



(19)



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 305 091**

(51) Int. Cl.:

C12N 15/10 (2006.01)

C12N 15/62 (2006.01)

C12N 15/67 (2006.01)

C12N 15/85 (2006.01)

C12N 15/90 (2006.01)

(12)

TRADUCCIÓN DE PATENTE EUROPEA

T3

(86) Número de solicitud europea: **01953251 .4**

(86) Fecha de presentación : **01.08.2001**

(87) Número de publicación de la solicitud: **1305411**

(87) Fecha de publicación de la solicitud: **02.05.2003**

(54) Título: **Método para producir dominios proteicos funcionales.**

(30) Prioridad: **01.08.2000 GB 0018876**

(45) Fecha de publicación de la mención BOPI:
01.11.2008

(45) Fecha de la publicación del folleto de la patente:
01.11.2008

(73) Titular/es: **Laboratoires Serono S.A.**
Centre Industriel
1267 Coinsins, Vaud, CH

(72) Inventor/es: **De Luca, Giampiero y**
Falciola, Luca

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

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

DESCRIPCIÓN

Método para producir dominios proteicos funcionales.

5 Esta invención se refiere a un método para la producción de proteínas, en particular de dominios proteicos funcionales.

Antecedentes de la invención

10 La funcionalidad de una proteína está estrechamente conectada con sus propiedades estructurales, que están organizadas en dos niveles diferentes: la estructura primaria (que corresponde a la secuencia de aminoácidos codificada por el gen), la estructura secundaria (localizaciones relativas preferidas en el núcleo de restos consecutivos secuencialmente), estructura terciaria (la posición relativa de todos los átomos de una cadena polipeptídica) y la estructura cuaternaria (la disposición de las diferentes subunidades de la proteína, cada una correspondiente a una cadena polipeptídica distinta, 15 en un complejo).

Además de estos niveles de organización, un polipéptido de interés puede corresponder a un dominio proteico, que puede definirse de dos maneras. Desde una perspectiva estructural, un dominio es una región de un polipéptido plegada de manera compacta que comprende una o más estructuras secundarias y/o terciarias. Las proteínas pequeñas (< 100 aminoácidos) están compuestas frecuentemente por un único dominio proteico estructural, mientras que las 20 proteínas más grandes consisten habitualmente en múltiples dominios proteicos estructurales. En la estructura tridimensional de una proteína, un dominio proteico estructural se observa como una unidad de polipéptido plegada de manera independiente que es muy distinta de otras partes de la proteína. Asociada con esta definición estructural, una segunda perspectiva de un dominio proteico es que es la parte más pequeña de una proteína que retiene una función, 25 que resulta de su interacción con una o más proteínas, ácidos nucleicos, azúcares, lípidos o cualquier otro compuesto orgánico o inorgánico. Un dominio funcional está compuesto generalmente por 50-350 aminoácidos y, por lo tanto, puede consistir en uno o más dominios estructurales.

En la naturaleza, las proteínas pueden contener uno o más dominios proteicos funcionales, unidos secuencialmente entre sí y organizados tridimensionalmente. Aunque es posible la asignación de una secuencia de aminoácidos 30 particular a una clase de dominios estructurales mediante RMN o cristalografía de rayos X, la demostración de que una secuencia de una proteína está asociada a una función se obtiene principalmente identificando las homologías de la secuencia con dominios proteicos funcionales conocidos y/o generando formas alteradas de la proteína original que se ensayan en un ensayo bioquímico o biológico relevante en estudios denominados de estructura-actividad o de 35 estructura-función.

Muchas proteínas comparten dominios proteicos funcionales y/o estructurales altamente homólogos, probablemente como consecuencia de un proceso denominado "exon shuffling" (combinación de exones de distintas unidades de transcripción). De acuerdo con esta teoría evolutiva, se ha creado un determinado número de genes mediante una 40 serie de duplicaciones, recombinaciones intrónicas, ensamblaje combinatorio y mutaciones de exones existentes que codifican "módulos" de la proteína (Patthy L, Gene (1999), 238(1):103-114).

La idea de que la evolución de las proteínas en eucariotas también está mediada por recombinación homóloga intrónica sexual de los genomas parentales, creando nuevos genes a partir de la combinación de exones asociados a 45 dominios proteicos específicos, también se acepta actualmente ya que una serie de estudios demuestran una relación entre la fase y la posición de intrones respecto a los límites estructurales entre dominios proteicos (de Souza SJ *et al.*, Proc Natl Acad Sci USA, (1996), 93(25): 14632-6). En un número significativo de acciones, es evidente que los módulos proteicos están asociados a uno o más exones limitados por intrones que tienen fase cero (el intrón no interrumpe un codón) o que tienen siempre la misma fase. De esta manera, los exones que codifican dichos dominios 50 "móviles" pueden combinarse más fácilmente en proteínas nuevas compuestas por mosaicos de dichos módulos proteicos sin ningún problema de marco (de Souza SJ *et al.*, Proc. Natl Acad Sci USA, (1998), 95(9): 5094-9; Kolkman JA y Stemmer WP, Nat Biotechnol. (2001), 19(5): 423-8).

Sin embargo, también se ha mostrado que un dominio proteico funcional puede ejercer actividades biológicas 55 claramente distintas de las ejercidas en el contexto de la proteína completa cuando se separa físicamente del resto del producto primario de la traducción, es decir, la proteína obtenida por la traducción directa del ARNm transcrito a partir de un gen. Dichos dominios proteicos funcionales se obtienen *in vivo* después de una escisión proteolítica, que puede realizarse por una endopeptidasa producida por la célula que codifica el producto primario de la traducción o por otras células, por ejemplo, cuando el producto primario de la traducción se secreta o expone en la membrana 60 celular. Dichos eventos están caracterizándose cada vez más frecuentemente y es evidente que los dominios proteicos funcionales producidos de esta manera pueden tener actividades fisiológicas importantes (Kiessling LL y Gordon EJ, Chem. Biol. (1998), 5(3): R49-R62; Halim NS, The Scientist (2000), 1 (16): 20; Blobel CP, Curr. Opin. Cell Biol. (2000), 12(5): 606-612).

65 Varias proteínas eucariotas valiosas comercialmente corresponden a estos dominios proteicos funcionales que, en un determinado número de casos, están codificados por un subconjunto de los exones codificadores que forman el gen completo.

Un ejemplo es la endostatina, un inhibidor endógeno de la angiogénesis y del crecimiento tumoral, que corresponde al fragmento proteolítico C-terminal del Colágeno XVIII 1alfa, una proteína colagenosa de la matriz extracelular. La endostatina está codificada esencialmente por los 3 exones del extremo 3' del gen del Colágeno XVIII 1alfa (COL18A1) pero sólo es completamente funcional cuando se libera mediante proteólisis del Colágeno XVIII.alfa-1, el producto primario de la traducción (O'Reilly MS *et al.*, Cell (1997), 88(2): 277-285).

Otro ejemplo es la citoquina inducida por la activación relacionada con el Factor de Necrosis Tumoral (TNF) (también denominada TRANCE, RANKL, OPGL u ODF), una proteína transmembrana de tipo II que está implicada en la ruta de señalización que activa la inducción rápida de genes que activan el desarrollo de los osteoclastos. TRANCE se forma como un producto primario de la traducción anclado a la membrana que se escinde por la metaloproteasa desintegrina TNF-alfa convertasa (TACE), generando un TRANCE soluble, una proteína completamente funcional que tiene una actividad potente en la supervivencia de las células dendríticas y osteoclastogénica. La secuencia de esta proteína soluble corresponde a la codificada por los últimos 3 exones del gen TRANCE (Lum L *et al.*, J. Biol. Chem. (1999), 274(19): 13613-8).

Para la producción a escala comercial de un dominio proteico funcional, existen dos alternativas principales, cada una de las cuales tiene sus inconvenientes.

Se podría intentar imitar a la naturaleza y producir en primer lugar el producto primario de la traducción, que se procesa posteriormente proteolíticamente para obtener la proteína deseada.

La realización de este proceso completo mediante tecnología de ADN recombinante es difícil técnicamente. No sólo debe expresarse el producto primario de la traducción sino que también tiene que identificarse la enzima proteolítica específica del dominio proteico funcional deseado, expresarse y permitir que interaccione con el producto primario de la traducción, en un modelo celular o en un sistema *in vitro*, para garantizar una escisión apropiada antes del procesamiento adicional.

Como segundo método, puede prepararse una construcción de expresión, que contiene sólo la secuencia codificadora de ADN del dominio proteico funcional aislada a partir del ARNm o del gen original. Incluso esta tecnología que se utiliza habitualmente requiere una serie de operaciones que pueden retrasar sensiblemente el desarrollo de un producto recombinante (Makrides SC, Protein Expr. Purif. (1999), 17(2): 183-202; Kaufman RJ, Mol. Biotechnol. (2001), 16(2): 151-60). Es necesario aislar la secuencia codificadora de interés a partir de la secuencia de ADNc completa, que tiene que modificarse de manera que pueda subclonarse más en un vector de expresión que contiene todos los elementos de control de la transcripción y de la traducción necesarios para la expresión correcta en la célula anfitriona. La construcción se utiliza entonces para transformar la célula anfitriona y, finalmente, se aplica un procedimiento de cribado a los transformantes para aislar los clones que expresan la proteína exógena, correctamente y a un nivel alto.

El aislamiento de los clones relevantes puede exigir mucho tiempo ya que los vectores de expresión habituales, además de los requerimientos listados previamente, necesitan recombinarse con la secuencia genómica de la célula anfitriona. Las construcciones de expresión mantenidas extracromosómicamente son inestables y sólo permiten una expresión transitoria de la proteína, frecuentemente insuficiente para permitir la producción a escala comercial.

Por lo tanto, es necesario un evento de recombinación que implique ADN exógeno y ADN genómico de la célula anfitriona codificador y no codificador para transmitir la construcción de expresión a todas las células generadas por los ciclos posteriores de replicación del ADN y división mitótica de la célula transformada originalmente. Este proceso es también completamente aleatorio y tendente a error, ya que ninguna característica específica del vector de expresión puede dirigir la incorporación completa de la secuencia exógena en el genoma de la célula anfitriona. Así, cualquier parte de él, que tenga incluso una homología muy baja con cualquier secuencia endógena, puede ser utilizado por la célula para eventos de recombinación no homólogos, que se sabe que son más frecuentes en varios órdenes de magnitud que la recombinación homóloga, resultando frecuentemente en la integración incompleta de las secuencias exógenas necesarias.

Estos problemas dan lugar al desarrollo de procedimientos de enriquecimiento genético y de selección cuyo propósito es eliminar los transformantes en los que la construcción de expresión se haya integrado de una forma incompleta. Debido a que algunas partes esenciales de la secuencia exógena pueden haberse perdido o alterado durante el proceso de recombinación, la secuencia codificadora de interés puede estar mutada, truncada o puede no expresarse. En cualquier caso, una gran mayoría de los transformantes, independientemente de la técnica utilizada para introducir el ADN y seleccionar las células, no produce la proteína esperada.

Finalmente, está bien establecido en la bibliografía que la expresión correcta de proteínas recombinantes en células eucariotas depende de muchos factores relacionados con la célula anfitriona específica. Características como la toxicidad de la proteína recombinante, el procesamiento y estabilidad del ARNm y otros eventos posteriores a la traducción están muy relacionados con el producto en sí mismo, con la secuencia codificadora y no codificadora en el vector de expresión y con la interacción de las secuencias exógenas con la base genómica de las células anfitrionas. De hecho, la inserción aleatoria de un gen recombinante completo, que puede contener muchas Kilobases de ADN, puede perturbar seriamente el genoma de la célula anfitriona, comprometiendo su estabilidad y viabilidad. Por lo tanto, incluso si las secuencias exógenas se han integrado completamente, algunos clones no pueden utilizarse para producir la proteína si dichas secuencias interrumpen secuencias genómicas importantes para el metabolismo y/o replicación de la célula.

Dichos clones pueden perderse debido a la presión selectiva en el cultivo celular o replicarse tan lentamente que es difícil obtener células suficientes que expresen de manera estable la proteína deseada con una eficacia satisfactoria.

Se han desarrollado estrategias alternativas con el fin de minimizar los inconvenientes de la integración de genes exógenos esencialmente incontrolada. Mayoritariamente, se basan en recombinación homóloga, una tecnología única que permite la inserción de una secuencia exógena específica en una secuencia genómica predeterminada de una célula de mamífero. Esta técnica se ha utilizado fundamentalmente para la caracterización de genes o secuencias reguladoras en modelos animales y celulares modificándolas, como se ha revisado recientemente (Muller U, Mech. Dev. (1999), 82 (1-2): 3-21; Sedivy JM y Dutriaux A, Trends Genet. (1999), 15(3): 88-90), generando genes interrumpidos, no funcionales o quiméricos. Se han introducido varios vectores y genes marcadores seleccionables, por ejemplo, en el genoma de células Madre Embrionarias (ES) de ratón, con el fin de estudiar los efectos de las alteraciones genéticas en varias características fenotípicas, como la regulación hormonal, fertilidad, respuesta inmunológica y desarrollo de órganos.

La viabilidad de las técnicas basadas en la recombinación homóloga para la producción de proteínas recombinantes se ha demostrado a nivel de productos primarios de la traducción completos (WO 91/09955, WO 95/31560), que se han obtenido poniendo un gen endógeno completo bajo el control de secuencias reguladoras de la transcripción exógenas. Alternativamente, pueden obtenerse proteínas truncadas mediante oligonucleótidos antisentido que, una vez transfectados en una célula, pueden emparejarse con el ARNm endógeno impidiendo que el ribosoma vaya a lo largo de la secuencia codificadora completa (WO 97/23244). Sin embargo, ninguno de los documentos de la bibliografía sugiere la utilización de la recombinación homóloga para expresar selectivamente uno o más exones de un gen endógeno que codifica un dominio proteico funcional comprendido en un producto primario de la traducción mediante la integración de secuencias reguladoras exógenas.

Compendio de la invención

Ahora se ha encontrado que, cuando el dominio proteico funcional corresponde al extremo N o C-terminal de un producto primario de la traducción, por ejemplo, codificado por al menos el exón o exones más próximos al extremo 5' ó 3', respectivamente, de un gen que contiene intrones, existe un método adicional mediante el cual puede producirse el dominio proteico funcional.

La presente invención se define por las reivindicaciones.

Así, la invención proporciona un método para producir una proteína, caracterizado porque dicha proteína es un fragmento de una proteína que es un producto primario de la traducción de un gen diana que es el extremo C o N terminal de dicho producto primario de la traducción de un gen diana y en el que la proteína tiene una actividad biológica, comprendiendo el método:

(i) crecer una célula anfitriona transfectada con una construcción de ADN que comprende:

- (a) una secuencia de ADN reguladora capaz de iniciar la transcripción y la traducción o de terminar la transcripción y la traducción del ADN que codifica dicho fragmento; y
- (b) una región de ADN de direccionamiento que comprende secuencias homólogas a una región genómica en 5' ó 3', respectivamente, respecto a la secuencia que codifica dicho fragmento, integrándose la construcción en el ADN genómico de la célula anfitriona en una posición determinada por el segmento de ADN de direccionamiento de manera que la expresión de dicho fragmento está bajo el control del ADN regulador; y, opcionalmente

(ii) cultivar la célula recombinante de manera homóloga; y, opcionalmente

(iii) recoger dicho fragmento

en el que dicho fragmento se selecciona del grupo que consiste en

(i) un dominio de unión extracelular de un receptor de membrana;

(ii) un fragmento proteolítico de una proteína estructural de la matriz extracelular que tiene propiedades antiangiogénicas; y

(iii) una proteína liberada proteolíticamente a partir de una proteína precursora inactiva.

La invención es útil para producir proteínas que forman dominios proteicos funcionales en las células en las que uno o más de los exones codificadores carecen, en un extremo, de la secuencia de ADN reguladora necesaria para su expresión y traducción directa como entidades moleculares distintas, y no como segmentos, de productos primarios de la traducción. En el término proteínas están incluidos los péptidos cortos y los oligopéptidos (que contienen, por ejemplo, hasta treinta restos de aminoácidos) así como los péptidos más largos (que contienen más de treinta restos de aminoácidos). Los péptidos y polipéptidos que pueden producirse mediante la invención incluyen los que se producen por la maduración proteolítica de un producto primario de la traducción.

En una realización preferida del método de la invención, el exón o exones codifican dicho fragmento que corresponde a un dominio proteico funcional cuya secuencia corresponde a la secuencia C o N-terminal de un producto primario de la traducción de un gen y dicho fragmento codificado por dicho exón o exones tiene una actividad biológica, comprendiendo el método crecer una célula anfitriona transfectada con una construcción de ADN que comprende:

(a) una secuencia de ADN reguladora capaz de iniciar la transcripción y la traducción o de terminar la transcripción y la traducción del exón o exones que codifican dicho fragmento correspondiente al dominio proteico funcional; y

(b) una región de ADN de direccionamiento que comprende secuencias homólogas a una región genómica en 5' ó 3', respectivamente, respecto al exón o exones que codifican dicho fragmento, integrándose la construcción en el ADN genómico de la célula anfitriona en una posición determinada por el segmento de ADN de direccionamiento de manera que la expresión del exón o exones está bajo el control del ADN regulador.

La célula anfitriona modificada con el método de la invención expresa selectivamente las secuencias exónicas presentes en el genoma que codifican un dominio proteico funcional como una molécula de ARNm nueva que se traduce directamente en el dominio proteico funcional, sin ninguna modificación proteolítica específica de tipo celular o tisular.

En las realizaciones preferidas, el péptido o polipéptido que forma el dominio proteico funcional está codificado por al menos el exón más próximo al extremo 5' ó 3' de un gen que contiene intrones. Así, en una realización preferida se proporciona un método para producir una proteína que es un dominio proteico funcional que corresponde al extremo C-terminal de un producto primario de la traducción de un gen diana y que está codificado por al menos el exón más próximo al extremo 3' de un gen diana que contiene intrones, comprendiendo el método:

(i) crecer una célula anfitriona transfectada con una construcción de ADN que está unida operativamente a dicho exón o exones después de la integración en el genoma de la célula anfitriona mediante recombinación homóloga, comprendiendo dicha construcción:

(a) una secuencia de ADN de direccionamiento que comprende secuencias homólogas a una región genómica en 5' respecto al exón o exones que codifican la proteína;

(b) un módulo de transcripción que consiste en una secuencia de ADN capaz de activar la transcripción de un ADN que codifica el dominio proteico funcional;

(c) un módulo de traducción que consiste en una secuencia de ADN capaz de iniciar la traducción del dominio proteico funcional; y opcionalmente

(d) un módulo de corte y empalme que comprende un sitio donador de corte y empalme sin emparejar en 5' que, cuando se complementa con el sitio aceptor de corte y empalme sin emparejar en 3' del exón endógeno que codifica el extremo N-terminal del dominio proteico funcional, permite el corte y empalme del transcrito primario, lo que resulta en la yuxtaposición "en marco" del módulo de traducción con la secuencia que codifica el dominio proteico funcional;

(ii) cultivar la célula recombinante de manera homóloga; y

(iii) recoger el dominio proteico funcional.

Adecuadamente, en esta realización de la invención, la construcción debería proporcionar secuencias exógenas que forman una unidad reguladora que permita el inicio correcto de la transcripción y la traducción del exón o exones que codifican el dominio proteico funcional. La región de ADN de direccionamiento estará formada por secuencias que pertenecen a la región genómica en 5' respecto a la secuencia que codifica el dominio proteico funcional (tal como el intrón adyacente del gen diana en su extremo 5'), con o sin secuencias que pertenecen al exón o exones que codifican el dominio proteico funcional. Si es necesario incrementar la longitud de la región de homología, también pueden incluirse secuencias de ADN que pertenecen a secuencias codificadoras y/o no codificadoras contiguas al gen diana (u opcionalmente de genes contiguos).

El codón de Metionina (ATG) necesario para empezar la traducción tiene que estar en un contexto de secuencia Kozak bueno y puede proporcionarse por el intrón en 5' respecto al exón o exones que codifican el dominio proteico funcional, por el exón que contiene el extremo N-terminal del dominio proteico funcional o estar unido operativamente a secuencias codificadoras endógenas mediante la adición de uno o más exones naturales o sintéticos, insertados entre el exón o exones que codifican el dominio proteico funcional y las secuencias exógenas comprendidas en la unidad reguladora con el fin de iniciar la transcripción. Si el codón de inicio de la traducción es endógeno, el módulo transcripcional comprenderá una región 5' no traducida apropiada.

El módulo de corte y empalme, que contiene un sitio donador de corte y empalme sin emparejar, está asociado al exón natural o sintético adyacente al exón que codifica el extremo N-terminal del dominio proteico funcional. El módulo de corte y empalme debe permitir que el codón de Metionina, una vez que se ha puesto en el mismo marco que el dominio proteico funcional después del proceso de corte y empalme, sea el resto N-terminal del dominio proteico

funcional. En esta realización, el gen diana que codifica el producto primario de la traducción puede o no expresarse ya a niveles significativos, pero la secuencia reguladora exógena debe permitir tener un sitio nuevo de inicio de la transcripción y de la traducción en 5' respecto al exón o exones que codifican el dominio proteico funcional.

Algunos ejemplos de cómo llevar a cabo esta realización de la invención se muestran en la Figura 1 A-F.

En otra realización preferida de la invención se proporciona un método para producir una proteína que es un dominio proteico funcional que corresponde al extremo N-terminal de un producto primario de la traducción de un gen diana y que está codificada por al menos el exón más próximo al extremo 5' de un gen diana que contiene intrones, comprendiendo el método:

(i) crecer una célula anfitriona transfectada con una construcción de ADN, que está unida operativamente a dicho exón o exones después de la integración en el genoma de la célula anfitriona por recombinación homóloga, comprendiendo dicha construcción

(a) una región de ADN de direccionamiento que comprende secuencias homólogas a una región genómica en 3' respecto al exón o exones que codifican la proteína;

(b) un módulo de transcripción que consiste en una secuencia de ADN capaz de terminar la transcripción del ADN genómico;

(c) un módulo de traducción que consiste en una secuencia de ADN capaz de terminar la traducción del dominio proteico funcional; y opcionalmente

(d) un módulo de corte y empalme que comprende un sitio aceptor de corte y empalme sin emparejar en 3' que, cuando se complementa con el sitio donador de corte y empalme sin emparejar en 5' del exón endógeno que codifica el extremo C-terminal del dominio proteico funcional, permite el corte y empalme correcto del transcrito primario, lo que resulta en la yuxtaposición "en marco" del módulo de traducción con la secuencia que codifica el dominio proteico funcional;

(ii) cultivar la célula recombinante de manera homóloga; y

(iii) recoger el dominio proteico funcional.

Adecuadamente, en esta realización de la invención, la construcción debe proporcionar secuencias exógenas que permitan la terminación correcta de la transcripción y de la traducción para el exón o exones que codifican la proteína que forma el dominio proteico funcional. La región de ADN de direccionamiento estará formada por secuencias que pertenecen a la región genómica en 3' respecto a la secuencia que codifica el dominio proteico funcional (tal como el intrón adyacente del gen diana en su extremo 3'), con o sin secuencias que pertenecen al exón que codifica el extremo C-terminal de la proteína o el intrón en su extremo 3', o ambos. Si es necesario incrementar la longitud de la región de homología, también pueden incluirse secuencias de ADN que pertenecen a secuencias codificadoras y/o no codificadoras contiguas al gen diana (u opcionalmente de genes contiguos).

El codón necesario para terminar la traducción podría proporcionarse por el intrón en 3' respecto al exón o exones que codifican el dominio proteico funcional o estar unidos operativamente a secuencias codificadoras endógenas mediante la adición de un exón o exones naturales o sintéticos, insertados entre el exón o exones que codifican el dominio proteico funcional y las secuencias exógenas introducidas por recombinación homóloga con el fin de terminar la transcripción. Si el codón de parada de la traducción es endógeno, el módulo transcripcional comprenderá una región 3' no traducida apropiada.

