(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 687 645**

(51) Int. CI.:

C12N 15/11 (2006.01)
G06F 19/18 (2011.01)
G06F 19/22 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **27.10.2004 PCT/US2004/035636**

(87) Fecha y número de publicación internacional: **12.05.2005 WO05042708**

(96) Fecha de presentación y número de la solicitud europea: **27.10.2004 E 04810056 (4)**

(97) Fecha y número de publicación de la concesión europea: **15.08.2018 EP 1692262**

(54) Título: **Método de diseño de ARNip para el silenciamiento de genes**

(30) Prioridad:

27.10.2003 US 515180 P
17.05.2004 US 572314 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
26.10.2018

(73) Titular/es:

MERCK SHARP & DOHME CORP. (100.0%)
126 East Lincoln Avenue
Rahway, NJ 07065, US

(72) Inventor/es:

JACKSON, AIMEE, L.;
BARTZ, STEVEN, R.;
BURCHARD, JULJA;
LINSLEY, PETER, S.;
GE, WEI y
CAVET, GUY, L.

(74) Agente/Representante:

VALLEJO LÓPEZ, Juan Pedro

DESCRIPCIÓN

Método de diseño de ARNip para el silenciamiento de genes

1. Campo de la invención

La presente divulgación se refiere a métodos para identificar motivos diana (*target*) de ARNip en un transcrito. La divulgación también se refiere a métodos para identificar genes inespecíficos (*off-target*) de un ARNip. Adicionalmente, la divulgación se refiere a métodos de diseño de ARNip con mayor especificidad y eficacia de silenciamiento. La divulgación también se refiere a una biblioteca de ARNip que comprende ARNip que presentan alta especificidad y eficacia de silenciamiento.

2. Antecedentes de la invención

El ARN de interferencia (ARNi) es un método poderoso para suprimir la expresión de genes en células de mamífero y ha generado muchas expectativas en la comunidad científica (Couzin, 2002, *Science* 298:2296-2297; McManus *et al*, 2002, *Nat. Rev. Genet.* **3**, 737-747; Hannon, G. J., 2002, *Nature* **418,** 244-251; Paddison *et al*, 2002, *Cancer Cell* **2**, 17-23). El ARN de interferencia se conserva a lo largo de la evolución, desde *C. elegans* a seres humanos, y se cree que actúa protegiendo a las células de la invasión por virus de ARN. Cuando una célula se infecta con un virus de ARNbc (bicatenario), una enzima de tipo RNasa lll, denominada Dicer, reconoce el ARNbc y lo dirige para la escisión. La enzima Dicer "trocea" el ARN en dúplex cortos de 21nt, denominados ARNip o ARN de interferencia pequeño, compuesto por 19nt de ribonucleótidos perfectamente emparejados con dos nucleótidos no emparejados en el extremo 3' de cada cadena. Estos dúplex pequeños se asocian con un complejo multiproteico denominado RISC y lo dirigen a transcritos de ARNm con similitud de secuencia con el ARNip. Como resultado, las nucleasas presentes en el complejo RISC escinden el transcrito de ARNm, anulando de este modo la expresión del producto génico. En el caso de infección vírica, este mecanismo daría como resultado la destrucción de transcritos víricos, impidiendo de este modo la síntesis vírica. Dado que los ARNip son bicatenarios, cualquier cadena tiene el potencial de asociarse con RISC y dirigir el silenciamiento de transcritos con similitud de secuencia.

El silenciamiento específico de genes promete la posibilidad de aprovechar los datos del genoma humano para esclarecer la función génica, identificando dianas farmacológicas y desarrollando terapias más específicas. Muchas de estas aplicaciones conllevan un alto grado de especificidad de los ARNip por sus dianas deseadas. La hibridación cruzada con transcritos que contienen identidad parcial con la secuencia de ARNip, puede dar lugar a fenotipos que reflejan el silenciamiento de transcritos no deseados, además del gen diana. Esto confundiría la identificación de los genes implicados en el fenotipo. Numerosos informes en la bibliografía proponen la especificidad exquisita de los ARNip, sugiriendo la necesidad de que exista identidad casi perfecta con la secuencia de ARNip (Elbashir *et al*, 2001. *EMBO J.* 20: 6877-6888; Tuschl *et al*, 1999, *Genes Dev.* 13: 3191-3197; Hutvagner *et al*, *Sciencexpress* 297: 2056-2060). Un informe reciente sugiere la necesidad de que exista complementariedad de secuencia perfecta para la escisión del transcrito dirigido al ARNip, mientras que la complementariedad parcial conducirá a la represión de la traducción sin degradación del transcrito, a la manera de los microARN (Hutvagner *et al*., Sciencexpress 297: 2056-2060).

La función biológica de los ARN reguladores pequeños, incluyendo los ARNip y los miARN, no se comprende bien. Una cuestión que prevalece se refiere al mecanismo por el cual se determinan las distintas rutas de silenciamiento de estas dos clases de ARN reguladores. Los miARN son ARN reguladores expresados a partir del genoma y se procesan a partir de estructuras precursoras de tipo tallo-bucle para producir ácidos nucleicos monocatenarios que se unen a secuencias en la UTR 3' del ARNm diana (Lee *et al*, 1993, *Cell* 75: 843-854; Reinhart *et al*, 2000, *Nature* 403: 901-906; Lee *et al*, 2001, *Science* 294: 862-864; Lau *et al*, 2001, *Science* 294: 858-862; Hutvagner *et al*, 2001, *Science* 293: 834-838). Los miARN se unen a secuencias de transcripción con una complementariedad solo parcial (Zeng *et al*, 2002, *Molec. Cell* 9:1327-1333) y reprimen la traducción sin afectar a los niveles estables de ARN (Lee *et al*, 1993, *Cell* 75:843-854; Wightman *et al*, 1993, *Cell* 75:855-862). Tanto los miARN como los ARNip se procesan mediante Dicer y se asocian con componentes del complejo de silenciamiento inducido por ARN (Hutvagner *et al*, 2001, *Science* 293: 834-838; Grishok *et al*, 2001, *Cell* 106: 23-34; Ketting *et al*, 2001, *Genes Dev.* 15: 2654-2659; Williams *et al*, 2002, *Proc. Natl. Acad. Sci. USA* 99: 6889-6894; Hammond *et al*, 2001, *Science* 293: 1146-1150; Mourlatos *et al*, 2002, *Genes Dev.* 16: 720-728). Un informe reciente (Hutvagner *et al*, 2002, *Sciencexpress* 297: 2056-2060) establece la hipótesis de que la regulación génica a través de la ruta del miARN frente a la ruta del ARNip, se determina exclusivamente por el grado de complementariedad con el transcrito diana. Se especula que los ARNip con identidad solo parcial con la diana de ARNm actuarán en la represión traduccional, similar a un miARN, en lugar de desencadenar la degradación del ARN.

También se ha demostrado que para el silenciamiento de genes *in vivo* puede utilizarse ARNip y ARNhp. La capacidad de utilizar ARNip y ARNhp para el silenciamiento de genes *in vivo* tiene el potencial de permitir la selección y el desarrollo de los ARNip para su uso terapéutico. Un reciente informe destaca el potencial de la aplicación terapéutica de los ARNip. La apoptosis mediada por Fas está implicada en un amplio espectro de enfermedades hepáticas, donde se podrían salvar vidas inhibiendo la muerte apoptótica de hepatocitos. Song (Song *et al*. 2003, *Nat. Medicine* 9, 347-351) inyectó a ratones, por vía intravenosa, ARNip dirigido contra el receptor de

Fas. El gen Fas se silenció en hepatocitos de ratón a niveles de ARNm y proteína, impidiendo la apoptosis y protegiendo a los ratones de hepatitis inducida por lesión hepática. Por tanto el silenciamiento de la expresión de Fas posee una promesa terapéutica para impedir el daño hepático protegiendo a los hepatocitos de la citotoxicidad. En otro ejemplo, se inyectó a ratones, por vía intraperitoneal, ARNip que se dirigía a TNF-a. La expresión del gen de TNF-a inducida por lipopolisacárido se inhibió y estos ratones se protegieron de la septicemia. En su conjunto, estos resultados sugieren que los ARNip pueden actuar *in vivo* y pueden tener potencial como fármacos terapéuticos (Sorensen *et al*, 2003, *J. Mol. Biol.* 327, 761-766).

Martinez *et al*. revelaron que el ARN de interferencia puede utilizarse para dirigir selectivamente mutaciones oncogénicas (Martinez *et al*, 2002, *Proc. Natl. Acad. Sci. USA* 99: 14849-14854). En este informe, se mostró que un ARNip que se dirigía a la región del mutante R248W de p53 que contiene la mutación puntual, silenciaba la expresión del mutante p53 pero no la del p53 de tipo silvestre (*wild-type*).

Wilda *et al*. revelaron que un ARNip que se dirigía al ARNm de la fusión M-BCR/ABL, podía utilizarse para reducir el ARNm de M-BCR/ABL y la oncoproteína de M-BRC/ABL en células leucémicas (Wilda *et al*, 2002, Oncogene 21: 5716-5724). Sin embargo, el informe también mostró que aplicando el ARNip en combinación con Imatinib, un inhibidor de tirosina quinasa ABL de molécula pequeña, a células leucémicas no aumentaba más la inducción de la apoptosis.

La patente de Estados Unidos n.º 6.506.559 desvela un procedimiento de ARN de interferencia para inhibir la expresión de un gen diana en una célula. El procedimiento comprende introducir, parcial o completamente, ARN bicatenario que tiene una secuencia en la región dúplex que es idéntica a una secuencia en el gen diana en la célula o en el medio extracelular. También se descubrió que secuencias de ARN con inserciones, deleciones y mutaciones puntuales individuales, con relación a la secuencia diana, eran eficaces para la inhibición de la expresión.

La publicación de solicitud de patente de Estados Unidos n.º US 2002/0086356, desvela ARN de interferencia en un sistema *in vitro* de *Drosophila* utilizando segmentos de ARN de 21-23 nucleótidos (nt) de longitud. La Publicación de Solicitud de Patente enseña que cuando estos fragmentos de 21-23 nt se purifican y se añaden de nuevo a extractos de *Drosophila*, actúan como mediadores en el ARN de interferencia específico de secuencia en ausencia de ARNbc largo. La publicación de solicitud de patente también enseña que también pueden utilizarse oligonucleótidos de la misma naturaleza o similar, sintetizados químicamente, para dirigir ARNm específicos a la degradación en células de mamífero.

La publicación PCT WO 02/44321 desvela que el ARN bicatenario (ARNbc) de 19 a 23nt de longitud induce el silenciamiento génico postranscripcional específico de secuencia en un sistema *in vitro* de *Drosophila*. La publicación PCT enseña que los ARN de interferencia pequeños (ARNip), generados mediante una reacción de procesamiento de tipo RNasa III, a partir de ARNbc largo o dúplex de ARNip, sintetizados químicamente, con extremos salientes en posición 3', actúan como mediadores eficaces en la escisión de ARN diana en el lisado y el sitio de escisión se localiza cerca del centro de la región que abarca el ARNip guía. La publicación PCT también proporciona pruebas de que la dirección del procesamiento de ARNbc determina si el ARN diana de sentido o antisentido idéntico, puede escindirse a través del complejo de ARNip producido.

La publicación de solicitud de patente de Estados Unidos n.º US 2002/016216 desvela un método para atenuar la expresión de un gen diana en células cultivadas, introduciendo en las células, una cantidad suficiente de ARN bicatenario (ARNbc) que contenía una secuencia de nucleótidos que se hibridaba, en condiciones rigurosas, con una secuencia de nucleótidos del gen diana, para atenuar la expresión del gen diana.

La publicación PCT WO 03/006477 desvela precursores de ARN modificados por ingeniería genética, que cuando se expresan en una célula, dicha célula los procesa para producir ARN de interferencia pequeños (ARNip) diana, que silencian selectivamente genes diana (escindiendo ARNm específicos) utilizando la ruta de ARN de interferencia (ARNi) de la propia célula. La publicación PCT enseña que introduciendo *in vivo* en las células, moléculas de ácido nucleico que codifican estos precursores de ARN modificados por ingeniería genética, con secuencias reguladoras apropiadas, la expresión de los precursores de ARN modificados por ingeniería genética, puede controlarse selectivamente tanto temporal como espacialmente, es decir, a tiempos concretos y/o en tejidos, órganos o células concretos.

Elbashir *et al.*, desvelaron un análisis sistemático de la longitud, estructura secundaria, esqueleto de azúcar y especificidad de secuencia del ARNip para el ARNi (Elbashir *et al*., 2001. EMBO J. 20: 6877-6888). Basándose en el análisis, Elbashir propuso normas para diseñar los ARNip.

Aza-Blanc *et al.,* publicaron correlaciones entre la eficacia del silenciamiento y el contenido de GC de las regiones 5' y 3' de la secuencia diana de 19 pb (Aza-Blanc *et al*, 2003, Mol. Cell 12: 627-637). Se descubrió que las secuencias que se dirigían a los ARNip con un extremo 5' rico en GC y un extremo 3' pobre en GC, eran las que tenían mejor rendimiento.

El documento WO 03/065281 desvela algoritmos estadísticos para predecir el plegamiento y la accesibilidad

específica y diseño de ácidos nucleicos.

Barash *et al*. desvelaron una estrategia hipergeométrica para descubrir supuestos sitios de unión al factor de transcripción *{Lecture Notes in Computer Science,* Springer Verlag (Berlín), vol. 2149, págs. 278-293, 1 de enero 2001).

Lim *et al* desvelaron un procedimiento informático para identificar genes de miARN en *C. elegans (Genes and Development,* 17 (8), 2003, 991-1008).

En el presente documento, el comentario o cita de una referencia no debe interpretarse como una admisión de que dicha referencia sea una técnica anterior a la presente invención.

3. Sumario de la invención

La invención es como se expone más adelante en las reivindicaciones del presente documento.

En un aspecto, la invención proporciona un método para seleccionar, a partir de una pluralidad de ARNip diferentes, uno o más ARNip para el silenciamiento de un gen diana en un organismo, dirigiéndose cada ARNip diferente en dicha pluralidad de ARNip diferentes, a una secuencia diana diferente en un transcrito de dicho gen diana, comprendiendo dicho método

(a) calcular una puntuación
para un motivo de secuencia dirigido correspondiente en dicho transcrito, para cada dicho ARNip diferente en dicha pluralidad de ARNip diferentes, en el que dicha puntuación se calcula utilizando una matriz de puntuación específica de posición (PSSM); en el que cada uno de dichos motivos de secuencia dirigidos comprende al menos una parte de la secuencia diana del ARNip correspondiente y/o una segunda secuencia en una región que flanquea dicha secuencia diana;
(b) clasificar dicha pluralidad de ARNip diferentes de acuerdo con dichas puntuaciones; y
(c) seleccionar uno o más ARNip de dichos ARNip clasificados;

en el que al menos una de las etapas (a), (b) o (c) se realiza mediante un ordenador adecuadamente programado.

En una realización preferida, cada motivo de secuencia comprende la secuencia diana del ARNip de direccionamiento.

En una realización, cada motivo de secuencia es una secuencia de nucleótidos de *L* nucleótidos, siendo *L* un número entero y la matriz de puntuación específica de posición es {log (e_{ij}/p_{ij})}, en la que e_{ij} es el peso del nucleótido *i* en la posición *j*, p_{ij} es el peso del nucleótido *i* en la posición *j* en una secuencia al azar, e *i = G, C, A, U(T), j = 1, ..., L*. En otra realización, cada motivo de secuencia es una secuencia de nucleótidos de *L* nucleótidos, siendo *L* un número entero y la matriz de puntuación específica de posición es {log (e_{ij}/p_{ij})}, en la que e_{ij} es el peso del nucleótido *i* en la posición *j*, p_{ij} es el peso del nucleótido *i* en la posición *j* en una secuencia al azar, e *i = G o C, A, U(T), j = 1, ..., L.*

En una realización, la puntuación de cada ARNip se calcula de acuerdo con la ecuación

$$\textit{Puntuación} = \sum_{t=1}^{L} \ln(e_t / p_t)$$

en la que e_t y p_t son, respectivamente, pesos del nucleótido en la posición *t* en el motivo de secuencia como se determina de acuerdo con la matriz de puntuación específica de posición y en una la secuencia al azar.

En otra realización, cada motivo de secuencia comprende la secuencia diana del ARNip de direccionamiento y al menos una secuencia flanqueante. Preferentemente, cada motivo de secuencia comprende la secuencia diana del ARNip de direccionamiento y una secuencia flanqueante en 5' y una secuencia flanqueante en 3'. En una realización, cada una de la secuencia flanqueante en 5' y secuencia flanqueante en 3', es una secuencia de *D* nucleótidos, siendo *D* un número entero. En una realización específica, cada secuencia diana es una secuencia de 19 nucleótidos, y cada una de la secuencia flanqueante en 5' y secuencia flanqueante en 3' es una secuencia de 10 nucleótidos. En otra realización específica, cada secuencia diana es una secuencia de 19 nucleótidos, y cada una de la secuencia flanqueante en 5' y secuencia flanqueante en 3' es una secuencia de 50 nucleótidos.

Preferentemente, el uno o más de los ARNip, consta de al menos 3 ARNip. En otra realización, el método comprende adicionalmente una etapa de des-solapamiento, que comprende seleccionar una pluralidad de los ARNip entre los al menos 3 ARNip, de tal manera que los ARNip en la pluralidad son suficientemente diferentes en una medición de diversidad de secuencia. En una realización, la medición de diversidad es una medición cuantificable, y la selección en la etapa de des-solapamiento comprende seleccionar ARNip que tienen una diferencia en la medición de diversidad de secuencia entre diferentes ARNip seleccionados por encima de un umbral determinado.

En una realización, la medición de diversidad de secuencia es el contenido global de GC de los ARNip. En una realización, el umbral determinado es 5 %. En otra realización, la medición de diversidad de secuencia es la distancia entre los ARNip a lo largo de la secuencia de transcripción. En una realización, el umbral es de 100 nucleótidos. En otra realización más, la medida de diversidad de secuencia es la identidad del dímero principal de los ARNip, donde a cada uno de los 16 posibles dímeros principales se le asigna una puntuación de 1-16, respectivamente. En una realización, el umbral es de 0,5.

En otra realización, el método comprende adicionalmente una etapa de selección de uno o más ARNip basándose en la especificidad de silenciamiento, comprendiendo la etapa de selección basándose en la especificidad de silenciamiento, (i) para cada uno de la pluralidad de los ARNip, predecir genes inespecíficos del ARNip de entre una pluralidad de genes, donde los genes inespecíficos son genes distintos al gen diana y son directamente silenciados por el ARNip; (ii) clasificar la pluralidad de los ARNip de acuerdo con sus respectivos números de genes inespecíficos; y (iii) seleccionar uno o más ARNip para los cuales el número de genes inespecíficos está por debajo de un umbral determinado.

En una realización, la predicción comprende (i1) evaluar la secuencia de cada uno de la pluralidad de genes basándose en un patrón de coincidencia de secuencia de ARNip predeterminado; y (i2) predecir el gen como un gen inespecífico si el gen comprende una secuencia que coincide con el ARNip basándose en el patrón de coincidencia de secuencia. En una realización, la etapa de evaluación comprende identificar un alineamiento del ARNip con una secuencia en un gen mediante un alineamiento FastA de baja rigurosidad.

En una realización, cada ARNip tiene *L* nucleótidos en su región dúplex y el patrón de coincidencia se representa mediante una matriz de puntuación específica de posición de coincidencia de posición (pmPSSM, siglas del inglés *position match position-specific score matrix*), constando la matriz de puntuación específica de posición de coincidencia de posición, de pesos de diferentes posiciones en un ARNip que coinciden con posiciones en la secuencia de transcripción en un transcrito inespecífico $\{P_j\}$, donde $j = 1, ..., L$, P_j es el peso de una coincidencia en la posición *j*.

En otra realización, la etapa (i1) comprende calcular una puntuación de coincidencia de posición, Puntuacióncp, de acuerdo con la ecuación

$$Puntuación cp = \sum_{i=1}^{L} \ln(E_i / 0{,}25)$$

en la que $E_i = P_i$ si la posición *i* es una coincidencia y $E_i = (1-P_i)/3$ si la posición *i* es una coincidencia errónea; y la etapa (i2) comprende predecir el gen como un gen inespecífico si la puntuación de coincidencia de posición es mayor que la de un umbral determinado.

En una realización preferida, *L* es 19 y la pmPSSM se da en la Tabla I.

Preferentemente, la pluralidad de genes comprende todos los genes exclusivos conocidos del organismo, distintos del gen diana.

En una realización, la matriz de puntuación específica de posición (PSSM) se determina mediante un método que comprende (aa) identificar una pluralidad de *N* ARNip que consta de ARNip que tienen una región dúplex de 19 nucleótidos y que tienen una eficacia de silenciamiento por encima de un umbral elegido; (bb) identificar, para cada ARNip, un motivo de secuencia funcional, comprendiendo el motivo de secuencia funcional una secuencia diana de 19 nucleótidos del ARNip y una secuencia flanqueante en 5' de 10 nucleótidos y una secuencia flanqueante en 3' de 10 nucleótidos; (cc) calcular una matriz de frecuencia $\{f_{ij}\}$, en la que $i = G, C, A, U(T); j = 1, 2, ..., L$, y en la que f_{ij} es la frecuencia del *i-gésimo* nucleótido en la *j-ésima* posición, basándose en los motivos de secuencia funcional de los ARNip de acuerdo con la ecuación

$$f_{ij} = \sum_{k=1}^{N} \delta_{ik}(j),$$

en la que

$$\delta_{ik}(j) = \begin{cases} 1, \text{si } k = i \\ 0, \text{si } k \neq i \end{cases},$$

y (d) determinar la PSSM calculando e_{ij} de acuerdo con la ecuación

$$e_{ij} = \frac{f_{ij}}{N}.$$

En otra realización, la matriz de puntuación específica de posición (PSSM) se obtiene mediante un método que comprende (aa) inicializar la PSSM con pesos al azar; (bb) seleccionar aleatoriamente un peso w_{ij} obtenido en (aa); (cc) cambiar el valor del peso seleccionado para generar una psPSSM de ensayo que comprenda el peso seleccionado que tenga el valor cambiado; (dd) calcular una puntuación para cada una de una pluralidad de motivos de secuencia funcional de ARNip utilizando la PSSM de ensayo de acuerdo con la ecuación

$$\text{Puntuación} = \sum_{k=1}^{L} \ln(w_k / p_k)$$

en la que w_k y p_k son, respectivamente, pesos de un nucleótido en la posición k en el motivo de secuencia funcional y en una secuencia aleatoria; (ee) calcular la correlación de la puntuación y una medida de una característica de un ARNip entre la pluralidad de los motivos de secuencia funcional de los ARNip; (ff) repetir las etapas (cc)-(ee) para una pluralidad de diferentes valores del peso seleccionado en un intervalo determinado y mantener el valor que corresponda a la mejor correlación para el peso seleccionado; y (gg) repetir las etapas (bb)-(ff) durante un número de veces elegido; determinando de este modo la PSSM.

En una realización, el método comprende adicionalmente seleccionar la pluralidad de motivos de secuencia funcional de ARNip mediante un método que comprende (i) identificar una pluralidad de ARNip que conste de los ARNip que tienen diferentes valores en la medida; (ii) identificar una pluralidad de cada uno de los motivos de secuencia funcional de ARNip correspondiente a un ARNip en la pluralidad de los ARNip. En una realización preferida, la característica es la eficacia de silenciamiento.

En una realización, la pluralidad de los N ARNip se dirige a una pluralidad de diferentes genes que tienen diferentes abundancias de transcrito en una célula.

En una realización, la etapa (b) se realiza seleccionando uno o más ARNip que tienen las puntuaciones más altas. En otra realización, la etapa (b) se realiza seleccionando uno o más ARNip que tienen una puntuación más próxima a un valor predeterminado, siendo el valor predeterminado el valor de la puntuación correspondiente a la mediana máxima de la eficacia de silenciamiento de una pluralidad de motivos de secuencia de ARNip. En una realización preferida, la pluralidad de motivos de secuencia de ARNip son motivos de secuencia en transcritos que tienen un nivel de abundancia menor de aproximadamente 3-5 copias por célula.

En otra realización, la etapa (b) se realiza seleccionando uno o más ARNip que tienen una puntuación dentro de un intervalo predeterminado, siendo el intervalo predeterminado un intervalo de puntuación correspondiente a una pluralidad de motivos de secuencia de ARNip que tienen un nivel de eficacia de silenciamiento determinado. En una realización, el porcentaje de eficacia de silenciamiento está por encima del 50 %, 75 % o 90 % a una dosis de ARNip de aproximadamente 100 nM.

En una realización preferida, la pluralidad de motivos de secuencia de ARNip, son motivos de secuencia en transcritos que tienen un nivel de abundancia menor de aproximadamente 3-5 copias por célula.

En otra realización preferida, la pluralidad de N ARNip comprende al menos 10, 50, 100, 200 o 500 ARNip diferentes.

En otra realización, la matriz de puntuación específica de posición (PSSM) comprende w_k, $k = 1, ..., L$, siendo w_k una diferencia en la probabilidad de encontrar el nucleótido G o C en una posición de secuencia k entre un primer tipo de ARNip y un segundo tipo de ARNip, y la puntuación de cada cadena se calcula de acuerdo con la ecuación

$$\text{Puntuación} = \sum_{k=1}^{L} w_k .$$

En una realización, el primer tipo de ARNip consta de uno o más ARNip que tienen eficacia de silenciamiento no menor que un primer umbral y el segundo tipo de ARNip consta de uno o más ARNip que tienen eficacia de silenciamiento menor que un segundo umbral.

En una realización, la diferencia en la probabilidad se describe mediante una suma de curvas gaussianas, representando cada una de las curvas gaussianas, la diferencia en la probabilidad de encontrar una G o C en una posición de secuencia diferente.

En una realización, el porcentaje del primer y segundo umbral es del 75 % a una dosis de ARNip de 100 nM.

También se desvela un método para seleccionar, a partir de una pluralidad de ARNip diferentes, uno o más ARNip para el silenciamiento de un gen diana en un organismo, direccionándose cada una de la pluralidad de ARNip diferentes, a una secuencia diana diferente en un transcrito del gen diana, comprendiendo el método (a) clasificar la pluralidad de ARNip diferentes de acuerdo con una composición de bases posicional de secuencias complementarias inversas de cadenas en sentido de los ARNip; y (b) seleccionar uno o más ARNip de los ARNip clasificados.

En una realización, la etapa de clasificación se realiza (a1) determinando una puntuación para cada ARNip diferente, donde la puntuación se calcula utilizando una matriz de puntuación específica de posición; y (a2) clasificando la pluralidad de ARNip diferentes de acuerdo con la puntuación.

En una realización, el ARNip tiene una secuencia de nucleótidos de *L* nucleótidos en su región dúplex, siendo *L* un número entero, donde la matriz de puntuación específica de posición comprende w_k, $k = 1, ..., L$, siendo w_k una diferencia en la probabilidad de encontrar el nucleótido *G* o *C* en una posición de secuencia *k* entre el complemento inverso de la secuencia en sentido de un primer tipo de ARNip y el complemento inverso de la secuencia en sentido de un segundo tipo de ARNip, y la puntuación de cada complemento inverso se calcula de acuerdo con la ecuación

$$Puntuación = \sum_{k=1}^{L} w_k \ .$$

En una realización, el primer tipo de ARNip consta de uno o más ARNip que tienen eficacia de silenciamiento no menor que un primer umbral y el segundo tipo de ARNip consta de uno o más ARNip que tienen eficacia de silenciamiento menor que un segundo umbral.

En otra realización, la diferencia en la probabilidad se describe mediante una suma de curvas gaussianas, representando cada una de las curvas gaussianas, la diferencia en la probabilidad de encontrar una G o C en una posición de secuencia diferente.

En una realización, el porcentaje del primer y segundo umbral es del 75 % a una dosis de ARNip de 100 nM.

También se desvela un método para seleccionar, a partir de una pluralidad de ARNip diferentes, uno o más ARNip para el silenciamiento de un gen diana en un organismo, direccionándose cada una de la pluralidad de ARNip diferentes, a una secuencia diana diferente en un transcrito del gen diana, comprendiendo el método: (i) para cada una de la pluralidad de ARNip diferentes, la predicción de genes inespecíficos del ARNip a partir de una pluralidad de genes, donde los genes inespecíficos son genes distintos al gen diana y se silencian directamente mediante el ARNip; (ii) clasificar la pluralidad de ARNip diferentes de acuerdo con el número de genes inespecíficos; y (iii) seleccionar uno o más ARNip para los cuales el número de genes inespecíficos está por debajo de un umbral determinado.

En una realización, la predicción comprende (i1) evaluar la secuencia de cada una de la pluralidad de genes basándose en un patrón de coincidencia de secuencia de ARNip predeterminado; y (i2) predecir un gen como un gen inespecífico si el gen comprende una secuencia que coincide con la del ARNip basándose en el patrón de coincidencia de secuencia.

En una realización, cada ARNip tiene *L* nucleótidos en su región dúplex, y el patrón de coincidencia de secuencia se representa mediante una matriz de puntuación específica de posición de coincidencia de posición (pmPSSM), constando la matriz de puntuación específica de posición de coincidencia de posición, de pesos de diferentes posiciones en un ARNip que coinciden con posiciones de secuencia de transcrito en un transcrito inespecífico $\{P_j\}$, donde $j = 1, ..., L$, P_j es el peso de una coincidencia en la posición *j*. En otra realización, la etapa (i1) comprende calcular una puntuación de coincidencia de posición, Puntuacióncp de acuerdo con la ecuación

$$Puntuacióncp = \sum_{i=1}^{L} \ln(E_i / 0{,}25)$$

en la que $E_i = P_i$ si la posición *i* es una coincidencia y $E_i = (1-P_i)/3$ si la posición *i* es una coincidencia errónea; y la etapa (i2) comprende predecir el gen como un gen inespecífico si la puntuación de coincidencia de posición es mayor que la de un umbral determinado.

En una realización preferida, *L* es 19, y la pmPSSM se da en la Tabla I.

En una realización, la pluralidad de genes comprende todos los genes exclusivos conocidos del organismo, distintos del gen diana.

También se desvela una biblioteca de ARNip, que comprende una pluralidad de ARNip para cada uno de una pluralidad de genes diferentes de un organismo, en el que cada ARNip consigue un silenciamiento de al menos 75 %, al menos 80 % o al menos 90 % de su gen diana. En una realización, la pluralidad de ARNip consta de al menos 3, al menos 5, o al menos 10 ARNip. En otra realización, la pluralidad de diferentes genes consta de al menos 10, al menos 100, al menos 500, al menos 1.000, al menos 10.000 o al menos 30.000 genes diferentes.

También se desvela un método para determinar una matriz de puntuación específica de posición de composición de bases (bsPSSM) $\{\log(e_{ij}/p_{ij})\}$ para representar patrones de composición de bases de motivos de secuencia funcional de ARNip de *L* nucleótidos en transcritos, donde *i = G, C, A, U(T)* y *j = 1,2, ..., L,* y cada motivo de secuencia funcional de ARNip comprende al menos una parte de la secuencia diana del ARNip de direccionamiento correspondiente y/o una secuencia en una región de secuencia que flanquea la secuencia diana, comprendiendo el método (a) identificar una pluralidad de *N* ARNip diferentes que consta de ARNip que tienen una eficacia de silenciamiento por encima de un umbral elegido; (b) identificar una pluralidad de *N* motivos de secuencia funcional de ARNip correspondientes, uno para cada ARNip diferente; (c) calcular una matriz de frecuencia $\{f_{ij}\}$, donde *i = G, C, A, U(T); j = 1, 2, ..., L*, y donde f_{ij} es la frecuencia del *i*gésimo nucleótido en la *j*ésima posición, en función de la pluralidad de los motivos de secuencia funcionales de ARNip de acuerdo con la ecuación

$$f_{ij} = \sum_{k=1}^{N} \delta_{ik}(j),$$

en la que $\delta_{ik}(j) = \begin{cases} 1, \text{si } k = i \\ 0, \text{si } k \neq i \end{cases}$, y (d) determinar la psPSSM calculando e_{ij} de acuerdo con la ecuación

$$e_{ij} = \frac{f_{ij}}{N}.$$

En una realización, cada motivo funcional de ARNip comprende la secuencia diana del ARNip de direccionamiento correspondiente y una o ambas secuencias flanqueantes de la secuencia diana.

En una realización, cada ARNip tiene *M* nucleótidos en su región dúplex, y cada motivo de secuencia funcional de ARNip consta de una secuencia diana de ARNip de *M* nucleótidos, una secuencia flanqueante en 5' de D_1 nucleótidos y una secuencia flanqueante en 3' de D_2 nucleótidos.

En una realización específica, cada ARNip tiene 19 nucleótidos en su región dúplex, y cada motivo de secuencia funcional de ARNip consta de una secuencia diana de ARNip de 19 nucleótidos, una secuencia flanqueante en 5' de 10 nucleótidos y una secuencia flanqueante en 3' de 10 nucleótidos. En otra realización específica, cada ARNip tiene 19 nucleótidos en su región dúplex, y cada motivo de secuencia funcional de ARNip consta de una secuencia diana de ARNip de 19 nucleótidos, una secuencia flanqueante en 5' de 50 nucleótidos y una secuencia flanqueante en 3' de 50 nucleótidos.

En una realización, la pluralidad de cada uno de los *N* ARNip se dirige a un gen cuya abundancia de transcritos está dentro de un intervalo determinado. En una realización, el intervalo es de al menos aproximadamente 5, 10 o 100 transcritos por célula. En otra realización, el intervalo es menor que aproximadamente 3-5 transcritos por célula.

En otra realización, el porcentaje del umbral de silenciamiento es 50 %, 75 % o 90 % a una dosis de ARNip de aproximadamente 100 nM. En otra realización adicional, la pluralidad de los *N* ARNip comprende 10, 50, 100, 200 o 500 ARNip diferentes.

También se desvela un método para determinar una matriz de puntuación específica de posición de composición de bases (bsPSSM, siglas del inglés *base composition position-specific score matrix*) $\{w_{ij}\}$ para representar un patrón de composición de bases que represente una pluralidad de motivos de secuencia funcional de ARNip diferentes de *L* nucleótidos, donde *i = G, C, A, U(T)* y *j = 1, 2, ..., L*, y cada motivo de secuencia funcional de ARNip comprende al menos una parte de la secuencia diana del ARNip de direccionamiento correspondiente y/o una secuencia en una región de secuencia que flanquea la secuencia diana de ARNip, comprendiendo el método (a) inicializar la bsPSSM con pesos al azar; (b) seleccionar al azar un peso w_{ij} obtenido en (a); (c) cambiar el valor del peso seleccionado para generar una psPSSM de ensayo que comprenda el peso seleccionado que tenga el valor cambiado; (d) calcular una puntuación para cada una de la pluralidad de motivos de secuencia funcional de los ARNip utilizando la bsPSSM de ensayo de acuerdo con la ecuación

$$\text{Puntuación} = \sum_{k=1}^{L} \ln(w_k / p_k)$$

en la que w_k y p_k son, respectivamente, pesos de un nucleótido en la posición k en el motivo de secuencia funcional y en una secuencia al azar; (e) calcular la correlación de la puntuación y una medida que caracterice a un ARNip entre la pluralidad de motivos de secuencia funcional de los ARNip; (f) repetir las etapas (c)-(e) para una pluralidad de diferentes valores del peso seleccionado en un intervalo determinado y conservar el valor que corresponda a la mejor correlación para el peso seleccionado; y (g) repetir las etapas (b)-(f) durante un número de veces elegido; determinando de este modo la psPSSM.

También se desvela un método para determinar una matriz de puntuación específica de posición de composición de bases (bsPSSM) $\{w_{ij}\}$ para representar un patrón de composición de bases que represente una pluralidad de motivos de secuencia funcional de ARNip diferentes de L nucleótidos, donde i = *G/C, A, U (T)* y j = *1, 2, ..., L*, y cada motivo de secuencia funcional de ARNip comprende al menos una parte de la secuencia diana del ARNip correspondiente y/o una secuencia en una región de secuencia que flanquea la secuencia diana de ARNip, comprendiendo el método (a) iniciar la bsPSSM con pesos al azar; (b) seleccionar al azar un peso w_{ij} obtenido en (a); (c) cambiar el valor del peso seleccionado para generar una psPSSM de ensayo que comprende el peso seleccionado que tiene el valor cambiado; (d) calcular una puntuación para cada uno de la pluralidad de motivos de secuencia funcional de ARNip utilizando la psPSSM de ensayo de acuerdo con la ecuación

$$\text{Puntuación} = \sum_{j=1}^{L} \ln(w_k / p_k)$$

en la que w_k y p_k son, respectivamente, pesos de un nucleótido en la posición k en el motivo de secuencia funcional y en una secuencia al azar; (e) calcular una correlación de la puntuación y una medida de una característica de un ARNip entre la pluralidad de motivos de secuencia funcional de los ARNip; (f) repetir las etapas (c)-(e) para una pluralidad de diferentes valores del peso seleccionado en un intervalo determinado y conservar el valor que corresponda a la mejor correlación para el peso seleccionado; y (g) repetir las etapas (b)-(f) durante un número de veces elegido; determinando de este modo la psPSSM.

En una realización, cada motivo funcional de ARNip comprende la secuencia diana del ARNip de direccionamiento correspondiente y una o las dos secuencias flanqueantes de la secuencia diana.

En otra realización, el método comprende adicionalmente seleccionar la pluralidad de motivos de secuencia funcional del ARNip mediante un método que comprende (i) identificar una pluralidad de ARNip que conste de los ARNip que tienen diferentes valores en la medida; (ii) identificar una pluralidad de cada uno de los motivos de secuencia funcional de ARNip correspondiente a un ARNip en la pluralidad de los ARNip.

En una realización, cada ARNip tiene M nucleótidos en su región dúplex, y cada motivo de secuencia funcional de ARNip consta de una secuencia diana de ARNip de M nucleótidos, una secuencia flanqueante en 5' de D_1 nucleótidos y una secuencia flanqueante en 3' de D_2 nucleótidos.

En una realización específica, cada ARNip tiene 19 nucleótidos en su región dúplex, y cada motivo de secuencia funcional de ARNip consta de una secuencia diana de ARNip de 19 nucleótidos, una secuencia flanqueante en 5' de 10 nucleótidos y una secuencia flanqueante en 3' de 10 nucleótidos. En otra realización específica, cada ARNip tiene 19 nucleótidos en su región dúplex, y cada motivo de secuencia funcional de ARNip consta de una secuencia diana de ARNip de 19 nucleótidos, una secuencia flanqueante en 5' de 50 nucleótidos y una secuencia flanqueante en 3' de 50 nucleótidos.

En una realización, la medida es la eficacia de silenciamiento.

En una realización, la pluralidad de cada uno de los N ARNip se dirige a un gen cuya abundancia de transcritos está dentro de un intervalo determinado. En una realización, el intervalo es de al menos aproximadamente 5, 10 o 100 transcritos por célula. En otra realización, el intervalo es menor que aproximadamente 3-5 transcritos por célula. En otra realización, el umbral es del 50 %, 75 % o 90 % a una dosis de ARNip de aproximadamente 100 nM.

En otra realización, el método comprende adicionalmente evaluar la psPSSM utilizando una curva ROC (siglas del inglés *receiver operating characteristic*) de la sensibilidad de la psPSSM frente a la no especificidad de la curva psPSSM, siendo la sensibilidad de la PSSM la proporción de positivos verdaderos detectada utilizando la psPSSM como una fracción de positivos verdaderos totales, y siendo la no especificidad de la PSSM la proporción de positivos falsos detectada utilizando la psPSSM como una fracción de positivos falsos totales.

En una realización, la pluralidad de motivos de secuencia funcional de ARNip consta de al menos 50, al menos 100, o

al menos 200 motivos de secuencia funcional de ARNip diferentes.

En otra realización adicional, el método comprende adicionalmente someter a ensayo la psPSSM utilizando otra pluralidad de motivos de secuencia funcional de ARNip.

También se desvela un método para determinar una matriz de puntuación específica de posición de coincidencia de posición (pmPSSM, siglas del inglés *position match position-specific score matrix*) *{E_i}* para representar un patrón de coincidencia de posición de un ARNip de *L* nucleótidos con su secuencia diana en un transcrito, donde E_i es una puntuación de una coincidencia en la posición *i*, *i = 1, 2, ..., L*, comprendiendo el método (a) identificar una pluralidad de *N* secuencias inespecíficas (*off-target*) de ARNip, donde cada secuencia inespecífica es una secuencia en la que el ARNip exhibe actividad de silenciamiento; (b) calcular una matriz de pesos de coincidencia de posición *{P_i}*, donde *i = 1,2, ..., L*, basándose en la pluralidad de *N* secuencias inespecíficas de ARNip de acuerdo con la ecuación

$$P_i = \frac{1}{N}\sum_{k=1}^{N}\delta_k(j),$$

en la que $\delta_k(j)$ es 1 si *k* es una coincidencia, y es 0 si *k* es un error de coincidencia; y (c) determinar la psPSSM calculando E_i de tal manera que $E_i = P_i$ si la posición *i* es una coincidencia y $E_i = (1-P_i)/3$ si la posición *i* es un error de coincidencia.

En una realización preferida, *L* = 19. En otra realización preferida, la matriz de peso de coincidencia de posición se indica en la Tabla I.

También se desvela un método para evaluar la actividad relativa de las dos cadenas de un ARNip en el silenciamiento de genes inespecíficos, que comprende comparar la composición de bases específicas de posición de la cadena en sentido del ARNip y la composición de bases específicas de posición de la cadena antisentido del ARNip o la cadena complementaria inversa de la cadena en sentido del ARNip, donde la cadena antisentido es la cadena guía para el direccionamiento de la secuencia diana deseada.

En una realización, la comparación se realiza mediante un método que comprende (a) determinar una puntuación para la cadena en sentido del ARNip, donde la puntuación se calcula utilizando una matriz de puntuación específica de posición; (b) determinar una puntación de la cadena antisentido del ARNip o la cadena complementaria inversa de la cadena en sentido del ARNip, utilizando la matriz de puntuación específica de posición; y (c) comparar la puntación de la cadena en sentido y la puntuación de la cadena antisentido o la cadena complementara inversa de la cadena en sentido, evaluando de este modo la preferencia de cadenas del ARNip.

En una realización, el ARNip tiene una secuencia de nucleótidos de *L* nucleótidos en su región dúplex, siendo *L* un número entero, donde la matriz de puntuación específica de posición es *{w_{ij}}*, donde w_{ij} es el peso del nucleótido *i* en la posición *j*, *i = G, C, A, U(T), j = 1, ..., L.*

En otra realización, el ARNip tiene una secuencia de nucleótidos de *L* nucleótidos en su región dúplex, siendo *L* un número entero, y la matriz de puntuación específica de posición es *{w_{ij}}*, donde w_{ij} es el peso del nucleótido *i* en la posición *j*, *i = G o C, A, U(T), j = 1, ..., L.*

En otra realización, la matriz de puntuación específica de posición se obtiene mediante un método que comprende (a) iniciar la matriz de puntuación específica de posición con pesos al azar; (b) seleccionar al azar un peso w_{ij} obtenido en (a); (c) cambiar el valor del peso seleccionado para generar una matriz de puntuación específica de posición de ensayo que comprende el peso seleccionado que tiene el valor cambiado; (d) calcular una puntuación para cada uno de una pluralidad de ARNip utilizando la matriz de puntación específica de posición de ensayo de acuerdo con la ecuación

$$\textit{Puntuación} = \sum_{j=1}^{L}\ln(w_j / p_j)$$

en la que w_j y p_j son respectivamente pesos de un nucleótido en la posición *j* en el ARNip y en una secuencia al azar; (e) calcular la correlación de la puntuación con una medida de una característica de un ARNip entre la pluralidad de los ARNip; (f) repetir las etapas (c)-(e) para una pluralidad de diferentes valores del peso seleccionado en un intervalo determinado y conservar el valor que corresponda a la mejor correlación para el peso seleccionado y (g) repetir las etapas (b)-(f) durante un número de veces elegido; determinando de este modo la matriz de puntación específica de posición.

En una realización, la medida es la eficacia de silenciamiento del ARNip.

En una realización, el ARNip tiene 19 nucleótidos en su región dúplex.

En otra realización, el ARNip tiene una secuencia de nucleótidos de nucleótidos de *L* nucleótidos en su región dúplex, siendo *L* un número entero, donde la matriz de puntuación específica de posición comprende w_k, $k = 1, ..., L$, siendo w_k una diferencia en la probabilidad de encontrar el nucleótido G o C en una posición de secuencia *k* entre un primer tipo de ARNip y un segundo tipo de ARNip, y la puntuación de cada cadena se calcula de acuerdo con la ecuación

$$Puntuación = \sum_{k=1}^{L} w_k \ .$$

En una realización, el primer tipo de ARNip consta de uno o más ARNip que tienen una eficacia de silenciamiento no menor que un primer umbral y el segundo tipo de ARNip consta de uno o más ARNip que tienen una eficacia de silenciamiento menor que un segundo umbral, y el ARNip se determina como que tiene preferencia antisentido si la puntuación determinada en la etapa (a) es mayor que la puntuación determinada en la etapa (b), o como que tiene preferencia en sentido si la puntuación determinada en la etapa (b) es mayor que la puntuación determinada en la etapa (a).

En otra realización, la diferencia en la probabilidad se describe mediante una suma de curvas gaussianas, representando cada una de las curvas gaussianas la diferencia en la probabilidad de encontrar una G o una C en una posición de secuencia diferente.

En una realización, el porcentaje del primer y segundo umbral es del 75 % a una dosis de ARNip de aproximadamente 100 nM.

También se desvela un sistema informático que comprende un procesador, y una memoria acoplada al procesador y que codifica uno o más programas, en el que el uno o más programas permiten que el procesador realice uno cualquiera de los métodos de la invención.

También se desvela un producto de programa informático para su uso junto con un ordenador que tiene un procesador y una memoria conectada al procesador, comprendiendo el producto de programa informático un medio de almacenamiento legible por ordenador que tiene un mecanismo de programa informático codificado en el mismo, en el que el mecanismo de programa informático puede cargarse en la memoria del ordenador y permitir que el ordenador lleve a cabo uno cualquiera de los métodos de la invención.

4. Breve descripción de las figuras

Las FIGS. 1A-C muestran que la composición de bases en, y alrededor de, una secuencia diana de ARNip afecta a la eficacia de silenciamiento del ARNip. Se ensayó un total de 377 ARNip mediante análisis Taqman con respecto a su capacidad para silenciar sus secuencias diana 24 h después de transfección en células HeLa. La mediana de silenciamiento diana fue de ~ 75 %. Este conjunto de datos se dividió en dos subconjuntos, uno que tenía una capacidad de silenciamiento menor que la mediana y otro que tenía una capacidad de silenciamiento igual a o mayor que la mediana (denominados ARNip “malos” y “buenos”, respectivamente). Aquí se muestra la diferencia de medias en una ventana de 5 (es decir, promediada sobre las 5 bases) en cuando al contenido en GC (FIG. 1A), contenido en A (FIG. 1B) y contenido en U (FIG. 1C) entre los ARNip buenos y malos en diferentes posiciones relativas en una secuencia diana.

FIGS. 2A-C (A) contenido en GC de ARNip buenos y malos; (B) contenido en A de ARNip buenos y malos; (C) contenido en U de ARNip buenos y malos. Las figuras muestran composiciones promedio de cada base. Por ejemplo, un valor de 0,5 en el eje y corresponde a un contenido de bases promedio de 50 %.

La FIG. 3 muestra el rendimiento de un modelo real de composición de bases de ARNip, utilizado en el método de diseño de ARNip de la invención. Los datos de eficacia de ARNip se subdividieron en dos pares de conjuntos de capacitación y ensayo. En cada uno de los conjuntos de capacitación se optimizaron las diferentes PSSM y se verificaron en los conjuntos de ensayo. El rendimiento de cada PSSM se evaluó con respecto a su capacidad para distinguir ARNip buenos (positivos verdaderos) y ARNip malos (positivos falsos) al seleccionar un número creciente de ARNip a partir de una lista clasificada por puntuación PSSM. Se muestran curvas ROC (por sus siglas del inglés *Receiver Operating Characteristics* que, en este contexto, se trata de curvas de eficacia diagnóstica) que demuestran el rendimiento de dos PSSM diferentes en sus respectivos conjuntos de capacitación y ensayo (líneas negras continuas y líneas grises discontinuas, respectivamente). El rendimiento esperado de la PSSM sobre datos aleatorizados se muestra por comparación (es decir, sin mejora en la capacidad de selección, línea de 45 º).

La FIG. 4 demuestra la capacidad predictiva de las PSSM en un conjunto de datos experimentales independientes. Se diseñaron nuevos ARNip para cinco genes por el método convencional como se describe en Elbashir *et al*, 2001, Nature 411:494-8, con la adición del método de predicción específico desvelado en esta solicitud y mediante el

método de predicción de especificidad y eficacia basado en PSSM de la invención. Se seleccionaron los tres mejores ARNip clasificados por gen para cada método y se adquirieron en Dharmacon. Los seis ARNip para cada uno de los cinco genes se ensayaron después para determinar su capacidad para silenciar sus secuencias diana. Se muestra un histograma del número de ARNip que silencian sus respectivos genes diana mediante una cantidad especificada. La curva continua representa el silenciamiento por ARNip diseñados por el método de la presente invención; la curva discontinua el representa el silenciamiento por ARNip diseñados por el método convencional y la curva gris de puntos representa el silenciamiento del conjunto de datos de los 377 ARNip.

Las FIGS. 5A-C muestran pesos medios de GC, A o U de los dos conjuntos de PSSM de composición de bases de capacitación y ensayo con los ARNip en el conjunto 1 y en el conjunto 2, respectivamente. La FIG. 5A representa pesos medios para GC, la FIG. 5B pesos medios para A y la FIG. 5C pesos medios para U. En la Tabla II se muestran los ARNip en el conjunto 1 y conjunto 2.

La FIG. 6 muestra un ejemplo de alineamientos de transcritos de genes inespecíficos en el núcleo de 19 meros de una secuencia oligonucleotídica de ARNip. Los genes inespecíficos se seleccionaron de la micromatriz Human 25k v2.2.1 seleccionando patrones cinéticos de abundancia de transcritos en consonancia con efectos directos de oligonucleótidos de ARNip. La columna de la izquierda enumera identificadores de secuencia de transcritos. Los alineamientos se generaron con el programa FASTA y se editaron a mano. Los recuadros y la zona gris demuestran el nivel más alto de similitud de secuencia en la mitad 3' del alineamiento.

La FIG. 7 muestra una matriz de puntuación específica de posición de coincidencia de posición para predecir efectos inespecíficos. La gráfica muestra el peso asociado con cada posición en una matriz que representa el alineamiento entre un oligonucleótido de ARNip y transcritos inespecíficos. El peso representa la probabilidad de que se observe una coincidencia en cada posición *i* a lo largo de un alineamiento entre un oligonucleótido de ARNip y un transcrito inespecífico observado.

La FIG. 8 muestra la optimización de la puntuación umbral para predecir efectos inespecíficos de los ARNip. Los valores de R^2 son el resultado de la correlación del número de alineamientos que puntúan por encima del umbral con el número de efectos inespecíficos observados.

La FIG. 9 muestra un flujograma de una realización a modo de ejemplo del método para seleccionar ARNip para su uso en el silenciamiento de un gen.

La FIG. 10 ilustra regiones de secuencia que pueden utilizarse para distinguir ARNip buenos y malos. Las PSSM se capacitaron en fragmentos de secuencia de más de 10 bases de longitud, desde 50 bases cadena arriba hasta 50 bases cadena abajo del ARNip de 19 meros y se ensayaron en conjuntos de ensayo independientes. El rendimiento de modelos capacitados en fragmentos de interés se comparó con el de modelos capacitados en secuencias aleatorias. La posición 1 corresponde a la primera base 5' en la región dúplex de un ARNip de 21 nt.

Las FIGS. 11A-B muestran modelos de curvas para PSSM. 11A: conjunto a modo de ejemplo de modelos de curvas para PSSM. 11B: rendimiento de los modelos en los conjuntos de capacitación y ensayo.

La FIG. 12 ilustra una realización a modo de ejemplo de un sistema informático útil para implementar los métodos de la presente invención.

