



(19)



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 292 110**

(51) Int. Cl.:
G06F 19/00 (2006.01)

(12)

TRADUCCIÓN DE PATENTE EUROPEA

T3

(86) Número de solicitud europea: **05712057 .8**

(86) Fecha de presentación : **26.01.2005**

(87) Número de publicación de la solicitud: **1709565**

(87) Fecha de publicación de la solicitud: **11.10.2006**

(54) Título: **Programa de ordenador para ayudar en la identificación de SNPS (polimorfismos de un solo nucleótido) con microformaciones.**

(30) Prioridad: **26.01.2004 US 539220 P**

(73) Titular/es: **Nimblegen Systems, Inc.**
1 Science Court
Madison, Wisconsin 53711, US

(45) Fecha de publicación de la mención BOPI:
01.03.2008

(72) Inventor/es: **Molla, Michael;**
Richmond, Todd;
Smith, Steve y
Albert, Thomas, J.

(45) Fecha de la publicación del folleto de la patente:
01.03.2008

(74) Agente: **Elzaburu Márquez, Alberto**

ES 2 292 110 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

Programa de ordenador para ayudar en la identificación de SNPS (polimorfismos de un solo nucleótido) con microformaciones.

Antecedentes de la invención

La llegada de la tecnología de microformaciones ("microarrays") de ADN hace posible construir una formación de centenares de miles de secuencias de ADN en un área muy pequeña, tal como del tamaño de una platina de microscopio. Véanse, por ejemplo, la patente de EE.UU. N° 6.375.903 y la patente de EE.UU. N° 5.143.854. La descripción de la patente de EE.UU. N° 6.375.903 permite la construcción de los llamados instrumentos sintetizadores de formaciones sin máscara (MAS = Maskless Array Synthesizer) en los que se usa la luz para dirigir síntesis de secuencias de ADN, siendo gobernada la dirección de la luz usando un dispositivo de microespejo digital (DMD = Digital Micromirror Device), conocido también como procesador digital de luz (DLP = Digital Light Processor). Usando un instrumento MAS, la selección de secuencias de ADN a construir en la microformación se efectúa bajo control de software, de modo que se pueden construir formaciones individualmente personalizadas a petición. En general, la tecnología de síntesis de microformaciones ADN basadas en los MAS permite la síntesis en paralelo de alrededor de más de 800.000 oligonucleótidos únicos en un área muy pequeña de una platina estándar de microscopio. Las microformaciones son sintetizadas, generalmente, usando la luz para dirigir qué oligonucleótidos son sintetizados en posiciones específicas de una formación, llamándose características estas posiciones. Típicamente, una secuencia de nucleótidos es sintetizada en cada característica de la formación, es decir, hay múltiples sondas en cada característica, pero todas esas sondas tienen la misma secuencia de nucleótidos. Para ciertas aplicaciones, sería ventajoso tener oligonucleótidos de diferentes secuencias presentes dentro de una característica de la formación, y poder controlar la relación y la dirección (5'-3', o 3'-5') de estos oligonucleótidos.

Un uso de las microformaciones es para efectuar análisis de secuencias de ADN aisladas, de organismos vivos. La ciencia ha puesto a disposición, ahora, secuencias de ADN generalizadas de genomas enteros de varios organismos importantes, incluidos los humanos. Una técnica que se puede usar para identificar una variante genética es secuenciar el ADN genómico de un individuo y, después, comparar esa secuencia con la secuencia de referencia de ese organismo. Se ha encontrado que muchas diferencias en la secuencia de ADN son presentadas como variaciones únicas en la secuencia de ADN, denominadas, a menudo, como polimorfismos de un solo nucleótido o SNP (Single Nucleotide Polymorphism). Efectuar una comparación de secuencias entre el genoma entero secuenciado de un individuo y el genoma de referencia de esa especie, es para identificar, por este mecanismo de fuerza bruta, los SNP para ese individuo. Sin embargo, este proceso es demasiado laborioso para ser práctico en la detección de SNP a gran escala. Por lo tanto, la identificación y análisis de los SNP es una tecnología a la que se ha dedicado mucha atención.

Breve resumen de la invención

La presente invención se resume como un método para identificar polimorfismos de un solo nucleótido a partir de datos de hibridación de microformaciones, siendo utilizados los datos como un procedimiento de resecuenciación. El método incluye las etapas de organizar los datos en grupos, cada uno de los cuales representa todas las variaciones observadas de la secuencia en una posición de un solo nucleótido; clasificar cada grupo como conformador o no-conformador dependiendo de si las sondas de mayor intensidad para la posición de un solo nucleótido se ajustan o no se ajustan a la secuencia de referencia en esa posición; correlacionar los conformadores y no-conformadores en un espacio de características; e identificar como SNP posibles solamente aquéllos conformadores que están situados en partes del espacio de características de los datos de microformaciones densamente pobladas con conformadores.