El módulo de corte y empalme, que contiene un sitio aceptor de corte y empalme sin emparejar, está asociado al exón natural o sintético próximo al exón que codifica el extremo C-terminal del dominio proteico funcional. El módulo de corte y empalme debe permitir que el codón de parada, una vez que se ha puesto en el mismo marco que el dominio proteico funcional después del proceso de corte y empalme, sea el codón de terminación del dominio proteico funcional. En esta realización, el gen diana que codifica el producto primario de la traducción tiene que expresarse ya a niveles significativos y la secuencia exógena es necesaria sólo para parar la transcripción y la traducción en una posición diferente.

Algunos ejemplos de cómo llevar a cabo esta realización se muestran en la fig. 1 G-L.

En un aspecto adicional, la invención proporciona construcciones que permiten a la célula anfitriona, una vez integradas correctamente en el genoma por recombinación homóloga, expresar un ARNm nuevo que incluye, entre todos los exones contenidos en el gen diana que codifican el producto primario de la traducción, sólo el o los que codifican un dominio proteico funcional.

La elección de la estrategia para la recombinación homóloga influye en la estructura final del gen diana, ya que la unidad reguladora puede insertarse en el intrón adyacente o reemplazar parte de este intrón y toda o parte de la región genómica que pertenece al mismo gen pero que no codifica el dominio proteico funcional.

La unidad reguladora, integrada por recombinación homóloga en el gen que comprende el exón o exones que codifican un dominio proteico funcional, contiene secuencias exógenas que incluyen un módulo de transcripción, un módulo de traducción y, opcionalmente, un módulo de corte y empalme. Se pretende que dichas secuencias exógenas proporcionen las secuencias necesarias para combinarse con las secuencias endógenas que engloban el dominio proteico funcional. Así, debería estar presente en el genoma de la célula anfitriona un nuevo gen recombinante, que contiene elementos endógenos de control de la transcripción y de la traducción en un extremo, exón o exones e intrón o intrones endógenos asociados al dominio proteico funcional en el centro y elementos exógenos de control de la transcripción y de la traducción en el otro extremo.

Los métodos de la invención están basados en la integración, por un único evento de recombinación homóloga, de una unidad reguladora en el gen eucariota que codifica el producto primario de la traducción al nivel de la región intrónica genómica inmediatamente adyacente a las secuencias exónicas que codifican el dominio proteico funcional. Dependiendo de la posición de las secuencias exónicas relevantes en el gen diana, dicho intrón es el que está inmediatamente en 5' ó 3' respecto a las secuencias exónicas cuando el dominio proteico funcional está, respectivamente, en el extremo C o N-terminal del producto primario de la traducción.

La presente invención proporciona algunas ventajas importantes respecto a la técnica anterior. La producción del dominio proteico funcional se consigue mediante los métodos conocidos en la técnica aislando la secuencia codificadora bien del dominio proteico funcional y expresándola bajo el control de una secuencia reguladora *ada* en sus extremos o del producto primario de la traducción y de la proteasa específica que genera el dominio proteico funcional, y expresando ambos.

La invención proporciona un método para producir dominios proteicos funcionales siempre que sean fragmentos C o N-terminales de un producto primario de la traducción y que se conozca la estructura exón/intrón del ADN genómico correspondiente. Este método implica la integración de secuencias reguladoras exógenas en 5' (si el dominio proteico funcional corresponde a un fragmento C-terminal) o 3' (si el dominio proteico funcional corresponde a un fragmento N-terminal) respecto al exón o exones que codifican dichos dominios proteicos en el ADN genómico. Los métodos de la invención permiten generar un gen recombinante que puede transcribirse y traducirse directamente por la célula en un dominio proteico funcional, a diferencia del gen que codifica el producto primario de la traducción.

Los métodos de la invención basados en las propiedades de la recombinación homóloga permiten una modificación controlada y precisa del genoma de la célula anfitriona con el fin de producir un dominio proteico funcional. La cantidad de la secuencia exógena que debe integrarse en el genoma de la célula anfitriona es muy limitada ya que se utiliza, como secuencia codificadora, la secuencia codificadora original presente en el genoma de la célula anfitriona. Además, sólo se integran los elementos adicionales necesarios tales como elementos de control de la transcripción y/o de la traducción.

La utilización de la secuencia de la célula anfitriona que codifica el dominio proteico funcional también proporciona las ventajas de eliminar cualquier alteración derivada de la recombinación de dicha secuencia codificadora y también utilizar los mismos procesos post-transcripcionales (por ejemplo, corte y empalme) y/o posteriores a la traducción (por ejemplo, glicosilación, fosforilación) que se aplican realmente *in vivo* para la maduración del dominio proteico funcional. La utilización de una única unidad reguladora elimina la necesidad de manipular el ADN complementario que codifica el producto primario de la traducción para aislar el segmento que codifica el dominio proteico funcional y adaptarlo al vector de expresión.

Finalmente, se ha demostrado que las construcciones de expresión genómicas (es decir, que contienen uno o más intrones sintéticos y/o naturales) se expresan más eficientemente que las construcciones idénticas que carecen de intrones (es decir, las que se obtienen generalmente utilizando las técnicas descritas en la bibliografía), debido probablemente al proceso de corte y empalme. Los métodos de la invención evitan la introducción de secuencias intrónicas exógenas mediante la utilización de intrones que interrumpen naturalmente la secuencia que codifica el dominio proteico funcional.

Descripción detallada de la invención

En los párrafos siguientes, se describirán apropiadamente los elementos básicos de la invención por referencia a la bibliografía sobre técnicas de recombinación homóloga (Muller U, *Mech. Dev.* (1999), 82(1-2): 3-21; Hasty P *et al.*, en "Gene targeting: a practical approach", ed. Joyner AL, pub. Oxford Univ. Press, 1999, 1-35) y técnicas de expresión de proteínas (Makrides SC., *Protein Expr. Purif.* (1999), 17(2): 183-202; Kaufman RJ, *Mol Biotechnol.* (2001), 16(2): 151-60).

Los dominios proteicos funcionales

La expresión "*dominio proteico funcional*" (FPD) significa un fragmento de una proteína que es el producto primario de la traducción de un gen, que ejerce una función biológica. El dominio proteico funcional, que puede estar compuesto por uno o más dominios proteicos estructurales (idénticos o diferentes entre sí), debería contener todas las características biofísicas necesarias para plegarse adecuadamente como una unidad plegada independiente y ejercer la actividad biológica esperada.

En el contexto del ADN genómico de una célula, un dominio proteico funcional está codificado por una parte de la región codificadora completa de un gen, mientras que el producto primario de la traducción está codificado por la secuencia codificadora completa de un gen. La secuencia genómica que codifica un dominio proteico funcional carece, en uno o los dos extremos, de secuencias reguladoras situadas apropiadamente, que son reconocidas por la maquinaria de expresión de la célula para generar un producto primario de la transcripción y de la traducción. Consecuentemente, un dominio proteico funcional corresponde a una parte de la región traducida comprendida en el ARNm transcrito por una célula y no puede ser el resultado de un evento alternativo de corte y empalme, que necesitaría una secuencia reguladora activa en ambos extremos.

Así, un gen que tiene un marco de lectura completo y funcional limitado por las secuencias reguladoras de la transcripción y de la traducción en sus extremos o un ARNm que codifica un dominio proteico funcional no existe, ya que no hay un ADN genómico que pueda ser transcrito y traducido directamente en una proteína correspondiente a un dominio proteico funcional. Una célula puede producir un dominio proteico funcional, a partir de un gen comprendido en su propio genoma y en el que se incluye la secuencia codificadora para dicha proteína, sólo después de la transcripción y la traducción del gen en un producto primario de la traducción, que se modifica posteriormente mediante un evento proteolítico específico.

En el sentido de la presente invención, los eventos proteolíticos que originan un dominio proteico funcional no son los que simplemente determinan la localización de una proteína, como cuando el péptido señal es reconocido y eliminado en proteínas extracelulares o transmembrana. Los eventos proteolíticos que originan un dominio proteico funcional tienen que ser, sin embargo, los que permitan a un dominio proteico funcional, mediante su separación del producto primario de la traducción, realizar una o más actividades biológicas proporcionando efectos fisiológicos distintos de los proporcionados por el producto primario de la traducción. Las actividades proteolíticas similares no son expresadas constitutivamente por cualquier célula, como en el caso de la enzima que elimina el péptido señal, sino que son expresadas específicamente sólo por determinados tipos celulares o sólo en determinadas condiciones metabólicas.

Dado que la función de una secuencia proteica resulta de su interacción con otras proteínas, ácidos nucleicos, lípidos, azúcares o cualquier otro ligando orgánico o inorgánico, un dominio proteico funcional, en el sentido de la presente invención, se define relacionando propiedades características con las propiedades de interacción, identificando tres grupos básicos.

Un primer grupo de dominios proteicos funcionales es en el que los efectos característicos se deben al hecho de que el dominio proteico funcional, aislado o en el contexto del producto primario de la traducción, mantiene las mismas propiedades de interacción, en términos de especificidad y afinidad por el ligando, pero que, cuando se separa del resto del producto primario de la traducción, no permite a la célula u organismo identificar la presencia de dicho ligando. Se proporcionan ejemplos por los dominios de unión extracelulares de receptores de membrana que, cuando se separan proteolíticamente de las partes transmembrana e intracelulares del receptor mediante proteasas extracelulares, sustraen al ligando de los receptores de membrana y bloquean la activación posterior de la vía de señalización intracelular.

Los dominios extracelulares separados son dominios proteicos funcionales en el sentido de la presente invención, ya que tienen el efecto de atrapar al ligando y silenciar la vía de señalización impidiendo la respuesta celular específica, en lugar de señalar a la célula la presencia del ligando como cuando están unidos al resto del producto primario de la traducción. Dichos dominios proteicos funcionales, denominados frecuentemente receptores señuelo, juegan un papel fisiológico importante, ya que proporcionan al organismo la posibilidad de regular de manera fina, por ejemplo, los efectos de las quimioquinas y citoquinas circulantes (Mantovani A *et al.*, Trends Immunol. (2001), 22(6): 328-36).

Un segundo grupo de dominios proteicos funcionales es en el que los efectos característicos se deben al hecho de que el dominio proteico funcional, cuando se aísla del contexto del producto primario de la traducción, reconoce con una alta afinidad un ligando no unido o poco unido por el producto primario de la traducción, determinando un efecto fisiológico inesperado. Los ejemplos de dichos dominios proteicos funcionales que tienen efectos fisiológicos importantes, tales como los fragmentos proteolíticos de proteínas estructurales de la matriz extracelular que tienen fuertes propiedades antiangiogénicas (Cao Y, Int. J. Biochem. Cell. Biol. (2001), 33(4): 357-69) se describen cada vez más frecuentemente en la bibliografía, mostrando que las proteínas también pueden funcionar como reservorios de dominios proteicos funcionales que son crípticos en el producto primario de la traducción hasta que un evento proteolítico, en el marco de un mecanismo fisiológico específico, les permite ser una forma aislada y funcional.

Un tercer grupo de dominios proteicos funcionales está representado por las proteínas liberadas proteolíticamente a partir de una proteína precursora inactiva. Muchas proteínas de señalización y secretadas ejercen su función fisiológica sólo después de haber sido separadas del producto primario de la traducción, como las citoquinas proinflamatorias (Dinarello CA, Chest (2000), 118 (2): 503-508).

Los dominios proteicos funcionales pueden identificarse utilizando diferentes procedimientos. Tradicionalmente, los dominios proteicos funcionales se identifican exponiendo una proteína a una serie de proteasas que tienen especificidad y ensayando la actividad de los fragmentos resultantes (Carrey E., en "Protein Structure: a practical approach", ed. Creighton T., Oxford Univ. Press (1989), 117-144). Este procedimiento ofrece las ventajas de rapidez, sencillez y sensibilidad pero está limitado por la cantidad de proteína nativa completa a digerir, por la especificidad de las proteasas y por la sensibilidad del ensayo utilizado para validar el fragmento proteolítico como un dominio proteico funcional.

Además, los avances realizados en el campo del aislamiento y secuenciación de proteínas (por ejemplo, espectroscopía de masa, geles bidimensionales) junto con los avances en el campo de la bioinformática, permiten el análisis paralelo de muchas muestras de proteínas, obteniendo información sobre la identidad y cantidad incluso de las especies moleculares menos representadas (Lottspeich F., *Angew. Chem. Int. Ed. Engl.* (1999); 38(17): 2476-2492).

- 5 Por ejemplo, ahora es posible identificar y aislar la mayoría, sino todas, de las proteínas presentes en una mezcla de proteínas, sin fragmentar u obtenidas después de una digestión controlada con una enzima proteolítica específica del contenido proteico completo en una muestra biológica, incluso sin un fraccionamiento preliminar (Spahr CS *et al.*, *Proteomics* (2001), 1(1): 93-107). Combinando estas tecnologías de digestión, detección, aislamiento, comparación de secuencias y análisis de proteínas con un ensayo apropiado basado en biología celular o bioquímica de alto rendimiento (Kuhlmann J, *Int. J. Clin. Pharm. and Ther.*, (1997), 35 (12): 541-552), se pueden identificar nuevos dominios

15 Se ha encontrado que actividades fisiológicas importantes están sometidas a una regulación fina por la actividad proteolítica ejercida por proteasas tales como las metaloproteinasas de la matriz (Raza SL y Cornelius LA, *J. Invest. Dermatol. Symp. Proc.* (2000), 5(1): 47-54) y caspasas (Los M *et al.*, *Trends Immunol.* (2001), 22(1): 31-4), que va más allá de la simple degradación de proteínas. Además, como las proteasas deberían suponer el 1,5-2% del contenido génico total en los seres humanos (Southan C, *J Pept Sci.* (2000), 6(9): 453-8) y los restos de los sitios de escisión de las proteasas asociados a proteínas específicas pueden identificarse cada vez más fácilmente (Turk BE *et al.*, *Nat. Biotech.* (2001), 19(7): 661-7), la variedad de proteasas y de proteínas que puede ser ensayada incrementará y consecuentemente la posibilidad de identificar dominios proteicos funcionales con un interés terapéutico y comercial.

- 20 Una vez que se ha identificado que una proteína funcional es el extremo N o C-terminal de un producto primario de la traducción y que se ha encontrado una correlación entre la organización del dominio proteico y la organización exón/intrón del gen correspondiente, pueden utilizarse construcciones que contienen la unidad reguladora y las regiones de direccionamiento apropiadas para generar células que expresan dicha especie proteica según los métodos de la invención.

- 30 La elección del exón o exones que se van a expresar como un dominio proteico funcional puede verse influida no sólo por las evidencias proporcionadas por el análisis de las muestras de las proteínas fragmentadas o no fragmentadas sino también por los resultados obtenidos combinando la información de la estructura tridimensional calculada de la proteína, la homología con otros dominios proteicos funcionales conocidos, estudios de mutagénesis y de estructura-función u otra simulación de modelos e informática.

- 35 Incluso si no se ha caracterizado inicialmente *in vivo* o *in vitro* que el dominio proteico funcional se corresponde exactamente con un número discreto de exones, la estructura y/o las homologías de la secuencia de la proteína pueden ser tales que puede identificarse que los elementos esenciales del dominio proteico funcional están codificados por un exón o exones específicos N o C-terminales. No es necesario que el dominio proteico funcional producido por los métodos de la invención tenga restos N o C-terminales idénticos a los identificados inicialmente *in vivo* o *in vitro* pero

- 40 debe ejercer una actividad comparable.
- La validación y caracterización adicionales de un fragmento proteolítico como un dominio proteico funcional pueden obtenerse mediante otro método de cribado funcional que implica técnicas de biología molecular como PCR aleatoria/mutagénesis por delección, mapeo de los sitios de escisión proteolítica, exposición en fagos, sistema de dos híbridos (WO 96/31625; WO 90/04788; Parry S *et al.*, *Biochem Biophys Res Commun.* (2001), 283(3): 715-20; Kawasaki M, e Inagaki F, *Biochem Biophys Res Commun.* (2001), 280(3): 842-4) o que implica métodos basados en no homología (Marcotte EM, *Curr Opin Struct Biol.* (2000), 10(3): 359-65; WO 00/11206). Antes de aplicar los métodos de la invención para producir un dominio funcional a gran escala, pueden producirse fragmentos de diferentes longitudes que pertenecen a un producto primario de la traducción en pequeña escala mediante las tecnologías habituales de expresión y ensayarlos en un ensayo para circunscribir la secuencia de proteína mínima o suficiente correspondiente a un dominio proteico funcional.

- 55 Finalmente, puede utilizarse un número creciente de algoritmos y programas informáticos disponibles en internet (Teichmann SA *et al.*, *Curr. Opin. Struct. Biol.* (2001), 11(3): 354-63; Skolnick J y Fetrow JS, *Trends Biotechnol.* (2000); 18(1) 34), separadamente o en combinación, para comparar una secuencia de proteínas o de ADN de función desconocida (como las obtenidas traduciendo EST o el resultado de programas de secuenciación genómica) con bases de datos de secuencias de proteínas o de ADN con una estructura y/o actividad biológica conocida. Este procedimiento *in silico* permite una buena aproximación de la posición y función de un dominio proteico englobado en una proteína de función desconocida o en un gen que todavía no se ha clonado. Por lo tanto, dichas herramientas bioinformáticas

- 60 pueden ayudar en la identificación de dominios proteicos funcionales que pueden expresarse mediante los métodos de la invención.
- Entre los dominios proteicos funcionales valiosos comercialmente, los que tienen un interés terapéutico representan un grupo preferido. La mayoría de las proteínas útiles terapéuticamente pueden agruparse en 3 clases: factores reguladores (incluyendo hormonas, citoquinas, linfoquinas, quimioquinas, receptores y otros factores reguladores del crecimiento y metabolismo celulares), productos sanguíneos (incluyendo factores sanguíneos derivados del suero y activadores enzimáticos del fibrinógeno) y anticuerpos monoclonales. Los productos primarios de la traducción que contienen dominios proteicos funcionales pueden estar codificados por genes que pertenecen a estas 3 clases, que pue-

den utilizarse como genes diana mediante los métodos de la invención para producir dominios proteicos funcionales utilizando exones endógenos.

Sin embargo, la bibliografía científica, de la que se han seleccionado las referencias de base y los ejemplos de la presente invención, muestra que los dominios proteicos funcionales que tienen un interés terapéutico pueden identificarse en los productos primarios de la traducción inicialmente no caracterizados como pertenecientes a estos grupos (tales como proteínas unidas a membrana, proteínas transmembrana, enzimas, proteínas de la matriz extracelular, proteínas de señalización intracelular y estructurales, proteínas nucleares) pero que también contienen dominios proteicos funcionales de interés terapéutico. Por lo tanto, también pueden utilizarse los genes correspondientes a estos productos primarios de la traducción como genes diana mediante los métodos de la invención para producir dominios proteicos funcionales utilizando exones endógenos.

La célula anfitriona

La invención se puede aplicar, en general, a secuencias de proteínas de origen eucariota, ya que la presencia de intrones está limitada esencialmente a los genes eucariotas. Por esta razón, la célula anfitriona será típicamente una célula eucariota, tal como una célula de mamífero, aunque pueden utilizarse otras células eucariotas derivadas de plantas, insectos, levaduras y hongos. Como la mayoría de los dominios proteicos funcionales de interés pertenecen a proteínas humanas, se prefieren las células anfitrionas humanas.

Puede utilizarse cualquier célula eucariota que contenga al menos una copia del gen que contiene intrones que codifica el producto primario de la traducción que contiene el dominio proteico funcional de interés, aunque las células eucariotas preferidas son células de mamífero diferenciadas y/o immortalizadas, en particular de origen humano, de simio y de roedor, como células CV1 de riñón de mono verde Africano transformadas con SV40 (más conocidas como células COS), células de Ovario de Hámster Chino (CHO), Riñón Embrionario Humano (HEK)-293, células de Riñón de Hámster (BHK), células de Riñón Canino Madin-Darby (MDCK) y cualquier otra línea celular eucariota madre, diferenciada o indiferenciada en la que estén presentes el exón o exones que codifican la proteína funcional de interés.

Antes de ser sometidas a los métodos de la invención, dichas células pueden estar ya modificadas a nivel genómico con otras construcciones virales o no virales, integradas mediante recombinación homóloga o no homóloga, que puede alterar la expresión y/o la estructura del gen diana o de otros genes. En el caso de líneas celulares transformadas o immortalizadas, cuando pueden estar presentes más de dos copias del gen diana, la unidad reguladora puede insertarse en una o más de las localizaciones posibles mediante ciclos sucesivos de recombinación homóloga.

La célula anfitriona debe elegirse teniendo también en cuenta la unidad reguladora exógena integrada en el extremo 5' ó 3' del exón o exones relevantes, de manera que dichas secuencias puedan ejercer completamente su actividad de una manera constitutiva o inducible. Por ejemplo, los métodos de la invención pueden aplicarse a híbridos de células somáticas de células immortalizadas, que pueden ser muy útiles para la expresión de dominios proteicos funcionales específicos, en particular, para péptidos o polipéptidos funcionales derivados de inmunoglobulinas o para cualquier otro dominio proteico funcional cuya transcripción se haya puesto, utilizando el método de la invención, bajo el control de elementos promotores y/o amplificadores específicos de inmunoglobulinas.

Debido a que las células primarias pueden modificarse mediante recombinación homóloga en cultivo con unas frecuencias comparables a las de las líneas celulares immortalizadas (Hatada S *et al.*, Proc Natl Acad Sci USA. (2000), 97(25): 13807-11), el método de la invención también es aplicable a células primarias, siempre que la producción del dominio proteico funcional se consiga en un entorno celular similar, por ejemplo, para propósitos de terapia génica. Obviamente, para la producción de un dominio proteico funcional de origen humano, las células anfitrionas serán células humanas.

Los criterios adicionales que pueden aplicarse para seleccionar el tipo de célula anfitriona para utilizarse en el método de la invención son la capacidad inherente del tipo de célula anfitriona para permitir la recombinación homóloga y el estado real de transcripción del gen diana. Una evaluación preliminar de estas características entre los tipos de célula candidatos puede ayudar en la elección del tipo celular en el que el método puede permitir un aislamiento más rápido y más sencillo de los clones que producen el dominio proteico funcional deseado.

La frecuencia de la recombinación homóloga se ha medido y se ha comparado en líneas celulares de fibroblastos de primate y murinos, mostrando diferencias importantes (Taghian DG y Nickoloff JA, Mol. Cell. Biol. (1997), 17(11): 6386-93). Ensayos similares, basados en extractos nucleares y/o transformación con un vector estándar de recombinación homóloga, pueden ayudar a cuantificar la actividad de recombinación de un tipo celular específico.

El efecto de la transcripción de genes diana en el direccionamiento de genes en células humanas cultivadas se ha evaluado comparando los efectos del direccionamiento génico en diferentes loci, en presencia o ausencia de un agente que estimula la transcripción de un sitio diana (Thyagarajan B *et al.*, Nucleic Acids Res. (1995), 23(14): 2784-90). El direccionamiento génico parece que está incrementado generalmente por la transcripción a través del sitio diana de una manera significativa.

Finalmente, la célula anfitriona debería expresar las enzimas que modifican el dominio proteico funcional después de la traducción que son distintas de las necesarias para generarlo a partir del producto primario de la traducción del gen diana. Por ejemplo, si los dominios proteicos funcionales corresponden al extremo N-terminal de un producto primario de la traducción secretado, la célula anfitriona debería permitir el procesamiento del péptido señal. En otras acciones específicas, la célula anfitriona debería permitir la glicosilación o fosforilación correcta del dominio proteico funcional. Si un tipo celular expresa el gen diana en las condiciones del cultivo celular pero no expresa la actividad proteolítica necesaria para generar el dominio proteico funcional y/o no expresa estos genes a un nivel suficiente para la explotación comercial, los métodos de la invención permiten generar clones que producen el dominio proteico funcional más eficientemente.