La FIG. 13 muestra una comparación de la distribución de las eficacias de silenciamiento de los ARNip entre los 30 ARNip diseñados utilizando el método de la invención (círculos negros) y de los ARNip diseñados utilizando el método convencional (círculos blancos). Eje x: 1, KIF14; 2, PLK; 3, IGF1R; 4, MAPK14; 5, KIF11. Eje y: nivel de ARN. Los ARNip diseñados utilizando el método convencional para los 5 genes presentaron una amplia distribución de habilidades de silenciamiento, mientras que los diseñados con el método de la invención mostraron un silenciamiento más constante dentro de cada gen, así como a través de los genes. Para la genómica funcional con ARNip es muy importante una distribución estrecha.

Las FIGS. 14A-B muestran una comparación del contenido en GC de los ARNip y sus complementos inversos con el contenido en GC de los ARNip malos. Los resultados indican que los ARNip malos tienen cadenas en sentido similares a las de los ARNip buenos, mientras que los ARNip buenos tienen cadenas en sentido similares a las de los ARNip malos. CI: complemento inverso de la secuencia diana de ARNip.
La FIG. 15 muestra que los ARNip menos eficaces tienen cadenas en sentido activas. El sesgo de cadena de 61 ARNip se predijo a partir de perfiles de expresión mediante el método de sesgado en 3' y a partir de la comparación de las puntuaciones PSSM de GC de los ARNip y sus complementos inversos. Las predicciones del sesgo de cadena se categorizaron por eficacia de silenciamiento del ARNip.

La FIG. 16 muestra que la eficacia de silenciamiento se relaciona con el nivel de expresión del transcrito. Se ensayó un total de 222 ARNip (3 ARNip por gen para 74 genes) mediante análisis de ADN ramificado (ADNr) o Taqman, para determinar su capacidad para silenciar sus secuencias diana 24 h después de la transfección en células HeLa. El porcentaje de silenciamiento (eje y) se representó como una función de la abundancia de transcritos (eje x) medida

como intensidad en la micromatriz. Se muestra la mediana de silenciamiento diana observado para los 3 ARNip por gen seleccionado mediante el algoritmo de diseño de ARNip anterior. Se muestra la dependencia de silenciamiento sobre el nivel de expresión de genes, como el promedio de intensidades de 2 tipos de matrices, para 74 genes. Se utilizaron ensayos TaqMan para 8 genes. Se muestran datos del análisis de ADNr para los 66 genes restantes.

La FIG. 17 muestra que la eficacia de silenciamiento de un ARNip está relacionada con su composición de bases. Mediante análisis de ADNr se ensayaron ARNip para genes poco expresados para determinar su capacidad para silenciar sus secuencias diana. Los datos se dividieron en subconjuntos que tenían un silenciamiento menor del 75 % y un silenciamiento igual a o mayor del 75 % (ARNip malos y buenos, respectivamente). Aquí se muestra la diferencia en contenido de GC entre los ARNip buenos y malos (eje y) en cada posición en la cadena en sentido de ARNip (eje x). El conjunto de datos incluye genes poco expresados y muy expresados de 570 ARNip seleccionados en los 33 genes poco expresados y 41 muy expresados mediante las reglas de Tuschl o selección aleatorizada. Las secuencias de ARNip se enumeran en la Tabla IV. El perfil de GC para los ARNip buenos en genes poco expresados (curva de puntos grises) muestra algunas preferencias de composición similares a las de los ARNip malos para genes bien expresados (curva negra), aunque también muestra algunas diferencias.

La FIG. 18 muestra la eficacia de los ARNip recién diseñados. Se diseñaron ARNip para 18 genes mal expresados mediante el método convencional y mediante el nuevo algoritmo. Ruta convencional: selección de una puntuación pssm máxima; filtro minimax para coincidencias inespecíficas largas. Ruta mejorada: selección de 1-3 G+C en las bases 2-7 de 19 meros en sentido, asimetría bases 1 y 19, -300 <puntuación pssm <+200 y puntuaciones blast menores de 16.200 bases en cada lado del oligonucleótido de 19 meros no se repiten o son secuencias de baja complejidad. Para cada método se seleccionaron los tres mejores ARNip clasificados por gen. Los seis ARNip de cada uno de los cinco genes se ensayaron después para determinar su capacidad para silenciar secuencias diana. Se muestra un histograma del número de ARNip que silencian sus genes dianas mediante una cantidad especificada. La curva con puntos, silenciamiento por ARNip diseñados por el nuevo algoritmo; la curva negra, silenciamiento por ARNip diseñados por el método convencional. Mediana de silenciamiento mejorado del 60 % (algoritmo convencional) al 80 % (algoritmo nuevo).

FIG. 19. Características de diseño de ARNip eficaces. Los estudios de criterios de diseño que se correlacionan con la eficacia del silenciamiento de ARNip han revelado diversas características que predicen la eficacia. Estas incluyen una asimetría de bases en los dos extremos para dirigir la cadena antisentido (guía) en RISC, una U en la posición 10 para la escisión eficaz del transcrito, un tramo con bajo contenido en GC que abarca el centro y el extremo 3’ de la cadena guía para una escisión mejorada y la región “semilla” en el extremo 5’ de la cadena antisentido implicada en la unión del transcrito. Las líneas grises sobre la región dúplex indican preferencias de secuencia, las líneas grises claras debajo de la región dúplex indican atributos funcionales.

La FIG. 20 muestra expresión frente a mediana de silenciamiento en 371 ARNip. Estos son ARNip del conjunto de capacitación original de los 377 ARNip. En el análisis no se incluyeron 6 ARNip ya que el nivel de expresión de su gen diana no estaba disponible.

5. Descripción detallada de la invención

La presente divulgación proporciona un método para identificar motivos diana de ARNip en un transcrito utilizando una estrategia de matriz de puntuación específica de posición. La divulgación también proporciona un método para identificar genes inespecíficos de un ARNip y para predecir la especificidad de un ARNip utilizando una estrategia de matriz de puntuación específica de posición. La divulgación proporciona además un método para diseñar ARNip con una mayor eficacia y especificidad de silenciamiento. La divulgación también proporciona una biblioteca de ARNip que comprende ARNip con alta eficacia y especificidad de silenciamiento.

En esta solicitud, a menudo se dice que un ARNip se dirige a un gen. Se entenderá que cuando se hace una afirmación de este tipo, significa que el ARNip está diseñado para dirigirse y causar la degradación de un transcrito del gen. Dicho gen también se denomina gen diana del ARNip, y la secuencia en el transcrito sobre la que actúa el ARNip se denomina secuencia diana. Por ejemplo, una secuencia de 19 nucleótidos en un transcrito que es idéntica a la secuencia de la secuencia de 19 nucleótidos en la cadena de sentido de la región dúplex de un ARNip, es la secuencia diana del ARNip. La cadena antisentido del ARNip, es decir, la cadena que actúa sobre la secuencia diana, también se denomina cadena guía. En el ejemplo anterior, la cadena antisentido de la región dúplex de 19 nucleótidos del ARNip es la cadena guía. En esta solicitud, a menudo se hace referencia a las características de un ARNip con referencia a su secuencia, por ejemplo, la composición de la base posicional. Se entenderá que, a menos que se indique específicamente lo contrario, dicha referencia se hace a la secuencia de la cadena de sentido del ARNip. En esta solicitud, a menudo se describe un nucleótido o una secuencia de nucleótidos en un ARNip con referencia al extremo 5 'o 3' del ARNip. Se entenderá que cuando se emplea dicha descripción, se refiere al extremo 5 'o 3' de la cadena de sentido del ARNip. También se entenderá que, cuando se hace una referencia al extremo 3 'del ARNip, se refiere a la región dúplex 3' del ARNip, es decir, los dos nucleótidos del saliente 3' no se incluyen en la numeración de los nucleótidos. En la solicitud, a un ARNip también se le denomina oligo (oligonucleótido). En esta solicitud, se analiza el diseño de ARNip en referencia al silenciamiento de una diana de cadena de sentido, es decir, una secuencia diana de transcripción correspondiente a la cadena de sentido del ARNip. Un experto en la materia

entenderá que los métodos de la invención también son aplicables al diseño de ARNip para silenciar una diana antisentido (véase, por ejemplo, Martínez et al, 2002, Cell 110: 563-574).

5.1. MÉTODOS DE IDENTIFICACIÓN DE MOTIVOS DE SECUENCIA EN UN GEN PARA EL DIRECCIONAMIENTO MEDIANTE UN ARN DE INTERFERENCIA PEQUEÑO

La invención proporciona un método de identificación de motivos de secuencia en un transcrito que puede ser dirigido por un ARNip para la degradación del transcrito, por ejemplo, un motivo de secuencia que probablemente es un sitio de direccionamiento de ARNip muy eficaz. Dicho motivo de secuencia también se denomina motivo susceptible a ARNip. El método también puede utilizarse para identificar un motivo de secuencia en un transcrito que puede ser menos deseable para el direccionamiento por un ARNip, por ejemplo, un motivo de secuencia que probablemente es un sitio de direccionamiento de ARNip menos eficaz. Dicho motivo de secuencia también se denomina motivo resistente a ARNip.

En una realización, se identifican rasgos de secuencia característicos de un motivo de secuencia funcional, por ejemplo, un motivo de secuencia susceptible a ARNip y se construye un perfil del motivo funcional, por ejemplo, una biblioteca de ARNip en la cual se ha determinado la eficacia del silenciamiento.

En una realización, la región de secuencia de interés se explora para identificar secuencias que coinciden con el perfil del motivo funcional.

5.1.1. PERFIL DE SECUENCIAS Y EFICACIA DEL SILENCIAMIENTO DIANA.

En una realización preferida, el perfil de un motivo de secuencia funcional se representa utilizando una matriz de puntuación específica de posición (PPSM). Se puede encontrar un análisis general de una PPSM, por ejemplo, en "Biological Sequence Analysis" de R. Durbin, S. Eddy, A. Krogh y G. Mitchison, Cambridge Univ. Press, 1998; y de Henikoff y col., 1994, J Mol Biol. 243: 574-8. Una PPSM es un descriptor de motivo de secuencia que captura las características de un motivo de secuencia funcional. En esta divulgación, se utiliza una PPSM para describir motivos de secuencia de la invención, por ejemplo, un motivo susceptible o resistente. Una PPSM de un motivo susceptible (resistente) a ARNip también se denomina PPSM susceptible (resistente). Un experto en la materia sabrá que una matriz de puntuación específica de posición también recibe el nombre de matriz de puntuación de posición específica, matriz de peso de posición (PWM, *position weight matrix*) o perfil.

En la presente invención, un motivo funcional puede comprender una o más secuencias en una secuencia diana de ARNip. Por ejemplo, la una o más secuencias en una secuencia diana de ARNip puede ser una secuencia en el extremo 5' de la secuencia diana, una secuencia en el extremo 3' de la secuencia diana. La una o más secuencias en una secuencia diana de ARNip también pueden ser dos tramos de secuencias, una en el extremo 5' de la secuencia diana y una en el extremo 3' de la secuencia diana. Un motivo funcional también puede comprender una o más secuencias en una región de secuencia que flanquea la secuencia diana de ARNip. Dicha una o más secuencias pueden estar directamente adyacentes a la secuencia diana de ARNip. Dicha una o más secuencias también pueden separarse de la secuencia diana de ARNip mediante una secuencia intermedia. La figura 10 ilustra algunos ejemplos de motivos funcionales.

En una realización, un motivo de secuencia funcional, por ejemplo, un motivo de secuencia susceptible o resistente, comprende al menos una porción de una secuencia dirigida por un ARNip. En una realización, el motivo funcional comprende un tramo contiguo de al menos 7 nucleótidos de la secuencia diana. En una realización preferida, el tramo contiguo está en una región 3' de la secuencia diana que comienza, por ejemplo, en las 3 bases en el extremo 3'. En otra realización, el tramo contiguo está en una región 5' de la secuencia diana. En otra realización, el motivo funcional comprende un tramo contiguo de al menos 3, 4, 5, 6 o 7 nucleótidos en una región 3' de la secuencia diana y comprende un tramo contiguo de al menos 3, 4, 5, 6, o 7 nucleótidos en una región 5' de la secuencia diana. En otra realización más, el motivo funcional comprende un tramo contiguo de al menos 11 nucleótidos en una región central de la secuencia diana. Los motivos de secuencia que comprenden una longitud menor que la longitud completa de la secuencia diana de ARNip pueden utilizarse para evaluar transcritos diana de ARNip que exhiben solo una secuencia parcial identificada en un ARNip (solicitud internacional Nº PCT / US2004 / 015439 de Jackson et al., presentada el 17 de mayo de 2004). En una realización preferida, el motivo funcional comprende la secuencia diana de ARNip de longitud completa.

El motivo funcional también puede comprender una secuencia flanqueante. Los inventores han descubierto que la secuencia de dicha región flanqueante juega un papel en la determinación de la eficacia del silenciamiento. En una realización, un motivo de secuencia funcional, por ejemplo, un motivo de secuencia susceptible o resistente, comprende al menos una parte de una secuencia dirigida por un ARNip y una o más secuencias en una o ambas regiones flanqueantes. Por lo tanto, un motivo de secuencia puede incluir una secuencia diana de ARNip de M nucleótidos, una secuencia flanqueante de D_1 nucleótidos en un lado de la secuencia diana de ARNip y una secuencia flanqueante de D_2 nucleótidos en el otro lado de la secuencia diana de ARNip donde M, D_1 y D_2 son números enteros apropiados. En una realización, $D_1 = D_2 = D$. En una realización, M = 19. En algunas realizaciones preferidas, D_1, D_2 o D tiene una longitud de al menos 5, 10, 20, 30, 50 nucleótidos. En una realización específica, un

motivo de secuencia susceptible o resistente consta de una secuencia diana de ARNip de 19 nucleótidos y una secuencia flanqueante de 10 nucleótidos en cualquier lado de la secuencia diana de ARNip. En otra realización específica, un motivo de secuencia susceptible o resistente consta de una secuencia diana de ARNip de 19 nucleótidos y una secuencia flanqueante de 50 nucleótidos en cualquier lado de la secuencia diana de ARNip.

En otra realización, un motivo de secuencia puede incluir una secuencia diana de ARNip de *M* nucleótidos y uno o más de lo siguiente: un tramo contiguo de D_1 nucleótidos que flanquean el extremo 5' de la secuencia diana, un tramo contiguo de D_2 nucleótidos que flanquea el extremo 3' de la secuencia diana, un tramo contiguo de D_3 nucleótidos que comienza aproximadamente a 35 nucleótidos cadena arriba del extremo 5 'de la secuencia diana, un tramo contiguo de D_4 nucleótidos que comienza aproximadamente a 25 nucleótidos aguas abajo del extremo 3' de la secuencia diana y un tramo contiguo de D_5 nucleótidos que comienza aproximadamente a 60 nucleótidos aguas abajo del extremo 3' de la secuencia diana, donde D_1, D_2, D_3 D_4 y D_5 son números enteros apropiados. En una realización, $D_1 = D_2 = D$. En algunas realizaciones preferidas, cada uno de D_1, D_2, D_3 D_4 y D_5 tiene una longitud de al menos 5, 10 o 20 nucleótidos. La longitud del motivo funcional es $L = M + D1 + D_2 + D_3 + D_4 + D_5$. En una realización específica, el motivo de secuencia incluye una secuencia diana de ARNip de 19 nucleótidos, un tramo contiguo de aproximadamente 10 nucleótidos que flanquea el extremo 5' de la secuencia diana, un tramo contiguo de aproximadamente 10 nucleótidos que flanquea el extremo 3' de la secuencia diana, un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 35 nucleótidos cadena arriba del extremo 5' de la secuencia diana, un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 25 nucleótidos cadena abajo del extremo 3' de la secuencia diana, y un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 60 nucleótidos cadena abajo del extremo 3' de la secuencia diana (véase la figura 10).

En otras realizaciones, un motivo de secuencia funcional, por ejemplo, un motivo de secuencia susceptible o resistente, comprende una o más secuencias en una o ambas regiones flanqueantes de una secuencia diana de ARNip pero no comprende ninguna secuencia diana de ARNip. En una realización, el motivo funcional comprende un tramo contiguo de aproximadamente 10 nucleótidos que flanquea el extremo 5' de la secuencia diana. En otra realización, el motivo funcional comprende un tramo contiguo de aproximadamente 10 nucleótidos que flanquea el extremo 3' de la secuencia diana. En una realización preferida, el motivo funcional comprende un tramo contiguo de aproximadamente 10 nucleótidos que flanquea el extremo 5' de la secuencia diana y un tramo contiguo de aproximadamente 10 nucleótidos que flanquea el extremo 3' de la secuencia diana. En una realización, el motivo funcional comprende un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 35 nucleótidos cadena arriba del extremo 5' de la secuencia diana. En otra realización, el motivo funcional comprende un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 25 nucleótidos cadena abajo del extremo 3' de la secuencia diana. En otra realización más, el motivo funcional comprende un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 60 nucleótidos cadena abajo del extremo 3' de la secuencia diana. En una realización preferida, el motivo funcional comprende un tramo contiguo de aproximadamente 10 nucleótidos que flanquea el extremo 5' de la secuencia diana, un tramo contiguo de aproximadamente 10 nucleótidos que flanquea el extremo 3' de la secuencia diana, un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 35 nucleótidos cadena arriba del extremo 5' de la secuencia diana, un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 25 nucleótidos cadena abajo del extremo 3' de la secuencia diana, y un tramo contiguo de aproximadamente 10 nucleótidos que comienza aproximadamente a 60 nucleótidos cadena abajo del extremo 3' de la secuencia diana. Por lo tanto, un motivo de secuencia puede incluir un tramo contiguo de D_1 nucleótidos que flanquea el extremo 5' de la secuencia diana, un tramo contiguo de D_2 nucleótidos que flanquea el extremo 3' de la secuencia diana, un tramo contiguo de D*3* nucleótidos que comienza aproximadamente A 35 nucleótidos cadena arriba del extremo 5' de la secuencia diana, un tramo contiguo de D_4 nucleótidos que comienza aproximadamente a 25 nucleótidos cadena abajo del extremo 3' de la secuencia diana, y un tramo contiguo de D_5 nucleótidos que comienza aproximadamente a 60 nucleótidos cadena abajo del extremo 3' de la secuencia diana, donde D_1, D_2, D_3 D_4 y D_5 son números enteros apropiados. En algunas realizaciones preferidas, cada uno de D_1, D_2, D_3 D_4 y D_5 tienen una longitud de al menos 5, 10 o 20 nucleótidos. La longitud del motivo funcional es $L = D1 + D_2 + D_3 + D_4 + D_5$.

En una realización, las características de un motivo de secuencia funcional se caracterizan utilizando la frecuencia de cada uno de G, C, A, U (o T) observada en cada posición a lo largo del motivo de secuencia. En la descripción, U (o T), o a veces simplemente U (T), se utiliza para indicar el nucleótido U o T. El conjunto de frecuencias forma una matriz de frecuencia, en la que cada elemento indica el número de veces que se ha observado un nucleótido determinado en una posición determinada. Una matriz de frecuencia que representa un motivo de secuencia de longitud *L* es una matriz de 4 • L $\{f_{ij}\}$, donde i = G, C, A, U(T); j = 1, 2, ..., L; por lo que f_{ij} es la frecuencia del *i-gésimo* nucleótido en la *j-ésima* posición. Una matriz de frecuencia de un motivo de secuencia se puede obtener o construir a partir de un conjunto de *N* secuencias diana de ARNip que exhibe una calidad deseada, por ejemplo, un nivel elegido de susceptibilidad o resistencia al silenciamiento de ARNip.

$$f_{ij} = \sum_{k=1}^{N} \delta_{ik}(j) \qquad (1)$$

en la que

$$\delta_{ik}(j) = \begin{cases} 1, \text{si } k = i \\ 0, \text{si } k \neq i \end{cases} \quad (2)$$

En realizaciones en las que un motivo de secuencia funcional consta de *M* nucleótidos en la secuencia diana de ARNip, una secuencia flanqueante de D_1 nucleótidos en un lado de la secuencia diana de ARNip y una secuencia flanqueante de D_2 nucleótidos en el otro lado de la secuencia diana de ARNip, $L = M + D_1 + D_2$. En realizaciones en las que el motivo funcional consta de *M* nucleótidos en la secuencia diana de ARNip, un tramo contiguo de D_1 nucleótidos que flanquea el extremo 5' de la secuencia diana, un tramo contiguo de D_2 nucleótidos que flanquea el extremo 3' de la secuencia diana, un tramo contiguo de D_3 nucleótidos que comienza aproximadamente a 35 nucleótidos cadena arriba del extremo 5' de la secuencia diana, un tramo contiguo de D_4 nucleótidos que comienza aproximadamente a 25 nucleótidos cadena abajo del extremo 3' de la secuencia diana, y un tramo contiguo de D_5 nucleótidos que comienza aproximadamente a 60 nucleótidos cadena abajo del extremo 3' de la secuencia diana, $L = D1 + D_2 + D_3 + D_4 + D_5$.

En otra realización, las características de un motivo de secuencia funcional se caracterizan utilizando un conjunto de pesos, uno para cada nucleótido que aparece en el motivo en una posición. En dicha realización, se puede utilizar una matriz de peso $\{e_{ij}\}$, donde *i* = G, C, A, U (T); *j* = 1, 2, ..., L, para representar un motivo de secuencia funcional de longitud *L*, donde e_{ij}, es el peso de encontrar el *i-gésimo* nucleótido en la *j-ésima* posición. En una realización, el peso e_{ij} es la probabilidad de encontrar el *i-gésimo* nucleótido en la *j-ésima* posición en el motivo de secuencia funcional. Cuando se utiliza una probabilidad para el peso, la matriz también se denomina matriz de probabilidad. Una matriz de probabilidad de un motivo de secuencia puede derivarse de una matriz de frecuencia de acuerdo con la ecuación

$$e_{ij} = \frac{f_{ij}}{N} \quad (3)$$

En una realización preferida, para caracterizar un motivo de secuencia funcional se utiliza una matriz de puntuación específica de posición (PSSM). La PSSM puede construirse utilizando los valores de verosimilitud logarítmica $\log(e_{ij}/p_{ij})$, donde e_{ij} es el peso de encontrar el nucleótido *i* en la posición *j,* y p_{ij} es el peso de encontrar el nucleótido *i* en la posición *j* en un secuencia aleatoria. En algunas realizaciones, la probabilidad de encontrar el *i-gésimo* nucleótido en la *j-ésima* posición en el motivo de secuencia funcional se utiliza como e_{ij}, la probabilidad de encontrar el nucleótido *i* en la posición *j* en una secuencia aleatoria se utiliza como p_{ij}. El peso o la probabilidad p_{ij} es un peso o una probabilidad "*a priori*". En algunas realizaciones, p_{ij} = 0,25 para cada posible nucleótido i $\in$ {G, C, A, U(T)} en cada posición *j*. Por lo tanto, para una secuencia dada de longitud *L*, la suma de cocientes de verosimilitud logarítmica en todas las posiciones puede utilizarse como puntuación para evaluar si es más o menos probable que la secuencia dada coincida con el motivo funcional en lugar de coincidir con una secuencia aleatoria:

$$Puntuación = \sum_{j=1}^{L} \ln(e_j / p_j) \quad (4)$$

en la que w_j y p_j son, respectivamente, pesos de un nucleótido en la posición *j* en el motivo de secuencia funcional y en una secuencia aleatoria. Por ejemplo, si dicha puntación es cero, la secuencia tiene la misma probabilidad de coincidir con el motivo de la secuencia que la de coincidir con una secuencia aleatoria. Es más probable que una secuencia coincida con el motivo de la secuencia si la relación es mayor que cero.

En otra realización, cuando no se van a distinguir dos o más nucleótidos diferentes, puede utilizarse una PSSM con una dimensión reducida. Por ejemplo, si las composiciones de bases relativas de G y C en un motivo de secuencia no se van a distinguir, una PSSM puede ser una matriz 3 • L $\{\log(E_{ij}/p_{ij}\}$, donde $i = G/C, A, U(T); j = 1, 2, ..., L$; donde E_{ij} es el peso, por ejemplo, la probabilidad, de encontrar el nucleótido *i* en la posición *j*, y p_{ij} es el peso, por ejemplo, la probabilidad, de encontrar el nucleótido *i* en la posición *j* en una secuencia aleatoria. Por lo tanto, en dichos casos, una PSSM tiene 3 conjuntos de pesos: específico de GC, específico de A y específico de U, por ejemplo, si la base en una posición es una G o una C, el logaritmo natural de la relación del peso GC y la probabilidad imparcial de encontrar una G o C en esa posición se utiliza como el peso específico de GC para la posición; y los logaritmos naturales de los pesos de A y T específicos de posición divididos entre la probabilidad imparcial de la base respectiva se utilizan como los pesos específicos de A y T para la posición, respectivamente. La puntación del cociente de verosimilitud logarítmica se representa mediante la Ecuación (5):

$$Puntuación = \sum_{j=1}^{L} \ln(E_j / p_j) \quad (5)$$

en la que E_j es el peso asignado a una base - A, U o G/C - en la posición *j*, y p_j = 0,25 para A o U y 0,5 para G/C.

En otra realización más, cuando las composiciones de bases relativas de G y C en un motivo de secuencia no se van a distinguir y las composiciones de bases relativas de A y T en el motivo de secuencia tampoco se van distinguir, una PSSM puede ser una matriz 1 • *L* {log(E_{ij}/p_{ij}}, donde *i* = G/C; *j = 1, 2, ..., L*; donde E_{ij} es el peso, por ejemplo, la probabilidad de encontrar el nucleótido i en la posición j, y p_{ij} es el peso, por ejemplo, la probabilidad de encontrar el nucleótido *i* en la posición *j* en una secuencia aleatoria. Por lo tanto, en dichos casos, una PSSM tiene 1 conjunto de pesos específicos de GC: si la base en una posición es una G o una C, el logaritmo natural de la relación de peso de GC y la probabilidad imparcial de encontrar una G o C en esa posición se utiliza como el peso específico de GC para la posición. La puntación del cociente de verosimilitud logarítmica se representa mediante la Ecuación (5), excepto que E_j es el peso asignado a una base - G/C - en la posición *j* y p_j = 0,50.

5.1.2. MÉTODOS DE DETERMINACIÓN DE UN PERFIL

La divulgación proporciona métodos de determinación de una PSSM de un motivo de secuencia funcional, basados en una pluralidad de ARNip para los cuales se ha determinado alguna cantidad o cantidades que caracterizan los ARNip. Por ejemplo, para determinar una PSSM de un motivo de secuencia susceptible o resistente a ARNip, puede utilizarse una pluralidad de ARNip, cuya eficacia de silenciamiento se ha determinado. En la divulgación, para simplificar, a menudo se utiliza la eficacia como una medida de clasificación der los ARNip. La eficacia de un ARNip se mide en ausencia de otros ARNip diseñados para silenciar el gen diana. Será obvio para un experto en la materia que los métodos de la invención son igualmente aplicables en casos en que los ARNip se clasifican basándose en otra medida. Dicha pluralidad de ARNip también se denomina biblioteca de ARNip. En casos en los que el motivo de secuencia funcional de interés comprende una o más secuencias en una o en las dos regiones flanqueantes, para determinar la PSSM del motivo funcional, puede utilizarse una pluralidad de motivos funcionales de ARNip, es decir, una secuencia que comprenda la secuencia diana de ARNip y las secuencias en la región o regiones flanqueante(s) en un transcrito. En una realización preferida, el motivo de secuencia funcional del ARNip consta de una secuencia diana de ARNip de 19 nucleótidos y de una secuencia flanqueante de 10 nucleótidos a cada lado de la secuencia diana de ARNip. Para simplificar, en esta divulgación, a menos que se especifique, la expresión "una biblioteca de ARNip" se utiliza a menudo para referirse tanto a una biblioteca de ARNip como a una biblioteca de motivos funcionales de ARNip. Se entenderá que, en los últimos casos, cuando se hace referencia a la eficacia de un ARNip, esto se refiere a la eficacia del ARNip que se dirige al motivo. Preferentemente, la pluralidad ARNip o motivos diana de ARNip, comprende al menos 10, 50, 100, 200, 500, 1000 o 10.000 ARNip o motivos diana de ARNip diferentes.

Cada ARNip diferente en la pluralidad o biblioteca de ARNip o motivos funcionales de ARNip, puede tener un nivel de eficacia diferente. En una realización, la pluralidad o biblioteca de ARNip consta de ARNip que tienen un nivel de eficacia elegido. En otra realización, la pluralidad o biblioteca de ARNip comprende ARNip que tienen niveles de eficacia diferentes. En dicha realización, los ARNip pueden agruparse en subconjuntos, constando cada uno de ellos de ARNip que tienen un nivel de eficacia elegido.

En una realización, se determina una PSSM de un motivo funcional de ARNip utilizando una pluralidad de ARNip que tienen una eficacia dada. En una realización, para determinar una PSSM de un motivo susceptible a ARNip, se utiliza una pluralidad de *N* ARNip que consta de ARNip que tienen una eficacia de silenciamiento por encima de un umbral elegido. La PSSM se determina basándose en la frecuencia de un nucleótido aparecido en una posición (véase la Sección 5.1.1). El umbral elegido puede ser de 50 %, 75 %, 80 % o 90 %. En otra realización, para determinar una PSSM de un motivo susceptible a ARNip, se utiliza una pluralidad de *N* ARNip que consta de ARNip que tienen una eficacia de silenciamiento por debajo de un umbral elegido. El umbral elegido puede ser de 5 %, 10 %, 20 %, 50 %, 75 % o 90 %. En una realización preferida, la PSSM tiene una dimensión reducida con un peso para G/C.

En realizaciones preferidas, una PSSM de un motivo susceptible o resistente se obtiene o construye utilizando una estrategia clasificadora con un conjunto de *N* secuencias. En dichas realizaciones, se utiliza una biblioteca de ARNip que comprende ARNip que tienen diferentes niveles de eficacia. En una realización, los ARNip de la biblioteca se pueden agrupar aleatoriamente en subconjuntos, constando cada uno de ellos de ARNip que tienen diferentes niveles de eficacia, un subconjunto se utiliza como un conjunto de capacitación para determinar una PSSM y el otro se utiliza como un conjunto de ensayo para validar la PSSM. Para dividir la biblioteca de ARNip existente en conjuntos de capacitación y de ensayo pueden utilizarse diferentes criterios. Para una biblioteca de ARNip, en la que una mayoría de los oligos de ARNip están diseñados con el método convencional, que requiere un dímero de AA inmediatamente antes de la secuencia oligonucleotídica de 19 meros, se utilizaron varias particiones y se combinaron más de una PSSM capacitada (en lugar de PSSM individuales) para asignar puntuaciones a los oligonucleótidos de ensayo. En la Tabla II se muestra una biblioteca de ARNip a modo de ejemplo y divisiones de la biblioteca en conjuntos de capacitación y ensayo.

En una realización preferida, el motivo de secuencia consta de 39 bases en la secuencia de transcripción, comenzando en 10 bases cadena arriba de la secuencia diana de ARNip de 19 meros y terminando en 10 bases cadena abajo de la secuencia de 19 meros. En la Sección 5.1.1., se describe la PSSM que caracteriza dicho motivo de secuencia.

En una realización preferida, la PSSM se determina mediante un proceso iterativo. Una PSSM se inicializa con

pesos aleatorios $\{e_{ii}\}$ o $\{E_{ij}\}$ dentro de un intervalo de búsqueda determinado para todas las bases en todas las posiciones. En otra realización preferida, la PSSM se inicializa a la diferencia media de composición de bases suavizada entre los ARNip buenos y malos en el conjunto de capacitación. Como ejemplo, una PSSM que describe un motivo de secuencia de 39 nucleótidos puede tener 117 elementos. En otra realización, los pesos se optimizan comparando la correlación de puntuaciones generadas con una cantidad de interés, por ejemplo, eficacia de silenciamiento, y seleccionando la PSSM cuya mejor puntuación se corresponda mejor con esa cantidad. La mejora en el rendimiento de la PSSM se puntúa comparando los valores de correlación antes y después de un cambio en los pesos en cualquier posición. En una realización, no hay ningún requisito mínimo para un cambio en la correlación. La mejora agregada se calcula como la diferencia entre la correlación final y la correlación inicial. En una realización, para una PSSM que caracteriza un motivo de secuencia de 39 meros, el umbral de mejora agregada después de 117 ciclos para la terminación de la optimización es una diferencia de 0,01.

En una realización, los pesos se optimizan para reflejar diferencias de composición de bases entre ARNip buenos, es decir, ARNip que tienen al menos una eficacia mediana, y ARNip malos, es decir, ARNip que tienen una eficacia inferior a la mediana, en el intervalo de valores permitidos para pesos. Si la PSSM se inicializa con una matriz de frecuencia, el intervalo de valores permitidos se corresponde con los elementos de matriz de frecuencia +/- 0,05. Si se utiliza una búsqueda imparcial, los intervalos de los valores permitidos para los pesos son de 0,45 a 0,55 para G/C y de 0,2 a 0,3 para A o U. En una realización, se permite que los pesos varíen de los valores iniciales en +/- 0,05. Si se utiliza una búsqueda imparcial, los pesos de la PSSM pueden ajustarse a valores iniciales aleatorios dentro del intervalo de búsqueda imparcial descrito anteriormente.

En una realización, la PSSM se determina mediante un procedimiento de optimización de mutación aleatoria en escalada (*hill-climbing*). En cada etapa del proceso, para la optimización, se selecciona una base al azar en una posición. Por ejemplo, para una PSSM que describe un motivo de secuencia de 39 nucleótidos, las 39 bases se convierten en un vector de 117 pesos: 39 pesos de G/C, 39 pesos de A y 39 pesos de U. En cada etapa, para la optimización se selecciona uno de estos 117 pesos y se ejecuta a través de todos los valores en el intervalo de búsqueda en esa etapa. Para cada valor en el intervalo de búsqueda, se calculan puntuaciones para un conjunto de capacitación de ARNip. Después, se calcula la correlación de estas puntuaciones con la eficacia de silenciamiento de los ARNip. El peso para esa posición que genera la mejor correlación entre las puntaciones y la eficacia del silenciamiento se conserva como el nuevo peso en esa posición.

En una realización, la medida utilizada para medir la efectividad de la capacitación y ensayo es la tasa de detección falsa (FDR, *false detection rate*) agregada, basada en la curva ROC, y se calcula como el promedio de las puntuaciones FDR de los mejores oligos de 33 % ordenadas por las puntuaciones dadas por la PSSM capacitada. Al calcular las puntuaciones FDR, los oligos con niveles de silenciamiento más bajos que la mediana, se consideran falsos, y los oligos con niveles de silenciamiento más altos que la mediana se consideran verdaderos. La "tasa de detección falsa" es el número de positivos falsos seleccionados dividido entre el número total de positivos verdaderos, medida en cada posición clasificada en una lista. La tasa de detección falsa puede ser una función de la fracción de todos los ARNip seleccionados. En una realización, el área bajo la curva al 33 % de la lista seleccionada como un solo número representa el rendimiento. En una realización, todos los ARNip al menos como la mediana se denominan "positivos" y todos los ARNip peores que la mediana se denominan "negativos". Por lo tanto, la mitad de los datos son positivos y la otra mitad son "positivos falsos". En una clasificación ideal, el área bajo la curva al 33 % o incluso al 50 % de la lista seleccionada debe ser 0. Por el contrario, una clasificación aleatoria haría que se seleccionaran los mismos números de positivos verdaderos y positivos falsos. Esto corresponde a un área bajo la curva del 0,17 al 33 % de la lista seleccionada, o del 0,25 al 50 % de la lista seleccionada.

Las correlaciones entre el % de silenciamiento y la puntuación de PSSM se calculan de acuerdo con el método conocido en la técnica (véase, por ejemplo, *Applied Multivariate Statistical Analysis*, 4ª edición, RA Johnson & EW Wichern, Prentice-hall, 1998)

El proceso continuó hasta que la mejora agregada en una pluralidad de iteraciones cayó por debajo de un umbral.

En una realización preferida, utilizando un conjunto de capacitación de ARNip, se obtiene una pluralidad de PSSM para un motivo de secuencia funcional. En esta divulgación, una pluralidad de PSSM también se denomina un "conjunto" de PSSM. Cada ronda de optimización puede detenerse en un óptimo local distinto al del óptimo global. El óptimo local particular alcanzado depende del historial de posiciones aleatorias seleccionadas para la optimización. Un umbral de mejora más alto puede no llevar una PSSM optimizada a un óptimo local más cercano al óptimo global. Por lo tanto, es más eficaz ejecutar optimizaciones múltiples que una optimización larga. Se descubrió que ejecuciones adicionales (por ejemplo, hasta 200) mejoraban el rendimiento. Se observó que, ejecutando más de 200 optimizaciones no se proporcionaban mejoras adicionales en el rendimiento. Empíricamente, la puntuación de los ARNip mediante el promedio de múltiples ejecuciones es menos eficaz que puntuar los ARNip candidatos en las PSSM generadas por cada ejecución y después sumar las puntaciones. Por lo tanto, en una realización, la pluralidad de PSSM se utiliza individualmente o se suma para generar una puntuación compuesta para cada coincidencia de secuencia. La pluralidad de matrices puede ensayarse individualmente o como un compuesto en un conjunto independiente de motivos diana de ARNip con eficacia de silenciamiento conocida para evaluar la utilidad para identificar motivos de secuencia y en el diseño de ARNip. En una realización preferida, la pluralidad de PSSM consta

de al menos 2, 10, 50, 100, 200 o 500 PSSM.

En una realización preferida, para obtener uno o más conjuntos de PSSM, se utiliza uno o más conjuntos de capacitación de ARNip diferentes. Estos diferentes conjuntos de PSSM pueden utilizarse conjuntamente en la determinación de la puntuación de un motivo de secuencia.

Los métodos de ponderación de secuencia se han utilizado en la técnica para reducir la redundancia y enfatizar la diversidad en aplicaciones de búsqueda y alineamiento de secuencias múltiples. Cada uno de estos métodos se basa en una idea de distancia entre una secuencia y una secuencia ancestral o generalizada. En el presente documento se presenta una estrategia diferente, en la que los pesos se basan en la diversidad observada en cada posición en el alineamiento y en la correlación entre la composición de bases y la eficacia observada de los ARNip, en lugar de en una medida de distancia de secuencia.

En otra realización más, las PSSM se generan mediante un método que crea la hipótesis de la dependencia de la composición de bases de cualquier posición en sus posiciones adyacentes, denominado "modelos de curvas".

En una realización, los modelos de curvas se generan como una suma de curvas normales (es decir, Gaussianas). Será obvio para un experto en la técnica que también puedan utilizarse otras funciones de curva adecuadas, por ejemplo, polinomios. Cada curva representa la probabilidad de encontrar una base particular en una región particular. El valor en cada posición en las curvas normales sumadas es el peso dado a esa posición para la base representada por la curva. Después, los pesos para cada base presente en cada posición en cada ARNip y sus secuencias flanqueantes, se suman para generar una puntuación de ARNip, es decir, la puntuación es Σ wi. El cálculo de la puntación también se puede describir como el producto puntual del contenido de bases en la secuencia con los pesos en el modelo de curva. Como tal, esta es una forma de representar la correlación de la secuencia de interés con el modelo.

Los modelos de curva pueden inicializarse para corresponder a los principales picos y valles presentes en la diferencia de composición de bases suavizada entre los ARNip buenos y malos, por ejemplo, como se describe en las Figs. 1A- C y 5A- C. En una realización, se obtienen modelos de curvas para G/C, A y U. En una realización, el modelo inicial se puede configurar para el modelo de curva G/C de 3 picos de la siguiente manera:

Pico 1

media: 1,5

desviación típica: 2

amplitud: 0,0455

La media, la desviación típica y la amplitud del pico 1, se configuran para corresponderse con el pico en la diferencia media en el contenido de GC entre los ARNip buenos y malos que aparecen en las bases 2 - 5 del sitio diana del ARNip en el conjunto 1 de ensayos de capacitación y de ensayo.

Pico 2

media: 11

desviación típica: 0,5

amplitud: 0,0337

La media, la desviación típica y la amplitud del pico 2 se configuran para corresponderse con el pico en la diferencia de medias en el contenido de GC entre los ARNip buenos y malos que aparecen en las bases 10-12 del sitio diana del ARNip en el conjunto 1 de ensayos de capacitación y de ensayo.

Pico 3

media: 18,5

desviación típica: 4

amplitud:-0,0548

La media, la desviación típica y la amplitud del pico 3 se configuran para corresponderse con el pico en la diferencia de medias en el contenido de GC entre los ARNip buenos y malos que aparecen en las bases 12-25 del sitio diana del ARNip en el conjunto 1 de ensayos de capacitación y de ensayo.

En un modelo de curva, se puede ajustar la altura máxima (amplitud), la posición central en la secuencia (media) y la anchura (desviación típica) de un pico. Los modelos de curvas se optimizan ajustando la amplitud, la media y la

desviación típica de cada pico sobre una cuadrícula de valores preestablecida. En una realización, los modelos de curvas se optimizan en varios conjuntos de capacitación y se ensayan en varios conjuntos de ensayo, por ejemplo, conjuntos de capacitación y conjuntos de ensayo como se describe en la Tabla II. Cada base, -G/C, A y U(o T), se optimiza por separado, y después se seleccionan combinaciones de modelos optimizados para obtener el mejor rendimiento.

Preferentemente, los criterios de optimización para los modelos de curvas son: (1) la fracción de oligos buenos en el 10 %, 15 %, 20 % y 33 % superior de las puntaciones, (2) la tasa de detección falsa en el 33 % y 50 % de los ARNip seleccionados, y (3) el coeficiente de correlación de silenciamiento de ARNip frente a puntuaciones de ARNip como una prueba decisiva.

Cuando el modelo se capacita, se explora una cuadrícula de valores posibles para la amplitud, la media y la desviación típica de cada pico. Adicionalmente se seleccionaron y examinaron los modelos con el valor superior o dentro del intervalo superior de valores de cualquiera de los criterios anteriores.

En una realización preferida, los modelos G/C se optimizan con 3 o 4 picos, los modelos A se optimizan con 3 picos, y los modelos U se optimizan con 5 picos. Los intervalos a modo de ejemplo de parámetros optimizados para modelos de curvas se muestran más adelante en el Ejemplo 3.

Preferentemente, se evalúa el rendimiento de la PSSM obtenida. En una realización, la PSSM se evalúa utilizando una curva ROC (*receiver operating characteristic*). Una curva ROC es un gráfico de la sensibilidad de una prueba de diagnóstico en función de la no especificidad. Una curva ROC indica las propiedades intrínsecas del rendimiento diagnóstico de una prueba y puede utilizarse para comparar ventajas relativas de procedimientos alternativos. En una realización, la sensibilidad de una PSSM se calcula como la proporción de positivos verdaderos detectados como una fracción del total de positivos verdaderos, mientras que la no especificidad de la PSSM se calcula como la proporción de positivos falsos detectados como una fracción del total de positivos falsos (véase, por ejemplo, G. Chambell, 1994, Statistics in Medicine 13: 499 - 508; Metz, 1986, Investigative Radiology 21: 720 - 733; Gribskov et al., 1996, Computers Chem. 20: 25 - 33). La FIG. 3 muestra curvas ROC de las dos PSSM seleccionadas para la mejor práctica actual de la invención.

En otra realización, el rendimiento de una PSSM se evalúa comparando una pluralidad de motivos de secuencia identificados utilizando la PSSM con una pluralidad de motivos de secuencia de referencia. La PSSM se utiliza para obtener la pluralidad de motivos de secuencia, por ejemplo, explorando uno o más transcritos e identificando motivos de secuencia que coinciden con la PSSM, por ejemplo, con una puntación por encima de un umbral. Preferentemente, la pluralidad comprende al menos 3, 5, 10, 20 o 50 motivos de secuencia diferentes. Los motivos de secuencia de referencia pueden ser de cualquier fuente adecuada. En una realización, se obtiene una pluralidad de motivos de secuencia de referencia utilizando un método estándar (por ejemplo, Elbashir et al., 2001, Nature, 411: 494-8). Después, para determinar si son idénticas, las dos pluralidades se comparan utilizando cualquier método convencional conocido en la técnica.

En una realización preferida, las dos pluralidades se comparan utilizando una prueba de suma de rangos de Wilcoxon. Una prueba de suma de rangos de Wilcoxon analiza si dos pluralidades de mediciones son idénticas (véase, por ejemplo, Snedecor y Cochran, Statistical Methods, octava edición, 1989, Iowa State University Press, págs. 142-144; McClave y Sincich, 2002, Statistics, novena edición, Prentice Hall, capítulo 14). La prueba de suma de rangos de Wilcoxon puede considerarse como un equivalente no paramétrico de la prueba de la *t* para datos independientes. Se utiliza para probar la hipótesis de que dos muestras independientes provienen de la misma población. Debido a que no es paramétrica, solo hace suposiciones limitadas sobre la distribución de los datos. Supone que la forma de la distribución es similar en los dos grupos. Es de particular relevancia si la prueba se va a utilizar como prueba de que la mediana es significativamente diferente entre los grupos.

La prueba clasifica todos los datos de ambos grupos. Al valor más pequeño se le asigna un rango de 1, al segundo más pequeño se le asigna un rango de 2, y así sucesivamente. Cuando los valores están vinculados, reciben un rango promedio. Los rangos de cada grupo se suman (de ahí el término prueba de suma de rangos). Las sumas de los rangos se comparan con los valores críticos tabulados para generar un valor de *p*. En una prueba de suma de rangos de Wilkoxon, *p*, una función de X, Y y α, es la probabilidad de observar un resultado igual o más extremo que el que usa los datos (X e Y) si la hipótesis nula es verdadera. El valor de *p* indica la importancia de probar la hipótesis nula de que las poblaciones que generan las dos muestras independientes, X e Y, son idénticas. X e Y son vectores pero pueden tener diferentes longitudes, es decir, las muestras pueden tener diferentes números de elementos. La hipótesis alternativa es que la mediana de la población X se desplaza desde la mediana de la población Y en una cantidad distinta de cero, α es un nivel de significancia determinado y es un escalar entre cero y uno. En alguna realización, el valor predeterminado de α se establece en 0,05. Si *p* es casi cero, la hipótesis nula puede rechazarse.

En una realización, el estrategia de PSSM de la presente invención se comparó con el método estándar (por ejemplo, Elbashir et al., 2001, Nature 411: 494-8) por su rendimiento en la identificación de los ARNip que tienen una alta eficacia. En la Figura 3 se muestran los resultados obtenidos con tres ARNip seleccionados por cada método.

Los ARNip seleccionados mediante el método que utiliza la PSSM mostraron una mejor eficacia media (88 % en comparación con 78 % para el ARNip con el método estándar) y fueron más uniformes en su rendimiento. La eficacia mínima mejoró mucho (75 % en comparación con 12 % para el método estándar). La distribución de las eficacias de silenciamiento de los ARNip diseñados utilizando el algoritmo basado en PSSM fue significativamente mejor que la de los ARNip diseñados utilizando el método estándar para los mismos genes (p = 0,004, prueba de suma de rangos de Wilcoxon).

5.1.3. MÉTODO ALTERNATIVO PARA EVALUAR LA EFICACIA DE SILENCIAMIENTO DE LOS ARNip

Las estrategias de matriz de puntuación específica de posición son el método preferido para representar motivos funcionales de ARNip, por ejemplo, motivos susceptibles y resistentes a ARNip. Sin embargo, la información representada por las PSSM también puede representarse por otros métodos que también proporcionan pesos para la composición de bases en posiciones particulares. Esta sección proporciona dichos métodos para evaluar motivos funcionales de ARNip.

5.1.3.1. MÉTODOS BASADOS EN VENTANAS DE SECUENCIA

Un método habitual para ponderar la composición de bases en las posiciones en una secuencia, es contar el número de bases o conjunto de bases particular en una "ventana" de posiciones de secuencia. Como alternativa, el recuento se representa como un porcentaje. El número de valores de dicha puntuación, denominada puntuación de ventana, depende del tamaño de la ventana. Por ejemplo, la puntuación de una ventana de tamaño 5 para el contenido de G/C puede dar valores de 0, 1, 2, 3, 4 o 5; o 0 %, 20 %, 40 %, 60 %, 80 % o 100 %.

Un método alternativo para puntuar una ventana es calcular la temperatura de fusión o ΔG del dúplex para las bases en esa ventana. Estas cantidades termodinámicas reflejan la composición de todas las bases en la ventana, así como su orden particular. Es muy obvio para un experto en la técnica que estas cantidades termodinámicas dependen directamente de la composición de bases de cada ventana, y que están controladas por el contenido de G/C de la ventana, mientras que muestran alguna variación con el orden de las bases.

En una realización, la información representada por las diferencias de composición de bases, por ejemplo, en las Figuras 1A, IB y 1C, se representa con ventanas de composición de bases correspondientes a las posiciones en los picos de composición aumentada o disminuida de una o más bases particulares. Estas ventanas pueden puntuarse por el contenido de la(s) base(es) particular(es), con una composición de bases aumentada o disminuida que corresponde a secuencias que son más o menos funcionales o resistentes para el direccionamiento de ARNip. Por ejemplo, para representar algunos de los motivos funcionales de ARNip reflejados en la Figura 1 A, puede utilizarse una ventana de 5 bases de mayor contenido de G/C desde la base -1 a la base 3 en relación al dúplex de ARNip de 19meros, y una ventana de 16 bases de menor contenido de G/C desde la base 14 a la base 29 en relación al dúplex de ARNip de 19 meros.

Las puntuaciones pueden utilizarse directamente como un clasificador: en el ejemplo de una ventana de 5 bases, un clasificador de 5 partes está disponible automáticamente. Las puntaciones también pueden compararse con un umbral calculado u obtenido empíricamente para utilizar la ventana como un clasificador de 2 partes. Las ventanas también pueden utilizarse en combinación. Las puntuaciones de cada secuencia en múltiples ventanas se pueden sumar con o sin normalización o ponderación. En una realización, las puntuaciones de cada ventana se normalizan restando la puntuación media en un conjunto de puntuaciones y después dividiendo entre la desviación típica en el conjunto de puntuaciones. En otra realización, las puntaciones se ponderan mediante el coeficiente de correlación de Pearson obtenido comparando la puntuación de la ventana con la eficacia medida de un conjunto de ARNip. En otra realización, las puntuaciones se normalizan, y después se ponderan antes de la suma.

Como ejemplo del uso de ventanas para representar motivos funcionales de ARNip, se consideró la siguiente lista de parámetros para la predicción de la eficacia de ARNip:

1. Parámetros directos.

ATG_Dist - distancia hasta el codón de inicio.

STOP_Dist: distancia hasta el final de la región codificante

Coding_Percent - ATG_Dist como porcentaje de la longitud de la región codificante

End_Dist - distancia hasta el final del transcrito

Total_Percent - posición de inicio como un porcentaje de la longitud de la secuencia de transcripción.

2. Parámetros basados en ventanas.

Se consideraron 119 bases en la secuencia de transcripción (19 meros más 50 bases cadena abajo y 50 bases cadena arriba). Se examinaron ventanas de 3-10 tamaños para cada posición desde el principio hasta el final del fragmento de 119 bases. Para cada posición de ventana se contaron los siguientes elementos:

a. Número de bases: A, C, G o U.

b. Número de pares de bases: M (A o C), R (A o G), W (A o U), S (C o G), Y (C o U) y K (G o U).

c. Números de varios dímeros ordenados: AC, AT, AG, MM, RY, KM, SW, etc.

d. Los tramos más largos de la base anterior o de unidades de dos bases.

3. Parámetros basados en motivos.
Estos parámetros también se basan en los fragmentos de 119 bases. Las letras incluyen las bases (A, C, G, U) y los pares de bases (M, R, W, S, Y, K).

(1) Un mero, dímeros o trímeros específicos de posición.

(2) Números de 1 a 7 meros en cuatro regiones grandes: 50 bases cadena arriba, el propio oligo de 19 meros, 50 bases cadena abajo, y toda la región de 119 meros.

4. Parámetros estructurales.

Los parámetros estructurales se basan en las siguientes regiones.

el propio oligo de 19 meros (prefijo: propio)

el oligo de 20 meros inmediato cadena arriba del oligo (prefijo: 20 arriba)

el oligo de 40 meros inmediato cadena arriba del oligo

el oligo de 60 meros inmediato cadena arriba del oligo

el oligo de 20 meros inmediato cadena abajo de oligo (prefijo: 20 abajo)

el oligo de 40 meros inmediato cadena abajo del oligo

el oligo de 60 meros inmediato cadena abajo del oligo

El emparejamiento de bases previsto se examinó con el programa informático *RNAStructure* y se calcularon los siguientes parámetros:

el recuento de bucles protuberantes (parámetro: protuberancia)

las bases totales en los bucles protuberantes (protuberantes_b)

el recuento de los bucles internos (internos)

las bases totales en los bucles internos (internos_b)

el recuento de horquillas (horquilla)

las bases totales en las horquillas (horquilla_b)

el recuento de otras regiones de motivo (otras)

las bases totales en las otras regiones de motivo (otras_b)

el total de bases emparejadas (total_emparejadas_b)

el total de bases no emparejadas (total_no emparejadas_b)

el tramo más largo de bases emparejadas (más largo_emparejadas_b)

el tramo más largo de bases no emparejadas (más largo _no emparejadas_b)

Por lo tanto, para cada ARNip, se calculó un total de 12*7 = 84 parámetros en relación con los motivos de la estructura secundaria.