El método está fundado en el aprendizaje mecanizado y es neutro en cuanto a la bioquímica de hibridaciones particulares y, por tanto, mejorará su predictibilidad con el tiempo. El método es capaz de discernir SNP reales, del ruido, en un solo experimento.

Otros objetos, ventajas y características de la presente invención serán evidentes en la siguiente memoria descriptiva.

Breve descripción de los dibujos

La Figura 1 es un diagrama de flujo que ilustra el flujo lógico de información en el presente método.

La Figura 2 es una ilustración del concepto del espacio de características, y de cómo se efectúa el análisis de datos en el espacio de características.

La Figura 3 es una representación gráfica del significado de la elección de un parámetro en el algoritmo usado para el aprendizaje mecanizado.

La Figura 4 es una representación gráfica de datos de la detección de SNP reales.

La Figura 5 es una representación gráfica de datos, que muestra la tasa de llamados frente a la tasa de detección para los SNP.

Descripción detallada de la invención

El método de resecuenciación genómica eficiente, descrito aquí, ha mostrado resultados significativos y utiliza tecnología de microformaciones de oligonucleótidos para su información de entrada. En particular, este tipo de resecuenciación comienza con una microformación, o chip, que consta de un teselado completo de la secuencia de referencia. Las microformaciones tienen sondas de oligonucleótidos dispuestas en áreas denominadas características. La terminología del teselado completo con respecto a un chip o medio de microformación es que el chip tiene, por lo menos, una característica que contiene sondas que corresponden exactamente a cada 25-mer en la secuencia de referencia. Además, el chip teselado también contiene, por cada base en esta secuencia de referencia de 25-mer, tres características, cada una de las cuales contiene sondas de no-coincidencia con un solo nucleótido sustituido, es decir, una sola falta de coincidencia. Cada sonda de no-coincidencia corresponde a cada posible variación de nucleótido o SNP en esta posición, que podría estar o no estar presente en una variación de la secuencia de referencia. Después, este chip de resecuenciación de teselado se usa en un proceso de hibridación frente al ADN genómico de individuos de la especie, y en el resultado de la hibridación medida, midiendo la intensidad óptica de fluorescencia desde el ADN etiquetado de la muestra coincidente a los conjuntos de sondas de la formación. En teoría, cada vez que un SNP está presente en una muestra genómica individual particular, la característica con una sonda no coincidente que representa esta variante de SNP debe tener una señal medida de intensidad mayor que la correspondiente característica para la sonda que coincide con la secuencia de referencia. Sin embargo, debido a la impredecibilidad de la intensidad de la señal, que varía el rendimiento de la hibridación, y a varias otras fuentes de ruido y error, este método da lugar, típicamente, a muchas posiciones de base cuyas identidades están predichas incorrectamente. Soluciones actuales a este problema requieren un reajuste extenso de parámetros que implica el análisis de cantidades muy grandes de datos. Este reajuste necesita ser repetido cada vez que se cambian las condiciones experimentales. Otra limitación de los métodos actuales es que, para tener una sola sonda representada por un número suficiente de píxeles, se debe usar un escáner de alta resolución muy costoso.

En este documento se presenta una técnica que usa aprendizaje mecanizado para diferenciar los SNP potenciales, de ruido de chips y variaciones en las condiciones de hibridación. A diferencia de otros métodos, el método de esta invención no requiere dicho escáner de alta resolución y, además, no requiere ningún reajuste exterior al único chip que se está analizando. Para organismos haploides, esto produce resultados similares a los resultados publicados en la tasa de identificación de SNP para el mejor conocido de los métodos estadísticos actuales. El método de esta invención usa solamente la intensidad media de señal de cada característica del chip, y no requiere datos del exterior del chip.