El módulo de transcripción

El módulo de transcripción, una primera secuencia de ADN comprendida en la unidad reguladora exógena, proporciona los elementos de control de la transcripción que no están en un extremo del exón o exones diana para obtener un transcrito primario que comprende, entre la totalidad de los exones que pertenecen al gen diana, sólo el o los que codifican el dominio proteico funcional. El módulo de transcripción debería integrarse en 5' respecto al exón más próximo al extremo 5' (en el caso de un dominio proteico funcional correspondiente al extremo C-terminal del producto primario de la traducción; Figura 1 A-F) o en 3' respecto al exón más próximo al extremo 3' (en el caso de un dominio proteico funcional correspondiente al extremo N-terminal del producto primario de la traducción; Figura 1 G-L) que codifica el dominio proteico funcional.

Dependiendo de la posición y de la secuencia del dominio proteico funcional, pueden incluirse diferentes clases de secuencia de ADN en el módulo de transcripción que se va a integrar en el gen diana, orientadas en la misma dirección: promotores, amplificadores, sitios de reconocimiento para los factores de transcripción, sitios de poliadenilación y cualquier otra clase de secuencia de ADN capaz de regular la transcripción de un ADN en una molécula de ARNm, incluyendo secuencias comprendidas en intrones opcionalmente añadidas a la construcción.

Los promotores se definen funcionalmente como sitios en los que el complejo proteico de la ARN Polimerasa II empieza la transcripción de un gen. Los amplificadores y otros sitios de reconocimiento para los factores de la transcripción pueden potenciar la actividad del promotor interaccionando con proteínas auxiliares y facilitando la agregación de un complejo de la transcripción activo. Un promotor, con o sin amplificadores, es un elemento necesario de la construcción de ADN insertada cuando el dominio proteico funcional está localizado en el extremo C-terminal del producto primario de la traducción.

La combinación de promotores y amplificadores utilizada preferiblemente es la que se sabe que activa, constitutivamente o después de una inducción, la expresión de un gen en la línea celular anfitriona cuando se lleva a cabo la recombinación homóloga. Por ejemplo, si la línea celular anfitriona consiste en células pituitarias que expresan naturalmente proteínas tales como la hormona de crecimiento y la prolactina, podría utilizarse el promotor de cualquiera de estos genes. También son adecuados los segmentos de ADN reguladores promiscuos o constitutivos que funcionan en la mayoría de los tipos celulares, tales como los elementos reguladores identificados en el Virus del Sarcoma de Rous (RSV), Virus de Simio 40 (SV40), Virus del Tumor Mamario de Ratón (MMTV), Virus Murino de la Leucemia de Moloney (MoMLV), Citomegalovirus (CMV), Sindbis (SG). Los ejemplos adicionales de promotores son los que regulan la transcripción de genes humanos tales como el Interferón alfa (IFN-alfa), proteínas de Choque por Calor (HSP), factor de Elongación 1.alfa (EF-1.alfa), Metalotioneína-I/II (MT-I/II), Ubiquitina C (UbC), Leucosialina (LS). Estos últimos promotores son útiles especialmente cuando las células anfitrionas están muy diferenciadas, como las células T (en las que es activo el promotor LS).

Los promotores inducibles son deseables para la producción de dominios proteicos funcionales que, por cualquier razón, pueden ser tóxicos y/o inhibidores del crecimiento para la célula anfitriona. Los ejemplos de promotores inducibles son MT-I/II (que contiene múltiples elementos de respuesta a metales activos en presencia de metales pesados) y *Lac* (un sistema operador-represor bacteriano, inducido por IPTG, que se ha adaptado a células de mamífero).

Cuando el dominio proteico funcional está localizado en el extremo N-terminal del producto primario de la traducción, el módulo de transcripción debería contener secuencias que permitan la terminación y modificación correctas del ARNm en su extremo 3', un proceso complejo que implica la escisión del transcrito primario y una reacción de poliadenilación acoplada. El resto poli(A) está presente en la mayoría de los ARNm de mamíferos y es esencial para la estabilidad del ARNm y para la eficacia de la traducción. Las señales para la poliadenilación están compuestas por una secuencia AATAAA, localizada 20-30 nucleótidos antes del extremo 5' del sitio de poliadenilación y un segmento rico en GT, inmediatamente después del extremo 3' del sitio de poliadenilación. Existen varias señales poli(A) eficaces que ya se han utilizado en vectores de expresión y que pueden utilizarse en la unidad reguladora, aisladas en genes eucariotas (hormona de crecimiento bovina, beta-globina de ratón) o virales (unidad de transcripción temprana de SV40, timidina quinasa del virus del Herpes simple).

Es evidente que, debido a que el nivel de expresión de un gen está determinado principalmente por el promotor y las demás regiones reguladoras de la transcripción en 5' respecto al gen, el gen diana modificado con una construcción que contiene sitios de terminación de la transcripción y de la traducción tiene que expresarse ya en la célula anfitriona a un nivel considerado como suficiente. En estas acciones, la elección de la célula anfitriona tiene que dirigirse a células que expresan ya fuertemente el gen que codifica el producto primario de la traducción.

En el módulo de transcripción puede estar presente, asociada con la señal poli(A), una secuencia adicional denominada el terminador de la transcripción, para asegurar que la transcripción no continúa en la secuencia genómica adyacente en 3' que no está relacionada con el dominio proteico funcional. Este evento podría dar lugar a dos resultados posibles: introducción de una secuencia innecesaria en el transcrito primario, que podría reducir o alterar la traducción del dominio proteico funcional deseado e inhibición de la actividad de un promotor en 3', que puede controlar un gen importante para la replicación o el metabolismo de la célula anfitriona. Incluso si no se ha definido un consenso claro por el análisis de diferentes ARNm, algunas de estas secuencias se han caracterizado bien en la bibliografía (Petitclerc D *et al.*, J. Biotechnol. (1995), 40(3): 169-78).

La construcción de ADN puede comprender adicionalmente otras secuencias de ADN que afectan la transcripción. Por ejemplo, la secuencia de ADN denominada dominios abiertos de cromatina (UCOE), si se inserta en la proximidad de un gen, puede permitir una mejor expresión de un gen poco expresado o silencioso protegiéndolo de secuencias reguladoras negativas potenciales cercanas al gen diana o forzando la apertura de dominios de cromatina cercanos. Se ha mostrado que varios de dichos elementos incrementan la expresión génica a partir de promotores heterólogos de una manera dependiente o independiente de tejido tanto en ratones transgénicos como en líneas celulares cultivadas (WO 00/05393).

El módulo de traducción

El módulo de traducción, una segunda secuencia de ADN comprendida en la unidad reguladora exógena, proporciona los elementos de control de la traducción que no están en un extremo del exón o exones diana en el marco correcto para obtener la traducción correcta y eficaz del transcrito primario en el dominio proteico funcional. El módulo de traducción debería integrarse entre el módulo de transcripción y el exón más próximo al extremo 5' (en el caso de un dominio proteico funcional correspondiente al extremo C-terminal del producto primario de la traducción) o en el exón más próximo al extremo 3' (en el caso de un dominio proteico funcional correspondiente al extremo N-terminal del producto primario de la traducción) que codifica el dominio proteico funcional y orientado en la misma dirección.

Dependiendo de la posición y de la secuencia del dominio proteico funcional, pueden incluirse diferentes clases de secuencias de ADN en el módulo de traducción: un codón de inicio de la traducción (que, junto con el contexto de nucleótidos circundante, forma la secuencia Kozac), un codón de parada de la traducción, una región 5'/3' no traducida y cualquier otra clase de secuencia de ADN capaz de regular la traducción de un ARNm en una proteína.

En el módulo de traducción tiene que introducirse un codón de inicio de la traducción, que es habitualmente ATG (que codifica Metionina) siempre que el dominio proteico funcional esté localizado en el extremo C-terminal del producto primario de la traducción y que la secuencia endógena (bien un intrón o un exón) no contenga un ATG en una posición conveniente para utilizarse para la traducción correcta del dominio proteico funcional. En estas acciones, el módulo de traducción debería contener un codón ATG exógeno englobado en una secuencia que está comprendida en el grupo de secuencias consenso definidas por Kozac para tener una eficacia óptima de inicio de la traducción. Dichos consensos (CC(A/G)CCATGG) emergieron a partir del análisis de la secuencia de inicio de la traducción de centenares de ARNm, aunque no todos los nucleótidos son igual de importantes: una o más citosinas pueden sustituirse con otro nucleótido pero la purina (A/G) debe conservarse (Kozac M, Gene (1999), 234(2): 187-208).

La región 5' no traducida (5'UTR) está relacionada con el codón de inicio de la traducción ya que es la secuencia que pertenece al transcrito primario en 5' respecto al codón de inicio de la traducción. Físicamente, es la secuencia entre el sitio de inicio de la transcripción (habitualmente 20-30 nucleótidos después del extremo 3' del promotor y constituida por un nucleótido G que va a ser modificado por las enzimas que forman una caperuza) y el codón de inicio de la traducción. Dependiendo de la secuencia del intrón diana de la recombinación homóloga y del exón adyacente que codifica el dominio proteico funcional, esta secuencia puede estar constituida completamente (si el ATG también se introduce por recombinación homóloga) o sólo parcialmente por secuencias exógenas. En la bibliografía no se ha definido una longitud específica o secuencia consenso para la región 5' no traducida pero, para minimizar las interferencias en la traducción correcta y eficaz del dominio proteico funcional, no debería exceder de 100-200 nucleótidos. Además, no debería contener ATG adicionales u otras secuencias (tales como regiones ricas en GC) que pueden emparejar y crear una estructura secundaria que pueda retrasar o parar la progresión del ribosoma durante la traducción. En determinadas acciones, por ejemplo cuando una 5'UTR es particularmente larga, puede incorporarse en la 5'UTR un sitio de entrada interno ribosomal (IRES) o un elemento amplificador de la traducción para facilitar la interacción del transcrito primario con las proteínas ribosomales e incrementar la eficacia de la traducción, como se ha mostrado para diferentes ARNm en varios tipos celulares de mamífero (Liu X *et al.*, Anal. Biochem. (2000), 280(1): 20-8).

Cuando el dominio proteico funcional está localizado en el extremo N-terminal del producto primario de la traducción, el módulo de traducción debería contener secuencias que permitan la terminación correcta de la traducción tales como un codón de terminación y una región 3' no traducida (3'UTR). Como se ha mencionado anteriormente, dependiendo de la secuencia del intrón diana de la recombinación homóloga y de la del exón adyacente que codifica el dominio proteico funcional, estas secuencias pueden estar constituidas completamente o parcialmente por una secuencia exógena pero, en la mayoría de los casos, el módulo de traducción proporcionará ambos elementos. Aparte del codón de parada bien establecido (TGA, TAA, TAG), la secuencia que rodea a este triplete puede tener algunos efectos en la eficacia de la terminación de la traducción. Si, por ejemplo, el nucleótido que sigue inmediatamente al codón de parada es una A o G, la terminación es más eficaz.

Como se ha indicado para la región 5' no traducida, para la región 3' no traducida, que consiste en el segmento del transcrito primario entre el codón de parada de la traducción y el sitio de poliadenilación, tampoco se describe en la bibliografía una longitud específica ni una secuencia consenso. Sin embargo, para una máxima eficacia de la transcripción, debería contener regiones desestabilizadoras, como secuencias ricas en A-T.

El codón de terminación o inicio de la traducción tiene que estar en el mismo marco de lectura que la secuencia que codifica el dominio proteico funcional. Si el intrón próximo al exón o exones que codifican el dominio proteico funcional contiene un trinucleótido correspondiente a un codón de parada o de inicio en el marco correcto y a una distancia que implica la adición de un número de aminoácidos compatible con la actividad del dominio proteico funcional, la integración de un módulo de transcripción apropiado, con un módulo de traducción que sólo contiene una región no traducida conveniente, puede permitir directamente la expresión correcta del dominio proteico funcional. De esta manera, estos trinucleótidos y la secuencia intrónica comprendida entre ellos y el extremo más próximo del exón adyacente puede ser completamente funcional ya que se convierten en parte de este exón.

La secuencia peptídica que resulta fusionada en el extremo N o C-terminal de la proteína funcional no debería interferir con la funcionalidad de esta secuencia proteica o, si lo hace, debería eliminarse o inactivarse fácilmente durante el proceso de purificación. En el caso de un dominio proteico funcional correspondiente al extremo C-terminal del producto primario de la traducción, también puede utilizarse el mismo método si el exón más próximo al extremo 5' contiene un ATG, que codifica normalmente una Metionina interna, en marco con el dominio proteico funcional, que limita su extremo N-terminal. La integración de un módulo de transcripción que activa la transcripción en 5' respecto a dicho ATG debería permitir la transcripción y traducción correctas del dominio proteico funcional.

Si embargo, la presencia de dichos trinucleótidos, que son habitualmente inactivos como sitios de inicio o terminación de la traducción ya que son eliminados durante el corte y empalme (si están comprendidos en un intrón) o anticipados por otros sitios de inicio de la traducción (sin están comprendidos en un exón), está limitada a un número pequeño de genes. Por lo tanto, el codón de inicio o parada de la traducción debería estar comprendido en un exón, bien sintético o natural, que se convierte en parte del módulo de traducción. Este módulo puede comprender uno o más exones exógenos (separados por secuencias intrónicas naturales o sintéticas) que pueden codificar adicionalmente una secuencia proteica bien homóloga o heteróloga respecto de secuencias proteicas comprendidas en el producto primario de la traducción codificado por el gen diana.

Sin embargo, como en la situación previa, esta secuencia proteica fusionada con la proteína funcional no debería interferir con la actividad del dominio proteico funcional o, si lo hace, debería eliminarse o inactivarse fácilmente, opcionalmente mediante actividades enzimáticas producidas por la célula anfitriona misma. Por ejemplo, si el dominio proteico funcional es un fragmento C-terminal de un producto primario de la traducción, el exón o exones pueden codificar un péptido señal (o el extremo N-terminal de un péptido señal) que, una vez puesto en el marco correcto entre el sitio de inicio de la traducción y la secuencia exónica endógena, permite la secreción del dominio proteico funcional en el medio de cultivo.

Alternativamente, esta secuencia proteica adicional puede ser simplemente un péptido espaciador o conector, que facilita opcionalmente la purificación y/o recogida del dominio proteico funcional. La secuencia proteica adicional también puede codificar el sitio de reconocimiento de una enzima proteolítica de manera que, si el dominio proteico funcional puede purificarse explotando la afinidad de la secuencia adicional por un sustrato inmovilizado en un soporte, puede eliminarse posteriormente utilizando proteasas disponibles comercialmente.

El módulo de corte y empalme

Como se ha indicado anteriormente, si la secuencia endógena que pertenece al gen diana no contiene un codón de terminación o inicio de la traducción en una posición apropiada para ser explotado (es decir, adyacente o interno al exón o exones que codifican el dominio proteico funcional y con el mismo marco del dominio proteico funcional), la construcción para la recombinación homóloga contendrá a uno de ellos en un exón exógeno, bien natural o sintético.

El exón o exones exógenos que pertenecen al módulo de traducción, sin embargo, tiene que situarse en marco con el exón o exones endógenos. Esta organización puede conseguirse utilizando un sitio de corte y empalme que complementa el sitio de corte y empalme del exón más próximo (Figura A, D, G, J) o eligiendo la secuencia de direccionamiento de manera que el exón se fusione de manera precisa con el exón más próximo (Figura B, E, H, K).

Por lo tanto, cuando el exón exógeno contiene un codón de inicio de la traducción, el módulo de corte y empalme será un sitio donador de corte y empalme en 5' localizado en el extremo 3' del módulo de traducción (Figura 1 A, D), que complementa el sitio aceptor de corte y empalme en 3' asociado al exón endógeno cercano que codifica el dominio proteico funcional. Alternativamente, cuando el exón sintético contiene un codón de parada de la traducción, el módulo de corte y empalme será un sitio aceptor de corte y empalme en 3' localizado en el extremo 5' del módulo de traducción (Figura 1 G, J), que complementa el sitio donador de corte y empalme en 5' asociado al exón endógeno cercano que codifica el dominio proteico funcional.

La unidad reguladora puede comprender, por lo tanto, un módulo de corte y empalme que permita la generación de un transcrito que codifica un dominio proteico funcional siempre que, después de la integración de la unidad reguladora, las secuencias intrónicas residuales separen el módulo de traducción del exón o exones que codifican el

dominio proteico funcional. Las secuencias de corte y empalme endógenas que dirigen habitualmente la fusión del exón o exones distintos de los que codifican el dominio proteico funcional con el o los que están bajo el control de la unidad reguladora exógena, deberían ser incapaces de reconstituir los eventos de corte y empalme del gen diana en la célula anfitriona. Esto se debe al hecho de que la unidad reguladora los reemplaza y/o está situada de manera que están demasiado lejos y no pueden ejercer su actividad eficazmente, permaneciendo sin emparejar.

Los sitios 5'/3' de corte y empalme son elementos fundamentales para la expresión de un gen, ya que permiten el corte y empalme apropiado de los transcritos primarios generados a partir del gen interrumpido por intrones. Estas secuencias están muy conservadas en muchos genes de vertebrados, especialmente en el extremo 5' y 3' de los intrones. La mayoría de los sitios donadores de corte y empalme en 5' en los Eucariotas superiores conforman la secuencia consenso AG||GTRAGT, en la que AG es el resto de dinucleótido conservado en el extremo 3' del exón, || indica el sitio de corte y empalme, GT es el resto de dinucleótido altamente conservado en el extremo 5' del intrón y R es una purina. Un sitio aceptor de corte y empalme en 3' está constituido esencialmente por la secuencia consenso YAG||G, en la que AG es el resto de dinucleótido altamente conservado en el extremo 3' del intrón (precedido típicamente por una cadena de pirimidina (Y)), || indica el sitio de corte y empalme y G es un nucleótido conservado en el extremo 5' del exón.

Aparte de la secuencia de corte y empalme consenso apropiada, el módulo de corte y empalme está seguido de secuencias intrónicas que provienen de la región de direccionamiento adyacente, de secuencias intrónicas naturales o sintéticas añadidas a la construcción entre el sitio de corte y empalme y la región de direccionamiento. Dichas secuencias intrónicas pueden contener, por ejemplo, otro elemento de secuencia conocido como el sitio de ramificación, que está localizado habitualmente a una distancia de 18-40 nucleótidos antes del extremo 5' del sitio de corte y empalme en 3'. Este sitio muestra la secuencia CTRACT, en la que N es cualquier nucleótido y A es el nucleótido que tiene un grupo -OH capaz de interactuar con el grupo fosfato del nucleótido G en el extremo 5' del intrón durante las etapas catalíticas del corte y empalme del ARNm. Se han revisado las secuencias consenso, el mecanismo, los factores y otras secuencias reguladoras implicadas en el corte y empalme del pre-ARNm de mamíferos, como los amplificadores de corte y empalme exónicos, (Long M *et al.*, Proc. Natl. Acad. Sci USA, (1998), 95(1): 219-223; Blencowe BJ, Trends Biochem. Sci. (2000), 25(3): 106).

Los avances en el conocimiento de los mecanismos de corte y empalme, aunque no permiten una definición más precisa de las secuencias consenso generales, pueden dar lugar a una selección de módulos de corte y empalme específicos en relación con el tipo celular que se va a modificar por recombinación homóloga. Por ejemplo, se ha encontrado que una serie de características y de restos de la secuencia es común en genes sometidos a corte y empalme específicamente en el cerebro, lo que indica que puede ser posible una regulación del corte y empalme específica del tipo celular seleccionando dichas secuencias (Brudno M *et al.*, Nucleic Acids Res. (2001), 29(11): 2338-48).

El módulo de corte y empalme está separado del elemento de control de corte y empalme endógeno, presente en el extremo cercano del exón adyacente que codifica el dominio proteico funcional en el gen diana, por una secuencia que contiene intrones que puede ser completamente endógena, completamente exógena o un híbrido de secuencias endógenas y exógenas. Como esta secuencia se eliminará finalmente por corte y empalme, la secuencia precisa puede no importar, siempre que no interfiera de manera adversa con la expresión apropiada del dominio proteico funcional deseado.

La secuencia de ADN de direccionamiento

La secuencia de ADN de direccionamiento es un elemento esencial de la construcción ya que es responsable de la integración y posicionamiento correctos de las secuencias exógenas (unidad reguladora, gen marcador positivo) en el genoma de la célula. Dichas secuencias, habitualmente clonadas o amplificadas por PCR a partir de ADN genómico, se definen funcionalmente como poseedoras de un nivel de homología con el ADN endógeno suficiente para dirigir los procesos moleculares que causan la recombinación homóloga (emparejamiento y desplazamiento de cadenas) en una región genómica específica.

La secuencia de ADN de direccionamiento puede estar constituida por un único segmento de ADN o dividida en dos segmentos de ADN, separados en la construcción por las secuencias exógenas correspondientes a la unidad reguladora y, opcionalmente, por el gen marcador positivo. Aunque se prefieren dos segmentos de direccionamiento con el fin de incrementar la eficacia y la precisión de la integración, la presente invención también engloba la utilización de un único segmento de direccionamiento. En su forma más sencilla, se utiliza un fragmento circular de ADN que contiene la unidad reguladora junto con el segmento de direccionamiento. De esta manera, el segmento de direccionamiento homólogo hibrida con su equivalente genómico y la unidad reguladora se inserta en el gen diana después del evento de entrecruzamiento.

El tamaño de cada uno de los segmentos de direccionamiento (es decir, las regiones de homología) no es crítico, aunque cuanto más cortas son las regiones, menos probable es que encuentren las regiones de homología apropiadas y recombinen en el punto deseado. Así, cuanto más cortas son las regiones de homología, menos eficaz es la recombinación homóloga, es decir, menor es el porcentaje de clones recombinados con éxito. Se ha sugerido que el requerimiento mínimo para la homología de secuencia es 25 pares de bases (Ayares D *et al.*, Proc Natl Acad Sci USA (1986), 83 (14): 5199-203). Se consiguen resultados óptimos cuando la región total de homología, incluyendo las dos regiones de direccionamiento, es grande, por ejemplo, una a cinco kilobases o más. Siempre que la unidad reguladora pueda

introducirse en el punto apropiado del genoma, no existe límite en el tamaño de los segmentos de direccionamiento, si éste no afecta la estabilidad del vector.

Las secuencias de ADN de direccionamiento contendrán, en muchas situaciones, el intrón 5' completo o un segmento de éste (si el dominio proteico funcional está en el extremo C-terminal del producto primario de la traducción) o 3' (si el dominio proteico funcional está en el extremo N-terminal del producto primario de la traducción) del exón o exones que codifican el dominio proteico funcional (Figura 1 A, D, G, J). La región de direccionamiento puede incluir otras partes del gen diana, incluyendo el exón o exones que codifican el dominio proteico funcional, además de este intrón específico (Figura 1 B, E, H, K). En algunos casos más, el intrón o intrones y el exón o exones próximos al exón o exones que codifican el dominio proteico funcional pueden eliminarse más o menos completamente y sustituirse por la unidad reguladora exógena (Figura 1 C, F, I, L), opcionalmente utilizando la secuencia localizada en un gen contiguo. Si la secuencia de ADN de direccionamiento está constituida por un único segmento de ADN, debería ser homóloga a un segmento del gen diana comprendido sólo en la parte del gen diana que incluye el exón o exones que codifican el dominio proteico funcional, a un segmento del gen diana comprendido sólo en la parte del gen diana que incluye el intrón próximo al exón o exones que codifican el dominio proteico funcional o a un segmento del gen diana que incluye secuencias que pertenecen a estas dos regiones. En cualquier caso, la integración de la construcción en el ADN genómico de la célula anfitriona por recombinación homóloga en una posición determinada por el segmento o segmentos de ADN de direccionamiento debería permitir la expresión del dominio proteico funcional bajo el control de la unidad reguladora.