5. Parámetros en predicciones inespecíficas (*off-target*).

Utilizando la puntación ponderada del programa FASTA, comentado en la Sección 5.2., la puntación minimax y la ΔG del dúplex prevista, comentada en la Sección 5.4, utilizando diferentes condiciones, se calcularon 10 parámetros diferentes.

Los parámetros se normalizaron y se ponderaron mediante el coeficiente de correlación de Pearson de las puntuaciones con la eficacia de silenciamiento de los ARNip examinados. Se utilizaron diversos métodos para seleccionar los parámetros con el mayor poder predictivo para la eficacia de los ARNip; los diversos métodos coincidieron en la selección de 1750 parámetros. 1190 de estos son parámetros de composición de base basados en ventana, 559 son parámetros de composición de bases basados en motivos, y solo se seleccionó 1 parámetro estructural. No se seleccionaron otros parámetros.

5.1.3.2. MÉTODOS DE PUNTUACIÓN DE FAMILIAS DE SECUENCIAS

Para representar motivos funcionales de ARNip, por ejemplo, motivos susceptibles o resistentes a ARNip, como alternativa a las PSSM, también pueden utilizarse patrones de secuencias consenso, modelos ocultos de Markov y redes neuronales.

En primer lugar, un motivo funcional de ARNip, por ejemplo, un motivo susceptible o resistente a ARNip, puede entenderse como una secuencia consenso imprecisa, para una familia de secuencias distantemente relacionadas, por ejemplo, la familia de sitios diana de ARNic funcional. La puntuación de secuencias para la similitud con un consenso familiar es muy conocida en la técnica (Gribskov, M., McLachlan, AD, y Esienberg, D. 1987. Profile analysis: detection ofdistantly related proteins. PNAS 84: 4355-4358; Gribskov, M ., Luthy, R., y Eisenberg, D. 1990. Profile analysis. Meth. Enzymol. 183: 146 - 159). Dichos métodos de puntuación se denominan comúnmente "perfiles", pero también pueden denominarse "moldes" o "patrones flexibles" o términos similares. Dichos métodos son descripciones más o menos estadísticas del consenso de un alineamiento de secuencias múltiples, utilizando puntuaciones específicas de posición para bases o aminoácidos particulares, así como para inserciones o deleciones en la secuencia. Los pesos se pueden derivar del grado de conservación en cada posición. Una diferencia entre los perfiles de consenso y las PSSM, como se utiliza el término en este texto, es que el espaciado puede ser flexible en los perfiles de consenso: las partes discontinuas de un motivo funcional de ARNip, por ejemplo, motivos susceptibles o resistentes a ARNip se pueden encontrar a diversas distancias entre sí, con inserciones o deleciones permitidas y puntuadas como son las bases.

Los modelos ocultos de Markov para perfiles, son modelos estadísticos que también representan el consenso de una familia de secuencias. Krogh y colegas (Krogh, A., Brown, M., Mian, IS, Sjolander, K. y Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. J. Mol Biol. 235:1501-1531) aplicaron técnicas de HMM para modelar perfiles de secuencia, adoptando técnicas de estudios de reconocimiento de voz (Rabiner, LR 1989. A tutorial on hidden Markov models and selected applications to speech recognition. Proc. IEEE 77:257-286). El uso de modelos ocultos de Markov para el análisis de secuencias biológicas es muy conocido actualmente en la técnica y se dispone fácilmente de las aplicaciones para el cálculo del modelo oculto de Markov, por ejemplo, el programa HMMER (http://hmmer.wustl.edu).

Los modelos ocultos de Markov para perfiles, se diferencian de los perfiles de consenso, como se describió anteriormente, en que los modelos ocultos de Markov para perfiles tienen una base probabilística formal para establecer los pesos para cada base, inserción o deleción en cada posición. Los modelos ocultos de Markov también pueden realizar el alineamiento de secuencias desconocidas para el descubrimiento de motivos, así como para determinar pesos específicos de posición para dichos motivos, mientras que los perfiles de consenso generalmente derivan de secuencias previamente alineadas.

Los perfiles de consenso y los modelos ocultos de Markov para perfiles pueden suponer que la composición de bases en una posición particular es independiente de la composición de bases de todas las otras posiciones. Esto es similar a las PSSM ascendentes aleatorias de esta invención, pero difiere de los modelos PSSM de ventanas y curvas.

Para capturar la dependencia de la composición de bases en una posición particular en la composición de posiciones vecinas, los modelos de Markov se pueden utilizar como cadenas de Markov de orden fijo y modelos de Markov interpolados. Salzberg y colegas aplicaron modelos de Markov interpolados para encontrar genes en genomas microbianos como una mejora sobre las cadenas de Markov de orden fijo (Salzberg, SL, Delcher, AL, Kasif, S. y White, O. 1998. Nucl. Acids Res. 26: 544-548). Una cadena de Markov de orden fijo predice cada base de una secuencia en función de un número fijo de bases que preceden a esa posición. El número de bases precedentes utilizado para predecir el siguiente se conoce como orden de la cadena de Markov. Los modelos de Markov interpolados utilizan un número flexible de bases precedentes para predecir la composición de bases en una

posición particular. Esto permite la capacitación en conjuntos de secuencias más pequeñas. Pueden estar disponibles datos predictivos suficientes para n-meros de varias longitudes en un conjunto de capacitación de modo que se puedan hacer algunas predicciones de bases sucesivas, mientras que pueden estar disponibles datos insuficientes para todos los oligómeros en cualquier longitud fija. Los modelos de Markov interpolados tienen así más libertad para utilizar oligómeros preferibles más largos para la predicción que las cadenas de Markov de orden fijo, cuando dichos oligómeros largos son suficientemente frecuentes en el conjunto de capacitación. Los modelos de Markov interpolados emplean una combinación ponderada de probabilidades de una pluralidad de longitudes de oligómeros para la clasificación de cada base.

Las cadenas de Markov de orden fijo y los modelos de Markov interpolados pueden representar motivos funcionales de ARNip, por ejemplo, motivos susceptibles o resistentes a ARNip en términos de dependencia de la composición de bases en una posición particular sobre la composición de las posiciones precedentes. Un proceso de construcción de modelos interpolados de Markov descubrirá los oligómeros más predictivos de motivos funcionales o no funcionales de ARNip.

Las redes neuronales también se emplean para puntuar secuencias por similitud con una familia de secuencias. Una red neuronal es una herramienta de análisis estadístico utilizada para construir un modelo a través de un proceso de aprendizaje iterativo. La red capacitada realizará después una tarea de clasificación, dependiente de la salida deseada y de la entrada de capacitación inicialmente asociada con esa salida. Por lo general, se suministra un programa de red neuronal o dispositivo informático con un conjunto de secuencias de capacitación y se configura un estado que representa esas secuencias. La red neuronal se ensaya después para determinar el rendimiento en un conjunto de secuencias de ensayo. Las redes neuronales se pueden utilizar para predecir y modelar motivos funcionales de ARNip, por ejemplo, motivos susceptibles y resistentes a ARNip. Una desventaja de las redes neuronales es que las características de secuencia reales de un motivo pueden ser difíciles o imposibles de determinar a partir del examen del estado de la red capacitada.

5.1.4. Métodos de identificación de motivos de secuencia en un gen para su direccionamiento por un ARNip

La divulgación proporciona un método para identificar uno o más motivos de secuencia en un transcrito que son motivos susceptibles o resistentes a ARNip. Por lo tanto, el método también proporciona ARNip funcionales o no funcionales correspondientes. En una realización, se explora la región de secuencia de interés para identificar secuencias que coincidan con el perfil de un motivo funcional. En una realización, se evalúa una pluralidad de posibles motivos de secuencia de ARNip que comprenden motivos de secuencia de ARNip que configuran la región en etapas de intervalos de bases predeterminados para identificar secuencias que coincidan con el perfil. En una realización preferida, se utilizan etapas de 1, 5, 10, 15 o 19 intervalos de bases. En una realización preferida, se explora toda la secuencia de transcripción. Para cada motivo de secuencia diferente se calcula una puntuación utilizando una PSSM como se describe en las Secciones 5.1.1.-5.1.3. Después, las secuencias se clasifican según la puntuación. Después, se selecciona una o más secuencias de la lista de clasificación. En una realización, los motivos de secuencia de ARNip que tienen las puntuaciones más altas se seleccionan como motivos susceptibles a ARNip. En otra realización, los motivos de secuencia de ARNip que tienen las puntuaciones más bajas se seleccionan como motivos resistentes a ARNip.

Los inventores han descubierto que la correlación entre la eficacia de silenciamiento y los perfiles de composición de bases de motivos funcionales de ARNip, puede depender de uno o más factores, por ejemplo, de la abundancia del transcrito diana. Por ejemplo, los inventores han descubierto que para silenciar genes poco expresados, p. ej., genes cuyos niveles de transcripción son inferiores a aproximadamente 5 copias por célula, los motivos funcionales de ARNip que tienen alta asimetría de contenido de GC en los dos extremos de la secuencia diana y que tienen un alto contenido de GC en las regiones de secuencia que flanquean la secuencia diana, tienen menor eficacia de silenciamiento que los motivos funcionales de ARNip que tienen moderada asimetría de contenido de GC en los dos extremos de la secuencia diana y bajo contenido de GC en las regiones flanqueantes. El efecto de la abundancia del transcrito diana sobre la eficacia de silenciamiento se ilustra en el Ejemplo 6.

Sin limitarse a ninguna teoría, los inventores llegan a la conclusión de que la eficacia de silenciamiento de un motivo funcional de ARNip particular es el resultado de la interacción de diversos procesos, incluyendo la formación de RISC y el desenrrollamiento del dúplex de ARNip, la difusión de RISC y de ARNm diana, la reacción del complejo RISC/diana, que puede incluir a difusión de RISC a lo largo del ARNm diana, la reacción de escisión, y la disociación de productos, etc. Por lo tanto, la abundancia del transcrito, el perfil de composición de bases del ARNip, el perfil de composición de bases de la secuencia diana y las secuencias flanqueantes, y la concentración del ARNip y RISC en una célula, pueden afectar a la eficacia de silenciamiento. Diferentes procesos pueden implicar diferentes regiones de secuencia de un ARNip o motivo de secuencia de un ARNip, es decir, diferentes regiones de secuencia de un ARNip o motivo de secuencia de secuencia de un ARNip pueden tener diferentes funciones en el reconocimiento, escisión y liberación de producto de un transcrito, los ARNip pueden diseñarse en función de criterios que tienen en cuenta una o más de dichas características. Por ejemplo, las bases próximas al extremo 5' de la cadena guía están implicadas en la unión a transcritos (transcritos tanto diana como inespecíficos), y se ha demostrado que es suficiente para la energía de unión al ARN diana. Un emparejamiento de bases más débil en el extremo 5' de la cadena antisentido (extremo 3' del dúplex) estimula la interacción preferencial de la cadena antisentido con RISC,

por ejemplo, facilitando el desenrrollamiento del dúplex de ARNip mediante un componente helicasa 5'-3' de RISC. Una preferencia por U en la posición 10 de la cadena en sentido de un ARNip se ha asociado con una eficacia de escisión mejorada por RISC como lo es en la mayoría de las endonucleasas. La secuencia con bajo contenido de GC que flanquea el sitio de escisión puede mejorar la accesibilidad del complejo RISC/nucleasa para la escisión, o la liberación del transcrito escindido, según estudios recientes que demuestran que los pares de bases formados por las regiones central y 3' de la cadena guía de ARNip proporcionan una geometría helicoidal necesaria para la catálisis. Por tanto, la invención proporciona un método de identificación de motivos de secuencia de ARNip (y por lo tanto los ARNip) obteniendo ARNip que tienen una composición de secuencia óptima en una o más regiones de secuencia de manera que estos ARNip son óptimos en uno o más procesos funcionales de ARNip. En una realización, el método comprende identificar motivos de secuencia de ARNip cuya secuencia global y/o diferentes regiones de secuencia tienen perfiles de composición deseados. El método puede utilizarse para identificar motivos de ARNip que tienen una composición de secuencia deseada en una región particular, por lo tanto, se optimizan para un proceso funcional. El método también se puede utilizar para identificar ARNip que tienen la composición de secuencia deseada en diversas regiones, por lo que se optimizan para una serie de procesos funcionales.

En una realización preferida, se obtiene un solo perfil funcional de ARNip, por ejemplo, un perfil representado por un conjunto de PSSM, por ejemplo, capacitando con datos de eficacia de silenciamiento una pluralidad de ARNip que se dirigen a genes que tienen diferentes abundancias de transcritos utilizando un método descrito en la Sección 5.1.2., o en la Sección 5.1.3., y se utiliza para evaluar motivos de secuencia de ARNip en transcripciones de genes que tienen abundancias en todos los intervalos. En una realización, los motivos de secuencia de ARNip en transcripciones génicas que tienen abundancias en cualquier intervalo, se evalúan basándose en el grado de similitud de sus perfiles de composición de bases en la secuencia con el perfil o perfiles representados por el conjunto de PSSM. En una realización, las puntuaciones de PSSM de motivos funcionales de ARNip para un gen de interés, se obtienen mediante un método descrito en la Sección 5.1.1. Basándose en los ARNip que se dirigen a genes que tienen niveles de expresión en diferentes intervalos, se determina un valor de referencia o un intervalo de valores de referencia predeterminado de la puntuación de PSSM. Más adelante se describen métodos para determinar el valor de referencia o intervalo de valores de referencia. Los motivos funcionales de ARNip en un gen particular se clasifican en función de la proximidad de sus puntuaciones al valor de referencia predeterminado o dentro del intervalo de referencia. Después, se seleccionan uno o más ARNip que tienen puntuaciones más próximas al valor predeterminado o dentro del intervalo de referencia.

El valor de referencia o el intervalo de referencia pueden determinarse de varias maneras. En una realización preferida, se evalúa la correlación de las puntuaciones de PSSM de una pluralidad de ARNip que tienen una o más características, por ejemplo, que tienen una eficacia particular en uno o más procesos funcionales de ARNip, con eficacia de silenciamiento. En una realización preferida, la característica es que la pluralidad de ARNip se dirige a genes poco expresados. El valor de la puntuación correspondiente a la mediana máxima de silenciamiento se utiliza como valor de referencia. En una realización específica, el valor de referencia es 0. Se seleccionan uno o más ARNip que tienen puntuaciones de PSSM más próximas a la puntuación de referencia.

En otra realización, el intervalo de puntuaciones correspondiente a ARNip que tienen un nivel de eficacia de silenciamiento determinado, por ejemplo, eficacia superior al 75 %, se utiliza como el intervalo para los valores de referencia. En una realización, se encuentra que los ARNip eficaces tienen puntuaciones entre -300 y +200 siempre que se controle el contenido de GC en las bases 2-7. En una realización específica, se utiliza un valor de referencia de entre -300 y +200. Se seleccionan uno o más ARNip que tienen puntuaciones de PSSM dentro del intervalo.

En otra realización preferida, como intervalo del valor de referencia, se utiliza un intervalo de puntuación particular dentro del intervalo de puntuaciones de PSSM de la pluralidad de ARNip que tienen una o más características, por ejemplo, que tienen una eficacia particular en uno o más procesos funcionales de ARNip. En una realización preferida, la característica es que la pluralidad de ARNip se dirige a genes poco expresados. En una realización, como intervalo del valor de referencia, se utiliza un cierto percentil en el intervalo de puntuaciones de PSSM, por ejemplo, 90 %, 80 %, 70% o 60 %. En una realización específica, el intervalo combinado de puntuación de PSSM en el conjunto de capacitación tiene un máximo de 200, teniendo un valor de 0 o menor el 97 % de las puntuaciones y por debajo de - 300 el 60 % de las puntuaciones.

En otra realización preferida adicional, como puntuación de referencia se utiliza una suma de puntuaciones de una pluralidad de conjuntos de PSSM (véase la Sección 5.1.2). En una realización específica, la pluralidad de conjuntos consta de los dos conjuntos de PSSM descritos anteriormente. Los dos conjuntos de PSSM difieren en la composición de bases preferida para los ARNip, en particular con respecto al contenido de GC de las secuencias de 19 meros y flanqueantes. Con una puntuación combinada de 0, los conjuntos de PSSM están en equilibrio en su preferencia por el ARNip.

En otra realización preferida, además de las puntuaciones de PSSM, los motivos de secuencia de ARNip también se clasifican según el contenido de GC en las posiciones correspondientes a las posiciones 2-7 de los ARNip correspondientes, y en la región se seleccionan uno o más motivos de secuencia de ARNip que tienen un contenido de GC de aproximadamente 0,15 a 0,5 (correspondiente a 1-3 G o C).

En otra realización preferida adicional, se seleccionan motivos de secuencia de ARNip que tienen una G o C en la posición correspondiente a la posición 1 del ARNip de 19 meros correspondiente y una A o T en la posición correspondiente a la posición 19 del correspondiente ARNip de 19 meros. En otra realización preferida adicional, se seleccionan motivos de ARNip en los que 200 bases a cada lado de la región diana de 19 meros, no son secuencias de repetición o de baja complejidad.

En una realización específica, los motivos de secuencia de ARNip se seleccionan de la siguiente manera: (1) se clasifican primero según el contenido de GC en las posiciones correspondientes a las posiciones 2-7 de los ARNip correspondientes, y en la región se seleccionan uno o más motivos de secuencia de ARNip que tienen un contenido en GC de aproximadamente 0,15 a 0,5 (correspondiente a 1-3 G o C); (2) a continuación, se seleccionan motivos de secuencia de ARNip que tengan una G o C en la posición correspondiente a la posición 1 del ARNip de 19 mero correspondiente y una A o T en la posición correspondiente a la posición 19 del ARNip de 19 meros correspondiente; (3) después se seleccionan ARNip que tengan puntaciones de PSSM en el intervalo de -300 a 200 o más próximas a 0; (4) después se selecciona diversas coincidencias de BLAST inespecíficas inferiores a 16; y (5) se seleccionan motivos de ARNip en los que 200 bases a cada lado de la región diana de 19 meros no son secuencias de repetición o de baja complejidad.

En otra realización, para cada uno de una pluralidad de intervalos de abundancia diferentes, se determina un valor de referencia o intervalo de referencia. La selección de motivos funcionales de ARNip en un gen de interés se realiza utilizando el valor de referencia o el intervalo de referencia apropiado para el intervalo de abundancia en el que se encuentra el gen de interés. En una realización, la pluralidad de diferentes intervalos de abundancia consta de dos intervalos: por debajo de aproximadamente 3-5 copias por célula, que corresponde a genes poco expresados, y por encima de 5 copias por célula, que corresponde a genes muy expresados. El valor de referencia o intervalo de referencia se puede determinar para cada intervalo de abundancia utilizando cualquiera de los métodos descritos anteriormente.

En otra realización, para una pluralidad de intervalos de abundancia de transcritos diferentes, se determina una pluralidad de perfiles de motivos funcionales de ARNip. Cada uno de dichos perfiles se determina en función de los datos de eficacia de silenciamiento de los ARNip que se dirigen a genes que tienen niveles de expresión en un intervalo determinado, es decir, genes cuyas abundancias de transcripción están dentro de un intervalo determinado, utilizando un método descrito en las Secciones 5.1.2 y 5.1.3., anteriores. En una realización, un conjunto de una o más PSSM para genes que tienen niveles de expresión en un intervalo determinado se capacitan como se describe en la Sección 5.1.2., utilizando ARNip que se dirigen a genes que tienen niveles de expresión en el intervalo. Las PSSM se utilizan después para identificar motivos funcionales de ARNip en un gen diana cuyo nivel de expresión está en el intervalo, por ejemplo, clasificándolos de acuerdo con las puntuaciones de PSSM obtenidas utilizando un método descrito en la Sección 5.1.1. En una realización preferida, los intervalos de abundancia de transcritos se dividen en dos intervalos: por debajo de aproximadamente 3-5 copias por célula, que corresponde a genes poco expresados, y por encima de 5 copias por célula, que corresponde a genes muy expresados. Se obtienen dos conjuntos de PSSM, uno para cada intervalo de abundancia. Los motivos funcionales de ARNip en un gen de interés se pueden identificar utilizando el conjunto de PSSM que es apropiado para la abundancia del gen de interés.

La divulgación también proporciona métodos para evaluar las eficacias de silenciamiento de motivos de secuencia de ARNip a diferentes concentraciones de ARNip. Por ejemplo, los métodos descritos anteriormente para evaluar la eficacia de silenciamiento de motivos de secuencia de ARNip en transcritos que tienen diferentes abundancias pueden utilizarse para dichos fines reemplazando el parámetro de abundancia con el parámetro de concentración. En una realización, se determina una pluralidad de perfiles de motivos funcionales de ARNip para una pluralidad de intervalos de concentración de ARNip diferentes. Cada uno de dichos perfiles se puede determinar basándose en datos de eficacia de silenciamiento de diferentes concentraciones de ARNip que se dirigen a genes que tienen un nivel de expresión diferente o que tienen un nivel de expresión en un intervalo diferente. En una realización, dichos perfiles se determinan para transcritos que tienen una abundancia determinad o que tienen una abundancia dentro de un intervalo de abundancias. Cada perfil de este tipo se puede determinar basándose en datos de eficacia de silenciamiento de diferentes concentraciones de ARNip que se dirigen a genes que tienen el nivel de expresión o que tienen un nivel de expresión en el intervalo. En una realización, una o más PSSM para un intervalo de concentración de ARNip determinado se capacitan basándose en datos de eficacia de silenciamiento de ARNip que tienen una concentración en el intervalo. Después, las PSSM pueden utilizarse para seleccionar ARNip que tienen alta eficiencia a una concentración que se encuentra en el intervalo de concentración. En una realización preferida, los intervalos de abundancia de transcritos se seleccionan para que estén por debajo de 5 copias por célula. En otra realización, los intervalos de abundancia de transcritos se seleccionan para que sean superiores a 5 copias por célula.

Por tanto la invención proporciona un método para seleccionar uno o más motivos funcionales de ARNip para el direccionamiento mediante los ARNip de una concentración determinada

Los métodos pueden utilizarse para identificar uno o más motivos funcionales de ARNip que pueden ser dirigidos por ARNip de una concentración determinada con eficacia de silenciamiento deseada. La concentración dada está preferentemente en el intervalo de bajo nanomolar a subnanomolar, más preferentemente en el intervalo de picomolar. En realizaciones específicas, la concentración dada es de 50 nmol, 20 nmol, 10 nmol, 5 nmol, 1 nmol, 0,5

nmol, 0,1 nmol, 0,05 nmol o 0,01 nmol. La eficacia de silenciamiento deseada es de al menos 50 %, 75 %, 90 % o 99 % a una concentración dada. Dichos métodos son particularmente útiles para diseñar ARNip terapéuticos. Para usos terapéuticos, a menudo es deseable identificar ARNip que puedan silenciar un gen diana con alta eficacia a concentraciones subnanomolares a picomolares. Por tanto, la invención también proporciona un método para el diseño de ARNip terapéuticos.

La divulgación también proporciona un método para determinar si un gen es adecuado para dirigirse por un ARNip terapéutico. En una realización, primero se determina la concentración deseada de ARNip y la eficacia de silenciamiento deseada. Utilizando un método de esta invención se evalúa una pluralidad de posibles motivos de secuencia de ARNip en el transcrito del gen. Se identifican uno o más motivos de secuencia de ARNip que exhiben la eficacia más alta, por ejemplo, que tienen puntuaciones de PSSM que satisfagan el criterio o criterios descritos anteriormente. El gen se determina como adecuado para dirigirse por un ARNip terapéutico si el uno o más motivos de secuencia de ARNip pueden ser dirigidos por los ARNip correspondientes con eficacia de silenciamiento superior o igual a la eficacia deseada. En una realización, la pluralidad de posibles motivos de secuencia de ARNip comprende motivos de secuencia de ARNip que abarcan o se extienden a lo largo de una parte de o en todo el transcrito en etapas de intervalos de bases predeterminados, por ejemplo en etapas de 1, 5, 10, 15 o 19 intervalos de bases. En una realización preferida, los motivos sucesivos de secuencias de ARNip solapantes se extienden a lo largo de toda la secuencia del transcrito. En otra realización preferida, los motivos sucesivos de secuencias de ARNip solapantes se extienden a lo largo de una región de o en toda la secuencia del transcrito a etapas de 1 intervalo de bases.

5.2. MÉTODOS DE IDENTIFICACIÓN DE GENES INESPECÍFICOS DE UN ARNip

La divulgación también proporciona un método para identificar genes inespecíficos de un ARNip. Como se utiliza en este documento, un gen "inespecífico" (no diana) es un gen que se silencia directamente mediante un ARNip que está diseñado para dirigirse a otro gen (véase la solicitud internacional Nº PCT / US2004 / 015439 de Jackson et al., Presentada el 17 de mayo de 2004). La cadena en sentido o la cadena antisentido del ARNip pueden silenciar un gen inespecífico.

5.2.1. PERFIL DE COINCIDENCIA DE SECUENCIAS Y SILENCIAMIENTO INESPECÍFICO

Los experimentos con micromatrices sugieren que la mayoría de los oligos de ARNip dan como resultado la regulación negativa de genes inespecíficos a través de interacciones directas entre un ARNip y los transcritos inespecíficos. Aunque la similitud de secuencia entre ARNbc y transcritos parece jugar un papel en la determinación de qué genes inespecíficos se ven afectados, las búsquedas de similitud de secuencias, incluso combinadas con modelos termodinámicos de hibridación, son insuficientes para predecir con precisión efectos inespecíficos. Sin embargo, el alineamiento de transcritos inespecíficos con secuencias de ARNip no válidas revela que algunas interacciones de emparejamiento de bases entre los dos parecen ser más importantes que otras (figura 6).

La divulgación proporciona un método para identificar posibles genes inespecíficos de un ARNip utilizando una PSSM que describe el patrón de coincidencia de secuencias entre un ARNip y una secuencia de un gen inespecífico (pmPSSM). En una realización, el patrón de coincidencia de secuencias se representa por pesos de diferentes posiciones en un ARNip para coincidir con las posiciones diana correspondientes en transcritos inespecíficos $\{P_i\}$, siendo P_i el peso de una coincidencia en la posición *i, i = 1, 2, ..., L*, siendo *L* la longitud del ARNip. Dicho patrón de coincidencia puede determinarse basándose en la frecuencia con la que se encuentra que cada posición en un ARNip coincide con transcritos inespecíficos afectados identificados como dianas directas del ARNip mediante regulación negativa simultánea con la diana deseada a través de análisis cinéticos de perfiles de expresión (véase la solicitud internacional Nº PCT/US2004/015439 de Jackson et al., presentada el 17 de mayo de 2004). Una pmPSSM puede ser $\{E_i\}$, en la que $E_i = P_i$ si la posición *i* en el alineamiento es una coincidencia y $E_i = (1- P_i)/3$ si la posición *i* es una coincidencia errónea. En la FIG. 7 se representa gráficamente una $\{P_i\}$ a modo de ejemplo para una secuencia de ARNip de 19 meros y se enumera en la Tabla I.

Tabla I Pesos de una pmPSSM a modo de ejemplo para ARNip de 21 nt que tienen una región dúplex de 19 nt

1	0,25
2	0,32
3	0,32
4	0,46
5	0,39
6	0,38
7	0,36
8	0,45
9	0,61
10	0,47
11	0,76
12	0,96

13	0,94
14	0,81
15	0,92
16	0,94
17	0,89
18	0,78
19	0,58

En una realización, para obtener una pmPSSM se utiliza el patrón de coincidencia de secuencia de transcritos inespecíficos. Los genes inespecíficos de un ARNip pueden identificarse utilizando un método divulgado en la solicitud internacional Nº PCT/US2004/015439 de Jackson et al., presentada el 17 de mayo de 2004. Por ejemplo, los genes inespecíficos de un ARNip se identifican basándose en cinética de silenciamiento (véase, por ejemplo, la solicitud internacional Nº PCT/US2004/015439 de Jackson et al., presentada el 17 de mayo de 2004). Después, puede generarse una pmPSSM utilizando la frecuencia de coincidencias encontradas en cada posición. En una realización, el alineamiento mostrado en la Fig. 6 y datos similares para otros ARNip, se combinaron para generar la matriz de puntuación específica de posición a modo de ejemplo para utilizar en la predicción de efectos inespecíficos.

El grado de coincidencia entre un ARNip y una secuencia en un transcrito puede evaluarse con la pmPSSM utilizando una puntuación (también denominada puntuación de coincidencia de posición, Puntuacióncp) de acuerdo con la siguiente ecuación

$$Puntuación = \sum_{i=1}^{L} \ln(E_i / 0{,}25) \qquad (6)$$

en la que *L* es la longitud del alineamiento, por ejemplo, 19. Una Puntuacióncp por encima de un umbral determinado identifica la secuencia como una posible secuencia inespecífica.

Los inventores han descubierto que para un ARNip determinado, el número de alineamientos con una puntuación por encima de un umbral es predictivo del número de efectos inespecíficos observados. El umbral de puntuación puede optimizarse maximizando la correlación entre el número previsto y observado de efectos diana (Fig. 8). El umbral optimizado puede utilizarse para favorecer la selección de ARNip con un número relativamente pequeño de efectos inespecíficos previsto.

5.2.2. MÉTODO DE IDENTIFICACIÓN DE GENES INESPECÍFICOS DE UN ARNip

Los genes inespecíficos de un ARNip determinado pueden identificarse identificando primero secuencias de transcrito inespecíficas que se alinean con el ARNip. Para el alineamiento por pares puede utilizarse cualquier método adecuado, tal como, pero sin limitación, BLAST y FASTA. La matriz de puntuación específica de posición se utiliza después para calcular puntuaciones de coincidencia de posición para estos alineamientos. En una realización preferida, los alineamientos se establecen con una búsqueda FASTA de baja rigurosidad y la puntuación para cada alineamiento se calcula de acuerdo con la ecuación 6. Una puntuación por encima de un umbral determinado identifica el transcrito que comprende la secuencia como un posible gen inespecífico.

La divulgación también proporciona un método para evaluar la especificidad de silenciamiento de un ARNip. En una realización, se identifican posibles genes inespecíficos del ARNip. El número total de dichos genes inespecíficos en el genoma o en una parte del genoma, se utiliza después como una medida de la especificidad de silenciamiento del ARNip.

5.3. MÉTODO PARA LA PREDICCIÓN DE PREFERENCIA DE CADENA DE LOS ARNip

La divulgación proporciona un método para predecir la preferencia de cadena y/o la eficacia y especificidad de los ARNip basándose en la composición de bases específica de posición de los ARNip. Los inventores han descubierto que se predice que un ARNip cuya puntuación PSSM de composición de bases (véase la Sección 5.1) es mayor que la puntuación PSSM de composición de bases (PSSM G/C) de su complemento inverso, tiene una cadena antisentido que es más activa que su cadena en sentido. Por el contrario, se predice que un ARNip cuya puntuación PSSM de composición bases es menor que la puntuación PSSM de composición de bases de su complemento inverso, tiene una cadena en sentido que es más activa que su cadena antisentido.

Se ha demostrado que la eficacia aumentada de un ARNip en el silenciamiento de un gen diana idéntico a una cadena en sentido corresponde a una mayor actividad de la cadena antisentido y a una menor actividad de la cadena en sentido. Los inventores han descubierto que la composición de bases de PSSM puede utilizarse para distinguir ARNip con cadenas en sentido fuertes como ARNip malos de ARNip con cadenas en sentido débiles como buenos ARNip. Se observó que los complementos inversos de los ARNip malos eran aún más diferentes de los propios ARNip malos que los buenos ARNip. En promedio, los complementos inversos de ARNip malos tenían un

contenido de G/C aún más fuerte en el extremo 5 'que los ARNip buenos y eran similares en contenido de G/C a los ARNip buenos en el extremo 3'. Por el contrario, los complementos inversos de los ARNip buenos se observaron sustancialmente más similares a los ARNip malos que a los ARNip buenos. En promedio, los complementos inversos de los ARNip buenos apenas difieren de los ARNip malos en el contenido de G/C en el extremo 5' y solo eran ligeramente menos ricos en G/C que los ARNip malos en el extremo 3'. Estos resultados indican que las PSSM G/C distinguen los ARNip con cadenas en sentido fuertes como ARNip malos de los ARNip con cadenas en sentido débiles como ARNip buenos.

La FIG 14A muestra la diferencia entre el contenido medio de G/C de los complementos inversos de los ARNip malos con el contenido medio de G/C de los propios ARNip malos, dentro de la región dúplex de ARNip de 19 meros. La diferencia entre el contenido medio de G/C de ARNip buenos y malos se muestra para comparación. Las curvas se suavizaron sobre una ventana de 5 (o parte de una ventana de 5, en los bordes de la secuencia).

La FIG 14B muestra la diferencia entre el contenido medio de G/C de los complementos inversos de los ARNip buenos con el contenido medio de G/C de los ARNip malos, dentro de la región dúplex de ARNip de 19 meros. La diferencia entre el contenido medio de G/C de los ARNip buenos y malos se muestra para comparación. Las curvas se suavizan sobre una ventana de 5 (o parte de una ventana de 5, en los bordes de la secuencia).

En la FIG. 15, los ARNip se agruparon por eficacia de silenciamiento medida, y se comparó la frecuencia de entradas activas en sentido por el método sesgado en 3' y el método de PSSM G/C. Aunque estas técnicas se basan en diferentes análisis, la concordancia es bastante buena. Ambas muestran que se predice que una mayor proporción de ARNip de silenciamiento bajo frente a ARNip de silenciamiento alto es activa en sentido. El coeficiente de correlación para (puntuación PSSM G/C de ARNip - puntuación PSSM G/C de complemento inverso) frente a $\log_{10}$ (puntuación de identidad en sentido/ puntuación de identidad antisentido) es de 0,59 para el conjunto de 61 ARNip agrupados en la FIG. 15.

Por lo tanto, en una realización, la invención proporciona un método para predecir la preferencia de cadena, es decir, cuál de las dos cadenas es más activa, de ARNip basados en la composición de bases específica de posición de los ARNip. En una realización, el método comprende evaluar la preferencia de cadena de un ARNip en el silenciamiento génico comparando las composiciones de bases de las cadenas en sentido y antisentido del ARNip. En otra realización, el método comprende evaluar la preferencia de cadena de un ARNip en el silenciamiento génico comparando las composiciones de bases de la cadena en sentido y el complemento inverso de la secuencia diana del ARNip.

En una realización, la secuencia de la cadena antisentido de un ARNip o el complemento inverso de la secuencia diana del ARNip en un transcrito se compara con la secuencia diana utilizando un estrategia PSSM (véase la Sección 5.1). Se puntúa un ARNip y su complemento inverso utilizando una PSSM basándose en una diferencia de contenido de G/C suavizada entre los ARNip buenos y malos dentro de la región dúplex como la matriz de peso. En una realización, se utiliza una matriz de peso de composición de bases, como se describe en la FIG. 14A, como la matriz de peso. En una realización preferida, la puntuación PSSM de cada cadena puede calcularse como el producto puntual del contenido de G/C de la cadena de ARNip con la matriz de diferencia del contenido de G/C (como el método de cálculo de puntuación de la curva modelo de PSSM). En una realización, un ARNip se identifica como activo en sentido si su puntuación de PSSM de complemento inverso supera su propia puntuación de PSSM.

En otra realización, el método sesgado en 3', como se describe en la solicitud internacional Nº PCT/US2004/015439 de Jackson et al., presentada el 17 de mayo de 2004, se utiliza junto con la puntuación PSSM para determinar la preferencia de cadena de un ARNip. En una realización de este tipo, un ARNip se identifica como activo en sentido por el método sesgado en 3' de determinación de preferencia de cadena, si la puntuación antisentido idéntica supera la puntuación en sentido idéntica.

El método basado en la comparación de PSSM G/C de los ARNip y sus complementos inversos para la predicción del sesgo de cadena se ensayó por comparación con la estimación del sesgo de la cadena de los perfiles de expresión de ARNip por el método sesgado de 3'.

La divulgación también proporciona un método para identificar ARNip que tienen una buena eficacia de silenciamiento. El método comprende identificar ARNip que tienen actividad de cadena antisentido dominante (ARNip "activos antisentido") como ARNip que tienen buena eficacia y especificidad de silenciamiento (para silenciar una diana idéntica en sentido). En una realización, el método descrito en la Sección 5.1. se utiliza para identificar ARNip que tienen una cadena en sentido bueno (es decir, identificar ARNip que tienen buena eficacia de silenciamiento hacia una diana idéntica antisentido). Dichos ARNip se eliminan después de los usos en el silenciamiento de dianas idénticos en sentido. El método también puede utilizarse para eliminar ARNip con actividad de cadena en sentido dominante (ARNip de "sentido-activo") ya que los ARNip tienen menos eficacia y especificidad para silenciar dianas idénticas en sentido. En una realización, el método descrito en la solicitud internacional Nº PCT/US2004/015439 de Jackson et al., presentada el 17 de mayo de 2004, se utiliza para determinar la preferencia de cadena de un ARNip.

Los complementos inversos de los ARNip malos, en promedio, parecen tener un perfil de contenido GC que difiere del de los ARNip malos de la misma manera que el perfil de contenido GC de los ARNip buenos difiere del de los ARNip malos. Sin embargo, los complementos inversos de los ARNip malos muestran diferencias incluso más extremas de los ARNip malos que de los ARNip malos.

Esta observación está de acuerdo con la evidencia en los perfiles de expresión de ARNip que muchos ARNip malos tienen cadenas en sentido activo.

La combinación de datos y análisis sugiere por tanto que los complementos inversos de ARNip malos forman un modelo alternativo, o quizás incluso más ventajoso, para ARNip efectivos que los ARNip buenos. De este modo, la invención también proporciona un método para seleccionar ARNip basándose en la composición de bases de la secuencia de un complemento inverso de la cadena en sentido de los ARNip. En una realización, se clasifica una pluralidad de ARNip diferentes diseñados para silenciar un gen diana en un organismo en una secuencia diana diferente en una transcripción del gen diana de acuerdo con la composición de bases posicional de las secuencias de complemento inverso de sus cadenas en sentido. Después, puede seleccionarse uno o más ARNip, cuya composición de bases posicional de secuencias complementarias inversas coincida con la composición de bases posicional de ARNip deseados Preferentemente, la clasificación de ARNip se lleva a cabo determinando primero una puntuación para cada ARNip diferente utilizando una matriz de puntuación específica de posición. Los ARNip se clasifican según la puntuación. Cualquier método descrito en la Sección 5.1., anterior, puede utilizarse para puntuar secuencias complementarias inversas. En una realización, para los ARNip que tienen una secuencia de nucleótidos de *L* nucleótidos en la región dúplex, siendo *L* un número entero, la matriz de puntuación específica de posición comprende una diferencia en la probabilidad de encontrar el nucleótido G o C en la posición de secuencia *k* entre el complemento inverso de un primer tipo de ARNip y el complemento inverso de un segundo tipo de ARNip designado como w_k , k = 1, ..., L. La puntuación para cada complemento inverso se calcula de acuerdo con la ecuación

$$Puntuación = \sum_{k=1}^{L} w_k \qquad (7)$$

El primer tipo de ARNip puede constar de uno o más ARNip que tienen eficacia de silenciamiento no inferior a un primer umbral, por ejemplo, 75 %, 80 % o 90 %, a una dosis adecuada, por ejemplo, 100 nM, y el segundo tipo de ARNip puede constar de uno o más ARNip que tienen eficacia de silenciamiento inferior a un segundo umbral, por ejemplo, 25 %, 50 % o 75 %, a una dosis adecuada, por ejemplo, 100 nM. En una realización preferida, la diferencia de probabilidad se describe mediante una suma de curvas gaussianas, representando cada una de dichas curvas gaussianas, la diferencia en la probabilidad de encontrar una G o C en una posición de secuencia diferente.
Los métodos de esta divulgación también pueden aplicarse a modelos en desarrollo, por ejemplo, PSSM, de motivos funcionales de ARNip capacitando matrices de puntuación específicas de posición para distinguir entre ARNip malos y sus complementos inversos (véase, por ejemplo, la Sección 5.1). Una restricción en este análisis es que los complementos inversos de los ARNip malos no tienen dianas designadas. De este modo, en una realización, las matrices de puntuación específicas de posición de las secuencias dúplex de ARNip de 19 meros se capacitan para distinguir entre ARNip malos y sus complementos inversos.

Se puede realizar una capacitación de secuencia flanqueante en genes inespecíficos en el caso de distinguir entre ARNip malos y sus complementos inversos, así como en el caso de distinguir entre dos grupos cualquiera de ARNip. En otras palabras, se puede suponer que la actividad inespecífica de los ARNip tiene los mismos requisitos de secuencia flanqueante que la actividad en la diana, ya que se cree que en ambos procesos, están implicados los mismos complejos de ARN-proteína.

Por lo tanto, si se utilizan los métodos de la aplicación inespecífica para identificar genes directamente regulados negativamente por un ARNip (es decir, mediante análisis cinético de regulación negativa para identificar un grupo de genes regulados negativamente con la misma semivida que la diana prevista), las regiones que flanquean el alineamiento del ARNip con los genes inespecíficos regulados directamente, pueden utilizarse para modelos de capacitación y ensayo de requisitos de secuencia flanqueante. Estos modelos pueden desarrollarse mediante cualquiera de los métodos de esta invención: PSSM de ascenso aleatorio, PSSM de modelo de curva, matrices de frecuencia de diferencia entre bueno-malo, matrices de frecuencia de composición buena y/o matrices de frecuencia de composición mala, etc.

5.4. MÉTODOS DE DISEÑO DE ARNip PARA EL SILENCIAMIENTO DE GENES

La divulgación proporciona un método para diseñar ARNip para el silenciamiento de genes. El método puede utilizarse para diseñar ARNip que tengan homología de secuencia completa con sus secuencias diana respectivas en un gen diana. El método también puede utilizarse para diseñar ARNip que tienen solo homología de secuencia parcial con un gen diana. Los métodos y composiciones para silenciar un gen diana utilizando un ARNip que tiene solo homología de secuencia parcial con su secuencia diana en un gen diana se describen en la solicitud internacional Nº PCT/US2004/015439 de Jackson et al., presentada el 17 de mayo de 2004. Por ejemplo, un ARNip que comprende una secuencia de nucleótidos contigua de cadena en sentido de 11-18 nucleótidos que es idéntica a

una secuencia de un transcrito del gen diana, pero el ARNip no tiene homología de longitud completa con ninguna secuencia en el transcrito, puede utilizarse para silenciar el transcrito . Dicha secuencia de nucleótidos contigua está preferentemente en la región central de las moléculas de ARNip. Una secuencia de nucleótidos contigua en la región central de un ARNip puede ser cualquier tramo continuo de secuencia de nucleótidos en el ARNip que no comience en el extremo 3'. Por ejemplo, una secuencia de nucleótidos contigua de 11 nucleótidos puede ser la secuencia de nucleótidos 2-12, 3-13, 4-14, 5-15, 6-16, 7-17, 8-18 o 9-19. En realizaciones preferidas, la secuencia de nucleótidos contigua tiene 11-16, 11-15, 14-15, 11, 12 o 13 nucleótidos de longitud. Como alternativa, para silenciar el transcrito también puede utilizarse un ARNip que comprenda una secuencia de nucleótidos contigua de cadena 3' en sentido de 9-18 nucleótidos que sea idéntica a una secuencia de un transcrito del gen diana pero cuyo ARNip no tenga identidad de secuencia de longitud completa con ninguna secuencia contigua en el transcrito. Una secuencia en 3' de 9 - 18 nucleótidos es un tramo continuo de nucleótidos que comienza en la primera base emparejada, es decir, no comprende el saliente en 3' de dos bases. En realizaciones preferidas, la secuencia de nucleótidos contigua tiene una longitud de 9-16, 9-15, 9-12, 11, 10 o 9 nucleótidos.

En realizaciones preferidas, el método de la Sección 5.1 se utiliza para identificar de entre una pluralidad de ARNip uno o más ARNip que tienen una alta eficacia de silenciamiento. En una realización, cada ARNip en la pluralidad de ARNip se evalúa respecto a la eficacia de silenciamiento mediante las PSSM de composición base. En una realización, esta etapa comprende calcular una o más puntuaciones de PSSM para cada ARNip. Después, la pluralidad de ARNip se clasifica según la puntuación, y se selecciona uno o más ARNip utilizando un método descrito en la Sección 5.1.4.

En otras realizaciones preferidas, para identificar de entre una pluralidad de ARNip uno o más ARNip que tengan alta especificidad de silenciamiento, se utiliza el método de la Sección 5.2. En una realización, se identifican los alineamientos de cada ARNip con secuencias en cada una de una pluralidad de transcritos no diana y se evalúan con la estrategia de pmPSSM (véase la Sección 5.2). Para cada uno de los alineamientos se calcula una Puntuacióncp. Una Puntuacióncp por encima de un umbral determinado identifica una secuencia como una posible secuencia inespecífica. Dicha Puntuacióncp también se denomina puntuación de alineamiento. Por ejemplo, cuando se utiliza FASTA para el alineamiento, una Puntuacióncp puede ser una puntuación de alineamiento FASTA ponderada. El transcrito que comprende la posible secuencia inespecífica se identifica como un posible transcrito inespecífico. El número total de dichos transcritos inespecíficos en el genoma o en una parte del genoma se utiliza como una medida de la especificidad de silenciamiento del ARNip. Después, pueden seleccionarse uno o más ARNip que tengan menos transcritos inespecíficos.

Los ARNip que tienen niveles deseados de eficacia y especificidad para un transcrito pueden evaluarse adicionalmente para determinar la diversidad de secuencia. En esta divulgación, la diversidad de secuencia también se denomina "variedad de secuencia" o simplemente "diversidad" o "variedad". La diversidad de secuencias puede representarse o medirse en función de algunas características de secuencia. Los ARNip pueden seleccionarse de manera que una pluralidad de ARNip que se dirigen a un gen comprenda ARNip que exhiban diferencia suficiente en una o más de dichas características de diversidad.

Preferentemente, las características de diversidad de secuencia utilizadas en el método desvelado en el presente documento son cuantificables. Por ejemplo, la diversidad de secuencias puede medirse basándose en el contenido de GC, en la ubicación de la secuencia diana de ARNip a lo largo del transcrito diana, o las dos bases cadena arriba del dúplex ARNip (es decir, el dímero principal, con 16 dímeros principales posibles diferentes). La diferencia de dos ARNip puede medirse como la diferencia entre valores de una medida de diversidad de secuencia. La diversidad o variedad de una pluralidad de ARNip puede representarse cuantitativamente mediante la diferencia mínima o el espaciamiento en una medida de diversidad de secuencia entre diferentes ARNip en la pluralidad.

En el método de diseño de ARNip desvelado, la etapa de selección de los ARNip para diversidad o variedad también se denomina etapa de "des-solapamiento". En una realización preferida, para una medida de diversidad de secuencia que es cuantificable, el des-solapamiento selecciona ARNip que tienen diferencias de una medida de diversidad de secuencia entre dos ARNip por encima de un umbral determinado. Por ejemplo, el des-solapamiento por posición establece una distancia mínima entre oligos seleccionados a lo largo de la secuencia del transcrito. En una realización, se seleccionan ARNip situados al menos a 100 bases de separación en el transcrito. El des-solapamiento por contenido de GC establece una diferencia mínima en el contenido de GC. En una realización, la diferencia mínima en contenido de GC es de 1 %, 2 % o 5 %. El des-solapamiento por dímeros principales establece la probabilidad de todos o de una parte de los 16 dímeros principales posibles entre los ARNip seleccionados. En una realización, a cada uno de los 16 dímeros posibles se le asigna una puntuación de 1 a 16, y se utiliza un 0, 5 para seleccionar todos los cebadores principales posibles con la misma probabilidad.

En algunas realizaciones, los candidatos se des-solapan preferentemente sobre el contenido de GC, con una separación mínima de 5 %, un número máximo de duplicados de cada valor de GC % de 100 y al menos 200 candidatos seleccionados; más preferentemente, se des-solapan sobre el contenido de GC con una separación mínima de 5 %, un número máximo de duplicados de cada valor de GC % de 80 y al menos 200 candidatos seleccionados; y aún más preferentemente, se des-solapan sobre el contenido de GC con una separación mínima de 5 %, un número máximo de duplicados de cada valor de GC % de 60 y al menos 200 candidatos seleccionados.

Los ARNip pueden seleccionarse además en función de criterios de selección adicionales.

En una realización, se eliminan secuencias de direccionamiento de ARNip no comunes a todas las formas de corte y empalme documentadas.

En otra realización, se eliminan secuencias de direccionamiento de ARNip que se solapan con elementos de repetición simples o intercalados.

En otra realización más, se seleccionan secuencias de direccionamiento de ARNip situadas al menos a 75 bases cadena abajo del codón de inicio de la traducción.

En otra realización, se eliminan secuencias de direccionamiento de ARNip que se solapan o cadena abajo del codón de terminación. Esto evita secuencias de direccionamiento ausentes en formas de poliadenilación alternativas no documentadas.

En otra realización más, se seleccionan ARNip con contenido de GC próximo al 50 %. En una realización, se eliminan ARNip con GC% <20 % y> 70 %. En otra realización, se retienen 10 % <GC % <90 %, 20 % <GC % <80 %, 25 % <GC % <75 %, 30 % <GC % <70 %.

En otra realización más, se eliminan secuencias de direccionamiento de ARNip que contienen 4 restos consecutivos de guanosina, citosina, adenina o uracilo. En otra realización más, se seleccionan ARNip que se dirigen a una secuencia con un resto de guanina o citosina en la primera posición en la región dúplex de 19 meros en el extremo 5'. Dichas secuencias diana de ARNip se transcriben eficazmente mediante la ARN polimerasa III.

En otra realización más, se eliminan los ARNip que se dirigen a una secuencia que contiene sitios de reconocimiento para una o más endonucleasas de restricción determinadas, por ejemplo, endonucleasas de restricción XhoI o EcoRI. Esta realización puede utilizarse para seleccionar secuencias de ARNip para la construcción de los vectores de ARNhp.

En otra realización más, se evalúa la energía de unión de los ARNip. Para un método a modo de ejemplo de determinación de energía de unión véase el documento WO 01/05935. En una realización preferida, la energía de unión se evalúa calculando ΔG de 21 meros del vecino más cercano.

En otra realización más, se evalúa la especificidad de unión de los ARNip. Para un método a modo de ejemplo de determinación de la especificidad de unión de un oligo de 21 meros véase el documento WO 01/05935. En una realización preferida, la especificidad de unión se evalúa calculando una puntuación minimax de 21 meros contra el conjunto de representantes de secuencia únicos de genes de un organismo, por ejemplo, el conjunto de secuencias únicas representativas para cada grupo de *Homo sapiens* Unigene construcción 161 (http: //www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene).

En otra realización más, el método para predecir la preferencia de cadena y/o la eficacia y especificidad de los ARNip en función de la composición de bases específica de posición de los ARNip como se describe en la Sección 5.3. puede utilizarse para evaluar los ARNip candidatos.

En la FIG. 9 se muestra un flujograma de una realización a modo de ejemplo del método utilizado para seleccionar los ARNip.

En la etapa 101, se seleccionan secuencias de ARNip que se dirigen a un transcrito. En una realización, se consideran todas las subsecuencias de 19 meros del transcrito. También se obtienen y se considerar secuencias flanqueantes apropiadas para cada secuencia de ARNip. Los ARNip se evalúan frente a los siguientes filtros: (1) eliminación de secuencias de direccionamiento a ARNip que no son comunes a todas las formas de corte y empalme documentadas; (2) eliminación de secuencias de direccionamiento de ARNip que se solapan con elementos de repetición simples o intercalados; (3) eliminación de secuencias de direccionamiento de ARNip situadas en las 75 bases cadena abajo del codón de inicio de la traducción; y (4) eliminación de ARNip solapantes o cadena abajo del codón de terminación.

Para la selección de ARNhp, también se realizan las siguientes etapas: (5) eliminación de la secuencia de direccionamiento de ARNip que contiene 4 restos consecutivos de guanosina, citosina, adenina o uracilo; (6) retención de ARNip dirigidos a una secuencia con un resto de guanina o citosina en la primera posición en la región dúplex de 19 meros en el extremo 5'; y (7) eliminación de ARNip dirigidos a una secuencia que contiene sitios de reconocimiento para una o más enzimas de restricción determinadas, por ejemplo, endonucleasas de restricción XhoI o EcoRI, si las secuencias de ARNip se utilizan en la construcción de los vectores s de ARNhp.