Para ilustrar esta técnica, imaginemos que la tarea inmediata es identificar los SNP en el contexto de un proceso particular de resecuenciación de ADN basado en microformaciones de oligonucleótidos. Este tipo de resecuenciación consiste en teselar totalmente (produciendo características que tienen sondas correspondientes a cada 25-mer) la secuencia de referencia del ADN de un organismo a través de una región de interés. Para cada una de estas características que tienen sondas correspondientes a la secuencia de referencia, también se crean, en la microformación, otras tres características que tienen sondas únicas no coincidentes. Cada una de las sondas no coincidentes inserta una base diferente de nucleótidos en su posición central. Por ejemplo, si el ADN de referencia del organismo incluye la secuencia:

CTGACATGCAGCTATGCATGCATGAA

Entonces, la característica que corresponde a la secuencia de referencia tendrá esta secuencia complementaria:

GACTGTACGTCGATACGTACGTACTT

Para efectuar la resecuenciación convencional de la microformación, la microformación incluiría, entonces, características con sondas que tienen cada una de las siguientes secuencias:

GACTGTACGTCGAAACGTACGTACTT

GACTGTACGTCGACACGTACGTACTT

GACTGTACGTCGAGACGTACGTACTT

Para los propósitos de esta descripción, un grupo de sondas, tales como las que representan todos los posibles SNP en una posición dada, se llamará un grupo de posiciones o grupo-p.

Los detalles del método de interpretar dicho chip de resecuenciación se pueden resumir como sigue: Dados los datos de un solo chip de resecuenciación, que representa, ya sea el genoma completo de un organismo, o alguna región o regiones de interés en dicho genoma, identificar, entre las posiciones en las cuales la secuencia de muestra parece diferir de la secuencia de referencia, cuáles de estas posiciones son probablemente SNP reales, en lugar de ruido, y devolver una salida que identifique estas posiciones junto con una medida de confianza para cada una.

Para entender la nueva solución descrita en este documento, consideremos que después que el chip ha sido expuesto a la muestra, cada una de las sondas tendrá una intensidad resultante. (Cuando nos referimos a las sondas, en esta descripción, nos referimos al conjunto de sondas de una característica, y los términos sondas y características se usan algunas veces intercambiabilmente en el resto de la descripción). Cada conjunto de grupos-p de 4 de tales intensidades

se denomina un ejemplo. Para la mayoría de estos ejemplos, la mayor de las 4 intensidades será la sonda de referencia, es decir, la sonda sin base no coincidente. En la siguiente Tabla 1 se presenta una muestra de datos de un proceso de resecuenciación. Las sondas de resecuenciación se muestran en la parte superior de la Tabla 1, seguidas por los datos de intensidades deducidos experimentalmente y el agrupamiento de los datos en un ejemplo. Como se muestra en la Tabla 1, se ilustran ejemplos para los cuales se encuentra la mayor intensidad para los conformadores de secuencias de referencia, como se esperaba. Para los propósitos del algoritmo de aprendizaje mecanizado, los conformadores también se llamarán ejemplos positivos. Cuando una de las sondas no coincidentes tiene la mayor intensidad de las sondas para ese nucleótido, el grupo-p se denomina un no-conformador. Para los propósitos de los algoritmos de aprendizaje mecanizado, los no-conformadores también se llamarán ejemplos negativos. Algunos de estos ejemplos negativos reflejan SNP reales en el ADN genómico del organismo individual. Sin embargo, muchos o la mayoría de ellos son los resultados de fallos de hibridación u otros tipos de ruido que no representan un SNP en la muestra. El objetivo es determinar la diferencia entre los SNP reales y el ruido.

La metodología usada para determinar esta diferencia es un proceso de dos etapas, resumido en la Tabla 2. Un punto clave es que éste es un proceso de aprendizaje de dos etapas. Primero, se efectúa un experimento de aprendizaje mecanizado estándar para producir y comprobar modelos de grupos-p conformadores frente a grupos-p no-conformadores que usan las características disponibles (de nuevo, característica es el término para un área de la microformación dedicada a una sola sonda). Después, esto se usa como un modelo intermediario para grupos-p cuya base de mayor intensidad representa exactamente la secuencia de la muestra. Como se ilustra en la Figura 1, los grupos-p que son identificados como exactos, pero indican algo más que la base de referencia en una posición dada, son denominados SNP candidatos. Se puede usar este modelo como tal intermediario porque se han desarrollado las siguientes dos reglas generalizadas, que se han mantenido para todos los datos que se han visto hasta ahora: 1) La inmensa mayoría de ejemplos negativos son debidos a ruido en los datos. 2) Los ejemplos resultantes de hibridaciones apropiadas estarán mucho más próximos entre sí en el espacio de características (que se describe más adelante) que a ejemplos que resultan de fallos de hibridación.

