruebas, se observará que:

- a través del algoritmo de k-NN (con un parámetro entrenado k = 3), se obtiene una tasa de error global del 12,2%;
- a través del algoritmo de bosque aleatorio (con parámetros entrenados *ntree* = 500, *mtry* = 3), se obtiene una tasa de error global del 15,8%;
- a través del algoritmo de SVM (con parámetros entrenados de *base radial*, γ = 0,06, *cost* = 4), se obtiene una tasa de error global del 9,7%.

Tal como se indicará a continuación en el presente documento, sin embargo, es importante evaluar no sólo la tasa de error global, sino también la tasa de falsos negativos y falsos positivos, aspirando a minimizar particularmente los falsos negativos.

Considerando las “matrices de confusión” de la primera fila de la figura 2, referente a la segunda serie de pruebas, se observará que:

- a través del algoritmo de k-NN (con un parámetro entrenado k = 3), se obtiene una tasa de error global del 12,2%;

- a través del algoritmo de bosque aleatorio (con parámetros entrenados *ntree* = 850, *mtry* = 2), se obtiene una tasa de error global del 15,5%;

- a través del algoritmo de SVM (con parámetros entrenados *de base radial*, γ = 0,17, *cost* = 4), se obtiene una tasa de error global del 13,3%.

Tal como se indicará a continuación en el presente documento, sin embargo, es importante evaluar no sólo la tasa de error global, sino también la tasa de falsos negativos y falsos positivos, aspirando a minimizar particularmente los falsos negativos.

Ahora, se ilustra con más detalle un ejemplo de implementación adicional, referente al aspecto ya mencionado relacionado con la reducción del número de genes de factores predictivos que va a usarse, es decir, con la selección adecuada de un grupo de genes de factores predictivos, en un conjunto de genes de factores predictivos conocido.

La reducción del número de genes de factores predictivos (es decir, las dimensiones de la base de marcadores), idealmente mientras se mantiene constante la exactitud de predicción, es ventajosa lo primero de todo porque simplifica el método diagnóstico y los kits de diagnóstico respectivos; luego, porque permite reducir el número de muestras en las que se realizan los procedimientos.

Además, en algunos casos, tal reducción puede permitir incluso una mejora en la exactitud de los resultados de predicción, reduciendo el “ruido” en la fluctuación de los resultados generados por el uso de un número mayor de genes de factores predictivos.

Por supuesto, la reducción del número de genes de factores predictivos implica una selección dirigida de los genes que tienen que ser parte del subconjunto óptimo (es decir, “grupo”).

Con respecto a esto, la característica de la presente divulgación de usar algoritmos predictivos entrenados, que operan en espacios vectoriales, permite ventajosamente aplicar técnicas estadístico-matemáticas, conocidas *per se* en el campo matemático, para una “reducción de espacios vectoriales”, que operan, comenzando desde un espacio vectorial inicial, independientemente del “significado” del contenido de los vectores.

Por tanto, según un ejemplo de implementación de la divulgación, la etapa de entrenamiento comprende la etapa adicional de determinar el grupo de genes de factores predictivos, en un conjunto de genes de factores predictivos, por medio de un algoritmo que pertenece a la clase de los “algoritmos para la reducción de la dimensión de espacios vectoriales”.

Por ejemplo, los algoritmos para la reducción de la dimensión de espacios vectoriales pueden basarse en los denominados métodos de “reducción de características”, aplicados en combinación con el algoritmo predictivo entrenado. Tal algoritmo de reducción analiza los resultados del algoritmo predictivo, durante la etapa de entrenamiento, y proporciona como salida un elemento de información indicativo del “peso” que cada uno de los genes de factores predictivos del conjunto inicial tiene para los fines de la determinación del resultado.

Esto proporciona un tipo de clasificación de genes de factores predictivos, desde el más relevante hasta el menos relevante, en cuanto a influencia sobre el resultado de predicción. Comenzando desde tal clasificación, se puede proceder mediante ensayo y error, ejecutando el algoritmo predictivo con un número de genes de factores predictivos cada vez más reducido, prestando atención para seleccionar, para un grupo que contiene N genes de factores predictivos, los primeros N genes de la clasificación determinados por el algoritmo de reducción. La reducción progresiva continúa hasta obtener un resultado óptimo, o en cualquier caso un resultado que se considere aceptable, según criterios preestablecidos.

Una vez que se ha identificado el número óptimo de genes de factores predictivos, entonces se selecciona el grupo respectivo de genes de factores predictivos, el método avanza entrenando el algoritmo predictivo y calibrando, finamente, los parámetros algorítmicos ajustables del mismo.