En la etapa 102, se evalúa el ARNip para determinar la eficacia de silenciamiento mediante PSSM de composición de bases. En una realización, la etapa 102 comprende calcular una primera puntuación de PSSM, es decir, la

puntuación de PSSM-1, y una segunda puntuación de PSSM, es decir, la puntuación de PSSM-2, para un ARNip. Las dos puntaciones se suman para calcular la puntuación combinada de PSSM-1 + PSSM-2 para el ARNip. En una realización, las PSSM utilizadas son aquellas cuyo rendimiento se muestra en la Figura 2. El ARNip se retiene si la puntuación combinada está por encima de un umbral determinado.

Después, la energía de unión del ARNip se evalúa calculando ΔG de 21 meros del vecino más cercano. Después, la especificidad de unión del ARNip se evalúa calculando una puntuación minimax de 21 meros contra el conjunto de secuencias únicas representativas de genes de un organismo, por ejemplo, el conjunto de secuencias únicas representativas para cada grupo de *Homo sapiens* Unigene construcción 161. Véase el documento WO 01/05935 para métodos de cálculo de ΔG y de puntuación minimax. En una realización, los parámetros para los alineamientos BLAST y los cálculos delta-G del vecino más cercano basados en los alineamientos BLAST, que se utilizan para calcular las puntuaciones minimax, son los siguientes: -p blastn -e 100 –F F -W 11 -b 200 -v 10000 –S 3; y delta-G: temperatura 66 °; sal 1 M; concentración 1 pM; tipo de ácido nucleico, ARN. En una realización, el ARNip se elimina si el (ΔG de 21 meros – minimax 21 meros) ≤0,5.

En la etapa 103, se explora el contenido global de GC de los ARNip. En una realización, se eliminan los ARNip con contenido de GC desviado significativamente del 50 %, por ejemplo, GC % <20 % y> 70 %.

En la etapa 104, se explora la diversidad o variedad de los ARNip. La posición simplemente se refiere a la posición del oligo en la secuencia de transcripción y se proporciona automáticamente identificando el oligo. La variedad se impone en una o más etapas de "des-solapamiento" en el método. En resumen, el des-solapamiento selecciona el espaciado por encima del umbral entre oligos seleccionados en algún parámetro calculable. Para des-solapar, los oligos se clasifican primero de acuerdo con algún parámetro pensado para distinguir mejor de los ejecutantes más pobres y después se seleccionan para el espaciado entre oligos de acuerdo con algún otro parámetro. Para comenzar, se selecciona el oligo mejor clasificado. Después, se examina la lista clasificada, y se selecciona el siguiente mejor oligo con al menos el espaciado mínimo requerido del oligo seleccionado. Después, también se selecciona el siguiente mejor oligo con al menos el espaciamiento mínimo entre los dos oligos seleccionados. El proceso continúa hasta que se selecciona el número deseado de oligos. En una realización, múltiples oligos pueden compartir el mismo valor si un parámetro tiene poco valor, y el número de oligos que comparten el mismo valor está limitado por un umbral establecido. En una realización, si se selecciona un número insuficiente de oligonucleótidos en un primer paso de des-solapamiento, el requisito de espaciamiento puede relajarse hasta que se seleccione el número deseado, o el conjunto de todos los oligos disponibles restantes.

Por ejemplo, el des-solapamiento por posición establece una distancia mínima entre oligos seleccionados a lo largo de la secuencia de transcripción. En una realización, los ARNip se clasifican por una puntuación de PSSM y se seleccionan los ARNip clasificados colocados al menos a 100 bases de distancia en el transcrito. El des-solapamiento por contenido de GC establece una diferencia mínima en el contenido de GC. En una realización, la diferencia mínima en contenido de GC es de 1 %, 2 % o 5 %. Se permiten duplicados para parámetros de pocos valores, tales como el GC % de un oligo de 19 meros. El des-solapamiento por dímeros principales establece la probabilidad de todos o de una parte de los 16 dímeros principales posibles entre los ARNip seleccionados. En una realización, a cada uno de los 16 dímeros posibles se le asigna una puntuación de 1 a 16, y se utiliza un 0, 5 para seleccionar todos los cebadores principales posibles con la misma probabilidad, es decir, para distribuir los ARNip candidatos sobre todos los posibles valores de dímeros principales.

El des-solapamiento con diferentes parámetros puede combinarse.

En la etapa 105, la actividad inespecífica de un ARNip se evalúa de acuerdo con el método descrito en la Sección 5.2. Los alineamientos de cada ARNip con secuencias en cada una de una pluralidad de transcritos no diana se identifican y evalúan con una pmPSSM utilizando una Puntuacióncp calculada de acuerdo con la ecuación (6). Una Puntuacióncp por encima de un umbral determinado identifica la secuencia como una posible secuencia inespecífica. El transcrito que comprende la posible secuencia inespecífica se identifica como un posible transcrito inespecífico. El número total de dichos transcritos inespecíficos en el genoma o en una parte del genoma se utiliza como una medida de la especificidad de silenciamiento del ARNip. Se seleccionan uno o más ARNip que tienen menos transcritos inespecíficos.

En una realización, se exploran transcritos de genes utilizando FASTA con los parámetros: KTUP 6 -r 3/-7 -g -6 -f -6 -d 14000 -b 14000 -E 7000. Como se describe en la Sección 5.2., para cada alineamiento se determina una Puntuacióncp. La puntuación FASTA ponderada se utiliza para: (1) cuantificar la coincidencia de secuencia más cercana al ARNip candidato; y (2) contar el total de coincidencias con el ARNip candidato con puntuaciones ponderadas por encima de un umbral. El número total de dichos genes inespecíficos en el genoma o en una parte del genoma se utiliza después como una medida de la especificidad de silenciamiento del ARNip.

En una realización preferida, los ARNip seleccionados se someten a una segunda ronda de selección por variedad (etapa 106), y se vuelven a clasificar por sus puntuaciones de PSSM de composición de bases (etapa 107). El número deseado de ARNip se retiene desde la parte superior de esta clasificación final (etapa 108).

La divulgación también proporciona un método para seleccionar una pluralidad de ARNip para cada uno de una pluralidad de genes diferentes, alcanzando cada ARNip al menos un 75 %, al menos un 80 %, o al menos un 90 % de silenciamiento de su gen diana. El método descrito anteriormente se utiliza para seleccionar una pluralidad de ARNip para cada uno de una pluralidad de genes. Preferentemente, la pluralidad de ARNip consta de al menos 3, 5 o 10 ARNip. Preferentemente, la pluralidad de genes diferentes consta de al menos 100, 500, 1.000, 5.000, 10 000 o 30 000 genes diferentes.

La divulgación también proporciona una biblioteca de ARNip que comprende una pluralidad de ARNip para cada uno de una pluralidad de genes diferentes, cada ARNip alcanza al menos un 75 %, al menos un 80 %, o al menos un 90 % de silenciamiento de su gen diana. Las condiciones estándar son ARNip 100 nM, silenciamiento ensayado por TaqMan 24 horas después de la transfección. Preferentemente, la pluralidad de ARNip consta de al menos 3, al menos 5, o al menos 10 ARNip. Preferentemente, la pluralidad de genes diferentes consta de al menos 100, 500, 1.000, 5.000, 10 000 o 30 000 genes diferentes.

5.5. MÉTODOS Y COMPOSICIONES PARA EL ARN DE INTERFERENCIA Y ENSAYOS CON CÉLULAS

Junto con la presente invención puede utilizarse cualquier método convencional para el silenciamiento génico, por ejemplo, para llevar a cabo nuestro silenciamiento génico utilizando ARNip diseñados por un método descrito en la presente invención (véase, por ejemplo, Guo et al., 1995, Cell 81: 611 - 620; Fire et al, 1998, Nature 391: 806 - 811; Grant, 1999, Cell 96: 303 - 306; Tabara y col., 1999, Cell 99: 123 - 132; Zamore et al, 2000, Cell 101: 25 - 33; Bass, 2000, Cell 101: 235 - 238; Petcherski et al., 2000, Nature 405: 364 - 368; Elbashir et al., Nature 411: 494 - 498; Paddison. et al, Proc. Natl. Acad. Sci. USA 99: 1443 - 1448). En una realización, el silenciamiento génico se induce presentando la célula con el ARNip, imitando el producto de escisión de Dicer (véase, por ejemplo, Elbashir et al., 2001, Nature 411, 494-498; Elbashir et al., 2001, Genes Dev. 15, 188-200). Los dúplex de ARNip sintéticos mantienen la capacidad de asociarse con RISC y el silenciamiento directo de transcritos de ARNm. Los ARNip pueden sintetizarse químicamente, o derivar de la escisión de ARN bicatenario mediante Dicer recombinante. Las células pueden transfectarse con el ARNip utilizando un método convencional conocido en la técnica.

En una realización, la transfección con ARNip se lleva a cabo de la siguiente manera: un día antes de la transfección, 100 microlitros de células elegidas, por ejemplo, células HeLa de cáncer de cuello uterino (ATCC, Cat. No. CCL-2), cultivadas en DMEM/suero bovino fetal al 10 % (Invitrogen, Carlsbad, CA) hasta una confluencia de aproximadamente el 90 %, se siembran en una placa de cultivo tisular de 96 pocillos (Coming, Coming, NY) a 1.500 células/pocillo. Para cada transfección 85 microlitros de OptiMEM (Invitrogen) se mezclan con 5 microlitros de ARNip diluido en serie (Dharma on, Denver) a partir de una reserva de 20 micro molar. Para cada transfección, se mezclaron 5 microlitros de OptiMEM con 5 microlitros de reactivo de Oligofectamina (Invitrogen) y se incubaron durante 5 minutos a temperatura ambiente. La mezcla de OptiMEM/Oligofectamina de 10 microlitros se dispensó en cada tubo con la mezcla de OptiMEM/ARNip, se mezcló y se incubó durante 15-20 minutos a temperatura ambiente. Diez microlitros de la mezcla de transfección se dividió en alícuotas en cada pocillo de la placa de 96 pocillos y se incubó durante 4 horas a 37 °C y con CO_2 al 5 %.

En una realización, el ARN de interferencia se lleva a cabo utilizando un conjunto de ARNip. En una realización preferida, para transfectar las células se utiliza un conjunto de ARNip que contiene al menos k (k = 2, 3, 4, 5, 6 o 10) ARNip diferentes que se dirigen a un gen diana en diferentes regiones de secuencia. En otra realización preferida, para supertransfectar las células se utiliza un conjunto de ARNip que contiene al menos k (k = 2, 3, 4, 5, 6 o 10) ARNip diferentes que se dirigen a dos o más genes diana diferentes. En una realización preferida, la concentración total de ARNip del conjunto es aproximadamente la misma que la concentración de un solo ARNip cuando se utiliza individualmente, por ejemplo, 100 nM. Preferentemente, la concentración total del conjunto de ARNip es una concentración óptima para silenciar el gen diana deseado. Una concentración óptima es una concentración cuyo aumento adicional no aumenta sustancialmente el nivel de silenciamiento. En una realización, la concentración óptima es una concentración cuyo aumento adicional no aumenta el nivel de silenciamiento en más de un 5 %, 10 % o 20 %. En una realización preferida, la composición del conjunto, que incluye el número de ARNip diferentes en el conjunto y la concentración de cada ARNip diferente, se elige de tal manera que el conjunto de ARNip cause un silenciamiento menor del 30 %, 20 %, 10 % o 5 %, 1 %, 0,1 % o 0,01 % de cualquier gen inespecífico. En otra realización preferida, la concentración de cada ARNip diferente en el conjunto de diferentes ARNip es aproximadamente la misma. Aún en otra realización preferida, las concentraciones respectivas de diferentes ARNip en el conjunto son diferentes entre sí en menos del 5 %, 10 %, 20 % o 50 %. Aún en otra realización preferida, al menos un ARNip en el conjunto de diferentes ARNip constituye más del 90 %, 80 %, 70 %, 50 % o 20 % de la concentración total de ARNip en el conjunto. Aún en otra realización preferida, ninguno de los ARNip en el conjunto de diferentes ARNip constituye más del 90 %, 80 %, 70 %, 50 % o 20 % de la concentración total de ARNip en el conjunto. Aún en otras realizaciones, cada ARNip en el conjunto tiene una concentración que es más baja que la concentración óptima cuando se utiliza individualmente. En una realización preferida, cada ARNip diferente en el conjunto tiene una concentración que es más baja que la concentración del ARNip que es eficaz para alcanzar un silenciamiento de al menos 30 %, 50 %, 75 %, 80 %, 85 %, 90 % o 95 % cuando se utiliza en ausencia de otros ARNip o en ausencia de otros ARNip diseñados para silenciar el gen. En otra realización preferida, cada ARNip diferente en el conjunto tiene una concentración que causa un silenciamiento del gen menor del 30 %, 20 %, 10 % o 5 % cuando se utiliza en ausencia de otros ARNip o en ausencia de otros ARNip diseñados para silenciar el gen. En

una realización preferida, cada ARNip tiene una concentración que causa un silenciamiento del gen diana menor del 30 %, 20 %, 10 % o 5 % cuando se utiliza solo, mientras que la pluralidad de ARNip causa un silenciamiento del gen diana de al menos 80 % o 90 %.

Otro método para el silenciamiento de genes es introducir en una célula un ARNhp, ARN de horquilla pequeña (véase, por ejemplo, Paddison et al., 2002, Genes Dev. 16, 948-958; Brummelkamp et al., 2002, Science 296, 550-553; Sui, G. et al. 2002, Proc. Natl. Acad. Sci. USA 99, 5515-5520), que puede procesarse en las células en ARNip. En este método, una secuencia de ARNip deseada, se expresa a partir de un plásmido (o virus) como una repetición invertida con una secuencia de bucle intermedia para formar una estructura en horquilla. El transcrito de ARN resultante que contiene la horquilla se procesa posteriormente por Dicer para producir los ARNip para el silenciamiento. Los ARNhp basados en plásmidos pueden expresarse de manera estable en las células, permitiendo el silenciamiento génico prolongado en las células tanto *in vitro* como *in vivo*, por ejemplo, en animales (véase, McCaffrey y col., 2002, Nature 418, 38-39; Xia et al., 2002, Nat. Biotech., 20, 1006 - 1010; Lewis et al., 2002, Nat. Genetics 32, 107 - 108; Rubinson et al., 2003, Nat. Genetics 33, 401 - 406; Tiscornia et al., 2003, Proc. Natl. Acad. Sci. USA 100, 1844-1848). Por lo tanto, en una realización, se utiliza un ARNhp basado en plásmido.

En una realización preferida, los ARNhp se expresan a partir de vectores recombinantes introducidos de manera transitoria o estable en el genoma (véase, por ejemplo, Paddison et al, 2002, Genes Dev 16: 948-958; Sui et al., 2002, Proc Natl Acad Sci Sci. USA 99: 5515 - 5520; Yu et al., 2002, Proc Natl Acad Sci USA 99: 6047 - 6052; Miyagishi et al., 2002, Nat Biotechnol 20: 497 - 500, Paul et al., 2002, Nat Biotechnol 20. : 505 - 508; Kwak y col., 2003, / Pharmacol Sci 93: 214 - 217; Brummelkamp y col., 2002, Science 296: 550 - 553; Boden y col., 2003, Nucleic Acids Res 31: 5033 - 5038; Kawasaki et al, 2003, Nucleic Acids Res 31: 700 - 707). El ARNip que altera el gen diana puede expresarse (a través de un ARNhp) mediante cualquier vector adecuado que codifique el ARNhp. El vector también puede codificar un marcador que puede utilizarse para seleccionar clones en los que el vector, o una parte suficiente del mismo, está integrado en el genoma del hospedador de tal manera que se expresa el ARNhp. Para suministrar el vector en las células puede utilizarse cualquier método convencional conocido en la técnica. En una realización, las células que expresan ARNhp se generan transfectando células adecuadas con un plásmido que contiene el vector. Las células pueden ser seleccionadas por el marcador apropiado. Después se recogen los clones y se prueban para eliminarlos. En una realización preferida, se introduce una pluralidad de vectores recombinantes en el genoma de manera que el nivel de expresión del ARNip puede estar por encima de un valor determinado. Dicha realización es particularmente útil para silenciar genes cuyo nivel de transcripción es bajo en la célula. Para administrar el vector en las células puede utilizarse cualquier método convencional conocido en la técnica. En una realización, las células que expresan ARNhp se generan transfectando células adecuadas con un plásmido que contiene el vector. El marcador apropiado puede seleccionar las células. Después se recogen los clones y se someten a ensayo para determinar la atenuación (*knockdown*). En una realización preferida, en el genoma se introduce una pluralidad de vectores recombinantes de tal manera que el nivel de expresión del ARNip puede estar por encima de un valor determinado. Dicha realización es particularmente útil para el silenciamiento de genes cuyo nivel de transcripción es bajo en la célula.

En una realización preferida, la expresión del ARNhp está bajo el control de un promotor inducible de tal manera que el silenciamiento de su gen diana puede activarse cuando se desee. La expresión inducible de un ARNip es particularmente útil para el direccionamiento a genes esenciales. En una realización, la expresión del ARNhp está bajo el control de un promotor regulado que permite ajustar el nivel de silenciamiento del gen diana. Esto permite identificar células en las que el gen diana está parcialmente inactivado (*knocked out*). Como se usa en este documento, un "promotor regulado" se refiere a un promotor que puede activarse cuando está presente un agente inductor apropiado. Un "agente inductor" puede ser cualquier molécula que pueda utilizarse para activar la transcripción activando el promotor regulado. Un agente inductor puede ser, pero sin limitación, un péptido o polipéptido, una hormona, o una pequeña molécula orgánica. También puede utilizarse un análogo de un agente inductor, es decir, una molécula que activa el promotor regulado como lo hace el agente inductor. El nivel de actividad del promotor regulado inducido por diferentes análogos puede ser diferente, lo que permite una mayor flexibilidad para ajustar el nivel de actividad del promotor regulado. El promotor regulado en el vector puede ser cualquier sistema de regulación de la transcripción de mamíferos conocido en la técnica (véase, por ejemplo, Gossen et al., 1995, Science 268: 1766 - 1769; Lucas et al., 1992, Annu. Rev. Biochem. 61: 1131). ; Li et al., 1996, Cell 85: 319-329; Saez et al., 2000, Proc. Natl. Acad. Sci. USA 97: 14512-14517; y Pollock et al., 2000, Proc. Natl. Acad. Sci. USA 97: 13221 - 13226). En realizaciones preferidas, el promotor regulado se regula de una manera dependiente de la dosificación y/o del análogo. En una realización, el nivel de actividad del promotor regulado se ajusta a un nivel deseado mediante un método que comprende ajustar la concentración del agente inductor al que responde el promotor regulado. El nivel de actividad deseado del promotor regulado, tal como se obtiene aplicando una concentración particular del agente inductor, puede determinarse basándose en el nivel de silenciamiento deseado del gen diana.

En una realización, se utiliza un sistema de expresión génica regulada por tetraciclina (véase, por ejemplo, Gossen et al, 1995, Science 268: 1766 - 1769, Patente de Estados Unidos Nº 6.004.941). Un sistema regulado por tet (tetraciclina) utiliza componentes del sistema represor/operativo/inductor tet de procariotas para regular la expresión génica en células eucariotas. De este modo, la invención proporciona métodos para utilizar el sistema regulador tet para regular la expresión de un ARNhp unido a una o más secuencias operativas tet. Los métodos conllevan

introducir, en una célula, un vector que codifique una proteína de fusión que active la transcripción. La proteína de fusión comprende un primer polipéptido que se une a una secuencia operativa tet en presencia de tetraciclina, o de un análogo de tetraciclina, que se une operativamente a un segundo polipéptido que activa la transcripción en las células. Al modular la concentración de una tetraciclina, o de un análogo de tetraciclina, se regula la expresión del ARNhp unido al operador tet.

En otras realizaciones, para regular la expresión del ARNhp, puede utilizarse un sistema de expresión génica regulado por ecdisona (véase, por ejemplo, Saez et al., 2000, Proc. Natl. Acad. Sci. USA 97: 14512-14517), o un sistema de expresión génica regulado por el elemento de respuesta a glucocorticoides MMTV (véase, por ejemplo, Lucas et al., 1992, Annu. Rev. Biochem. 61: 1131).

En una realización, se utiliza el vector pRETRO-SUPER (pRS) que codifica un marcador de resistencia a puromicina e impulsa la expresión de ARNhp a partir de un promotor H1 (ARN Pol III). El plásmido pRS-ARNhp puede generarse mediante cualquier método convencional conocido en la técnica. En una realización, el plásmido pRS-ARNhp se desconvoluciona del conjunto de plásmidos de la biblioteca para un gen elegido mediante la transformación de bacterias con el conjunto y buscando clones que solo contengan el plásmido de interés. Preferentemente, se utiliza una secuencia de ARNip de 19 meros junto con cebadores directos e inversos adecuados para PCR específica de secuencia. Los plásmidos se identifican por PCR específica de secuencia y se confirman mediante secuenciación. Las células que expresan el ARNhp se generan transfectando células adecuadas con el plásmido pRS- ARNhp. Las células se seleccionan mediante el marcador apropiado, por ejemplo, puromicina, y se mantienen hasta que aparecen las colonias. Después se recogen los clones y se someten a ensayo para determinar la atenuación génica. En otra realización, un ARNhp se expresa mediante un plásmido, por ejemplo, un plásmido pRS- ARNhp. La atenuación génica por el plásmido pRS- ARNhp puede realizarse transfectando células utilizando Lipofectamine 2000 (Invitrogen).

En otro método más, los ARNip pueden suministrarse *in vivo* a un órgano o a un tejido de un animal, tal como un ser humano, (véase, por ejemplo, Song et al., 2003, Nat. Medicine 9, 347-351; Sorensen et al., 2003, J. Mol Biol. 327, 761-766; Lewis et al., 2002, Nat. Genetics 32, 107-108). En este método, al animal se le inyecta por vía intravenosa una solución de ARNip. Después, el ARNip puede alcanzar un órgano o un tejido de interés y reducir eficazmente la expresión del gen diana en el órgano o el tejido del animal.

Los ARNip también pueden suministrarse a un órgano o a un tejido utilizando una estrategia de terapia génica. Para suministrar el ARNip puede utilizarse cualquiera de los métodos de terapia génica disponibles en la técnica. Para revisiones generales de los métodos de terapia génica, véase Goldspiel et al., 1993, Clinical Pharmacy 12: 488-505; Wu y Wu, 1991, Biotherapy 3: 87-95; Tolstoshev, 1993, Ann. Rev. Pharmacol. Toxicol. 32: 573 - 596; Mulligan, 1993, Science 260: 926 - 932; y Morgan y Anderson, 1993, Ann. Rev. Biochem. 62: 191 - 217; mayo de 1993, TIBTECH 11(5): 155-215). En una realización preferida, el agente terapéutico comprende, como parte de un vector de expresión, un ácido nucleico que codifica el ARNip. En particular, dicho ácido nucleico tiene un promotor unido operativamente a la región codificante de ARNip, en la que el promotor es inducible o constitutivo y, opcionalmente, específico de tejido. En otra realización particular, se utiliza una molécula de ácido nucleico en la que la secuencia codificante de ARNip está flanqueada por regiones que promueven la recombinación homóloga en un sitio deseado en el genoma (véase, por ejemplo, Koller y Smithies, 1989, Proc. Natl. Acad. Sci. Sci. USA 86: 8932 - 8935; Zijlstra et al., 1989, Nature 342: 435 - 438).

En una realización específica, el ácido nucleico se administra directamente *in vivo*. Esto puede efectuarse mediante cualquiera de los numerosos métodos conocidos en la técnica, por ejemplo, construyéndolo como parte de un vector de expresión de ácido nucleico apropiado y administrándolo de manera que se convierta en intracelular, por ejemplo, mediante infección utilizando un vector retrovírico defectuoso o atenuado u otro vector vírico (véase la Patente de Estados Unidos Nº 4.980.286), o mediante inyección directa de ADN desprotegido (*naked*) o utilizando bombardeo con micropartículas (por ejemplo, una pistola de genes, Biolistic, Dupont), o revistiendo con lípidos o receptores de superficie celular o agentes transfectantes, encapsulación en liposomas, micropartículas o microcápsulas, o administrándolo enlazado a un péptido que se sabe que entra en el núcleo, administrándolo enlazado a un ligando sujeto a endocitosis mediada por receptor (véase, por ejemplo, Wu y Wu, 1987, J. Biol. Chem. 262: 4429-4432) (que puede utilizarse para dirigir tipos de células que expresan específicamente los receptores), etc. En otra realización, se puede formar un complejo de ácido nucleico-ligando en el que el ligando comprende un péptido vírico fusogénico que altera endosomas, permitiendo que el ácido nucleico impida la degradación lisosómica. En otra realización adicional, el ácido nucleico puede dirigirse *in vivo* para la captación y expresión específicas de células, dirigiéndose a un receptor específico (véanse, por ejemplo, las Publicaciones PCT WO 92/06180 del 16 de abril de 1992 (Wu et al.); WO 92/22635 del 23 de diciembre de 1992 (Wilson et al.); WO 92/20316 del 26 de noviembre de 1992 (Findeis et al.); WO 93/14188 del 22 de julio de 1993 (Clarke et al.), WO 93/20221 del 14 de octubre de 1993 (Young)).

Como alternativa, el ácido nucleico puede introducirse intracelularmente e incorporarse en el ADN de la célula hospedadora para su expresión, mediante recombinación homóloga (Koller y Smithies, 1989, Proc. Natl. Acad. Sci. USA 86: 8932 - 8935; Zijlstra et al., 1989, Nature 342: 435 - 438).

En una realización específica, se utiliza un vector vírico que contiene el ácido nucleico que codifica el ARNip. Por

ejemplo, puede utilizarse un vector retrovírico (véase Miller et al., 1993, Meth. Enzymol. 217: 581-599). Estos vectores retrovíricos se han modificado para delecionar secuencias retrovíricas que no son necesarias para el empaquetamiento del genoma vírico y la integración en el ADN de la célula hospedadora. El ácido nucleico que codifica el ARNip que se utilizará en la terapia génica se clona en el vector, lo que facilita el suministro del gen a un paciente. Se pueden encontrar más detalles sobre vectores retrovíricos en Boesen et al., 1994, Biotherapy 6: 291-302, que describen el uso de un vector retrovírico para suministrar el gen *mdr1* a células madre hematopoyéticas con el fin de hacer que las células madre sean más resistentes a la quimioterapia. Otras referencias que ilustran el uso de vectores retrovíricos en terapia génica son: Clowes et al., 1994, J. Clin. Invertir. 93: 644-651; Kiem et al., 1994, Blood 83: 1467-1473; Salmons y Gunzberg, 1993, Human Gene Therapy 4: 129-141; y Grossman y Wilson, 1993, Curr. Opin. Genet y Devel. 3:110-114.

Los adenovirus son otros vectores víricos que pueden utilizarse en la terapia génica. Los adenovirus son vehículos especialmente atractivos para suministrar genes al epitelio respiratorio. Los adenovirus infectan de manera natural el epitelio respiratorio donde causan una enfermedad leve. Otras dianas para los sistemas de suministro basados en adenovirus son el hígado, el sistema nervioso central, las células endoteliales y el músculo. Los adenovirus tienen la ventaja de que tienen la capacidad de infectar células que no se dividen. Kozarsky y Wilson (1993, Current Opinion in Genetics and Development 3: 499-503) presentan una revisión de la terapia génica basada en adenovirus. Bout et al. (1994, Human Gene Therapy 5: 3-10) demostraron el uso de vectores de adenovirus para transferir genes al epitelio respiratorio de monos *rhesus*. Se pueden encontrar otros ejemplos del uso de adenovirus en la terapia génica en Rosenfeld et al., 1991, Science 252: 431-434; Rosenfeld et al., 1992, Cell 68: 143 - 155; y Mastrangeli et al., 1993, J. Clin. Invertir. 91: 225-234. En terapia génica también pueden utilizarse virus adenoasociados (AAV, por sus siglas en inglés) (Walsh et al., 1993, Proc. Soc. Exp. Biol. Med. 204: 289-300).

El grado de silenciamiento puede determinarse utilizando cualquier método convencional de cuantificación de ARN o proteínas, conocido en la técnica. Por ejemplo, la cuantificación del ARN puede realizarse utilizando PCR en tiempo real, por ejemplo, utilizando el reactivo de ensayo desarrollado previamente por TaqMan de AP Biosystems (nº 4319442). La sonda de cebador para el gen apropiado puede diseñarse utilizando cualquier método convencional conocido en la técnica, por ejemplo, utilizando el programa informático Primer Express. Los valores de ARN pueden normalizarse a ARN para la actina (n. ° 4326315). Los niveles de proteína pueden cuantificarse mediante citometría de flujo después de la tinción con un anticuerpo apropiado y un anticuerpo secundario marcado. Los niveles de proteína también pueden cuantificarse mediante transferencia Western de lisados celulares con anticuerpos monoclonales apropiados, seguido de análisis de obtención de imágenes Kodak de inmunotransferencia quimioluminiscente. Los niveles de proteína también pueden normalizarse a los niveles de actina.

Los efectos del silenciamiento génico en una célula pueden evaluarse mediante cualquier ensayo conocido. Por ejemplo, el crecimiento celular puede analizar utilizando cualquier ensayo adecuado de proliferación o inhibición del crecimiento conocido en la técnica. En una realización preferida, se utiliza un ensayo de proliferación con MTT (véase, por ejemplo, van de Loosdrechet, et al., 1994, J. Immunol. Methods 174: 311 - 320; Ohno et al., 1991, J. Immunol. Methods 145: 199-203; Ferrari et al., 1990, J. Immunol. Methods 131: 165-172; Alley et al., 1988, Cancer Res. 48: 589-601; Carmichael et al., 1987, Cancer Res. 47: 936- 942; Gerlier et al., 1986, J. Immunol. Methods 65: 55-63; Mosmann, 1983, J. hnmunological Methods 65: 55-63) para analizar el efecto de uno o más agentes en la inhibición del crecimiento de células. Las células se tratan con concentraciones elegidas de uno o más agentes candidatos durante un período de tiempo elegido, por ejemplo, durante 4 a 72 horas. Las células se incuban después con una cantidad adecuada de bromuro de 3-(4,5-dimetiltiazol-2-il)-2,5-difeniltetrazolio (MTT) durante un período de tiempo elegido, por ejemplo, 1-8 horas, de tal manera que las células viables transformen el MTT en un depósito intracelular de formazán insoluble. Después de eliminar el exceso de MTT contenido en el sobrenadante, se añade un disolvente de MTT adecuado, por ejemplo, una solución de DMSO, para disolver el formazán. La concentración de MTT, que es proporcional al número de células viables, se mide después determinando la densidad óptica, por ejemplo, a 570 nm. Se puede ensayar una pluralidad de diferentes concentraciones del agente candidato para permitir la determinación de las concentraciones del agente o agentes candidatos que causan una inhibición del 50 %.

En otra realización preferida, se utiliza un ensayo con alamarBlue™ de proliferación celular para explorar uno o más agentes candidatos que puedan utilizarse para inhibir el crecimiento de células (véase, por ejemplo, Page et al., 1993, Int. J. Oncol. 3: 473-476).) Un ensayo con alamarBlue™ mide la respiración celular y la utiliza como una medida del número de células vivas. El entorno interno de las células en proliferación es más reducido que el de las células que no proliferan. Por ejemplo, las proporciones de NADPH / NADP, FADH / FAD, FMNH / FMN y NADH / NAF aumentan durante la proliferación. El alamarBlue puede reducirse mediante estos productos intermedios metabólicos y, por lo tanto, puede utilizarse para controlar la proliferación celular. El número de células de una muestra tratada según lo medido por alamarBlue, puede expresarse en porcentaje con relación al de una muestra de control no tratada. La reducción de alamarBlue puede medirse mediante absorción o espectroscopía de fluorescencia. En una realización, la reducción de alamarBlue se determina por la absorbancia y se calcula como porcentaje reducido utilizando la ecuación:

$$\% \text{ Reducido} = \frac{(\varepsilon_{ox}\lambda_2)(A\lambda_1) - (\varepsilon_{ox}\lambda_1)(A\lambda_2)}{(\varepsilon_{red}\lambda_1)(A'\lambda_2) - (\varepsilon_{red}\lambda_2)(A'\lambda_1)} \times 100 \qquad (8)$$

en la que:

λ_1 = 570 nm

λ_2 = 600 nm

($\varepsilon_{red}\lambda_1$) = 155,677 (Coeficiente de extinción molar de alamarBlue reducido a 570 nm)

($\varepsilon_{red}\lambda_2$) = 14,652 (Coeficiente de extinción molar de alamarBlue reducido a 600 nm)

($\varepsilon_{ox}\lambda_1$) = 80,586 (Coeficiente de extinción molar de alamarBlue oxidado a 570 nm)
($\varepsilon_{rox}\lambda_2$) = 117,216 (Coeficiente de extinción molar de alamarBlue oxidado a 600 nm)

(A λ_1) = Absorbancia de los pocillos de ensayo a 570 nm

(A λ_2) = Absorbancia de los pocillos de ensayo a 600 nm

(A'λ_1) = Absorbancia de los pocillos de control negativos que contienen medio y alamarBlue, pero en los que no se han añadido células, a 570 nm.

(A'λ_2)= Absorbancia de los pocillos de control negativos que contienen medio y alamarBlue, pero en los que no se han añadido células, a 600 nm, Preferentemente, el % Reducido de los pocillos que no contienen células se restó del % Reducido de pocillos que contenían muestras para determinar el % Reducido por encima del fondo.+-

El análisis del ciclo celular puede llevarse a cabo utilizando métodos convencionales conocidos en la técnica. En una realización, el sobrenadante de cada pocillo se combina con las células que se han recogido mediante tripsinización. La mezcla se centrifuga después a una velocidad adecuada. A continuación, las células se fijan, por ejemplo, con etanol al 70 % enfriado con hielo, durante un período de tiempo adecuado, por ejemplo, ~ 30 minutos. Las células fijadas pueden lavarse una vez con PBS y suspenderse de nuevo, por ejemplo, en 0,5 ml de PBS que contiene yoduro de propidio (10 microgramos/ml) y RNasa A (1 mg/ml), y se incuban a una temperatura adecuada, por ejemplo, a 37 °C., durante un período de tiempo adecuado, por ejemplo, 30 min. Después, se realiza análisis de citometría de flujo utilizando un citómetro de flujo. En una realización, la población de células Sub-G1 se utiliza como una medida de muerte celular. Por ejemplo, se dice que las células se han sensibilizado a un agente si la población de células Sub-G1 de la muestra tratada con el agente es mayor que la población de células Sub-G1 de la muestra no tratada con el agente.

5.6. SISTEMAS Y MÉTODOS DE IMPLEMENTACIÓN.

Los métodos analíticos de la presente invención pueden implementarse preferentemente utilizando un sistema informático, tal como el sistema informático descrito en esta sección, de acuerdo con los siguientes programas y métodos. Dicho sistema informático también puede almacenar y tratar preferentemente señales medidas obtenidas en diversos experimentos que pueden utilizarse mediante un sistema informático implementado con los métodos analíticos de esta invención. En consecuencia, dichos sistemas informáticos también se consideran parte de la presente invención.

En la FIG. 12 se ilustra un ejemplo de un sistema informático adecuado para implementar los métodos analíticos de esta invención. En esta figura se ilustra el sistema informático 1201 que comprende componentes internos y que está vinculado a componentes externos. Los componentes internos de este sistema informático incluyen uno o más elementos procesadores 1202 interconectados con una memoria principal 1203. Por ejemplo, el sistema informático 1201 puede ser un procesador basado en Intel Pentium IV® de 2 GHz o mayor velocidad de reloj y con una memoria principal de 256 MB o mayor. En una realización preferida, el sistema informático 1201 es un conjunto de una pluralidad de ordenadores que comprende un "nodo" principal y ocho "nodos" hermanos, teniendo cada nodo una unidad de procesamiento central ("CPU", *Central Processing Unit*). Además, el conjunto también comprende al menos 128 MB de memoria de acceso aleatorio ("RAM", *Random Access Memory*) en el nodo principal y al menos 256 MB de RAM en cada uno de los ocho nodos hermanos. Por lo tanto, los sistemas informáticos de la presente invención no están limitados a los que consisten en una sola unidad de memoria o una sola unidad procesadora.

Los componentes externos pueden incluir un almacenamiento masivo 1204. Este almacenamiento masivo puede ser uno o más discos duros que generalmente se empaquetan junto con el procesador y la memoria. Dicho disco duro tiene normalmente una capacidad de almacenamiento de 10 GB o mayor, y más preferentemente, tiene una capacidad de almacenamiento de al menos 40 GB. Por ejemplo, en una realización preferida, descrita

anteriormente, en la que un sistema informático de la invención comprende varios nodos, cada nodo puede tener su propio disco duro. El nodo principal tiene preferentemente un disco duro con una capacidad de almacenamiento de al menos 10 GB, mientras que cada nodo hermano tiene preferentemente un disco duro con una capacidad de almacenamiento de al menos 40 GB. Un sistema informático de la invención puede comprender además otras unidades de almacenamiento masivo que incluyen, por ejemplo, una o más unidades de disquete, una o más unidades de CD-ROM, una o más unidades de DVD o una o más unidades de cinta de audio digital (DAT, *Digital Audio Tape)*.

Otros componentes externos incluyen normalmente un dispositivo de interfaz de usuario 1205, que más normalmente es un monitor y un teclado junto con un dispositivo de entrada gráfica 1206 tal como un "ratón". El sistema informático también está por lo general vinculado a un enlace de red 1207 que puede ser, por ejemplo, parte de una red de área local ("LAN", *Local Area Network*), a otros sistemas informáticos locales y/o a parte de una red de área amplia ("WAN", *Wide Area Network*), tal como Internet, que está conectado a otros sistemas informáticos remotos. Por ejemplo, en la realización preferida, comentada anteriormente, en la que el sistema informático comprende una pluralidad de nodos, cada nodo está preferentemente conectado a una red, preferentemente una red NFS (sistema de archivos en red, *Network File System)*, de modo que los nodos del sistema informático se comunican entre sí y, opcionalmente, con otros sistemas informáticos por medio de la red y, por lo tanto, pueden compartir datos y tareas de procesamiento entre sí.

Cargados en la memoria durante el funcionamiento de dicho sistema informático hay varios componentes de programación (*software*) que también se muestran esquemáticamente en la FIG. 12. Los componentes de programación comprenden tanto componentes de programación, que son estándar en la técnica, como componentes que son especiales para la presente invención. Por lo general, estos componentes de programación se almacenan en un almacenamiento masivo tal como el disco duro 1204, pero también pueden almacenarse en otros medios legibles por ordenador, incluyendo, por ejemplo, uno o más disquetes, uno o más CD-ROM, uno o más DVD o uno o más DAT. El componente de programación 1210 representa un sistema operativo que se encarga de gestionar el sistema informático y sus interconexiones de red. El sistema operativo puede ser, por ejemplo, de la familia de Microsoft Windows™ tal como Windows 95, Windows 98, Windows NT, Windows 2000 o Windows XP. Como alternativa, el programa operativo puede ser un sistema operativo Macintosh, un sistema operativo UNIX o un sistema operativo LINUX. Los componentes de programación 1211 comprenden lenguajes y funciones habituales que están presentes preferentemente en el sistema para ayudar a los programas que implementan métodos específicos para la presente invención. Los lenguajes que pueden utilizarse para programar los métodos analíticos de la invención incluyen, por ejemplo, C y C ++, FORTRAN, PERL, HTML, JAVA y cualquiera de los lenguajes de comandos Shell UNIX o LINUX, tal como el lenguaje script de shell C. Los métodos de la invención también pueden programarse o modelarse en paquetes de programación matemática que permiten la entrada simbólica de ecuaciones y la especificación de alto nivel de procesamiento, incluidos los algoritmos específicos que se utilizarán, liberando así a un usuario de la necesidad de programar procesalmente ecuaciones y algoritmos individuales. Dichos paquetes incluyen, por ejemplo, Matlab de Mathworks (Natick, MA), Mathematica de Wolfram Research (Champaign, IL) o S-Plus de MathSoft (Seattle, WA).
El componente de programación 1212 comprende cualquier método analítico de la presente invención descrito anteriormente, preferentemente programado en un paquete de lenguaje o símbolos de procedimiento. Por ejemplo, el componente de programación 1212 incluye preferentemente programas que hacen que el procesador implemente pasos para aceptar una pluralidad de señales medidas y almacenar las señales medidas en la memoria. Por ejemplo, el sistema informático puede aceptar señales medidas introducidas manualmente por un usuario (por ejemplo, mediante la interfaz de usuario). Sin embargo, más preferentemente, los programas hacen que el sistema informático recupere las señales medidas de una base de datos. Dicha base de datos puede almacenarse en un almacenamiento masivo (por ejemplo, un disco duro) u otro medio legible por ordenador y cargarse en la memoria del ordenador, o el sistema informático puede acceder al compendio por medio de la red 1207.

Además de los ejemplos de estructuras de programas y de los sistemas informáticos descritos en este documento, otras estructuras de programas y sistemas informáticos alternativos serán fácilmente evidentes para los expertos en la materia. Por lo tanto, dichos sistemas alternativos, que no se apartan del sistema informático y de las estructuras de programas descritos anteriormente, están destinados a incluirse en las reivindicaciones adjuntas.

6. EJEMPLOS

Como ilustración de la presente invención se presentan los siguientes ejemplos y, de ninguna manera, pretenden limitarla.

6.1. Ejemplo 1: Diseño de ARNip para obtener una eficacia de silenciamiento alta

Se construyó una biblioteca de ARNip dirigidos a más de 700 genes. Los ARNip de la biblioteca se diseñaron utilizando una estrategia "convencional", basada en una combinación de principios de diseño limitados disponibles en la bibliografía científica (Elbashir et al., 2001, Nature 411: 494-8) y un método para predecir efectos diana por puntuación de similitud de secuencia como se describe en la Sección 5.2. Se analizó un conjunto de 377 ARNip mediante análisis de Taqman para determinar su capacidad para silenciar a sus respectivos genes diana. El

conjunto de 377 ARNip se enumera en la Tabla II. La Tabla II enumera la siguiente información de los 377 ARNip: el número ID del ARNip, el número de registro del gen diana, la posición inicial de la secuencia diana, la secuencia diana, el % de silenciamiento, el conjunto al que pertenece (es decir, capacitación o ensayo) en el Conjunto 1, el conjunto al que pertenece en el Conjunto 2, y la SEC ID NO. Los resultados de este análisis mostraron que la mayoría de los ARNip silenciaron satisfactoriamente a sus genes diana (mediana de silenciamiento, ~75 %), pero los ARNip individuales aún mostraban un amplio intervalo de rendimiento de silenciamiento. La buena (o mala) capacidad de silenciamiento no se asoció consecuentemente con ninguna base particular en ninguna posición, con un contenido global de GC, con la posición de la secuencia de ARNip dentro del transcrito diana, o con el corte y empalme alternativo de transcritos diana.

Utilizando una estrategia clasificadora se exploró la posible relación entre el silenciamiento del gen diana y la composición de bases, la termodinámica y la estructura secundaria del ARNip y las secuencias diana. Los ARNip se dividieron en grupos que contenían los que tenían una capacidad de silenciamiento inferior a la mediana (ARNip "malos") y los que tenían una capacidad de silenciamiento mediana o mejor (ARNip "buenos"). Se evaluaron diversas medidas para determinar su capacidad para distinguir ARNip buenos y malos, incluida la composición de bases en ventanas de la secuencia dúplex de ARNip de 19 meros y la región diana flanqueante, predicciones de estructuras secundarias por diversos programas y propiedades termodinámicas. Estas pruebas revelaron que la eficacia del ARNip se correlacionaba bien con el ARNip y la composición de bases del gen diana, pero mal con las predicciones de la estructura secundaria y las propiedades termodinámicas. En particular, el contenido de GC de los ARNip buenos difería sustancialmente del de los ARNip malos de una manera específica de la posición (Figuras 1-3). Por ejemplo, no se observó que los dúplex de ARNip buenos estuvieran asociados con ninguna secuencia particular, pero tendían a ser ricos en GC en el extremo 5' y pobres en GC en el extremo 3'. Los datos indican que un dúplex de ARNip bueno estimula la interacción preferencial de la cadena antisentido al ser pobre en GC en su extremo 3' y desmotiva la interacción de la cadena en sentido al ser rica en GC en su extremo 5'. Los datos demuestran además que las preferencias de secuencia específicas de posición se extienden más allá de los límites de la secuencia diana de ARNip en la(s) secuencia(s) adyacente(s). Esto sugiere que durante el silenciamiento del ARN, las etapas que no sean desenrollar el dúplex de ARNip se ven afectadas por las preferencias de composición de bases específicas de posición.

La diferencia de contenido de GC entre los ARNip buenos y malos, mostrada en las Figs. 1 y 2, se utilizó para desarrollar métodos para seleccionar ARNip buenos. Los mejores resultados se obtuvieron con una estrategia de matriz de puntuación específica de posición (PSSM). La PSSM proporciona pesos para GC, A o U en cada posición en la cadena en sentido de la secuencia del gen diana desde 10 bases cadena arriba del inicio hasta 10 bases cadena abajo del extremo del dúplex de ARNip. Los datos de eficacia de ARNip se dividieron en dos conjuntos, uno para utilizarlo en un ensayo de capacitación y el otro en un ensayo independiente. Se utilizó un algoritmo de búsqueda de mutación ascendente aleatoria para optimizar simultáneamente los pesos de cada base en cada posición de la PSSM. El criterio de optimización fue el coeficiente de correlación entre el silenciamiento del ARNip diana y su puntuación PSSM. Se promediaron varias ejecuciones de optimización en el conjunto de datos de capacitación para completar cada PSSM. Después, cada PSSM se analizó en el conjunto (ensayo) independiente de los ARNip. En la Figura 2 se muestra el rendimiento de dos PSSM en sus conjuntos de datos de capacitación y ensayo.

Se desarrolló un método de diseño de ARNip basado en una matriz de puntuación específica de posición (PSSM). Se utilizó un esquema de puntuación para predecir la eficacia de los oligos de ARNip. La puntuación es una suma ponderada de 39 bases (10 bases cadena arriba del oligo de 19 meros, 19 bases en el propio ARNip, y 10 bases cadena abajo) calculada de la siguiente manera:

$$Puntuación = \sum_{i=1}^{39} \ln(E_i / p_i)$$

en la que P_i es igual a la probabilidad aleatoria de cualquier base, es decir, 0,25, y E_i el peso asignado a la base A, U, G o C en la posición *i*. Por lo tanto, es necesario asignar y optimizar un total de 117 pesos (39 posiciones por 3 tipos de bases, G o C, A, U).

Para optimizar los pesos, se utilizó un algoritmo de búsqueda de mutación aleatoria en escalada (RMHC, siglas del inglés *Random-Mutation Hill Climbing*) basándose en un conjunto de oligos de capacitación y los perfiles resultantes se aplicaron a un conjunto de ensayo, siendo los criterios de optimización el coeficiente de correlación entre los niveles de atenuación (*knock-down,* KD) de los oligos y las puntuaciones de PSSM calculadas. El parámetro para medir la efectividad de la capacitación y ensayo es la tasa de detección falsa (FDR) agregada basada en la curva ROC, y se calcula como el promedio de las puntuaciones de FDR del 33 % de los oligos principales ordenados por las puntuaciones dadas por el indicador de capacitación. Al calcular las puntuaciones de FDR, aquellos oligos con niveles de silenciamiento inferiores a la mediana se consideran falsos, y aquellos con niveles de silenciamiento superiores a la mediana se consideran verdaderos.

Se utilizaron diferentes criterios para dividir los datos de rendimiento de ARNip existentes en conjuntos de capacitación y ensayo. El mayor obstáculo para una división ideal es que la gran mayoría de los oligonucleótidos de

ARNip están diseñados con el método convencional, que requiere un dímero AA inmediatamente antes de la secuencia del oligo de 19 meros. Más tarde se descubrió que esta limitación era perjudicial en lugar de útil para el proceso de diseño y se anuló. Para limitar la influencia de esto en el procedimiento de capacitación, se utilizaron diversas divisiones y para asignar puntuaciones a los oligos de ensayo se combinó más de un indicador de capacitación, es decir, PSSM (en lugar de indicadores únicos).

Finalmente, se construyó un procedimiento de diseño de oligo de ARNip de vanguardia (también conocido como "procedimiento en fase de desarrollo", (*pipeline*)). Este incorpora el procedimiento de predicción inespecífico y dos conjuntos de indicadores de eficacia del oligo de ARNip capacitados y ensayados en diferentes conjuntos de datos. Se seleccionó y analizó un total de 30 oligos de ARNip (6 oligos para cada uno de los 5 genes). Los resultados fueron significativamente mejores que cualquiera de los de la fase en desarrollo anteriormente existentes.

Los resultados iniciales de capacitación y ensayo mostraron que la PSSM es muy eficaz para predecir la eficacia diana los oligos de ARNip. Normalmente, las puntuaciones FDR agregada para la capacitación están entre 0,02 y 0,08, y las de ensayo entre 0,05 y 0,10. Como referencia, las predicciones aleatorias tienen una FDR agregada promedio de 0,17, con una desviación típica de 0,02 (datos calculados con 10 000 predicciones generadas aleatoriamente). La Fig. 3 ilustra curvas ROC típicas, generadas por un conjunto de aproximadamente 200 indicadores optimizados aleatoriamente. Se puede observar que el rendimiento del conjunto de capacitación es mejor que el del ensayo, lo que apenas es sorprendente. Ambas curvas son significativamente mejores que al azar.

La Fig. 5 ilustra los perfiles de secuencia resultantes de la capacitación y ensayo en varios conjuntos de oligómeros diferentes. Este perfil ilustra que las bases G o C son muy preferidas al principio, es decir, en el extremo 5', y muy desfavorecidas al final, es decir, en el extremo 3', de la secuencia de 19 meros. Para confirmar esta observación, se calcularon los niveles promedio de atenuación de oligos que comenzaban y terminaban con G/C o A/U, y aquellos oligos que comenzaban con G/C y terminaban con A/U tenían mejor rendimiento, muy superior al de las otras tres categorías. Simplemente comparando los pesos en diferentes posiciones, un oligo de 19 meros que tenía una secuencia de GCGTTAATGTGATAATATA (SEQ ID NO: 1) y los oligos que eran más similares a esta secuencia, se identificaban como un ARNip que podía tener una alta eficacia de silenciamiento.

El método de diseño incorporó ambas PSSM mostradas en la FIG. 3 porque la combinación dio un mejor rendimiento en comparación con el uso de una sola PSSM. El método de diseño de ARNip mejorado seleccionó oligonucleótidos en base a 4 principios: composición de bases, identidad inespecífica, posición en el transcrito y variedad de secuencias. Se eliminaron ciertos oligonucleótidos que contenían la secuencia de características tales como regiones no traducidas, repeticiones o ejecuciones homopoliméricas. Los oligonucleótidos restantes se clasificaron por sus puntuaciones de PSSM. Los oligonucleótidos de alto rango se seleccionaron por variedad en contenido de GC, en posición de inicio y en las dos bases cadena arriba del dúplex de ARNip de 19 meros. Los oligonucleótidos seleccionados se filtraron después para la actividad prevista inespecífica, que se calculó como una puntuación de alineamiento FASTA ponderada de posición. Los oligonucleótidos restantes se clasificaron según las puntuaciones de PSSM, se sometieron a una segunda ronda de selección de variedad y finalmente se volvieron a clasificar según sus puntuaciones de PSSM. El número deseado de ARNip se retuvo desde la parte superior de esta clasificación final.

El método mejorado se comparó con el método convencional mediante ensayos paralelos de nuevos ARNip seleccionados por cada uno de ellos. En la Figura 3 se muestran los resultados obtenidos con tres ARNip seleccionados por cada método. Los ARNip diseñados por el algoritmo mejorado mostraron una mejor eficacia media (88 %, en comparación con 78 % para ARNip según el método convencional) y fueron más uniformes en su rendimiento. La distribución de las eficacias de silenciamiento del algoritmo mejorado para los ARNip fue significativamente mejor que la del método convencional para los ARNip para los mismos genes ($p = 0,004$, prueba de suma de rangos de Wilcoxon).