Siguiendo estas reglas, un área de espacio de características con gran densidad de ejemplos positivos es improbable que contenga sondas que sean fallos de hibridación. De hecho, la probabilidad de que cualquier ejemplo dado en un área sea un error de hibridación puede ser estimada aproximadamente por la densidad de ejemplos negativos en esa área de espacio de características. Esto se ilustra en la Figura 2, donde los conformadores y los no-conformadores están situados en su geometría real en el espacio de características de la microformación, y la identificación de un SNP probablemente real está ilustrada por el hecho de que está rodeado en el espacio físico por conformadores. En otras palabras, el no-conformador en las proximidades de otros no-conformadores no es creíble, pero un no-conformador en las proximidades de conformadores es creíble. Efectuando esta estimación para cada uno de los ejemplos negativos en el conjunto de ensayo, se encuentra una probabilidad aproximada de que el no-conformador sea el resultado de un error de hibridación, en lugar de un SNP real en el genoma del individuo. Los ejemplos negativos con muy baja probabilidad de ser errores de hibridación e, inversamente, alta probabilidad de ser una reflexión correcta de la secuencia subyacente, son predichos, por este método, como que son SNP. Así pues, el modelo estará dominado por las características que identifican un grupo-p fiable. Teóricamente, en casos en que no es cierta la primera suposición, el conjunto de ensayo podría ser complementado, simplemente, con datos de otros chips que contengan pocos o ningún SNP.

La solución de la presente invención emplea aprendizaje supervisado, que es el aprendizaje a partir de un conjunto de ejemplos etiquetados, para categorizar ejemplos futuros que no están etiquetados. Otra característica importante del software de la invención es que, aunque este tipo de aprendizaje requiere, a menudo, ejemplos etiquetados humanamente, dicho software no necesita que el usuario etiquete ningún ejemplo. Las categorías son determinadas automáticamente a partir de los datos.

Para descubrir cuáles de estos ejemplos negativos son probablemente SNP reales, se aplica al problema un número de algoritmos de aprendizaje mecanizado, que incluyen Redes Neuronales Artificiales, Árboles de Decisión, y K Vecinos Más Próximos. En experimentos no descritos en este documento, se encuentra que el mejor realizador era al algoritmo de K Vecinos Más Próximos. Los K Vecinos Más Próximos es un algoritmo de clasificación muy conocido dentro de la comunidad de Aprendizaje Mecanizado. Se denomina un algoritmo de clasificación porque se usa para clasificar ejemplos basados en características conocidas. Este algoritmo particular consiste en trazar cada ejemplo en un espacio de características y, después, para cada uno de estos ejemplos, encontrar los otros K ejemplos más próximos a él en este espacio de características. Las categorías de estos K vecinos dictan la predicción. Si es mayor que algún umbral P de estos vecinos, es un ejemplo positivo; la predicción es positiva. De lo contrario, la predicción es negativa. El número de ejemplos positivos entre los K vecinos también se puede usar como una medida de confianza de la predicción.

El valor apropiado para K y definiciones útiles de proximidad y espacio de características tienden a variar entre tareas de aprendizaje. En este caso, el espacio de características de esta invención es el espacio de ejemplos de 5 dimensiones, donde 4 de las dimensiones corresponden a las intensidades de las 4 sondas del ejemplo, y la quinta dimensión es la identidad de la base en la posición central de la sonda con la mayor intensidad del ejemplo. La proximidad entre dos sondas se define como infinita en casos en que los dos ejemplos difieren en la quinta dimensión. En todos los demás casos, se define como:

$$\text{proximidad} = \sum_{i=1}^4 |\text{grupo-p}_{1i} - \text{grupo-p}_{2i}|$$

Donde grupo- p_1 = ejemplo 1, y grupo- p_2 = ejemplo 2, y grupo- p_{Ni} = la intensidad de la sonda más intensa de orden i en el ejemplo grupo- p_N .

Para escoger una buena K , se ensayan varios valores entre uno y 250 para ver cuántos falsos positivos necesitarían haber sido registrados para identificar todos los SNP reales. Los resultados de este ensayo se muestran en la Figura 3. Aunque, como se ilustra en la Figura 3, parece que hay un amplio margen de valores apropiados para K , se ha escogido $K = 100$.

Se usa un algoritmo de línea de base para comparar la sonda de mayor intensidad con la segunda sonda de mayor intensidad. Este algoritmo se describe en la Tabla 3. Si la relación está por encima de un valor umbral, se supone que la base representada por la sonda de mayor intensidad es la base de la secuencia. Si este grupo- p es un no-conformador, el algoritmo de línea de base se llama SNP candidato.