Dichas estrategias de direccionamiento son factibles, siempre que la integración de la construcción no dé lugar a ninguna modificación del orden y/o secuencia de los elementos genómicos que codifican el dominio proteico funcional o de otras secuencias genómicas que son necesarias para la viabilidad o metabolismo celular. Si el producto primario de la traducción del gen diana es en sí mismo esencial para la viabilidad o metabolismo celular, la copia de dicho gen no modificada por recombinación homóloga debería tener un nivel de expresión suficiente para mantener a la célula activa metabólicamente.

El conocimiento de la secuencia y de la estructura completa del gen, desde el promotor hasta el sitio de poliadetilación, facilita la elección de la estrategia de direccionamiento más apropiada, pero puede ser posible generar la construcción empezando justo desde la secuencia y la estructura de los exones e intrones asociados con el dominio proteico funcional. Esto puede ocurrir, en particular, en el caso de los genes humanos para los que sólo está disponible un clon genómico que contiene una secuencia candidata buena para el extremo 5' ó 3', en particular, cuando se compara con el gen de ratón correspondiente ya secuenciado y caracterizado. Tomando como base la homología con otros genes en el mismo u otros organismos y predicciones informáticas (Rogic *S et al.*, Genome Res. (2001), 11(5): 817-32), pueden definirse en el clon genómico las uniones relevantes exón/intrón junto con las secuencias circundantes que limitan en un extremo el dominio proteico funcional. Por lo tanto, pueden seleccionarse fragmentos del clon genómico ya que tienen una longitud suficiente para generar las secuencias de direccionamiento necesarias para guiar a la unidad reguladora a la localización genómica correcta.

La secuencia de ADN de direccionamiento puede contener secuencias homólogas al gen diana que son contiguas o no contiguas en el genoma diana. En el primer caso, se producirá la simple inserción de la secuencia exógena, bien internamente o en un extremo del intrón próximo al exón o exones que codifican el dominio proteico funcional (Figura 1 A, B, D, E, G, H, J, K). En el segundo caso, la parte del gen diana que separa las secuencias no contiguas próximas al exón o exones que codifican el dominio proteico funcional (intrón o intrones y/o exón o exones) se delecionarán y reemplazarán por las secuencias exógenas como consecuencia de la recombinación homóloga (Figura 1 C, F, I, L). En ambos tipos de vectores, la recombinación homóloga está dirigida por secuencias de ADN del ADN exógeno entrante que pueden alinearse directamente con secuencias homólogas en el gen diana que codifica el dominio proteico funcional. La disposición lineal particular de las secuencias homólogas en el ADN exógeno determinará la posición y la orientación de la unidad reguladora al nivel de la región intrónica próxima al exón o exones que codifican el dominio proteico funcional.

Siempre que los segmentos de direccionamiento se obtienen por PCR y/o derivan de una célula no isogénica con la célula anfitriona, deben ser secuenciados cuando se prepara la construcción para identificar cualquier diferencia en la secuencia (debida al origen de la región de direccionamiento) que puede alterar significativamente la secuencia y expresión esperadas del dominio proteico funcional.

Mientras que el tipo de construcción de direccionamiento elegido determinará la naturaleza de la modificación genómica del gen diana que resulta en un gen recombinante que codifica un dominio proteico funcional, la eficacia real con la que una construcción dada puede utilizarse para obtener líneas celulares diana depende principalmente de la construcción de direccionamiento.

En particular, se ha mostrado que la frecuencia absoluta de direccionamiento que puede conseguirse con la construcción depende de un número de factores, incluyendo la longitud de las secuencias homólogas en la construcción de direccionamiento, el grado de homología entre las secuencias de la construcción de direccionamiento y el gen diana y la región genómica diana particular. Se ha mostrado que la frecuencia de direccionamiento incrementa al incrementar la longitud de la homología de secuencia entre el vector de direccionamiento y el locus, hasta que se alcanza un plató en la frecuencia de direccionamiento entre 10-14 kb, sin una diferencia consistente en la frecuencia entre los vectores de inserción y de reemplazamiento (Deng C y Capecchi MR, Mol Cell Biol. (1992), 12(8): 3365-71). Este plató puede

reflejar un límite en el tamaño de fragmentos de ADN intactos que pueden introducirse en las células, en lugar de un límite en el efecto de la longitud de la homología en la frecuencia de direccionamiento.

Respecto al nivel de homología requerido para dirigir la integración correcta de la construcción por recombinación homóloga, la secuencia de ADN de direccionamiento debe hibridar con secuencias endógenas en las condiciones de hibridación astringentes descritas en la bibliografía (Sambrook *et al.*, "Molecular Cloning: A Laboratory Manual", Cold Spring Harbor Press, 1989), por ejemplo, utilizando las condiciones de lavado siguientes: 2x SCC, 0,1% SDS, temperatura ambiente dos veces, 30 minutos cada vez; después 2 x SCC, 0,1% SDS, 50 C una vez durante 30 minutos; después 2 x SCC, temperatura ambiente dos veces, 10 minutos cada vez.

Pueden identificarse secuencias homólogas que contienen como máximo aproximadamente 25-30% de emparejamientos erróneos de pares de bases. Más preferiblemente, las cadenas de ácido nucleico homólogas contienen 15-25% de emparejamientos erróneos de pares de bases, incluso más preferiblemente 5-15% de emparejamientos erróneos de pares de bases. Estos grados de homología pueden seleccionarse utilizando condiciones de lavado más astringentes para la identificación de clones de bibliotecas génicas (u otras fuentes de material genético), como es muy conocido en la técnica.

El gen marcador seleccionable

La construcción también puede comprender uno o más de: un gen de selección positiva, un gen amplificable y un gen de selección negativa. La construcción utilizada para la recombinación homóloga puede contener secuencias exógenas distintas de las que constituyen la unidad reguladora. En particular, puede añadirse a la construcción uno o más de un gen de selección positiva, un gen amplificable o un gen de selección negativa para facilitar la identificación de los clones transformados que tienen la unidad reguladora integrada correctamente en el genoma. Por esta razón, el gen o genes amplificables y/o marcadores están situados en la construcción entre las regiones de direccionamiento, habitualmente entre el módulo de transcripción y una de las regiones de direccionamiento (Figura 1 D-F, J-L). Sea cual sea el gen marcador seleccionable utilizado, debe constituir una unidad de transcripción y traducción distinta de la del dominio proteico funcional y, para evitar cualquier interferencia con el último, puede separarse de la unidad reguladora por secuencias que evitan cualquier evento de "read-through" (transcripción mas allá del gen), tal como los terminadores de la transcripción descritos anteriormente. Opcionalmente, el gen marcador positivo puede carecer de secuencia señal de la transcripción y/o traducción en un extremo y ésta será proporcionada por el gen diana cuando se integre apropiadamente, generando un gen de fusión.

Un gen marcador seleccionable positivo es capaz de hacer que la célula anfitriona transfectada sea resistente a un entorno que normalmente es tóxico. Los ejemplos de dichos genes son adenosina desaminasa (ADA), aminoglicósido fosfotransferasa (neo), dihidrofolato reductasa (DHFR), higromicina-B-fosfotransferasa (HPH), timidina quinasa (tk), xantina-guanina fosforibosiltransferasa (gpt), gen de resistencia a múltiples fármacos (MDR), ornitina descarboxilasa (ODC) y N-(fosfonacetil)-L-aspartato (CAD).

Además, o como alternativa del gen marcador seleccionable positivo, en la construcción también puede incluirse opcionalmente un gen amplificable. Los genes amplificables son genes que dan lugar a un incremento del número de copias cuando están bajo presión selectiva. El número de copias de un gen situado adyacente al gen amplificable, incluyendo el gen nuevo que codifica el dominio proteico funcional, también incrementará. Los genes amplificables que pueden utilizarse incluyen DHFR, MDR, ODC, ADA y CAD. Los miembros del grupo de genes marcadores seleccionables positivamente y los del grupo de genes amplificables se superponen de manera que, en teoría, en lugar de utilizar dos genes, uno para la selección positiva y otro para la amplificación, podría utilizarse un gen para ambos propósitos. Sin embargo, como la mayoría de las líneas celulares contienen copias endógenas de estos genes amplificables, las células serán ya de alguna manera resistentes a las condiciones de selección y puede ser difícil distinguir a las células que tienen ADN transfectado de las que no reciben el ADN transfectado. Así, en los casos en los que se desea un gen amplificable, también debería incluirse en la construcción un gen de selección positiva que sea dominante, tal como HPH, gpt, neo y tk (en células tk⁻). Para algunas aplicaciones puede ser posible o preferible omitir el marcador amplificable, incluso si un incremento del número de copias del gen nuevo que codifica el dominio proteico funcional puede proporcionar finalmente una cantidad mayor de esta proteína. La amplificación puede no ser necesaria, por ejemplo, cuando la unidad reguladora dirige la transcripción y la traducción del dominio proteico funcional muy eficazmente. También es posible eliminar el gen de selección positiva y seleccionar las células solamente mediante cribado para determinar la producción de la proteína o ARNm deseado. Sin embargo, en la mayoría de los casos se prefiere incluir al menos el gen de selección positiva.

En la construcción también puede estar presente un gen marcador seleccionable negativo, externamente a las secuencias de direccionamiento. Dicho gen no se expresa en las células en las que la construcción de ADN se inserta apropiadamente por recombinación homóloga ya que se elimina, pero se expresa en las células en las que la construcción de ADN no se inserta apropiadamente, tal como por integración aleatoria. Si el vector se inserta correctamente por recombinación homóloga, recombinará en las regiones de homología, causando la pérdida de secuencias fuera de estas regiones. Uno de dichos genes es el gen de la timidina quinasa del Virus del Herpes Simple (HSVtk). El HSVtk tiene una menor astringencia para los nucleótidos y es capaz de fosforilar análogos de nucleótidos que las células de mamífero normales son incapaces de fosforilar. Si el HSVtk está presente en las células, los análogos de nucleótidos tales como aciclovir y ganciclovir son fosforilados e incorporados en el ADN de la célula anfitriona matando así a la célula.

Sea cual sea el gen marcador utilizado, está reconocido en la bibliografía que dicho gen puede eliminarse posteriormente utilizando una recombinasa específica de sitio (como Flp o Cre), presente en el genoma de la célula, co-transfectada con la construcción para la recombinación homóloga o introducida posteriormente por transformación o por cualquier tecnología recombinante. Dichos procedimientos, que pueden ser necesarios si el gen marcador afecta la transcripción de los genes cercanos, se han descrito y pueden utilizarse opcionalmente para activar o inactivar elementos de la unidad reguladora (Gorman C y Bullock C, Curr Opin Biotechnol. (2000), 11(5): 455-60; Kuhn R y Schwenk F, Curr. Opin. Immunol. (1997), 9(2): 183-8).

La construcción

La construcción, o vector de direccionamiento, que comprende las secuencias de ADN, como se ha descrito anteriormente, se introduce, en una forma linearizada o circular, en una célula anfitriona de manera que la región de ADN de direccionamiento pueda hibridar con la secuencia genómica homóloga y permitir la recombinación homóloga entre las secuencias endógenas y exógenas, que están integradas de manera estable en el genoma de la célula anfitriona.

El ensamblaje de la construcción debe tener en cuenta la orientación del exón o exones que codifican el dominio proteico funcional. Así, la secuencia de direccionamiento debe clonarse en la construcción para permitir a los elementos comprendidos en la unidad reguladora integrarse por recombinación homóloga en la misma orientación, determinando el enlace operativo entre la unidad reguladora exógena y la región genómica endógena que comprende el exón o exones que codifican el dominio proteico funcional. Por el contrario, el gen o genes marcadores seleccionables pueden tener cualquier orientación pero, como se ha dicho antes, no deben interferir con la actividad de la unidad reguladora exógena y, por lo tanto, se sitúan habitualmente el módulo de transcripción y la secuencia de direccionamiento más próxima (Figura 1 D-F, J-L).

Las secuencias de ADN que pertenecen a la construcción que son distintas de la unidad reguladora y del gen o de los genes marcadores seleccionables, que pueden integrarse opcionalmente por recombinación homóloga (relacionadas con la replicación o selección en células bacterianas, por ejemplo), tampoco deben afectar la transcripción o traducción del dominio proteico funcional.

Producción de un dominio proteico funcional

Una vez que se conoce la secuencia de aminoácidos y la estructura génica de un dominio proteico funcional, su producción puede obtenerse utilizando los métodos de la invención utilizando el procedimiento siguiente:

- Identificación de la región génica diana en la que la integración (por inserción o por reemplazamiento) de la unidad reguladora permitirá la expresión precisa y eficaz del dominio proteico funcional;

- Construcción del vector de direccionamiento que contiene las secuencias de direccionamiento, la unidad reguladora y, opcionalmente, un gen o genes amplificables y/o seleccionables;

- Selección de una célula anfitriona conveniente que contiene las secuencias de direccionamiento y el exón o exones que codifican el dominio proteico funcional;

- Transformación de la célula anfitriona con la construcción utilizando técnicas conocidas (lipofección, electroporación, precipitación con fosfato de calcio);

- Identificación de los transformantes que contienen el gen recombinante como se esperaba después de la recombinación homóloga, utilizando tecnologías como selección positiva y/o negativa, amplificación, análisis de sitios de restricción, PCR transcripción inversa/genómica, transferencia Southern o secuenciación de ADN;

- Expansión de los transformantes seleccionados y selección de los clones que expresan correctamente el dominio proteico funcional, aplicando técnicas de análisis de ARNm (extensión de cebadores, transferencia Northern) y análisis de proteínas (transferencia Western, ELISA, secuenciación de proteínas, mapeo de epítopos, gel de poliacrilamida bidimensional, purificación por afinidad, ensayos enzimáticos, análisis de espectro CD/RMN, filtración en gel por HPLC, Espectrometría de Masa) a muestras recogidas y procesadas a partir de células cultivadas y/o del medio de cultivo utilizando tecnologías de separación habituales (lisis/disrupción celular, extracción, precipitación, cromatografía);

- Expansión adicional de los clones seleccionados en cultivo para permitir un análisis cuantitativo para elegir el clon recombinante homólogo útil para la producción comercial y para desarrollar un protocolo eficaz para expresar y purificar el dominio proteico funcional recogido a partir de los cultivos celulares expandidos.

Descripción de las figuras

La Figura 1 muestra algunas maneras posibles mediante las que una construcción de ADN de la invención puede recombinar de manera homóloga con un gen diana para obtener un gen nuevo que expresa los exones que codifican un dominio proteico funcional.

HS1 y HS2 son los segmentos de la construcción homólogos al gen diana que contiene los exones que codifican el dominio proteico funcional que se utilizan como secuencias de ADN de direccionamiento. En esta representación esquemática, el gen diana contiene 4 exones (EX1, EX2, EX3, EX4) y tres intrones (IN1, IN2, IN3). Los exones que codifican el dominio proteico funcional son los indicados con una barra negra y localizados bien en el extremo 3' (en A-F) o en el extremo 5' (en G-L) del gen diana. En D-F y J-L, la construcción también contiene un gen marcador (MK). La unidad reguladora está compuesta por el módulo de transcripción (TS) y el módulo de traducción (TL) y, siempre que es necesario eliminar por corte y empalme las secuencias intrónicas entre el módulo de traducción y el exón más próximo que codifica el dominio proteico funcional, un módulo de corte y empalme exógeno que contiene bien un sitio donador de corte y empalme (SD en A y D) o un sitio aceptor de corte y empalme (SA en G y J).

Las construcciones de ADN pueden dirigir la inserción de la unidad reguladora y, opcionalmente, de un gen marcador en la región intrónica adyacente a los exones que codifican el dominio proteico funcional, de manera que la transcripción y la traducción del dominio proteico funcional empieza (como se representa en A-F por \rightarrow) o para (como se representa en G-L por el símbolo \hookrightarrow). Los elementos reguladores de comienzo y parada de la transcripción/traducción endógenos y del gen marcador se representan simbólicamente por \rightarrow y \hookrightarrow , respectivamente.

El módulo de corte y empalme puede estar localizado en el extremo 3' (A y D) o en el extremo 5' (G y J) del módulo de traducción. Por ejemplo, si la inserción de la unidad reguladora exógena divide el intrón (IN2) en dos partes (IN2' e IN2''), las secuencias intrónicas intermedias (IN2'' en A y D; IN2' en G y J) se eliminan por corte y empalme utilizando el módulo de corte y empalme, de manera que EX3/EX4 (en A y D) o EX1/EX2 (en G y J) se transcriben y traducen selectivamente para obtener el dominio proteico funcional. Alternativamente, las construcciones de ADN pueden dirigir la inserción de la unidad reguladora y, opcionalmente, de un gen marcador entre el intrón y el exón de interés (IN2 y EX3 en B y E; IN2 y EX2 en H y K), sin dividir el intrón, de manera que EX3/EX4 (en B y E) o EX1/EX2 (en H y K) se transcriben y traducen selectivamente para obtener el dominio proteico funcional.

Incluso alternativamente, las construcciones de ADN pueden dirigir el reemplazamiento de un segmento génico endógeno adyacente a los exones que codifican el dominio proteico funcional por la unidad reguladora exógena y, opcionalmente, de un gen marcador. Como en la situación anterior, una secuencia de direccionamiento pertenece necesariamente a un exón que codifica el dominio proteico funcional (EX3 en B, C, E y F; EX2 en H, I, K y L), mientras que la otra secuencia de direccionamiento corresponde a la secuencia endógena no comprendida en la región genómica para el dominio proteico funcional. El resultado es el reemplazamiento del segmento entre las dos regiones de homología (IN2 y parte de EX2 en C y F; IN2 y parte de EX3 en I y L) por la secuencia reguladora exógena, mientras que se generan nuevas secuencias intrónicas (IN1/EX2 en C y F; EX3/IN3 en I y L).

La Figura 2 muestra un alineamiento de proteína TRANCE de ratón y humana (SEQ ID NO: 1 y SEQ ID NO: 2) en la región procesada proteolíticamente para generar sTRANCE (= significa identidad, + significa homología). Los restos que se sabe que son posiblemente el extremo N-terminal de sTRANCE de ratón (Schlondorff J *et al.*, J Biol Chem. (2001), 276(18): 14665-74) están indicados con §. La correspondencia entre el extremo N-terminal de sTRANCE, la proteína TRANCE de ratón (IC, TM y ECD significan dominio intracelular, transmembrana y extracelular, respectivamente), el gen TRANCE de ratón (EX significa exón, IN intrón, seguido del número relativo) y el cromosoma 13 humano (la numeración de los nucleótidos está referida a GenBank versión NT_009935.3) se indica con las líneas punteadas. La numeración de las secuencias de las proteínas humana y de ratón corresponde a las publicaciones originales (Anderson DM *et al.*, Nature (1997), 390(6656): 175-9; Wong BR *et al.*, J Biol Chem. (1997), 272(40): 25190-4).

La Figura 3 muestra la secuencia proteica del dominio NC1 del colágeno humano XVIII 1 alfa con los exones codificadores correspondientes (SEQ ID NO: 3). La numeración corresponde a la variante larga completa de la secuencia como se ha publicado (Oh SP *et al.*, Genomics 1994 Feb; 19(3): 494-9). Los restos caracterizados en la bibliografía como extremo N-terminal posible de los dominios proteicos funcionales de la endostatina de ratón o humana están indicados con §.

Figura 4. (A) muestra la posición de la secuencia de direccionamiento en 5' y 3' utilizada para construir pEnd-HR#1 y pEnd-HR#2 en el gen humano COL18A1. También se indican los cebadores diseñados para amplificar las secuencias genómicas de COL18A1 necesarias para construir los vectores de direccionamiento utilizando el clon original de GenBank AL163302, la longitud de los exones y de las secuencias amplificadas. (B) muestra una representación simplificada de pEnd-HR#1 y pEnd-HR#2 indicando la posición relativa de la unidad reguladora, de las secuencias de direccionamiento (líneas punteadas) y de los genes marcadores positivos y negativos. La caja muestra la secuencia que codifica el Péptido Señal de Ig de ratón (mIgSP; nucleótidos 1-56 del registro de GenBank M13329), seguido del sitio donador de corte y empalme (SD; secuencia subrayada) incluyendo el último nucleótido de esta secuencia codificadora y los primeros seis nucleótidos del intrón, elegidos entre los restos que estimulan el corte y empalme (SEQ ID NO: 4). El sitio de corte y empalme se indica con el símbolo \llcorner . El último codón TCA original del Péptido Señal de Ig de ratón se modificó a TCG para proporcionar un extremo 3' mejor del exón mIgSP para su corte y empalme correcto pero esta mutación no cambia el aminoácido correspondiente (Serina).

La Figura 5 muestra un mapa simplificado del plásmido pBS-EF1alfa-mIgSP-SD con los sitios de restricción relevantes.

La Figura 6 muestra un mapa simplificado del plásmido pGEM-3Z-mPGK-TK-HR con los sitios de restricción relevantes.

La Figura 7 muestra un mapa simplificado del plásmido pEnd-HR#1. También se indica la posición relativa de las regiones de direccionamiento en 5' y 3' (5'HR y 3'HR) de los genes para la selección positiva (SEL+) y para la selección negativa (SEL-) y la de la unidad reguladora (RU) que contiene el promotor EF1-alfa, el péptido señal de Ig de ratón y el sitio donador de corte y empalme. El sitio de restricción único *NotI* utilizado para linearizar el plásmido antes de la transfección está localizado en el extremo de 3'HR. El plásmido pEnd-HR#2 se diferencia de pEnd-HR#1 a nivel de la longitud y posición en el ADN genómico de las secuencias de direccionamiento (Figura 4A) y es ligeramente mayor (16,9 Kb).

La Figura 8 muestra la secuencia del ARNm nuevo expresado en las células en las que el plásmido pEnd-HR#1 se ha integrado correctamente en el gen COL18A1 humano, en particular, el exón mIgSP y las secuencias codificadoras endógenas. Los codones de comienzo y de parada se indican en negrita. El ARNm completo contiene 2.210 nucleótidos (sólo se muestra la parte codificadora del exón 41), mientras que el dominio proteico funcional está codificado como una proteína que contiene 275 aminoácidos (19 de ellos pertenecen al exón mIgSP y 256 de ellos pertenecen a la secuencia codificadora de los exones 38-41 de COL18A1 humano).

La Figura 9 muestra la secuencia del ARNm nuevo expresado en las células en las que el plásmido pEnd-HR#2 se ha integrado correctamente en el gen COL18A1 humano, en particular, el exón mIgSP y las secuencias codificadoras endógenas. Los codones de comienzo y de parada se indican en negrita. El ARNm completo contiene 1.964 nucleótidos (sólo se muestra la parte codificadora del exón 41), mientras que el dominio proteico funcional está codificado como una proteína que contiene 193 aminoácidos (19 de ellos pertenecen a mIgSP y 174 de ellos pertenecen a la secuencia codificadora de los exones 39-41 de COL18A1 humano).

Figura 10. (A) muestra un mapa simplificado de un fragmento pEAK-HR#2. El intrón 37 de COL18A1 y las secuencias codificadoras de los exones 39-41 se fusionaron con un epítopo FLAG y con un sitio de poliadenilación. El pEAK-HR#1 contenía el intrón 37 en lugar del intrón 38 y el exón 38 entre el intrón 37 y los exones 39-41. (B) muestra una transferencia Western realizada utilizando extractos celulares completos ensayados con un anticuerpo policlonal de conejo frente a endostatinas humanas diluido 1/100 (Chemicon AB 1878). (C) muestra una transferencia Western realizada utilizando 1,5 mililitros de medio condicionado de las mismas células transfectadas ensayados mediante inmunoprecipitación con un anticuerpo monoclonal anti-FLAG-M2 de ratón diluido 1/1.000 (Sigma F3165) y 20 microlitros de M2-FLAG-agarosa (Sigma A1205). Se diluyeron 1/10.000 anticuerpos secundarios unidos a anticuerpos marcados con peroxidasa de rábano frente a anti-conejo o anti-ratón de cabra (Amersham-Pharmacia) y se detectaron con el reactivo ECL Western Pico (Pierce). El peso molecular esperado de los dominios proteicos funcionales para pEnd-HR#1 y pEnd-HR#2 es respectivamente 31 Kd y 22 Kd (antes de la eliminación del péptido señal, como en B) ó 29 Kd y 19 Kd (después de la eliminación del péptido señal y de la secreción en el medio de cultivo, como en C). El desplazamiento del peso molecular se debe a la glicosilación a nivel de la secuencia codificada por el exón 38.