Los resultados de ensayo de 30 oligos experimentales utilizando la nueva fase de desarrollo demostraron ser satisfactorios. La Tabla III enumera los 30 ARNip. Anteriormente, un diseño de ARNip con el método convencional, tenía un nivel medio de silenciamiento del 75 %. De los 30 oligos experimentales, 28 tuvieron niveles de silenciamiento iguales o mejores que 75 %, 26 mejores que o iguales a 80 % y 37 % mejores que 90 %, en comparación con solo un 10 % mejor que 90 % con el método convencional. Dos genes diana (KIF14 e IGF1R) habían sido muy difíciles de silenciar por los ARNip, logrando anteriormente oligos previos, niveles de silenciamiento solo del 40 % al 70 % y no superiores al 80 %. Los 12 nuevos oligos dirigidos a estos genes lograron niveles de silenciamiento de al menos 80 % y 6 lograron niveles de 90 %. Los dos oligos de entre los 30 oligos que tenían un nivel de silenciamiento menor del 75 %, resultaron estar dirigidos a un exón que era exclusivo de una secuencia de transcripción diana, pero faltaba en todas las demás formas de corte y empalme alternativas del mismo gen. Por lo tanto, la anomalía de estos dos oligos se debió a una secuencia de entrada incorrecta en lugar de al método de PSSM. Por lo tanto, cuando se proporcionan las secuencias de entrada apropiadas, la nueva fase de desarrollo parece que tiene la capacidad de escoger oligos que puedan atenuar genes diana en al menos un 75 % para el 100 % de los genes diana.

Tabla II Biblioteca de 377 ARNip

BioID	número de registro	posición de inicio	secuencia de 19 meros	% de silenciamiento	Conjunto 1	Conjunto 2	SEQ ID NO
31	NM_000075	437	TGTTGTCCGGCTGATGGAC	27,0	Capacitación	Capacitación	2
36	NM_001813	1036	ACTCTTACTGCTCTCCAGT	86,1	Ensayo	Capacitación	3
37	NM_001813	1278	CTTAACACGGATGCTGGTG	60,1	Ensayo	Capacitación	4
38	NM_001813	3427	GGAGAGCTTTCTAGGACCT	88,0	Ensayo	Capacitación	5
39	NM_004073	192	AGTCATCCCGCAGAGCCGC	55,0	Capacitación	Capacitación	6
40	NM_004073	1745	ATCGTAGTGCTTGTACTTA	70,0	Capacitación	Capacitación	7
41	NM_004073	717	GGAGACGTACCGCTGCATC	65,0	Capacitación	Capacitación	8
42	AK092024	437	GCAGTGATTGCTCAGCAGC	93,0	Capacitación	Capacitación	9
43	NM_030932	935	GAGTTTACCGACCACCAAG	81,0	Capacitación	Capacitación	10
44	NM_030932	1186	TGCGGATGCCATTCAGTGG	35,0	Capacitación	Capacitación	11
45	NM_030932	1620	CACGGTTGGCAGAGTCTAT	73,0	Capacitación	Capacitación	12
49	U53530	169	GCAAGTTGAGCTCTACCGC	59,0	Capacitación	Capacitación	13
50	U53530	190	TGGCCAGCGCTTACTGGAA	75,0	Capacitación	Capacitación	14
64	NM_006101	1623	GTTCAAAAGCTGGATGATC	79,0	Ensayo	Capacitación	15
65	NM_006101	186	GGCCTCTATACCCCTCAAA	74,4	Ensayo	Capacitación	16
66	NM_006101	968	AGAACCGAATCGTCTAGAG	80,3	Ensayo	Capacitación	17
67	NM_000859	253	CACGATGCATAGCCATCCT	25,0	Capacitación	Capacitación	18
68	NM_000859	1075	CAGAGACAGAATCTACACT	45,0	Capacitación	Capacitación	19
69	NM_000859	1720	CAACAGAAGGTTGTCTTGT	50,0	Capacitación	Capacitación	20
70	NM_000859	2572	TTGTGTGTGGGACCGTAAT	80,0	Capacitación	Capacitación	21
71	NM_000875	276	GCTCACGGTCATTACCGAG	63,9	Capacitación	Capacitación	22
72	NM_000875	441	CCTGAGGAACATTACTCGG	0,0	Capacitación	Capacitación	23
73	NM_000875	483	TGCTGACCTCTGTTACCTC	50,0	Capacitación	Capacitación	24
74	NM_000875	777	CGACACGGCCTGTGTAGCT	58,0	Capacitación	Capacitación	25
75	NM_000875	987	CGGCAGCCAGAGCATGTAC	63,0	Capacitación	Capacitación	26
76	NM_000875	1320	CCAGAACTTGCAGCAACTG	70,0	Capacitación	Capacitación	27
81	NM_000875	351	CCTCACGGTCATCCGCGGC	0,0	Capacitación	Capacitación	28
83	NM_000875	387	CTACGCCCTGGTCATCTTC	32,0	Capacitación	Capacitación	29
84	NM_000875	417	TCTCAAGGATATTGGGCTT	54,0	Capacitación	Capacitación	30
85	NM_000875	423	GGATATTGGGCTTTACAAC	71,0	Capacitación	Capacitación	31
86	NM_000875	450	CATTACTCGGGGGGCCATC	53,0	Capacitación	Capacitación	32
87	NM_000875	481	AATGCTGACCTCTGTTACC	54,6	Capacitación	Capacitación	33
117	NM_004523	1689	CTGGATCGTAAGAAGGCAG	74,7	Capacitación	Ensayo	34
118	NM_004523	484	TGGAAGGTGAAAGGTCACC	16,0	Capacitación	Ensayo	35
119	NM_004523	802	GGACAACTGCAGCTACTCT	84,1	Capacitación	Ensayo	36
139	NM_002358	219	TACGGACTCACCTTGCTTG	83,0	Capacitación	Capacitación	37
144	NM_001315	779	GTATATACATTCAGCTGAC	78,5	Capacitación		38
145	NM_001315	1080	GGAACACCCCCCGCTTATC	27,2	Capacitación		39
146	NM_001315	1317	GTGGCCGATCCTTATGATC	81,3	Capacitación		40
152	NM_001315	607	ATGTGATTGGTCTGTTGGA	95,0	Capacitación		41
153	NM_001315	1395	GTCATCAGCTTTGTGCCAC	92,0	Capacitación		42
154	NM_001315	799	TAATTCACAGGGACCTAAA	82,0	Capacitación		43
155	NM_001315	1277	TGCCTACTTTGCTCAGTAC	95,0	Capacitación		44
193	NM_001315	565	CCTACAGAGAACTGCGGTT	90,0	Capacitación		45
190	NM_001315	763	TTCTCCGAGGTCTAAAGTA	87,0	Capacitación		46

192	NM_001315	1314	CCAGTGGCCGATCCTTATG	89,0	Capacitación	47
194	NM_001315	1491	GGCCTTTTCACGGGAACTC	97,0	Capacitación	48
201	NM_016195	2044	CTGAAGAAGCTACTGCTTG	80,3	Ensayo	Capacitación 49
202	NM_016195	4053	GACATGCGAATGACACTAG	75,9	Ensayo	Capacitación 50
203	NM_016195	3710	AGAGGAACTCTCTGCAAGC	84,7	Ensayo	Capacitación 51
204	NM_014875	4478	AAACTGGGAGGCTACTTAC	93,0	Ensayo	Capacitación 52
205	NM_014875	1297	ACTGACAACAAAGTGCAGC	37,0	Ensayo	Capacitación 53
206	NM_014875	5130	CTCACATTGTCCACCAGGA	91,6	Ensayo	Capacitación 54
210	NM_004523	4394	GACCTGTGCCTTTTAGAGA	63,7	Capacitación	Ensayo 55
211	NM_004523	2117	GACTTCATTGACAGTGGCC	71,0	Capacitación	Ensayo 56
212	NM_004523	799	AAAGGACAACTGCAGCTAC	49,0	Capacitación	Ensayo 57
213	NM_000314	2753	TGGAGGGGAATGCTCAGAA	40,0	Capacitación	Capacitación 58
214	NM_000314	2510	TAAAGATGGCACTTTCCCG	79,0	Capacitación	Capacitación 59
215	NM_000314	2935	AAGGCAGCTAAAGGAAGTG	55,0	Capacitación	Capacitación 60
234	NM_007054	963	TATTGGGCCAGCAGATTAC	76,9	Capacitación	Capacitación 61
235	NM_007054	593	TTATGACGCTAGGCCACAA	74,4	Capacitación	Capacitación 62
236	NM_007054	1926	GGAGAAAGATCCCTTTGAG	78,3	Capacitación	Capacitación 63
237	NM_006845	324	ACAAAAACGGAGATCCGTC	72,2	Capacitación	Capacitación 64
238	NM_006845	2206	ATAAGCAGCAAGAAACGGC	30,9	Capacitación	Capacitación 65
239	NM_006845	766	GAATTTCGGGCTACTTTGG	65,8	Capacitación	Capacitación 66
240	NM_005163	454	CGCACCTTCCATGTGGAGA	86,8	Capacitación	Capacitación 67
241	NM_005163	1777	AGACGTTTTTGTGCTGTGG	76,0	Capacitación	Capacitación 68
242	NM_005163	1026	GCTGGAGAACCTCATGCTG	87,8	Capacitación	Capacitación 69
243	NM_005733	2139	CTCTACCACTGAAGAGTTG	90,7	Capacitación	Capacitación 70
244	NM_005733	1106	AAGTGGGTCGTAAGAACCA	82,5	Capacitación	Capacitación 71
245	NM_005733	696	GAAGCTGTCCCTGCTAAAT	93,4	Capacitación	Capacitación 72
246	NM_001813	3928	GAAGAGATCCCAGTGCTTC	86,8	Ensayo	Capacitación 73
247	NM_001813	4456	TCTGAAAGTGACCAGCTCA	82,5	Ensayo	Capacitación 74
248	NM_001813	2293	GAAAATGAAGCTTTGCGGG	78,4	Ensayo	Capacitación 75
249	NM_005030	1135	AAGAAGAACCAGTGGTTCG	83,0	Ensayo	Ensayo 76
250	NM_005030	572	CCGAGTTATTCATCGAGAC	93,6	Ensayo	Ensayo 77
251	NM_005030	832	AAGAGACCTACCTCCGGAT	85,0	Ensayo	Ensayo 78
255	NM_001315	3050	AATATCCTCAGGGGTGGAG	36,0	Capacitación	79
256	NM_001315	1526	GTGCCTCTTGTTGCAGAGA	88,0	Capacitación	80
257	NM_001315	521	GAAGCTCTCCAGACCATTT	96,0	Capacitación	81
261	NM_006218	456	AGAAGCTGTGGATCTTAGG	65,3	Ensayo	Capacitación 82
262	NM_006218	3144	TGATGCACATCATGGTGGC	68,9	Ensayo	Capacitación 83
263	NM_006218	2293	CTAGGAAACCTCAGGCTTA	94,7	Ensayo	Capacitación 84
264	NM_000075	1073	GCGAATCTCTGCCTTTCGA	79,0	Capacitación	Capacitación 85
265	NM_000075	685	CAGTCAAGCTGGCTGACTT	78,0	Capacitación	Capacitación 86
266	NM_000075	581	GGATCTGATGCGCCAGTTT	77,0	Capacitación	Capacitación 87
288	NM_020242	1829	GCACAACTCCTGCAAATTC	87,4	Capacitación	Capacitación 88
289	NM_020242	3566	GATGGAAGAGCCTCTAAGA	82,7	Capacitación	Capacitación 89
290	NM_020242	2631	ACGAAAAGCTGCTTGAGAG	73,4	Capacitación	Capacitación 90
291	NM_004073	570	GAAGACCATCTGTGGCACC	65,0	Capacitación	Capacitación 91
292	NM_004073	1977	TCAGGGACCAGCTTTACTG	60,0	Capacitación	Capacitación 92
293	NM_004073	958	GTTACCAAGAGCCTCTTTG	75,0	Capacitación	Capacitación 93

294	NM_005026	3279	AACCAAAGTGAACTGGCTG	56,3	Capacitación	Capacitación	94
295	NM_005026	2121	GATCGGCCACTTCCTTTTC	70,9	Capacitación	Capacitación	95
296	NM_005026	4004	AGAGATCTGGGCCTCATGT	67,3	Capacitación	Capacitación	96
303	NM_000051	5373	AGTTCGATCAGCAGCTGTT	60,9	Capacitación	Capacitación	97
304	NM_000051	3471	TAGATTGTTCCAGGACACG	71,2	Capacitación	Capacitación	98
305	NM_000051	7140	GAAGTTGGATGCCAGCTGT	56,3	Capacitación	Capacitación	99
309	NM_004064	1755	TGGTGATCACTCCAGGTAG	25,3	Capacitación	Capacitación	100
310	NM_004064	1505	TGTCCCTTTCAGAGACAGC	5,0	Capacitación	Capacitación	101
311	NM_004064	1049	GACGTCAAACGTAAACAGC	50,2	Capacitación	Capacitación	102
312	NM_006219	1049	AAGTTCATGTCAGGGCTGG	76,6	Ensayo	Capacitación	103
313	NM_006219	2631	CAAAGATGCCCTTCTGAAC	88,9	Ensayo	Capacitación	104
314	NM_006219	453	AATGCGCAAATTCAGCGAG	32,9	Ensayo	Capacitación	105
339	NM_003600	437	GCACAAAAGCTTGTCTCCA	96,0	Ensayo	Capacitación	106
340	NM_003600	1071	TTGCAGATTTTGGGTGGTC	37,0	Ensayo	Capacitación	107
341	NM_003600	1459	ACAGTCTTAGGAATCGTGC	61,1	Ensayo	Capacitación	108
342	NM_004958	1476	AGGACTTCGCCCATAAGAG	61,8	Ensayo	Capacitación	109
343	NM_004958	5773	CAACCTCCAGGATACACTC	80,9	Ensayo	Capacitación	110
344	NM_004958	7886	CCAACTTTCTAGCTGCTGT	71,1	Ensayo	Capacitación	111
348	NM_004856	1999	GAATGTGAGCGTAGAGTGG	92,2	Capacitación	Capacitación	112
349	NM_004856	1516	CCATTGGTTACTGACGTGG	87,7	Capacitación	Capacitación	113
350	NM_004856	845	AACCCAAACCTCCACAATC	71,8	Capacitación	Capacitación	114
369	XM_294563	117	GAAAGAAGCAGTTGACCTC	59,9	Capacitación	Capacitación	115
370	XM_294563	2006	CTAAAAGCTGGGTGGACTC	69,4	Capacitación	Capacitación	116
371	XM_294563	389	GAAAGCACCTCTTTGTGTG	64,2	Capacitación	Capacitación	117
399	NM_000546	1286	TGAGGCCTTGGAACTCAAG	17,8			118
400	NM_000546	2066	CCTCTTGGTCGACCTTAGT	74,5			119
401	NM_000546	1546	GCACCCAGGACTTCCATTT	93,2			120
417	NM_001184	3790	GAAACTGCAGCTATCTTCC	75,8	Capacitación	Capacitación	121
418	NM_001184	7717	GTTACAATGAGGCTGATGC	73,0	Capacitación	Capacitación	122
419	NM_001184	5953	TCACGACTCGCTGAACTGT	78,8	Capacitación	Capacitación	123
453	NM_005978	323	GACCGACCCTGAAGCAGAA	91,3	Ensayo	Ensayo	124
454	NM_005978	254	TTCCAGGAGTATGCTGTTT	74,4	Ensayo	Ensayo	125
455	NM_005978	145	GGAACTTCTGCACAAGGAG	96,5	Ensayo	Ensayo	126
465	NM_000551	495	TGTTGACGGACAGCCTATT	75,5	Ensayo	Capacitación	127
466	NM_000551	1056	GGCATTGGCATCTGCTTTT	89,7	Ensayo	Capacitación	128
467	NM_000551	3147	GTGAATGAGACACTCCAGT	82,2	Ensayo	Capacitación	129
468	NM_002658	1944	GAGCTGGTGTCTGATTGTT	82,8	Ensayo	Capacitación	130
469	NM_002658	1765	GTGTAAGCAGCTGAGGTCT	44,4	Ensayo	Capacitación	131
470	NM_002658	232	CTGCCCAAAGAAATTCGGA	47,8	Ensayo	Capacitación	132
507	NM_003391	792	ATTTGCCCGCGCATTTGTG	27,2	Ensayo	Capacitación	133
508	NM_003391	2171	AGAAGATGAATGGTCTGGC	69,4	Ensayo	Capacitación	134
509	NM_003391	981	AACGGGCGATTATCTCTGG	43,3	Ensayo	Capacitación	135
540	NM_002387	3490	GACTTAGAGCTGGGAATCT	83,7	Ensayo	Capacitación	136
541	NM_002387	4098	AGTTGAGGAGGTTTCTGCA	86,1	Ensayo	Capacitación	137
542	NM_002387	1930	GGATTATATCCAGCAGCTC	82,3	Ensayo	Capacitación	138
585	NM_014885	509	GTGGCTGGATTCATGTTCC	81,5	Capacitación	Capacitación	139
586	NM_014885	798	CAAGGCATCCGTTATATCT	84,7	Capacitación	Capacitación	140

587	NM_014885	270	ACCAGGATTTGGAGTGGAT	84,7	Capacitación	Capacitación 141
639	NM_001274	250	CTGAAGAAGCAGTCGCAGT	77,7		142
640	NM_001274	858	ATCGATTCTGCTCCTCTAG	86,2		143
641	NM_001274	1332	TGCCTGAAAGAGACTTGTG	85,4		144
651	NM_001259	807	TCTTGGACGTGATTGGACT	89,8	Capacitación	Capacitación 145
652	NM_001259	1036	AGAAAACCTGGATTCCCAC	88,9	Capacitación	Capacitación 146
653	NM_001259	556	ACCACAGAACATTCTGGTG	89,3	Capacitación	Capacitación 147
672	NM_003161	2211	GAAAGCCAGACAACTTCTG	87,1	Ensayo	Capacitación 148
673	NM_003161	1223	CTCTCAGTGAAAGTGCCAA	91,2	Ensayo	Capacitación 149
674	NM_003161	604	GACACTGCCTGCTTTTACT	98,1	Ensayo	Capacitación 150
678	NM_004972	3526	AAGAACCTGGTGAAAGTCC	57,2	Capacitación	Capacitación 151
679	NM_004972	4877	GAAGTGCAGCAGGTTAAGA	54,8	Capacitación	Capacitación 152
680	NM_004972	1509	AGCCGAGTTGTAACTATCC	74,9	Capacitación	Capacitación 153
684	NM_007194	1245	GATCACAGTGGCAATGGAA	80,9		154
685	NM_007194	1432	AAACTCTTGGAAGTGGTGC	39,2		155
686	NM_007194	2269	ATGAATCCACAGCTCTACC	44,6		156
687	NM_007313	3866	GAATGGAAGCCTGAACTGA	92,4	Ensayo	Capacitación 157
688	NM_007313	2451	AGACATCATGGAGTCCAGC	5,0	Ensayo	Capacitación 158
689	NM_007313	1296	CAAGTTCTCCATCAAGTCC	91,1	Ensayo	Capacitación 159
711	NM_139049	129	GGAATAGTATGCGCAGCTT	92,5	Ensayo	Capacitación 160
712	NM_139049	369	GTGATTCAGATGGAGCTAG	89,0	Ensayo	Capacitación 161
713	NM_139049	969	CACCCGTACATCAATGTCT	77,0	Ensayo	Capacitación 162
858	NM_001253	522	TCATTGGAAGAACAGCGGC	0,0	Ensayo	Capacitación 163
859	NM_001253	2571	AAGAAGACGTTCAGCGACA	93,5	Ensayo	Capacitación 164
860	NM_001253	911	AAAAAGCCTGCCCTTGGTT	88,1	Ensayo	Capacitación 165
1110	NM_006101	1847	CTTGCAACGTCTGTTAGAG	72,3	Ensayo	Capacitación 166
1111	NM_006101	999	CTGAAGGCTTCCTTACAAG	82,9	Ensayo	Capacitación 167
1112	NM_006101	1278	CAGAAGTTGTGGAATGAGG	79,1	Ensayo	Capacitación 168
1182	NM_016231	1302	GCAATGAGGACAGCTTGTG	79,8	Ensayo	Capacitación 169
1183	NM_016231	1829	TGTAGCTTTCCACTGGAGT	79,3	Ensayo	Capacitación 170
1184	NM_016231	1019	TCTCCTTGTGAACAGCAAC	62,5	Ensayo	Capacitación 171
1212	NM_001654	1072	AGTGAAGAACCTGGGGTAC	79,3	Ensayo	Capacitación 172
1213	NM_001654	595	GTTCCACCAGCATTGTTCC	86,2	Ensayo	Capacitación 173
1214	NM_001654	1258	GAATGAGATGCAGGTGCTC	86,9	Ensayo	Capacitación 174
1287	NM_005417	2425	CAATTCGTCGGAGGCATCA	73,9	Ensayo	Capacitación 175
1288	NM_005417	1077	GGGGAGTTTGCTGGACTTT	66,4	Ensayo	Capacitación 176
1289	NM_005417	3338	GCAGTGCCTGCCTATGAAA	68,2	Ensayo	Capacitación 177
1290	NM_001982	3223	CTAGACCTAGACCTAGACT	63,5	Ensayo	Capacitación 178
1291	NM_001982	3658	GAGGATGTCAACGGTTATG	49,4	Ensayo	Capacitación 179
1292	NM_001982	2289	CAAAGTCTTGGCCAGAATC	45,3	Ensayo	Capacitación 180
1293	NM_005400	249	GATCGAGCTGGCTGTCTTT	85,4	Ensayo	Capacitación 181
1294	NM_005400	1326	GGTCTTAAAGAAGGACGTC	63,4	Ensayo	Capacitación 182
1295	NM_005400	1848	TGAGGACGACCTATTTGAG	0,0	Ensayo	Capacitación 183
1317	NM_002086	465	TGAGCTGGTGGATTATCAC	85,5	Ensayo	Ensayo 184
1318	NM_002086	183	CTGGTACAAGGCAGAGCTT	95,5	Ensayo	Ensayo 185
1319	NM_002086	720	CCGGAACGTCTAAGAGTCA	92,3	Ensayo	Ensayo 186
1332	NM_006219	2925	TACAGAAAAGTTTGGCCGG	20,1	Ensayo	Capacitación 187

1333	NM_006219	2346	AATGAAGCCTTTGTGGCTG	22,4	Ensayo	Capacitación 188
1334	NM_006219	2044	GTGCACATTCCTGCTGTCT	79,0	Ensayo	Capacitación 189
1335	NM_003600	1618	CCTCCCTATTCAGAAAGCT	84,2	Ensayo	Capacitación 190
1336	NM_003600	650	GACTTTGAAATTGGTCGCC	52,1	Ensayo	Capacitación 191
1337	NM_003600	538	CACCCAAAAGAGCAAGCAG	96,3	Ensayo	Capacitación 192
1338	XM_294563	2703	TAAGCCTGGTGGTGATCTT	78,1	Capacitación	Capacitación 193
1339	XM_294563	1701	AAGGTCTTTACGCCAGTAC	29,5	Capacitación	Capacitación 194
1340	XM_294563	789	GGAATGTATCCGAGCACTG	73,5	Capacitación	Capacitación 195
1386	NM_033360	493	GGACTCTGAAGATGTACCT	91,0	Ensayo	Capacitación 196
1387	NM_033360	897	GGCATACTAGTACAAGTGG	84,8	Ensayo	Capacitación 197
1388	NM_033360	704	GAAAAGACTCCTGGCTGTG	0,0	Ensayo	Capacitación 198
1389	NM_024408	4735	CTTTGAATGCCAGGGGAAC	91,6	Ensayo	Capacitación 199
1390	NM_024408	2674	CCAAGGAACCTGCTTTGAT	96,4	Ensayo	Capacitación 200
1391	NM_024408	5159	GACTCAGACCACTGCTTCA	95,8	Ensayo	Capacitación 201
1392	NM_000435	6045	GCTGCTGTTGGACCACTTT	0,0	Ensayo	Capacitación 202
1393	NM_000435	5495	TGCCAACTGAAGAGGATGA	0,0	Ensayo	Capacitación 203
1394	NM_000435	4869	TGATCACTGCTTCCCCGAT	0,0	Ensayo	Capacitación 204
1410	AF308602	770	ATATCGACGATTGTCCAGG	36,7	Ensayo	Capacitación 205
1411	AF308602	3939	AGGCAAGCCCTGCAAGAAT	81,3	Ensayo	Capacitación 206
1412	AF308602	1644	CACTTACACCTGTGTGTGC	81,3	Ensayo	Capacitación 207
1581	NM_005633	3593	TATCAGACCGGACCTCTAT	70,8	Ensayo	Capacitación 208
1582	NM_005633	364	ATTGACCACCAGGTTTCTG	1,4	Ensayo	Capacitación 209
1583	NM_005633	3926	CTTACAAAAGGGAGCACAC	66,9	Ensayo	Capacitación 210
1620	NM_002388	1097	GTCTCAGCTTCTGCGGTAT	95,0	Ensayo	Capacitación 211
1621	NM_002388	286	AGGATTTTGTGGCCTCCAT	94,6	Ensayo	Capacitación 212
1622	NM_002388	2268	TCCAGGTTGAAGGCATTCA	92,5	Ensayo	Capacitación 213
1629	NM_012193	3191	TTGGCAAAGGCTCCTTGTA	80,0	Ensayo	Ensayo 214
1630	NM_012193	5335	CCATCTGCTTGAGCTACTT	85,0	Ensayo	Ensayo 215
1631	NM_012193	2781	GTTGACTTACCTGACGGAC	43,1	Ensayo	Ensayo 216
1632	NM_004380	3708	GACATCCCGAGTCTATAAG	85,3	Ensayo	Capacitación 217
1633	NM_004380	339	TGGAGGAGAATTAGGCCTT	81,1	Ensayo	Capacitación 218
1634	NM_004380	5079	GCACAAGGAGGTCTTCTTC	79,0	Ensayo	Capacitación 219
1641	NM_017412	2331	CAGATCACTCCAGGCATAG	97,3	Ensayo	Capacitación 220
1643	NM_017412	2783	ATGTGTGGTGACTGCTTTG	95,7	Ensayo	Capacitación 221
1695	NM_001903	2137	TGACATCATTGTGCTGGCC	38,4	Ensayo	Capacitación 222
1696	NM_001903	655	CGTTCCGATCCTCTATACT	97,9	Ensayo	Capacitación 223
1697	NM_001903	3117	TGACCAAAGATGACCTGTG	40,1	Ensayo	Capacitación 224
1815	NM_020168	3064	GAGAAAGAATGGGGTCGGT	85,0	Capacitación	Capacitación 225
1816	NM_020168	681	CGACATCCAGAAGTTGTCA	86,1	Capacitación	Capacitación 226
1817	NM_020168	1917	TGAGGAGCAGATTGCCACT	72,1	Capacitación	Capacitación 227
2502	NM_000271	237	GAGGTACAATTGCGAATAT	87,0	Capacitación	Capacitación 228
2503	NM_000271	559	TACTACGTCGGACAGAGTT	76,0	Capacitación	Capacitación 229
2504	NM_000271	1783	AACTACAATAACGCCACTG	39,0	Capacitación	Capacitación 230
2505	NM_000271	2976	GCCACAGTCGTCTTGCTGT	84,0	Capacitación	Capacitación 231
2512	NM_005030	245	GGGCGGCTTTGCCAAGTGC	88,6	Ensayo	Ensayo 232
2513	NM_005030	1381	CACGCCTCATCCTCTACAA	90,5	Ensayo	Ensayo 233
2514	NM_005030	834	GAGACCTACCTCCGGATCA	91,0	Ensayo	Ensayo 234

2521	NM_000314	1316	CCCACCACAGCTAGAACTT	93,0	Capacitación	Capacitación	235
2522	NM_000314	1534	CTATTCCCAGTCAGAGGCG	89,0	Capacitación	Capacitación	236
2523	NM_000314	2083	CAGTAGAGGAGCCGTCAAA	90,0	Capacitación	Capacitación	237
2524	NM_006622	1928	CAGTTCACTATTACGCAGA	65,0	Capacitación	Capacitación	238
2525	NM_006622	586	TGTTACGAGATGACAGATT	73,0	Capacitación	Capacitación	239
2526	NM_006622	1252	AACCCAGAGGATCGTCCCA	70,0	Capacitación	Capacitación	240
2527	NM_139164	200	CTGTTTGGAGAAAACCCTC	79,0	Capacitación	Capacitación	241
2528	NM_139164	568	GACAACCCAAACCAGAGTC	71,0	Capacitación	Capacitación	242
2529	NM_139164	488	GTCTTGACTGGGATGAAAA	66,0	Capacitación	Capacitación	243
2530	NM_139164	578	ACCAGAGTCTTTTGACAGG	82,0	Capacitación	Capacitación	244
2546	NM_014875	1090	TAGACCACCCATTGCTTCC	63,5	Ensayo	Capacitación	245
2547	NM_014875	1739	AGAGCCTTCGAAGGCTTCA	73,2	Ensayo	Capacitación	246
2548	NM_014875	3563	GACCATAGCATCCGCCATG	87,1	Ensayo	Capacitación	247
2602	NM_002387	2655	TAGCTCTGCTAGAGGAGGA	71,0	Ensayo	Capacitación	248
2603	NM_002387	1418	ACAGAACGGCTGAATAGCC	43,5	Ensayo	Capacitación	249
2604	NM_002387	941	GAGAATGAGAGCCTGACTG	81,0	Ensayo	Capacitación	250
2605	NM_016231	1683	GGAAACAGAGTGCCTCTCT	55,3	Ensayo	Capacitación	251
2606	NM_016231	915	CCACTCAGCTCAGATCATG	82,3	Ensayo	Capacitación	252
2607	NM_016231	737	TCTGGTCTCTTGCAAAAGG	30,3	Ensayo	Capacitación	253
2611	NM_004380	4230	ATTTTTGCGGCGCCAGAAT	79,0	Ensayo	Capacitación	254
2612	NM_004380	2197	GAAAAACGGAGGTCGCGTT	85,9	Ensayo	Capacitación	255
2613	NM_004380	5701	GAAAACAAATGCCCCGTGC	55,4	Ensayo	Capacitación	256
2614	NM_005978	276	TGGCACTCATCACTGTCAT	91,8	Ensayo	Ensayo	257
2615	NM_005978	229	TGAGAACAGTGACCAGCAG	91,9	Ensayo	Ensayo	258
2616	NM_005978	369	GGGCCCAGGACTGTTGATG	94,5	Ensayo	Ensayo	259
2617	NM_017412	3128	AGAGATGGGCATTGTTTCC	94,3	Ensayo	Capacitación	260
2618	NM_017412	814	GCTCATGGAGATGTTTGGT	88,7	Ensayo	Capacitación	261
2619	NM_017412	1459	AGCATTGCTGTTTCACGCC	93,1	Ensayo	Capacitación	262
2620	NM_001654	1902	TTGAGCTGCTGCAACGGTC	67,2	Ensayo	Capacitación	263
2621	NM_001654	1006	GTCCCCACATTCCAAGTCA	90,0	Ensayo	Capacitación	264
2622	NM_001654	2327	CCTCTCTGGAATTTGTGCC	85,7	Ensayo	Capacitación	265
2623	NM_002658	202	CAAGTACTTCTCCAACATT	87,2	Ensayo	Capacitación	266
2624	NM_002658	181	TGGAGGAACATGTGTGTCC	0,0	Ensayo	Capacitación	267
2625	NM_002658	436	TTACTGCAGGAACCCAGAC	0,0	Ensayo	Capacitación	268
2629	NM_006218	1334	TGGCTTTGAATCTTTGGCC	3,5	Ensayo	Capacitación	269
2630	NM_006218	2613	AGGTGCACTGCAGTTCAAC	53,8	Ensayo	Capacitación	270
2631	NM_006218	1910	TTCAGCTAGTACAGGTCCT	78,0	Ensayo	Capacitación	271
2632	NM_003161	1834	TTGATTCCTCGCGACATCT	88,3	Ensayo	Capacitación	272
2633	NM_003161	1555	GCTTTTCCCATGATCTCCA	90,7	Ensayo	Capacitación	273
2634	NM_003161	217	CTTGGCATGGAACATTGTG	61,4	Ensayo	Capacitación	274
2635	NM_003391	2072	GCCTCAGAAAGGGATTGCT	79,1	Ensayo	Capacitación	275
2636	NM_003391	1318	GCTCTGGATGTGCACACAT	60,5	Ensayo	Capacitación	276
2637	NM_003391	1734	GTGTCTCAAAGGAGCTTTC	87,1	Ensayo	Capacitación	277
2641	AF308602	4260	ATTCAACGGGCTCTTGTGC	0,0	Ensayo	Capacitación	278
2642	AF308602	1974	GATCGATGGCTACGAGTGT	84,0	Ensayo	Capacitación	279
2643	AF308602	5142	CATCCCCTACAAGATCGAG	41,6	Ensayo	Capacitación	280
2644	NM_024408	8232	GCAACTTTGGTCTCCTTTC	91,0	Ensayo	Capacitación	281

2645	NM_024408	10503	GCAATTGGCTGTGATGCTC	86,6	Ensayo	Capacitación 282
2646	NM_024408	8643	GAGACAAGTTAACTCGTGC	89,4	Ensayo	Capacitación 283
2647	NM_007313	4222	TCCTGGCAAGAAAGCTTGA	65,6	Ensayo	Capacitación 284
2648	NM_007313	3237	AAACCTCTACACGTTCTGC	53,5	Ensayo	Capacitación 285
2649	NM_007313	302	CTAAAGGTGAAAAGCTCCG	67,8	Ensayo	Capacitación 286
2650	NM_000551	631	GATCTGGAAGACCACCCAA	70,9	Ensayo	Capacitación 287
2651	NM_000551	4678	CAGAACCCAAAAGGGTAAG	0,0	Ensayo	Capacitación 288
2652	NM_000551	4382	AGGAAATAGGCAGGGTGTG	4,3	Ensayo	Capacitación 289
2653	NM_001903	1888	AGCAGTGCTGATGATAAGG	89,1	Ensayo	Capacitación 290
2654	NM_001903	2606	AAGCCATTGGTGAAGAGAG	91,9	Ensayo	Capacitación 291
2655	NM_001903	1583	TGTGTCATTGCTCTCCAAG	90,3	Ensayo	Capacitación 292
2656	NM_002388	842	GCAGATGAGCAAGGATGCT	86,8	Ensayo	Capacitación 293
2657	NM_002388	1754	GTACATCCATGTGGCCAAA	94,6	Ensayo	Capacitación 294
2658	NM_002388	2642	TGGGTCATGAAAGCTGCCA	93,1	Ensayo	Capacitación 295
2662	NM_005633	3251	GAACACCGTTAACACCTCC	31,2	Ensayo	Capacitación 296
2663	NM_005633	2899	ATAACAGGAGAGATCCAGC	21,7	Ensayo	Capacitación 297
2664	NM_005633	2607	TGGTGTCCTTGAGGTTGTC	75,1	Ensayo	Capacitación 298
2665	NM_033360	329	ACCTGTCTCTTGGATATTC	81,4	Ensayo	Capacitación 299
2666	NM_033360	529	TAAATGTGATTTGCCTTCT	47,8	Ensayo	Capacitación 300
2667	NM_033360	585	GAAGTTATGGAATTCCTTT	94,2	Ensayo	Capacitación 301
2668	NM_139049	745	CACCATGTCCTGAATTCAT	80,7	Ensayo	Capacitación 302
2669	NM_139049	433	TCAAGCACCTTCATTCTGC	42,6	Ensayo	Capacitación 303
2670	NM_139049	550	CGAGTTTTATGATGACGCC	79,9	Ensayo	Capacitación 304
2671	NM_002086	555	ATACGTCCAGGCCCTCTTT	87,9	Ensayo	Ensayo 305
2672	NM_002086	392	TGCAGCACTTCAAGGTGCT	36,9	Ensayo	Ensayo 306
2673	NM_002086	675	CGGGCAGACCGGCATGTTT	92,6	Ensayo	Ensayo 307
2674	NM_004958	5024	GACATGAGAACCTGGCTCA	77,8	Ensayo	Capacitación 308
2675	NM_004958	2155	CTTGCAGGCCTTGTTTGTG	83,2	Ensayo	Capacitación 309
2676	NM_004958	6955	TAATACAGCTGGGGACGAC	52,3	Ensayo	Capacitación 310
2677	NM_012193	467	AGAACCTCGGCTACAACGT	71,5	Ensayo	Ensayo 311
2678	NM_012193	473	TCGGCTACAACGTGACCAA	51,3	Ensayo	Ensayo 312
2679	NM_012193	449	TCCGCATCTCCATGTGCCA	37,5	Ensayo	Ensayo 313
2680	NM_005400	665	TCACAAAGTGTGCTGGGTT	43,9	Ensayo	Capacitación 314
2681	NM_005400	2178	CCAGGAGGAATTCAAAGGT	41,6	Ensayo	Capacitación 315
2682	NM_005400	1022	GCTCACCATCTGAGGAAGA	64,2	Ensayo	Capacitación 316
2686	NM_001982	948	TGACAGTGGAGCCTGTGTA	65,8	Ensayo	Capacitación 317
2687	NM_001982	1800	CTTTCTGAATGGGGAGCCT	61,7	Ensayo	Capacitación 318
2688	NM_001982	2860	TACACACACCAGAGTGATG	0,0	Ensayo	Capacitación 319
2692	NM_016195	5331	ATGAAGGAGAGTGATCACC	10,5	Ensayo	Capacitación 320
2693	NM_016195	4829	AATGGCAGTGAAACACCCT	67,3	Ensayo	Capacitación 321
2694	NM_016195	1480	AAGTTTGTGTCCCAGACAC	80,5	Ensayo	Capacitación 322
2695	NM_000435	2107	AATGGCTTCCGCTGCCTCT	0,0	Ensayo	Capacitación 323
2696	NM_000435	5193	GAACATGGCCAAGGGTGAG	15,5	Ensayo	Capacitación 324
2697	NM_000435	7273	GAGTCTGGGACCTCCTTCT	0,0	Ensayo	Capacitación 325
2802	NM_004523	46	CCAGGGAGACTCCGGCCCC	6,7	Capacitación	Ensayo 326
2803	NM_004523	132	GGGACCGTCATGGCGTCGC	8,2	Capacitación	Ensayo 327
2804	NM_004523	221	ATTTAATTTGGCAGAGCGG	0,0	Capacitación	Ensayo 328

2805NM_004523	322	GCTCAAGGAAAACATACAC	76,2	Capacitación Ensayo	329
2806NM_004523	365	TACTAAACAGATTGATGTT	77,9	Capacitación Ensayo	330
2807NM_004523	581	TACTGATAATGGTACTGAA	93,8	Capacitación Ensayo	331
2808NM_004523	716	AGGAGTGATAATTAAAGGT	84,8	Capacitación Ensayo	332
2809NM_004523	852	GTTTTCTCTGTTACAATAC	85,4	Capacitación Ensayo	333
2810NM_004523	995	TGGAAATATAAATCAATCC	0,0	Capacitación Ensayo	334
2811NM_004523	1085	ACTAACTAGAATCCTCCAG	0,0	Capacitación Ensayo	335
2812NM_004523	1174	AAACTCTGAGTACATTGGA	81,9	Capacitación Ensayo	336
2813NM_004523	1375	TAACTGTTCAAGAAGAGCA	14,1	Capacitación Ensayo	337
2814NM_004523	1570	AAGAAGAATATATCACATC	0,0	Capacitación Ensayo	338
2815NM_004523	1706	AGTTGACCAACACAATGCA	86,0	Capacitación Ensayo	339
2816NM_004523	2197	TACATGAACTACAAGAAAA	90,0	Capacitación Ensayo	340
2817NM_004523	2858	GACTAAGCTTAATTGCTTT	87,0	Capacitación Ensayo	341
2818NM_004523	3089	GGGGCAGTATACTGAAGAA	64,5	Capacitación Ensayo	342
2819NM_004523	3878	TTCTTGTATATTATTAAGT	0,0	Capacitación Ensayo	343
2820NM_004523	4455	TCTATAATTTATATTCTTT	9,3	Capacitación Ensayo	344
2821NM_004523	4648	TACAAAGAATAAATTTTCT	23,5	Capacitación Ensayo	345
2823NM_005030	45	CAGCGCAGCTTCGGGAGCA	72,1	Capacitación Ensayo	346
2824NM_005030	131	CGGAGTTGCAGCTCCCGGA	85,7	Capacitación Ensayo	347
2825NM_005030	303	GGCAAGATTGTGCCTAAGT	80,1	Capacitación Ensayo	348
2826NM_005030	346	GGGAGAAGATGTCCATGGA	100,0	Capacitación Ensayo	349
2827NM_005030	432	GACTTCGTGTTCGTGGTGT	89,3	Capacitación Ensayo	350
2828NM_005030	519	GCCCGATACTACCTACGGC	86,2	Capacitación Ensayo	351
2829NM_005030	648	GGACTGGCAACCAAAGTCG	86,7	Capacitación Ensayo	352
2830NM_005030	777	TGTATCATGTATACCTTGT	84,3	Capacitación Ensayo	353
2831NM_005030	821	TTCTTGCCTAAAAGAGACC	26,8	Capacitación Ensayo	354
2832NM_005030	907	TCCAGAAGATGCTTCAGAC	90,8	Capacitación Ensayo	355
2833NM_005030	952	ACGAGCTGCTTAATGACGA	87,7	Capacitación Ensayo	356
2834NM_005030	1038	TCGATTGCTCCCAGCAGCC	31,4	Capacitación Ensayo	357
2835NM_005030	1082	CACAGTCCTCAATAAAGGC	62,9	Capacitación Ensayo	358
2836NM_005030	1214	CAATGCCTCCAAGCCCTCG	0,0	Capacitación Ensayo	359
2837NM_005030	1300	AGTGGGTGGACTATTCGGA	84,9	Capacitación Ensayo	360
2838NM_005030	1515	TACATGAGCGAGCACTTGC	20,3	Capacitación Ensayo	361
2839NM_005030	1860	CTCAAGGCCTCCTAATAGC	74,2	Capacitación Ensayo	362
2840NM_005030	1946	CCGCGGTGCCATGTCTGCA	79,7	Capacitación Ensayo	363
2841NM_005030	2075	CCCCTCCCCCTCAACCCCA	34,6	Capacitación Ensayo	364
3041NM_014875	4629	ATTTTCTAGAAAACGGTAA	91,8		365
3042NM_014875	77	GAGGGGCGAAGTTTCGGCA	71,2		366
3043NM_014875	243	CTGGGACCGGGAAGCCGGA	0,0		367
3044NM_014875	5094	CTTCTACTTCTGTTGGCAG	85,9		368
3045NM_014875	4354	ACTTACTATTCAGACTGCA	85,7		369
3046NM_014875	524	GCCCTCACCCACAGTAGCC	68,1		370
3047NM_014875	5349	CAGAGGAATGCACACCCAG	73,6		371
3048NM_014875	4824	GATTGATTAGATCTCTTGA	91,3		372
3049NM_014875	3014	GTGAGTATTATCCCAGTTG	41,5		373
3050NM_014875	2959	ATCTGGGGTGCTGATTGCT	46,3		374
3051NM_014875	1514	GTGACAGTGGCAGTACGCG	67,7		375

3052NM_014875	1114	TCAGACTGAAGTTGTTAGA	80,8	376
3053NM_014875	2079	GTTGGCTAGAATTGGGAAA	91,8	377
3054NM_014875	3560	GAAGACCATAGCATCCGCC	74,8	378

Tabla III 30 ARNip diseñados utilizando el método de este ejemplo

BioID	Registro	Nombre del gen	Secuencia (cadena en sentido)	% de Silenciamiento	SEQ ID NO
3844	NM_014875	KIF14	CAGGTAAAGTCAGAGACAT	87	379
3845	NM_014875	KIF14	GGGATTGACGGCAGTAAGA	89	380
3846	NM_014875	KIF14	CACTGAATGTGGGAGGTGA	92	381
3847	NM_014875	KIF14	GTCTGGGTGGAAATTCAAA	93	382
3848	NM_014875	KIF14	CATCTTTGCTGAATCGAAA	86	383
3849	NM 014875	KIF14	CAGGGATGCTGTTTGGATA	95	384
3850	NM_005030	PLK	CCCTGTGTGGGACTCCTAA	87	385
3851	NM_005030	PLK	GGTGTTCGCGGGCAAGATT	86	386
3852	NM_005030	PLK	CGCCTCATCCTCTACAATG	88	387
3853	NM_005030	PLK	GTTCTTTACTTCTGGCTAT	97	388
3854	NM_005030	PLK	CTCCTTAAATATTTCCGCA	92	389
3855	NM 005030	PLK	CTGAGCCTGAGGCCCGATA	75	390
3856	NM_000875	IGF1R	CAAATTATGTGTTTCCGAA	90	391
3857	NM_000875	IGF1R	CGCATGTGCTGGCAGTATA	84	392
3858	NM_000875	IGF1R	CCGAAGATTTCACAGTCAA	79	393
3859	NM_000875	IGF1R	ACCATTGATTCTGTTACTT	86	394
3860	NM_000875	IGF1R	ACCGCAAAGTCTTTGAGAA	88	395
3861	NM 000875	IGF1R	GTCCTGACATGCTGTTTGA	79	396
3862	NM_001315	MAPK14	GGAATTCAATGATGTGTAT	85	397
3863	NM_001315	MAPK14	GCTGTTGACTGGAAGAACA	84	398
3864	NM_001315	MAPK14	CTCCTGAGATCATGCTGAA	81	399
3865	NM_001315	MAPK14	CCATTTCAGTCCATCATTC	88	400
3866	NM_001315	MAPK14	CAGATTATGCGTCTGACAG	25	401
3867	NM 001315	MAPKI4	CGCTTATCTCATTAACAGG	14	402
3871	NM_004523	KIF11	GAGCCCAGATCAACCTTTA	87	403
3872	NM_004523	KIF11	CTGACAAGAGCTCAAGGAA	89	404
3873	NM_004523	KIF11	GGCATTAACACACTGGAGA	92	405
3874	NM_004523	KIF11	GATGGCAGCTCAAAGCAAA	93	406
3875	NM_004523	KIF11	CAGCAGAAATCTAAGGATA	86	407
3876	NM 004523	KIF11	CGTTCTGGAGCTGTTGATA	95	408

6.2. Ejemplo 2: Selección de ARNip con respecto a su especificidad de silenciamiento

Se ha demostrado la importancia de los efectos inespecíficos de las secuencias de ARNip y ARNhc. Los experimentos con micromatrices sugieren que la mayoría de los oligos de ARNip dan como resultado la regulación negativa de genes inespecíficos a través de interacciones directas entre un ARNip y los transcritos inespecíficos. Aunque la similitud de secuencia entre ARNbc y transcritos parece jugar un papel en la determinación de qué genes inespecíficos se ven afectados, las búsquedas de similitud de secuencias, incluso combinadas con modelos termodinámicos de hibridación, son insuficientes para predecir con precisión efectos inespecíficos. Sin embargo, el alineamiento de transcritos inespecíficos con secuencias de ARNip no válidas revela que algunas interacciones de emparejamiento de bases entre los dos parecen ser más importantes que otras (figura 6).

La Figura 6 muestra un ejemplo de alineamientos de transcritos de genes diana en el oligonucleótido de 19 meros núcleo de una secuencia oligonucleotídica de ARNip. Los genes inespecíficos se seleccionaron de la micromatriz Human 25k v2.2.1 seleccionando patrones cinéticos de la abundancia de transcritos en consonancia con los efectos directos de los oligos de ARNip. Los alineamientos se generaron con FASTA y se editaron a mano. Los recuadros negros y el área gris demuestran el mayor nivel de similitud de secuencia en la mitad 3' del alineamiento.

El alineamiento mostrado en la Fig. 6 y datos similares para otros ARNip, se combinaron para generar una matriz de puntuación específica de posición para su uso en la predicción de efectos inespecíficos. La matriz, que refleja la frecuencia con la que se encuentra que cada posición en el oligo, coincide con transcritos inespecíficos afectados,

se representa en la figura 7.

La matriz de puntuación específica de posición se utiliza para calcular puntuaciones para alineamientos entre una secuencia de ARNi candidata y secuencias de transcrito inespecíficas. Los alineamientos de interés se establecen con una búsqueda FASTA de baja rigurosidad y la puntuación para cada alineamiento se calcula con la ecuación 6

$$Puntuación = \sum_{i=1}^{n} \ln(E_i / 0{,}25)$$

en la que: n es la longitud del alineamiento (generalmente 19); $E_i = P_i$ de la Fig. 7 si la posición *i* en el alineamiento es una coincidencia y $E_i = (1-P_i)/3$ si la posición *i* es una coincidencia errónea. Se observó que el número de alineamientos para un ARNip determinado que puntúa por encima de un umbral es indicativo del número de efectos inespecíficos observados. El umbral de la puntuación se optimizó para maximizar la correlación entre el número de efectos previstos y observados (Fig. 8). El procedimiento en fase de desarrollo de selección utiliza el umbral optimizado para favorecer secuencias con un número relativamente pequeño de efectos inespecíficos previstos.

6.3. Ejemplo 3: PSSMS por modelos de curvas

También se generaron PSSM mediante un método que creo la hipótesis de la dependencia de la composición de bases de cualquier posición en sus posiciones adyacentes, denominado "modelos de curvas".

Los modelos de curvas se generaron como una suma de curvas normales. Cada curva representa la probabilidad de encontrar una base particular en una región particular. El valor en cada posición en las curvas normales sumadas es el peso determinado en esa posición para la base representada por la curva. Los pesos para cada base presente en cada posición en cada ARNip y sus secuencias flanqueantes, se sumaron para generar una puntuación de ARNip, es decir, la puntuación es Σw_i. El cálculo de la puntuación también puede describirse como el producto puntual del contenido de bases en la secuencia con los pesos en el modelo de curva. Como tal, es una forma de representar la correlación de la secuencia de interés con el modelo.

Los modelos de curvas pueden inicializarse para corresponderse con los picos y valles principales presentes en la diferencia de composición de bases suavizada entre los ARNip buenos y malos, por ejemplo, como se describe en las Figs. 1A-C y 5A-C. Para el modelo de curva G/C de 3 picos, el modelo inicial puede configurarse de la siguiente manera:

Pico 1

media: 1,5

desviación típica: 2

amplitud: 0,0455

La media, la desviación típica y la amplitud del pico 1, se configuran para corresponderse con el pico en la diferencia media en el contenido de GC entre los ARNip buenos y malos que aparecen en las bases 2 - 5 del sitio diana del ARNip en el conjunto 1 de ensayos de capacitación y de ensayo.

Pico 2

media: 11

desviación típica: 0,5

amplitud: 0,0337

La media, la desviación típica y la amplitud del pico 2 se configuran para corresponderse con el pico en la diferencia de medias en el contenido de GC entre los ARNip buenos y malos que aparecen en las bases 10-12 del sitio diana del ARNip en el conjunto 1 de ensayos de capacitación y de ensayo.

Pico 3

media: 18,5

desviación típica: 4

amplitud:-0,0548

La media, la desviación típica y la amplitud del pico 3 se configuran para corresponderse con el pico en la diferencia de medias en el contenido de GC entre los ARNip buenos y malos que aparecen en las bases 12-25 del sitio diana del ARNip en el conjunto 1 de ensayos de capacitación y de ensayo.

En un modelo de curva, se puede ajustar la altura máxima (amplitud), la posición central en la secuencia (media) y la

anchura (desviación típica) de un pico. Los modelos de curvas se optimizaron ajustando la amplitud, la media y la desviación típica de cada pico sobre una cuadrícula de valores preestablecida. En una realización, los modelos de curvas se optimizaron en varios conjuntos de capacitación y se ensayaron en varios conjuntos de ensayo, por ejemplo, conjuntos de capacitación y conjuntos de ensayo como se describe en la Tabla II. Cada base, -G/C, A o U, se optimizó por separado, y después se exploraron combinaciones de modelos optimizados para obtener el mejor rendimiento.

Los criterios de optimización para los modelos de curvas fueron: (1) la fracción de oligos buenos en el 10 %, 15 %, 20 % y 33 % superior de las puntaciones, (2) la tasa de detección falsa en el 33 % y 50 % de los ARNip seleccionados, y (3) el coeficiente de correlación de silenciamiento de ARNip frente a puntuaciones de ARNip se utilizó como una prueba decisiva.

Cuando el modelo se capacitó, se exploró una cuadrícula de valores posibles para la amplitud, la media y la desviación típica de cada pico. Adicionalmente se seleccionaron y examinaron los modelos con el valor superior o dentro del intervalo superior de valores de cualquiera de los criterios anteriores.

Los modelos G/C se optimizaron con 3 o 4 picos. Los modelos A se optimizaron con 3 picos. Los modelos U se optimizaron con 5 picos.