Ejemplo

Para evaluar el algoritmo, se escoge una tarea realista útil. Una cepa de virus SARS (Severe Acute Respiratory Syndrome = Síndrome Respiratorio Agudo Grave) ha sido completamente secuenciada mediante secuenciación capilar estándar. Se suministró una cepa diferente de muestra. Esta muestra difería en secuencia genética, de la referencia, en un grado desconocido. La tarea fue identificar los SNP candidatos entre estas variaciones. Las predicciones serían evaluadas subsiguientemente usando otra secuenciación capilar y varios otros métodos de laboratorio.

Usando la secuencia de referencia, se diseñó un chip de resecuenciación que incluía las cepas directas e inversas de este virus. Después se expuso este chip a ácidos nucleicos de la muestra del nuevo virus. Después se usó el algoritmo de la invención para predecir los SNP en este chip. Una vez obtenidos estos resultados, se combinaron las predicciones directas e inversas para cada posición de SNP posible, promediando las dos predicciones.

El algoritmo de la invención se comportó muy bien en esta tarea. De las 24.900 posiciones de secuencias representadas por grupos- p en este chip, 442 fueron no-conformadoras. De estas 442, la solución de la invención identificó 36 como SNP candidatos. Subsiguientes experimentos de laboratorio confirmaron 24 SNP, todos los cuales fueron identificados por el software de la invención. Este resultado total está resumido en la parte delantera de una curva Característica-Operativa-Receptora (ROC = Receiver-Operating-Characteristic), de la Figura 4.

Una curva ROC es una gráfica de positivos verdaderos frente a positivos falsos. Se obtiene ejecutando un algoritmo en varios umbrales. En el caso del algoritmo de los K vecinos más próximos, este umbral es el porcentaje de vecinos que necesitan ser etiquetados como ejemplos positivos para que un grupo- p no-conformador sea clasificado como un SNP. En este caso, los positivos falsos son no-SNP incorrectamente clasificados como SNP. Los positivos verdaderos son SNP correctamente identificados por el algoritmo de la invención. Una curva perfecta de aprendizaje alcanzaría la esquina superior izquierda, puesto que significaría que el algoritmo es capaz de identificar todos los positivos verdaderos sin producir ningún positivo falso. Aunque la curva del algoritmo de la invención no alcanza completamente esta esquina, obsérvese que domina el algoritmo de línea de base.

Tomando como base los resultados de este experimento, parece que el método descrito en este documento es muy superior al algoritmo de línea de base descrito anteriormente, y es un método fiable y eficiente para la identificación de los SNP. Sin embargo, recuérdese que esta solución clasifica algunos grupos- p como no-llamados, a saber, aquellos grupos cuyos vecinos son predominantemente no-conformadores. El porcentaje de grupos- p que son llamados (ya sean SNP o no-SNP) también se conoce como la tasa de llamados. Si esta tasa es demasiado baja, el procedimiento es inútil porque el chip no puede ser interpretado. Por tanto, para que sea útil, el software también debe dar lugar a una buena tasa de llamados.

Para ajustar la tasa de llamados, se puede usar el valor umbral, como se describió anteriormente. Como se ilustra en la Figura 5, la solución de la invención mantiene satisfactoriamente la tasa de llamados por encima del 97% para umbrales entre el 94% y el 98%, que identifican fácilmente todos los SNP de la muestra y desclasifica solamente un pequeño número de no-SNP.

REIVINDICACIONES

1. Un método implementado en ordenador, para identificar polimorfismos de un solo nucleótido a partir de datos de hibridación de microformaciones para un proceso de resecuenciación, comprendiendo el método las etapas de:

organizar los datos en grupos de posiciones que corresponden a variaciones de la secuencia en una posición de un solo nucleótido;

clasificar los grupos de posiciones como un conformador o no-conformador, dependiendo de si las sondas de mayor intensidad para la posición de un solo nucleótido se conforman o no se conforman a la secuencia de referencia en esa posición;

correlacionar los conformadores y no-conformadores en un espacio de características que representa la geometría física de la microformación; e

identificar como SNP posibles, solamente aquéllos no-conformadores que están situados en el espacio de características adyacente a los conformadores.

2. Un método como el reivindicado en la reivindicación 1, en el que la identificación de los SNP incluye un algoritmo de aprendizaje mecanizado para situar los conjuntos de datos en el espacio de características.

3. Un método como el reivindicado en la reivindicación 2, en el que el algoritmo de aprendizaje mecanizado es el algoritmo de K vecinos más próximos.

