(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 618 632**

(51) Int. CI.:

C12N 15/62 (2006.01)
C12N 15/82 (2006.01)
C07K 14/195 (2006.01)
C12N 1/21 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **28.09.2012 PCT/EP2012/069219**

(87) Fecha y número de publicación internacional: **04.04.2013 WO2013045632**

(96) Fecha de presentación y número de la solicitud europea: **28.09.2012 E 12772263 (5)**

(97) Fecha y número de publicación de la concesión europea: **14.12.2016 EP 2761006**

(54) Título: **Inteínas divididas y usos de éstas**

(30) Prioridad:

28.09.2011 US 201161540101 P
13.06.2012 EP 12171848

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
21.06.2017

(73) Titular/es:

ZERA INTEIN PROTEIN SOLUTIONS, S.L. (100.0%)
Calle Cavallers, 50
08034 Barcelona, ES

(72) Inventor/es:

PALLISSE BERGWERF, ROSER;
SCHMIDT, STEFAN ROBERT;
MARCO FELIU, DÍDAC y
CARVAJAL VALLEJOS, PATRICIA KARINA

(74) Agente/Representante:

ARIAS SANZ, Juan

DESCRIPCIÓN

Inteínas divididas y usos de éstas

Campo de la invención

La presente invención se refiere generalmente a inteínas divididas robustas y a usos de éstas, por ejemplo, en la purificación y en la ingeniería de proteínas.

Antecedentes

Las inteínas son elementos proteicos internos que se auto-escinden de su proteína huésped y catalizan la ligación de las secuencias flanqueantes (exteínas) con un enlace peptídico. La escisión de inteínas es un proceso posterior a la traducción que no requiere enzimas o cofactores auxiliares. Este proceso de auto-escisión se denomina "corte y empalme de proteínas", por analogía con el corte y empalme de intrones de ARN de preARNm (Perler F et al., Nucl Acids Res. 22:1125-1127 (1994)). Los segmentos se denominan "inteína" para la secuencia proteica interna, y "exteína" para la secuencia proteica externa, con las exteínas aguas arriba denominadas "N-exteínas" y las exteínas aguas abajo denominadas "C-exteínas". Los productos del proceso de corte y empalme de proteínas son dos proteínas estables: la proteína madura y la inteína.

Estructura de mini-inteínas e inteínas grandes

Las inteínas se clasifican en dos grupos: grande y mínima (mini) (Liu XQ, Ann Rev Genet 34:61-76 (2000)). Las inteínas grandes contienen un dominio de endonucleasa de direccionamiento que está ausente en las mini-inteínas. Se han preparado por ingeniería mini-inteínas con corte y empalme eficiente a partir de inteínas grandes mediante la deleción del dominio de endonucleasa central, demostrando que el dominio de endonucleasa no está implicado en el corte y empalme de proteínas (Chong S. y Xu M., J Biol Chem. 272:15587-15589 (1997); Derbyshire V. et al., Proc Natl Acad Sci USA. 94:11466-11471 (1997); y Shingledecker K. et al. Gene. 207:187-195 (1998)).

Todas las inteínas conocidas comparten un grado bajo de similitud de secuencia, con residuos conservados sólo en los extremos N y C. La mayor parte de las inteínas empiezan con Ser o Cys y terminan en His-Asn o en His-Gln. El primer aminoácido de la C-exteína es una Ser, Thr, o Cys, invariable, pero el residuo que precede la inteína en la N-exteína no está conservado (Perler F. 2002, Nucl. Acids Res. 30: 383-384). Sin embargo, se encontró que los residuos próximos a la unión inteína-corte y empalme en ambas exteínas N y C-terminales aceleraban o atenuaban el corte y empalme de proteínas (Amitai G et al. 2009, Proc. Natl. Acad. Sci. USA. 106:11005-11010).

Mecanismos de corte y empalme en cis y trans de inteínas

Las inteínas pueden clasificarse por su mecanismo de corte y empalme. Las inteínas de clase 1, que son el grupo de inteínas más estudiado, tienen un proceso rápido de cuatro ataques nucleofílicos, mediados por tres de los cuatro residuos conservados de unión de corte y empalme. En la etapa 1, el proceso de corte y empalme empieza con un desplazamiento acilo del residuo de serina o cisteína localizado en la primera posición del dominio de corte y empalme N-terminal. Esto forma un enlace (tio)éster en la unión N-exteína/inteína. En la etapa 2, el enlace (tio)éster es atacado por el grupo OH o SH del primer residuo en la C-exteína (Cys, Ser, o Thr). Esto da lugar a una transesterificación, que transfiere la N-exteína a la cadena lateral del primer residuo de la C-exteína. En la etapa 3, la ciclación del residuo conservado Asn o Gln localizado en la última posición del dominio de corte y empalme C-terminal une las exteínas por un enlace (tio)éster. Finalmente, la etapa 4 es una reorganización del enlace (tio)éster a un enlace peptídico por un desplazamiento acilo espontáneo S-N o O-N. Los aminoácidos importantes implicados directamente o indirectamente en la reacción de corte y empalme se muestran en la figura 3A.

La escisión específica de sitio de las uniones inteína-exteína en las inteínas de la clase 1 puede conseguirse por mutación de los residuos conservados de la inteína. La mutación del residuo Asn o Gln en el extremo C de la inteína suprime las etapas 3 y 4 de la reacción de corte y empalme y resulta sólo en la escisión N-terminal. Como la etapa 1 todavía ocurre, el enlace (tio)éster puede hidrolizarse espontáneamente, separando la N-exteína de la parte inteína/C-exteína. El residuo de serina o cisteína localizado en la primera posición del dominio de corte y empalme N-terminal se requiere para la escisión N-terminal (véase la figura 3C). La mutación de este primer residuo conservado de la inteína suprime las etapas 1, 2, y 4 de la reacción de corte y empalme y da lugar sólo a la escisión C-terminal. En dicha inteína mutada, la ciclación de Asn (etapa 3) todavía ocurre, para separar la C-exteína de la parte N-exteína/inteína. La Asn (o Gln), y los residuos de His localizados respectivamente en las posiciones última (X_N) y penúltima (X_{N-1}) del dominio de corte y empalme C-terminal se requieren para la escisión N-terminal (véase la figura 3B). La escisión controlable de las inteínas con corte y empalme en cis modificadas se ha adaptado para un amplio rango de aplicaciones útiles en biología molecular y biotecnología.

Inteínas divididas naturales

Las inteínas también pueden existir como dos fragmentos codificados por dos genes que se transcriben y traducen de forma separada. Estas inteínas denominadas divididas se auto-asocian y catalizan actividad de corte y empalme de proteínas en trans.

Las inteínas divididas se han identificado en diversas cianobacterias y archaea (Caspi et al., Mol Microbiol. 50:1569-1577 (2003); Choi J. et al., J Mol Biol. 356:1093-1106 (2006.); Dassa B. et al., Biochemistry. 46:322-330 (2007.); Liu X. y Yang J., J Biol Chem. 278:26315-26318 (2003); Wu H. et al., Proc Natl Acad Sci USA. 95:9226-9231 (1998.); y Zettler J. et al., FEBS Letters. 583:909-914 (2009)), pero no se han encontrado en eucariotas hasta ahora. Recientemente, un análisis bioinformático de datos metagenómicos medioambientales reveló 26 loci diferentes con una nueva organización genómica. En cada locus, una región codificadora de enzima conservada está interrumpida por una inteína dividida, con un gen de endonucleasa independiente insertado entre las secciones que codifican los subdominios de la inteína. Entre ellos, se ensamblaron completamente cinco loci: las ADN helicasas (gp41-1, gp41-8); Inosina-5'-monofosfato deshidrogenasa (IMPDH-1); y subunidades catalíticas de ribonucleótido reductasa (NrdA-2 y NrdJ-1). Esta organización génica fracturada parece estar presente principalmente en fagos (Dassa et al., Nucleic Acids Research. 37:2560-2573 (2009)).

La inteína dividida Npu DnaE se caracterizó como que tiene la tasa más alta reportada para la reacción de corte y empalme en trans de proteínas. Además, la reacción de corte y empalme de proteínas de Npu DnaE se considera robusta y con un alto rendimiento respecto a diferentes secuencias de exteína, temperaturas de 6 a 37°C, y la presencia de hasta 6M Urea (Zettler J. et al., FEBS Letters. 583:909-914 (2009); Iwai I. et al., FEBSLetters 580:1853-1858 (2006)). Como se esperaba, cuando se introdujo la mutación Cysl Ala en el dominio N de estas inteínas, el desplazamiento inicial de acilo N a S y, por lo tanto, el corte y empalme de proteínas se bloqueó. Desafortunadamente, la reacción de escisión C-terminal también se inhibió casi completamente. La dependencia de la ciclación de asparagina en la unión de corte y empalme C-terminal en el desplazamiento de acilo en el enlace peptídico escindible N-terminal parece ser una propiedad única común a los alelos de inteína dividida DnaE naturales (Zettler J. et al. FEBS Letters. 583:909-914 (2009)).

Aplicaciones de inteínas en biotecnología

Las inteínas son herramientas valiosas en un amplio rango de aplicaciones biotecnológicas. La ligación de péptidos y proteínas usando la actividad de corte y empalme natural de las inteínas se conoce como ligación de proteínas mediada por inteínas (IPL), o ligación de proteínas expresadas (EPL), y está bien establecida en métodos de biología molecular y biotecnología (Evans T. et al., Biopolymers 51:333-342 (1999); Muir T. et al., Proc Natl Acad Sci USA. 95:6705-6710 (1998); y Severinov K. y Muir T., J Biol Chem. 273:16205-16209 (1998)). Además, las inteínas se han usado para la purificación de proteínas por la escisión específica de sitio sólo en el límite inteína-proteína diana (Lu W. et al, J Chromatography A. 1218:2553-2560 (2011)). El uso de procedimientos mediados por inteínas en la bioseparación está bien establecido a escala de laboratorio y está atrayendo un interés creciente en biotecnología a gran escala. El potencial de estas técnicas de purificación de proteínas para la producción de proteínas a gran escala es claro, pero deben desarrollarse los sistemas de purificación de proteínas mediados por inteínas en condiciones industriales a mayor escala. Otras aplicaciones son marcaje segmental de proteínas para análisis por RMN, ciclación de proteínas, expresión controlada de proteínas tóxicas, conjugación de puntos cuánticos a proteínas e incorporación de aminoácidos no canónicos, (Arnold U., Biotechnol Lett. 31:1129-1139 (2009); Charalambous A. et al., J Nanobiotechnology 7:9 (2009); Oeemig J. et al., FEBS Letters 583:1451-1456 (2009); Seyedsayamdost M. et al., Nat Protoc. 2:1225-1235 (2007); Züger S. y Iwai H., Nat Biotechnol. 23:736-740 (2005); y Evans T. et al., Annu Rev Plant Biol. 56:375-392 (2005)). En estudios de investigación básica, las inteínas se han usado para monitorizar las interacciones proteína-proteína in vivo, específicamente la translocación de proteínas en orgánulos celulares, ligación de polipéptido exógeno a proteínas de membrana en células vivas o fotocontrol de la actividad de proteínas (Chong S. y Xu M., Homing endonucleases and inteins. Vol 16. Springer, Berlin Heidelberg, Nueva York, 273-292 (2005); Ozawa T. y Umezawa Y., Homing endonucleases and inteins. Vol 16. Springer, Berlin Heidelberg, Nueva York, 307-323 (2005); Ozawa T. et al., Nat Biotechnol. 21:287-293 (2003); Dhar T. y Mootz H., Chem Commun. 47:3063-3065 (2011); y Binschik J. et al., Angewandte Chemie International Ed. 50(14):3249-3252 (2011)). La mayor parte de las inteínas usadas en biotecnología derivan de organismos procariotas, o son variantes preparadas por ingeniería de la inteína VMA1 de *S. cerevisiae* (Elleuche y Pöggeler 2010 Appl. Microbiol Biotechnol 78:479-489).

Con el fin de hacer uso de dichas técnicas en procesos biológicos a gran escala, deben identificarse inteínas con propiedades robustas y métodos para usar las mismas. Las inteínas y métodos para usar dichas inteínas que se describen en la presente memoria abordan esta necesidad proporcionando inteínas altamente activas que funcionan en un gran intervalo de temperatura, en presencia de sales, y cuando se fusionan a polipéptidos de secuencias variables.

Compendio breve de la invención

La presente invención proporciona inteínas divididas robustas y métodos para usar las mismas. Las inteínas divididas son activas en un gran intervalo de temperatura, en un amplio intervalo de pH, y en presencia de sales caotrópicas. También muestran una alta tolerancia a variabilidad de secuencia en polipéptidos heterólogos fusionados. Estas características hacen que las inteínas divididas sean especialmente útiles en técnicas de purificación e ingeniería de proteínas.

En particular, se proporcionan proteínas de fusión que comprenden (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y (ii) un polipéptido

heterólogo, en el que el polipéptido heterólogo es C-terminal respecto al dominio de inteína. En algunas realizaciones, el último aminoácido del dominio de inteína es asparagina o glutamina. En algunas realizaciones, el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina, por ejemplo, una alanina. En algunas realizaciones, el penúltimo aminoácido del dominio de inteína es un aminoácido distinto de histidina. En algunas realizaciones, el polipéptido heterólogo está unido directamente al dominio de inteína a través de un enlace peptídico. En algunas realizaciones, el primer aminoácido del polipéptido heterólogo es serina, cisteína, o treonina. En algunas realizaciones, el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina, por ejemplo, una alanina y el primer aminoácido del polipéptido heterólogo es distinto de serina, treonina o cisteína, por ejemplo, alanina. En algunas realizaciones, la proteína de fusión comprende además un conector entre el polipéptido heterólogo y el dominio de inteína. En algunas realizaciones, el primer aminoácido del conector es serina, cisteína, o treonina. En algunas realizaciones, el primer aminoácido del conector es un aminoácido distinto de serina, cisteína, o treonina, es decir, una alanina. En algunas realizaciones, el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina, por ejemplo, una alanina y el primer aminoácido del conector es un aminoácido distinto de serina, treonina o cisteína, por ejemplo, una alanina. En algunas realizaciones, el conector comprende 1-5 aminoácidos de una secuencia de exteína nativa. También se proporcionan proteínas de fusión que comprenden un dominio de inteína que tiene una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y (ii) un polipéptido heterólogo, en el que el polipéptido heterólogo es C-terminal respecto al dominio de inteína.

Además, se proporcionan proteínas de fusión que comprenden (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y (ii) un polipéptido heterólogo, en el que el polipéptido heterólogo es N-terminal respecto al dominio de inteína. En algunas realizaciones, el primer aminoácido del dominio de inteína es una cisteína. En algunas realizaciones, el primer aminoácido del dominio de inteína es un aminoácido distinto de serina o cisteína, por ejemplo, una alanina. En algunas realizaciones, el polipéptido heterólogo está unido directamente al dominio de inteína a través de un enlace peptídico. En algunas realizaciones, la proteína de fusión comprende además un conector entre el polipéptido heterólogo y el dominio de inteína. En algunas realizaciones, el conector comprende 1-5 aminoácidos de una secuencia de exteína nativa. También se proporcionan proteínas de fusión que comprenden un dominio de inteína que tiene una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y un polipéptido heterólogo, en el que el polipéptido heterólogo es N-terminal respecto al dominio de inteína.

Además, se proporcionan proteínas de fusión que comprenden un primer dominio de inteína, un segundo dominio de inteína, y un polipéptido heterólogo. Además, se proporcionan proteínas de fusión que comprenden un primer dominio de inteína, un segundo dominio de inteína, y un polipéptido heterólogo en el que el polipéptido heterólogo es N-terminal respecto al primer dominio de inteína, y el polipéptido heterólogo es C-terminal respecto al segundo dominio de inteína. Además, se proporcionan proteínas de fusión que comprenden un primer dominio de inteína, un segundo dominio de inteína, y un polipéptido heterólogo en el que el polipéptido heterólogo es N-terminal respecto al primer dominio de inteína (dominio de corte y empalme N-terminal), y el polipéptido heterólogo es C-terminal respecto al segundo dominio de inteína (dominio de corte y empalme C-terminal). En algunas realizaciones, (a) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:3 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:7; (b) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:12 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:16; (c) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:20 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:24; (d) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:34 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:38; o (d) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:64 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:65. En algunas realizaciones, el primer aminoácido del polipéptido heterólogo es serina, cisteína, o treonina. En algunas realizaciones, la proteína de fusión comprende además un conector entre el polipéptido heterólogo y el segundo dominio de inteína, en el que el primer aminoácido del conector es serina, cisteína, o treonina. En algunas realizaciones, el primer aminoácido del conector es serina.

En la presente memoria también se proporcionan polinucleótidos que codifican las proteínas de fusión según la invención.

También se proporcionan composiciones que comprenden proteínas de fusión. Dichas composiciones son útiles, por ejemplo, para reacciones de escisión C-terminal, reacciones de escisión N-terminal, reacciones de corte y empalme en trans, y métodos de ciclación de proteínas.