A continuación se enumeran intervalos de optimización a modo de ejemplo de los modelos:

Modelos G/C de 3 picos:

pico 1:

amplitudes: gc1 = 0 - 0,091
medias: gc1 = -2,5 - 1,5
desviaciones típicas: gc1 = 2 , 5 - 4

pico 2:

amplitudes: gc2 = 0,0337 - 0,1011
medias: gc2 =11-11,5
desviaciones típicas: gc2 = 0 , 5 - 0 , 9

pico 3:

amplitudes: gc3 = -0,1644 - -0,0822
medias: gc3 = 18,75-20,75
desviaciones típicas: gc3 =2,5-3,5

Modelos G/C de 4 picos:

pico 0:

amplitudes: gc0 = 0 - 0,091
medias: gc0 = -5,5 - -3,5
desviaciones típicas: gc0 = 1 - 2,5

pico 1:

amplitudes: gc1 = 0 - 0,091
medias: gc1 = -2,5 - 1,5
desviaciones típicas: gc1 =2,5-4

pico 2:

amplitudes: gc2 = 0,0337 - 0,1011
medias: gc2 =11-11,5
desviaciones típicas: gc2 = 0,5-0,9

pico 3:

amplitudes: gc3 = -0,1644 - -0,0822
medias: gc3 = 18,75-20,75
Desviaciones típicas: gc3 =2,5-3,5

Modelos U de 5 picos:

U pico 1:

amplitudes: u1 = -0,2 - 0,0
medias: u1 = 1 -2
desviaciones típicas: u1 = ,75 - 1,5

U pico 2:

amplitudes: u2 = 0,0 - 0,16
medias: u2 = 5 - 6
desviaciones típicas: u2 = ,75 - 1,5

U pico 3:

amplitudes: u3 = 0,0 - 0,1
medias: u3 = 10 - 11
desviaciones típicas: u3 = 1 - 2

U pico 4:

amplitudes: u4 = 0,0 - 0,16
medias: u4 = 13 - 14
desviaciones típicas: u4 = ,75 - 1,5

U pico 5:

amplitudes: u5 = 0,0 - 0,16
medias: u5 = 17 - 18
desviaciones típicas: u5 = 1 - 3

Modelo A de 3 picos:

A pico 1:

amplitudes: a1 = 0,0442 - 0,2210
medias: a1 = 5,5 - 6,5
desviaciones típicas: a1 = 1 - 2

A pico 2:

amplitudes: a2 = -,05 - 0
medias: a2 = 10 - 12,5
Desviaciones típicas: a2 = 2,5 - 4,5

A pico 3:

amplitudes: a3 = 0,0442 - 0,2210
medias: a3 = 18-20
desviaciones típicas: a3 = 4 - 6

En la Fig. 11A se muestra un conjunto a modo de ejemplo de modelos de curvas de PSSM. En la Fig. 11B se muestra el rendimiento de los modelos en los conjuntos de capacitación y de ensayo.

6.4. Ejemplo 4: Modelos de composición de bases para la predicción de la preferencia de cadenas de ARNip

La diferencia media en el contenido de G/C entre los ARNip buenos y malos, proporciona un modelo para las PSSM G/C que puede utilizarse para clasificar motivos funcionales y resistentes a ARNip. Como se sabe que ambas cadenas del ARNip pueden ser activas (véase, por ejemplo, Elbashir et al., 2001, Genes Dev. 15: 188-200), fue interesante descubrir como de bien se ajustaba el contenido de G/C de las dos cadenas en sentido y antisentido de los ARNip al modelo de contenido de G/C del motivo dina funcional de ARNip resultante de la diferencia media en el contenido de G/C entre los ARNip buenos y malos. Para ello, se examinaron los complementos inversos de los ARNip buenos y malos. Estos complementos inversos corresponden a los supuestos sitios diana de coincidencia perfecta para las cadenas en sentido de los dúplex de ARNip. Los complementos inversos se compararon con los

ARNip buenos y malos reales, representados por los sitios diana reales, de coincidencia perfecta, de las cadenas antisentido de los dúplex de ARNip

La Fig. 14A muestra la diferencia entre el contenido medio de G/C de los complementos inversos de los ARNip malos con el contenido medio de G/C de los propios ARNip malos, dentro de la región dúplex de ARNip de 19 meros. Para comparar, se muestra la diferencia entre el contenido medio de G/C de los ARNip buenos y malos. Las curvas se suavizaron sobre una ventana de 5 (o parte de una ventana de 5, en los bordes de la secuencia).

La figura 14B muestra la diferencia entre el contenido medio de G/C de los complementos inversos de los ARNip buenos con el contenido medio de G/C de los ARNip malos, dentro de la región dúplex de ARNip de 19 meros. Para comparar, se muestra la diferencia entre el contenido medio de G/C de los ARNip buenos y malos. Las curvas se suavizaron sobre una ventana de 5 (o parte de una ventana de 5, en los bordes de la secuencia).

Se observó que los complementos inversos de los ARNip malos eran aún más diferentes de los propios ARNip malos que de los ARNip buenos. En promedio, los complementos inversos de los ARNip malos tenían un contenido de G/C aún más fuerte en el extremo 5' que los ARNip buenos y fueron similares en contenido de G/C a los ARNip buenos en el extremo 3'. Por el contrario, se observó que los complementos inversos de los ARNip buenos eran sustancialmente más similares a los ARNip malos que a los ARNip buenos. En promedio, los complementos inversos de los ARNip buenos apenas diferían de los ARNip malos en el contenido de G/C en el extremo 5' y solo eran ligeramente menos ricos en G/C que los ARNip malos en el extremo 3'.

Estos resultados parecen implicar que las PSSM G/C están distinguiendo ARNip con cadenas en sentido fuertes como ARNip malos de ARNip con cadenas en sentido débiles como ARNip buenos. Se predice que un ARNip cuya puntuación de PSSM G/C es mayor que la puntuación de PSSM G/C de su complemento inverso tiene una cadena antisentido que es más activa que su cadena en sentido. Por el contrario, se predice que un ARNip cuya puntuación de PSSM G/C es menor que la puntuación de PSSM G/C de su complemento inverso tiene una cadena en sentido que es más activa que su cadena antisentido.

Se ha demostrado que el aumento de la eficacia corresponde a una mayor actividad de la cadena antisentido y a una menor actividad de la cadena en sentido. Por lo tanto, las PSSM G/C de esta invención parecen distinguir ARNip buenos con mayor eficacia debido a la actividad dominante de la cadena antisentido (ARNip “antisentido-activos”) de ARNip con actividad dominante de la cadena en sentido (ARNip "en sentido- activos").

La importancia de comparar las PSSM G/C de los ARNip y sus complementos inversos para la predicción del sesgo de cadenas se sometió a ensayo por comparación con la estimación del sesgo de la cadena de perfiles de expresión de ARNip por el método de sesgado en 3'.

Los ARNip y sus complementos inversos se puntuaron utilizando como matriz de peso la diferencia de contenido de G/C suavizada entre los ARNip buenos y malos en el oligo de 19 meros, mostrado en la FIG. 14A. La puntuación de PSSM G/C de cada cadena fue el producto puntual del contenido de G/C de la cadena de ARNip con la matriz de diferencia de contenido de G/C, siguiendo el método de cálculo de puntuación de las PSSM de modelos de curva.

Los ARNip se denominaron en sentido-activo por el método de sesgado en 3’ de análisis de perfil de expresión si la puntuación idéntica antisentido superaba la puntuación idéntica en sentido. Los ARNip se denominaron en sentido-activo por el método de PSSM G/C si la puntuación de PSSM G/C de su complemento inverso superaba su propia puntuación de PSSM G/C

En la Fig. 15, los ARNip se combinaron según la eficacia de silenciamiento medida, y se comparó la frecuencia de los identificados como de sentido-activo según los métodos de perfil de expresión y PSSM G/C. Aunque estas técnicas se basan en análisis distintos, la concordancia es bastante buena. Las dos muestran que se predice que una mayor proporción de ARNip de silenciamiento bajo frente a ARNip de silenciamiento alto es de sentido activo. El coeficiente de correlación para (puntuación de PSSM G/C de ARNip - puntuación de PSSM G/C de complemento inverso) frente a $\log_{10}$ (puntuación de identidad en sentido/puntuación de identidad antisentido) es de 0,59 para el conjunto de 61 ARNip combinados en la FIG. 15.

6.5. Ejemplo 5: Diseño de ARNip para el silenciamiento de genes que tienen nivel bajos de transcritos

En los ejemplos anteriores, se describió un algoritmo de diseño de ARNip mejorado que permite la selección de ARNip con mayor capacidad silenciadora y más uniforme. A pesar de esta mejora drástica, algunos genes siguen siendo difíciles de silenciar con alta eficacia. Se observó una tendencia general hacia un silenciamiento más pobre para genes poco expresados (intensidad inferior a -0,5 en micromatriz, <5 copias por célula, Figura 16). Este ejemplo describe la identificación de los parámetros que afectan a la eficacia del silenciamiento de los ARNip en genes poco expresados.

Se seleccionaron veinticuatro genes poco expresados para el análisis detallado de los parámetros que afectan a la eficacia del silenciamiento del ARNip. Se evaluaron diversos criterios con respecto a su capacidad para distinguir

ARNip buenos y malos, que incluyen la composición de bases de la secuencia dúplex de ARNip de 19 meros y la región diana flanqueante. Además, se consideró la contribución del contenido de GC del transcrito diana. Estos ensayos revelaron que la eficacia del ARNip se correlacionaba bien con la composición de bases del gen diana y del ARNip. En particular, el contenido de GC de los ARNip buenos difirió sustancialmente del de los ARNip malos de una manera específica de la región (Figura 17). Las secuencias de los ARNip utilizadas en la generación de la Figura 17 se enumeran en la Tabla IV. Los dúplex de ARNip bueno tendían a ser pobres en GC en las posiciones 2-7 del extremo 5 'de la cadena en sentido, y pobres en GC en el extremo 3' (posiciones 18-19). Además, la eficacia del ARNip se correlacionó con un bajo contenido en GC en la secuencia de transcripción que flanquea el sitio de unión de ARNip. La necesidad de un bajo contenido en GC como determinante de la eficacia del ARNip puede explicar la dificultad de silenciar los transcritos mal expresados, ya que estos transcritos tienden a ser generalmente ricos en GC. La composición de bases del dúplex de ARNip también afectó al silenciamiento de genes mal expresados. En particular, el contenido de GC de los ARNip buenos difirió sustancialmente del de los ARNip malos de una manera específica de la región (Figura 17). Los dúplex de ARNip buenos tendían a ser ricos en GC en la primera posición, pobres en GC en las posiciones 2-7 del extremo 5 'de la cadena en sentido y pobres en GC en el extremo 3' (posiciones 18-19). De los criterios examinados, un bajo contenido de GC en las posiciones 2-7 de la cadena en sentido (Figura 17, línea discontinua) produjo la mayor mejora en la eficacia del silenciamiento. Esto está en consonancia con la región del ARNip implicada en la etapa de catálisis del silenciamiento del transcrito. El bajo contenido de GC en esta región puede proporcionar accesibilidad o geometría helicoidal óptima para una escisión mejorada. La necesidad de un bajo contenido de GC en esta región del ARNip también puede seleccionar sitios diana que contienen un bajo contenido de GC que flanquea el sitio de unión, lo que también se correlaciona con la eficacia del silenciamiento.

La composición de bases para ARNip buenos en genes poco expresados difiere algo de nuestros criterios de composición de base previamente obtenidos para ARNip buenos en genes bien expresados (Figura 17, línea continua). Los ARNip buenos en ambos tipos de genes muestran una preferencia por un nivel alto de GC en la posición 1 y bajo en el extremo 3'. Sin embargo, los ARNip para genes bien expresados muestran una asimetría extrema en el contenido de GC entre los dos extremos, mientras que los ARNip para genes poco expresados prefieren una asimetría más moderada. Nuestro algoritmo de diseño anterior busca maximizar la asimetría, de acuerdo con las características observadas en ARNip buenos para genes bien expresados. Nuestros resultados actuales indican que la composición de bases de más de una región del ARNip puede influir en la eficacia. Las diferentes regiones del ARNip pueden ser más críticas para el silenciamiento de diferentes dianas, dependiendo quizá de las características del transcrito diana, tales como el nivel de expresión o el contenido global de GC. De acuerdo con esta idea, diferentes algoritmos de diseño disponibles en el comercia, funcionan bien en diferentes subconjuntos de genes (datos no mostrados).

Se desarrolló un nuevo algoritmo de diseño de ARNip basado en la composición de GC procedente de genes poco expresados. El nuevo algoritmo incluye los siguientes ajustes al algoritmo anterior:

(1) selección para 1-3 G + C en sentido de 19 meros en bases 2-7,

(2) asimetría en bases 1 y 19 del oligo de 19 meros (posición 1, G o C, posición 19, A o T),

(3) -300 <puntuación pssm <+200,

(4) la mayor coincidencia inespecífica con BLAST no supera 16, y

(5) las 200 bases a cada lado del oligo de 19 meros no son secuencias de repetición o de baja complejidad.

El nuevo algoritmo se comparó con el algoritmo descrito en los ejemplos previos, mediante ensayos en paralelo de nuevos ARNip seleccionados por cada uno. En la Figura 18 se muestran los resultados obtenidos con tres ARNip seleccionados para cada método. Los ARNip diseñados por el nuevo algoritmo del presente ejemplo mostraron una mejor eficacia media (80 %, en comparación con 60 % según el método convencional para ARNip) y fueron más uniformes en cuanto a su rendimiento. La distribución de las eficacias de silenciamiento de los ARNip obtenidos por el nuevo algoritmo fue significativamente mejor que la del algoritmo anterior para los mismos genes ($p = 10^{-5}$, suma de rangos de Wilcoxon). Los ARNip diseñados utilizando el nuevo algoritmo de diseño también parecen ser eficaces para silenciar transcritos más altamente expresados, según un análisis de 12 genes altamente expresados.

Los nuevos criterios de diseño pueden capturar características importantes para la funcionalidad del ARNip en general (Figura 19), y destaca que diferentes regiones de los ARNip tienen diferentes funciones en el reconocimiento de transcritos, escisión y liberación de productos. Las bases próximas al extremo 5' de la cadena guía están implicadas en la unión del transcrito (transcritos tanto diana como inespecíficos), y recientemente se ha demostrado que son suficientes para la energía de unión al ARN diana. Los criterios de diseño también son coherentes con los datos disponibles sobre cómo los ARNip interactúan con RISC, el complejo proteína-ARN que actúa como intermediario en el silenciamiento de ARN. Estos estudios muestran que el emparejamiento de bases más débil en el extremo 5' de la cadena antisentido (extremo 3' del dúplex) estimula la interacción preferencial de la cadena antisentido con RISC, quizá facilitando el desenrrollamiento del dúplex de ARNip a través de un componente de

helicasa en 5'-3' de RISC. Como en el diseño anterior, nuestro nuevo diseño mantiene la asimetría de la composición base que estimula la interacción preferencial de la cadena antisentido. Esto sugiere que la ineficacia previa de silenciamiento de transcritos poco expresados no se debe a la asociación ineficaz con RISC, sino que probablemente se debe a un direccionamiento ineficaz del complejo RISC en el transcrito diana, o a una escisión y liberación ineficaces del transcrito diana. Los diseños descritos en estos ejemplos incluyen una preferencia por U en la posición 10 de la cadena en sentido, que se ha asociado con una eficacia de escisión mejorada por RISC tal como ocurre en la mayoría de las endonucleasas. La preferencia observada por el bajo contenido de GC que flanquea el sitio de escisión puede mejorar la accesibilidad del complejo RISC/nucleasa para la escisión, o la liberación del transcrito escindido, coherente con estudios recientes que demuestran que los pares de bases formados por las regiones central y 3' de la cadena guía del ARNip proporcionan una geometría helicoidal necesaria para la catálisis. Los nuevos criterios de diseño pueden aumentar la eficacia de estas etapas y de etapas adicionales en la ruta de ARNi, proporcionando de este modo un silenciamiento eficaz de transcritos a diferentes niveles de expresión.

Tabla IV ARNip para las Figura 17

NÚMERO DE REGISTRO	**GEN**	**Secuencia de ARNip**	**SEQ ID NO**
AK092024_NM_030932	DIAPH3	GCAGTGATTGCTCAGCAGC	409
AK092024_NM_030932	DIAPH3	GAGTTTACCGACCACCAAG	410
AK092024_NM_030932	DIAPH3	CACGGTTGGCAGAGTCTAT	411
AK092024_NM_030932	DIAPH3	TGCGGATGCCATTCAGTGG	412
NM_014875	KIF14	AAACTGGGAGGCTACTTAC	413
NM_014875	KIF14	CTCACATTGTCCACCAGGA	414
NM_014875	KIF14	GACCATAGCATCCGCCATG	415
NM_014875	KIF14	AGAGCCTTCGAAGGCTTCA	416
NM_014875	KIF14	TAGACCACCCATTGCTTCC	417
NM_014875	KIF14	ACTGACAACAAAGTGCAGC	418
U53530	DNCH1	TGGCCAGCGCTTACTGGAA	419
U53530	DNCH1	GCAAGTTGAGCTCTACCGC	420
NM_000859	HMGCR	TTGTGTGTGGGACCGTAAT	421
NM_000859	HMGCR	CAACAGAAGGTTGTCTTGT	422
NM_000859	HMGCR	CAGAGACAGAATCTACACT	423
NM_000859	HMGCR	CACGATGCATAGCCATCCT	424
NM_000271	NPC1	GAGGTACAATTGCGAATAT	425
NM_000271	NPC1	GCCACAGTCGTCTTGCTGT	426
NM_000271	NPC1	TACTACGTCGGACAGAGTT	427
NM_000271	NPC1	AACTACAATAACGCCACTG	428
NM_004523	KNSL1	TACTGATAATGGTACTGAA	429
NM_004523	KNSL1	TACATGAACTACAAGAAAA	430
NM_004523	KNSL1	GACTAAGCTTAATTGCTTT	431
NM_004523	KNSL1	AGTTGACCAACACAATGCA	432
NM_004523	KNSL1	GTTTTCTCTGTTACAATAC	433
NM_004523	KNSL1	AGGAGTGATAATTAAAGGT	434
NM_004523	KNSL1	AAACTCTGAGTACATTGGA	435
NM_004523	KNSL1	TACTAAACAGATTGATGTT	436
NM_004523	KNSL1	GCTCAAGGAAAACATACAC	437
NM_004523	KNSL1	CTGGATCGTAAGAAGGCAG	438
NM_004523	KNSL1	GACTTCATTGACAGTGGCC	439
NM_004523	KNSL1	GGACAACTGCAGCTACTCT	440
NM_004523	KNSL1	GGGGCAGTATACTGAAGAA	441
NM_004523	KNSL1	GACCTGTGCCTTTTAGAGA	442

NM_004523	KNSL1	AAAGGACAACTGCAGCTAC	443
NM_004523	KNSL1	TACAAAGAATAAATTTTCT	444
NM_004523	KNSL1	TGGAAGGTGAAAGGTCACC	445
NM_004523	KNSL1	TAACTGTTCAAGAAGAGCA	446
NM_004523	KNSL1	TCTATAATTTATATTCTTT	447
NM_004523	KNSL1	GGGACCGTCATGGCGTCGC	448
NM_004523	KNSL1	CCAGGGAGACTCCGGCCCC	449
NM_004523	KNSL1	ATTTAATTTGGCAGAGCGG	450
NM_004523	KNSL1	TGGAAATATAAATCAATCC	451
NM_004523	KNSL1	ACTAACTAGAATCCTCCAG	452
NM_004523	KNSL1	AAGAAGAATATATCACATC	453
NM_004523	KNSL1	TTCTTGTATATTATTAAGT	454
NM_004064	CDKN1B	GACGTCAAACGTAAACAGC	455
NM_004064	CDKN1B	TGGTGATCACTCCAGGTAG	456
NM_004064	CDKN1B	TGTCCCTTTCAGAGACAGC	457
NM_004073	CNK	GTTACCAAGAGCCTCTTTG	458
NM_004073	CNK	ATCGTAGTGCTTGTACTTA	459
NM_004073	CNK	GAAGACCATCTGTGGCACC	460
NM_004073	CNK	GGAGACGTACCGCTGCATC	461
NM_004073	CNK	TCAGGGACCAGCTTTACTG	462
NM_004073	CNK	AGTCATCCCGCAGAGCCGC	463
NM_001315	MAPK14	GGCCTTTTCACGGGAACTC	464
NM_001315	MAPK14	GAAGCTCTCCAGACCATTT	465
NM_001315	MAPK14	TGCCTACTTTGCTCAGTAC	466
NM_001315	MAPK14	ATGTGATTGGTCTGTTGGA	467
NM_001315	MAPK14	GTCATCAGCTTTGTGCCAC	468
NM_001315	MAPK14	CCTACAGAGAACTGCGGTT	469
NM_001315	MAPK14	CCAGTGGCCGATCCTTATG	470
NM_001315	MAPK14	GTGCCTCTTGTTGCAGAGA	471
NM_001315	MAPK14	TTCTCCGAGGTCTAAAGTA	472
NM_001315	MAPK14	TAATTCACAGGGACCTAAA	473
NM_001315	MAPK14	GTGGCCGATCCTTATGATC	474
NM_001315	MAPK14	GTATATACATTCAGCTGAC	475
NM_001315	MAPK14	AATATCCTCAGGGGTGGAG	476
NM_001315	MAPK14	GGAACACCCCCCGCTTATC	477
NM_006101	HEC	CTGAAGGCTTCCTTACAAG	478
NM_006101	HEC	AGAACCGAATCGTCTAGAG	479
NM_006101	HEC	CAGAAGTTGTGGAATGAGG	480
NM_006101	HEC	GTTCAAAAGCTGGATGATC	481
NM_006101	HEC	GGCCTCTATACCCCTCAAA	482
NM_006101	HEC	CTTGCAACGTCTGTTAGAG	483
NM_000314	PTEN	CCCACCACAGCTAGAACTT	484
NM_000314	PTEN	CAGTAGAGGAGCCGTCAAA	485
NM_000314	PTEN	CTATTCCCAGTCAGAGGCG	486
NM_000314	PTEN	TAAAGATGGCACTTTCCCG	487
NM_000314	PTEN	AAGGCAGCTAAAGGAAGTG	488
NM_000314	PTEN	TGGAGGGGAATGCTCAGAA	489

NM_000075	CDK4	GCGAATCTCTGCCTTTCGA	490
NM_000075	CDK4	CAGTCAAGCTGGCTGACTT	491
NM_000075	CDK4	GGATCTGATGCGCCAGTTT	492
NM_000075	CDK4	TGTTGTCCGGCTGATGGAC	493
NM_006622	SNK	TGTTACGAGATGACAGATT	494
NM_006622	SNK	AACCCAGAGGATCGTCCCA	495
NM_006622	SNK	CAGTTCACTATTACGCAGA	496
NM_139164	STARD4	ACCAGAGTCTTTTGACAGG	497
NM_139164	STARD4	CTGTTTGGAGAAAACCCTC	498
NM_139164	STARD4	GACAACCCAAACCAGAGTC	499
NM_139164	STARD4	GTCTTGACTGGGATGAAAA	500
NM_005030	PLK	GGGAGAAGATGTCCATGGA	501
NM_005030	PLK	CCGAGTTATTCATCGAGAC	502
NM_005030	PLK	GAGACCTACCTCCGGATCA	503
NM_005030	PLK	TCCAGAAGATGCTTCAGAC	504
NM_005030	PLK	CACGCCTCATCCTCTACAA	505
NM_005030	PLK	GACTTCGTGTTCGTGGTGT	506
NM_005030	PLK	GGGCGGCTTTGCCAAGTGC	507
NM_005030	PLK	ACGAGCTGCTTAATGACGA	508
NM_005030	PLK	GGACTGGCAACCAAAGTCG	509
NM_005030	PLK	GCCCGATACTACCTACGGC	510
NM_005030	PLK	CGGAGTTGCAGCTCCCGGA	511
NM_005030	PLK	AAGAGACCTACCTCCGGAT	512
NM_005030	PLK	AGTGGGTGGACTATTCGGA	513
NM_005030	PLK	TGTATCATGTATACCTTGT	514
NM_005030	PLK	AAGAAGAACCAGTGGTTCG	515
NM_005030	PLK	GGCAAGATTGTGCCTAAGT	516
NM_005030	PLK	CCGCGGTGCCATGTCTGCA	517
NM_005030	PLK	CTCAAGGCCTCCTAATAGC	518
NM_005030	PLK	CAGCGCAGCTTCGGGAGCA	519
NM_005030	PLK	CACAGTCCTCAATAAAGGC	520
NM_005030	PLK	CCCCTCCCCCTCAACCCCA	521
NM_005030	PLK	TCGATTGCTCCCAGCAGCC	522
NM_005030	PLK	TTCTTGCCTAAAAGAGACC	523
NM_005030	PLK	TACATGAGCGAGCACTTGC	524
NM_005030	PLK	CAATGCCTCCAAGCCCTCG	525
NM_000875	IGF1R	GGATATTGGGCTTTACAAC	526
NM_000875	IGF1R	CTTGCAGCAACTGTGGGAC	527
NM_000875	IGF1R	GCTCACGGTCATTACCGAG	528
NM_000875	IGF1R	GATGATTCAGATGGCCGGA	529
NM_000875	IGF1R	CGACACGGCCTGTGTAGCT	530
NM_000875	IGF1R	AATGCTGACCTCTGTTACC	531
NM_000875	IGF1R	TCTCAAGGATATTGGGCTT	532
NM_000875	IGF1R	CATTACTCGGGGGGCCATC	533
NM_000875	IGF1R	TGCTGACCTCTGTTACCTC	534
NM_000875	IGF1R	CTACGCCCTGGTCATCTTC	535
NM_000875	IGF1R	CCTCACGGTCATCCGCGGC	536
NM_000875	IGF1R	CCTGAGGAACATTACTCGG	537

NM_001813	CENPE	GGAGAGCTTTCTAGGACCT	538
NM_001813	CENPE	GAAGAGATCCCAGTGCTTC	539
NM_001813	CENPE	ACTCTTACTGCTCTCCAGT	540
NM_001813	CENPE	TCTGAAAGTGACCAGCTCA	541
NM_001813	CENPE	GAAAATGAAGCTTTGCGGG	542
NM_001813	CENPE	CTTAACACGGATGCTGGTG	543
NM_004958	FRAP1	CTTGCAGGCCTTGTTTGTG	544
NM_004958	FRAP1	CAACCTCCAGGATACACTC	545
NM_004958	FRAP1	GACATGAGAACCTGGCTCA	546
NM_004958	FRAP1	CCAACTTTCTAGCTGCTGT	547
NM_004958	FRAP1	AGGACTTCGCCCATAAGAG	548
NM_004958	FRAP1	TAATACAGCTGGGGACGAC	549
NM_005163	AKT1	GCTGGAGAACCTCATGCTG	550
NM_005163	AKT1	CGCACCTTCCATGTGGAGA	551
NM_005163	AKT1	AGACGTTTTTGTGCTGTGG	552
NM_002358	MAD2L1	TACGGACTCACCTTGCTTG	553
NM_000551	VHL	GGCATTGGCATCTGCTTTT	554
NM_000551	VHL	GTGAATGAGACACTCCAGT	555
NM_000551	VHL	TGTTGACGGACAGCCTATT	556
NM_000551	VHL	GATCTGGAAGACCACCCAA	557
NM_000551	VHL	AGGAAATAGGCAGGGTGTG	558
NM_000551	VHL	CAGAACCCAAAAGGGTAAG	559
NM_001654	ARAF1	GTCCCCACATTCCAAGTCA	560
NM_001654	ARAF1	GAATGAGATGCAGGTGCTC	561
NM_001654	ARAF1	GTTCCACCAGCATTGTTCC	562
NM_001654	ARAF1	CCTCTCTGGAATTTGTGCC	563
NM_001654	ARAF1	AGTGAAGAACCTGGGGTAC	564
NM_001654	ARAF1	TTGAGCTGCTGCAACGGTC	565
NM_000435	NOTCH3	GAACATGGCCAAGGGTGAG	566
NM_000435	NOTCH3	GAGTCTGGGACCTCCTTCT	567
NM_000435	NOTCH3	AATGGCTTCCGCTGCCTCT	568
NM_000435	NOTCH3	TGATCACTGCTTCCCCGAT	569
NM_000435	NOTCH3	TGCCAACTGAAGAGGATGA	570
NM_000435	NOTCH3	GCTGCTGTTGGACCACTTT	571
NM_024408	NOTCH2	CCAAGGAACCTGCTTTGAT	572
NM_024408	NOTCH2	GACTCAGACCACTGCTTCA	573
NM_024408	NOTCH2	CTTTGAATGCCAGGGGAAC	574
NM_024408	NOTCH2	GCAACTTTGGTCTCCTTTC	575
NM_024408	NOTCH2	GAGACAAGTTAACTCGTGC	576
NM_024408	NOTCH2	GCAATTGGCTGTGATGCTC	577
NM_012193	FZD4	CCATCTGCTTGAGCTACTT	578
NM_012193	FZD4	TTGGCAAAGGCTCCTTGTA	579
NM_012193	FZD4	AGAACCTCGGCTACAACGT	580
NM_012193	FZD4	TCGGCTACAACGTGACCAA	581
NM_012193	FZD4	GTTGACTTACCTGACGGAC	582
NM_012193	FZD4	TCCGCATCTCCATGTGCCA	583
NM_007313	ABL1	GAATGGAAGCCTGAACTGA	584

NM_007313	ABL1	CAAGTTCTCCATCAAGTCC	585
NM_007313	ABL1	CTAAAGGTGAAAAGCTCCG	586
NM_007313	ABL1	TCCTGGCAAGAAAGCTTGA	587
NM_007313	ABL1	AAACCTCTACACGTTCTGC	588
NM_007313	ABL1	AGACATCATGGAGTCCAGC	589
NM_017412	FZD3	CAGATCACTCCAGGCATAG	590
NM_017412	FZD3	ATGTGTGGTGACTGCTTTG	591
NM_017412	FZD3	AGAGATGGGCATTGTTTCC	592
NM_017412	FZD3	AGCATTGCTGTTTCACGCC	593
NM_017412	FZD3	GCTCATGGAGATGTTTGGT	594
NM_005633	SOS1	TGGTGTCCTTGAGGTTGTC	595
NM_005633	SOS1	TATCAGACCGGACCTCTAT	596
NM_005633	SOS1	CTTACAAAAGGGAGCACAC	597
NM_005633	SOS1	GAACACCGTTAACACCTCC	598
NM_005633	SOS1	ATAACAGGAGAGATCCAGC	599
NM_005633	SOS1	ATTGACCACCAGGTTTCTG	600
NM_005417	SRC	CAATTCGTCGGAGGCATCA	601
NM_005417	SRC	GCAGTGCCTGCCTATGAAA	602
NM_005417	SRC	GGGGAGTTTGCTGGACTTT	603
NM_005400	PRKCE	GATCGAGCTGGCTGTCTTT	604
NM_005400	PRKCE	GCTCACCATCTGAGGAAGA	605
NM_005400	PRKCE	GGTCTTAAAGAAGGACGTC	606
NM_005400	PRKCE	TCACAAAGTGTGCTGGGTT	607
NM_005400	PRKCE	CCAGGAGGAATTCAAAGGT	608
NM_005400	PRKCE	TGAGGACGACCTATTTGAG	609
NM_002388	MCM3	GTCTCAGCTTCTGCGGTAT	610
NM_002388	MCM3	GTACATCCATGTGGCCAAA	611
NM_002388	MCM3	AGGATTTTGTGGCCTCCAT	612
NM_002388	MCM3	TGGGTCATGAAAGCTGCCA	613
NM_002388	MCM3	TCCAGGTTGAAGGCATTCA	614
NM_002388	MCM3	GCAGATGAGCAAGGATGCT	615
NM_004380	CREBBP	GAAAAACGGAGGTCGCGTT	616
NM_004380	CREBBP	GACATCCCGAGTCTATAAG	617
NM_004380	CREBBP	TGGAGGAGAATTAGGCCTT	618
NM_004380	CREBBP	ATTTTTGCGGCGCCAGAAT	619
NM_004380	CREBBP	GCACAAGGAGGTCTTCTTC	620
NM_004380	CREBBP	GAAAACAAATGCCCCGTGC	621
NM_006219	PIK3CB	CAAAGATGCCCTTCTGAAC	622
NM_006219	PIK3CB	GTGCACATTCCTGCTGTCT	623
NM_006219	PIK3CB	AAGTTCATGTCAGGGCTGG	624
NM_006219	PIK3CB	AATGCGCAAATTCAGCGAG	625
NM_006219	PIK3CB	AATGAAGCCTTTGTGGCTG	626
NM_006219	PIK3CB	TACAGAAAAGTTTGGCCGG	627
NM_006218	PIK3CA	CTAGGAAACCTCAGGCTTA	628
NM_006218	PIK3CA	TTCAGCTAGTACAGGTCCT	629

NM_006218	PIK3CA	TGATGCACATCATGGTGGC	630
NM_006218	PIK3CA	AGAAGCTGTGGATCTTAGG	631
NM_006218	PIK3CA	AGGTGCACTGCAGTTCAAC	632
NM_006218	PIK3CA	TGGCTTTGAATCTTTGGCC	633
NM_002086	GRB2	CTGGTACAAGGCAGAGCTT	634
NM_002086	GRB2	CGGGCAGACCGGCATGTTT	635
NM_002086	GRB2	CCGGAACGTCTAAGAGTCA	636
NM_002086	GRB2	ATACGTCCAGGCCCTCTTT	637
NM_002086	GRB2	TGAGCTGGTGGATTATCAC	638
NM_002086	GRB2	TGCAGCACTTCAAGGTGCT	639
NM_001982	ERBB3	TGACAGTGGAGCCTGTGTA	640
NM_001982	ERBB3	CTAGACCTAGACCTAGACT	641
NM_001982	ERBB3	CTTTCTGAATGGGGAGCCT	642
NM_001982	ERBB3	GAGGATGTCAACGGTTATG	643
NM_001982	ERBB3	CAAAGTCTTGGCCAGAATC	644
NM_001982	ERBB3	TACACACACCAGAGTGATG	645
NM_001903	CTNNA1	CGTTCCGATCCTCTATACT	646
NM_001903	CTNNA1	AAGCCATTGGTGAAGAGAG	647
NM_001903	CTNNA1	TGTGTCATTGCTCTCCAAG	648
NM_001903	CTNNA1	AGCAGTGCTGATGATAAGG	649
NM_001903	CTNNA1	TGACCAAAGATGACCTGTG	650
NM_001903	CTNNA1	TGACATCATTGTGCTGGCC	651
NM_003600	STK6	CACCCAAAAGAGCAAGCAG	652
NM_003600	STK6	GCACAAAAGCTTGTCTCCA	653
NM_003600	STK6	CCTCCCTATTCAGAAAGCT	654
NM_003600	STK6	ACAGTCTTAGGAATCGTGC	655
NM_003600	STK6	GACTTTGAAATTGGTCGCC	656
NM_003600	STK6	TTGCAGATTTTGGGTGGTC	657
NM_003161	RPS6KB1	GACACTGCCTGCTTTTACT	658
NM_003161	RPS6KB1	CTCTCAGTGAAAGTGCCAA	659
NM_003161	RPS6KB1	GCTTTTCCCATGATCTCCA	660
NM_003161	RPS6KB1	TTGATTCCTCGCGACATCT	661
NM_003161	RPS6KB1	GAAAGCCAGACAACTTCTG	662
NM_003161	RPS6KB1	CTTGGCATGGAACATTGTG	663
AF308602	NOTCH1	GATCGATGGCTACGAGTGT	664
AF308602	NOTCH1	CACTTACACCTGTGTGTGC	665
AF308602	NOTCH1	AGGCAAGCCCTGCAAGAAT	666
AF308602	NOTCH1	CATCCCCTACAAGATCGAG	667
AF308602	NOTCH1	ATATCGACGATTGTCCAGG	668
AF308602	NOTCH1	ATTCAACGGGCTCTTGTGC	669
NM_016231	NLK	CCACTCAGCTCAGATCATG	670
NM_016231	NLK	GCAATGAGGACAGCTTGTG	671
NM_016231	NLK	TGTAGCTTTCCACTGGAGT	672
NM_016231	NLK	TCTCCTTGTGAACAGCAAC	673
NM_016231	NLK	GGAAACAGAGTGCCTCTCT	674

NM_016231	NLK	TCTGGTCTCTTGCAAAAGG	675
NM_ 001253	CDC5L	AAGAAGACGTTCAGCGACA	676
NM_001253	CDC5L	AAAAAGCCTGCCCTTGGTT	677
NM_001253	CDC5L	TCATTGGAAGAACAGCGGC	678
NM_003391	WNT2	GTGTCTCAAAGGAGCTTTC	679
NM_003391	WNT2	GCCTCAGAAAGGGATTGCT	680
NM_003391	WNT2	AGAAGATGAATGGTCTGGC	681
NM_003391	WNT2	GCTCTGGATGTGCACACAT	682
NM_003391	WNT2	AACGGGCGATTATCTCTGG	683
NM_003391	WNT2	ATTTGCCCGCGCATTTGTG	684
NM_002387	MCC	AGTTGAGGAGGTTTCTGCA	685
NM_002387	MCC	GACTTAGAGCTGGGAATCT	686
NM_002387	MCC	GGATTATATCCAGCAGCTC	687
NM_002387	MCC	GAGAATGAGAGCCTGACTG	688
NM_002387	MCC	TAGCTCTGCTAGAGGAGGA	689
NM_002387	MCC	ACAGAACGGCTGAATAGCC	690
NM_005978	S100A2	GGAACTTCTGCACAAGGAG	691
NM_005978	S100A2	GGGCCCAGGACTGTTGATG	692
NM_005978	S100A2	TGAGAACAGTGACCAGCAG	693
NM_005978	S100A2	TGGCACTCATCACTGTCAT	694
NM_005978	S100A2	GACCGACCCTGAAGCAGAA	695
NM_005978	S100A2	TTCCAGGAGTATGCTGTTT	696
NM_033360	KRAS2	GAAGTTATGGAATTCCTTT	697
NM_033360	KRAS2	GGACTCTGAAGATGTACCT	698
NM_033360	KRAS2	GGCATACTAGTACAAGTGG	699
NM_033360	KRAS2	ACCTGTCTCTTGGATATTC	700
NM_033360	KRAS2	ACCTGTCTCTTGGATATTC	701
NM_033360	KRAS2	GAAAAGACTCCTGGCTGTG	702
NM_139049	MAPK8	GAAAAGACTCCTGGCTGTG	703
NM_139049	MAPK8	GTGATTCAGATGGAGCTAG	704
NM_139049	MAPK8	CACCATGTCCTGAATTCAT	705
NM_139049	MAPK8	CGAGTTTTATGATGACGCC	706
NM_139049	MAPK8	CACCCGTACATCAATGTCT	707
NM_139049	MAPK8	TCAAGCACCTTCATTCTGC	708
NM_002658	PLAU	CAAGTACTTCTCCAACATT	709
NM_002658	PLAU	GAGCTGGTGTCTGATTGTT	710
NM_002658	PLAU	CTGCCCAAAGAAATTCGGA	711
NM_002658	PLAU	GTGTAAGCAGCTGAGGTCT	712
NM_002658	PLAU	TGGAGGAACATGTGTGTCC	713
NM_002658	PLAU	TTACTGCAGGAACCCAGAC	714
NM_016195	MPHOSPH1	AGAGGAACTCTCTGCAAGC	715
NM_016195	MPHOSPH1	AAGTTTGTGTCCCAGACAC	716
NM_016195	MPHOSPH1	CTGAAGAAGCTACTGCTTG	717
NM_016195	MPHOSPH1	GACATGCGAATGACACTAG	718
NM_016195	MPHOSPH1	AATGGCAGTGAAACACCCT	719

NM_016195	MPHOSPH1	ATGAAGGAGAGTGATCACC	720
NM_020168	PAK6	CGACATCCAGAAGTTGTCA	721
NM_020168	PAK6	GAGAAAGAATGGGGTCGGT	722
NM_020168	PAK6	TGAGGAGCAGATTGCCACT	723
NM_000051	ATM	TAGATTGTTCCAGGACACG	724
NM_000051	ATM	AGTTCGATCAGCAGCTGTT	725
NM_000051	ATM	GAAGTTGGATGCCAGCTGT	726
NM_001259	CDK6	TCTTGGACGTGATTGGACT	727
NM_001259	CDK6	ACCACAGAACATTCTGGTG	728
NM_001259	CDK6	AGAAAACCTGGATTCCCAC	729
NM_004856	KNSL5	GAATGTGAGCGTAGAGTGG	730
NM_004856	KNSL5	CCATTGGTTACTGACGTGG	731
NM_004856	KNSL5	AACCCAAACCTCCACAATC	732
NM_006845	KNSL6	ACAAAAACGGAGATCCGTC	733
NM_006845	KNSL6	GAATTTCGGGCTACTTTGG	734
NM_006845	KNSL6	ATAAGCAGCAAGAAACGGC	735
NM_004972	JAK2	AGCCGAGTTGTAACTATCC	736
NM_004972	JAK2	AAGAACCTGGTGAAAGTCC	737
NM_004972	JAK2	GAAGTGCAGCAGGTTAAGA	738
NM_005026	PIK3CD	GATCGGCCACTTCCTTTTC	739
NM_005026	PIK3CD	AGAGATCTGGGCCTCATGT	740
NM_005026	PIK3CD	AACCAAAGTGAACTGGCTG	741
NM_014885	APC10	CAAGGCATCCGTTATATCT	742
NM_014885	APC10	ACCAGGATTTGGAGTGGAT	743
NM_014885	APC10	GTGGCTGGATTCATGTTCC	744
NM_005733	RAB6KIFL	GAAGCTGTCCCTGCTAAAT	745
NM_005733	RAB6KIFL	CTCTACCACTGAAGAGTTG	746
NM_005733	RAB6KIFL	AAGTGGGTCGTAAGAACCA	747
NM_007054	KIF3A	GGAGAAAGATCCCTTTGAG	748
NM_007054	KIF3A	TATTGGGCCAGCAGATTAC	749
NM_007054	KIF3A	TTATGACGCTAGGCCACAA	750
NM_020242	KNSL7	GCACAACTCCTGCAAATTC	751
NM_020242	KNSL7	GATGGAAGAGCCTCTAAGA	752
NM_020242	KNSL7	ACGAAAAGCTGCTTGAGAG	753
NM_001184	ATR	TCACGACTCGCTGAACTGT	754
NM_001184	ATR	GAAACTGCAGCTATCTTCC	755
NM_001184	ATR	GTTACAATGAGGCTGATGC	756
NM_014875	KIF14	ATTTTCTAGAAAACGGTAA	757
NM_014875	KIF14	GAGGGGCGAAGTTTCGGCA	758
NM_014875	KIF14	CTGGGACCGGGAAGCCGGA	759
NM_014875	KIF14	CTTCTACTTCTGTTGGCAG	760
NM_014875	KIF14	ACTTACTATTCAGACTGCA	761
NM_014875	KIF14	GCCCTCACCCACAGTAGCC	762
NM_014875	KIF14	CAGAGGAATGCACACCCAG	763
NM_014875	KIF14	GATTGATTAGATCTCTTGA	764

NM_014875	KIF14	GTGAGTATTATCCCAGTTG	765
NM_014875	KIF14	ATCTGGGGTGCTGATTGCT	766
NM_014875	KIF14	GTGACAGTGGCAGTACGCG	767
NM_014875	KIF14	TCAGACTGAAGTTGTTAGA	768
NM_014875	KIF14	GTTGGCTAGAATTGGGAAA	769
NM_014875	KIF14	GAAGACCATAGCATCCGCC	770
NM_001274	CHEK1	TGCCTGAAAGAGACTTGTG	771
NM_001274	CHEK1	ATCGATTCTGCTCCTCTAG	772
NM_001274	CHEK1	CTGAAGAAGCAGTCGCAGT	773
NM_007194	CHEK2	GATCACAGTGGCAATGGAA	774
NM_007194	CHEK2	ATGAATCCACAGCTCTACC	775
NM_007194	CHEK2	AAACTCTTGGAAGTGGTGC	776
NM_000546	TP53	GCACCCAGGACTTCCATTT	777
NM_000546	TP53	CCTCTTGGTCGACCTTAGT	778
NM_000546	TP53	TGAGGCCTTGGAACTCAAG	779
NM_005400	PRKCE	AGCGCCTGGGCCTGGATGA	780
NM_005400	PRKCE	ACCGGGCAGCATCGTCTCC	781
NM_005400	PRKCE	CAGCGGCCAGAGAAGGAAA	782
NM_005400	PRKCE	CAGAAGGAAGAGTGTATGT	783
NM_005400	PRKCE	TGCAGTGTAAAGTCTGCAA	784
NM_005400	PRKCE	GCGCATCGGCCAAACGGCC	785
NM_005400	PRKCE	ATTGCAGAGACTTCATCTG	786
NM_005400	PRKCE	GAAGAGCCGGTACTCACCC	787
NM_005400	PRKCE	AGTACTGGCCGACCTGGGC	788
NM_005400	PRKCE	GGATGCAGAAGGTCACTGC	789
NM_005400	PRKCE	CGTGAGCTTGAAGCCCACA	790
NM_005400	PRKCE	CACAAAGTGTGCTGGGTTA	791
NM_005400	PRKCE	GACGAAGCAATTGTAAAGC	792
NM_005400	PRKCE	CACCCTTCAAACCACGCAT	793
NM_005400	PRKCE	GTCAGCATCTTGAAAGCTT	794
NM_005400	PRKCE	CAACCGAGGAGAGGAGCAC	795
NM_005400	PRKCE	TACATTGCCCTCAATGTGG	796
NM_005400	PRKCE	GAGGAATCGCCAAAGTACT	797
NM_005400	PRKCE	GGGATTTGAAACTGGACAA	798
NM_006218	PIK3CA	TTACACGTTCATGTGCTGG	799
NM_006218	PIK3CA	CACAATCCATGAACAGCAT	800
NM_006218	PIK3CA	CAATCAAACCTGAACAGGC	801
NM_006218	PIK3CA	CAGTTCAACAGCCACACAC	802
NM_006218	PIK3CA	GTGTTACAAGGCTTATCTA	803
NM_006218	PIK3CA	GATCCTATGGTTCGAGGTT	804
NM_006218	PIK3CA	CTCCAAATAATGACAAGCA	805
NM_006218	PIK3CA	ACTTTGCCTTTCCATTTGC	806
NM_006218	PIK3CA	AGAATATCAGGGCAAGTAC	807
NM_006218	PIK3CA	TTGGATCTTCCACACAATT	808
NM_006218	PIK3CA	AGTAGGCAACCGTGAAGAA	809

NM_006218	PIK3CA	CAGGGCTTGCTGTCTCCTC	810
NM_006218	PIK3CA	GAGCCCAAGAATGCACAAA	811
NM_006218	PIK3CA	GCCAGAACAAGTAATTGCT	812
NM_006218	PIK3CA	GGATGCCCTACAGGGCTTG	813
NM_006218	PIK3CA	TCAAATTATTCGTATTATG	814
NM_006218	PIK3CA	GAATTGGAGATCGTCACAA	815
NM_006218	PIK3CA	TGAGGTGGTGCGAAATTCT	816
NM_006218	PIK3CA	GATTTACGGCAAGATATGC	817
NM_006218	PIK3CA	TGATGAATACTTCCTAGAA	818
NM_001982	ERBB3	GCTGCTGGGACTATGCCCA	819
NM_001982	ERBB3	ATCTGCACAATTGATGTCT	820
NM_001982	ERBB3	CTTTGAACTGGACCAAGGT	821
NM_001982	ERBB3	CATCATGCCCACTGCAGGC	822
NM_001982	ERBB3	AACTTTCCAGCTGGAACCC	823
NM_001982	ERBB3	TGAAGGAAATTAGTGCTGG	824
NM_001982	ERBB3	AATTCGCCAGCGGTTCAGG	825
NM_001982	ERBB3	ACCAGAGCTTCAAGACTGT	826
NM_001982	ERBB3	GAGGCTACAGACTCTGCCT	827
NM_001982	ERBB3	TGGAGCCAGAACTAGACCT	828
NM_001982	ERBB3	ACACTGTACAAGCTCTACG	829
NM_001982	ERBB3	TAATGGTCACTGCTTTGGG	830
NM_001982	ERBB3	ACAGGCACTCCTGGAGATA	831
NM_001982	ERBB3	GTTTAGGACAAACACTGGT	832
NM_001982	ERBB3	GATTACTGGCATAGCAGGC	833
NM_001982	ERBB3	ATGAATACATGAACCGGAG	834
NM_001982	ERBB3	CACTTAATCGGCCACGTGG	835
NM_001982	ERBB3	GGCCTGTCCTCCTGACAAG	836
NM_001982	ERBB3	TCTGCGGAGTCATGAGGGC	837
NM_001982	ERBB3	TAGACCTAGACTTGGAAGC	838
NM_004283	RAB3D	GATTTCAGGTCTCCCTGTC	839
NM_004283	RAB3D	GCCACAGTGGTTATCTCCA	840
NM_004283	RAB3D	GCAATCCCTTCCCTCCTGT	841
NM_004283	RAB3D	TCTCTGATCCTGAAGTGAA	842
NM_004283	RAB3D	CATCAATGTGAAGCAGGTC	843
NM_004283	RAB3D	CATGAGCTTGCTGCTTTCC	844
NM_004283	RAB3D	AACGTGTTGTGCCTGCTGA	845
NM_004283	RAB3D	CTGCTTTCCAGGGTGTGTT	846
NM_004283	RAB3D	GCGGCCAGGGCCAAGCCGC	847
NM_004283	RAB3D	CTTCTAGCTTAGAACCATT	848
NM_004283	RAB3D	CAGGGTGTGTTGAGGGTGG	849
NM_004283	RAB3D	CTCTTTCTCAGGTCCTGCA	850
NM_004283	RAB3D	CTTGTGCCAAGATGGCATC	851
NM_004283	RAB3D	GCACCATCACCACGGCCTA	852
NM_004283	RAB3D	CGCGGACGACTCCTTCACT	853
NM_004283	RAB3D	TCATCCAGGGAAGGCGGCG	854

NM_004283	RAB3D	GACACTGACGTGCATGAGC	855
NM_004283	RAB3D	CCCTCCCAGGCCCTGTTTA	856
NM_004283	RAB3D	AGGTCTTCGAGCGCCTGGT	857
NM_004283	RAB3D	CCTCTTTCTCAGGTCCTGC	858
NM_003620	PPM1D	TTGCCCGGGAGCACTTGTG	859
NM_003620	PPM1D	CGTGTGCGACGGGCACGGC	860
NM_003620	PPM1D	ATTAGGTCTTAAAGTAGTT	861
NM_003620	PPM1D	AGCCCTGACTTTAAGGATA	862
NM_003620	PPM1D	TGTGGAGCCCGAACCGACG	863
NM_003620	PPM1D	GCGACGGGCACGGCGGGCG	864
NM_003620	PPM1D	GATTATATGGGTATATATT	865
NM_003620	PPM1D	TTAGAAGGAGCACAGTTAT	866
NM_003620	PPM1D	CCGGCCAGCCGGCCATGGC	867
NM_003620	PPM1D	GAGCAGATAACACTAGTGC	868
NM_003620	PPM1D	AGATGCCATCTCAATGTGC	869
NM_003620	PPM1D	GCGGCACAGTTTGCCCGGG	870
NM_003620	PPM1D	CGTAGCAATGCCTTCTCAG	871
NM_003620	PPM1D	TATATGGGTATATATTCAT	872
NM_003620	PPM1D	GCTGCTAATTCCCAACATT	873
NM_003620	PPM1D	ACAACTGCCAGTGTGGTCA	874
NM_003620	PPM1D	TTGACCCTCAGAAGCACAA	875
NM_003620	PPM1D	GTCTTAAAGTAGTTACTCC	876
NM_003620	PPM1D	ATGCTCCGAGCAGATAACA	877
NM_003620	PPM1D	GCGCCTAGTGTGTCTCCCG	878
NM_022048	CSNK1G1	TAGCCATCCAGCTGCTTTC	879
NM_022048	CSNK1G1	TTCTCATTGGAAGGGACTC	880
NM_022048	CSNK1G1	CACGCATCTTGGCAAAGAG	881
NM_022048	CSNK1G1	TAGCTTGGAGGACTTGTTT	882
NM_022048	CSNK1G1	ACTCAATTGTACCTGCAGC	883
NM_022048	CSNK1G1	CTAAGTGCTGCTGTTTCTT	884
NM_022048	CSNK1G1	GCAAAGCCGGAGAGATGAT	885
NM_022048	CSNK1G1	CCTCTTCACAGACCTCTTT	886
NM_022048	CSNK1G1	GAAGGGACTCCTCTTTGGG	887
NM_022048	CSNK1G1	GAGAGCTCAGATTAGGTAA	888
NM_022048	CSNK1G1	CACGTAGATTCTGGTGCAT	889
NM_022048	CSNK1G1	ATGAGTATTTACGGACCCT	890
NM_022048	CSNK1G1	GGTGGGACCCAACTTCAGG	891
NM_022048	CSNK1G1	AGAGCTGAATGTTGATGAT	892
NM_022048	CSNK1G1	GATTCTGGTGCATCTGCAA	893
NM_022048	CSNK1G1	AACTTCAGGGTTGGCAAGA	894
NM_022048	CSNK1G1	TCTCGAATGGAATACGTGC	895
NM_022048	CSNK1G1	CCGAGGAGAGTGGGAAATT	896
NM_022048	CSNK1G1	GGGAGCCCACTCCAATGCA	897
NM_022048	CSNK1G1	GTCAAGCCAGAGAACTTCC	898
NM_000082	CKN1	TTAGCAGTTTCCTGGTCTC	899