Figura 11. (A) muestra un mapa simplificado de la región codificadora de 1,0 Kb contenida en la expresada por las células humanas 293-EBNA después de la integración de pEnd-HR#1. Se indica la posición de los cebadores utilizados para identificar el transcrito. (B) muestra un mapa simplificado de la región genómica de 2,4 Kb de COL18A1 en células humanas 293-EBNA después de la integración de pEnd-HR#1. Se indica la posición de los cebadores y los sitios de restricción utilizados para caracterizar los clones.

La Figura 12 muestra un gel de agarosa en el que se separan los productos de la amplificación, obtenidos amplificando el ADNc de células seleccionadas con o-1165 y o-1175 como cebadores. Cada grupo corresponde a células obtenidas a partir de una única placa.

La Figura 13 muestra el análisis de restricción de los dos segmentos amplificados obtenidos a partir del grupo 4 de clones positivos para pEnd-HR#1, FragA y FragB. Las longitudes esperadas de los fragmentos amplificados obtenidos a partir de FragA son 501 (A), 577 (B), 618 (C) y 851 (D) bases. Las longitudes esperadas de los fragmentos amplificados obtenidos a partir de FragB son 255 (A), 331 (B), 372 (C) y 605 (D) bases.

Figura 14. (A) muestra los fragmentos de ADN amplificados utilizando los cebadores específicos de exón o-1165 y o-1166 utilizando ADN genómico extraído a partir del grupo 4 original de células 293-EBNA transfectadas con pEnd-HR#1 (1), a partir de grupos de clones aislados adicionalmente a partir de este último grupo (2-4) o a partir de células 293-EBNA no transfectadas (5). (B) fragmentos de ADN amplificados utilizando los cebadores específicos de intrón o-1121 y o-1168 utilizando ADN genómico extraído a partir de diferentes grupos de clones aislados a partir del grupo 4 original de células 293-EBNA transfectadas con pEnd-HR#1 (2-4) o a partir de células 293-EBNA no transfectadas (5). Los grupos de clones de los carriles 2-4 todavía expresan el transcrito esperado.

Figura 15. (A) fragmentos de ADN amplificados utilizando ADN genómico de células 293-EBNA y cebadores específicos de secuencias presentes en el intrón 37 (1) o ADN genómico del grupo de clones positivos para pEnd-HR#1, identificados previamente por RT-PCR y cebador que hibrida en el exón de mIgSP y el exón 38 (2), junto con el patrón obtenido de cada uno de los dos fragmentos con una serie de enzimas de restricción. (B) Tabla que resume la longitud esperada de los fragmentos de ADN.

La invención se describirá ahora con referencia a los ejemplos siguientes, que no deben interpretarse como limitativos de la invención de ninguna manera.

Ejemplos

TRANCE soluble

El gen TRANCE de ratón contiene 5 exones, el primero codifica esencialmente el dominio intracelular y transmembrana de la proteína, mientras que la mayor parte del dominio extracelular está codificado por los 4 exones restantes. En particular, el segmento que codifica específicamente el dominio proteico funcional *in vivo* (TRANCE soluble o sTRANCE) está codificado completamente por los exones 3º, 4º y 5º (Lum L. *et al.*, J Biol Chem. (1999), 274(19): 13613-8; Kodaira K. *et al.*, Gene (1999), 230(1), 121-127). La estructura del gen humano correspondiente no se conoce todavía pero un segmento genómico humano asociado al cromosoma 13 humano (registro de GenBank NT_009935) contiene la secuencia codificadora de la proteína TRANCE humana dividida en dos segmentos que son muy similares a los exones del gen TRANCE de ratón en términos de secuencia y longitud. La longitud de las secuencias intrónicas también parece similar en los dos genes (Figura 2).

Recientemente, se han caracterizado formas de sTRANCE que tienen secuencias N-terminales ligeramente diferentes (Schlondorff J *et al.*, J Biol Chem. (2001), 276(18): 14665-74), lo que sugiere que un dominio proteico funcional puede reducirse a la secuencia codificadora de los exones 4-5. También es interesante indicar la baja homología entre TRANCE de ratón y humano en el área que rodea la secuencia N-terminal posible de sTRANCE.

Si se desea la producción de un dominio proteico funcional correspondiente a TRANCE soluble, puede modificarse una célula anfitriona de ratón o humana, por ejemplo, utilizando dos segmentos contiguos o no contiguos al intrón 2 (20 Kb de longitud) como secuencias de direccionamiento para la recombinación homóloga, que expresan los exones 3-5. La unidad reguladora contendrá un módulo de transcripción, que contiene secuencias promotoras y de amplificación activas en células humanas y un módulo de traducción que contiene un exón sintético, con 5'UTR conveniente, un codón Met y un sitio donador de corte y empalme en 5'.

Alternativamente, a la vista de la homología entre la secuencia de ADNc y de proteína de TRANCE humano y de ratón y de las formas recientemente identificadas de sTRANCE, la secuencia correspondiente al intrón 3 también puede dirigirse a células de ratón o humanas, que sólo expresa los exones 4 y 5. La construcción puede tener una unidad reguladora similar a la utilizada para expresar los exones 3-5 pero también puede utilizarse una construcción simplificada que sólo incluye un módulo de transcripción apropiado y una región 5' no traducida como módulo de transcripción, sin un módulo de corte y empalme, ya que hay una Metionina conservada justo en el comienzo del exón 4 tanto en la secuencia genómica de TRANCE humana como de ratón.

El método de la invención es útil para producir el dominio extracelular de otras proteínas que pertenecen a la familia del TNF como CD40L, CD70, FasL, que tienen una estructura génica similar (Kodaira K. *et al.*, Gene (1999), 230(1), 121-127; Locksley RM *et al.*, Cell. (2001), 104(4): 487-501).

Factores antiangiogénicos derivados del Colágeno XVIII 1.alfa (Endostatina)

a) La estrategia de direccionamiento

La endostatina pertenece a un número creciente de dominios proteicos funcionales relacionados con la angiogénesis que se generan *in vivo* como fragmentos proteolíticos de un producto primario de la traducción secretado que no tiene ninguna actividad relacionada con la angiogénesis. Como se ha revisado recientemente (Cao Y, Int. J. Biochem. Cell Biol. (2001), 33(4): 357-69), los inhibidores conocidos de la angiogénesis como PEX, endostatina o restina son los fragmentos C-terminales de, respectivamente, MMP-2, colágeno XVIII 1 alfa y colágeno XV, mientras que Fn-f y vasostatina son los fragmentos N-terminales de, respectivamente, fibronectina y calreticulina.

En particular, muchas proteínas que pertenecen a la familia de las proteínas de colágeno y que actúan como elementos estructurales de la matriz extracelular son modificadas proteolíticamente para obtener factores angiostáticos, de los que el más estudiado se conoce como endostatina. Este dominio proteico funcional está codificado por el gen del colágeno XVIII 1 alfa (COL18A1) y, como un fragmento altamente similar denominado restina codificado por el gen del colágeno XV (COL15), representa el extremo C-terminal del dominio no colagenoso (NC1) del producto primario de la traducción (John H *et al.*, Biochemistry (1999), 38(2): 10217-24; Sasaki T *et al.*, J Mol Biol. (2000), 301(5): 1179-90).

El gen COL18A1 humano contiene 41 exones, mientras que el gen COL18A1 de ratón contiene 43 exones pero el dominio NC1 está codificado por los últimos seis exones en ambos organismos. Puede realizarse una distinción más entre los exones del gen COL18A1 humano que codifican el dominio NC1 y los asociados con el dominio de multimerización (exones 36-37), una región bisagra (exón 38) y un dominio central de endostatina (exones 39-41).

La endostatina se caracterizó inicialmente en ratón como un fragmento que contiene 183 aminoácidos correspondientes a los últimos 9 aminoácidos codificados por el exón 40 más los aminoácidos codificados por los exones 41-43 (O'Reilly MS *et al.*, Cell (1997) 88(2): 277-285). Sin embargo, en muestras humanas no se ha encontrado un

fragmento correspondiente que contenga los últimos 9 aminoácidos codificados por el exón 38 más los aminoácidos codificados por los exones 39-41. Muchos estudios han mostrado que la región bisagra es particularmente sensible a varias proteasas (Felbor U *et al.*, EMBO J. (2000), 19(6): 1187-94; Ferreras M *et al.*, FEBS Lett. (2000), 486(3): 247-51; John H *et al.*, Biochemistry (1999), 38(32): 10217-24; Wen W *et al.*, Cancer Res. (1999), 59(24): 6052-6056), lo que da lugar a una serie de fragmentos que tienen una secuencia N-terminal diferente codificada por el exón 38 (Figura 3). La bibliografía también muestra que los primeros aminoácidos codificados por el exón 39 están estructurados (Hohenester E *et al.*, EMBO J. (1998), 17(6): 1656-64) y que las proteínas que tienen un extremo N-terminal al menos 4 aminoácidos más largo o más corto comparado con el extremo N-terminal de la proteína codificada por los exones 39-41 son, respectivamente, activas e inactivas (Yamaguchi *et al.*, EMBO J. (1999), 18(16): 4414-4423; Standker L *et al.*, FEBS Lett. (1997), 420(2-3): 129-133). Finalmente, los fragmentos aislados o quiméricos derivados a partir de diferentes secuencias codificadas por los exones 39-41 tienen propiedades variables relacionadas con la migración y la proliferación celulares (WO 00/63249, WO 00/67771).

Puede concluirse que los exones 39-41 de COL18A1 humano codifican una unidad de plegamiento autónoma de la proteína del colágeno humano XVIII 1alfa correspondiente al dominio proteico funcional central *bona fide* proporcionando las propiedades angiostáticas de la endostatina natural. Además, la bibliografía también muestra que no se espera que la adición opcional de secuencias que pertenecen a los exones 36-38 o incluso secuencias heterólogas en el extremo N-terminal de dicho dominio proteico funcional, si tienen una longitud limitada (Yamaguchi N *et al.*, EMBO J. (1999), 18(16): 4414-4423; Blezinger P *et al.*, Nat. Biotechnol. (1999), 17(4): 343-348), interfiera con las propiedades angiostáticas asociadas con dichos dominios proteicos funcionales semejantes a la endostatina.

La secuencia del ADN genómico del gen COL18A1 humano, que está localizada en el cromosoma 21 (Hattori M *et al.*, Nature (2000), 405(6784): 311-9), está incluida en un clon genómico de 340 Kilobases accesible a través de GenBank (número de registro AL163302) y los cebadores que permiten la amplificación específica de los segmentos incluidos en este clon pueden diseñarse fácilmente. Por lo tanto, los dominios proteicos funcionales que tienen las propiedades angiostáticas de la endostatina humana pueden producirse modificando células humanas, según la invención, utilizando un vector que contiene una unidad reguladora y secuencias de direccionamiento que permiten la integración de la unidad reguladora a nivel bien del intrón 37 o del intrón 38 del gen COL18A1 humano por recombinación homóloga.

Se ensamblaron dos construcciones diferentes (pEnd-HR#1 y pEnd-HR#2) utilizando la misma unidad reguladora y diferentes secuencias de direccionamiento pertenecientes al gen COL18A1 humano (Figura 4A). La construcción pEnd-HR#1 permite el reemplazamiento, por recombinación homóloga, del extremo 3' del intrón 36, exón 37 completo y extremo 5' del intrón 37 por una unidad reguladora que promueve la expresión de los exones 38-41. La construcción pEnd-HR#2 permite el reemplazamiento, por recombinación homóloga, del extremo 3' del intrón 37 y del exón 38 completo por una unidad reguladora que promueve la expresión de los exones 39-41.

La clonación de los segmentos de ADN, la construcción de los plásmidos y la transfección, así como la selección y análisis de las células se realizaron utilizando técnicas estándar descritas en la bibliografía (Ausubel FM *et al.*, "Current Protocols in Molecular Biology" pub. John Wiley & Sons Inc., 1999; Sambrook *et al.*, "Molecular Cloning: A Laboratory Manual", Cold Spring Harbor Press, 1989; Hasty P *et al.*, en "Gene targeting: a practical approach", ed. Joyner AL, pub. Oxford Univ. Press (1999), 1-35). Todos los plásmidos se mantuvieron y propagaron antes de la transfección en células humanas utilizando la cepa de *E. coli* habitual DH5alfa o XL1 blue.

b) Construcción de los vectores de direccionamiento

La clonación de los fragmentos genómicos de COL18A1 necesarios para generar el vector de direccionamiento se realizó amplificando, utilizando PCR, las regiones de homología apropiadas del clon de GenBank AL163302, en particular, en el segmento de 9,7 Kilobases comprendido entre el exón 32 y la región 3' no traducida del exón 41 (Figura 4A).

El cebador o-1124 de 37 bases de longitud (SEQ ID NO: 5) contiene, en el extremo 5', una secuencia de 10 bases de longitud que incluye un sitio de restricción *Sal* I, mientras que las 27 bases en el extremo 3' corresponden a los nucleótidos 202790-202816 del clon AL163302. Esta última secuencia permite la hibridación de o-1124 en el exón 32 del gen COL18A1 humano y se utilizó como cebador directo para amplificar la región de direccionamiento 5' de ambas construcciones.

El cebador o-1125 de 36 bases de longitud (SEQ ID NO: 6) contiene, en el extremo 5', una secuencia de 9 bases de longitud que incluye un sitio de restricción *Bam* HI, mientras que las 27 bases en el extremo 3' son complementarias a los nucleótidos 206301-206327 del clon AL163302. Esta última secuencia permite la hibridación de o-1125 en el intrón 36 del gen COL18A1 humano y se utilizó como cebador inverso para amplificar la región de direccionamiento 5' de la construcción pEnd-HR#1.

El cebador o-1121 de 35 bases de longitud (SEQ ID NO: 7) contiene, en el extremo 5', una secuencia de 9 bases de longitud que incluye un sitio de restricción *Bam* HI, mientras que las 26 bases en el extremo 3' son complementarias a los nucleótidos 208099-208125 del clon AL163302. Esta última secuencia permite la hibridación de o-1121 en el intrón 37 del gen COL18A1 humano y se utilizó como cebador inverso para amplificar la región de direccionamiento 5' de la construcción pEnd-HR#2.

ES 2 305 091 T3

El cebador o-1116 de 35 bases de longitud (SEQ ID NO: 8) contiene, en el extremo 5', una secuencia de 10 bases de longitud que incluye un sitio de restricción *Xba*I, mientras que las 25 bases en el extremo 3' corresponden a los nucleótidos 206382-206406 del clon AL163302. Esta última secuencia permite la hibridación de o-1116 en el intrón 37 del gen COL18A1 humano y se utilizó como cebador directo para amplificar la región de direccionamiento 3' de la construcción pEnd-HR#1.

El cebador o-1117 de 40 bases de longitud (SEQ ID NO: 9) contiene, en el extremo 5', una secuencia de 16 bases de longitud que incluye un sitio de restricción *Not*I, mientras que las 24 bases en el extremo 3' son complementarias a los nucleótidos 208098-208121 del clon AL163302. Esta última secuencia permite la hibridación de o-1117 en el intrón 37 del gen COL18A1 humano y se utilizó como cebador inverso para amplificar la región de direccionamiento 3' de la construcción pEnd-HR#1.

El cebador o-1126 de 34 bases de longitud (SEQ ID NO: 10) contiene, en el extremo 5', una secuencia de 9 bases de longitud que incluye un sitio de restricción *Xba*I, mientras que las 25 bases en el extremo 3' corresponden a los nucleótidos 208381-208405 del clon AL163302. Esta última secuencia permite la hibridación de o-1126 en el intrón 38 del gen COL18A1 humano y se utilizó como cebador directo para amplificar la región de direccionamiento 3' de la construcción pEnd-HR#2.

El cebador o-1123 de 43 bases de longitud (SEQ ID NO: 11) contiene, en el extremo 5', una secuencia de 17 bases de longitud que incluye un sitio de restricción *Not*I, mientras que las 26 bases en el extremo 3' son complementarias a los nucleótidos 209828-209853 del clon AL163302. Esta última secuencia permite la hibridación de o-1123 en el intrón 39 del gen COL18A1 humano y se utilizó como cebador inverso para amplificar la región de direccionamiento 3' de la construcción pEnd-HR#2.

Utilizando el clon genómico AL163302 como molde y o-1124 y o-1125 como cebadores, la PCR da lugar a la generación del fragmento de ADN de direccionamiento 5' de la construcción pEnd-HR#1. Este fragmento de 3,5 Kb incluye el extremo 3' del exón 32, los intrones 32-35 y los exones 33-36 completos, el extremo 5' del intrón 36 y los sitios de restricción únicos *Sal*I y *Bam*HI, respectivamente, en los extremos 5' y 3'.

Utilizando el clon genómico AL163302 como molde y o-1124 y o-1121 como cebadores, la PCR da lugar a la generación del fragmento de direccionamiento 5' de la construcción pEnd-HR#2. Este fragmento de 5.354 pares de bases incluye el extremo 3' del exón 32, los intrones 32-36 y los exones 33-37 completos, el extremo 5' del intrón 37, con los sitios de restricción únicos *Sal*I y *Bam*HI, respectivamente, en los extremos 5' y 3'.

Utilizando el clon genómico AL163302 como molde y o-1116 y o-1117 como cebadores, la PCR da lugar a la generación del fragmento de ADN de direccionamiento 3' de la construcción pEnd-HR#1. Este fragmento de 1,7 Kb incluye una región central del intrón 37 y los sitios de restricción únicos *Xba*I y *Not*I, respectivamente, en los extremos 5' y 3'.

Utilizando el clon genómico AL163302 como molde y o-1126 y o-1123 como cebadores, la PCR da lugar a la generación del fragmento de direccionamiento 3' de la construcción pEnd-HR#2. Este fragmento de 1,5 Kb incluye el intrón 38 y el exón 39 completos y el extremo 5' del intrón 39 con los sitios de restricción únicos *Xba*I y *Not*I, respectivamente, en los extremos 5' y 3'.

Como los productos de PCR son particularmente largos, se prefieren enzimas y procedimientos específicos conocidos en la técnica para amplificar el fragmento necesario para generar los vectores de direccionamiento. Los kits para una PCR de alta fidelidad y de amplio intervalo están disponibles comercialmente, como el kit de PCR Herculase (Stratagene).

Cada región de homología 5' y 3' amplificada se clonó, respectivamente, entre los sitios *Bam*HI y *Sal*I, y los sitios *Xba*I y *Not*I de un plásmido pBluescript-KSII (pBS-KSII; Stratagene). Los fragmentos genómicos amplificados y clonados se analizaron entonces por mapeo de restricción y secuenciación parcial de ADN para confirmar su identidad.

La construcción de la unidad reguladora se realizó ensamblando secuencias de ADN conocidas en la bibliografía (Figura 4B). El módulo de transcripción se eligió entre los promotores humanos o no humanos activos constitutivamente a niveles altos en células humanas cultivadas. Un ejemplo es el promotor del gen del factor de elongación-1 alfa humano (EF-1) que se ha clonado (Uetsuki T, *et al.*, J Biol Chem. (1989), 264(10): 5791-8) que se ha demostrado que es muy eficaz en un intervalo amplio de células anfitrionas (Mizushima S y Nagata S, Nucleic Acids Res. (1990), 18(17): 5322). Este promotor de 1,2 Kb, presente en muchos plásmidos disponibles comercialmente (InVitrogen), contiene una caja TATA seguida de un sitio de comienzo de la transcripción, que comienza la transcripción de un exón no traducido de 22 bases de longitud y de un intrón de 0,9 Kb que tiene un efecto amplificador de la transcripción ya que contiene varias secuencias Sp1 y Ap1.

El módulo de traducción que contiene un sitio de inicio de la traducción se combinó con un módulo de corte y empalme con una secuencia consenso para un sitio donador de corte y empalme. Además, con el fin de facilitar el aislamiento del dominio proteico funcional localizado en el extremo 3' de un gen, se incluyó una secuencia que codifica un péptido señal en marco con los exones que codifican el dominio proteico funcional entre el sitio de inicio de la traducción y la secuencia consenso de corte y empalme. Como péptido señal se eligió la secuencia del Péptido

Señal de Ig de ratón (mIgSP) (número de registro de GenBank M13329), mientras que el consenso apropiado de corte y empalme se eligió entre los que se han mostrado recientemente como funcionales en células humanas (Long M *et al.*, Proc Natl Acad Sci USA (1998), 95(1): 219-223; Blencowe BJ, Trends Biochem. Sci. (2000), 25(3): 106-110).

Un fragmento del promotor de EF-1 alfa humano (correspondiente a los nucleótidos 373-1.561 en el registro J04617 de GenBank) se clonó entre el *ClaI* y *NheI* de pBS-KSII generando el vector pBS-EF1 alfa. El exón que codifica mIgSP y el sitio donador de corte y empalme se combinaron en un segmento de ADN sintético de 0,2 Kb que contiene, en el extremo 5', un sitio de restricción *XbaI* y, en el extremo 3', sitios de restricción *NheI* y *NotI*. Este fragmento se clonó en pBS-EF1alfa entre los sitios *XbaI* y *NotI*, tal que está localizado en 3' respecto al promotor EF1alfa humano, generando el plásmido pBS-EF1alfa-mIgSP-SD (Figura 5).

El núcleo del vector de recombinación homóloga se construyó empezando desde el plásmido pGEM-3Z (Promega), al que se añadieron como sigue marcadores de selección positiva y negativa, las regiones de direccionamiento y los módulos de transcripción, traducción y corte y empalme.

El gen para la selección negativa para recombinación homóloga fue Timidina Quinasa de HSV-1 (HSV-TK; un fragmento de 1,8 Kb de longitud derivado del genoma completo del herpesvirus 1 humano, depositado en GenBank como NC_001806), bajo el control del promotor de la fosfoglicerato quinasa (mPGK) activo ubicuamente y la señal de poliadenilación (respectivamente, fragmentos de 508 y 480 bases de longitud derivados del plásmido depositado en GenBank como X76683).

El sitio de clonación múltiple en 3' respecto a HSV-TK se modificó para permitir la clonación de todos los demás elementos. Se introdujeron dos nuevos sitios de restricción únicos *NotI* y *ClaI* por clonación de dos oligonucleótidos hibridados, mientras que se eliminaron dos sitios *XbaI* cercanos al promotor mPGK y un sitio de poliadenilación por digestión secuencial y religación, generando el plásmido pGEM-3Z-mPGK-TK-HR (Figura 6). Este último plásmido se utilizó para clonar, para cada construcción, primero la región de direccionamiento 5', después el marcador de selección positiva y finalmente la región de direccionamiento 3' junto con el módulo de transcripción/traducción.

La región de direccionamiento 5' de cada construcción se cortó del vector pBS-KSII utilizando los sitios de restricción *Bam* HI y *Sal* I y se subclonaron en pGEM-3Z-PGK-TK-HR entre los sitios *Bam* HI y *Xho* I, resultando en los plásmidos pGEM-3Z-PGK-TK-5'HR#1 y pGEM-3Z-mPGK-TK-5'HR#2. El gen de resistencia a la higromicina se eligió como el gen de selección positiva y se obtuvo del plásmido pHyEGFP (Clontech), en el que el gen de resistencia se expresa bajo el control de un promotor viral (CMV) como una proteína de fusión con la Proteína Verde Fluorescente (GFP). El plásmido comercial se modificó para eliminar los dos sitios *NotI* adyacentes cortando con *NotI* y rellenando con la enzima Klenow (Life Technologies). El casete CMV-HyEGFP-poliA se clonó como un fragmento *Cla* I-*Bgl* II en los plásmidos pGEM-3Z-PGK-TK-5'HR#1 y pGEM-3Z-mPGK-TK-5'HR#2 utilizando los sitios de restricción *ClaI* y *Bam*HI, generando los plásmidos pGEM-3Z-PGK-TK-HYG-5'HR#1 y pGEM-3Z-mPGK-TK-HYG-5'HR#2.