También se proporcionan células huésped que comprenden las proteínas, proteínas de fusión, polinucleótidos o composiciones.

Se proporcionan métodos para usar polipéptidos y proteínas de fusión proporcionados en la presente memoria, por ejemplo, en reacciones de escisión C-terminal, reacciones de escisión N-terminal, reacciones de corte y empalme en trans, y ciclación de proteínas. Dichos métodos pueden ocurrir a temperaturas de aproximadamente 0°C a aproximadamente 60^0C a un pH de aproximadamente 6 a aproximadamente 10, y/o en presencia de aproximadamente 0,5 M a aproximadamente 6 M urea.

En algunas realizaciones, la constante de velocidad de la reacción de las reacciones proporcionadas en la presente memoria es al menos aproximadamente 1 x 10^{-1} s^{-1}, o al menos aproximadamente 2 x 10^{-1} s^{-1}. En algunas realizaciones, la vida media de la velocidad de reacción es menor de aproximadamente 100 segundos, menor de aproximadamente 50 segundos, o menor de aproximadamente 25 segundos o menor de aproximadamente 15 segundos.

Las reacciones pueden iniciarse, por ejemplo, por un desplazamiento en temperatura o pH o mezclando proteínas.

La invención también proporciona un vector que comprende un polinucleótido que codifica un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y al menos un sitio de clonación aguas abajo de dicho polinucleótido que permite la clonación de un polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el dominio de inteína y el polipéptido codificado por el polinucleótido de interés.

La invención también proporciona un vector que comprende un polinucleótido que codifica un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y al menos un sitio de clonación aguas arriba de dicho polinucleótido que permite la clonación de un polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el polipéptido codificado por el polinucleótido de interés y el dominio de inteína.

La invención también proporciona un vector que comprende un polinucleótido que codifica un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65, y al menos un sitio de clonación aguas abajo de dicho polinucleótido que permite la clonación de un polinucleótido de interés, y un polinucleótido aguas abajo del sitio de clonación, que codifica un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64, de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el polipéptido codificado por el polinucleótido de interés y los dos dominios de inteína, en el que

a. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:7, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:3;

b. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:16, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:12;

c. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:24, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:20;

d. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:38, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:34.

La invención también proporciona un vector que comprende un polinucleótido que codifica un primer dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65, un primer sitio de clonación aguas abajo de dicho polinucleótido que codifica un primer dominio de inteína, un polinucleótido que codifica un segundo dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y un segundo sitio de clonación aguas arriba de dicho polinucleótido que codifica un segundo dominio de inteína, en el que el primer sitio de clonación permite la clonación de un primer polinucleótido de interés y el segundo sitio de clonación permite la clonación de un segundo polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende, en dicho orden, el polipéptido codificado por el segundo polinucleótido de interés, el segundo dominio de inteína, el primer dominio de inteína y el polipéptido codificado por el segundo polinucleótido de interés y en el que

a. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:7, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:3;

b. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:16, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:12;

c. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:24, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:20;

d. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:38, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:34 o

e. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:65, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:64.

Breve descripción de los dibujos/figuras

Fig. 1. **(A)** Representación esquemática de una reacción de corte y empalme en trans. El fragmento inicial en el extremo N (F1) y el fragmento en el extremo C (F2) de la inteína dividida se muestran en la parte superior. Las regiones que se unirán después de la reacción de corte y empalme en trans se indican en gris. Los 3 fragmentos que resultan de la reacción de corte y empalme en trans se muestran en la parte inferior. F3 representa el producto de corte y empalme (en gris), y F4 y F5 corresponden a los subproductos de la inteína dividida N y C, respectivamente. La etiqueta de purificación StreptagII (ST), proteína de cabeza D del fago bacteriófago λ (gpD), los cinco aminoácidos naturales flanqueantes que pertenecen a la exteína del extremo N (E^N), la inteína dividida del extremo N (IN), la etiqueta de purificación hexa-histidina (H6), la inteína dividida del extremo C (I^C), los cinco aminoácidos naturales flanqueantes que pertenecen a la exteína del extremo C (E^C), tiorredoxina (Trx), y la unión de las exteínas del extremo N y C (E^{NC}). **(B)** Curso de tiempo de la reacción de corte y empalme en trans de la inteína dividida Gp41.1 (G1) analizado por tinción con azul de Coomassie de un gel de SDS-PAGE. Los fragmentos iniciales en el extremo N y C (F1 y F2, respectivamente) se muestran en el carril 1. La reacción de corte y empalme en trans después de incubación a 25°C durante 10s, 20s, 30s, 1 min, 5 min, 30 min, 1h y 2h se muestra en los carriles 2-9. La posición de F1, F2, F3, F4 y F5 se indica por flechas.

Fig. 2. **(A)** Reacción de auto-escisión C-terminal de la inteína dividida Gp41.1 (G1). El primer aminoácido (cisteína) de la inteína del extremo N se ha sustituido con una alanina (C1A). La representación esquemática de una reacción de auto-escisión general se muestra a la izquierda. La incubación del fragmento del extremo N mutado (F1) y el fragmento del extremo C (F2) de la inteína dividida resulta en la escisión de F2 para generar el fragmento que contiene Trx deseado (F3) y el subproducto de la inteína dividida del extremo C (F4). El fragmento liberado en la reacción de auto-escisión se indica en gris. Un curso de tiempo de la reacción de auto-escisión de la inteína dividida Gp41.1 (G1) analizado por tinción con azul de Coomassie de un gel de SDS-PAGE se muestra a la derecha. Los fragmentos iniciales del extremo N y C (F1 y F2, respectivamente) se muestran en el carril 1. La reacción de auto-escisión después de incubación a 37°C durante 1 min, 5 min, 35 min, 1h, 2h y 4h, se muestra en los carriles 2 a 7. La posición de F1, F2, F3, y F4 se indica por flechas. (**B)** Reacción de auto-escisión C-terminal de la inteína dividida Gp41.1 (G1). La representación esquemática de una reacción de auto-escisión general cuando la inteína del extremo N contiene la sustitución C1A y los 5 aminoácidos naturales correspondientes a la E^C están ausentes de F2 se muestra a la izquierda. Una reacción de auto-escisión de la inteína dividida Gp41.1 (G1) analizado por tinción con azul de Coomassie de un gel de SDS-PAGE se muestra a la derecha. Los fragmentos iniciales del extremo N y C (F1 y F2, respectivamente) se muestran en el carril 2. El carril 3 muestra la reacción de auto-escisión después de incubación a 37°C durante 23h. La posición de F1, F2, F3, y F4 se indica por flechas. La etiqueta de purificación StreptagII (ST), proteína de cabeza D del fago bacteriófago λ (gpD), los cinco aminoácidos naturales flanqueantes que pertenecen a la exteína del extremo N (E^N), la inteína dividida del extremo N (IN), la etiqueta de purificación hexa-histidina (H6), la inteína dividida del extremo C (I^C), los cinco aminoácidos naturales flanqueantes que pertenecen a la exteína del extremo C (E^C) y tiorredoxina (Trx).

Fig. 3. Representación esquemática que muestra construcciones de **(A)** corte y empalme en trans, **(B)** auto-escisión C-terminal, y **(C)** auto-escisión N-terminal. Los aminoácidos naturales flanqueantes que pertenecen a la exteína del extremo N (E^N), la inteína dividida del extremo N (IN), la inteína dividida del extremo C (I^C), los aminoácidos naturales flanqueantes que pertenecen a la exteína del extremo C (E^C). Se indican los aminoácidos clave implicados directamente o indirectamente en la reacción correspondiente.

Descripción detallada de la invención

Lo siguiente proporciona una descripción de inteínas divididas que son útiles en varias aplicaciones de ingeniería de proteínas. Las inteínas divididas contienen secuencias Gp41.1, Gp41.8, NrdA2, NrdJ1 o IMPDH1 fusionadas a proteínas heterólogas y pueden usarse, por ejemplo, en la síntesis, escisión, purificación, ligación, ciclación de proteínas, y regulación y/o monitorización de la actividad de proteínas.

Los encabezamientos de sección usados en la presente memoria tienen sólo propósitos organizativos y no deben considerarse de ninguna manera como limitantes de la materia sujeto descrita.

I. Definiciones

A no ser que se defina expresamente otra cosa, los términos usados en la presente memoria deben entenderse según su significado ordinario en la técnica. Los términos usados en el singular o referidos como "un" o "una" también incluyen el plural y vice versa, a no ser que se especifique otra cosa o se indique por el contexto. Las técnicas y procedimientos estándar se realizan generalmente según métodos convencionales en la técnica y varias referencias generales (véase, generalmente, Sambrook *et al.* Molecular Cloning: A Laboratory Manual, 2ª ed. (1989) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., que se incorpora en la presente memoria por referencia), que se proporcionan a lo largo de este documento.

Los términos "polipéptido", "péptido", y "proteína" se usan indistintamente en la presente memoria para hacer referencia a polímeros de aminoácidos de cualquier longitud. El polímero puede ser lineal o ramificado, puede comprender aminoácidos modificados, y puede estar interrumpido por no aminoácidos. Los términos también

engloban un polímero de aminoácidos que se ha modificado naturalmente o por intervención; por ejemplo, formación de enlace disulfuro, glicosilación, lipidación, acetilación, fosforilación, o cualquier otra manipulación o modificación, tal como conjugación, con un componente de marcaje. También se incluyen en la definición, por ejemplo, polipéptidos que contienen uno o más análogos de un aminoácido (incluyendo, por ejemplo, aminoácidos no naturales, *etc.*), así como otras modificaciones conocidas en la técnica.

Un "polipéptido de fusión" es un polipéptido comprendido por al menos dos polipéptidos y opcionalmente una secuencia conectora para unir de forma operativa los dos polipéptidos en un polipéptido continuo. Los dos polipéptidos unidos en un polipéptido de fusión derivan típicamente de dos fuentes independientes, y, por lo tanto, un polipéptido de fusión comprende dos polipéptidos unidos que normalmente no se encuentran unidos en la naturaleza. Los dos polipéptidos pueden estar unidos de forma operativa por un enlace peptídico o pueden estar unidos indirectamente a través de un conector descrito en la presente memoria o conocido de otra forma en la técnica.

Un "ácido nucleico", "polinucleótido", o "molécula de ácido nucleico" es un compuesto polimérico comprendido por subunidades unidas covalentemente denominadas nucleótidos. El ácido nucleico incluye ácido polirribonucleico (ARN) y ácido polidesoxirribonucleico (ADN), ambos de los cuales pueden ser monocatenarios o bicatenarios. El ADN incluye ADNc, ADN genómico, ADN sintético, y ADN semi-sintético.

Los términos "idéntico" o porcentaje de "identidad" en el contexto de dos o más ácidos nucleicos o polipéptidos, se refieren a dos o más secuencias o subsecuencias que son iguales o que tienen un porcentaje especificado de residuos de nucleótidos o aminoácidos que son iguales, cuando se compara y alinea (introduciendo huecos, si es necesario) para una máxima correspondencia, sin considerar ninguna sustitución de aminoácidos conservativa como parte de la identidad de secuencia. El porcentaje de identidad puede medirse usando software de comparación de secuencias o algoritmos o por inspección visual. En la técnica se conocen varios algoritmos y software que pueden usarse para obtener alineamientos de secuencias de aminoácidos o nucleótidos. Uno de dichos ejemplos no limitativos de un algoritmo de alineamiento de secuencias es el algoritmo descrito en Karlin et al, 1990, Proc. Natl. Acad. Sci., 87:2264-2268, según se modifica en Karlin et al., 1993, Proc. Natl. Acad. Sci., 90:5873-5877, e incorporado en los programas NBLAST y XBLAST (Altschul et al., 1991, Nucleic Acids Res., 25:3389-3402). En determinadas realizaciones, puede usarse Gapped BLAST como se describe en Altschul et al., 1997, Nucleic Acids Res. 25:3389-3402. BLAST-2, WU-BLAST-2 (Altschul et al., 1996, Methods in Enzymology, 266:460-480), ALIGN, ALIGN-2 (Genentech, South San Francisco, California) o Megalign (DNASTAR) son programas de software públicamente disponibles que pueden usarse para alinear secuencias. En determinadas realizaciones, el porcentaje de identidad entre dos secuencias de nucleótidos se determina usando el programa GAP en software GCG (por ejemplo, usando una matriz NWSgapdna.CMP y un peso de hueco de 40, 50, 60, 70, ó 90 y un peso de longitud de 1, 2, 3, 4, 5, ó 6). En determinadas realizaciones alternativas, el programa GAP en el paquete de software GCG, que incorpora el algoritmo de Needleman y Wunsch (J. Mol. Biol. 48:444-453 (1970)) puede usarse para determinar el porcentaje de identidad entre dos secuencias de aminoácidos (por ejemplo, usando bien una matriz Blossum 62 o una matriz PAM250, y un peso de hueco de 16, 14, 12, 10, 8, 6, ó 4 y un peso de longitud de 1, 2, 3, 4, 5). Alternativamente, en determinadas realizaciones, el porcentaje de identidad entre secuencias de nucleótidos o aminoácidos se determina usando el algoritmo de Myers y Miller (CABIOS, 4:11-17 (1989)). Por ejemplo, el porcentaje de identidad puede determinarse usando el programa ALIGN (versión 2.0) y usando una PAM120 con tabla de residuo, una penalización por longitud de hueco de 12 y una penalización por hueco de 4. Los parámetros apropiados para el alineamiento máximo por software de alineamiento particular pueden determinarse por un experto en la técnica. En determinadas realizaciones, se usan los parámetros por defecto del software de alineamiento. En determinadas realizaciones, el porcentaje de identidad "X" de una primera secuencia de aminoácidos respecto a una segunda secuencia de aminoácidos se calcula como 100 x (Y/Z), en el que Y es el número de residuos de aminoácidos puntuado como concordancias idénticas en el alineamiento de la primera y segunda secuencias (según se alinean por inspección visual o un programa particular de alineamiento de secuencias) y Z es el número total de residuos en la segunda secuencia. Si la segunda secuencia es más larga que la primera secuencia, entonces el porcentaje de identidad puede determinarse sólo en la región de superposición entre dicha primera y segunda secuencias. En este caso, puede usarse la misma fórmula que anteriormente pero usando como valor Z la longitud de la región en la que se superponen la primera y segunda secuencia, teniendo dicha región una longitud que es sustancialmente la misma que la longitud de la primera secuencia.

Como un ejemplo no limitativo, si cualquier polinucleótido particular tiene un determinado porcentaje de identidad de secuencia (por ejemplo, es al menos 80% idéntico, al menos 85% idéntico, al menos 90% idéntico, y en algunas realizaciones, al menos 95%, 96%, 97%, 98%, ó 99% idéntico) a una secuencia de referencia puede determinarse, en determinadas realizaciones, usando el programa Bestfit (Wisconsin Sequence Analysis Package, Versión 8 para Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711). Bestfit usa el algoritmo de homología local de Smith y Waterman, Advances in Applied Mathematics 2: 482 489 (1981), para encontrar el mejor segmento de homología entre dos secuencias. Cuando se usa Bestfit o cualquier otro programa de alineamiento de secuencias para determinar si una secuencia particular es, por ejemplo, 95% idéntica a una secuencia de referencia según la presente invención, los parámetros se ajustan de manera que el porcentaje de identidad se calcula sobre la longitud completa de la secuencia de nucleótidos de referencia y que se permiten huecos en la homología de hasta 5% del número total de nucleótidos en la secuencia de referencia.

En algunas realizaciones, dos ácidos nucleicos o polipéptidos de la invención son sustancialmente idénticos, lo que significa que tienen al menos 70%, al menos 75%, al menos 80%, al menos 85%, al menos 90%, y en algunas realizaciones al menos 95%, 96%, 97%, 98%, 99% de identidad de residuos de nucleótidos o aminoácidos, cuando se comparan y alinean para máxima correspondencia, según se mide usando un algoritmo de comparación de secuencias o por inspección visual. La identidad puede existir sobre una región de las secuencias que tiene al menos aproximadamente 10, aproximadamente 20, aproximadamente 40-60 residuos de longitud o cualquier valor de número entero entre éstos, y puede ser sobre una región más larga de 60-80 residuos, por ejemplo, al menos aproximadamente 90-100 residuos, y en algunas realizaciones, las secuencias son sustancialmente idénticas sobre la longitud completa de las secuencias que se están comparando, tal como la región codificadora de una secuencia de nucleótidos, por ejemplo.

El término "vector" significa una construcción, que es capaz de administrar, y opcionalmente expresar, uno o más genes o secuencias e interés en una célula huésped. Los ejemplos de vectores incluyen, pero no están limitados a, vectores virales, vectores de expresión de ADN o ARN desnudo, vectores de plásmido, cósmido o fago, vectores de expresión de ADN o ARN asociados con agentes de condensación catiónicos, vectores de expresión de ADN o ARN encapsulados en liposomas, y determinadas células eucariotas, tales como células productoras. Los vectores pueden ser estables y pueden ser auto-replicantes. Un "vector de expresión" es un vector que es capaz de dirigir la expresión de genes con los que está asociado de forma operativa.