NM_000082	CKN1	ATGTGAGAAGAGCATCAGG	900
NM_000082	CKN1	AGCAGTGTGTTCCATTGGC	901
NM_000082	CKN1	GGATCCTGTTCTCACATTC	902
NM_000082	CKN1	CAGCAGTGATGAAGAAGGA	903
NM_000082	CKN1	GATAACTATGCTTAAGGGA	904
NM_000082	CKN1	TGGACTTCACCTCCTCACT	905
NM_000082	CKN1	TTGAAGTCTGGATCCTGTT	906
NM_000082	CKN1	AGGAACTTTATAGTGGTAG	907
NM_000082	CKN1	AAGTGATGGACTTCACCTC	908
NM_000082	CKN1	TGTTTATACAGTTTACTCA	909
NM_000082	CKN1	GAAGGGAGATACATGTTAT	910
NM_000082	CKN1	GGGTTTGGAGGACCCTCTT	911
NM_000082	CKN1	ATATGTCTCCAGTCTCCAC	912
NM_000082	CKN1	GATGGACTTCACCTCCTCA	913
NM_000082	CKN1	TGAAAGTATGGGATACAAA	914
NM_000082	CKN1	ATGTAAAGCAGTGTGTTCC	915
NM_000082	CKN1	TCTACAGGGTCACAGACAA	916
NM_000082	CKN1	GAGGCCATCAGTATTGACT	917
NM_000082	CKN1	ACTGTTTGGTAGCAGTTGG	918
NM_002843	PTPRJ	AGGAGGAGGCGAAGGAGAC	919
NM_002843	PTPRJ	CTACGTCACCACCACGGAG	920
NM_002843	PTPRJ	TCGCCTAATTCCAAAGGAA	921
NM_002843	PTPRJ	CAAGTATGTAGTAAAGCAT	922
NM_002843	PTPRJ	AAGCTGGTCACCCTTCTGC	923
NM_002843	PTPRJ	CACAGAAGGTGGCTTGGAT	924
NM_002843	PTPRJ	TGGAATCTAGCCGATGGAA	925
NM_002843	PTPRJ	ATAAACAGAATGGAACTGG	926
NM_002843	PTPRJ	CCTGGAGAGCTGCTCCTCT	927
NM_002843	PTPRJ	AACTTTAAGTTGGCAGAAC	928
NM_002843	PTPRJ	ACACAGTGGAGATCTTTGC	929
NM_002843	PTPRJ	CAGTACACACGGCCCAGCA	930
NM_002843	PTPRJ	TTGAACAGGGAAGAACCAA	931
NM_002843	PTPRJ	ATTATGTTGACTAAATGTG	932
NM_002843	PTPRJ	TGACTCAAGACTCAAGACT	933
NM_002843	PTPRJ	AACTTTCGGTCCAGACCCA	934
NM_002843	PTPRJ	GGCCAGACCACGGTGTTCC	935
NM_002843	PTPRJ	TCACTGGAACCTGGCCGGA	936
NM_002843	PTPRJ	ACACAGGAGGGAGCTGGCA	937
NM_002843	PTPRJ	TGTTCTCATTTGATCAGGG	938
NM_004037	AMPD2	TCATCCGGGAGAAGTACAT	939
NM_004037	AMPD2	ACCCAACTATACCAAGGAA	940
NM_004037	AMPD2	CCTGCATGAACCAGAAGCA	941
NM_004037	AMPD2	CTGCGGGAGGTCTTTGAGA	942
NM_004037	AMPD2	GCCTCTTTGATGTGTACCG	943
NM_004037	AMPD2	GACAACATGAGAAATCGTG	944
NM_004037	AMPD2	GCCACCCAGTGAAAGCAAA	945

NM_004037	AMPD2	CAGGAACACTTTCCATCGC	946
NM_004037	AMPD2	TGTGGGAGAGGCAGCTGCC	947
NM_004037	AMPD2	GCCGTGAACAGACGCTGCG	948
NM_004037	AMPD2	AAATATCCCTTTAAGAAGC	949
NM_004037	AMPD2	GTAAAGAGCCACTGGCTGG	950
NM_004037	AMPD2	CGTCCTGCATGAACCAGAA	951
NM_004037	AMPD2	GCTCAGCAACAACAGCCTC	952
NM_004037	AMPD2	CACATCATCAAGGAGGTGA	953
NM_004037	AMPD2	CTCATTGTTGTTTGGGCTC	954
NM_004037	AMPD2	AAGCTCAGCTCCTGCGATA	955
NM_004037	AMPD2	TGCGATATGTGTGAGCTGG	956
NM_004037	AMPD2	CTGGGCCCATCCACCACCT	957
NM_004037	AMPD2	GAAGGACCAGCTAGCCTGG	958
NM_016218	POLK	TATTTCATTTCTTGTCAAT	959
NM_016218	POLK	GACGAGGGATGGAGAGAGG	960
NM_016218	POLK	AGTAGATTGTATAGCTTTA	961
NM_016218	POLK	TATAGATAACTCATCTAAA	962
NM_016218	POLK	AAGAACTTTGCAGTGAGCT	963
NM_016218	POLK	GAATTAGAACAAAGCCGAA	964
NM_016218	POLK	TGTGCTATCAATGAGTTCT	965
NM_016218	POLK	ACACCTGACGAGGGATGGA	966
NM_016218	POLK	TGCATCTACAGTTTCATCT	967
NM_016218	POLK	ACACACCTGACGAGGGATG	968
NM_016218	POLK	TGGATAGCACAAAGGAGAA	969
NM_016218	POLK	AGGGTGCATCAGTCTGGAA	970
NM_016218	POLK	TATAGCTTTAGTAGATACT	971
NM_016218	POLK	TGTTTCTACTGCAGAAGAA	972
NM_016218	POLK	GTTGTTTCTACTGCAGAAG	973
NM_016218	POLK	CTGACAAAGATAAGTTTGT	974
NM_016218	POLK	GCATCAGTCTGGAAGCCTT	975
NM_016218	POLK	CTCAGGATCTACAGAAAGA	976
NM_016218	POLK	AAGGAGATTTGGTGTTCGT	977
NM_016218	POLK	TAGTGCACATTGACATGGA	978

REIVINDICACIONES

1. Un método para seleccionar, a partir de una pluralidad de ARNip diferentes, uno o más ARNip para silenciar un gen diana en un organismo, dirigiéndose cada ARNip diferente en dicha pluralidad de ARNip diferentes, a una secuencia diana diferente en un transcrito de dicho gen diana, comprendiendo dicho método

(a) calcular una puntuación para un motivo de secuencia dirigido correspondiente en dicho transcrito, para cada dicho ARNip diferente en dicha pluralidad de ARNip diferentes, en donde dicha puntuación se calcula utilizando una matriz de puntuación específica de posición (PSSM); en donde cada uno de dichos motivos de secuencia dirigidos comprende al menos una parte de la secuencia diana del ARNip correspondiente y/o una segunda secuencia en una región que flanquea dicha secuencia diana;
(b) clasificar dicha pluralidad de ARNip diferentes de acuerdo con dichas puntuaciones; y
(c) seleccionar uno o más ARNip de dichos ARNip clasificados;

en el que al menos una de las etapas (a), (b) o (c) se realiza mediante un ordenador programado adecuadamente.

2. El método de la reivindicación 1, en el que cada uno de dichos motivos de secuencia dirigidos comprende dicha secuencia diana de dicho ARNip correspondiente.

3. El método de la reivindicación 2, en el que cada uno de dichos motivos de secuencia dirigidos es una secuencia de nucleótidos de *L* nucleótidos, siendo *L* un número entero y en el que dicha PSSM es {log (e_{ij}/p_{ij})}, donde e_{ij} es el peso del nucleótido *i* en la posición *j*, p_{ij} es el peso del nucleótido *i* en la posición *j* en una secuencia al azar, e *i = G, C, A, U(T), j = 1, ..., L*.

4. El método de la reivindicación 2, en el que cada uno de dichos motivos de secuencia dirigidos es una secuencia de nucleótidos de *L* nucleótidos, siendo *L* un número entero y en donde dicha PSSM es {log (e_{ij}/p_{ij})}, donde e_{ij} es el peso del nucleótido *i* en la posición *j*, p_{ij} es el peso del nucleótido *i* en la posición *j* en una secuencia al azar, e *i = G o C, A, U(T), j = 1, ..., L*.

5. El método de la reivindicación 4, en el que dicha puntuación para cada dicho ARNip diferente se calcula de acuerdo con la ecuación

$$\textit{Puntuación} = \sum_{t=1}^{L} \ln(e_t / p_t)$$

en la que e_t representa el peso del nucleótido en la posición *t* en cada uno de dichos motivos de secuencia dirigidos, como se determina de acuerdo con dicha PPSM, y p_t representa el peso del nucleótido en la posición *t* en una secuencia aleatoria.

6. El método de una cualquiera de las reivindicaciones 1-5, en el que cada uno de dichos motivos de secuencia dirigidos comprende dicha secuencia diana de dicho ARNip correspondiente y al menos una secuencia flanqueante.

7. El método de la reivindicación 6, en el que cada uno de dichos motivos de secuencia dirigidos comprende dicha secuencia diana de dicho ARNip correspondiente y una secuencia flanqueante en 5' y una secuencia flanqueante en 3'.

8. El método de la reivindicación 7, en el que cada una de dicha secuencia flanqueante en 5' y dicha secuencia flanqueante en 3', es una secuencia de *D* nucleótidos, siendo *D* un número entero.

9. El método de la reivindicación 8, en el que cada una de dicha secuencia diana es una secuencia de 19 nucleótidos y cada una de dichas secuencia flanqueante en 5' y secuencia flanqueante en 3' es una secuencia de 10 nucleótidos.

10. El método de la reivindicación 7, en el que cada una de dicha secuencia diana es una secuencia de 19 nucleótidos y cada una de dichas secuencia flanqueante en 5' y secuencia flanqueante en 3' es una secuencia de 50 nucleótidos.

11. El método de la reivindicación 9, en el que dichos uno o más ARNip consisten en al menos 3 ARNip.

12. El método de la reivindicación 11, que adicionalmente comprende una etapa de seleccionar adicionalmente una pluralidad de ARNip entre dichos al menos 3 ARNip, de tal manera que los ARNip en dicha pluralidad de ARNip son suficientemente diferentes en una medida de diversidad de secuencia.

13. El método de la reivindicación 12, en el que dicha medida de diversidad de secuencia es una medida cuantificable, y dicha etapa de selección adicional comprende seleccionar ARNip que tengan una diferencia en dicha medida de diversidad de secuencia entre diferentes ARNip seleccionados por encima de un umbral determinado.

14. El método de la reivindicación 13, en el que dicha medida de diversidad de secuencia es el contenido global de GC de cada uno de dichos ARNip.

15. El método de la reivindicación 14, en el que dicho umbral determinado es del 5 %.

16. El método de la reivindicación 13, en el que dicha medida de diversidad de secuencia es la distancia entre los ARNip a lo largo de la secuencia de transcripción.

17. El método de la reivindicación 16, en el que dicho umbral determinado es de 100 nucleótidos.

18. El método de la reivindicación 13, en el que dicha medida de diversidad de secuencia es la identidad del dímero principal de dichos ARNip, en el que a cada uno de los 16 posibles dímeros principales se le asigna una puntuación de 1-16, respectivamente.

19. El método de la reivindicación 18, en el que dicho umbral determinado es de 0,5, en el que todos los dímeros principales se seleccionan con la misma probabilidad.

20. El método de la reivindicación 9, en el que dicha PSSM se obtiene mediante un método que comprende:

(aa) identificar una pluralidad de *N* ARNip que consta de ARNip que tienen una región dúplex de 19 nucleótidos y que tienen una eficacia de silenciamiento por encima de un umbral elegido;
(bb) identificar, para cada uno de dichos *N* ARNip un motivo de secuencia funcional, comprendiendo dicho motivo de secuencia funcional una secuencia diana de 19 nucleótidos de cada uno de dichos *N* ARNip y una secuencia flanqueante en 5' de 10 nucleótidos y una secuencia flanqueante en 3' de 10 nucleótidos;
(cc) calcular una matriz de frecuencia $\{f_{ij}\}$, en la que $i = G, C, A, U(T)$; $j = 1, 2, ..., L$, y en la que f_{ij} es la frecuencia del *i-gésimo* nucleótido en la *j-ésima* posición, basándose en dicho motivo de secuencia funcional de acuerdo con la ecuación

$$f_{ij} = \sum_{k=1}^{N} \delta_{ik}(j),$$

en la que

$$\delta_{ik}(j) = \begin{cases} 1, \text{si } k = i \\ 0, \text{si } k \neq i \end{cases}$$

, en la que *k* es la identidad del nucleótido en la *j-ésima* posición en dicho motivo de secuencia funcional, y

(d) determinar dicha PSSM calculando e_{ij} de acuerdo con la ecuación

$$e_{ij} = \frac{f_{ij}}{N}.$$

21. El método de la reivindicación 20, en el que dicha pluralidad de *N* ARNip se dirigen a una pluralidad de genes diferentes que tienen diferentes abundancias de transcritos en una célula.

22. El método de una cualquiera de las reivindicaciones 1-21, en el que dicha etapa de clasificación se lleva a cabo determinando una puntuación para cada uno de dichos ARNip diferentes, en el que dicha etapa (b) se lleva a cabo seleccionando uno o más ARNip que tienen las puntuaciones más altas.

23. El método de una cualquiera de las reivindicaciones 1-21, en el que dicha etapa de clasificación se lleva a cabo determinando una puntuación para cada uno de dichos ARNip diferentes, en el que dicha etapa (b) se lleva a cabo seleccionando uno o más ARNip que tienen dicha puntuación más próxima a un valor predeterminado, en donde dicho valor predeterminado es el valor de puntuación correspondiente a la mediana máxima de la eficacia de silenciamiento de una pluralidad de motivos de secuencia de ARNip.

24. El método de la reivindicación 23, en el que dicha pluralidad de motivos de secuencia de ARNip son motivos de secuencia en transcritos que tienen niveles de abundancia menores de 3 o menores de 5 copias por célula.

25. El método de una cualquiera de las reivindicaciones 1-21, en el que dicha etapa de clasificación se lleva a cabo determinando una puntuación para cada uno de dichos ARNip diferentes, en el que dicha etapa (b) se lleva a cabo seleccionando uno o más ARNip que tienen dicha puntuación dentro de un intervalo predeterminado, en donde dicho intervalo predeterminado es un intervalo de puntuación correspondiente a una pluralidad de motivos de secuencia de ARNip que tienen cada uno un nivel de eficacia de silenciamiento determinado.

26. El método de la reivindicación 25, en el que dicha eficacia de silenciamiento está por encima del 50 %, 75 % o 90 % a una concentración de ARNip de 100 nM.

27. El método de las reivindicaciones 25 o 26, en el que dicha pluralidad de motivos de secuencia de ARNip son motivos de secuencia en transcritos que tienen niveles de abundancia menores de 3 o menores de 5 copias por célula.

28. El método de una cualquiera de las reivindicaciones 20-27, en el que dicha pluralidad de *N* ARNip comprende al menos 10, 50, 100, 200 o 500 ARNip diferentes.

29. El método de una cualquiera de las reivindicaciones 1-10, en el que dicha PSSM se obtiene mediante un método que comprende

(aa) inicializar dicha PSSM con pesos al azar;
(bb) seleccionar aleatoriamente un peso w_{ij} obtenido en (aa);
(cc) cambiar el valor de dicho peso seleccionado para generar una PSSM de ensayo que comprenda dicho peso seleccionado que tenga dicho valor cambiado;
(dd) calcular una puntuación de ensayo para cada motivo de secuencia funcional de ARNip en una pluralidad de motivos secuencia funcional de ARNip utilizando dicha PSSM de ensayo de acuerdo con la ecuación

$$Puntuación\ de\ ensayo = \sum_{k=1}^{L} \ln(w_k / p_k)$$

en la que dichas w_k y p_k son, respectivamente, pesos de un nucleótido en la posición k en dicho motivo de secuencia funcional y en una secuencia aleatoria;
(ee) calcular la correlación de dicha puntuación de ensayo y una medida de una característica de un ARNip entre dicha pluralidad de motivos de secuencia funcional de ARNip;
(ff) repetir las etapas (cc)-(ee) para una pluralidad de diferentes valores de dicho peso seleccionado en un intervalo determinado y mantener el valor que corresponda a la mejor correlación para dicho peso seleccionado; y
(gg) repetir las etapas (bb)-(ff) durante un número de veces elegido; determinando de este modo dicha PSSM.

30. El método de la reivindicación 29, que adicionalmente comprende seleccionar dicha pluralidad de motivos de secuencia funcional de ARNip mediante un método que comprende:

(i) identificar una pluralidad de ARNip que consiste en ARNip que tienen diferentes valores en dicha medida; y
(ii) identificar una pluralidad de motivos de secuencia funcional de ARNip correspondiendo cada uno a un ARNip en dicha pluralidad de ARNip.

31. El método de la reivindicación 30, en el que dicha característica es eficacia de silenciamiento.

32. El método de la reivindicación 31, en el que dicha pluralidad de *N* ARNip se dirige a una pluralidad de genes diferentes que tienen diferentes abundancias de transcritos en una célula.

33. El método de una cualquiera de las reivindicaciones 30-32, en el que dicha etapa (b) se lleva a cabo seleccionando uno o más ARNip que tienen las puntuaciones más altas.

34. El método de una cualquiera de las reivindicaciones 30-32, en el que dicha etapa (b) se lleva a cabo seleccionando uno o más ARNip que tienen una puntuación más próxima a un valor predeterminado, en donde dicho valor predeterminado es el valor de puntuación correspondiente a la mediana máxima de la eficacia de silenciamiento de una pluralidad de motivos de secuencia de ARNip.

35. El método de una cualquiera de las reivindicaciones 30-34, en el que dicha pluralidad de motivos de secuencia funcional de ARNip son motivos de secuencia en transcritos que tienen niveles de abundancia menores de 3 o menores de 5 copias por célula.

36. El método de una cualquiera de las reivindicaciones 30-32, en el que dicha etapa (b) se lleva a cabo seleccionando uno o más ARNip que tienen una puntación dentro de un intervalo predeterminado, en donde dicho intervalo predeterminado es un intervalo de puntuación correspondiente a una pluralidad de motivos de secuencia de ARNip que tienen cada uno un nivel de eficacia de silenciamiento determinado.

37. El método de la reivindicación 36, en el que dicha eficacia de silenciamiento está por encima del 50 %, 75 % o 90 % a una concentración de ARNip de 100 nM.

38. El método de la reivindicación 37, en el que dicha pluralidad de motivos de secuencia funcional de ARNip son motivos de secuencia en transcritos que tienen niveles de abundancia menores de 3 o menores de 5 copias por célula.

39. El método de una cualquiera de las reivindicaciones 31-38, en el que dicha pluralidad de ARNip comprende al menos 10, 50, 100, 200 o 500 ARNip diferentes.

40. Un sistema informático que comprende

un procesador, y
una memoria acoplada a dicho procesador y que codifica uno o más programas,

en el que dicho uno o más programas permiten que el procesador lleve a cabo el método de una cualquiera de las reivindicaciones 1-20 y 29.

41. Un producto de programa informático para su uso junto con un ordenador que tiene un procesador y una memoria conectada al procesador, comprendiendo dicho producto de programa informático un medio de almacenamiento legible por ordenador que tiene un mecanismo de programa informático codificado en el mismo, en donde dicho mecanismo de programa informático puede cargarse en la memoria del ordenador y permitir que el ordenador lleve a cabo el método de una cualquiera de las reivindicaciones 1-20 y 29.

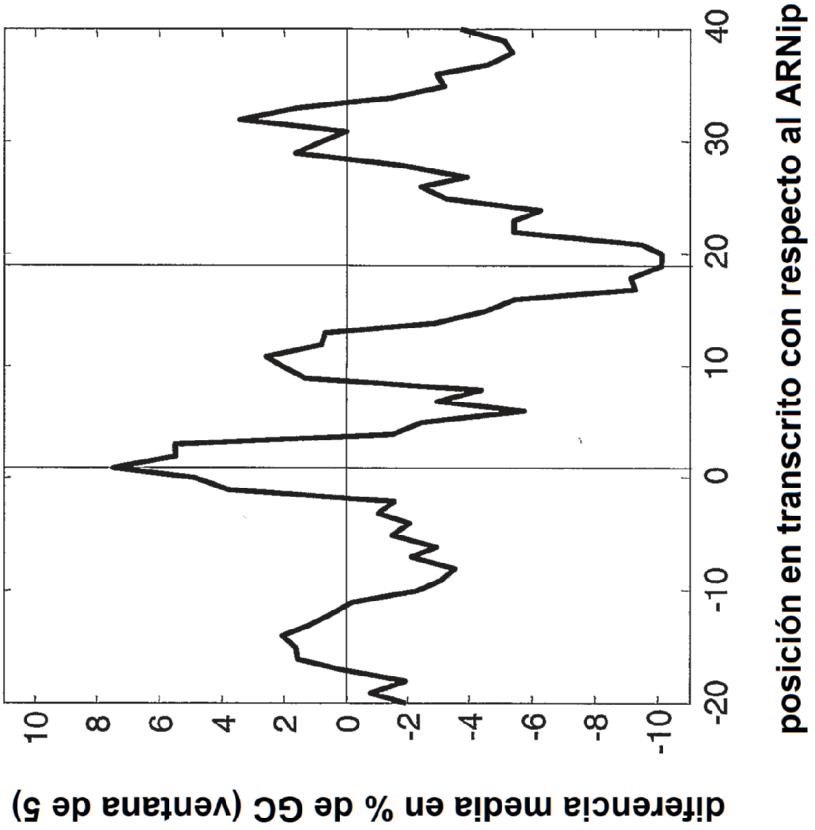

diferencia media en % de GC (ventana de 5)
10
8
6
4
2
0
-2
-4
-6
-8
-10
-20
-10
0
10
20
30
40
posición en transcrito con respecto al ARNip

FIG. 1A

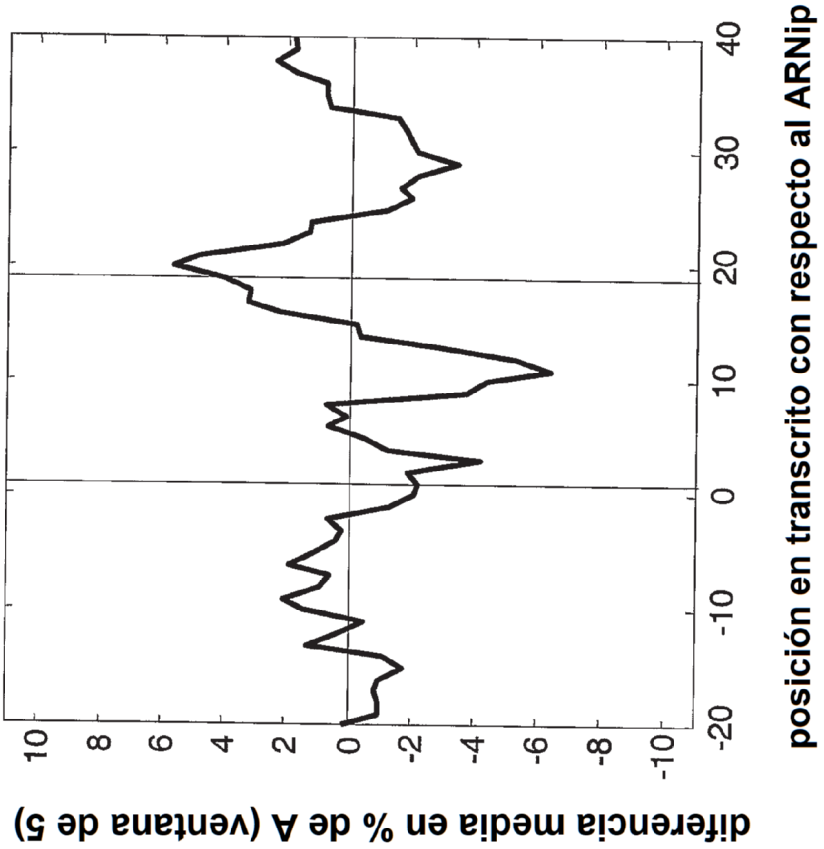

diferencia media en % de A (ventana de 5)
10
8
6
4
2
0
-2
-4
-6
-8
-10
-20
-10
0
10
20
30
40
posición en transcrito con respecto al ARNip

FIG. 1B

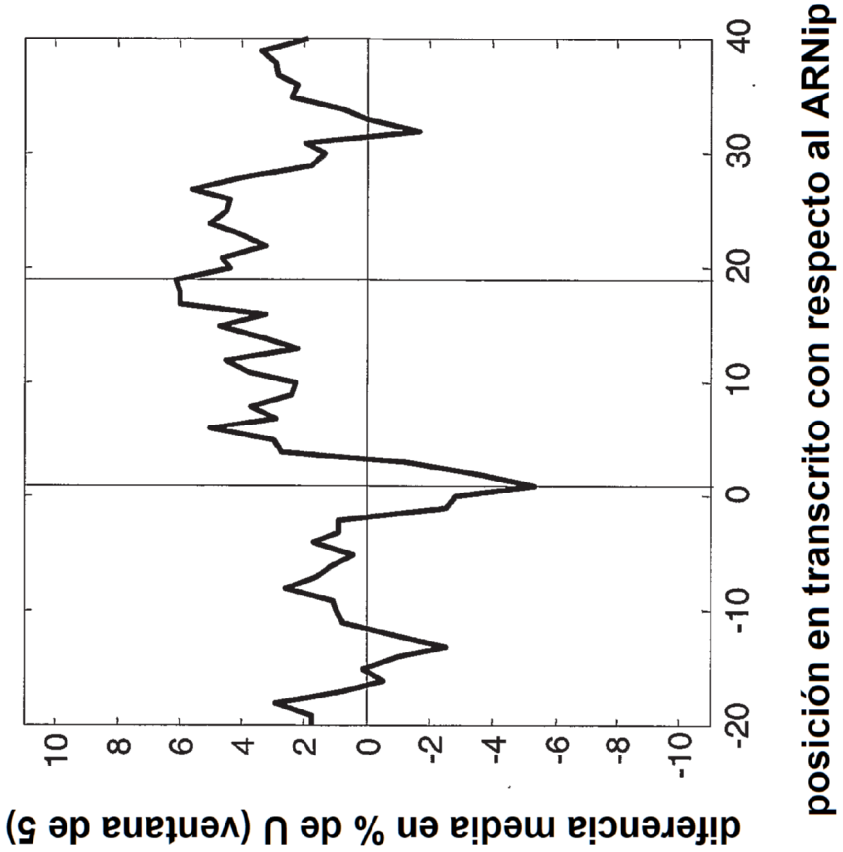

diferencia media en % de U (ventana de 5)
10
8
6
4
2
0
-2
-4
-6
-8
-10
-20
-10
0
10
20
30
40
posición en transcrito con respecto al ARNip

FIG. 1C

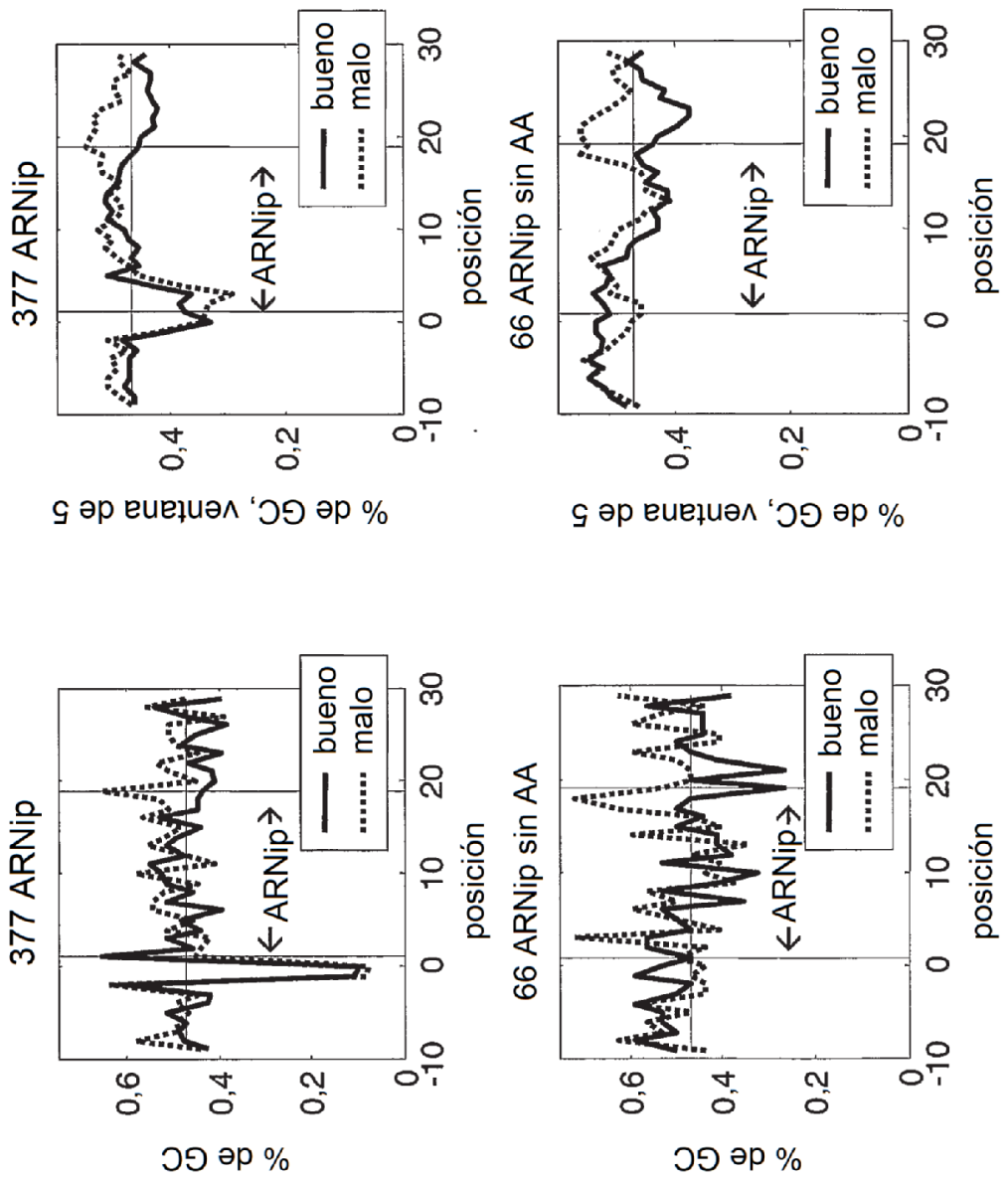

377 ARNip
66 ARNip sin AA
% de GC
% de GC, ventana de 5
posición
←ARNip→
bueno
malo
0,6
0,4
0,2
0
-10
0
10
20
30

FIG. 2A

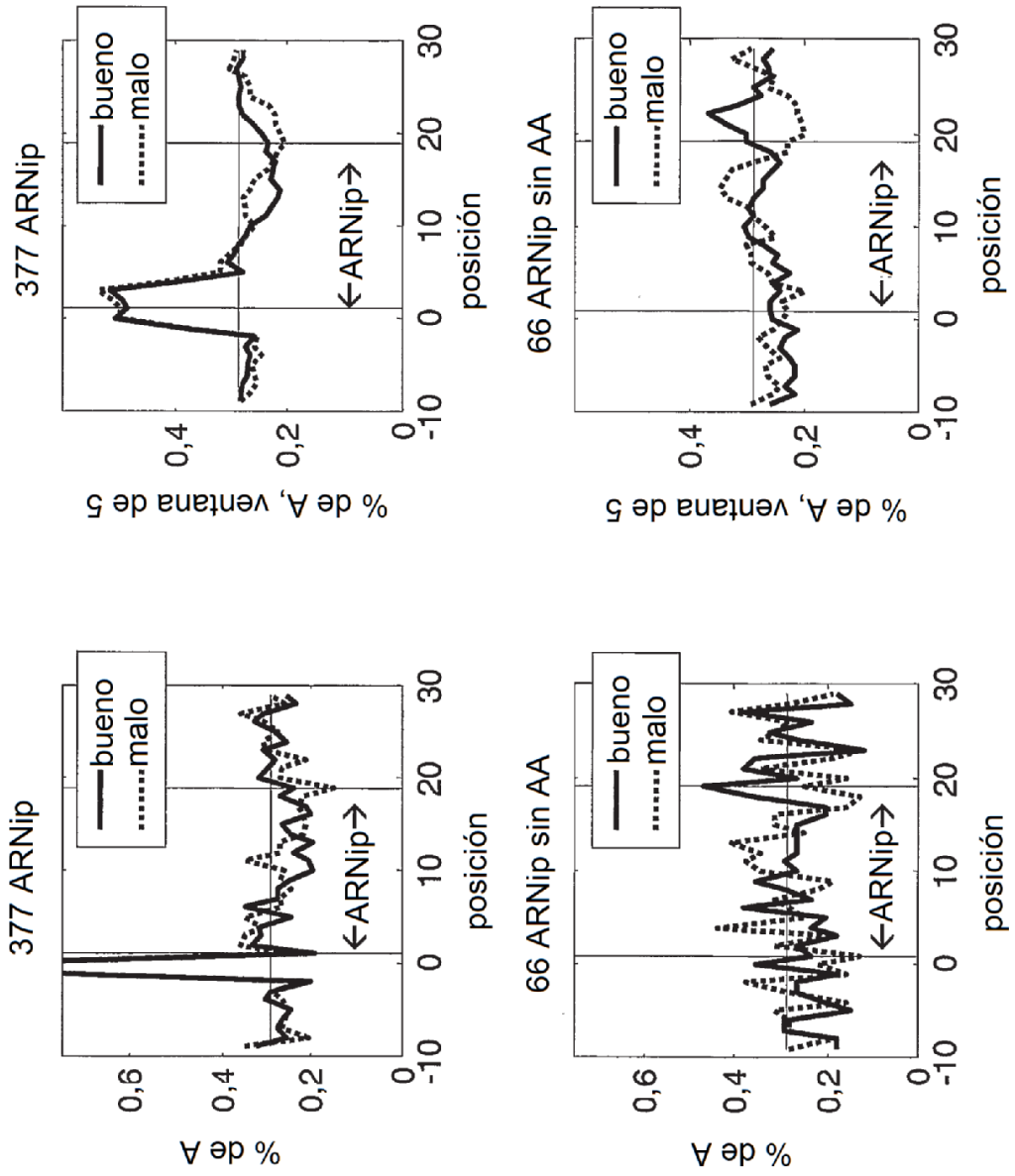

377 ARNip
bueno
malo
←ARNip→
posición
% de A, ventana de 5
66 ARNip sin AA
% de A

FIG. 2B

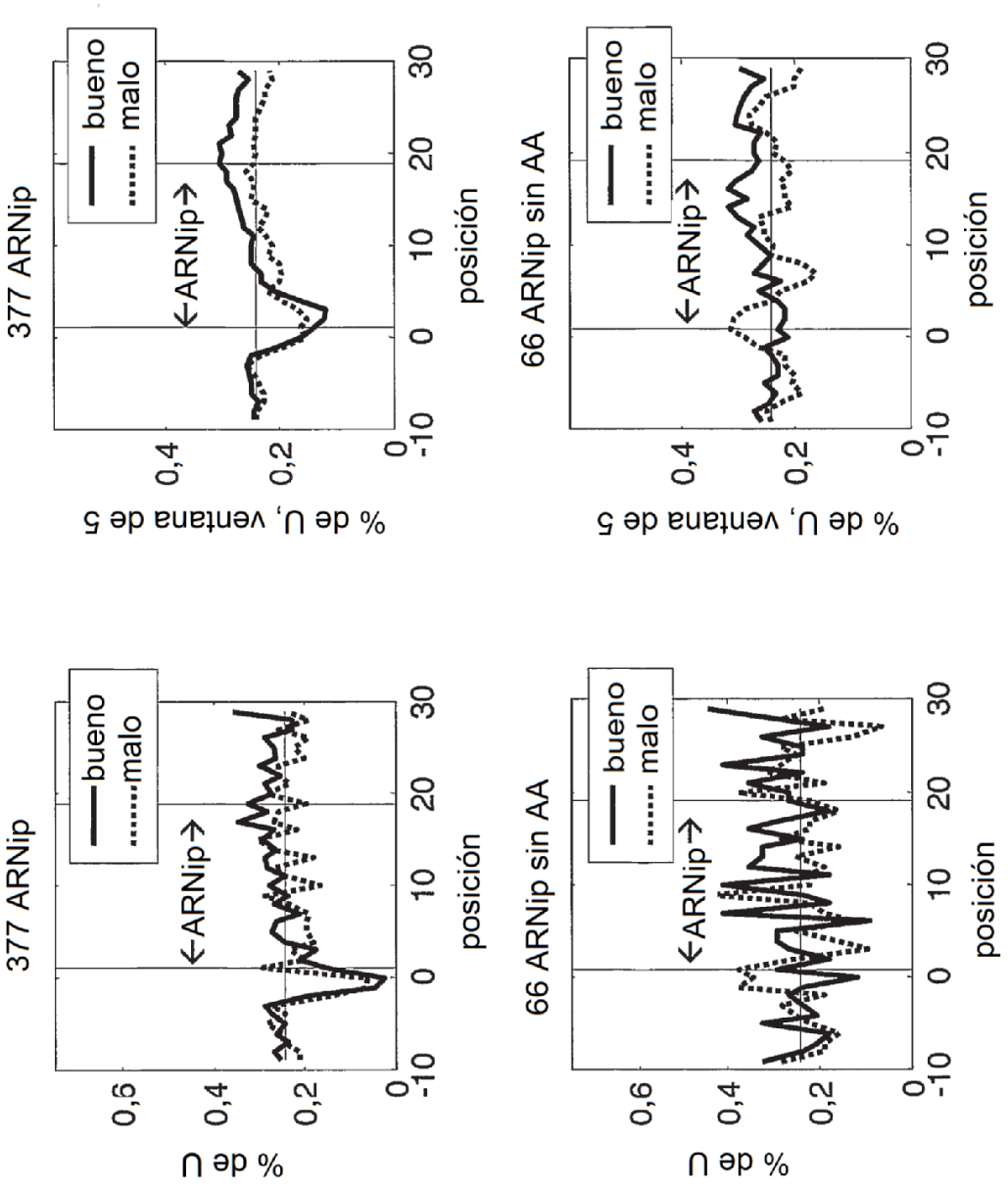

377 ARNip
66 ARNip sin AA
bueno
malo
←ARNip→
% de U
% de U, ventana de 5
posición
0
0,2
0,4
0,6
-10
0
10
20
30

FIG. 2C

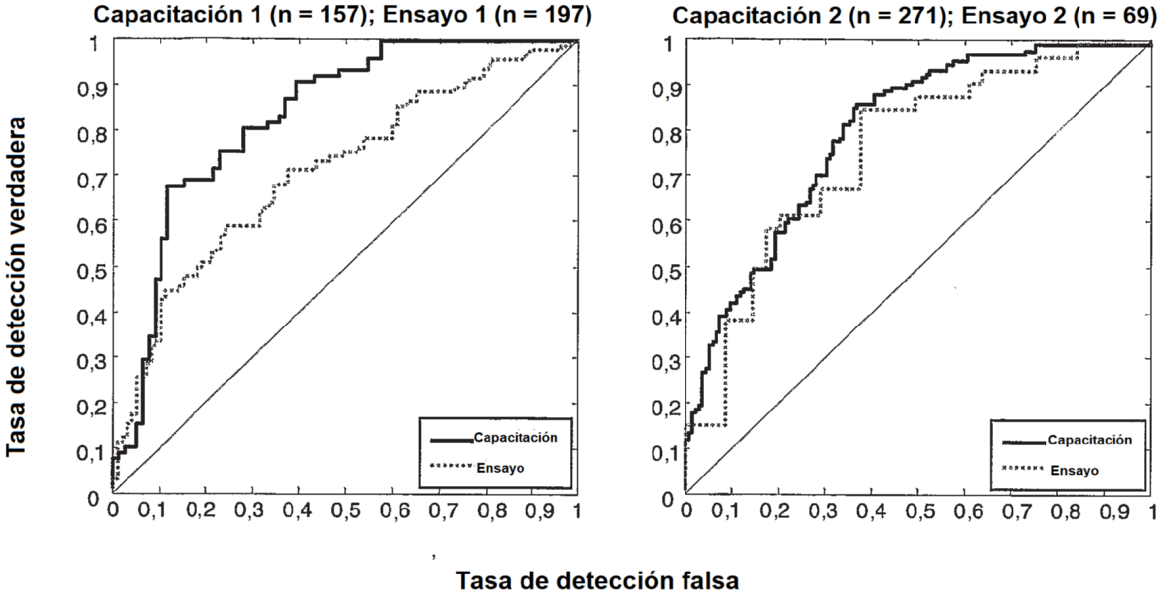

Capacitación 1 (n = 157); Ensayo 1 (n = 197)
Capacitación 2 (n = 271); Ensayo 2 (n = 69)
Tasa de detección verdadera
Tasa de detección falsa
Capacitación
Ensayo
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1

FIG. 3

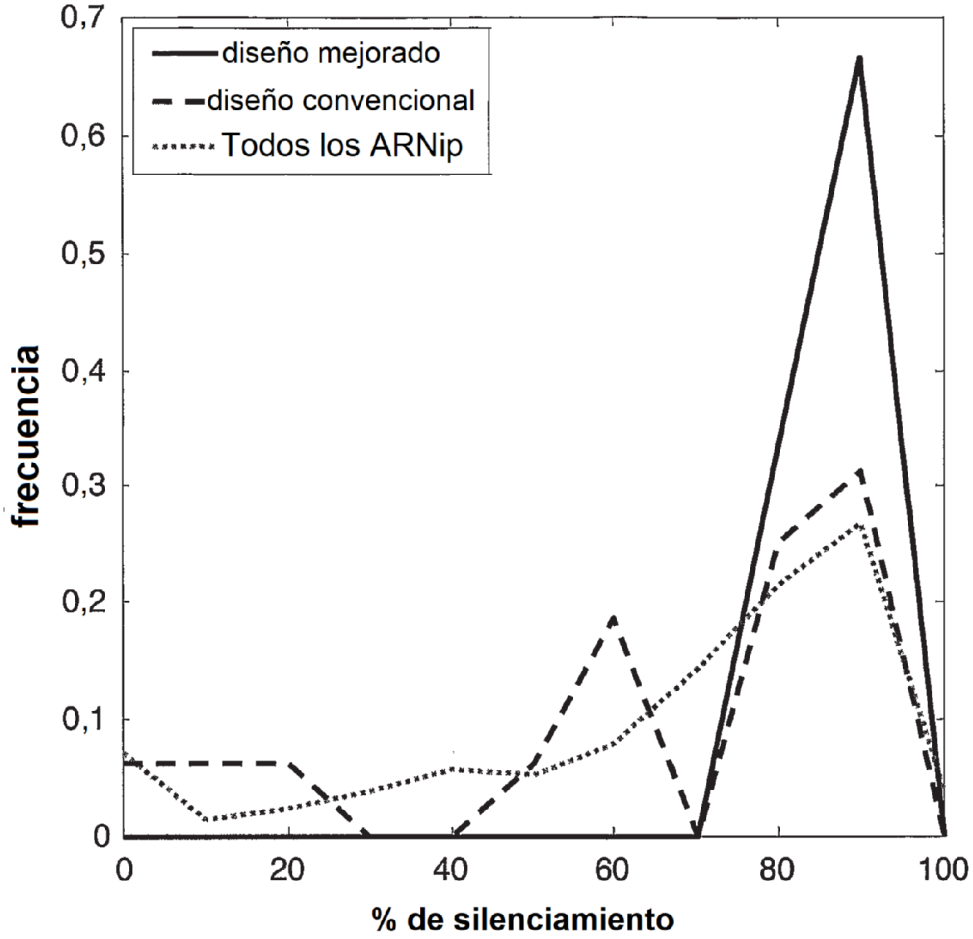
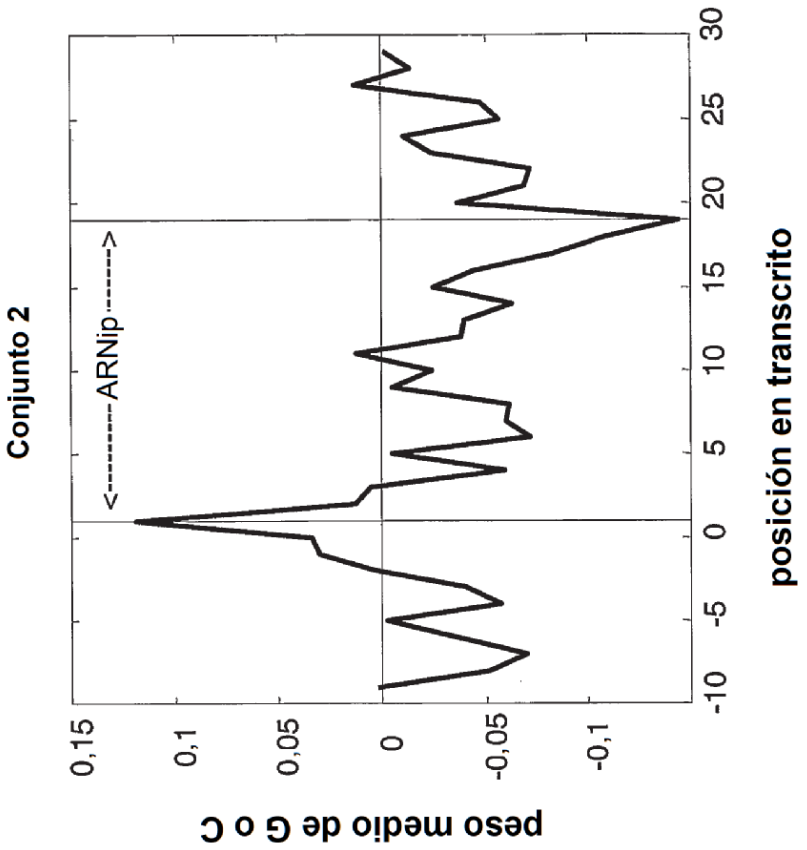

0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
frecuencia
diseño mejorado
diseño convencional
Todos los ARNip
0
20
40
60
80
100
% de silenciamiento

FIG. 4


Conjunto 2
<---------ARNip---------->
peso medio de G o C
posición en transcrito
0,15
0,1
0,05
0
-0,05
-0,1
-10
-5
0
5
10
15
20
25
30

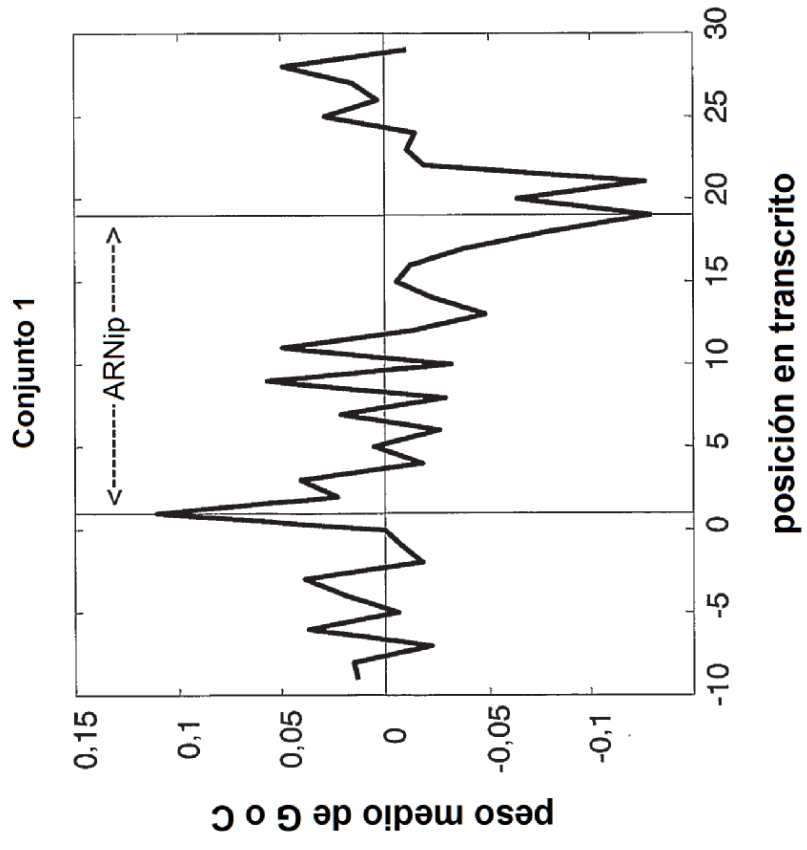

Conjunto 1
<---------ARNip---------->
peso medio de G o C
posición en transcrito
0,15
0,1
0,05
0
-0,05
-0,1
-10
-5
0
5
10
15
20
25
30