La región de direccionamiento 3' de cada construcción se cortó del vector pBS-KSII como un fragmento *XbaI*-*NotI* y se clonó en el plásmido pBS-EF1alfa-mIgSP-SD, utilizando los sitios *Nhe* I y *Not* I después del extremo 3' del exón que codifica el péptido señal mIgSP seguido del sitio donador de corte y empalme consenso. Los plásmidos resultantes pBS-EF1alfa-mIgSP-SD-3'HR#1 y pBS-EF1alfa-mIgSP-SD-3'HR#2 contienen el segmento EF1-mIgSP fusionado con la región de direccionamiento 3' entre los sitios *ClaI* y *NotI*. Dicho fragmento *ClaI*-*NotI* se introdujo finalmente entre los sitios *Cla* I y *Not* I de pGEM-3Z-PGK-TK-HYG-5'HR#1 y pGEM-3Z-PGK-TK-HYG-5'HR#2, después del extremo 3' del casete de selección positiva, obteniendo el vector final pEnd-HR#1 (Figura 7) y pEnd-HR#2. Estos plásmidos pueden linearizarse en el sitio único *NotI* localizado en el extremo 3' de la región de direccionamiento 3' antes de transfectarse en células para dirigir las secuencias exógenas a las localizaciones específicas en el gen COL18A1 humano. Una vez integrada, se espera que la unidad reguladora dirija la transcripción de un ARNm que codifica mIgSP fusionado en marco con los exones 38-41 de COL18A1 para pEnd-HR#1 (SEQ ID NO: 12; SEQ ID NO: 13; Figura 8) o los exones 39-41 de COL18A1 para pEnd-HR#2 (SEQ ID NO: 14; SEQ ID NO: 15; Figura 9).

La actividad de la unidad reguladora exógena incluida en pEnd-HR#1 y pEnd-HR#2 se ensayó de manera preliminar transfectando temporalmente plásmidos pEAK que contienen la secuencia codificadora de endostatina (Edge BioSystems). Las regiones reguladoras se clonaron en 5' respecto a un fragmento de ADN que contiene la secuencia intrónica que se transcribirá y se cortará y empalmará después de la recombinación homóloga en el gen COL18A1 humano, las secuencias codificadoras en los exones después del extremo 3' del gen COL18A1 humano fusionadas en marco con un epítipo heterólogo que facilita la identificación de la proteína incluso en pequeñas cantidades, un codón de parada y un sitio de poliadenilación presente en el vector (Figura 10A). Las dos construcciones (pEAK-pEnd-HR#1 y pEAK-pEnd-HR#2) se transfectaron en células humanas 293-EBNA (utilizadas posteriormente para la recombinación homóloga), que no expresan el gen COL18A1 (Yamaguchi N *et al.*, EMBO J. (1999), 18(16): 4414-4423).

El ARNm, así como las proteínas secretadas e intracelulares, de las células transfectadas se ensayaron para verificar la transcripción, el corte y empalme y la traducción correctas de la construcción.

El análisis se realizó en primer lugar por RT-PCR, amplificando el ADNc obtenido a partir de las células transfectadas y después por transferencia Western utilizando anticuerpos disponibles comercialmente para la endostatina

(Chemicon Inc.) y para el epítipo FLAG (Amersham-Pharmacia). En particular, el análisis de proteínas muestra que el promotor EF-1 alfa es activo en 293-EBNA y, después de la transcripción y corte y empalme, el ARNm final se traduce en un dominio proteico funcional del tamaño esperado (Figura 10B). Esta proteína se secreta entonces en el medio de cultivo gracias al péptido señal exógeno (Figura 10C), como se muestra en la bibliografía con una construcción que
5 contiene sólo secuencias exónicas de COL18A1 (Blezyer P *et al.*, Nat. Biotechnol. (1999), 17(4): 343-348).

c) *Transfección de los vectores de direccionamiento y selección de clones*

Los vectores pEnd-HR#1 y pEnd-HR#2 pueden utilizarse para transfectar tipos celulares humanos en los que las
10 secuencias reguladoras exógenas son activas, o inducibles, independientemente de si el gen COL18A1 endógeno se expresa ya. Este gen está altamente expresado en los tejidos vasculares del hígado, corazón y riñón, así como en hepatocitos (Saarela J *et al.*, Am. J. Pathol. (1998), 153(2): 611-626). Por lo tanto, las líneas celulares derivadas de estos tipos celulares pueden utilizarse para aplicar los métodos de la invención aunque también pueden utilizarse otros
15 tipos celulares que no expresan COL18A1 ya que, incluso si la estructura de la cromatina en este locus puede reprimir opcionalmente la transcripción, las secuencias reguladoras suficientemente fuertes y activas ubicuamente o inducibles pueden superar dichas limitaciones. Consecuentemente, la elección del tipo celular puede extenderse a líneas celulares humanas immortalizadas que pueden transfectarse y expandirse fácilmente, como células HT1080, W138, HepG2 ó 293.

Como se ha mostrado anteriormente, las células embrionarias humanas derivadas de riñón 293-EBNA se toman
20 como un ejemplo de una línea celular humana immortalizada que puede modificarse eficazmente con el método de la invención para obtener la expresión selectiva de los exones en el gen COL18A1 humano que codifican un dominio proteico funcional antiangiogénico. Estas células, que están disponibles comercialmente (Invitrogen), que expresan el antígeno 1 nuclear del virus de Epstein-Barr (EBNA-1), pueden transfectarse eficazmente utilizando electroporación y
25 crecerse en Medio de Eagle modificado por Dulbecco (DMEM) estándar que contiene 10% de Suero Fetal de Ternera, 4,5 gramos/litro de glucosa y antibióticos (100 microgramos/mililitro de penicilina y estreptomycin; Gibco-BRL).

La transfección por electroporación se llevó a cabo utilizando un aparato Multiporator (Eppendorf), equipado
30 con electrodos a una distancia de 4 milímetros, en las condiciones recomendadas (60 microSiemens y 500 Voltios) y utilizando el tampón hipo-osmolar optimizado para las células de mamífero proporcionado por el fabricante. Se electroporaron ocho alicuotas de células 293-EBNA en la fase logarítmica de crecimiento (cada una con $2,5 \times 10^6$ células) con 12 microgramos de plásmido linearizado con *NotI*, bien pEnd-HR#1 o pEnd-HR#2, en 800 microlitros de tampón. Después del pulso, se plaquéó un total de 2×10^7 células transfectadas con el mismo vector linearizado en
35 cuatro placas de cultivo tisular con un diámetro de 150 milímetros (NUNC) previamente recubiertas con D-poliisina (SIGMA). Después de 72 horas, se inició la selección utilizando 250 microgramos/mililitro de higromicina (Life Technology), cambiando el medio cada dos días. Después de cuatro días más de selección, se aplicó la selección negativa para la recombinación homóloga incubando las células con un medio de cultivo tisular que contiene 10 micromolar de Ganciclovir (Cymevene; Roche) además de la higromicina.

Aproximadamente 25 días después de la transfección, las células seleccionadas se aislaron utilizando puntas de
40 pipeta bajo el microscopio como clones aislados. Se generó un grupo de clones para cada placa (aproximadamente 300 clones/grupo) y los cuatro grupos se mantuvieron bajo selección positiva hasta 4-5 semanas para expandirlos y obtener una cantidad de material suficiente para análisis adicionales antes de proceder a la selección de grupos más pequeños de clones.

d) *Identificación y análisis de los clones que expresan endostatina*

Pueden aplicarse muchos métodos a los clones que resultan de la selección positiva-negativa con el fin de identificar
50 las células en las que se expresa un dominio proteico funcional antiangiogénico como resultado de la integración, por recombinación homóloga, de la unidad reguladora exógena en el gen COL18A1 humano. Los experimentos siguientes se llevaron a cabo para confirmar la integración correcta de la unidad reguladora y la expresión específica de los exones que codifican el dominio proteico funcional antiangiogénico.

El análisis de la estructura del gen COL18A1 humano y de los transcritos en los clones seleccionados se llevó
55 a cabo en el ADN genómico y en el ARNm extraídos a partir de grupos positivos de células utilizando tecnologías descritas en la bibliografía conocidas por el experto en la técnica. Se prefirió la amplificación selectiva de segmentos de ADN utilizando cebadores de ADN (PCR) como un primer método ya que es más rápido y requiere menos material biológico, con el fin de identificar secuencias que estarían comprendidas en el ADN genómico y en el ARNm después del direccionamiento y corte y empalme correctos (Figura 11A-B). Algunos de los segmentos amplificados también
60 se clonaron y se sometieron a secuenciación de ADN para una confirmación adicional de su identidad de secuencia.

La PCR se realizó en un termociclador Applied Biosystem 9700 utilizando los kits comerciales (Qiagen) HotStar-TAQ PCR (cuando el material inicial es ADN genómico o ADNc) o OneStep HotStar RT-PCR (cuando el material
65 inicial es ARN total), esencialmente como describe el fabricante. El volumen final de las reacciones fue 25 ó 50 microlitros. Después de terminar la reacción de PCR, 10 ó 20 microlitros de la mezcla de reacción se corrieron en geles de agarosa y se cribaron para detectar la presencia de los fragmentos de PCR esperados que indicarían la integración apropiada de la secuencia exógena en la posición diana del gen COL18A1 o la presencia del ARNm sometido a corte y empalme esperado que codifica el dominio proteico funcional antiangiogénico.

ES 2 305 091 T3

Los grupos de clones transfectados con pEnd-HR#1 o pEnd-HR#2 se cribaron en primer lugar para detectar la presencia de los nuevos transcritos. El ARNm extraído a partir de los clones se transcribió de manera inversa y se amplificó por PCR (método en dos etapas) o se transcribió directamente de manera inversa y se amplificó en el mismo tubo (método en una etapa). Los cebadores se diseñaron para hibridar con el molde en diferentes secuencias exónicas, manteniendo el mismo cebador directo que hibrida en el exón mIgSP y utilizando diferentes cebadores inversos que hibridan con el exón humano endógeno de COL18A1 (Figura 11A).

Para el método en dos etapas, un microgramo de ARN total se transcribió de manera inversa con cebadores oligo-dT₁₈ utilizando el kit Superscript-II cDNA (Life technologies), como describe el fabricante, y 5 unidades de Transcriptasa Inversa MMLV (Promega Biotech). Después de una incubación de 45 minutos a 37°C, se añadió ARNasa H (1 unidad) para eliminar el ARN emparejado con el ADN extendido a partir de los oligonucleótidos y la incubación a 37°C se prolongó durante 15 minutos más. El ADN complementario (ADNc) resultante se diluyó finalmente hasta una concentración de 10 nanogramos/microlitro utilizando agua sin ARNasa. Después, se llevó a cabo una PCR utilizando 20 nanogramos de ADNc cebado con oligo-dT y 0,5 microMolar de cada cebador y aplicando el programa siguiente:

1x95°C durante 15 minutos;

5x de 95°C durante 45 segundos, 60°C a 56°C (descendiendo 1°C en cada ciclo), 72°C durante 1 minuto;

35x95°C durante 45 segundos, 54°C durante 30 segundos y 72°C durante 1 minuto;

1x72°C durante 10 minutos.

Para el método en una etapa, la RT-PCR OneStep HotStar se llevó a cabo utilizando 500 nanogramos de ARN total y 0,5 microMolar de cada cebador. La reacción de transcripción inversa se llevó a cabo a 50°C durante 30 minutos según las condiciones proporcionadas por el fabricante y después se aplicó el programa siguiente:

1x95°C durante 15 minutos;

35x95°C durante 30 segundos, 57°C durante 30 segundos, 72°C durante 1 minuto;

1x72°C durante 10 minutos.

Inicialmente, se utilizó el oligonucleótido que hibrida con la región codificadora del exón mIgSP exógeno (o-1165; SEQ ID NO: 16) junto con un cebador que hibrida en el exón 40 de COL18A1 endógeno (o-1175; SEQ ID NO: 17), que se debería expresar después de la integración de cualquiera de las dos construcciones en el gen COL18A1 humano. Mediante la amplificación del ADNc generado utilizando el ARNm extraído de los grupos de los clones seleccionados que fueron transfectados con pEnd-HR#1 o pEnd-HR#2 fue posible identificar un grupo de clones para cada construcción de direccionamiento (el grupo 4 contiene pEnd-HR#1 y el grupo 3 pEnd-HR#2) que expresa una molécula de ARNm que tiene tanto las secuencias exónicas exógenas como endógenas separadas por el número de nucleótidos esperado después del direccionamiento, transcripción y corte y empalme correctos (577 bases para pEnd-HR#1, 331 bases para pEnd-HR#2). La PCR fue negativa para las células no transfectadas (Figura 12).

En el grupo de clones pEnd-HR#1 positivos fue posible identificar una banda menor que se corresponde aparentemente con el tamaño esperado para pEnd-HR#2. Por lo tanto, se realizó un análisis por PCR detallado de cada una de las dos especies moleculares utilizando el cebador más externo o-1165 (SEQ ID NO: 16) y o-1131 (SEQ ID NO: 18) y el grupo 4 de clones pEnd-HR#1 positivos, se amplificaron dos bandas más grandes (FragA y FragB), que tienen la misma diferencia de longitud observada utilizando el cebador inverso más interno o-1175. Dichos fragmentos se aislaron del gel, se clonaron y se utilizaron independientemente como molde con o-1165 como cebador directo y otros oligonucleótidos como cebadores inversos anidados: o-1164 (SEQ ID NO: 19), o-1175 (SEQ ID NO: 17), o-1179 (SEQ ID NO: 20).

Si la banda menor fuera el resultado de un corte y empalme alternativo, los fragmentos amplificados obtenidos utilizando FragA y FragB como molde se diferenciarían en un segmento correspondiente a la longitud del exón ausente. Como todas las parejas de cebadores amplifican fragmentos que se diferencian en aproximadamente 250 pares de bases entre FragA y FragB, puede concluirse que la banda menor identificada originalmente en el grupo 4 corresponde realmente a un transcrito en el que el exón 38 (246 bases), comprendido en todos los fragmentos, se eliminó por corte y empalme debido a un evento de corte y empalme alternativo (Figura 13). Este corte y empalme irregular, que se confirmó posteriormente por la secuenciación de los fragmentos clonados, da lugar, sin embargo, a un transcrito idéntico al obtenido transfectando las células con pEnd-HR#2, ya que los exones 38 y 39 tienen el mismo marco (Figura 8).

Finalmente, se obtuvo por RT-PCR un fragmento de 0,9 Kb en el ARNm del grupo 4 de células transfectadas con pEnd-HR#1 utilizando el cebador específico de mIgSP o-1165 y el cebador o-1193 (SEQ ID NO: 21) que es específico para la región 5' no traducida del exón 41. Este fragmento se clonó y se secuenció. La secuencia correspondió al esperado (Figura 8), confirmando adicionalmente que la construcción de ADN se había integrado correctamente para dirigir la transcripción de los exones diana de COL18A1 específicos para endostatina.

ES 2 305 091 T3

Se realizó un análisis adicional a nivel genómico en el grupo de clones transfectados con pEnd-HR#1 y que expresan el transcrito correcto utilizando PCR con oligonucleótidos que hibridan con secuencias exónicas, como en el caso de o-1165 (SEQ ID NO: 16) y o-1166 (SEQ ID NO: 22) o intrónicas, como en el caso de o-1168 (SEQ ID NO: 23) y o-1121 (SEQ ID NO: 7) (Figura 11B).

La PCR se llevó a cabo utilizando 200 nanogramos de ADN genómico (aislado a partir de los grupos positivos de clones de pEnd-HR#1 identificados por RT-PCR o a partir de células 293-EBNA no transformadas) y 0,5 micromolar de cada cebador, aplicando el programa siguiente:

1x95°C durante 15 minutos;

5x95°C durante 30 segundos, 62°C-58°C durante 30 segundos (descendiendo 1°C en cada ciclo), 72°C durante 2 minutos 15 segundos;

25x95°C durante 30 segundos, 57°C durante 30 segundos, 72°C durante 2 minutos 15 segundos;

1x72°C durante 10 minutos.

Los cebadores que hibridan en las regiones exónicas (una insertada por recombinación homóloga, la otra endógena) fueron capaces de amplificar un fragmento de la longitud esperada (1.820 bases) en todos los grupos de clones generados a partir del grupo positivo original pero no en las células no transfectadas (Figura 14A) confirmando la integración del exón de mIgSP. Los cebadores que hibridan en las regiones intrónicas fueron capaces de amplificar un fragmento de la longitud esperada (1.784 bases) tanto en los clones transfectados como no transfectados, confirmando la integridad de la estructura génica (Figura 14B).

Se obtuvo una evidencia adicional de que la estructura génica es la esperada después de la integración de la unidad reguladora contenida en el vector de direccionamiento pEnd-HR#1 mediante la digestión de los fragmentos obtenidos utilizando los cebadores específicos de exones e intrones con una serie de enzimas de restricción para verificar que los subfragmentos resultantes tenían la longitud esperada. Todas las enzimas ensayadas proporcionaron el patrón de restricción esperado (Figura 15 A, B).

El análisis por PCR, realizado en el ARNm y el ADN genómico extraídos a partir de las células modificadas por los métodos de la invención, permitió la identificación de clones en los que la integración de una unidad reguladora en el gen COL18A1 humano causa la expresión específica de los exones que codifican el dominio proteico funcional que determina las propiedades angiostáticas de la endostatina. Por lo tanto, el análisis puede ir más allá mediante el aislamiento y la caracterización del clon que mejor expresa el dominio proteico funcional deseado. Este análisis adicional puede llevarse a cabo hibridando sondas específicas de endostatina con ARN total (transferencia Northern) o ADN genómico (transferencia Southern) obtenido a partir de estos clones después de haberlos expandido lo suficiente. Por ejemplo, la transferencia Southern debería permitir la identificación de una banda de 2,2 Kb ó 1,9 Kb en el ARN aislado a partir de clones transfectados con, respectivamente, pEnd-HR#1 o pEnd-HR#2. Si el ADN genómico de las células positivas y de las células no transfectadas se digiere con *NheI* y *SpeI*, se separa en un gel de agarosa, se transfiere a un filtro y se ensaya con un fragmento radiactivo correspondiente a la región genómica de COL18A1 humano que incluye los exones 32-36 y los intrones 32-36, el patrón de hibridación sería diferente ya que el fragmento de 12,4 Kb, que incluye COL18A1 humano desde el intrón 31 hasta el final, visible en las células no transfectadas, estaría reemplazado por un fragmento más corto en las células positivas (4,4 Kb en las células transfectadas con pEnd-HR#1) debido a los sitios adicionales *NheI* y *SpeI* en el gen para la selección positiva.

A nivel de proteínas, antes de una expansión, recogida y purificación adicionales de los dominios proteicos funcionales angiostáticos, el cribado puede realizarse utilizando tecnologías basadas en anticuerpos (ELISA, Transferencia Western, inmunoprecipitación) con el fin de identificar los clones que tienen los niveles más altos de producción.

La actividad antiangiogénica de los dominios proteicos funcionales purificados adicionalmente puede determinarse mediante uno de los muchos métodos descritos en la bibliografía basados en células endoteliales. Los extractos de proteínas, preparaciones purificadas o medios de cultivo obtenidos a partir de los clones positivos pueden ensayarse en un ensayo de migración de células endoteliales, utilizando endostatina recombinante o purificada humana como estándar. Uno de los ensayos más habituales emplea células endoteliales de la vena umbilical humana (HUVEC) que están disponibles comercialmente (Clonetics) y pueden cultivarse para construir un ensayo de migración fiable (Yamaguchi N *et al.*, EMBO J. (1999), 18(6): 4414-4423).

Trx 80

La tiorredoxina humana (Trx) es una enzima que cataliza reducciones disulfuro intracelulares. Una forma truncada de tiorredoxina (Trx80) que contiene los 80-84 restos N-terminales y que carece de actividad enzimática, es escindida y secretada por líneas celulares de monocitos y, en sí misma, es una citoquina mitogénica potente que estimula el crecimiento de células mononucleares de sangre periférica humana sin estimular (Pekkari K *et al.*, J. Biol. Chem. (2000), 275(48): 37474-80). Además, el medio de monocitos CD14(+) humanos purificado se activó específicamente hacia la diferenciación con Trx80 como se determina por la expresión incrementada de CD14, CD40, CD54, CD86. Trx80 también induce la secreción de IL-12 a partir de monocitos CD40(+) en cultivos de células mononucleares de

sangre periférica humana, un efecto incrementado por IL-2 que induce la secreción de interferón-gamma en cultivos de PBMC (Pekkari K *et al.*, Blood. 2001 Mayo 15; 97(10): 3184-90). Aunque la Trx puede tener alguna actividad citoquina con interleuquinas después de la secreción sin líder (Bertini R *et al.*, J Exp Med. (1999), 189(11): 1783-9), los efectos obtenidos con Trx80 no son reproducibles utilizando la proteína completa.

El gen de Trx humano (registros de GenBank X54539 y X54540) contiene cinco exones que codifican una proteína que contiene 105 restos (Kaghad M *et al.*, Gene (1994), 140(2): 273-8). Los 4 primeros exones, todos con marco cero, codifican un total de 85 restos y se corresponden esencialmente con Trx80. La secuencia de 1,3 Kb que contiene el intrón 4, los exones 4-5 y parte del intrón 3 (registro de GenBank X70288) puede utilizarse para construir un vector de direccionamiento que permita la integración de una unidad reguladora para terminar la transcripción y la traducción, junto con un sitio aceptor de corte y empalme, a nivel del intrón 4.

Citoquinas derivadas de la Tirosina-ARNt sintetasa humana

Las sintetasas de aminoacil-ARNt catalizan la aminoacilación de ARN de transferencia (ARNt). Aunque la tirosil-ARNt sintetasa humana es inactiva como molécula de señalización celular, puede secretarse y escindirse en dos citoquinas distintas en condiciones apoptóticas, probablemente por la leucocito elastasa, una proteasa extracelular (Wakasugi K y Schimmel P, Science (1999), 284(5411): 147-51). El fragmento N-terminal en el que se encuentra el sitio catalítico actúa como una citoquina semejante a la interleuquina-8. El dominio C-terminal es una citoquina semejante al polipéptido II activador de monocitos y endotelios (EMAP II), que tiene una actividad potente de quimiotaxis de leucocitos y monocitos y que estimula la producción de mieloperoxidasa y del factor de necrosis tumoral alfa.

El sitio de escisión posible de la proteína, que contiene 528 restos, está localizado en el resto 360 aunque los fragmentos obtenidos mediante la escisión en el resto 344 también son activos. Utilizando la secuencia codificadora conocida de la tirosil-ARNt sintetasa (registro de GenBank BC001933) para investigar el genoma humano, puede encontrarse un clon genómico que contiene secuencias interrumpidas correspondientes a la secuencia que codifica la tirosil-ARNt sintetasa (registro de Genbank AL356780). En particular, la secuencia que codifica los aminoácidos 303-348 y 349-380 corresponde, respectivamente, a las secuencias 98110-97970 y 96712-96615 del clon humano, que tiene una orientación de la numeración contraria.

En esta situación, el segmento del clon comprendido entre 98110 y 96615, así como otras secuencias circundantes disponibles en el clon, pueden utilizarse para dirigir una unidad reguladora iniciando o terminando la transcripción y la traducción entre 97970 y 96712, dependiendo de si se desea la expresión de la citoquina del extremo N-terminal (aminoácidos 1-348) o C-terminal (aminoácidos 349-528).