"Promotor" se refiere a un fragmento de ADN capaz de controlar la expresión de una secuencia codificadora o ARN funcional. En general, una región codificadora está localizada en 3' respecto a un promotor. Los promotores pueden derivar en su totalidad de un gen nativo, o pueden estar compuestos por diferentes elementos derivados de diferentes promotores encontrados en la naturaleza, o incluso comprender segmentos de ADN sintéticos. Los expertos en la técnica entienden que diferentes promotores pueden dirigir la expresión de un gen en diferentes tejidos o tipos celulares, o en diferentes estadios de desarrollo, o en respuesta a diferentes condiciones ambientales o fisiológicas. Los promotores que causan que un gen se exprese en la mayor parte de los tipos celulares la mayor parte del tiempo se refieren comúnmente como "promotores constitutivos". Se reconoce además que ya que en la mayor parte de los casos los límites exactos de secuencias reguladoras no se han definido completamente, los fragmentos de ADN de diferentes longitudes pueden tener una actividad promotora idéntica. Un promotor está generalmente limitado en su extremo 3' por el sitio de inicio de la transcripción y se extiende aguas arriba (dirección 5') para incluir el número mínimo de bases o elementos necesarios para iniciar la transcripción a niveles detectables por encima del fondo. En el promotor se encontrará un sitio de inicio de la transcripción (definido convenientemente, por ejemplo, mediante mapeo con nucleasa S1), así como dominios de unión de proteínas (secuencias consenso) responsables de la unión de la ARN polimerasa.

El término "heterólogo" tal y como se usa en la presente memoria se refiere a un elemento de un vector, plásmido o célula huésped que deriva de una fuente distinta de la fuente endógena. Así, por ejemplo, una secuencia heteróloga (por ejemplo, una secuencia de polinucleótido o una secuencia de polipéptido) podría ser una secuencia que deriva de un gen o plásmido diferente del mismo huésped, de una cepa diferente de célula huésped, o de un organismo de un grupo taxonómico diferente (por ejemplo, diferente reino, filo, clase, orden, familia, género, o especie, o cualquier subgrupo en una de estas clasificaciones). El término "heterólogo" también se usa como sinónimo en la presente memoria del término "exógeno".

Una "región codificadora" de ADN o ARN es una molécula de ADN o ARN que se transcribe y/o traduce en un polipéptido en una célula in vitro o in vivo cuando se pone bajo el control de secuencias reguladoras apropiadas. "Regiones reguladoras adecuadas" se refiere a regiones de ácido nucleico localizadas aguas arriba (secuencias no codificadoras 5'), en, o aguas abajo (secuencias no codificadoras 3') de una región codificadora, y que influyen en la transcripción, procesamiento o estabilidad del ARN, o traducción de la región codificadora asociada. Las regiones reguladoras pueden incluir promotores, secuencias líder de la traducción, sitio de procesamiento de ARN, sitio de unión de efector y estructura tallo-bucle. Los límites de la región codificadora están determinados por un codón de inicio en el extremo 5' (amino) y un codón de parada de la traducción en el extremo 3' (carboxilo). Una región codificadora puede incluir, pero no está limitada a, regiones procariotas, ADNc de ARNm, moléculas de ADN genómico, moléculas de ADN sintético, o moléculas de ARN. Si la región codificadora se pretende para la expresión en una célula eucariota, una señal de poliadenilación y una secuencia de terminación de la transcripción estarán localizadas habitualmente en 3' respecto a la región codificadora.

"Marco de lectura abierto" se abrevia ORF y significa una longitud de ácido nucleico, bien ADN, ADNc o ARN, que comprende una señal de inicio de la traducción o codón de inicio, tal como un ATG o AUG, y un codón de terminación y puede traducirse potencialmente en una secuencia de polipéptido.

Una región codificadora está "bajo el control" de elementos de control de la transcripción y traducción en una célula cuando la ARN polimerasa transcribe la región codificadora en ARNm, que entonces se somete a corte y empalme de ARN en trans (si la región codificadora contiene intrones) y se traduce en la proteína codificada por la región codificadora.

"Regiones de control de la transcripción y traducción" son regiones reguladoras de ADN, tales como promotores, potenciadores, terminadores, y semejantes, que proporcionan la expresión de una región codificadora en una célula huésped. En células eucariotas, las señales de poliadenilación son regiones de control.

Los términos "asociado de forma operativa" y "unido de forma operativa" se refieren a la asociación de dos moléculas de manera que la función de una se ve influida por la otra. Por ejemplo, un promotor está asociado de forma operativa con una región codificadora cuando es capaz de influir en la expresión de esa región codificadora (es decir, que la región codificadora está bajo el control transcripcional del promotor). Las regiones codificadoras pueden estar asociadas de forma operativa con regiones reguladoras en orientación con sentido o antisentido. Dos moléculas están "unidas de forma operativa" ya estén unidas directamente (por ejemplo, una proteína de fusión) o indirectamente (por ejemplo a través de un conector).

Tal y como se usa en la presente memoria, el término "expresión" se refiere a la transcripción de ARN (por ejemplo, ARNm) a partir de un molde de ácido nucleico y/o la traducción de ARNm en un polipéptido. El término "expresión incrementada" se pretende que incluya una alteración en la expresión génica a nivel de una producción incrementada de ARNm y/o a nivel de expresión de polipéptido, resultando generalmente en una cantidad incrementada de un producto génico o proteína. En algunos casos, "expresión incrementada" se usa indistintamente con el término "sobreexpresión" o "sobreexpresado".

II. Inteínas

Una inteína es un elemento proteico que es capaz de auto-escindirse de una proteína huésped y catalizar la ligación de las secuencias flanqueantes con un enlace peptídico. Una inteína dividida es cualquier inteína en la que el dominio N-terminal de la inteína y el dominio C-terminal de la inteína no están unidos directamente a través de un enlace peptídico. Las inteínas divididas naturales se han identificado en cianobacterias y archaea, pero las inteínas divididas también pueden crearse artificialmente separando la secuencia de una inteína en dos partes. Las inteínas divididas descritas en la presente memoria proporcionan ventajas sobre las inteínas divididas conocidas ya que funcionan sobre un intervalo de temperatura grande y en presencia de sales. También se escinden a velocidades que son más rápidas que otras inteínas divididas conocidas. Además, las inteínas divididas descritas en la presente memoria son tolerantes a variación de secuencia tanto en la inteína como en la exteína y/o secuencias de polipéptido heterólogas. Las inteínas divididas descritas en la presente memoria proporcionan ventajas sobre las inteínas divididas conocidas ya que pueden realizar la auto-escisión C-terminal independientemente del primer aminoácido de la C-Exteína.

Las inteínas divididas usadas en la presente memoria pueden comprender los seis restos de corte y empalme de proteínas conservados de la familia HINT (Hog/Inteína). Las secuencias de dichos restos conservados pueden usarse para predecir qué aminoácidos en un dominio de inteína están más estrictamente conservados y qué aminoácidos están menos estrictamente conservados. Las mutaciones de los aminoácidos más estrictamente conservados pueden reducir la eficacia de la escisión de la inteína.

Un "dominio N-terminal de inteína" se refiere a una secuencia de inteína que comprende una secuencia de aminoácidos N-terminal que es funcional para reacciones de corte y empalme en trans y/o reacciones de auto-escisión N-terminal. Un dominio N-terminal de inteína puede retirarse por corte y empalme cuando ocurre el corte y empalme en trans. Los ensayos adecuados para determinar si una secuencia de inteína es un dominio N-terminal pueden encontrarse, por ejemplo, en el ejemplo 1 de la presente invención, que proporciona un ensayo para medir la actividad de corte y empalme en trans o en el ejemplo 6, que proporciona un ensayo para detectar la auto-escisión N-terminal.

El dominio N-terminal de inteína puede comprender uno o más de los restos N1, N2, N3, y/o N4 de la familia HINT (Hog/Inteína). Así, por ejemplo, un dominio N-terminal de inteína puede comprender los restos N1 y N3.

En algunas realizaciones, el dominio N-terminal de inteína comprende una secuencia de caja N1 (caja A). La caja N1 es una secuencia no estrictamente conservada. La caja N1 puede comprender, por ejemplo, la secuencia ChsXcplhXTXXG (SEQ ID NO:44), en la que h es un aminoácido hidrofóbico, s es un aminoácido pequeño, c es un aminoácido cargado, p es un aminoácido polar, y 1 es un aminoácido grande. En algunas realizaciones, el dominio N-terminal de inteína comprende la secuencia $X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10}X_{11}X_{12}X_{13}$ (SEQ ID NO:45), en la que X_1 es C; X_2 es L, F, o V; X_3 es S, T, V, o A; X_4 es L, P, G, o Y; X_5 es D, E, K, o G; X_6 es T o A; X_7 es E, Q, L, M, K, o T; X_8 es I o V; X_9 es L, Q, V, N, K, D, o T; X_{10} es T, I, o V; X_{11} es V, P, Q, N, E, K, o L; X_{12} es E, Q, G, N, Y, I, o E; y X_{13} es Y, G, K, P, o D. En algunas realizaciones, el dominio N-terminal de inteína comprende la secuencia $X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10}X_{11}X_{12}X_{13}$ (SEQ ID NO:46), en la que X_1 es C; X_2 es L, F, o V; X_3 es S, T, V, o A; X_4 es L, P, o G; X_5 es D, K, o G; X_6 es T o A; X_7 es Q, L, M, K, o T; X_8 es I o V; X_9 es Q, V, N, K, D, o T; X_{10} es T, I, o V; X_{11} es P, Q, N, E, K, o L; X_{12} es E, Q, G, N, Y, I, o E, y X_{13} es G, K, P, o D.

Tomando como base las propiedades químicas de los aminoácidos, pueden agruparse como: (i) cargados (D, E, K, R, H), (ii) ácidos (D, E), (iii) básicos (K, R, H), (iv) pequeños (V, C, S, T, P, G, D, A), (v) polares (N,Q, S, T), (vi) grandes (E, Q, R, K, H, Y, W, F, M, L, I) , (vii) hidrofóbicos (V, I, L, M, F, Y, W, A) y (viii) nucleofílicos (S, T, C).

En algunas realizaciones, el dominio N-terminal de inteína caja N1 comprende una secuencia que es al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los aminoácidos 1 a 13 de SEQ ID NO:3.

En algunas realizaciones, un dominio N-terminal de inteína caja N1 comprende una secuencia que es al menos aproximadamente 30%, al menos aproximadamente 35%, al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los aminoácidos 1 a 13 de SEQ ID NO:12.

En algunas realizaciones, un dominio C-terminal de inteína caja N1 comprende una secuencia que es al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los aminoácidos 1 a 13 de SEQ ID NO:34.

En algunas realizaciones, el dominio N-terminal de inteína caja N1 comprende una secuencia que es al menos aproximadamente 30%, al menos aproximadamente 35%, al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los aminoácidos 1 a 13 de SEQ ID NO:64.

En algunas realizaciones, un dominio N-terminal de inteína caja N1 comprende una secuencia que es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los aminoácidos 1 a 13 de SEQ ID NO:20.

En algunas realizaciones, el dominio N-terminal de inteína comprende una secuencia de caja N3 (caja B). La caja N3 es una secuencia no estrictamente conservada. La caja N3 puede comprender, por ejemplo, la secuencia GXXhXhTXaHXhhTX (SEQ ID NO:47), en la que h es un aminoácido hidrofóbico y a es un aminoácido ácido. En algunas realizaciones, el dominio N-terminal de inteína comprende la secuencia $X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10}X_{11}X_{12}X_{13}$ X_{14} X_{15} (SEQ ID NO:48), en la que X_1 es G o A; X_2 es S, K, Q, N, o F; X_3 es L, E, K, o R; X_4 es I, L, o V; X_5 es R, I, V, o N; X_6 es A, C, V, o E; X_7 es T, S, o D; X_8 es K, E, A, P, o N; X_9 es D, E, N, o I; X_{10} es H; X_{11} es K, L, Q, o M; X_{12} es F, V, o I; X_{13} es M, P, F, Y, o A; X_{14} es T; y X_{15} es V, Q, K, o L. En algunas realizaciones, el dominio N-terminal de inteína comprende la secuencia $X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10}X_{11}X_{12}X_{13}X_{14}X_{15}$ (SEQ ID NO:49), en la que X_1 es G o A; X_2 es K, Q, N, o F; X_3 es E, K, o R; X_4 es I, L, o V; X_5 es R, I, V, o N; X_6 es C, V, o E; X_7 es T, S, o D; X_8 es E, A, P, o N; X_9 es D, E, N, o I; X_{10} es H; X_{11} es K, L, Q, o M; X_{12} es F, V, o I; X_{13} es P, F, Y, o A; X_{14} es T; y X_{15} es Q, K, o L.

El primer aminoácido en un dominio N-terminal de inteína está altamente conservado y es importante para la reacción de corte y empalme de proteínas. Por lo tanto, en algunas realizaciones, el primer aminoácido en un dominio N-terminal de inteína es una cisteína. En algunas realizaciones, el primer aminoácido en un dominio N-terminal de inteína es una serina. En otras realizaciones, el primer aminoácido en un dominio N-terminal de inteína puede mutarse a un aminoácido que evita o disminuya la escisión entre un polipéptido heterólogo o N-exteína y la inteína. Así, en algunas realizaciones, el primer aminoácido en un dominio N-terminal de inteína es un aminoácido distinto de serina o cisteína. Por ejemplo, el primer aminoácido en un dominio N-terminal de inteína puede ser una alanina.

En algunas realizaciones, el dominio N-terminal de inteína es aproximadamente 50 a aproximadamente 150 aminoácidos. En algunas realizaciones, el dominio N-terminal de inteína es aproximadamente 60 a aproximadamente 140 aminoácidos. En algunas realizaciones, el dominio N-terminal de inteína es aproximadamente 75 a aproximadamente 125 aminoácidos. En algunas realizaciones, el dominio N-terminal de inteína es aproximadamente 70 a aproximadamente 80, aproximadamente 80 a aproximadamente 90, aproximadamente 90 a aproximadamente 100, aproximadamente 100 a aproximadamente 110, aproximadamente 110 a aproximadamente 120, o aproximadamente 120 a aproximadamente 130 aminoácidos.

En algunas realizaciones, un dominio N-terminal de inteína comprende los aminoácidos de una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64.

En algunas realizaciones, un dominio N-terminal de inteína comprende una secuencia que es al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos

aproximadamente 95%, al menos aproximadamente 96%, al menos aproximadamente 97%, al menos aproximadamente 98%, o al menos aproximadamente 99% idéntica a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64.

En alguna realización, el dominio N-terminal de inteína comprende la secuencia correspondiente al dominio N-terminal de gp41-1 (SEQ ID NO:79), gp41-2 (SEQ ID NO:80), gp41-3 (SEQ ID NO:81), gp41-4 (SEQ ID NO:82), gp41-5 (SEQ ID NO:83), gp41-6 (SEQ ID NO:84), gp41-7 (SEQ ID NO:85), gp41-8 (SEQ ID NO:86), IMPDH-1 (SEQ ID NO:87), NrdA-1 (SEQ ID NO:88), NrdA-2 (SEQ ID NO:89), NrdA-4 (SEQ ID NO:90), NrdA-5 (SEQ ID NO:91), NrdA-6 (SEQ ID NO:92), NrdJ-1 (SEQ ID NO:93) y NrdJ-2 (SEQ ID NO:94).

En algunas realizaciones, un dominio N-terminal de inteína contiene al menos aproximadamente 10, al menos aproximadamente 20, al menos aproximadamente 30, al menos aproximadamente 40, o al menos aproximadamente 50 aminoácidos de una secuencia seleccionada del grupo que consiste en SEQ ID NO: 3, 12, 20, 34 y 64. En algunas realizaciones, un dominio N-terminal de inteína contiene al menos aproximadamente 10, al menos aproximadamente 20, al menos aproximadamente 30, al menos aproximadamente 40, o al menos aproximadamente 50 aminoácidos consecutivos de una secuencia seleccionada del grupo que consiste en SEQ ID NO: 3, 12, 20, 34 y 64. En algunas realizaciones, un dominio N-terminal de inteína contiene una deleción de no más de aproximadamente 5, aproximadamente 10, aproximadamente 15, aproximadamente 20, o aproximadamente 25 aminoácidos de una secuencia seleccionada del grupo que consiste en SEQ ID NO: 3, 12, 20, 34 y 64. En algunas realizaciones, un dominio N-terminal de inteína contiene una deleción de no más de aproximadamente 5, aproximadamente 10, aproximadamente 15, aproximadamente 20, o aproximadamente 25 aminoácidos consecutivos de una secuencia seleccionada del grupo que consiste en SEQ ID NO: 3, 12, 20, 34 y 64.

En algunas realizaciones, un dominio N-terminal de inteína comprende una secuencia que es al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO:3.

En algunas realizaciones, un dominio N-terminal de inteína comprende una secuencia que es al menos aproximadamente 30%, al menos aproximadamente 35%, al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO:12.

En algunas realizaciones, un dominio N-terminal de inteína comprende una secuencia que es al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO: 20.