Según una realización adicional de la divulgación, se lleva a cabo la selección del grupo de genes de factores predictivos usando un paradigma de “selección de características”, basándose en el uso del algoritmo de bosque aleatorio para seleccionar las variables más relevantes. En esta realización, puede usarse una técnica de “muestreo aleatorio” de datos para seleccionar el modelo con una menor tasa de error según la variación en el número de las variables.

En la segunda fila de la figura 1, y en las filas segunda y tercera de la figura 2, se exponen resultados de series de pruebas adicionales llevadas a cabo aplicando las técnicas de reducción mencionadas anteriormente.

Particularmente, en la segunda fila de la figura 1, con referencia a la primera serie de pruebas, debe indicarse que se obtienen buenos resultados, en cuanto a tasa de error, con grupos de 4 genes de factores predictivos (con

bosque aleatorio y SVM) o de 6 genes de factores predictivos (con k-NN), tal como se indica en la figura. Particularmente, se obtuvieron los siguientes resultados:

- k-NN (k = 3), 6 genes de factores predictivos (indicados en la figura como "*características*"), tasa de error global = 12,2%
- bosque aleatorio (*ntree* = 500, *mtry* = 1), 4 genes de factores predictivos (indicados en la figura como "*características"*), tasa de error global = 14,6%
- SVM (*base radial*, γ = 0,5, *cost* = 4), 4 genes de factores predictivos (indicados en la figura como "*características"*), tasa de error global = 9,7%,

cuyos resultados son comparables con, y algunas veces incluso mejores que, los resultados expuestos en la primera fila de la figura 1, relativos al uso de 8 genes de factores predictivos.

A continuación en el presente documento se exponen algunas observaciones sintéticas adicionales sobre los resultados obtenidos.

Los mejores rendimientos (en cuanto a una menor tasa de error global) se obtuvieron mediante el algoritmo predictivo SVM entrenado en 82 muestras; tal resultado se validó también en el caso de usar sólo 4 genes de factores predictivos, seleccionados adecuadamente tal como se indicó anteriormente y se ilustra en la figura 1.

Generalmente, las diferentes aplicaciones del método, con los diferentes algoritmos predictivos, mostraron una aparición de falsos positivos (FPR) mucho mayor que la aparición de falsos negativos (FNR). Tal como se señalará a continuación en el presente documento, esto es una ventaja en el campo de diagnóstico.

En las series de pruebas llevadas a cabo, se obtuvo la tasa de error más baja usando el algoritmo predictivo SVM. Por otro lado, la tasa de falsos negativos FNR más baja siempre se ha obtenido usando el algoritmo de bosque aleatorio predictivo, que, sin embargo, produjo también la peor tasa de error global.

La disponibilidad de diferentes algoritmos predictivos, con diferentes ventajas e inconvenientes, permite articular de manera óptima la aplicación del método de la presente invención, en una de las diferentes realizaciones posibles, según el criterio usado para evaluar los resultados.

Tal como se expuso anteriormente, el método de la presente invención permite satisfacer la necesidad de proporcionar una herramienta de diagnóstico que sea fiable y está disponible a un coste accesible a los laboratorios diagnósticos de organismos públicos y privados, así como que sea competitiva con otras herramientas existentes.

Particularmente, se ha observado que el método puede distinguir con exactitud casos negativos de casos positivos.

Además, el método descrito anteriormente muestra una aparición de falsos positivos (es decir, proporciona un resultado indicativo de la enfermedad, cuando realmente el sujeto está sano) mucho mayor que los falsos negativos (es decir, sujetos diagnosticados como sanos, pero que, en evaluaciones adicionales, resultan estar enfermos).

Este hecho, que es una ventaja en el campo de diagnóstico, no se detecta mediante la observación de la tasa de error global (OE), pero resulta evidente a partir del análisis de las matrices de confusión.

El factor predictivo descrito en el presente documento permite clasificar la enfermedad con una exactitud aceptable en 3 clases de gravedad posibles, es decir, identificar los grados de pronóstico de clase de Gleason 6, Gleason 7 y Gleason 8, en los que la clase de Gleason es la clasificación usada para describir un cáncer de próstata teniendo en cuenta las características citológicas de las células y su organización.

Además, el método de la presente invención también demostró ser una herramienta de pronóstico válida, para la evaluación predictiva del posible desarrollo de la enfermedad con el tiempo.

Una ventaja particular relacionada con el método descrito es la posibilidad de usar, para los ensayos moleculares, muestras de archivo (o "repertorio") almacenadas adecuadamente, por ejemplo en parafina, sin excluir sin embargo el uso de muestras "recientes".