Sitio de unión a antígenos de una cadena pesada de inmunoglobulina

En algunas realizaciones, la unidad reguladora puede contener una secuencia capaz de terminar la transcripción y la traducción en una posición correspondiente al extremo 5' del dominio proteico funcional. Dicho método puede aplicarse siempre que el dominio proteico funcional de interés esté localizado en el primer o primeros exones del gen diana y que el gen diana se exprese a niveles altos constitutivamente o, después de una inducción, en la célula a la que se aplica el método.

Un ejemplo está representado por los sitios de unión de los anticuerpos, que están localizados en el extremo N-terminal de una molécula de inmunoglobulina. Los sitios de unión a antígenos de los anticuerpos convencionales están formados principalmente por los bucles hipervariables tanto de los dominios variables de las cadenas pesada como ligera. Sin embargo, los sitios de unión a antígenos funcionales también pueden estar formados por los dominios variables de la cadena pesada (VH) solos, como en los camellos y camélidos, en los que los anticuerpos contienen sólo dos dominios variables de cadena pesada y carecen de cadenas ligeras. Los análisis de las diferencias de la secuencia de aminoácidos entre los VH de estos anticuerpos de camello que tienen sólo cadenas pesadas y los dominios VH de los anticuerpos humanos convencionales ha ayudado en el diseño de un dominio VH humano alterado. Se ha visto que este VH camelizado es, como el VH de camello, una unidad de reconocimiento pequeña, robusta y eficaz formada por un único dominio de inmunoglobulina (Ig) (Riechmann L. *et al.*, J Immunol Methods. 1999 Dic 10; 231(1-2): 25-38; Davies J. *et al.*, Biotechnology (NY). 1995 Mayo; 13(5): 475-9).

El exón que codifica el dominio VH de una IgG resulta de la reorganización y mutación que se produce durante el desarrollo de la célula B. Cuando una célula de mieloma se fusiona con una célula B que codifica un anticuerpo que tiene una alta afinidad por un antígeno, las células de hibridoma resultantes transcriben y traducen activamente el gen completo de la IgG, aunque estas células también pueden integrar secuencias exógenas por recombinación homóloga con una eficacia alta (Schulman MJ *et al.*, Mol Cell Biol. (1990) 10(9): 4466-4472). Si se desea obtener sólo el dominio VH como un dominio proteico funcional, el gen de IgG puede modificarse utilizando el método de la invención integrando, por recombinación homóloga, una unidad reguladora que contiene los módulos de terminación de la transcripción y de la traducción en el intrón posterior al exón relevante.

REIVINDICACIONES

1. Un método para producir una proteína **caracterizado** porque dicha proteína es un fragmento de una proteína que es un producto primario de la traducción de un gen diana que es el extremo C o N-terminal de dicho producto primario de la traducción de un gen diana y en el que dicho fragmento ejerce una función biológica, comprendiendo el método:

(i) crecer una célula anfitriona transfectada con una construcción de ADN que comprende:

(a) una secuencia de ADN reguladora capaz de iniciar la transcripción y la traducción o de terminar la transcripción y la traducción del ADN que codifica dicho fragmento; y

(b) una región de ADN de direccionamiento que comprende secuencias homólogas a una región del gen diana en 5' ó 3', respectivamente, respecto a la secuencia que codifica dicho fragmento,

integrándose la construcción en el ADN genómico de la célula anfitriona en una posición determinada por el segmento de ADN de direccionamiento de manera que la expresión de dicho fragmento está bajo el control del ADN regulador; y, opcionalmente

(ii) cultivar la célula recombinante de manera homóloga; y, opcionalmente

(iii) recoger dicho fragmento;

en el que dicho fragmento se selecciona del grupo que consiste en:

(i) un dominio de unión extracelular de un receptor de membrana;

(ii) un fragmento proteolítico de una proteína estructural de la matriz extracelular que tiene propiedades antiangiogénicas; y

(iii) una proteína liberada proteolíticamente a partir de una proteína precursora inactiva.

2. El método de la reivindicación 1, en el que dicho fragmento está codificado por exones y en el que dicho segmento de ADN de direccionamiento comprende secuencias homólogas a una región genómica en 5' ó 3' respecto al exón o exones que codifican la proteína.

3. El método de cualquiera de las reivindicaciones 1 ó 2, en el que:

(i) dicho fragmento corresponde al extremo C-terminal de un producto primario de la traducción de un gen diana y está codificado por al menos el exón más próximo al extremo 3' de un gen diana que contiene intrones;

(ii) dicha región de ADN de direccionamiento comprende secuencias homólogas a una región genómica en 5' respecto al exón o exones que codifican el fragmento;

(iii) dicha construcción de ADN está unida operativamente a dicho exón o exones después de la integración en el genoma de la célula anfitriona por recombinación homóloga;

(iv) dicho ADN regulador comprende:

(a) un módulo de transcripción que consiste en una secuencia de ADN capaz de activar la transcripción de un ADN que codifica dicho fragmento; y

(b) un módulo de traducción que consiste en una secuencia de ADN capaz de iniciar la traducción de dicho fragmento; y

(v) dicha construcción de ADN comprende opcionalmente un módulo de corte y empalme que comprende un sitio donador de corte y empalme sin emparejar en 5' que, cuando se complementa con el sitio aceptor de corte y empalme sin emparejar en 3' del exón o exones endógenos que codifican el extremo N-terminal de dicho fragmento, permite el corte y empalme del transcrito primario, lo que resulta en la yuxtaposición "en marco" del módulo de traducción respecto a la secuencia que codifica dicho fragmento.

4. El método de la reivindicación 3 en el que el módulo de traducción comprende una región 5' no traducida y en el que el codón de inicio de la traducción está comprendido en el exón que codifica el extremo N-terminal de dicho fragmento.

5. El método de la reivindicación 3 en el que el módulo de traducción comprende uno o más exones exógenos que contienen una región 5' no traducida y un codón de inicio de la traducción.

6. El método de la reivindicación 5 en el que dicho exón o exones exógenos se ponen “en marco” con la secuencia que codifica dicho fragmento por el módulo de corte y empalme, convirtiéndose en el exón o exones N-terminales de dicho fragmento.

5 7. El método de la reivindicación 5 en el que dicho exón o exones exógenos se ponen “en marco” con la secuencia que codifica dicho fragmento eligiendo la secuencia de direccionamiento de manera tal que dicho exón o exones exógenos se fusionen con el exón que codifica el extremo N-terminal de dicho fragmento.

10 8. El método de cualquiera de las reivindicaciones 5 a 7 en el que dicho exón o exones exógenos codifican un péptido señal o el extremo N-terminal de un péptido señal.

9. El método de cualquiera de las reivindicaciones 1 ó 2, en el que:

15 (i) dicho fragmento corresponde al extremo N-terminal de un producto primario de la traducción de un gen diana y está codificado por al menos el exón más próximo al extremo 5' de un gen diana que contiene intrones;

(ii) dicha región de ADN de direccionamiento comprende secuencias homólogas a una región genómica en 3' respecto al exón o exones que codifican el fragmento;

20 (iii) dicha construcción de ADN está unida operativamente a dicho exón o exones después de la integración en el genoma de la célula anfitriona por recombinación homóloga;

(iv) dicho ADN regulador comprende:

25 (a) un módulo de transcripción que consiste en una secuencia de ADN capaz de terminar la transcripción del ADN genómico;

(b) un módulo de traducción que consiste en una secuencia de ADN capaz de terminar la traducción de dicha proteína; y

30 (v) dicha construcción de ADN comprende opcionalmente un módulo de corte y empalme que comprende un sitio aceptor de corte y empalme en 3' sin emparejar que, cuando se complementa con el sitio donador de corte y empalme en 5' sin emparejar del exón o exones endógenos que codifican el extremo C-terminal de dicho fragmento, permite el corte y empalme correcto del transcrito primario, lo que resulta en la yuxtaposición “en marco” del módulo de traducción respecto a la secuencia que codifica dicho fragmento.

10. El método de la reivindicación 9 en el que el módulo de traducción comprende un codón de terminación de la traducción y una región 3' no traducida.

40 11. El método de cualquiera de las reivindicaciones 9 ó 10 en el que el módulo de traducción comprende uno o más exones exógenos que contienen una región 3' no traducida y un codón de terminación de la traducción que se pone “en marco” con la secuencia que codifica dicho fragmento por el módulo de corte y empalme, convirtiéndose en el codón de terminación de la traducción de dicho fragmento.

45 12. Los métodos de cualquiera de las reivindicaciones 5, 6, 7, 8 u 11, en los que dicho exón o exones exógenos, cuando se ponen en marco con el exón o exones que codifican dicho fragmento, codifican una secuencia de proteína heteróloga a secuencias de proteínas comprendidas en el producto primario de la traducción codificado por el gen diana.

50 13. Los métodos de cualquiera de las reivindicaciones 5, 6, 7, 8 u 11, en los que dicho exón o exones exógenos, cuando se ponen en marco con el exón o exones que codifican dicho fragmento, codifican una secuencia de proteína homóloga a secuencias de proteínas comprendidas en el producto primario de la traducción codificado por el gen diana.

55 14. El método de cualquiera de las reivindicaciones 12 ó 13, en el que dicho exón o exones exógenos codifican adicionalmente el sitio de reconocimiento de una enzima proteolítica.

15. El método de la reivindicación 14 en el que dicha enzima proteolítica es expresada por la célula anfitriona.

60 16. El método de cualquiera de las reivindicaciones 14 ó 15 en el que dicha enzima proteolítica es una proteasa disponible comercialmente.

17. Un método según una cualquiera de las reivindicaciones 1 a 16 en el que la región de direccionamiento consiste en dos segmentos de direccionamiento.

65 18. Un método según una cualquiera de las reivindicaciones 1 a 17 en el que la construcción comprende además uno o más genes marcadores seleccionables y/o amplificables.

ES 2 305 091 T3

19. Un método según una cualquiera de las reivindicaciones anteriores en el que el producto primario de la traducción del gen diana que contiene el exón o exones que codifican el fragmento se selecciona del grupo que consiste en una hormona, una citoquina, una linfoquina, una quimioquina, una proteína unida a membrana, una proteína transmembrana, una proteína de la matriz extracelular, una proteína intracelular, una proteína nuclear, un factor regulador del crecimiento y metabolismo celular, una enzima, un receptor, un producto sanguíneo o un anticuerpo monoclonal.

20. Un método según la reivindicación 19 en el que el fragmento se selecciona del grupo que consiste en TRANCE soluble, Endostatina, Trx80, citoquinas derivadas de la tirosil-ARNt sintetasa o un sitio de unión a antígenos de una inmunoglobulina.

FIGURA 1

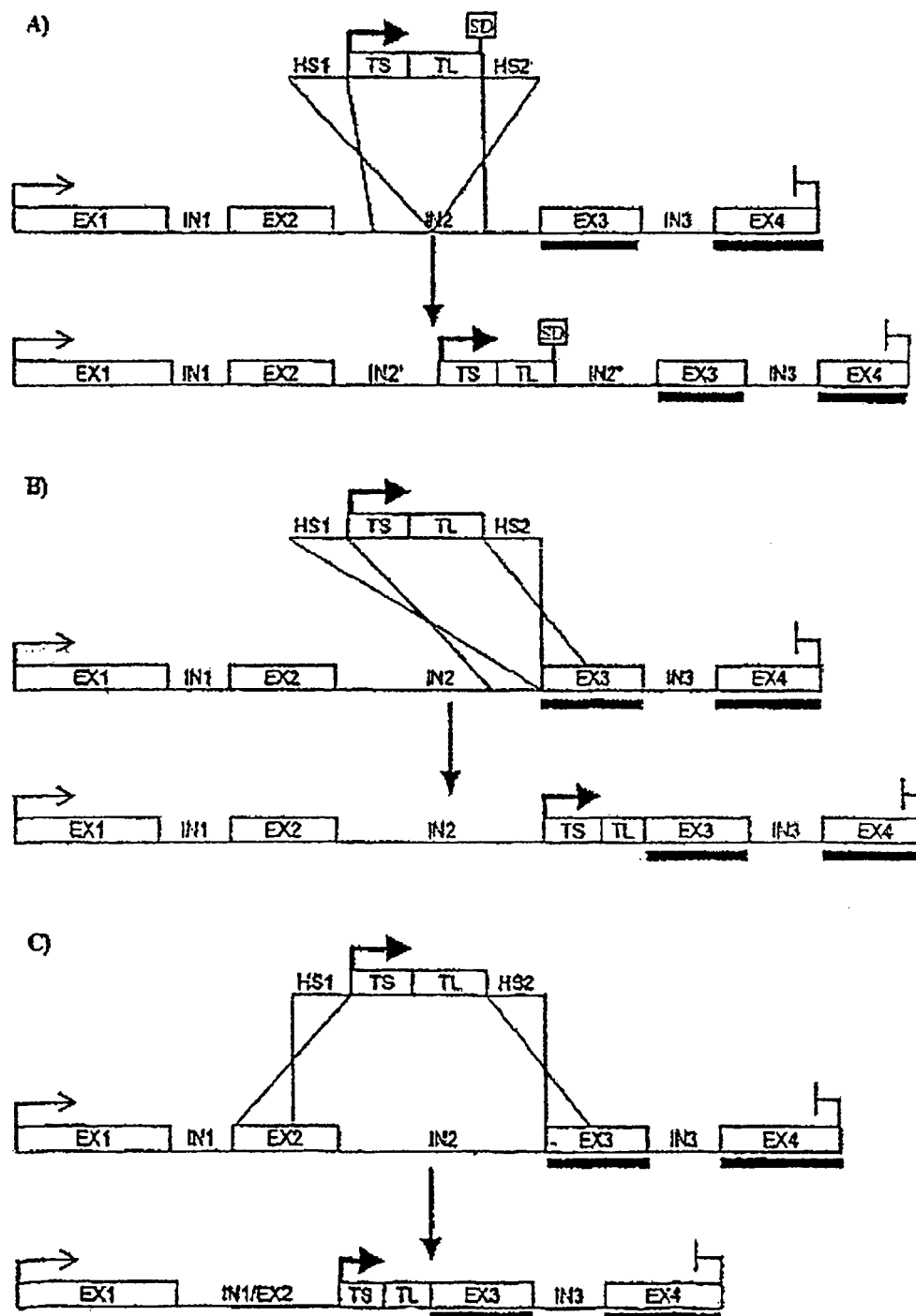
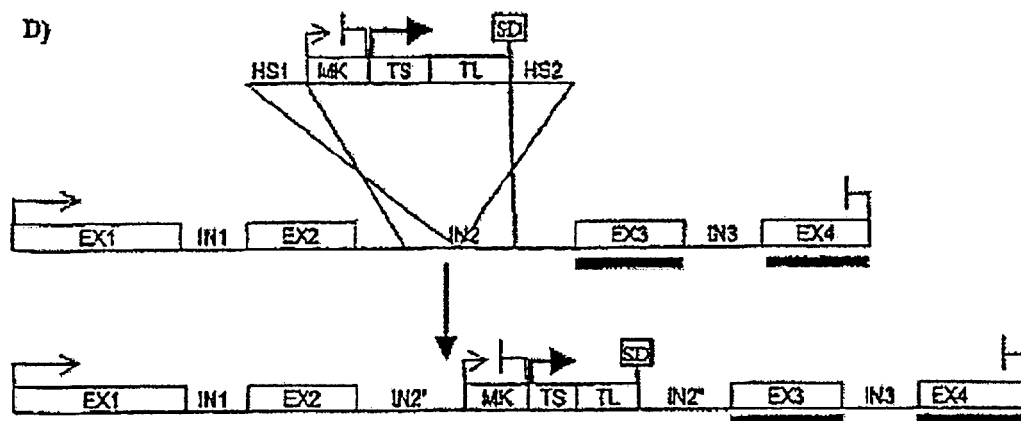
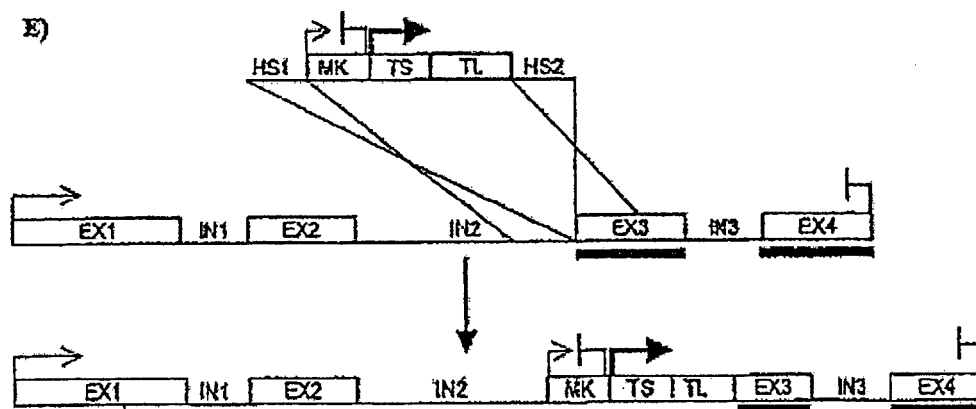


FIGURA 1

D)



E)



F)

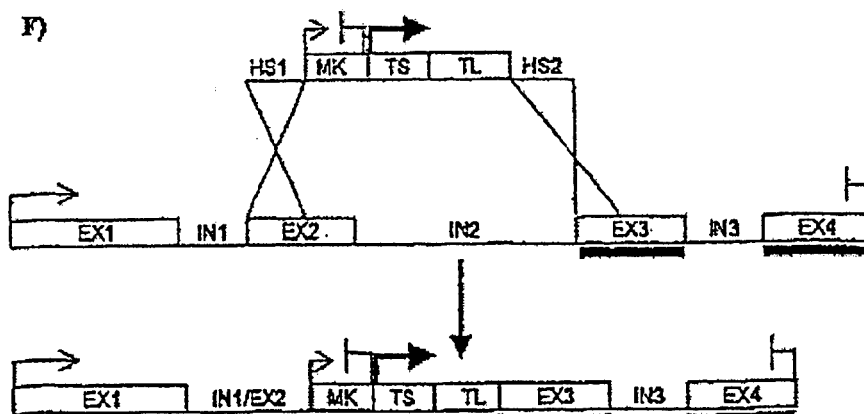
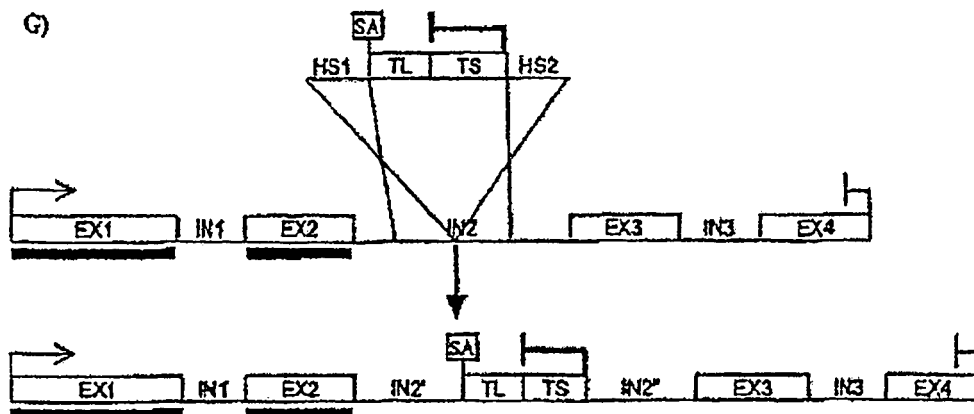
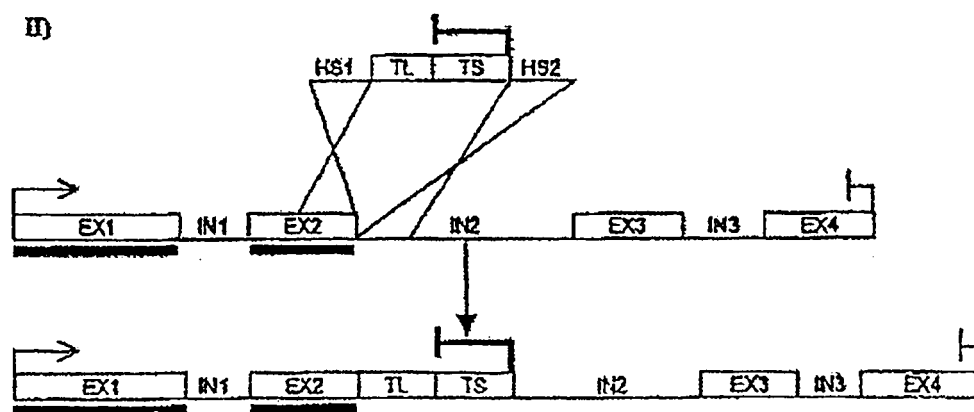


FIGURA 1

G)



II)



D)