En algunas realizaciones, un dominio N-terminal de inteína comprende una secuencia que es al menos aproximadamente 30%, al menos aproximadamente 35%, al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO: 64.

En algunas realizaciones, un dominio N-terminal de inteína comprende una secuencia que es al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO: 34.

Un "dominio C-terminal de inteína" se refiere a una secuencia de inteína que comprende una secuencia de aminoácidos C-terminal que es funcional para reacciones de corte y empalme en trans y/o reacciones de auto-escisión C-terminal. Un dominio C-terminal de inteína puede retirarse por corte y empalme cuando ocurre el corte y empalme en trans. Los ensayos adecuados para determinar si un polipéptido dado es un dominio C-terminal de inteína pueden encontrarse en el ejemplo 1 de la presente invención, que proporciona un ensayo para medir la actividad de corte y empalme en trans en presencia de un dominio N-terminal de inteína funcional o en el ejemplo 5, que proporciona un ensayo para detectar la auto-escisión C-terminal de una proteína de fusión que comprende una inteína C-terminal y un polipéptido heterólogo en presencia de un dominio N-terminal de inteína que porta una mutación en el primer residuo de cisteína.

El dominio C-terminal de inteína puede comprender los restos C1 y/o C2 de la familia HINT (Hog/Inteína).

En algunas realizaciones, el dominio C-terminal de inteína comprende una secuencia de caja C2 (caja F). La caja C2 es una secuencia no estrictamente conservada. La caja C2 puede comprender, por ejemplo, la secuencia

XhhDIpVXXpHXFX (SEQ ID NO: 50), en la que h es un aminoácido hidrofóbico y p es un aminoácido polar. En algunas realizaciones, el dominio C-terminal de inteína comprende la secuencia $X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10}X_{11}X_{12}X_{13}X_{14}$ (SEQ ID NO:51), en la que X_1 es N, E, L, K, Q, D, P, o R; X_2 es V, L, o T; X_3 es Y, I, V, H, o F; X_4 es D; X_5 es I o L; X_6 es G, E, T, Q, o K; X_7 es V o T; X_8 es E, S, T, D, N, o K; X_9 es R, G, D, N, Q, S, o K; X_{10} es D, E, N, T, o K; X_{11} es H, R, S, I, o N; X_{12} es N, L, S, I, o N; X_{13} es F, Y, L, o I; y X_{14} es A, Y, F, N, C, o S. En algunas realizaciones, el dominio C-terminal de inteína comprende la secuencia $X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10}X_{11}X_{12}X_{13}$ X_{14}(SEQ ID NO:52), en la que X_1 es E, L, K, Q, D, P, o R; X_2 es V, L, o T; X_3 es Y, I, V, H, o F; X_4 es D; X_5 es I o L; X_6 es G, E, T, Q, o K; X_7 es V o T; X_8 es E, S, T, D, N, o K; X_9 es G, D, N, Q, S, o K; X_{10} es D, E, N, T, o K; X_{11} es H, R, S, I, o N; X_{12} es N, L, S, I, o N; X_{13} es F, Y, L, o I; y X_{14} es A, Y, F, N, C, o S.

En algunas realizaciones, el dominio C-terminal de inteína comprende parte de la secuencia de la caja C1 (caja G). La caja C1 es una secuencia no estrictamente conservada. La caja C1 puede comprender, por ejemplo, la secuencia hNXIhXHNn (SEQ ID NO: 53), en la que h es un aminoácido hidrofóbico y n es un aminoácido nucleofílico. En algunas realizaciones, el dominio C-terminal de inteína comprende la secuencia $X_1X_2X_3X_4X_5X_6X_7X_8X_9$ (SEQ ID NO: 54), en la que X_1 es L, A, V, I, o C; X_2 es N o R; X_3 es G, D, A, o N; X_4 es I, F, o T; X_5 es L, I, o V; X_6 es V, I, T, o A; X_7 es H o S; X_8 es N; y X_9 es S, T, o C. En algunas realizaciones, el dominio C-terminal de inteína comprende la secuencia $X_1X_2X_3X_4X_5X_6X_7X_8X_9$ (SEQ ID NO:55), en la que X_1 es A, V, I, o C; X_2 es N o R; X_3 es G, D, A, o N; X_4 es I, F, o T; X_5 es L o V; X_6 es V, I, o T; X_7 es H; X_8 es N; y X_9 es S, T, o C. En la secuencia de la caja C1, los aminoácidos de X_1 a X_8 corresponden a la secuencia de inteína, y X_9 corresponde al primer aminoácido de la exteína.

En algunas realizaciones, un dominio C-terminal de inteína de caja C1 comprende una secuencia que es al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los últimos 8 aminoácidos de SEQ ID NO:7.

En algunas realizaciones, un dominio C-terminal de inteína de caja C1 comprende una secuencia que es al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los últimos 8 aminoácidos de SEQ ID NO:16.

En algunas realizaciones, un dominio C-terminal de inteína de caja C1 comprende una secuencia que es al menos aproximadamente 45%, al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 65%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los últimos 8 aminoácidos de SEQ ID NO: 38.

En algunas realizaciones, un dominio C-terminal de inteína de caja C1 comprende una secuencia que es al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los últimos 8 aminoácidos de SEQ ID NO:65.

En algunas realizaciones, un dominio C-terminal de inteína de caja C1 comprende una secuencia que es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia correspondiente a los últimos 8 aminoácidos de SEQ ID NO:24.

Los últimos dos aminoácidos en un dominio C-terminal de inteína están altamente conservados y son importantes para la reacción de corte y empalme de la proteína. Por lo tanto, en algunas realizaciones, el último aminoácido en un dominio C-terminal de inteína es una asparagina. En algunas realizaciones, el último aminoácido en un dominio C-terminal de inteína es una glutamina. En algunas realizaciones, el penúltimo aminoácido en un dominio C-terminal de inteína es una histidina. En otras realizaciones, el último y/o penúltimo aminoácido en un dominio C-terminal de inteína puede mutarse a un aminoácido que evita o disminuya la escisión entre un polipéptido heterólogo o exteína y la inteína. Así, en algunas realizaciones, el último aminoácido en un dominio C-terminal de inteína es un aminoácido distinto de asparagina o glutamina. En algunas realizaciones, el penúltimo aminoácido en un dominio C-terminal de inteína es un aminoácido distinto de histidina. En algunas realizaciones, el último aminoácido en un dominio C-terminal de inteína es un aminoácido distinto de asparagina o glutamina y el primer aminoácido en un dominio C-terminal de exteína es un aminoácido distinto de serina. Por ejemplo, el último aminoácido en un dominio C-terminal de inteína y/o el primer aminoácido en un dominio C-terminal de exteína puede ser una alanina.

En algunas realizaciones, el dominio C-terminal de inteína es aproximadamente 10 a aproximadamente 80 aminoácidos. En algunas realizaciones, el dominio C-terminal de inteína es aproximadamente 20 a aproximadamente 70 aminoácidos. En algunas realizaciones, el dominio C-terminal de inteína es aproximadamente 30 a aproximadamente 60 aminoácidos. En algunas realizaciones, el dominio C-terminal de inteína es aproximadamente 25 a aproximadamente 35, aproximadamente 30 a aproximadamente 40, aproximadamente 35 a

aproximadamente 45, aproximadamente 40 a aproximadamente 50, aproximadamente 45 a aproximadamente 55, o aproximadamente 55 a aproximadamente 65 aminoácidos.

En algunas realizaciones, un dominio C-terminal de inteína comprende los aminoácidos de una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65. En algunas realizaciones, un dominio C-terminal de inteína comprende una secuencia que es al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, al menos aproximadamente 96%, al menos aproximadamente 97%, al menos aproximadamente 98%, o al menos aproximadamente 99% idéntica a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65. En algunas realizaciones, un dominio C-terminal de inteína contiene al menos aproximadamente 10, al menos aproximadamente 20, al menos aproximadamente 30, al menos aproximadamente 40, o al menos aproximadamente 50 aminoácidos de una secuencia seleccionada del grupo que consiste en SEQ ID NO: 7, 16, 24, 38 y 65. En algunas realizaciones, un dominio C-terminal de inteína contiene al menos aproximadamente 10, al menos aproximadamente 20, al menos aproximadamente 30, al menos aproximadamente 40, o al menos aproximadamente 50 aminoácidos consecutivos de una secuencia seleccionada del grupo que consiste en SEQ ID NO: 7, 16, 24, 38 y 65. En algunas realizaciones, un dominio C-terminal de inteína contiene una deleción de no más de aproximadamente 5, aproximadamente 10, aproximadamente 15, aproximadamente 20, o aproximadamente 25 aminoácidos de una secuencia seleccionada del grupo que consiste en SEQ ID NO: 7, 16, 24, 38 y 65. En algunas realizaciones, un dominio C-terminal de inteína contiene una deleción de no más de aproximadamente 5, aproximadamente 10, aproximadamente 15, aproximadamente 20, o aproximadamente 25 aminoácidos consecutivos de una secuencia seleccionada del grupo que consiste en SEQ ID NO: 7, 16, 24, 38 y 65.

En algunas realizaciones, un dominio C-terminal de inteína comprende una secuencia que es al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO: 7.

En algunas realizaciones, un dominio C-terminal de inteína comprende una secuencia que es al menos aproximadamente 30%, al menos aproximadamente 35%, al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO:16.

En algunas realizaciones, un dominio C-terminal de inteína comprende una secuencia que es al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO:38.

En algunas realizaciones, un dominio C-terminal de inteína comprende una secuencia que es al menos aproximadamente 30%, al menos aproximadamente 35%, al menos aproximadamente 40%, al menos aproximadamente 45%, es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO:65.

En algunas realizaciones, un dominio C-terminal de inteína comprende una secuencia que es al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a una secuencia de SEQ ID NO:24.

En alguna realización, el dominio C-terminal de inteína comprende la secuencia correspondiente al dominio C-terminal de gp41-1 (SEQ ID NO:95), gp41-2 (SEQ ID NO:96), gp41-3 (SEQ ID NO:97), gp41-8 (SEQ ID NO:98), gp41-8 (SEQ ID NO:99), IMPDH-1 (SEQ ID NO:100), IMPDH-2 (SEQ ID NO:101), IMPDH-3 (SEQ ID NO:102) NrdA-2 (SEQ ID NO:103) NrdA-3 (SEQ ID NO:104), NrdA-5 (SEQ ID NO:105), NrdA-6 (SEQ ID NO:106) NrdA-7 (SEQ ID NO:107), NrdJ-1 (SEQ ID NO:108).

En algunas realizaciones, el dominio N-terminal de inteína y el dominio C-terminal de inteína tienen cargas opuestas. Así, en algunas realizaciones, el dominio N-terminal de inteína está cargado negativamente, y el dominio C-terminal de inteína está cargado positivamente. En otras realizaciones, el dominio N-terminal de inteína está cargado positivamente, y el dominio C-terminal de inteína está cargado negativamente.

Tabla 1: Secuencias de los dominios N- y C-terminales de las inteínas usadas en la presente invención. Las secuencias subrayadas corresponden a las cajas N1 de los dominios de inteína N-terminales. Las secuencias con

doble subrayado corresponden a las cajas C1 de los dominios de inteína C-terminales (que carecen del primer aminoácido de la exteína).

Dominio de inteína	SEQ ID NO:	Secuencia
Dominio N-terminal de GP41.1	3	1 CLDLKTQVQT PQGMKEISNI QVGDLVLSNT GYNEVLNVFP KSKKKSYKIT LEDGKEIICS 61 EEHLFPTQTG EMNISGGLKE GMCLYVKE
Dominio N-terminal de GP41.8	12	1 CLSLDTMVVT NGKAIEIRDV KVGDWLESEC GPVQVTEVLP IIKQPVFEIV LKSGKKIRVS 61 ANHKFPTKDG LKTINSGLKV GDFLRSRA
Dominio N-terminal de NrdJ1	20	1 CLVGSSEIIT RNYGKTTIKE VVEIFDNDKN IQVLAFNTHT DNIEWAPIKA AQLTRPNAEL 61 VELEINTLHG VKTIRCTPDH PVYTKNRDYV RADELTDDDE LVVAI
Dominio N-terminal de IMPDH1	34	1 CFVPGTLVNT ENGLKKIEEI KVGDKVFSHT GKLQEVVDTL IFDRDEEIIS INGIDCTKNH 61 EFYVIDKENA NRVNEDNIHL FARWVHAEEL DMKKHLLIEL E
Dominio N-terminal de NrdA-2	64	1 CLTGDAKIDV LIDNIPISQI SLEEVVNLFN EGKEIYVLSY NIDTKEVEYK EISDAGLISE 61 SAEVLEIIDE ETGQKIVCTP DHKVYTLNRG YVSAKDLKED DELVFS
Dominio N-terminal de DNA-E	28	1 CLSYETEILT VEYGLLPIGK IVEKRIECTV YSVDNNGNIY TQPVAQWHDR GEQEVFEYCL 61 EDGSLIRATK DHKFMTVDGQ MLPIDEIFER ELDLMRVDNL PN
Dominio C-terminal de GP41.1	7	1 MMLKKILKIE ELDERELIDI EVSGNHLFYA NDILTHN
Dominio C-terminal de GP41.8	7,3	1 MCEIFENEID WDEIASIEYV GVEETIDINV TNDRLFFANG ILTHN
Dominio C-terminal de NrdJ1	24	1 MEAKTYIGKL KSRKIVSNED TYDIQTSTHN FFANDILVHN
Dominio C-terminal de IMPDH1	17,2	1 MKFKLKEITS IETKHYKGKV HDLTVNQDHS YNVRGTVVHN
Dominio C-terminal de NrdA-2	65	1 MGLKIIKRES KEPVFDITVK DNSNFFANNI LVHN
Dominio C-terminal de DNA-E	31	1 MIKIATRKYL GKQNVYDIGV ERDHNFALKN GFIASN

Entre las varias cajas identificadas en secuencias de proteínas inteínas e inteínas divididas (N1, N2, C1 y C2), C1 es la caja más conservada y está implicada directamente en la reacción de corte y empalme en trans. El papel central de C1 se considera una característica importante en la clasificación y agrupamiento de las inteínas divididas.

En algunas realizaciones, un dominio C-terminal de inteína contiene una caja C1 que es al menos aproximadamente 60%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95%, o al menos aproximadamente 99% idéntica a la secuencia ANDILTHNS (SEQ ID NO:78), correspondiente a la caja C1 del dominio C-terminal de la inteína dividida gp-41-1.

Como para el resto del documento, la identidad se calcula como el porcentaje de aminoácidos idénticos comparado con el número total de aminoácidos, cuando

Nombre la inteína dividida	% de identidad frente a la caja C1 de gp41-1
gp41-2	87,5
gp41-8	87,5
gp41-9	100
NrdA2	75
NrdA3	87,5
NrdA6	62,5
NrdA7	87,5
NrdJ1	87,5
Npu DNA-E	25

Tabla 2: Porcentaje de identidad de las cajas C1 (también conocidas como cajas G) de varias inteínas divididas respecto a la caja C1 de gp41-1 (ANDILTHNS, SEQ ID NO:78). Como para el resto de la patente, la identidad se calcula como el porcentaje de aminoácidos idénticos respecto al número total de aminoácidos. Para propósitos de claridad, la identidad se calcula entre dos secuencias que tienen la misma longitud. La inteína DNA-E conocida en la técnica anterior se diferencia del resto de las inteínas en que la caja C1 no muestra una identidad sustancial con la caja C1 de gp41-1.

III. Proteínas de fusión de inteínas

En la presente memoria también se describen proteínas de fusión que comprenden inteínas divididas. El dominio N-terminal de inteína y/o el dominio C-terminal de inteína pueden fusionarse bien directamente (es decir, a través de un enlace peptídico) o indirectamente (es decir, a través de una secuencia de aminoácidos conectora) a un polipéptido heterólogo.

Así, en algunas realizaciones, un polipéptido heterólogo se fusiona bien directamente o indirectamente al extremo N de un dominio N-terminal de inteína. Dichos polipéptidos también pueden comprender opcionalmente aminoácidos o polipéptidos heterólogos adicionales fusionados bien directamente o indirectamente al extremo C del dominio N-terminal de inteína (por ejemplo, etiquetas de expresión o purificación) o fusionados directamente o indirectamente al extremo N del polipéptido heterólogo.

En algunas realizaciones, un polipéptido heterólogo se fusiona bien directamente o indirectamente al extremo C de un dominio C-terminal de inteína. Dichos polipéptidos también pueden comprender opcionalmente aminoácidos o polipéptidos heterólogos adicionales fusionados bien directamente o indirectamente al extremo N del dominio C-terminal de inteína (por ejemplo, etiquetas de expresión o purificación) o fusionados directamente o indirectamente al extremo C del polipéptido heterólogo.

En algunas realizaciones, una proteína de fusión que comprende un polipéptido heterólogo fusionado al extremo C de un dominio C-terminal de inteína y una proteína de fusión que comprende un polipéptido heterólogo fusionado al extremo N de un dominio N-terminal de inteína se expresan como dos polipéptidos separados.