4. Un método como el reivindicado en la reivindicación 1, en el que el espacio de características es un espacio de cinco dimensiones constituido por cuatro intensidades de las cuatro sondas en el espacio de posiciones, y la identidad de la base en la posición central de la sonda.

Figura 1. Reparto de Identificación de SNPs como una Tarea de Clasificación

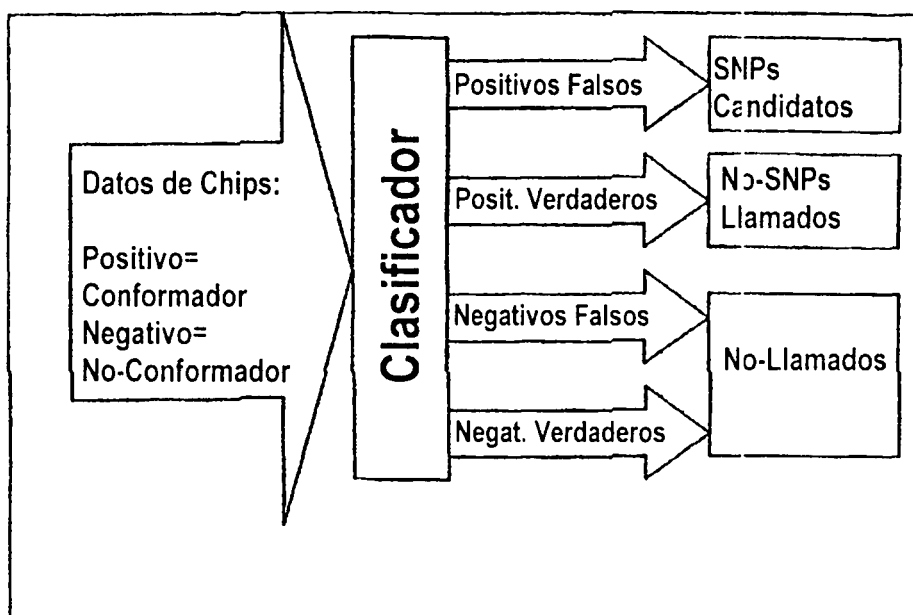


FIG. 1

Figura 2. Aprendizaje por Categoría Representante

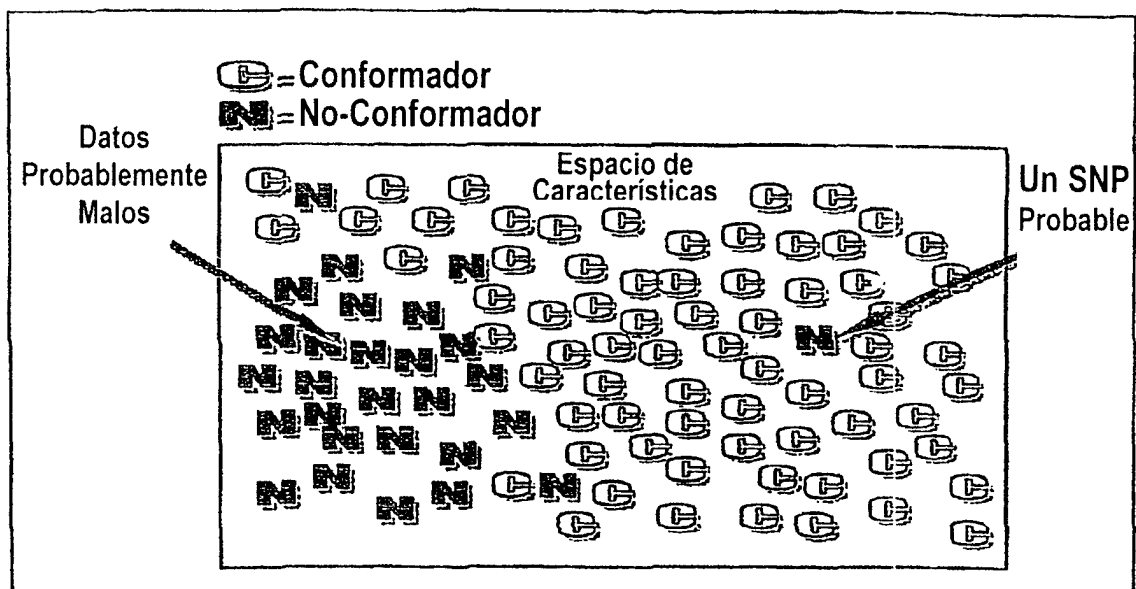


FIG. 2

Figura 3. Escogiendo K

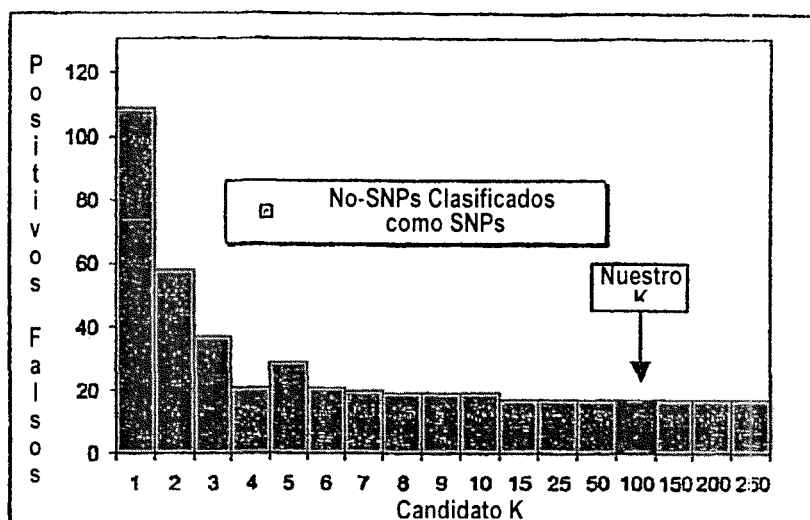


FIG. 3

Figura 4: Curva para Detección de SNP de SARS

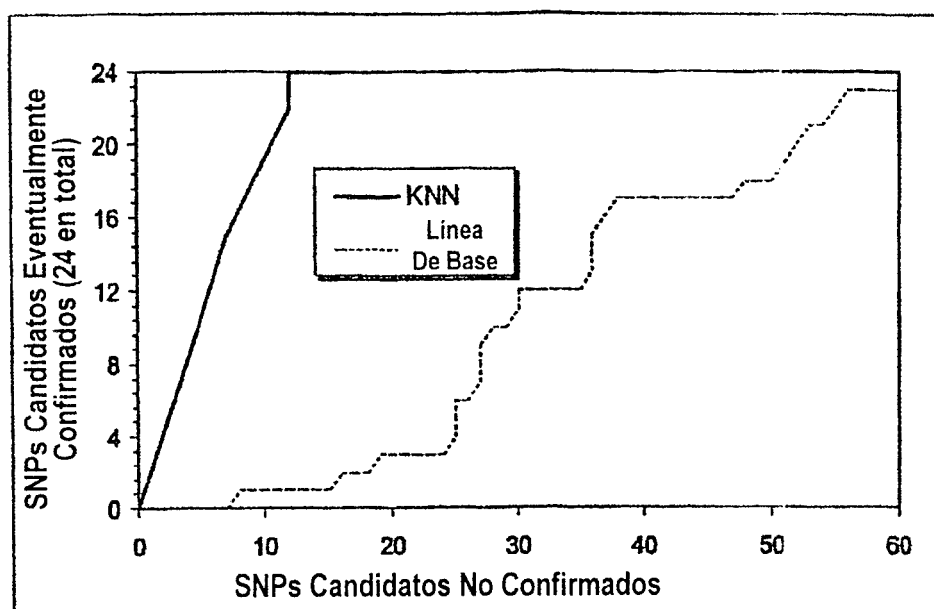


FIG. 4

Figura 5: Tasa de Llamados frente a Tasa de Detección de SNPs

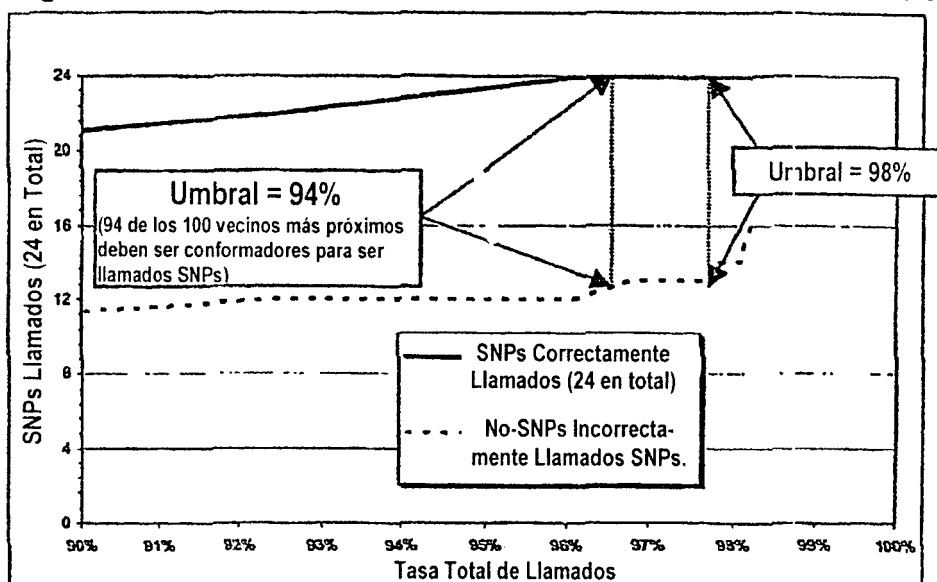


FIG. 5

Tabla 1. Las Características Usadas para Describir los Grupos-p

Secuencia de Referencia : AGCTTT**AGCATATATCGATCCTAGCATACG**ATCCTTATACITACATTACCCT...

Sondas de Resecuenciación (Sondas de Referencia en Negrita)

Grupo-p 7:	AAGCATATAT C AATCCTAGCATAC	← Sonda 7A
	AAGCATATAT C ATCCTAGCATAC	← Sonda 7C
	AAGCATATAT C GATCCTAGCATAC	← Sonda 7G
	AAGCATATAT C TATCCTAGCATAC	← Sonda 7T