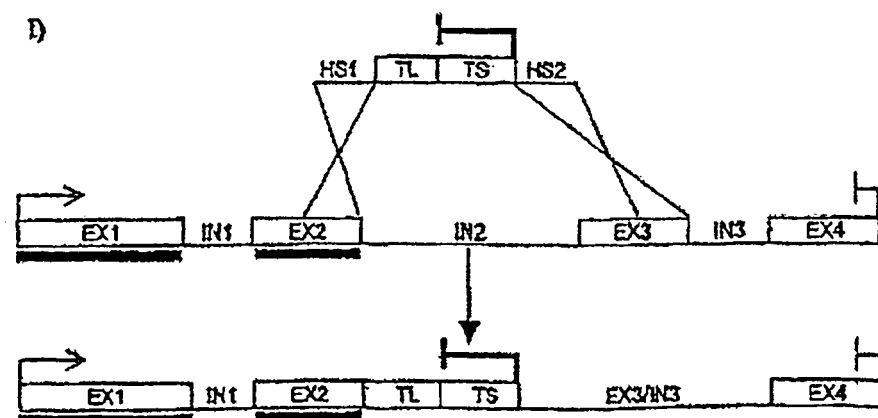


FIGURA 1

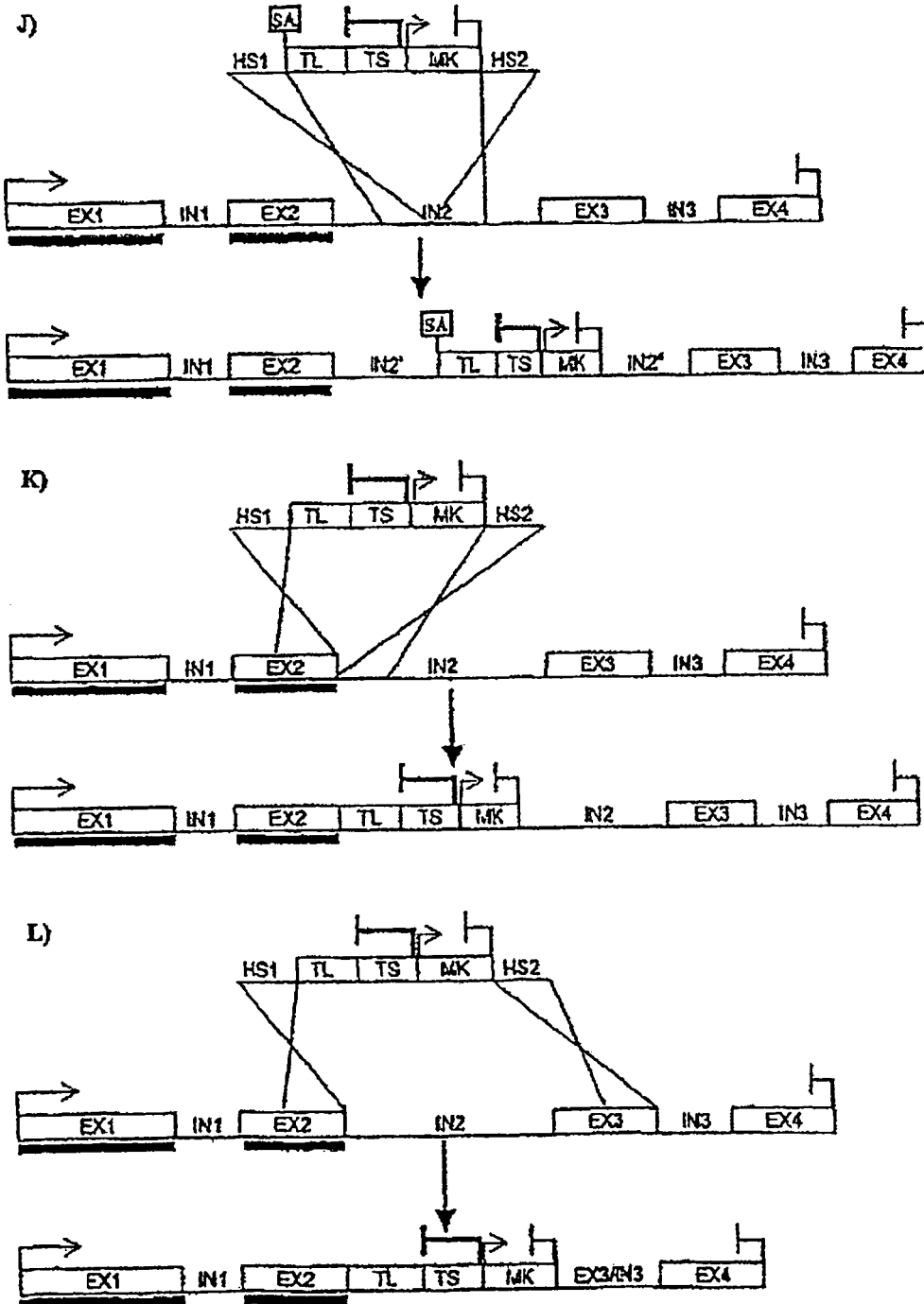


FIGURA 2

gen TRANCE de ratón

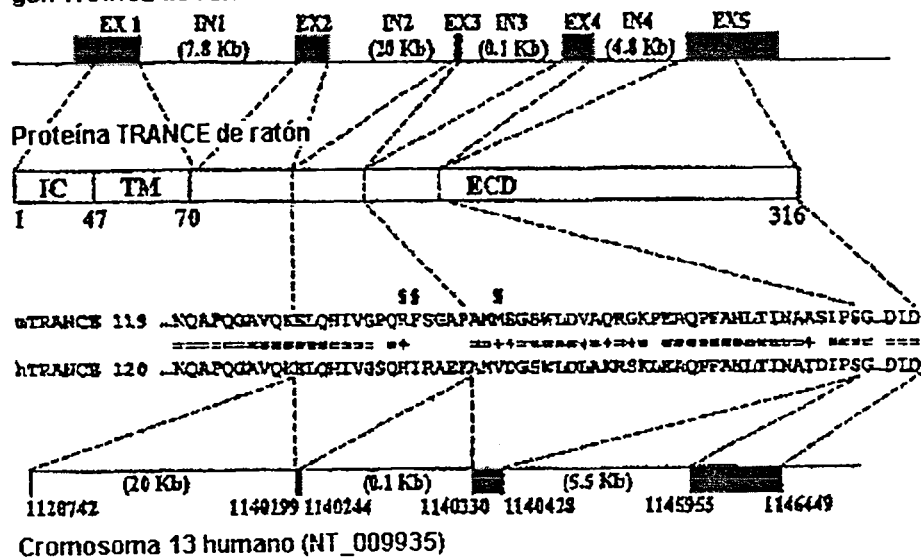


FIGURA 3

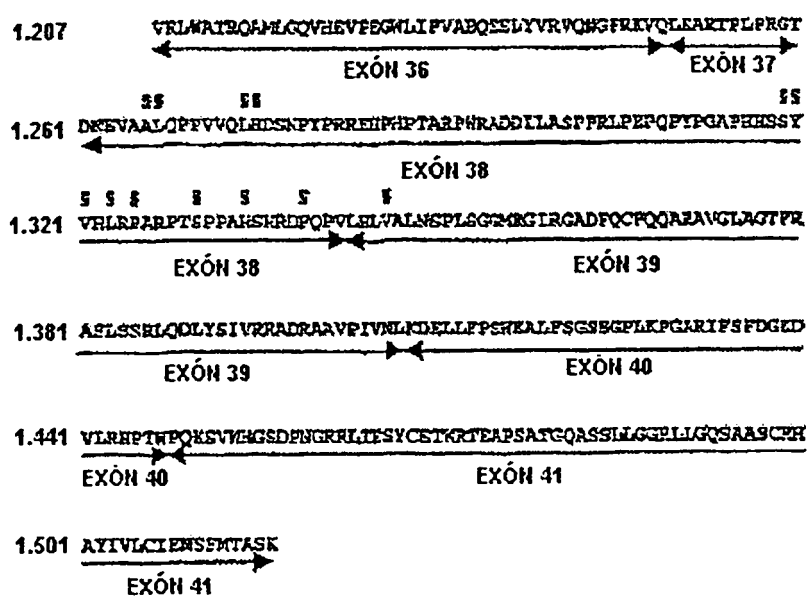
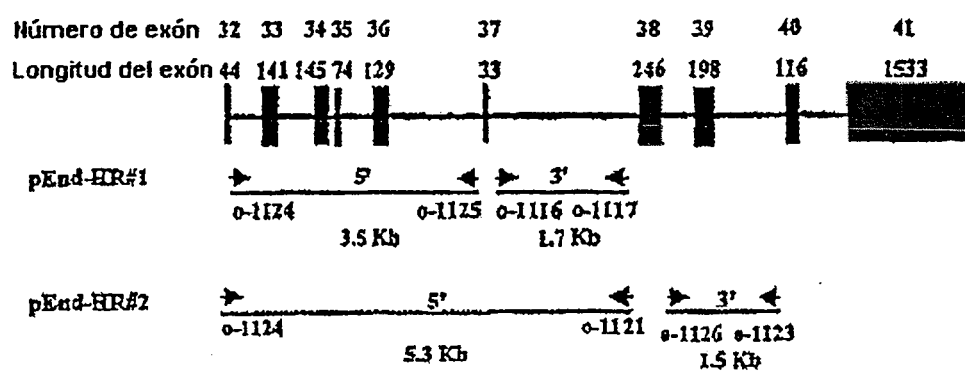


FIGURA 4

A)



B)

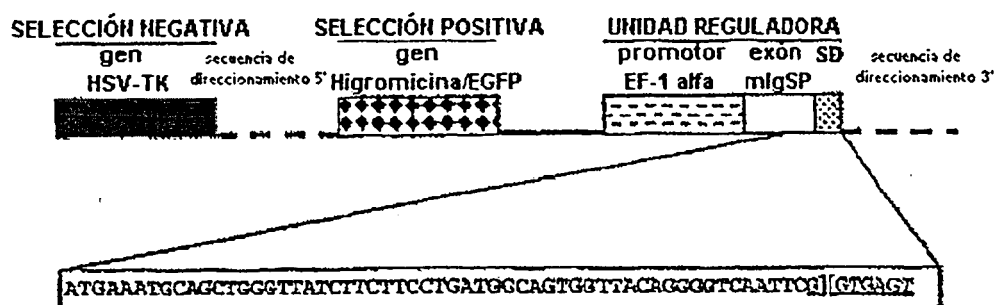


FIGURA 5

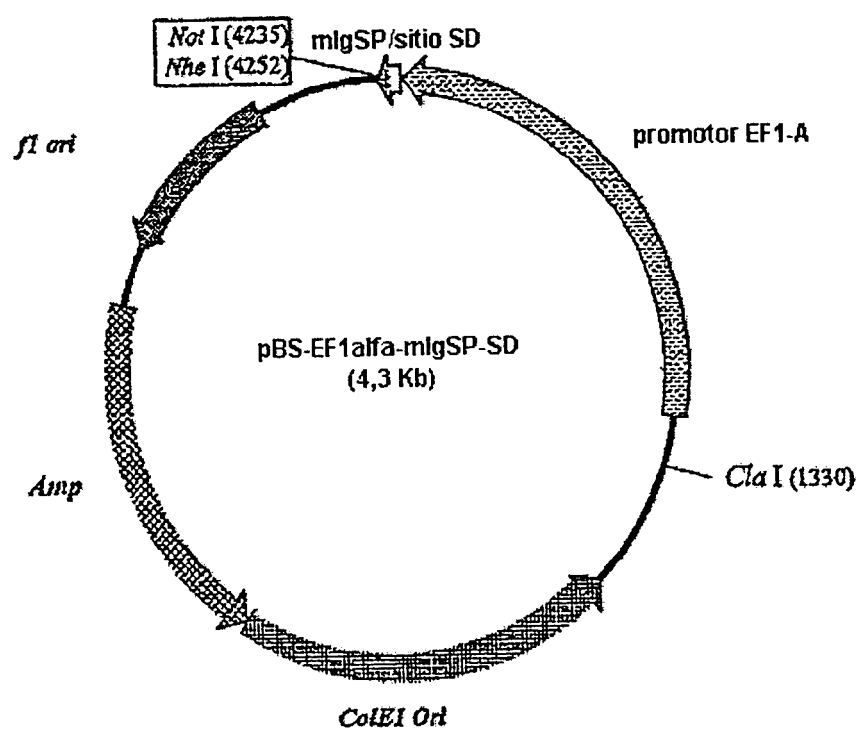


FIGURA 6

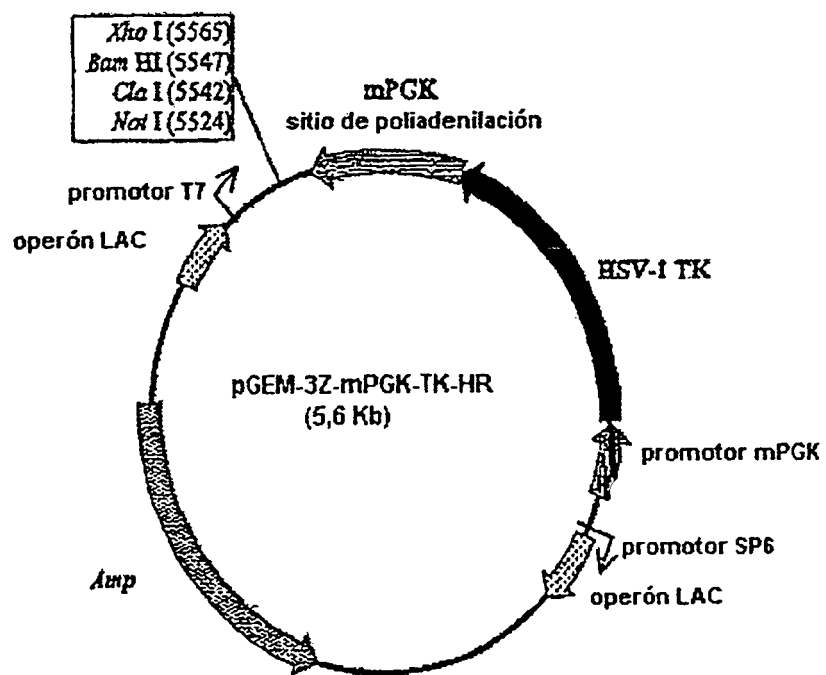
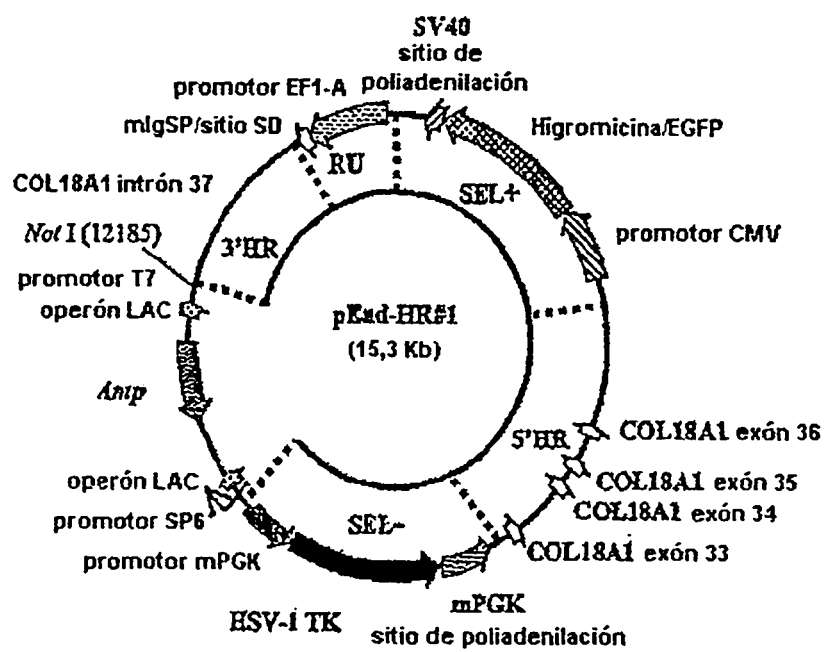


FIGURA 7



1
← CCGCTGAGCTCAGAGTGCCTACAGACTCAGCTAGATCGGGCTGTGAC

exón mlgSP

1
M K C S N V I F F L M A Y
51
GTTCTTACAATGAAGTCAGCTGGGTATCTCTCTCTCTCATGCGCAGTGG

exón mlgSP

14
V T G V N S D H E V A A L Q P E V
181
TTACAGGCGGTGAATTCGGCACTAAGAATGGCGCGCCCTTCAGGCCCGCTG

exón mlgSP exón 38

31
V Q L H D S N F Y P R R E H P H P
161
CTCCAGCTGCAAGCACGACCCTTACCGGGGGGCGSAGCACCGCCGACCG

exón 38

48
T A R P W R A D D I L A S P P H
201
CAACGGGGCGCGCCCTGCGGGCGAGTGACATCTCTGGCGAGCGCCCGCTGCG

exón 38

64
L P E F Q P Y P G A P K H S S Y V
251
TGCGCGAGCGCCAGCGCTAACCGCGAGCCCGCGACCGACCGCTCCCTAGTGC

exón 38

81
H L R P A R P T S P P A H S H R D
301
CACTTCGGCGCGCGCGAGCCGACGAGCGCGCGCGCGCGCGCGCGCGCGCG

exón 38

98
F Q P V L H L V A L H S P L S G
351
CTTCAGCGCGCTGCTCCACCTGCTTGGCGTCAACAGCGCCCGCTTCAGCGG

exón 38 exón 39

114
G H E G I E G A D E Q C F Q Q A R
401
GCAATGCGCGCATCGCGCGCGCGCGCGCTCCAGCTCTTCAGCGAGCGCGG

exón 39

121
A V G L A G T P R A F L S S R L Q
451
GCGTTCGGGCTGCGCGCGCGCGCTTCGGCGGCTCTCTGCTCTGCGCGCTGCA

exón 39

148
D L T S I V R R A D R A A V P I
501
GGACCTCTACAGCATCTGTGGCTGGTGGCGACCGCGCGCGCGCGCGCGCG

exón 39

164
Y H L K D E L L P F S H E A D F S
551
TCAACCTCAAGGACAGCTCTCTTTCGCGCTGGGAGGCTCTCTCTCTCA

exón 39 exón 40

181
G S E G P L K P G A R I Z S P D G
601
GGCTCTGAGGCTCGCTCAAGCGCGCGCGCGCGCATCTCTCTCTCTCAAGG

exón 40

199
K O V L R H P T W P Q K S V N H
651
CAAGGAGCTCTGAGGCGACCGCGCGCTGCGCGCGCGCGCGCGCGCGCGCG

exón 40 exón 41

214
G S P P H G R X L T E S Y C B T W
701
GCTCGGACCG

exón 41

FIGURA 8 (cont.)

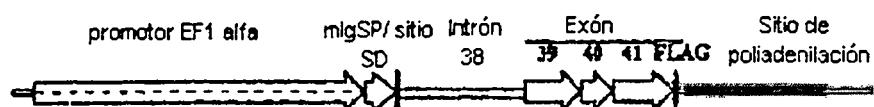
231	R T E A F S A T G Q A S S L L G G
751	<u>CCGACCGAGGCTCCCTCGGACACGGGCGCGGCTCTCCCTGCTCGGGG</u>
	exón 41
248	R L L G Q S A A S C H H A Y I V
801	<u>CAAGCTCTCGGGCGAGCTCGGCGGCTGCGCATGACGCTACATCTTC</u>
	exón 41
264	L C I E N S F H T A S Xparada
821	<u>TCTGCATTGAGAGACGCTTCATGACTGCTCCAAAGTAGGCGCGCTGGA</u>
	exón 41

FIGURA 9

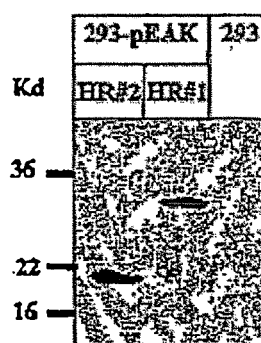


FIGURA 10

A)



B)



C)

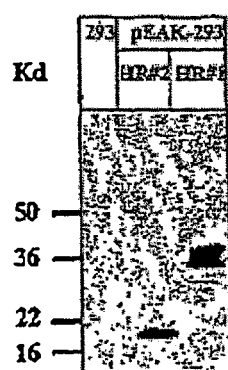
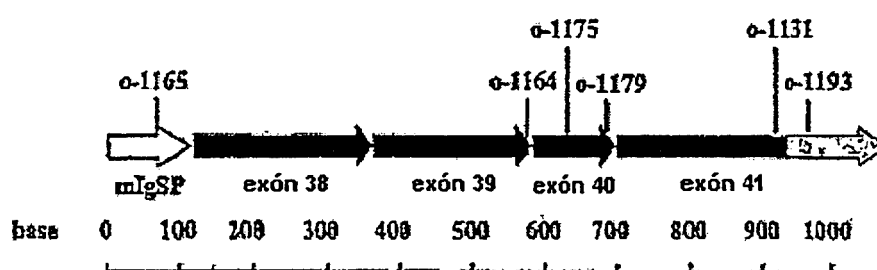


FIGURA 11

A)



B)

exón 38

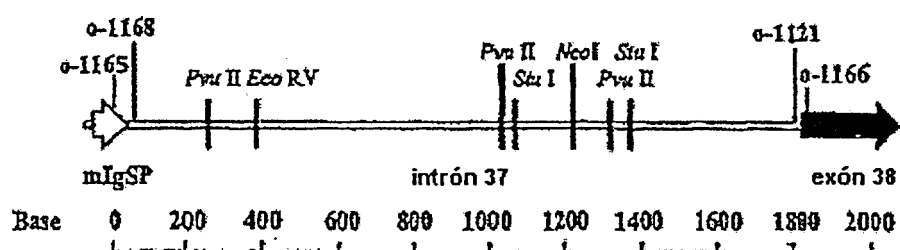


FIGURA 12

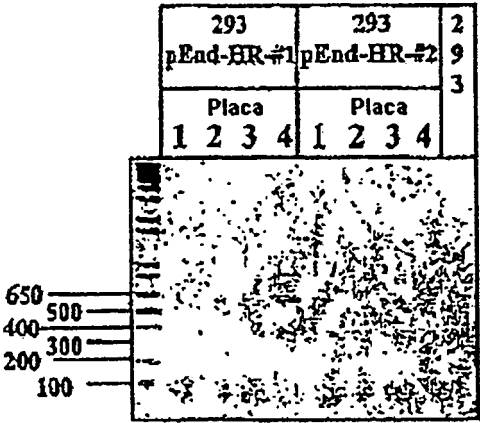
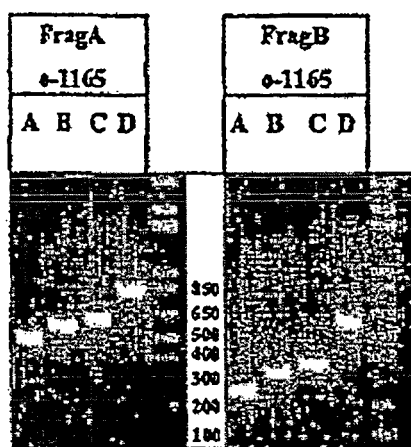


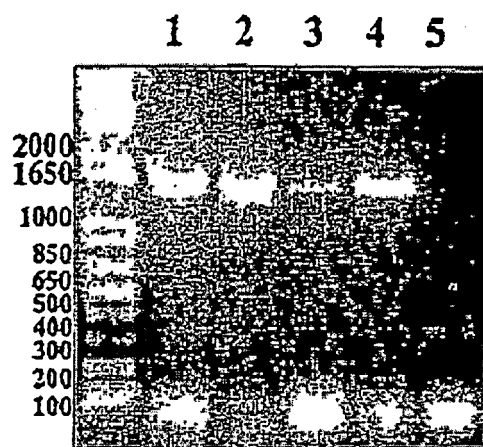
FIGURA 13



A = o-1164
B = o-1175
C = o-1179
D = o-1131

FIGURA 14

A)



B)

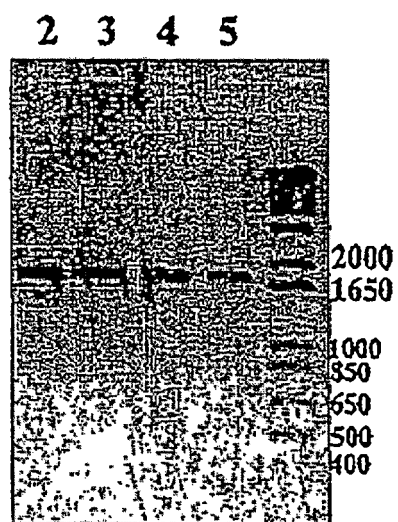
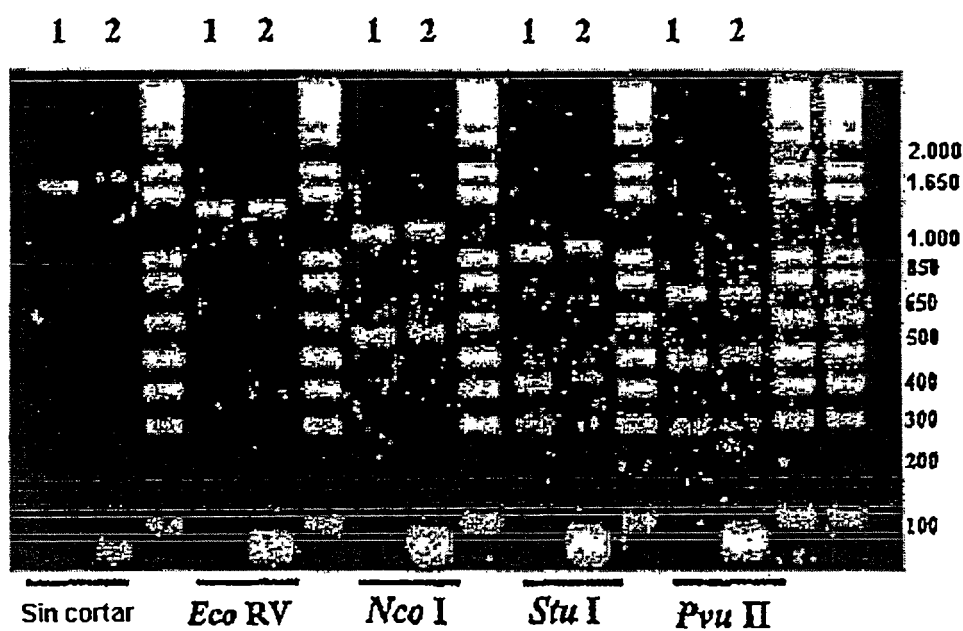


FIGURA 15

A)



B)

	(1)	(2)
Sin cortar	1.784	1.821
<i>Eco RV</i>	1.418, 366	1.427, 394
<i>Nco I</i>	1.192, 592	1.220, 601
<i>Stu I</i>	1.052, 425, 307	1.080, 434, 307
<i>Pvu II</i>	781, 497, 294, 212	781, 506, 294, 240