En algunas realizaciones, una proteína de fusión que comprende un polipéptido heterólogo fusionado al extremo C de un dominio C-terminal de inteína y una proteína de fusión que comprende un polipéptido heterólogo fusionado al extremo N de un dominio N-terminal de inteína se expresan como un único polipéptido. La proteína de fusión que comprende un polipéptido heterólogo fusionado al extremo C de un dominio C-terminal de inteína puede separarse de la proteína de fusión que comprende un polipéptido heterólogo fusionado al extremo N de un dominio N-terminal de inteína por aproximadamente 1 a aproximadamente 1.000, aproximadamente 1 a aproximadamente 500, aproximadamente 1 a aproximadamente 250, aproximadamente 1 a aproximadamente 200, aproximadamente 1 a aproximadamente 150, aproximadamente 1 a aproximadamente 100, o aproximadamente 1 a aproximadamente 50 aminoácidos.

En algunas realizaciones, una proteína de fusión comprende un dominio C-terminal de inteína fusionado al extremo N de un polipéptido heterólogo. En una realización preferida, la proteína de fusión comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y (ii) un polipéptido heterólogo, en el que el polipéptido heterólogo es C-terminal respecto al dominio de inteína. En una realización aún más preferida, el último aminoácido del dominio de inteína es glutamina o asparagina. En otra realización más, el primer aminoácido del polipéptido heterólogo se selecciona del grupo que consiste en Met, Cys, Thr, Arg, Lys, Ser, Gin, His, Ala, Tyr, Phe, Asn, Trp, Val, Leu, Asp, Ile, Gly, Glu o Pro. En otra realización, el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina y el primer aminoácido del polipéptido heterólogo se selecciona del grupo que consiste en Met, Cys, Thr, Arg, Lys, Ser, Gln, His, Ala, Tyr, Phe, Asn, Trp, Val, Leu, Asp, Ile, Gly, Glu o Pro. En otra realización, el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina y el primer aminoácido del polipéptido heterólogo es un aminoácido distinto de serina, cisteína o treonina.

En algunas realizaciones, una proteína de fusión comprende un dominio N-terminal de inteína fusionado al extremo C de un polipéptido heterólogo. En otra realización, la proteína de fusión comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y (ii) un polipéptido heterólogo, en el que el polipéptido heterólogo es N-terminal respecto al dominio de inteína. En una realización preferida, el primer aminoácido del dominio de inteína es una serina o cisteína. En otra realización más, el primer aminoácido del dominio de inteína es un aminoácido distinto de serina o cisteína.

El polipéptido heterólogo puede ser, por ejemplo, una enzima, una hormona, tal como calcitonina, eritropoyetina, trombopoyetina, hormona de crecimiento humana, factor de crecimiento epidérmico, y semejantes, un interferón, una citoquina, una proteína que tiene uso terapéutico, nutracéutico, agrícola, o industrial. Los polipéptidos heterólogos adicionales pueden ser enzimas, anticuerpos, fragmentos de anticuerpo, y proteínas farmacéuticas. Un polipéptido heterólogo también puede ser un fragmento de polipéptido.

El polipéptido heterólogo también puede ser, por ejemplo, una cadena de anticuerpo, anticuerpos de dominio único, anticuerpo de cadena pesada de camélido (VHH o nanocuerpos), o un anticuerpo recombinante desarrollado usando combinaciones de dominios de anticuerpos, tal como formatos monovalente (fragmento variable (Fv), fragmento de anticuerpo Fv estabilizado por disulfuro (dsFv), scFv, fragmento de anticuerpo de cadena única (scAb) y Fab), divalente (minicuerpo, fragmento divalente, $F(ab')_2$ y $(scFv)_2$) y multivalente (tetracuerpo, triacuerpo y $F(ab')_3$) (Figura 3 de Vijayalakshmi B et al. Methods Volumen 56, Número 2, febrero 2012, 116-129).

En algunas realizaciones, el primer aminoácido del polipéptido heterólogo es una serina, cisteína, o treonina. En algunas realizaciones, el primer aminoácido del polipéptido heterólogo no es una serina, cisteína, o treonina.

En algunas realizaciones, una proteína de fusión que comprende un polipéptido heterólogo y una inteína o dominio de inteína comprende además secuencias adicionales tales como etiquetas de purificación o etiquetas de expresión. Dichas etiquetas de expresión y/o purificación incluyen, por ejemplo, etiquetas Strep, His, y Myc.

En algunas realizaciones, la proteína de fusión comprende además una secuencia que incrementa la solubilidad de la proteína, por ejemplo, una proteína D de cabeza de fago bacteriófago lambda (gpD), tiorredoxina (Tx) o GST.

En algunas realizaciones, una proteína de fusión que comprende un polipéptido heterólogo y un dominio de inteína N- y/o C-terminal puede comprender un resto químico adicional que incluye, entre otros, grupos fluorescentes, biotina, polietilen glicol (PEG), análogos de aminoácidos, aminoácidos no naturales, grupos fosfato, grupos glicosilo, marcadores radioisótopos, y moléculas farmacéuticas. En otras realizaciones, el polipéptido heterólogo puede comprender uno o más grupos químicamente reactivos que incluyen, entre otros, cetona, aldehído, residuos de Cys y residuos de Lys.

En algunas realizaciones, la proteína de fusión comprende un conector entre el polipéptido heterólogo y la secuencia de inteína. Así, la proteína de fusión puede comprender un conector entre el extremo C de la proteína heteróloga y el extremo N del dominio N-terminal de la inteína. La proteína de fusión también puede comprender un conector entre el extremo N de la proteína heteróloga y el extremo C del dominio C-terminal de la inteína. El conector, puede tener una longitud, por ejemplo, de 1-10 aminoácidos. El conector puede tener una longitud de 1-5 aminoácidos. Así, el conector puede contener 1, 2, 3, 4, ó 5 aminoácidos. En algunas realizaciones, el conector puede comprender una secuencia de exteína.

En algunas realizaciones, el primer aminoácido del conector que contacta el polipéptido heterólogo y el extremo C del dominio C-terminal de una inteína se selecciona del grupo que consiste en Met, Cys, Thr, Arg, Lys, Ser, Gln, His, Ala, Tyr, Phe, Asn, Trp, Val, Leu, Asp, Ile, Gly, Glu o Pro. En otra realización más, el primer aminoácido del conector que contacta el polipéptido heterólogo y el extremo C del dominio C-terminal de una inteína puede comprender una serina, cisteína, o treonina. Una serina, cisteína, o treonina adyacente al extremo C del dominio C-terminal de una inteína puede incrementar la eficiencia de la escisión C-terminal (es decir, la escisión entre el dominio C-terminal de inteína y la serina, cisteína, o treonina del polipéptido heterólogo). En algunas realizaciones, el primer, segundo, tercer, cuarto, y/o quinto aminoácido del conector es una serina, cisteína, o treonina.

En algunas realizaciones, el conector puede comprender una secuencia de exteína nativa. Tal y como se usa en la presente memoria, el término "exteína" se refiere a la secuencia que se encuentra naturalmente próxima a una inteína o dominio de inteína. Así, un polipéptido heterólogo, que es un polipéptido que no se encuentra naturalmente próximo a una inteína o dominio de inteína, no es una exteína. En algunas realizaciones, la exteína comprende una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 4, 8, 13, 17, 21, 25, 35, y 39. En algunas realizaciones, un conector que comprende aminoácidos de una exteína comprende, por ejemplo, los primeros (es decir, N-terminales) 1-5 aminoácidos de una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 4, 8, 13, 17, 21, 25, 35, y 39. En algunas realizaciones, el conector comprende 1, 2, 3, 4, ó 5 aminoácidos de una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 4, 8, 13, 17, 21, 25, 35, y 39. En algunas realizaciones, una proteína de fusión comprende un dominio de inteína y un dominio de exteína que se encuentran juntos naturalmente. En otras realizaciones, una proteína de fusión comprende un dominio de inteína y un dominio de exteína que no se encuentran juntos naturalmente, es decir, un dominio de exteína heterólogo. Como ejemplo, una proteína de fusión puede comprender un dominio de inteína Gp41.1 y un dominio de exteína heterólogo tal como un dominio de exteína IMPDH.

III. Polinucleótidos que codifican fusiones de inteína y expresión de fusiones de inteína

En la presente memoria también se describen polinucleótidos que codifican fusiones de inteína. Los polinucleótidos pueden estar en la forma de ARN o ADN. El ADN incluye ADNc, ADN genómico, y ADN sintético; y puede ser bicatenario o monocatenario, y si es monocatenario puede ser la cadena codificadora o no codificadora (anti-sentido). En determinadas realizaciones, los polinucleótidos se aíslan. En determinadas realizaciones, los polinucleótidos son sustancialmente puros.

Dichos polinucleótidos pueden, por ejemplo, incorporarse en un vector de expresión para producir proteínas de fusión de inteína. Los vectores de expresión son construcciones de ADN replicables que tienen fragmentos derivados de ADN sintético o ADNc que codifican una proteína de fusión de inteína, unidos de forma operativa a elementos reguladores de la transcripción o traducción adecuados. Los elementos reguladores de la transcripción o traducción pueden derivar, por ejemplo, de genes de mamífero, microbianos, virales, o de insecto. Una unidad transcripcional comprende generalmente un ensamblaje de (1) un elemento o elementos genéticos que tienen un papel regulador en la expresión génica, por ejemplo, promotores o potenciadores de la transcripción, (2) una secuencia estructural o codificadora que se transcribe en ARNm y se traduce en proteína, y (3) secuencias de inicio y terminación de la transcripción y traducción apropiadas, como se describe con detalle más adelante. Dichos elementos reguladores pueden incluir una secuencia operadora para controlar la transcripción. La capacidad de replicarse en un huésped, habitualmente conferida por un origen de replicación, y un gen de selección para facilitar el reconocimiento de transformantes pueden incorporarse adicionalmente. Las regiones de ADN se unen de forma operativa cuando están funcionalmente relacionadas entre sí. Por ejemplo, el ADN para un péptido señal está unido de forma operativa a ADN para un polipéptido si se expresa como un precursor que participa en la secreción del polipéptido; un promotor está unido de forma operativa a una secuencia codificadora si controla la transcripción de la secuencia; o un sitio de unión a ribosoma está unido de forma operativa a una secuencia codificadora si está posicionado de manera que permita la traducción.

La elección de la secuencia de control de la expresión y el vector de expresión dependerá de la elección del huésped. Puede emplearse una amplia variedad de combinaciones huésped/vector de expresión. Los vectores de expresión útiles para huéspedes eucariotas, incluye, por ejemplo, vectores que comprenden secuencias de control de la expresión de SV40, virus de papiloma bovino, adenovirus y citomegalovirus. Los vectores de expresión útiles para huéspedes bacterianos incluyen plásmidos bacterianos conocidos, tales como plásmidos de *Esherichia coli,* incluyendo pCR 1, pBR322, pMB9 y sus derivados, plásmidos con un rango de huésped más amplio, tal como M13 y fagos de ADN monocatenario filamentosos.

En algunas realizaciones, un vector que comprende un polinucleótido que codifica una inteína comprende además un sitio de clonación múltiple. Un sitio de clonación múltiple es una secuencia de polinucleótido que comprende uno o más sitios de restricción únicos. Los ejemplos no limitativos de los sitios de restricción incluyen EcoRI, SacI, KpnI, SmaI, XmaI, BamHI, XbaI, HincII, PstI, SphI, HindIII, AvaI, o cualquier combinación de éstos.

Los sitios de clonación múltiples pueden usarse en vectores que comprenden un polinucleótido que codifica una inteína para simplificar la inserción de un polinucleótido que codifica un polipéptido heterólogo en el vector de manera que el vector puede usarse para expresar una proteína de fusión que comprende la inteína y el polipéptido heterólogo. Así, por ejemplo, un vector puede comprender una secuencia que codifica un dominio C-terminal de inteína aguas arriba de un sitio de clonación múltiple de manera que una secuencia que codifica un polipéptido heterólogo pueda insertarse fácilmente aguas abajo del dominio C-terminal de inteína. Un vector también puede comprender una secuencia que codifica un dominio N-terminal de inteína aguas abajo de un sitio de clonación múltiple de manera que una secuencia que codifica un polipéptido heterólogo pueda insertarse fácilmente aguas arriba del dominio N-terminal de inteína.

Así, por ejemplo, un vector puede comprender una secuencia que codifica un dominio C-terminal de inteína aguas arriba de un sitio de clonación múltiple, que a su vez está aguas arriba de una secuencia que codifica un dominio N-

terminal de inteína de manera que una secuencia que codifica un polipéptido heterólogo pueda insertarse fácilmente aguas abajo del dominio C-terminal de inteína y aguas arriba del dominio N-terminal de inteína.

Un vector que comprende un polinucleótido con un sitio de clonación múltiple aguas arriba de un dominio N-terminal de inteína puede combinarse en un kit con un vector que comprende un polinucleótido con un sitio de clonación múltiple aguas abajo de un dominio C-terminal de inteína. En algunas realizaciones, un único vector comprende un polinucleótido con un sitio de clonación múltiple aguas arriba de un dominio N-terminal de inteína y un polinucleótido con un sitio de clonación múltiple aguas abajo de un dominio C-terminal de inteína. En dichos vectores, cada uno de los polinucleótidos con un sitio de clonación múltiple aguas arriba de un dominio N-terminal de inteína y el polinucleótido con un sitio de clonación múltiple aguas abajo de un dominio C-terminal de inteína pueden unirse de forma operativa a secuencias reguladoras, y las secuencias reguladoras pueden ser iguales o diferentes.

Los vectores pueden comprender al menos un promotor. El promotor puede ser cualquier secuencia que es adecuada para dirigir la expresión de un dominio de inteína o fusión de inteína.

Diferentes huéspedes tienen frecuentemente preferencias para un codón particular para usarse para codificar un residuo de aminoácido particular. Dichas preferencias de codones son muy conocidas y una secuencia de ADN que codifica una secuencia de proteína de fusión deseada puede alterarse, usando mutagénesis in vitro, por ejemplo, de manera que los codones preferidos del huésped se utilizan para un huésped particular en el que se quiere expresar la proteína de fusión.

También se contempla una molécula de ácido nucleico recombinante tal como una molécula de ADN, que comprende un vector o construcción génica que contiene una o más secuencias reguladoras (elementos de control) tales como un promotor adecuado para dirigir la expresión del gen en un organismo de célula huésped bacteriano o eucariota compatible unidas de forma operativa a un segmento de ácido nucleico exógeno (por ejemplo, un segmento o secuencia de ADN) que define un gen que codifica una proteína de fusión contemplada, como se ha discutido anteriormente. Más particularmente, también se contempla una molécula de ADN recombinante que comprende un vector génico que comprende un promotor para dirigir la expresión de la proteína de fusión en células de un organismo huésped unido de forma operativa a un segmento de ADN que define un gen que codifica un dominio de inteína unido a un polipéptido heterólogo. Esta molécula de ADN recombinante, después de transfección y expresión adecuada en una célula huésped, proporciona una proteína de fusión contemplada.

Como es muy conocido en la técnica, siempre que el ácido nucleico requerido, ilustrativamente secuencia de ADN, esté presente, (incluyendo señales de inicio y parada), pueden estar presentes habitualmente pares de bases adicionales en cualquier extremo del segmento de ADN, y este segmento todavía puede utilizarse para expresar la proteína. Esto, por supuesto, presume la ausencia en el segmento de una secuencia de ADN unida de forma operativa que reprime la expresión, expresa un producto adicional que consume la proteína de fusión deseada que se desea expresar, expresa un producto que consume un producto de la reacción deseado producido por esa proteína de fusión deseada, o interfiere de otra manera con la expresión del gen del segmento de ADN.

Así, siempre que el segmento de ADN carece de dichas secuencias de ADN de interferencia, un segmento de ADN de la invención puede tener una longitud de aproximadamente 500 a aproximadamente 15.000 pares de bases. El tamaño máximo de una molécula de ADN recombinante, particularmente un vector de expresión, está gobernado en gran medida por conveniencia y el tamaño del vector que puede ser acomodado por una célula huésped, una vez están presentes todas las secuencias de ADN mínimas requeridas para la replicación y expresión, cuando se desea. Los tamaños mínimos de vectores son muy conocidos.

Un segmento de ADN que codifica una proteína de fusión puede sintetizarse por técnicas químicas, por ejemplo, el método fosfotriéster de Matteucci et al., 1981 J. Am. Chem. Soc., 103:3185. Por supuesto, mediante la síntesis química de la secuencia codificadora, puede hacerse cualquier modificación deseada simplemente sustituyendo las bases apropiadas por aquellas que codifican la secuencia de residuos de aminoácidos nativos.

Los segmentos de ADN que contienen un gen que codifica la proteína de fusión también pueden obtenerse de moléculas de ADN recombinante (vectores plasmídicos) que contienen ese gen.