FIG. 5A

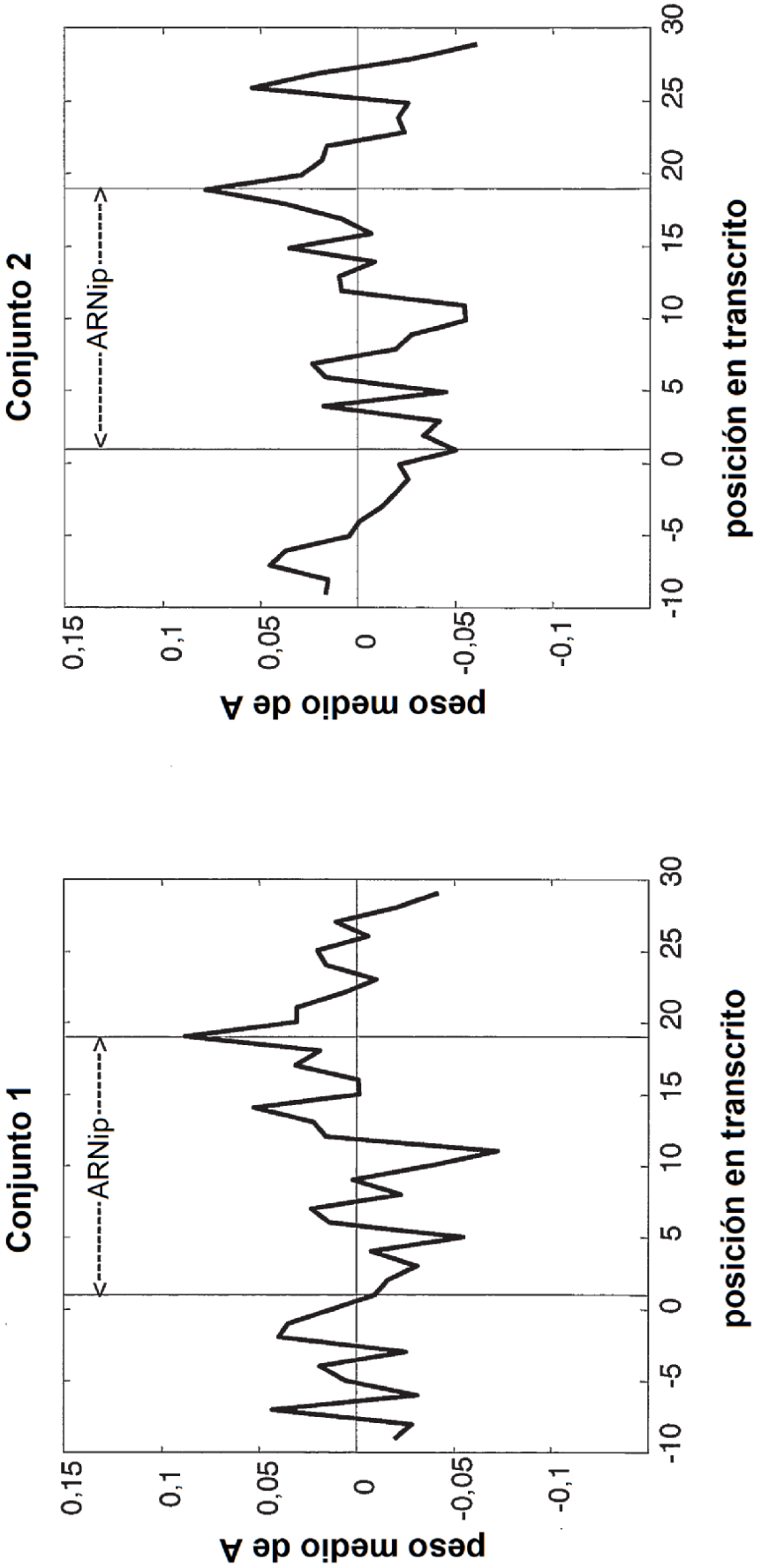
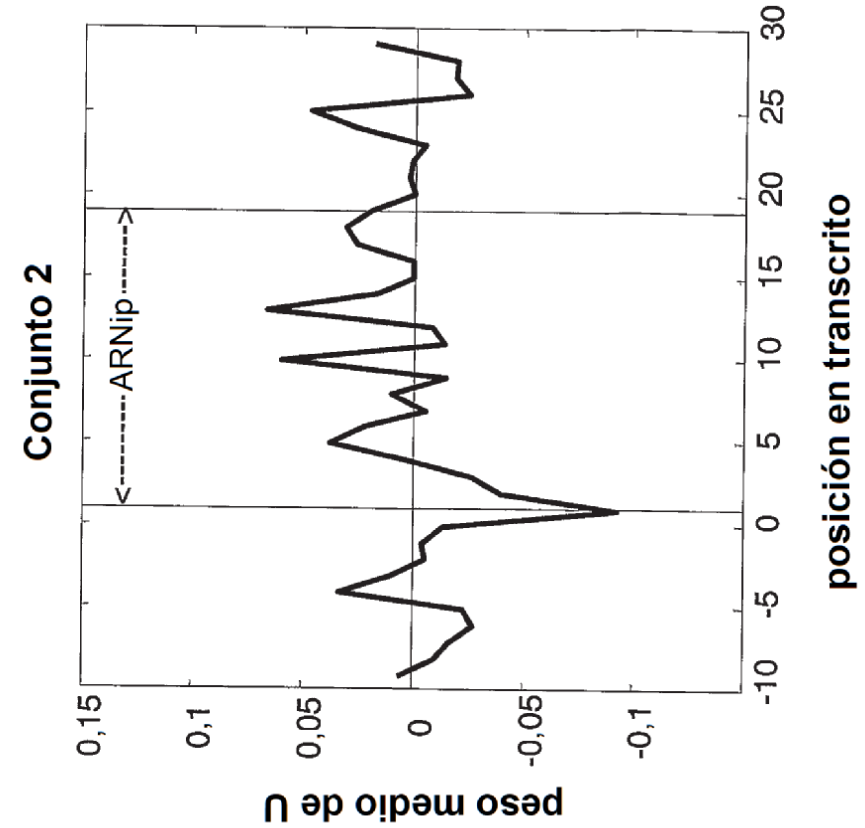

Conjunto 2
ARNip
peso medio de A
posición en transcrito
0,15
0,1
0,05
0
-0,05
-0,1
-10
-5
0
5
10
15
20
25
30
Conjunto 1
ARNip
peso medio de A
posición en transcrito

FIG. 5B


Conjunto 2
ARNip
peso medio de U
0,15
0,1
0,05
0
-0,05
-0,1
-10
-5
0
5
10
15
20
25
30
posición en transcrito

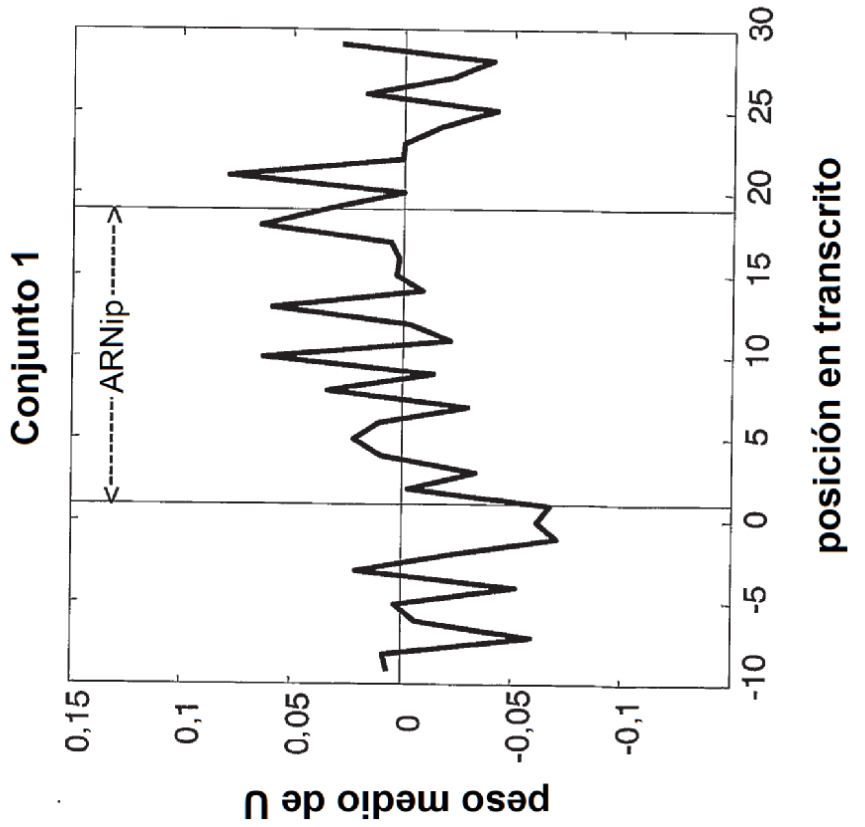

Conjunto 1
ARNip
peso medio de U
0,15
0,1
0,05
0
-0,05
-0,1
-10
-5
0
5
10
15
20
25
30
posición en transcrito

FIG. 5C

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
NM_018461	C	C	A	A	A	A	G	G	T	C	A	T	T	A	C	T	C	A	G
NM_018130	G	A	T	C	C	T	G	A	T	C	A	T	T	A	T	C	T	A	G
NM_005861	C	T	C	C	A	G	G	C	T	C	A	T	T	G	C	C	G	C	G
Cóntig46927_RC	A	T	T	T	T	T	G	G	T	G	A	T	C	A	C	C	G	A	G
Cóntig52414_RC	A	T	C	C	G	G	A	G	T	T	A	T	T	A	C	G	A	A	G
NM_001887	A	A	C	C	G	T	G	C	C	T	A	T	T	A	C	C	A	G	C
NM_004034	A	C	G	A	G	T	G	G	A	G	A	T	T	A	C	C	G	A	A
NM_001156	A	C	G	A	G	T	G	G	A	G	A	T	T	A	C	C	G	A	A
NM_002570	T	G	T	T	C	T	A	A	T	T	T	T	T	A	C	C	G	A	T
Cóntig54761_RC	T	A	A	A	A	A	T	C	T	T	T	T	T	A	C	C	G	A	A
Cóntig53248_RC	G	T	G	G	C	C	T	T	T	T	T	T	T	A	C	C	G	A	T
Cóntig55337_RC	T	G	A	G	A	A	A	A	A	A	G	T	T	A	C	C	G	A	A

FIG. 6

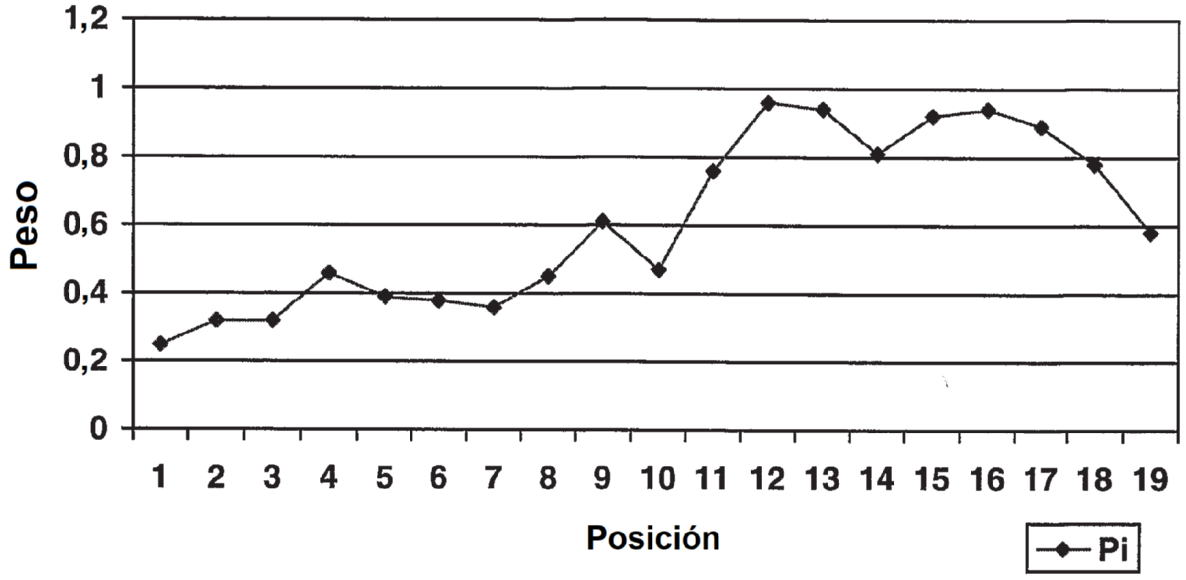

1,2
1
0,8
0,6
0,4
0,2
0
Peso
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Posición
Pi

FIG. 7

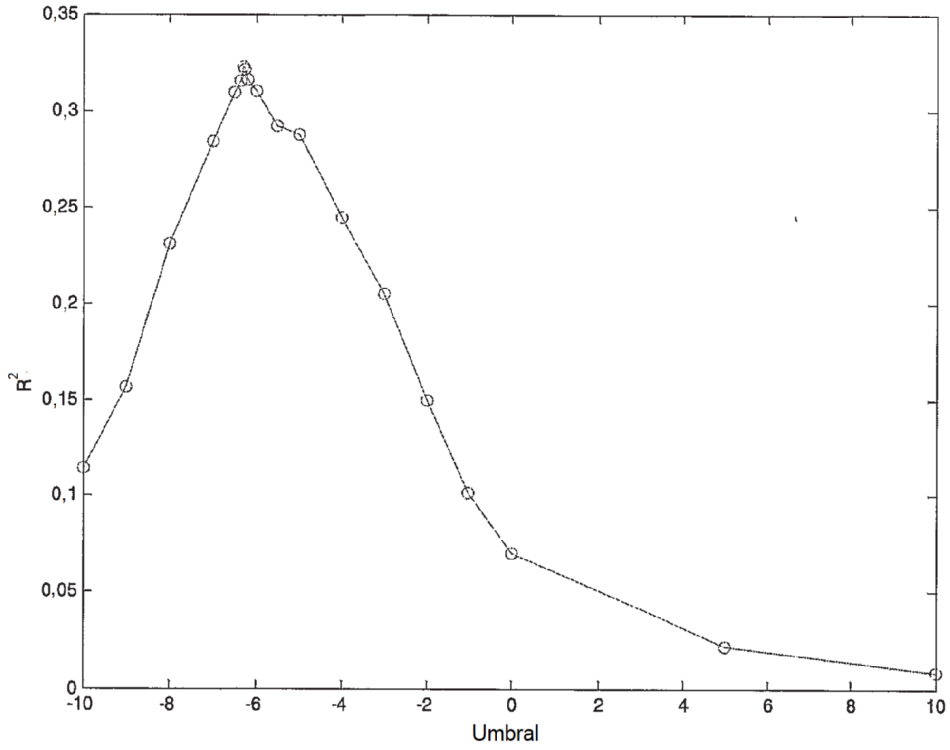


FIG. 8

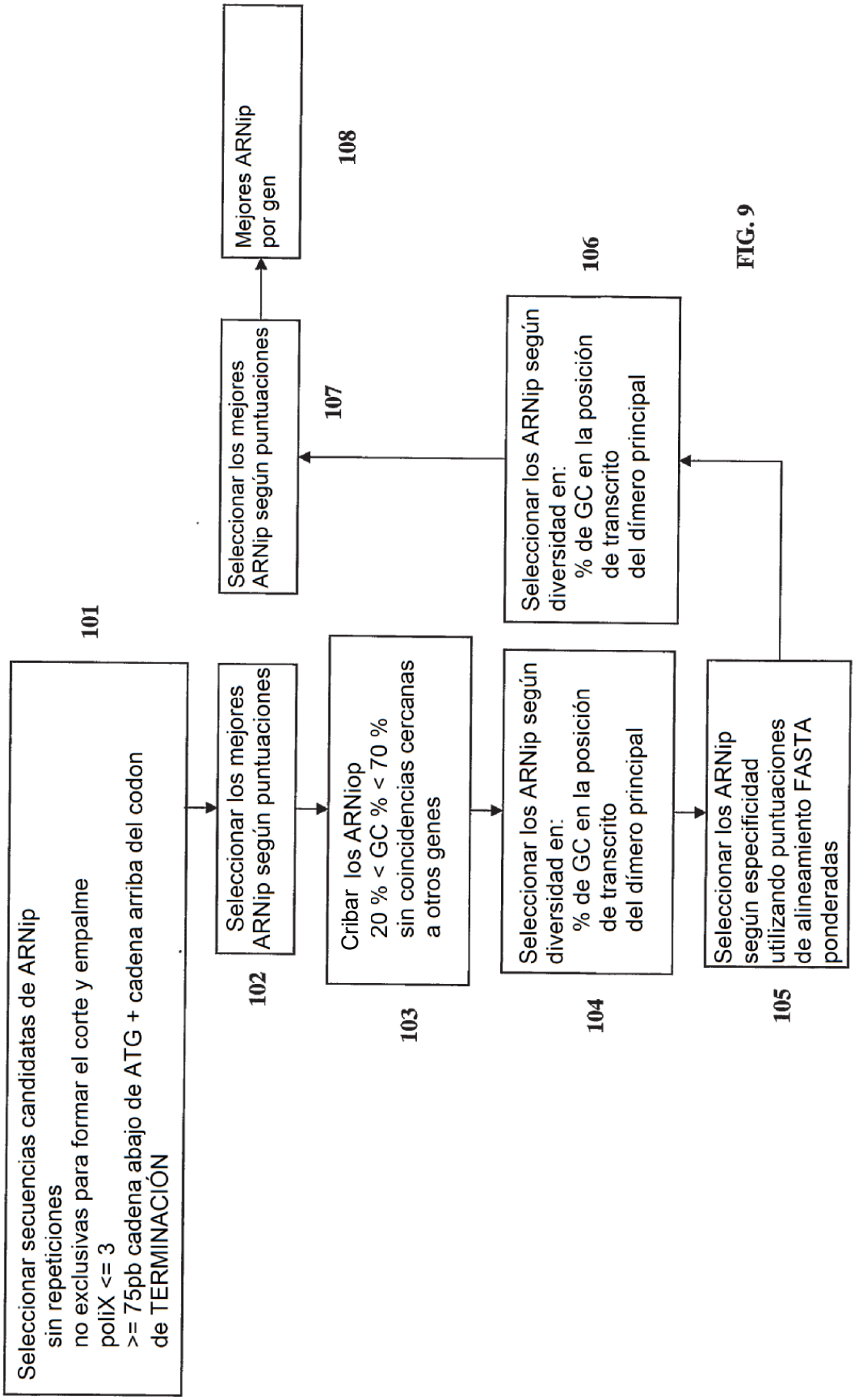

Seleccionar secuencias candidatas de ARNip
sin repeticiones
no exclusivas para formar el corte y empalme
poliX <= 3
>= 75pb cadena abajo de ATG + cadena arriba del codon de TERMINACIÓN
101
Seleccionar los mejores ARNip según puntuaciones
102
Cribar los ARNiop
20 % < GC % < 70 %
sin coincidencias cercanas a otros genes
103
Seleccionar los ARNip según diversidad en:
% de GC en la posición de transcrito
del dímero principal
104
Seleccionar los ARNip según especificidad utilizando puntuaciones de alineamiento FASTA ponderadas
105
Seleccionar los ARNip según diversidad en:
% de GC en la posición de transcrito
del dímero principal
106
Seleccionar los mejores ARNip según puntuaciones
107
Mejores ARNip por gen
108

FIG. 9

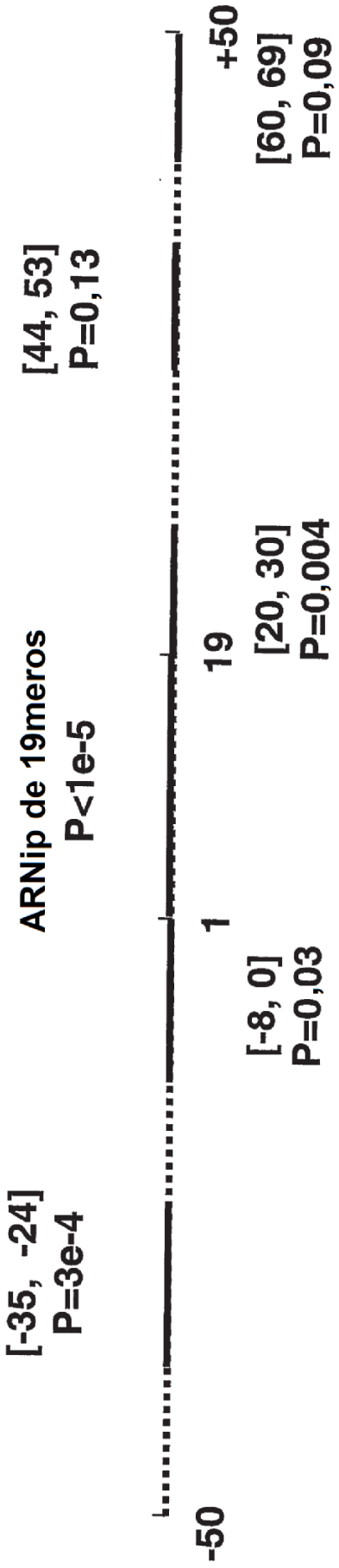

-50
[-35, -24]
P=3e-4
ARNip de 19meros
P<1e-5
1
[-8, 0]
P=0,03
19
[20, 30]
P=0,004
[44, 53]
P=0,13
+50
[60, 69]
P=0,09

FIG. 10

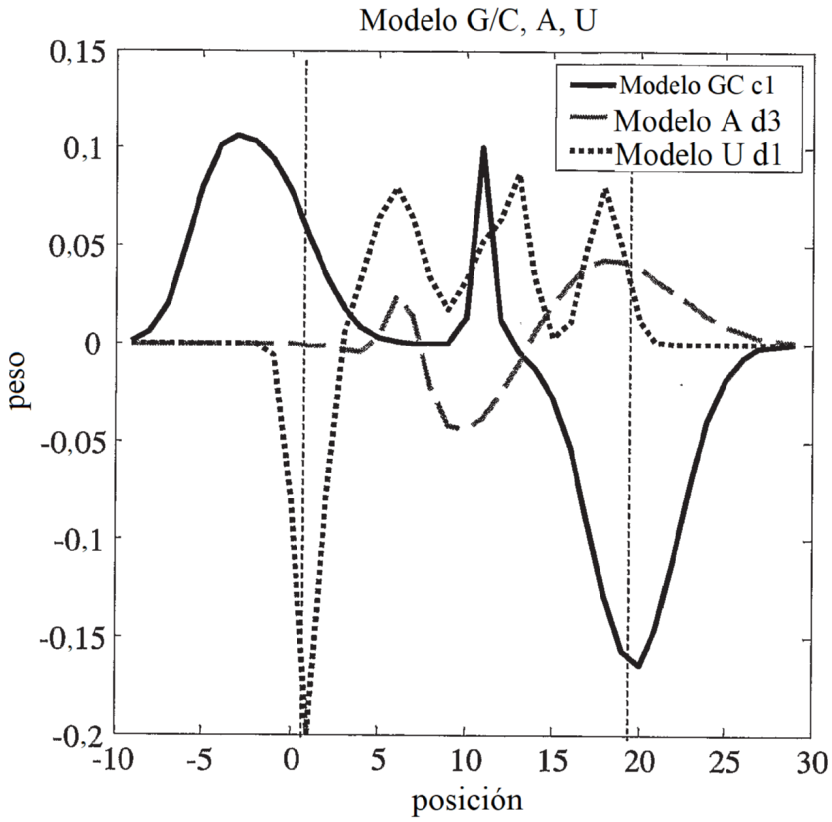

Modelo G/C, A, U
Modelo GC c1
Modelo A d3
Modelo U d1
peso
posición
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
-0,2
-10
-5
0
5
10
15
20
25
30

FIG. 11A

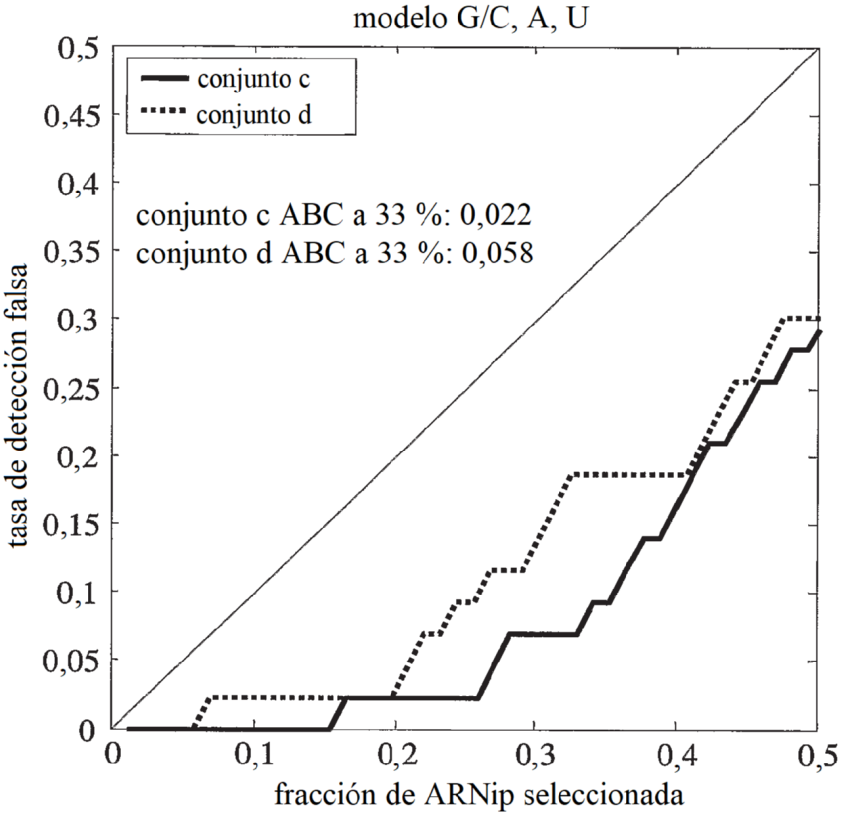

modelo G/C, A, U
conjunto c
conjunto d
conjunto c ABC a 33 %: 0,022
conjunto d ABC a 33 %: 0,058
tasa de detección falsa
fracción de ARNip seleccionada
0
0,05
0,1
0,15
0,2
0,25
0,3
0,35
0,4
0,45
0,5
0
0,1
0,2
0,3
0,4
0,5

FIG. 11B

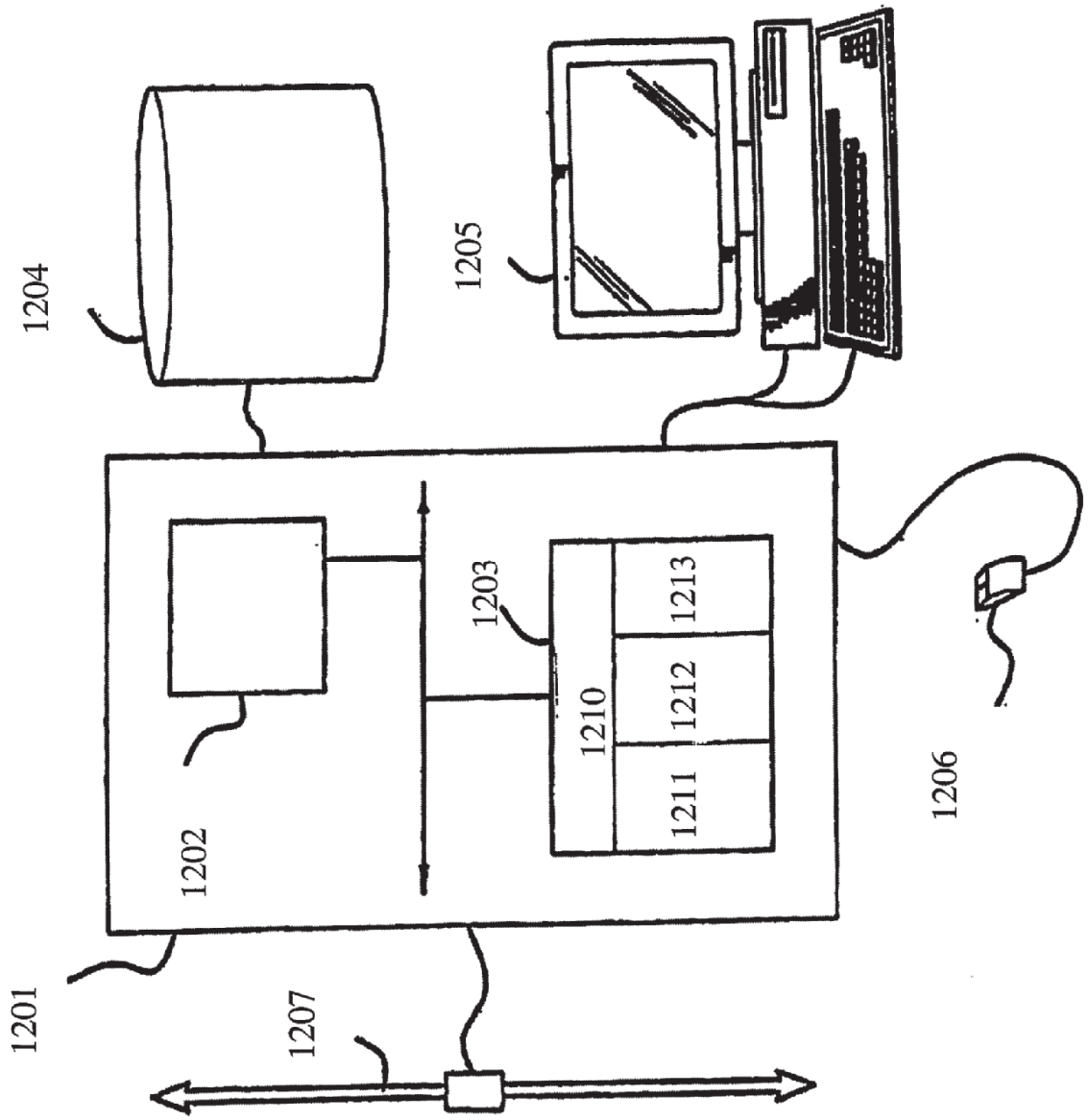


FIG. 12

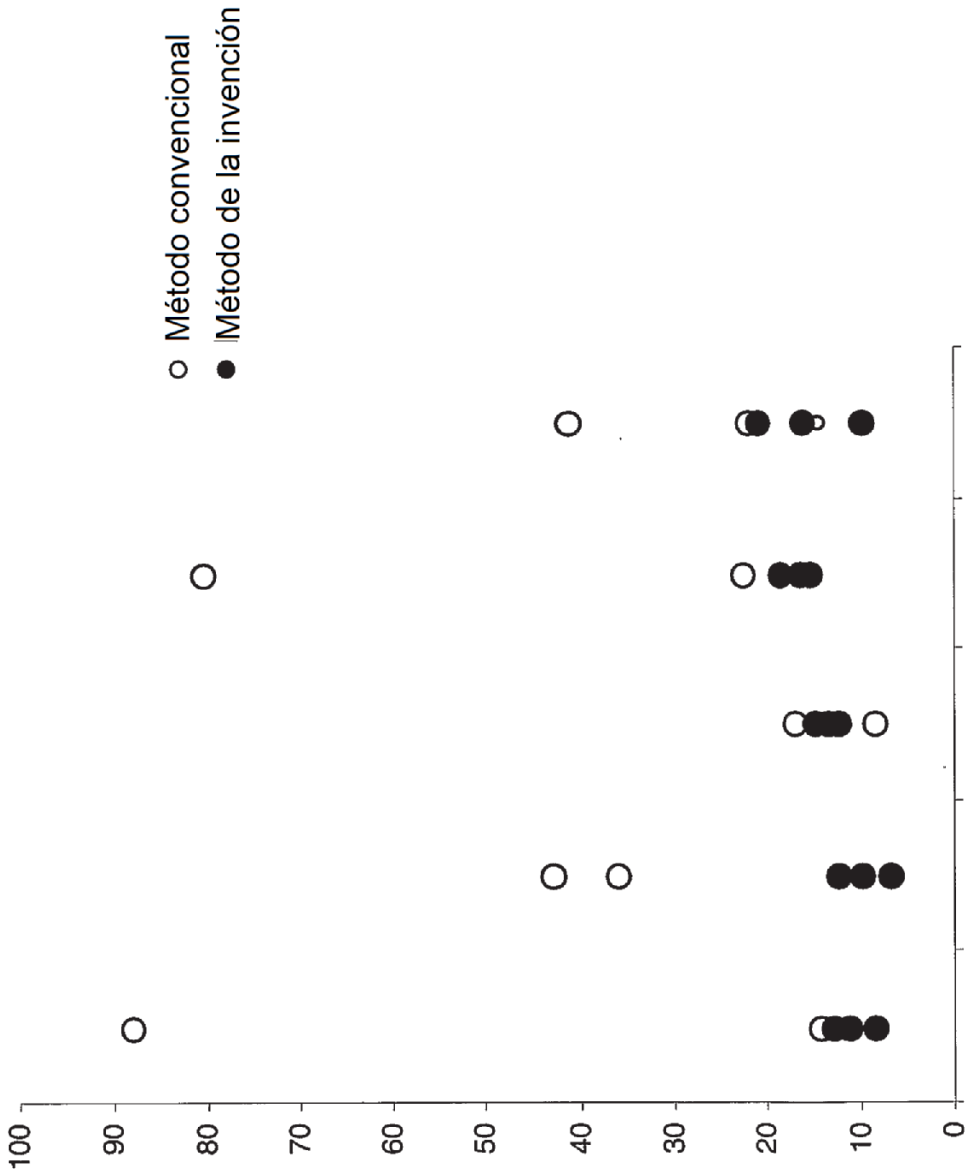

Método convencional
Método de la invención
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5

FIG. 13

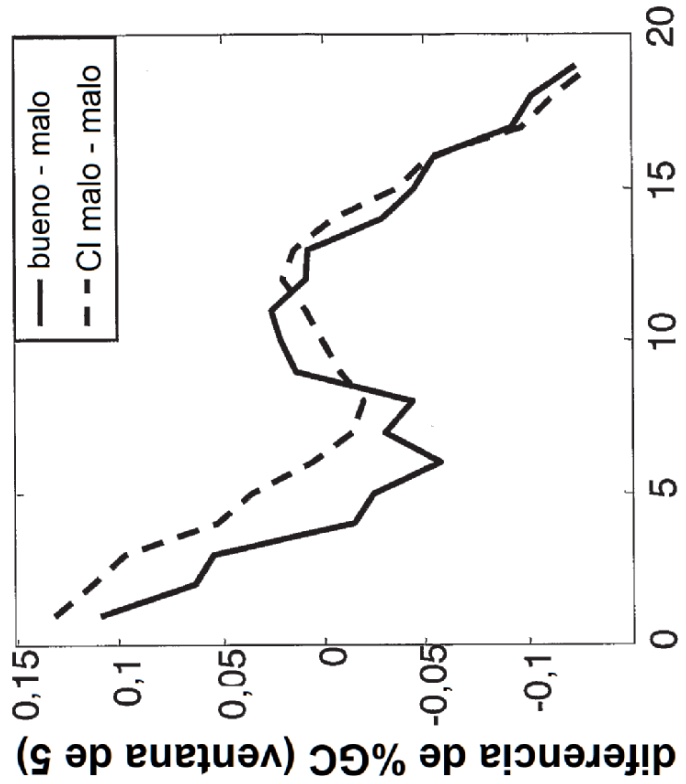


FIG. 14B

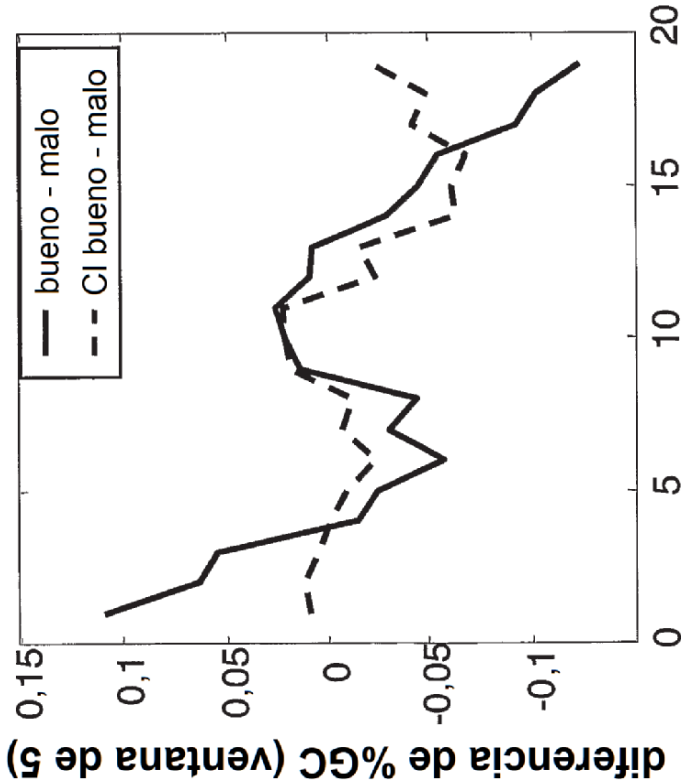


FIG. 14A

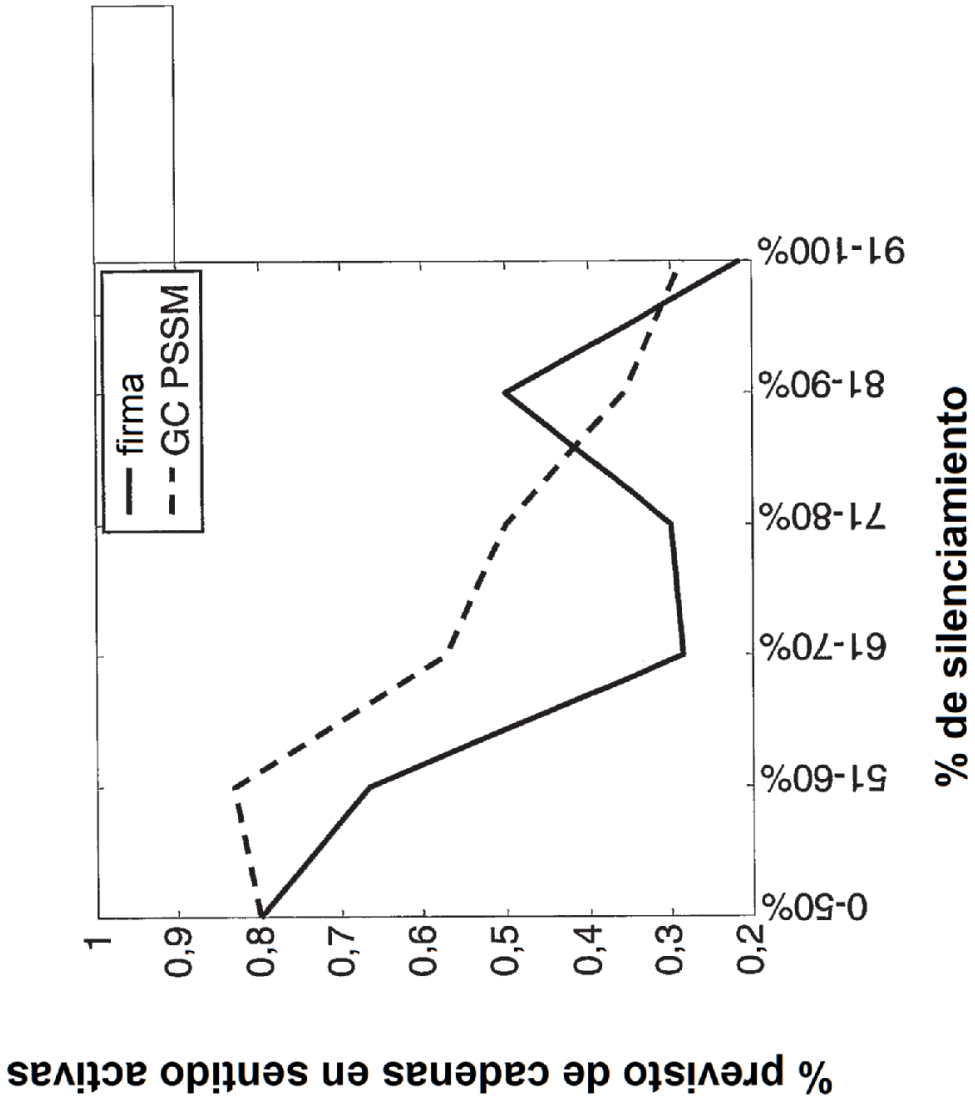

firma
GC PSSM
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
% previsto de cadenas en sentido activas
0-50%
51-60%
61-70%
71-80%
81-90%
91-100%
% de silenciamiento

FIG. 15

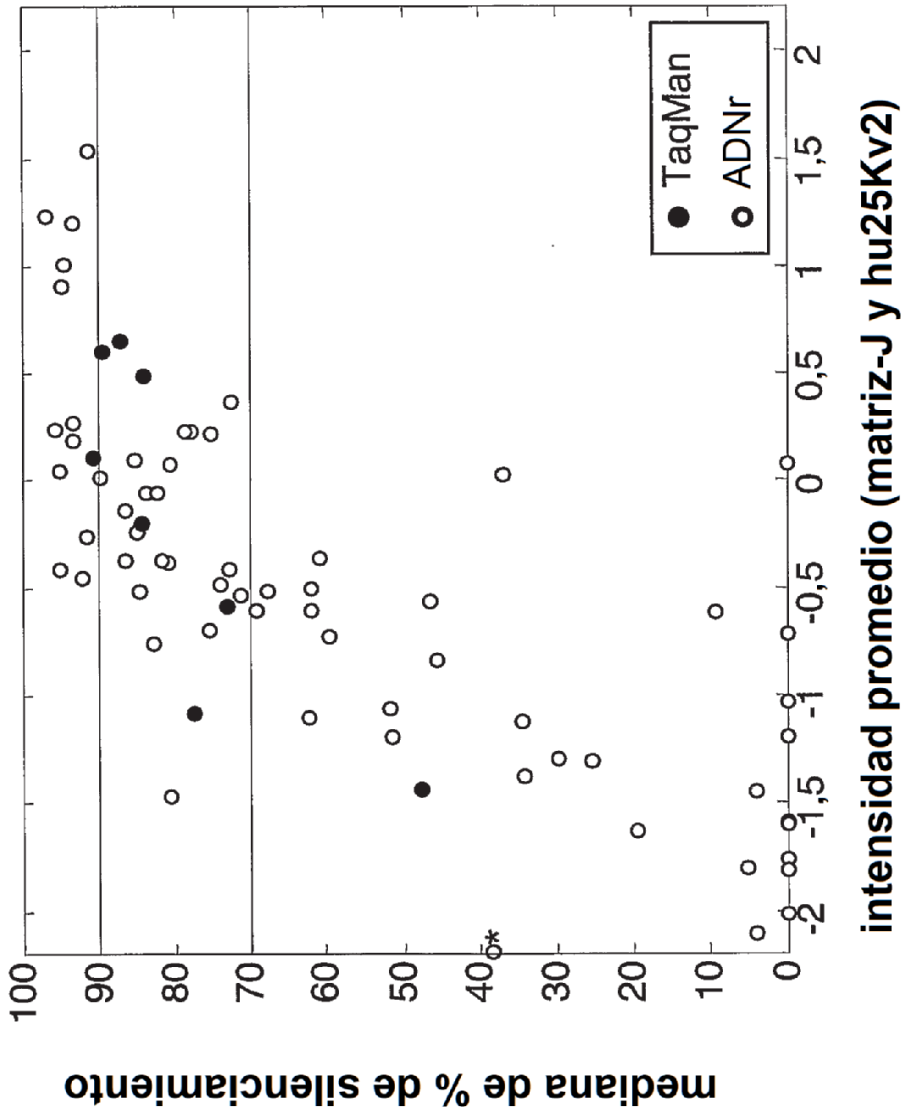

mediana de % de silenciamiento
intensidad promedio (matriz-J y hu25Kv2)
TaqMan
ADNr
100
90
80
70
60
50
40
30
20
10
0
-2
-1,5
-1
-0,5
0
0,5
1
1,5
2

FIG. 16

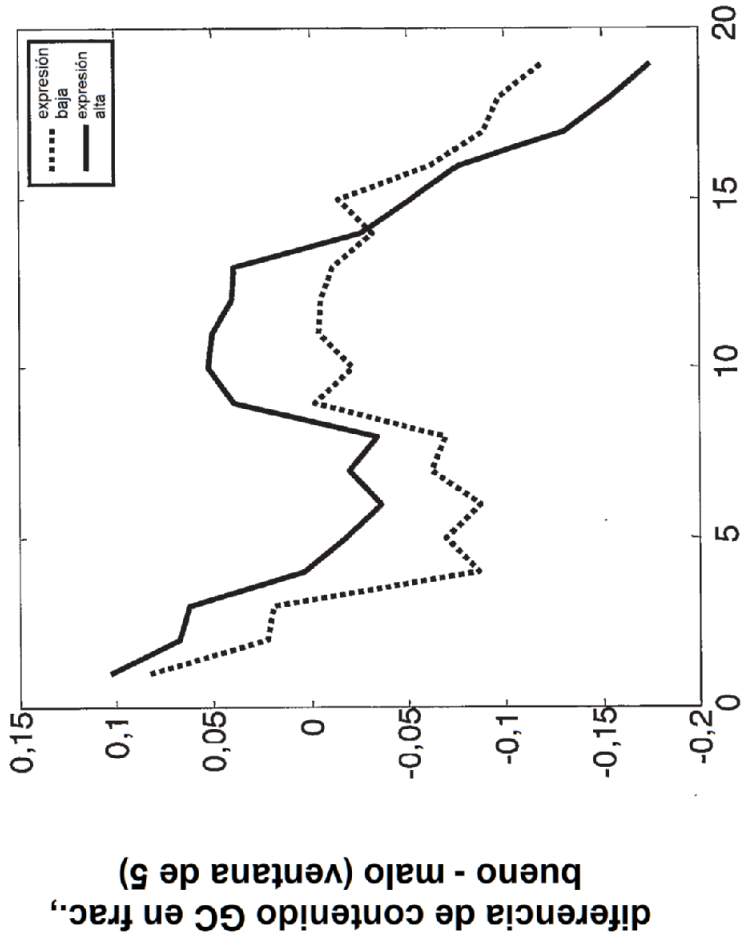

expresión baja
expresión alta
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
-0,2
0
5
10
15
20
posición
diferencia de contenido GC en frac.,
bueno - malo (ventana de 5)

FIG. 17

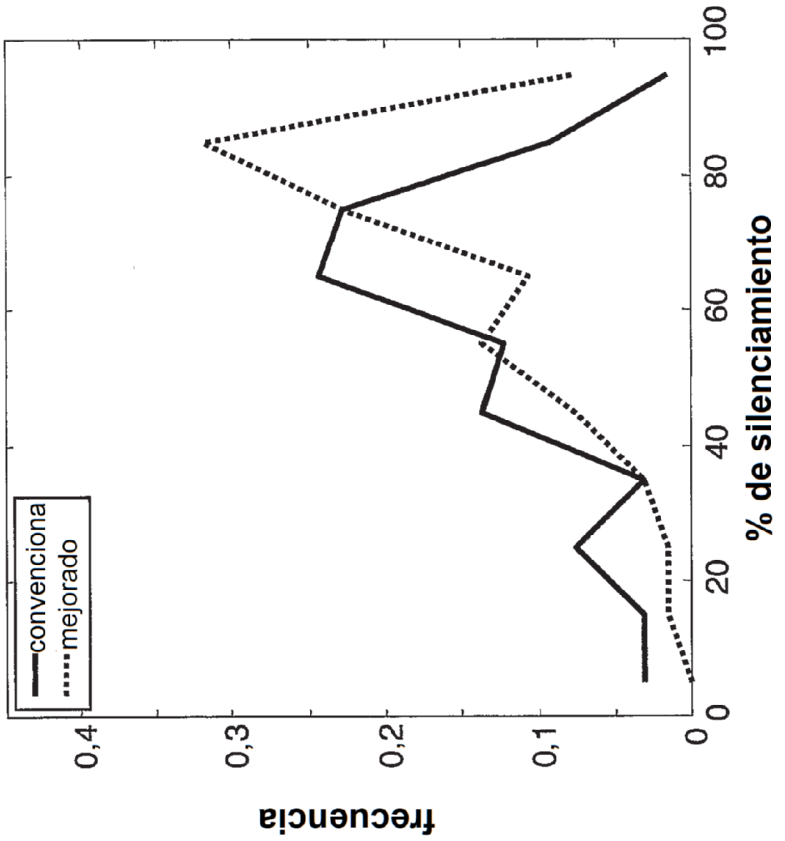

convenciona
mejorado
0,4
0,3
0,2
0,1
0
frecuencia
0
20
40
60
80
100
% de silenciamiento

FIG. 18

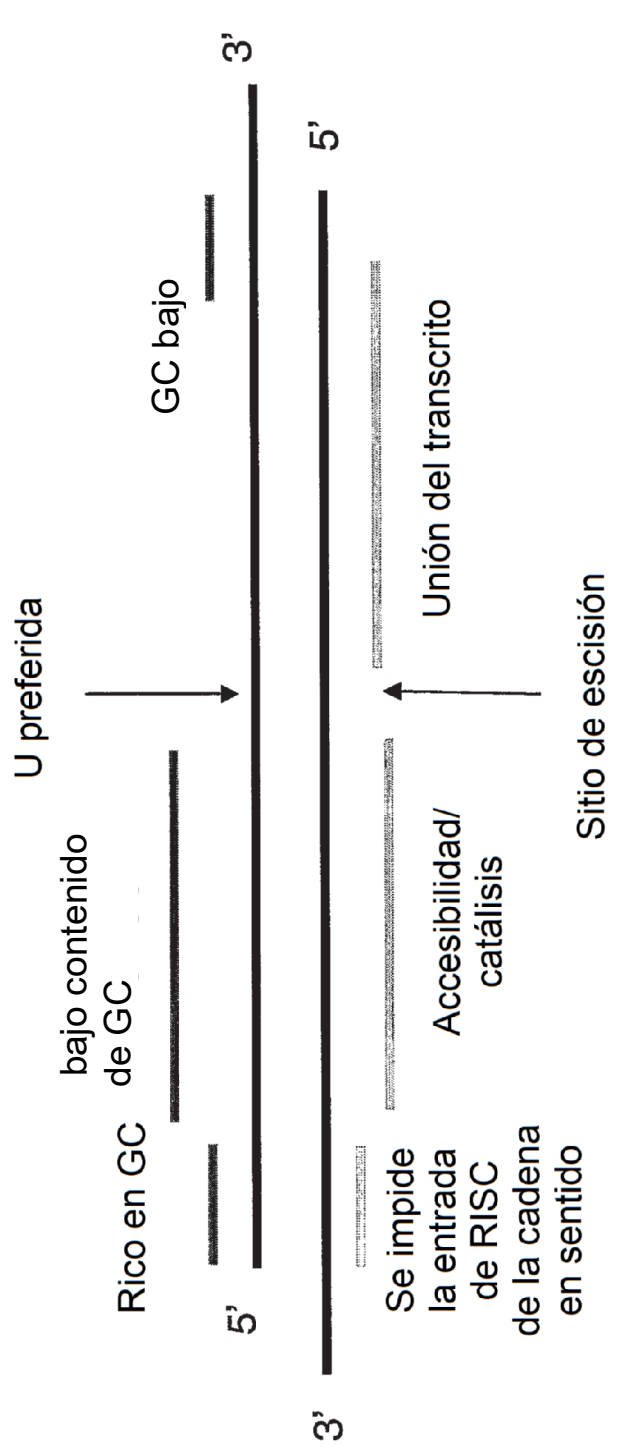

U preferida
bajo contenido de GC
Rico en GC
GC bajo
5'
3'
5'
3'
Se impide la entrada de RISC de la cadena en sentido
Accesibilidad/ catálisis
Unión del transcrito
Sitio de escisión

FIG. 19

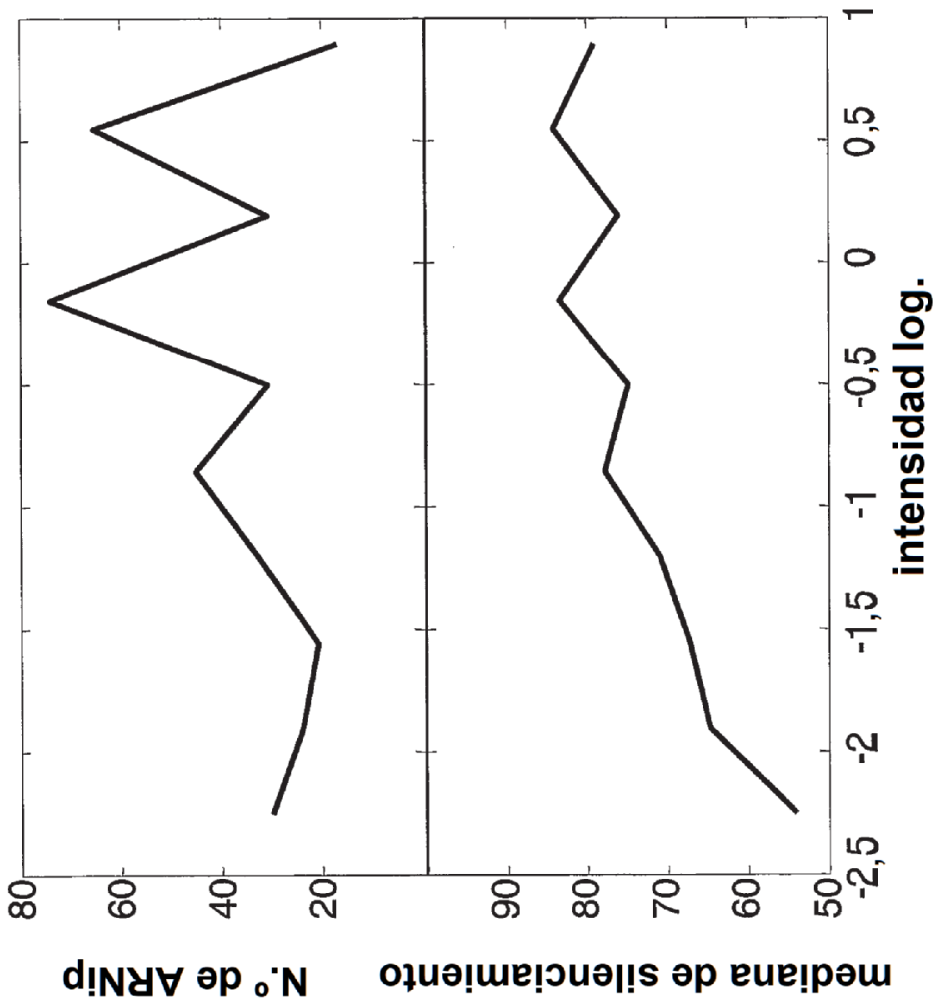

N.º de ARNip
80
60
40
20
mediana de silenciamiento
90
80
70
60
50
-2,5
-2
-1,5
-1
-0,5
0
0,5
1
intensidad log.

FIG. 20