Grupo-p 8:	AGCATATAT C ATCCTAGCATACG	← Sonda 8A
	AGCATATAT C GATCCTAGCATACG	← Sonda 8C
	AGCATATAT C GTATCCTAGCATACG	← Sonda 8G
	AGCATATAT C TATCCTAGCATACG	← Sonda 8T
Grupo-p 9:	GCATATAT C GATCCTAGCATACGA	← Sonda 9A
	GCATATAT C GTATCCTAGCATACGA	← Sonda 9C
	GCATATAT C GATCCTAGCATACGA	← Sonda 9G
	GCATATAT C GTATCCTAGCATACGA	← Sonda 9T

Intensidades Resultantes (obtenidas exponiendo el chip a la muestra)

Sonda	Intensidad
...	...
7A	1543
7C	342
7G	3354
7T	737
8A	1456
8C	2432
8G	212
8T	334
9A	332
9C	456
9G	232
9T	2443
...	...

← Obsérvese que, aunque la sonda de referencia del grupo-p 8 es 8A, la sonda de mayor intensidad de este grupo es 8C. Dicho grupo-p se llama no-conformador.

El Conjunto de Características:

Cada grupo-p produce un ejemplo. Las características son la base de referencia y las cuatro intensidades clasificadas. (obsérvese que el aprendiz no da información acerca de qué sonda real tiene la mayor intensidad). Aunque la sonda de mayor intensidad de este grupo-p sea o no sea la base de referencia, es decir, aunque este grupo-p sea o no sea un conformador, es la categoría del ejemplo,

Ejemplo	Base de Referencia	Intensidad 1	Intensidad 2	Intensidad 3	Intensidad 4	¿Conformador? (categoría)