Un vector que dirige la expresión de un gen de proteína de fusión en una célula huésped se refiere en la presente memoria como un "vector de expresión". Un vector de expresión contiene elementos de control de la expresión incluyendo el promotor. El gen que codifica la proteína de fusión está unido de forma operativa al vector de expresión para permitir que la secuencia promotora dirija la unión de la ARN polimerasa y la expresión del gen que codifica la proteína de fusión. En la expresión del gen que codifica el polipéptido son útiles los promotores que son inducibles, virales, sintéticos, constitutivos como se describe por Paszkowski et al., 1989 EMBO J., 3:2719 y Odell et al., 1985 Nature, 313:810, así como regulados temporalmente, regulados espacialmente, y regulados espaciotemporalmente como se proporciona en Chua et al., 1989 Science, 244:174-181.

En la presente memoria se contemplan los vectores de expresión compatibles con células eucariotas, tales como los compatibles con células de procariotas (*E. coli*), mamíferos, algas o insectos y semejantes. Dichos vectores de expresión también pueden usarse para formar moléculas de ADN recombinante de la presente invención. Los vectores de expresión de células procariotas y eucariotas son muy conocidos en la técnica y están disponibles en

varias fuentes comerciales. Normalmente, dichos vectores contienen uno o más sitios de restricción convenientes para la inserción del segmento de ADN y secuencias promotoras deseadas. Opcionalmente, dichos vectores contienen un marcador seleccionable específico para uso en células procariotas o eucariotas.

La elección de qué vector de expresión y finalmente a qué promotor se une de forma operativa un gen que codifica una proteína de fusión depende directamente de las propiedades funcionales deseadas, por ejemplo, la localización y curso de tiempo de la expresión de la proteína, y la célula huésped que se va a transformar. Éstas son limitaciones muy conocidas inherentes a la técnica de la construcción de moléculas de ADN recombinante. Sin embargo, un vector útil en la práctica de la presente invención puede dirigir la replicación, y preferiblemente también la expresión (para un vector de expresión) del gen de la proteína de fusión incluido en el segmento de ADN al que está unido de forma operativa.

Las proteínas de fusión de inteínas divididas pueden expresarse en cualquier tipo celular. Por ejemplo, las proteínas de fusión de inteínas divididas pueden expresarse en procariotas, plantas (por ejemplo, monocotiledóneas o dicotiledóneas), animales, insectos, hongos, o levaduras (por ejemplo, *Saccharomyces* o *Pichia*). Las células adecuadas incluyen, como ejemplo, plantas (por ejemplo, tomate, tabaco, arabidopsis, alfalfa), células de mamífero (por ejemplo, células CHO, COS y 293T), hongos filamentosos (por ejemplo, Tricoderma resei y Aspergillus sp.), y células de insecto. Los ejemplos de líneas celulares huésped de mamíferos adecuadas incluyen las líneas COS-7 de células de riñón de mono, descritas por Gluzman (Cell 23:175, 1981), y otras líneas celulares capaces de expresar un vector apropiado incluyendo, por ejemplo, células L, líneas celulares C127, 3T3, de ovario de hámster chino (CHO), HeLa y BHK. Los sistemas de baculovirus para la producción de proteínas heterólogas en células de insecto están revisados por Luckow y Summers, Bio/Technology 6:47 (1988). Las proteínas de fusión de inteínas divididas pueden purificarse de dichas células usando técnicas que son conocidas en la técnica. Además, las proteínas de fusión de inteínas divididas pueden producirse en sistemas de transcripción/traducción sin células.

IV. Composiciones que comprenden las fusiones de inteína

La invención también se refiere a composiciones y kits de partes que contienen las proteínas de fusión de la invención. El término "composición", tal y como se usa en la presente memoria, se refiere a una combinación de uno o más componentes en la que los componentes pueden:

(i) proporcionarse como formulaciones separadas (es decir, independientemente una de otra), que se juntan posteriormente para uso conjunto entre sí; o

(ii) envasarse y presentarse conjuntamente como componentes separados de un "envase de combinación" para uso conjunto entre sí.

En una realización, la composición o kit de partes comprende componentes adecuados para la escisión C-terminal de un polipéptido que está conectado con el extremo C del dominio C-terminal de una inteína. Estas composiciones comprenden

(i) un primer componente que es una proteína de fusión que comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y (ii) un polipéptido heterólogo, en el que el polipéptido heterólogo es C-terminal respecto al dominio de inteína y

(ii) un segundo componente que se selecciona del grupo que consiste en una proteína de fusión que comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y (ii) un polipéptido heterólogo, en el que el polipéptido heterólogo es N-terminal respecto al dominio de inteína y un dominio N-terminal de inteína en el que el primer aminoácido del dominio de inteína es un aminoácido distinto de serina o cisteína.

en la que

a. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:7 y el dominio de inteína de la proteína de fusión que forma el segundo componente o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:3;

b. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:16 y el dominio de inteína de la proteína de fusión del segundo componente o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:12;

c. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:24 y el dominio de inteína de la proteína de fusión que forma el segundo componente o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:20;

d. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:38 y el dominio de inteína de la proteína de fusión que forma el segundo componente o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:34 o

e. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:65 y el dominio de inteína de la proteína de fusión que forma el segundo componente o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:64.

En otra realización, el polipéptido heterólogo y el dominio de inteína que forman la proteína de fusión que forma el primer componente de la composición del kit de partes están bien conectados directamente por un enlace peptídico o por un conector. En otra realización, el último aminoácido del dominio C-terminal de inteína es glutamina o asparagina.

En otra realización, el segundo componente se selecciona del grupo que consiste en un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64, en el que el primer aminoácido del dominio de inteína es un aminoácido distinto de serina o cisteína. (no tengo claro que sea necesario que el dominio N de la inteína esté fusionado a una proteína heteróloga para funcionar)

En otra realización, la composición o kit de partes de la invención comprende componentes adecuados para la escisión N-terminal de un polipéptido que está conectado con el extremo N del dominio N-terminal de una inteína. Estas composiciones comprenden

(i) un primer componente que es una proteína de fusión que comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y (ii) un polipéptido heterólogo, en el que el polipéptido heterólogo es N-terminal respecto al dominio de inteína y

(ii) un segundo componente que se selecciona del grupo que consiste en la proteína de fusión que comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y (ii) un polipéptido heterólogo, en el que el polipéptido heterólogo es C-terminal respecto al dominio de inteína y un dominio C-terminal de inteína y en el que el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina y en el que el primer aminoácido del polipéptido heterólogo o del conector es un aminoácido distinto de serina, cisteína, o treonina.

en la que

a. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:3 y el dominio de inteína de la proteína de fusión que forma el segundo componente o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:7;

b. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:12 y el dominio de inteína del segundo componente o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:16;

c. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:20 y el dominio de inteína de la proteína de fusión que forma el segundo componente o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:24;

d. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:34 y el dominio de inteína de la proteína de fusión que forma el segundo componente o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:38 o

e. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:64 y el dominio de inteína de la proteína de fusión que forma el segundo componente o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:65.

En una realización preferida, el polipéptido heterólogo y el dominio de inteína que forman el primer componente de la composición o kit de partes están bien conectados directamente por un enlace peptídico o por un conector. En otra realización, el primer aminoácido del dominio de inteína es una serina o cisteína.

En otra realización, el segundo componente se selecciona del grupo que consiste en un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65, en el que el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina.

En otra realización, la composición o kit de partes según la invención comprende reactivos adecuados para unir covalentemente el extremo N de un primer polipéptido al extremo C de un segundo polipéptido, comprendiendo dicha composición

(i) una proteína de fusión que comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y (ii) un segundo polipéptido heterólogo, en el que el polipéptido heterólogo es C-terminal respecto al dominio de inteína y

(ii) una proteína de fusión que comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y (ii) un primer polipéptido heterólogo, en el que el polipéptido heterólogo es N-terminal respecto al dominio de inteína

en la que

a. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:7 y el dominio de inteína de la proteína de fusión que forma el segundo componente es al menos 75% idéntico a SEQ ID NO:3;

b. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:16 y el dominio de inteína de la proteína de fusión que forma el segundo componente es al menos 75% idéntico a SEQ ID NO:12;

c. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:24 y el dominio de inteína de la proteína de fusión que forma el segundo componente es al menos 75% idéntico a SEQ ID NO:20;

d. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:38 y el dominio de inteína de la proteína de fusión que forma el segundo componente es al menos 75% idéntico a SEQ ID NO:34; o

e. el dominio de inteína de la proteína de fusión que forma el primer componente es al menos 75% idéntico a SEQ ID NO:65 y el dominio de inteína de la proteína de fusión que forma el segundo componente es al menos 75% idéntico a SEQ ID NO:64.

En una realización preferida, el polipéptido heterólogo y el dominio de inteína que forman parte de la proteína de fusión que es el primer componente de la composición están bien conectados directamente por un enlace peptídico o por un conector. En una realización más preferida, el último aminoácido del dominio de inteína en la proteína de fusión que es el primer componente de la invención es glutamina o asparagina.

En otra realización preferida, el polipéptido heterólogo y el dominio de inteína que forman parte de la proteína de fusión que es el segundo componente de la composición están bien conectados directamente por un enlace peptídico o por un conector. En una realización más preferida, el primer aminoácido del dominio de inteína en la proteína de fusión que es el segundo componente de la invención es serina o cisteína.

La relación de los componentes en las composiciones es adecuada para el procesamiento eficiente de las proteínas de fusión. Las relaciones adecuadas del primer y segundo componentes incluyen, sin limitación 1.000:1, 100:1; 10:1, 1:1, 1:10, 1:100 y 1:1.000.

V. Métodos para usar las fusiones de inteína

Las inteínas divididas y proteínas de fusión que comprenden inteínas divididas descritas en la presente memoria pueden usarse, por ejemplo, para escindir, unir (corte y empalme) y/o ciclar secuencias de polipéptido. Las secuencias de inteína catalizan estas reacciones, que pueden ocurrir en ausencia de cualesquiera otras enzimas, aditivos químicos, o tratamientos.

En algunas realizaciones, un polipéptido heterólogo puede escindirse de un dominio de inteína. Por ejemplo, un polipéptido heterólogo puede escindirse del extremo C de un dominio C-terminal de inteína usando una inteína dividida que comprende una proteína de fusión que comprende una secuencia de aminoácidos que evita o disminuye la escisión entre un polipéptido heterólogo y un dominio N-terminal de inteína. En una realización preferida, el primer aminoácido del dominio N-terminal de inteína es un aminoácido distinto de cisteína o serina, por ejemplo, alanina. El primer aminoácido del polipéptido heterólogo puede seleccionarse con el fin de incrementar el rendimiento de la reacción, bien porque resulta en una vida media incrementada de la reacción o porque resulta en un valor k incrementado. Así, en una realización preferida, el primer aminoácido del polipéptido heterólogo o del conector que conecta el dominio C-terminal de la inteína y el polipéptido heterólogo se selecciona del grupo que consiste en Met, Cys, Thr, Arg, Lys, Ser, Gln, His, Ala, Tyr, Phe, Asn, Trp, Val, Leu, Asp, Ile, Gly, Glu o Pro.

Además, un polipéptido heterólogo puede escindirse del extremo N de un dominio N-terminal de inteína usando una inteína dividida que comprende una proteína de fusión que contiene una secuencia de aminoácidos que disminuye la escisión entre un polipéptido heterólogo y un dominio C-terminal de inteína. En una realización preferida, el último aminoácido del dominio C-terminal de inteína es un aminoácido distinto de glutamina o asparagina, por ejemplo, alanina.

En algunas realizaciones, un primer polipéptido puede unirse (con corte y empalme) a un segundo polipéptido poniendo en contacto una proteína de fusión que comprende el primer polipéptido y un dominio N-terminal de inteína con una proteína de fusión que comprende el segundo polipéptido y un dominio C-terminal de inteína. El extremo C del primer polipéptido se unirá al extremo N del segundo polipéptido.

En algunas realizaciones, las inteínas divididas pueden usarse para ciclar un polipéptido que comprende un dominio C-terminal de inteína en el extremo N del polipéptido y un dominio N-terminal de inteína en el extremo C del polipéptido.

En algunas realizaciones, la reacción ocurre a aproximadamente 0°C, a aproximadamente 60°C. En algunas realizaciones, la reacción ocurre a aproximadamente 0°C, aproximadamente 4°C, aproximadamente 8°C, aproximadamente 12°C, aproximadamente 20°C, aproximadamente 25°C, aproximadamente 30°C, aproximadamente 32°C, aproximadamente 34°C, aproximadamente 37°C, aproximadamente 40°C, aproximadamente 45°C, aproximadamente 50°C, aproximadamente 55°C, o aproximadamente 60°C.

En algunas realizaciones, la reacción ocurre a un pH de aproximadamente 5 a aproximadamente 10. En algunas realizaciones, la reacción ocurre a un pH de aproximadamente 6, aproximadamente 6,5, aproximadamente 7, aproximadamente 7,5, aproximadamente 8, aproximadamente 8,5, aproximadamente 9, aproximadamente 9,5 o aproximadamente 10.

En algunas realizaciones, la reacción ocurre en presencia de un agente desnaturalizante, por ejemplo, para incrementar la solubilidad de la proteína. En algunas realizaciones, la reacción ocurre en presencia de urea. En algunas realizaciones, la reacción ocurre en presencia de no más de aproximadamente 6,5M, aproximadamente 6M, aproximadamente 5M, aproximadamente 4,5M, aproximadamente 4M, aproximadamente 3,5M, aproximadamente 3M, aproximadamente 2,5M, aproximadamente 2M, aproximadamente 1,5M, aproximadamente 1M, o aproximadamente 0,5M urea. En algunas realizaciones, la reacción ocurre en presencia de aproximadamente 0,5 a aproximadamente 6M, aproximadamente 0,5 a aproximadamente 4M, aproximadamente 1 a aproximadamente 4 M, aproximadamente 2 a aproximadamente 4M, o aproximadamente 3 a aproximadamente 4M urea. En algunas realizaciones, la reacción ocurre en presencia de aproximadamente 0,5 a aproximadamente 2M, o aproximadamente 0,5 a 1 M urea.

Los métodos descritos en la presente memoria demuestran que las inteínas divididas pueden tener una actividad robusta. Así, en algunas realizaciones la constante de la velocidad de la reacción es al menos aproximadamente 0,5 x 10^{-1} s^{-1}, 1 x 10^{-1} s^{-1}, 1,5 x 10^{-1} s^{-1}, 0,5 x 10^{-2} s^{-1}, aproximadamente 1 x 10^{-2} s^{-1}, aproximadamente 1,5 x 10^{-2} s^{-1}, aproximadamente 2,0 x 10^{-2} s^{-1}, aproximadamente 2,5 x 10^{-2} s^{-1}, o aproximadamente 3 x 10^{-2} s^{-1} cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares. Además, la vida media de la velocidad de reacción puede ser menor de aproximadamente 150, aproximadamente 100, aproximadamente 50, aproximadamente 40, aproximadamente 45, aproximadamente 30, aproximadamente 25, aproximadamente 20, o aproximadamente 15 segundos cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares. En algunas realizaciones, la constante de velocidad de la reacción de corte y empalme en trans es al menos aproximadamente 0,5 x 10-^{1}s^{-1}, 1 x 10^{-1} s^{-1}, 1,5 x 10^{-1} s^{-1}, 0,5 x 10^{-2} s^{-1}, aproximadamente 1 x 10^{-2} s^{-1}, aproximadamente 1,5 x 10^{-2} s^{-1}, aproximadamente 2,0 x 10^{-2} s^{-1}, aproximadamente 2,5 x 10^{-2} s^{-1}, o aproximadamente 3 x 10^{-2} s^{-1} cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares. Además, la vida media de la velocidad de reacción puede ser menor de aproximadamente 150, aproximadamente 100, aproximadamente 50, aproximadamente 40, aproximadamente 45, aproximadamente 30, aproximadamente 25, aproximadamente 20, o aproximadamente 15 segundos cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares. En algunas realizaciones, la constante de la velocidad de la reacción de escisión C es al menos aproximadamente 1 x 10^{4} s^{-1} , 3 x 10^{-4} s^{-1} , 6 x 10^{-4} s^{-1}, 9 x 10^{-4} s^{-1} 1 x 10^{-5} s^{-1}, aproximadamente 3 x 10^{-5} s^{-1}, aproximadamente 6 x 10^{-5} s^{-1}, o aproximadamente 9 x 10^{-5} s^{-1}, cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares. Además, la vida media de la velocidad de reacción puede ser menor de aproximadamente 150, aproximadamente 100, aproximadamente 50, aproximadamente 40, aproximadamente 45, aproximadamente 30, aproximadamente 25, aproximadamente 20, o aproximadamente 15 minutos cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares (la escisión C es más lenta).

En algunas realizaciones, la reacción resulta en un rendimiento de al menos aproximadamente 50%, al menos aproximadamente 55%, al menos aproximadamente 60%, al menos aproximadamente 65%, de al menos aproximadamente 70%, al menos aproximadamente 75%, al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95% cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares. En algunas realizaciones, la reacción resulta en un rendimiento de al menos aproximadamente 80%, al menos aproximadamente 85%, al menos aproximadamente 90%, al menos aproximadamente 95% en 5 minutos cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares. En algunas realizaciones, la reacción resulta en un rendimiento de aproximadamente 75 a aproximadamente 80%, 80% a aproximadamente 85%, 85% a aproximadamente 90%, o aproximadamente 90 a 95% en 5 minutos cuando se mezclan un dominio N-terminal y dominio C-terminal de inteína en concentraciones equimolares.