Ventajosamente, esto permite seguir la evolución de la enfermedad con el tiempo, y permite llevar a cabo análisis estadísticos y epidemiológicos muy amplios.

REIVINDICACIONES

1. Método para el diagnóstico de cánceres de próstata, que comprende las etapas de:

a. determinar una pluralidad de valores de expresión de cada uno de un grupo de genes predictivos, en una muestra de tejido prostático aislada,

b. proporcionar a un algoritmo predictivo entrenado, que funciona en espacios vectoriales, un vector de entrada que contiene en elementos vectoriales respectivos cada valor de expresión de dicha pluralidad de valores de expresión en dicha muestra de tejido prostático que va a analizarse;

c. procesar el vector de entrada, mediante el algoritmo predictivo entrenado;

d. generar un resultado representativo de dicho diagnóstico para dicha muestra de tejido prostático;

en el que:

el algoritmo predictivo entrenado es una máquina de soporte vectorial, y el grupo de genes predictivos se compone de (i) genes que codifican para HMBS, CLU, GAS I y ODC;

comprendiendo además dicho método, antes de la etapa a., la etapa de:

- entrenar el algoritmo predictivo, usando, como entrada, valores de expresión conocidos en muestras de tejido prostático que corresponden al diagnóstico de cáncer de próstata.

2. Método según la reivindicación 1, en el que la etapa a. se lleva a cabo usando PCR en tiempo real.

3. Método según la reivindicación anterior, en el que en dicha etapa a. se usan los siguientes cebadores:

Gas1 fwd 5'- CGCACCGTCATTGAGGAC

Gas1 rev 5'- CACGCAGTCGTTGAGCAG

Clu fwd 5'- CCTCACTTCTTCTTTCCCAAG

Clu rev 5'- GTACGGAGAGAAGGGCATC

Odc fwd 5'- CAGTCTGTCGTCTCAGTGTG

Odc rev 5'- TTCGCCCGTTCCAAAAGGAG

Hmbs fwd 5'- CGCTGCATCGCTGAAAGG

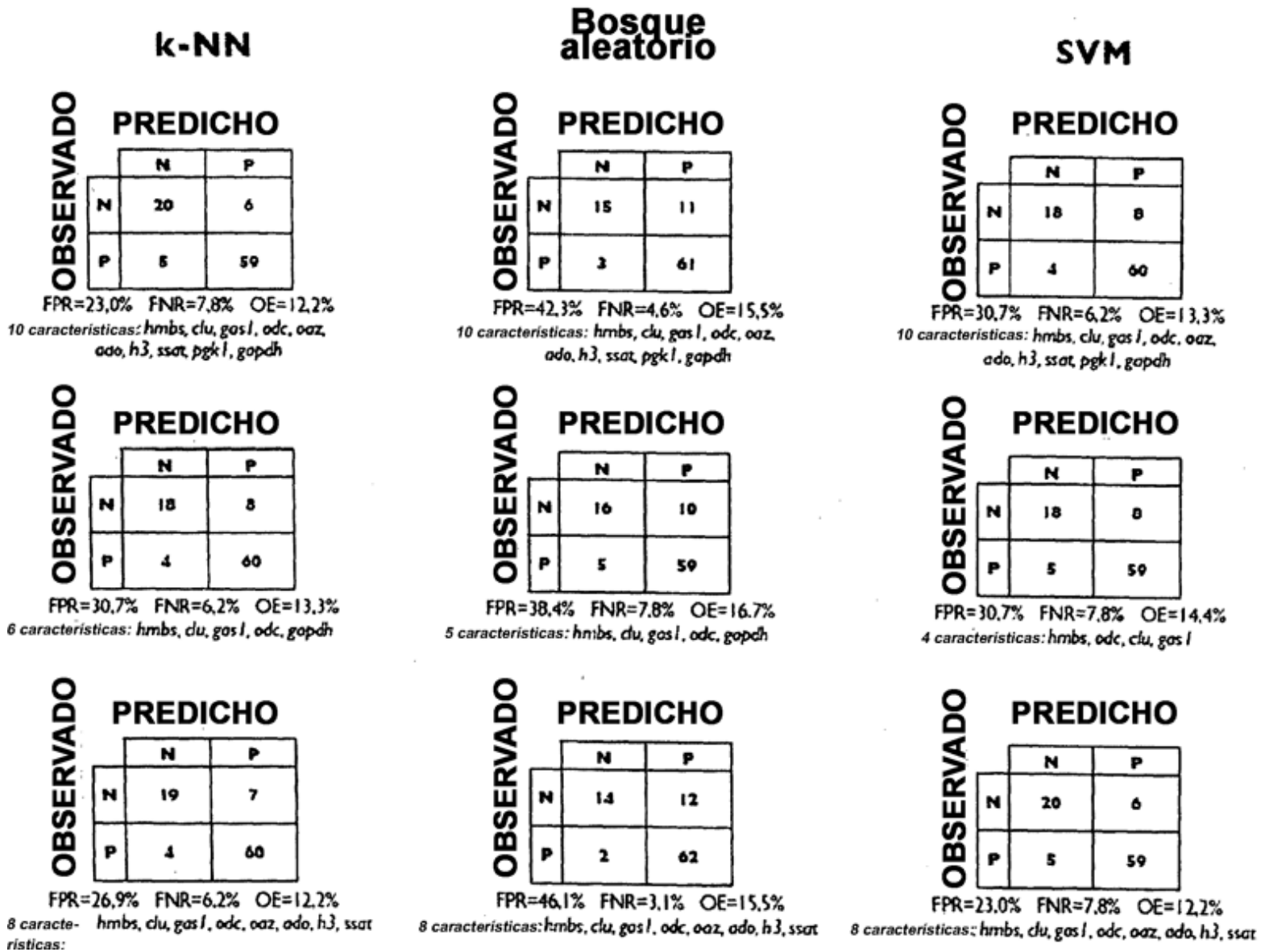
Hmbs rev 5'- ACGGCTACTGGCACACTG.