...	...	...	...	...	...	...
7	G	3354	1543	737	342	VERDADERO
8	A	2435	1456	334	212	FALSO
9	T	2443	456	332	232	VERDADERO
...	...	...	...	...	...	...

Tabla 2. Nuestro Algoritmo

Llamar un <i>ejemplo</i> , a cada grupo-p
Para cada ejemplo
Entrenar el clasificador en el resto de los datos
Predecir la <i>categoría</i> del ejemplo
Si la predicción = <i>no-conformador</i>
Clasificar este ejemplo como un <i>no-llamado</i>
Si no
Si la categoría real de este ejemplo = <i>conformador</i>
Clasificar este ejemplo como un <i>no-SNP</i>
Si no
Clasificar este ejemplo como un <i>SNP candidato</i>

Tabla 3. Un Algoritmo de Línea de Base

Llamar un <i>ejemplo</i> , a cada grupo-p
Para cada ejemplo
Hacer <i>IntensidadMax</i> = Intensidad de la base de mayor intensidad en este ejemplo
Hacer <i>IntensidadSegunda</i> = Intensidad de la segunda base de mayor intensidad en este ejemplo
Si $(IntensidadMax / IntensidadSegunda) < \text{umbral}$
Clasificar este ejemplo como un <i>no-llamado</i>
Si no
Si la categoría real de este ejemplo = <i>conformador</i>
Clasificar este ejemplo como un <i>no-SNP</i>
Si no
Clasificar este ejemplo como un <i>SNP candidato</i>