En algunas realizaciones, una reacción de inteína (por ejemplo, escisión, unión (corte y empalme), ciclación) puede iniciarse poniendo en contacto una proteína de fusión que comprende un dominio N-terminal de inteína y, opcionalmente, un polipéptido heterólogo con una proteína de fusión que comprende un dominio C-terminal de inteína y, opcionalmente, un polipéptido heterólogo. En algunas realizaciones, una reacción de inteína puede

iniciarse cambiando las condiciones, por ejemplo, la temperatura o pH, a las que se incuba una proteína de fusión de inteína dividida o una combinación de proteínas de fusión de inteína dividida. En algunas realizaciones, una escisión C-terminal se inicia por un cambio de pH o temperatura.

En algunas realizaciones, una reacción de inteína se inicia poniendo en contacto proteínas de fusión con DTT u otro nucleófilo fuerte. En algunas realizaciones, se usa DTT para potenciar una reacción. En algunas realizaciones, una escisión N-terminal se inicia por un nucleófilo fuerte, por ejemplo, DTT.

Otra manera de inducir el corte y empalme o escisión de proteínas es por contacto con un péptido o agente peptidomimético que activa el corte y empalme o escisión. Otra manera de inducir el corte y empalme o escisión de proteínas es por la eliminación de un péptido o agente peptidomimético que bloquea o inhibe el corte y empalme o escisión.

En algunas realizaciones, la proteína de fusión puede unirse a una resina, por ejemplo, para el propósito de la separación o purificación de la proteína tal como las proporcionadas, por ejemplo, en Lu et al., Journal of Chromatography A 1218: 2553-2560 (2011) y Elleuche y Poggeler, Appl. Microbiol. Biotechnol 87:479-489 (2010), que se incorporan en la presente memoria por referencia. Además, una proteína de fusión puede estar en disolución, unida a lechos o columna de afinidad, anclada a una membrana celular o superficie de fago. Los agentes de unión por afinidad pueden incluir etiquetas de His, dominios de unión de quitina, una proteína de unión a maltosa, o una glutatión-S-transferasa, por ejemplo. La proteína de fusión puede estar dentro o fuera de una célula.

En algunas realizaciones, las reacciones de inteína pueden usarse en la purificación de proteínas (por ejemplo, usando etiquetas cromatográficas o etiquetas no cromatográficas y/o en procesos a gran escala), en la circularización de proteínas, en la polimerización de proteínas, y en la producción de selenoproteínas, por ejemplo, como se describe en Elleuche y Poggeler, Appl. Microbiol. Biotechnol 87:479-489 (2010), y Evans T. et al., Biopolymers 51:333-342 (1999), que se incorporan en la presente memoria por referencia en su totalidad. La alta eficiencia de las inteínas proporcionadas en la presente memoria las hace particularmente idóneas para aplicaciones industriales a gran escala.

En algunas realizaciones, una reacción de inteína puede usarse para producir un polipéptido diana. El polipéptido diana puede ser un polipéptido de fusión que contiene dos secuencias que previamente no estaban unidas. El polipéptido diana también puede ser un polipéptido que se escinde de una secuencia a la que estaba unido previamente.

Pueden realizarse múltiples reacciones de corte y empalme en tándem y en cualquier orden para organizar y reorganizar secuencias de polipéptido o para unir múltiples o diferentes polipéptidos, según se desee.

VI. Vectores para la generación de proteínas de fusión de inteína

La invención también proporciona vectores adecuados para la generación de proteínas de fusión de inteína que comprenden un polinucleótido que codifica un dominio de inteína y uno o más sitios de clonación que permiten la inserción de un polinucleótido que codifica un polipéptido heterólogo en una posición que resulta en la expresión de una proteína de fusión que comprende el dominio de inteína y el polipéptido heterólogo.

Por lo tanto, en otro aspecto, la invención se refiere a un vector que comprende un polinucleótido que codifica un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y al menos un sitio de clonación aguas abajo de dicho polinucleótido que permite la clonación de un polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el dominio de inteína y el polipéptido codificado por el polinucleótido de interés.

En una realización, el polinucleótido que codifica un dominio de inteína que muestra al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 codifica un dominio de inteína en el que el último aminoácido es glutamina o asparagina. En otra realización, el polinucleótido que codifica un dominio de inteína que muestra al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65, codifica un dominio de inteína en el que el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina. En otra realización, el vector comprende además un polinucleótido que codifica un polipéptido que forma un péptido conector entre el dominio de inteína y el polipéptido codificado por el péptido heterólogo. En una realización preferida, cuando el polinucleótido que codifica el dominio de inteína que muestra al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 codifica un dominio de inteína en el que el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina, entonces el polinucleótido codifica una región conectora en la que el primer aminoácido de dicho conector es un aminoácido distinto de serina, treonina o serina.

En otro aspecto, la invención se refiere a un vector que comprende un polinucleótido que codifica un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y al menos un sitio de clonación aguas arriba de dicho polinucleótido que permite la clonación de un polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el polipéptido codificado por el polinucleótido de interés y el dominio de inteína.

En una realización, el polinucleótido que codifica un dominio de inteína que muestra al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 codifica un dominio de inteína en el que el primer aminoácido es serina o cisteína. En otra realización, el polinucleótido que codifica un dominio de inteína que muestra al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65, codifica un dominio de inteína en el que el primer aminoácido del dominio de inteína es un aminoácido distinto de serina o cisteína.

En otra realización, la invención se refiere a un vector que es útil para clonar un polinucleótido que codifica una proteína de interés y para producir dicho polipéptido que puede ciclarse. Así, la invención se refiere a un vector que comprende un polinucleótido que codifica un primer dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65, y al menos un sitio de clonación aguas abajo de dicho polinucleótido que permite la clonación de un polinucleótido de interés, y un polinucleótido aguas abajo del sitio de clonación, que codifica un segundo dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64, de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el polipéptido codificado por el polinucleótido de interés y el primer y segundo dominios de inteína.

En una realización preferida, el último aminoácido del primer dominio de inteína es glutamina o asparagina. En otra realización, el penúltimo aminoácido del primer dominio de inteína es histidina. En otra realización más, el primer aminoácido del segundo dominio de inteína es serina o cisteína.

En una realización, la invención se refiere a un vector que es útil para preparar proteínas de fusión que comprenden los dominios N-terminal y C-terminal y dos regiones de un polipéptido que se van a conectar por una reacción de corte y empalme en trans entre ambos dominios de inteína. Así, en otro aspecto, la invención se refiere a un vector que comprende:

(i) un polinucleótido que codifica un primer dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65,

(ii) un primer sitio de clonación aguas abajo de dicho polinucleótido que codifica un primer dominio de inteína,

(iii) un polinucleótido que codifica un segundo dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y

(iv) un segundo sitio de clonación aguas arriba de dicho polinucleótido que codifica un segundo dominio de inteína,

en el que el primer sitio de clonación permite la clonación de un primer polinucleótido de interés y el segundo sitio de clonación permite la clonación de un segundo polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende, en dicho orden, el polipéptido codificado por el segundo polinucleótido de interés, el segundo dominio de inteína, el primer dominio de inteína y el polipéptido codificado por el segundo polinucleótido de interés y en el que

a. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:7, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:3;

b. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:16, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:12;

c. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:24, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:20;

d. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:38, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:34 o

e. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:65, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:64.

En otra realización, el vector comprende además un polinucleótido que codifica un primer conector peptídico que conecta el segundo dominio de inteína y el polipéptido codificado por el segundo polinucleótido de interés y/o que comprende además un polinucleótido que codifica un segundo conector peptídico que conecta el primer dominio de inteína y el polipéptido codificado por el primer polinucleótido de interés.

En otra realización, el primer aminoácido del segundo dominio de inteína es cisteína o serina, en el que el último aminoácido del primer dominio de inteína es glutamina o asparagina, en el que el penúltimo aminoácido del primer dominio de inteína es histidina y/o en el que el primer aminoácido del segundo polipéptido de interés o del primer conector peptídico es cisteína, serina o treonina.

Tal y como se usa en esta invención, el término "vector" se refiere a un vehículo mediante el cual un polinucleótido o una molécula de ADN puede manipularse o introducirse en una célula. El vector puede ser un polinucleótido lineal o

circular, o puede ser un polinucleótido de gran tamaño o cualquier otro tipo de construcción, tal como ADN o ARN de un genoma viral, un virión o cualquier otra construcción biológica que permita la manipulación de ADN o la introducción de éste en la célula. Se entiende que las expresiones "vector recombinante" y "sistema recombinante" pueden usarse indistintamente con el término "vector". Los expertos en la técnica observarán que no hay limitación en los términos del tipo de vector que pueden usarse, ya que dicho vector puede ser un vector de clonación adecuado para la propagación y para obtener los polinucleótidos o construcciones génicas o vectores de expresión adecuados en diferentes organismos heterólogos adecuados para la purificación de las proteínas de fusión. Así, los vectores adecuados según esta invención incluyen vectores de expresión en procariotas, tales como pUC18, pUC19, Bluescript y los derivados de éste, mp18, mp19, pBR322, pMB9, CoIE1, pCR1, RP4, fagos y vectores "lanzadera", tales como pSA3 y pAT28, vectores de expresión en levaduras, tales como los vectores del tipo plásmido de 2 micrómetros, plásmidos de integración, vectores YEP, plásmidos de centrómeros y similares, vectores de expresión en células de insecto, tales como los vectores de la serie pAC y la serie pVL, vectores de expresión en plantas, tales como los vectores de la serie pIBI, pEarleyGate, pAVA, pCAMBIA, pGSA, pGWB, pMDC, pMY, pORE y similares, y vectores de expresión en células de eucariotas superiores basados en vectores virales (adenovirus, virus asociados con adenovirus, así como retrovirus y lentivirus) y vectores no virales, tales como pSilencer 4.1-CMV (Ambion), pcDNA3, pcDNA3.1/hyg, pHCMV/Zeo, pCR3.1, pEF1/His, pIND/GS, pRc/HCMV2, pSV40/Zeo2, pTRACER-HCMV, pUB6/V5-His, pVAX1, pZeoSV2, pCI, pSVL y pKSV-10, pBPV-1, pML2d y pTDT1.

En una forma preferida de realización, el vector comprende además, en la posición 3' respecto al polinucleótido que codifica el dominio de inteína, uno o varios sitios que permiten la clonación de los polinucleótidos que codifican un polipéptido heterólogo. Preferiblemente, los sitios de clonación están agrupados de manera que se forma un sitio de clonación múltiple, como aparecen frecuentemente en los vectores de clonación. Así, el término "sitio de clonación múltiple", tal y como se usa en esta invención, se refiere a una secuencia de ácido nucleico que comprende una serie de dos o más secuencias diana de endonucleasa de restricción que están localizadas cerca una de otra. Los sitios de clonación múltiple incluyen dianas de endonucleasa de restricción que permiten la inserción de fragmentos con extremos romos, extremos 5' cohesivos o extremos 3' cohesivos. La inserción de polinucleótidos de interés se realiza usando métodos estándar de biología molecular, como se describe, por ejemplo, por Sambrook et al. (Sambrook et al. Molecular Cloning: A Laboratory Manual, Cold Spring Harbour Laboratory Press, 1989) y/o Ausubel et al. (Current Protocols in Molecular Biology, Greene Pub. Associates and Wiley- Interscience (1988, incluyendo todas las actualizaciones hasta la fecha).

Como será evidente para el experto en la técnica a partir de la descripción de la presente memoria, la presente descripción es útil para producir construcciones de expresión, es decir, en las que ácidos nucleicos están unidos de forma operativa a promotores adecuados.

Los sistemas de expresión sin células están contemplados por la presente descripción. Por ejemplo, un ácido nucleico se une de forma operativa a un promotor adecuado, por ejemplo, un promotor T7, y la construcción de expresión resultante se expone a condiciones suficientes para la transcripción y traducción. Los vectores de expresión típicos para la expresión in vitro o expresión sin células se han descrito e incluyen, pero no están limitados a, los sistemas TNT T7 y TNT T3 (Promega), los vectores pEXP1-DEST y pEXP2-DEST (Invitrogen).

Están disponibles muchos vectores para expresión en células. Los componentes del vector incluyen generalmente, pero no están limitados a, uno o más de los siguientes: una secuencia señal, una secuencia que codifica un o unos polipéptidos, un elemento potenciador, un promotor, y una secuencia de terminación de la transcripción. El experto en la técnica estará al tanto de secuencias adecuadas para la expresión de una proteína. Por ejemplo, las secuencias señal ejemplares incluyen señales de secreción de procariotas (por ejemplo, pelB, fosfatasa alcalina, penicilinasa, lpp, o enterotoxina estable al calor II), señales de secreción de levaduras (por ejemplo, líder de invertasa, un factor líder, o líder de fosfatasa ácida) o señales de secreción de mamíferos (por ejemplo, señal del herpes simple gD).

Los promotores ejemplares incluyen aquellos activos en procariotas (por ejemplo, promotor phoA, sistemas de promotor de beta-lactamasa y lactosa, fosfatasa alcalina, un sistema de promotor de triptófano (trp), y promotores híbridos tales como el promotor tac). Estos promotores son útiles para la expresión en procariotas incluyendo eubacterias, tales como organismos Gram-negativos o Gram-positivos, por ejemplo, Enterobacteriaceae tales como Escherichia, por ejemplo, E. coli, Enterobacter, Erwinia, Klebsiella, Proteus, Salmonella, por ejemplo, Salmonella typhimurium, Serratia, por ejemplo, Serratia marcescans, y Shigella, así como Bacilli tales como B. subtilis y B. licheniformis, Pseudomonas tales como P. aeruginosa, y Streptomyces. En un ejemplo, el huésped es E. coli. Un huésped de clonación preferido de E. coli es E. coli 294 (ATCC 31,446), aunque son adecuadas otras cepas tales como E. coli B, E. coli X 1776 (ATCC 31,537), y E. coli W3110 (ATCC 27,325), DH5a o DH10B.

Los promotores ejemplares activos en células de mamífero incluyen promotor temprano inmediato de citomegalovirus (CMV-IE), promotor del factor de elongación 1-oc humano (EF1), promotores de ARN pequeño nuclear (U1 a y Ulb), un promotor de la cadena pesada de miosina, promotor del virus de simio 40 (SV40), promotor del virus del sarcoma de Rous (RSV), promotor tardío principal de adenovirus, promotor de beta-actina; elemento regulador híbrido que comprende un potenciador de CMV/promotor de beta-actina o un promotor de inmunoglobulina o fragmento activo de ésta. Los ejemplos de líneas celulares huésped de mamíferos útiles son línea CV1 de riñón de mono transformada con SV40 (COS-7, ATCC CRL 1651); línea de riñón embrionario humano (293 o células 293

subclonadas para crecimiento en un cultivo en suspensión); células de riñón de cría de hámster (BHK, ATCC CCL 10); o células de ovario de hámster chino (CHO).

Los promotores típicos adecuados para expresión en células de levadura tales como por ejemplo una célula de levadura seleccionada del grupo que comprende *Pichia pastoris, Saccharomyces cerevisiae* y *S. pombe,* incluyen, pero no están limitados a, el promotor ADH1, el promotor GAL1, el promotor GAL4, el promotor CUPI, el promotor PH05, el promotor nmt, el promotor RPR1, o el promotor TEF1.

Los promotores típicos adecuados para expresión en células de insecto incluyen, pero no están limitados a, el promotor OPEI2, el promotor de actina de insecto aislado de Bombyx muri, el promotor dsh de Drosophila sp. y el promotor de metalotioneína inducible. Las células de insecto ejemplares para la expresión de proteínas recombinantes incluyen una célula de insecto seleccionada del grupo que comprende, células BT1 -TN-5B1-4, y células de Spodoptera frugiperda (por ejemplo, células sf19, células sf21). Los insectos adecuados para la expresión de los fragmentos de ácido nucleico incluyen, pero no están limitados a, Drosophila sp. También se contempla el uso de *S. frugiperda*.

El vector según la presente invención puede comprender además un polinucleótido que codifica una proteína marcadora. Las proteínas marcadoras adecuadas para la presente descripción incluyen aquellas que confieren resistencia a antibióticos o resistencia a otro compuesto tóxico. Los ejemplos de proteínas marcadoras que confieren resistencia a antibióticos incluyen neomicina fosfotransferasa que fosforila la neomicina y kanamicina, o hpt, que fosforila la higromicina, o proteínas que confieren resistencia, por ejemplo, a bleomicina, estreptomicina, tetraciclina, cloranfenicol, ampicilina, gentamicina, geneticina (G418), espectinomicina o blasticidina). En un ejemplo, la proteína confiere resistencia a cloranfenicol. Por ejemplo, la proteína es un gen de E. coli designado CmR, por ejemplo, como se describe en Nilsen et al, J. Bacteriol, 178: 3188-3193, 1996.