4. Método según cualquiera de las reivindicaciones anteriores, en el que dicha muestra aislada es una muestra que, después de su aislamiento, se incrustó en parafina y que, antes de su uso, se desparafina de manera adecuada y además se procesa opcionalmente.

FIG. 1

k-NN
PREDICHO
OBSERVADO
N
P
17
5
5
55
FPR=22,7% FNR=8,3% OE=12,2%
8 características: hmbs, clu, gasl, odc, oaz, odo, h3, ssat
PREDICHO
OBSERVADO
N
P
17
5
5
55
FPR=22,7% FNR=8,3% OE=12,2%
6 características: hmbs, clu, gasl, odc, oaz, odo
Bosque aleatorio
PREDICHO
OBSERVADO
N
P
13
9
4
56
FPR=40,9% FNR=6,7% OE=15,85%
8 características: hmbs, clu, gasl, odc, oaz, odo, h3, ssat
PREDICHO
OBSERVADO
N
P
13
9
3
57
FPR=40,9% FNR=5% OE=14,6%
4 características: hmbs, clu, gasl, odc
SVM
PREDICHO
OBSERVADO
N
P
18
4
4
56
FPR=18,1% FNR=6,7% OE=9,7%
8 características: hmbs, clu, gasl, odc, oaz, odo, h3, ssat
PREDICHO
OBSERVADO
N
P
17
5
4
56
FPR=18,1% FNR=6,7% OE=9,7%
4 características: hmbs, clu, gasl, odc

FIG. 2


k-NN
Bosque aleatorio
SVM
OBSERVADO
PREDICHO
N
P
20
6
5
59
FPR=23,0% FNR=7,8% OE=12,2%
10 características: hmbs, clu, gas1, odc, oaz, ado, h3, ssat, pgk1, gapdh
15
11
3
61
FPR=42,3% FNR=4,6% OE=15,5%
10 características: hmbs, clu, gas1, odc, oaz, ado, h3, ssat, pgk1, gapdh
18
8
4
60
FPR=30,7% FNR=6,2% OE=13,3%
10 características: hmbs, clu, gas1, odc, oaz, ado, h3, ssat, pgk1, gapdh
18
8
4
60
FPR=30,7% FNR=6,2% OE=13,3%
6 características: hmbs, clu, gas1, odc, gapdh
16
10
5
59
FPR=38,4% FNR=7,8% OE=16,7%
5 características: hmbs, clu, gas1, odc, gapdh
18
8
5
59
FPR=30,7% FNR=7,8% OE=14,4%
4 características: hmbs, odc, clu, gas1
19
7
4
60
FPR=26,9% FNR=6,2% OE=12,2%
8 características: hmbs, clu, gas1, odc, oaz, ado, h3, ssat
14
12
2
62
FPR=46,1% FNR=3,1% OE=15,5%
8 características: hmbs, clu, gas1, odc, oaz, ado, h3, ssat
20
6
5
59
FPR=23,0% FNR=7,8% OE=12,2%
8 características: hmbs, clu, gas1, odc, oaz, ado, h3, ssat