Alternativamente, la proteína marcadora complementa una auxotrofía en una célula. Por ejemplo, una célula eucariota que carece de la expresión de HPRT se transforma con una construcción de expresión que comprende un ácido nucleico que codifica HPRT. La expresión del gen informador resulta en que la célula es capaz de crecer en medio HAT, mientas las células que no expresan el gen informador no son capaces de crecer en estas condiciones.

Alternativamente, en el caso de una célula de levadura, la proteína marcadora es, por ejemplo, LEU2 o LYS2 o TRP. Dicho gen informador es capaz de complementar una célula de levadura que es auxotrófica para el gen relevante, y, como consecuencia, incapaz de producir el aminoácido relevante.

En otro ejemplo, la proteína marcadora que es directamente detectable, por ejemplo, es una proteína fluorescente. En la técnica se conocen varios genes informadores fluorescentes e incluyen, por ejemplo, aquellos que codifican la proteína verde fluorescente (GFP), proteína verde fluorescente potenciada (eGFP), proteína verde fluorescente desplazada al rojo (RFP), proteína morada fluorescente (CFP), proteína amarilla fluorescente (YFP), proteína roja fluorescente monomérica de discosoma (dsRED), o dsRED2; proteína naranja fluorescente monomérica o GFP monomérica de Aequorea coerulescens. Estas proteínas permiten la selección de una célula que expresa la proteína marcadora usando técnicas estándar, por ejemplo, separación celular activada por fluorescencia (FACS).

En un ejemplo adicional, un gen marcador es una enzima que cataliza una reacción detectable. Los genes informadores enzimáticos ejemplares incluyen, por ejemplo, beta-galactosidasa, fosfatasa alcalina, luciferasa de luciérnaga o luciferasa de Renilla. Por ejemplo, la expresión de beta-galactosidasa se detecta por la adición del sustrato 5-bromo-4-cloro-3-indolil-beta-D-galactopiranósido (x-gal), que es hidrolizado por la beta-galactosidasa para producir un precipitado de color azul. Alternativamente, la expresión bien de luciferasa de luciérnaga o luciferasa de Renilla se detecta por la adición de un sustrato que en presencia de la proteína relevante es luminiscente y es detectable, por ejemplo, usando un espectrofotómetro.

La clonación del polinucleótido que codifica el polipéptido de interés en el vector de la invención puede llevarse a cabo usando técnicas estándar muy conocidas para el experto en la técnica. Por ejemplo, el polinucleótido que codifica el polipéptido de interés se produce usando la reacción en cadena de la polimerasa (PCR). Los métodos para realizar PCR son conocidos en la técnica. En el caso de los anticuerpos, puede usarse PCR para amplificar regiones variables, opcionalmente unidas a una o más regiones constantes, por ejemplo, de un sujeto o de una biblioteca o después de cribar una biblioteca. Los cebadores para amplificar dichos ácidos nucleicos que codifican regiones de anticuerpo son conocidos en la técnica (por ejemplo, como se describe en US6.096.551 y WOOO/70023). En un ejemplo adicional, el ácido nucleico puede producirse/aislarse usando digestión con endonucleasas de restricción según métodos estándar en la técnica.

Los métodos para unir ácidos nucleicos serán evidentes para el experto en la técnica y se describen, por ejemplo, en Sambrook et al. Molecular Cloning: A Laboratory Manual, Cold Spring Harbour Laboratory Press, 1989 y/o Ausubel et al. (editores), Current Protocols in Molecular Biology, Greene Pub. Associates and Wiley-Interscience (1988, incluyendo todas las actualizaciones hasta la fecha). En un ejemplo, el método usa una ligasa, por ejemplo ADN ligasa T4, para unir ácidos nucleicos.

En una forma ejemplar de la descripción, se usa clonación independiente de ligasa para unir ácidos nucleicos.

En una forma de clonación independiente de ligasa, se incluyen regiones monocatenarias complementarias en dos ácidos nucleicos que se quieren unir. Estos ácidos nucleicos se hibridan entre sí y el ácido nucleico resultante se transforma en una célula, en la que las enzimas endógenas reparan cualquier hueco que permanece y forman un único ácido nucleico contiguo.

En otra forma de clonación independiente de ligasa, se usan una o más enzimas para aumentar la formación de una única molécula de ácido nucleico. Por ejemplo, US7575860 describe una técnica en la que se usa una polimerasa que tiene actividad endonucleasa 3'-5' (por ejemplo, de virus Vaccinia) para unir los dos ácidos nucleicos. Por ejemplo, los ácidos nucleicos que se van a unir comprenden regiones que son sustancialmente idénticas o son idénticas. Estas regiones pueden tener una longitud de entre 5 a 50 nucleótidos, por ejemplo, aproximadamente 12 a 15 nucleótidos de longitud, tal como aproximadamente 15 nucleótidos de longitud. Los ácidos nucleicos que se van a unir se ponen en contacto con una polimerasa que tiene actividad exonucleasa 3'-5'. Las polimerasas ejemplares incluyen la ADN polimerasa de vaccinia, la ADN polimerasa T4 y el fragmento Klenow de la ADN polimerasa I de E. Coli. En un ejemplo, el ácido nucleico se pone además en contacto con una proteína de unión a ADN monocatenario, tal como, proteínas de unión a monocatenario de vaccinia y E. coli, proteína ICP8 del virus del herpes simple, y proteína A de replicación de levadura y humana (por ejemplo, yRPA y hRPA). Los kits para realizar este tipo de clonación independiente de ligasa están disponibles comercialmente en Clontech con el nombre comercial In-Fusion(R).

Los métodos de clonación independiente de ligasa adicionales son conocidos en la técnica e incluyen, por ejemplo, clonación independiente de ligación (LIC; por ejemplo, como se describe en Aslanidis et al, Nucl. Acids Res., 18: 6069), clonación mediada por exonucleasa T7 (US5580759), clonación basada en productos de PCR con extremos cohesivos (Liu et al, Nucleic Acids Res 24: 2458-2459, 1996), clonación basada en escisión de uracilo (Nisson et al, PCR Meth. Appl 7: 120-123, 1991), clonación independiente de ligasa basada en fósforotioato (por ejemplo, como se describe por Blanusa et al, Anal. Biochem, 406: 141-146, 2010).

El ácido nucleico resultante puede introducirse en células usando un método estándar en la técnica, por ejemplo, como se discute más adelante.

En un ejemplo, se usa recombinación para unir ácidos nucleicos. Por ejemplo, dos ácidos nucleicos que se van a unir comprenden ambos una región (por ejemplo, con una longitud de 100 nucleótidos ó 50 nucleótidos ó 20 nucleótidos ó 10 nucleótidos) que son idénticas o sustancialmente idénticas. Los ácidos nucleicos se introducen en células capaces de recombinación homóloga y se seleccionan las células en las que ha ocurrido la recombinación homóloga, por ejemplo, seleccionando para la expresión de la proteína marcadora.

La invención se describe en la presente memoria mediante los ejemplos siguientes que se pretenden como meramente ilustrativos y no limitativos del alcance de la invención.

Ejemplos

Ejemplo 1: Actividad de corte y empalme en trans de las inteínas divididas Gp41.1, Gp41.8, NrdJ1, IMPDH1 comparado con Npu DnaE

Las reacciones de corte y empalme en trans in vitro se realizaron con construcciones que contenían secuencias de inteínas divididas de Gp41.1 (G1), Gp41.8 (G8), NrdJ1(N1), y IMPDH1(I1). La inteína dividida Npu DnaE (DE), que se ha caracterizado como una inteína robusta y con alto rendimiento (Zettler J. et al, FEBS Letters 583:909-914 (2009)), se seleccionó como un control. La numeración, abreviatura, secuencia y peso molecular de estas inteínas se presentan en la Tabla 3 siguiente. Como se muestra en la Figura 1A, el fragmento N-terminal de cada construcción de inteína dividida consistió en (i) una etiqueta de purificación StreptagII (ST), (ii) la proteína de cabeza D del fago bacteriófago λ (gpD), que puede incrementar la solubilidad de la proteína, (iii) los cinco aminoácidos flanqueantes naturales que pertenecen a la N-exteína (E^N), (iv) el fragmento del extremo N de inteína dividida (I^N), y (v) la etiqueta hexa-histidina de purificación (H_6). El fragmento C-terminal de cada construcción de inteína dividida también se muestra en la Figura 1A y consistió en (i) el fragmento del extremo C de inteína dividida (I^C), (ii) los cinco aminoácidos flanqueantes naturales que pertenecen a la C-exteína (E^C), (iii) tiorredoxina, que puede incrementar la solubilidad de la proteína y ayudar en el plegamiento de la proteína, y (iv) la etiqueta hexa-histidina de purificación (H_6).

Todas estas proteínas de fusión se expresaron independientemente en *E. coli,* y las formas solubles se purificaron. Se mezclaron concentraciones equimolares (5-15 μM) de parejas de inteína dividida N- y C-terminal ($G1^N$+$G1^C$+, $G8^N$+$G8^C$, $N1^N$+$N1^C$, y I^N+I^C). Después de incubar a 25°C, la reacción de corte y empalme en trans se paró a diferentes puntos de tiempo hirviendo durante 5 minutos inmediatamente después de la adición de tampón de muestra con SDS. La reacción de corte y empalme en trans se resume en la Figura 1A.

Tabla 3: Resumen de la numeración, abreviatura, secuencia y pesos moleculares de las inteínas divididas de las proteínas de fusión que contienen las inteínas divididas. Las secuencias de Streptag II e His están subrayadas. El conector entre las inteínas divididas y las proteínas de interés (gpD o Trx) se indica en negrita e itálica. La secuencia de la exteína en el conector está enmarcada entre paréntesis.

Abr.	**Secuencia de proteína**	**PM kDa**
$G1^N$	MASWSHPQFEKAS-gpD-***GS[TRSGY]***-$Gp41.1^N$-GGHHHHHH (SEQ ID NO:2)	24,2
$G8^N$	MASWSHPQFEKAS-gpD-***GS[SQLNR]***-$Gp41.8^N$-GGHHHHHH (SEQ ID NO:11)	24,2
$N1^N$	MASWSHPQFEKAS-gpD-***GS[GTNPC]***-$NrdJ1^N$-GGHHHHHH (SEQ ID NO:19)	26,2
11^N	MASWSHPQFEKAS-gpD-***GS[GIGGG]***-$IMPDH1^N$-GGHHHHHH (SEQ ID NO:33)	25,8
DE^N	MASWSHPQFEKAS-gpD-***GS***-$DnaE^N$(SEQ ID NO:27)	24,7
$G1^C$	$Gp41.1^C$-***[SSSDV]GT-T***rx-EFRSHHHHHH (SEQ ID NO:6)	18,8
$G8^C$	$Gp41.8^C$-***[SAVEE]GT-****Trx*-EFRSHHHHHH (SEQ ID NO:15)	19,1
$N1^C$	$NrdJI^C$-***[SEIVL]GT-****Trx*-EFRSHHHHHH (SEQ ID NO:23)	18,6
$I1^C$	$IMPDH1^C$-***[SICST]*****GT-**Trx-EFRSHHHHHH (SEQ ID NO:37)	18,6
DE^C	$DnaE^C$-***[CFN]GT-****Trx*-EFRSHHHHHH (SEQ ID NO:30)	17,9
$G1^{N(C1A)}$	MASWSHPQFEKAS-gpD-***GS[TRSGY]***-$Gp41.1^{N(C1A)}$-GGHHHHHH (SEQ ID NO:56)	24,2
$G8^{N(C1A)}$	MASWSHPQFEKAS-gpD-***GS[SQLNR]***-$Gp41.8^{N(C1A)}$-GGHHHHHH (SEQ ID NO:57)	24,2
$N1^{N(C1A)}$	MASWSHPQFEKAS-gpD-***GS[GTNPC]***-$NrdJ1^{N(C1A)}$-GGHHHHHH (SEQ ID NO:58)	26,2

Abr.	Secuencia de proteína	PM kDa
$I1^{N(C1A)}$	MASWSHPQFEKAS-gpD-***GS[GIGGG]***-$IMPDH1^{N(C1A)}$-GGHHHHHH (SEQ ID NO:59)	25,8
$G1^{C(\Delta ext)}$	Gp41-1^{C}-*GT*-Trx-EFRSHHHHHH (SEQ ID NO:60)	18,3
$G1^{C(S)}$	Gp41-1^{C}-***[S]GT* Trx-**EFRSHHHHHH (SEQ ID NO:66)	18,3
$G8^{C(\Delta ext)}$	Gp41-8^{C}-***GT*-Trx-**EFRSHHHHHH (SEQ ID NO:61)	18,6
$N1^{C(\Delta ext)}$	NrdJ-1^{C}-***GT*-**Trx-EFRSHHHHHH (SEQ ID NO:62)	18,2
$I1^{C(\Delta ext)}$	IMPDH-1^{C}-***GT***-Trx-EFRSHHHHHH (SEQ ID NO:63)	18,1
$G1^{N(\Delta ext)}$	MASWSHPQFEKAS-gpD-***GS***-Gp41.1^{N}-GGHHHHHH (SEQ ID NO:67)	24,2
$G1^{C(N\rightarrow A)}$	Gp41.$1^{C(N\rightarrow A)}$-***[SSSDV]GT*-**Trx-EFRSHHHHHH (SEQ ID NO:68)	18,8
$G8^{C(N\rightarrow A)}$	Gp41.$8^{C(N\rightarrow A)}$-***[SAVEE]GT***-*T*rx-EFRSHHHHHH (SEQ ID NO:69)	19,1
$N1^{C(N\rightarrow A)}$	NrdJ$1^{C(N\rightarrow A)}$-***[SEIVL]GT***-Trx-EFRSHHHHHH (SEQ ID NO:70)	18,6
$I1^{C(N\rightarrow A)}$	IMPDH$1^{C(N\rightarrow A)}$***[SICST]*GT-**Trx-EFRSHHHHHH (SEQ ID NO:71)	18,6
$G1^{C(N/S\rightarrow A)}$	Gp41.$1^{C(N/S\rightarrow A)}$-***[ASSDV]GT***-Trx-EFRSHHHHHH (SEQ ID NO:72)	18,8
$G8^{C(NS\rightarrow A)}$	Gp41.$8^{C(N/S\rightarrow A)}$-***[AAVEE]GT***-Trx-EFRSHHHHHH (SEQ ID NO:73)	19,1
$N1^{C(N/S\rightarrow A)}$	NrdJ$1^{C(N/S\rightarrow A)}$-***[AEIVL]GT***- Trx-EFRSHHHHHH (SEQ ID NO:74)	18,6
$I1^{C(N/S\rightarrow A)}$	IMPDH$1^{C(N/S\rightarrow A)}$-***[AICST]GT***-Tr**x-**EFRSHHHHHH (SEQ ID NO:75)	18,6

P=Número de proteína, Abr=Abreviatura, MW=peso molecular, Gp41.1^{N}: fragmento N-terminal de la inteína dividida Gp41.1 (SEQ ID NO:3), Gp41.8^{N}: fragmento N-terminal de la inteína dividida Gp41.8 (SEQ ID NO: 12), NrdJ1^{N}: fragmento N-terminal de la inteína dividida Nrdj1 (SEQ ID NO: 20), IMPDH1^{N}: fragmento N-terminal de la inteína dividida IMPDH1 (SEQ ID NO: 34), DnaE^{N}: fragmento N-terminal de la inteína dividida DnaE (SEQ ID NO: 28), Gp41.1^{C}: fragmento C-terminal de la inteína dividida Gp41.1 (SEQ ID NO:7), Gp41.8^{C}: fragmento C-terminal de la inteína dividida Gp41.8 (SEQ ID NO: 16), NrdJ1^{C}: fragmento C-terminal de la inteína dividida Nrdj1 (SEQ ID NO: 24), IMPDH1^{C}: fragmento C-terminal de la inteína dividida IMPDH1 (SEQ ID NO: 38), DnaE^{C}: fragmento C-terminal de la inteína dividida DnaE (SEQ ID NO: 31), Trx: tiorredoxina de *E.coli* (SEQ ID NO:77); gpD: proteína de cabeza D del bacteriófago λ (SEQ ID NO:76).

Los experimentos se realizaron con todas las cuatro inteínas divididas (Gp41.1 (G1), Gp41.8(G8), NrdJ1(N1), y IMPDH1(I1)). Los resultados ejemplares obtenidos usando Gp41.1 (G1) se muestran en la Figura 1B. Los fragmentos N- y C-terminales iniciales (Figura 1B, carril 1: F1 y F2, respectivamente) reaccionaron muy rápidamente para producir el producto sometido a corte y empalme y los subproductos de la inteína dividida N y C (Figura 1B, carril 2-9: F3, F4, y F5, respectivamente). La velocidad de la reacción de corte y empalme en trans se calculó como la constante de velocidad "k", que es directamente proporcional a la velocidad de la reacción de corte y empalme en trans. También se calculó la vida media de la reacción "$t_{1/2}$", que representa el tiempo necesario para que la mitad de los precursores (F1 o F2) en una reacción de corte y empalme se consuma.

Sorprendentemente, todas las inteínas divididas analizadas (G1, G8, N1 y I1) fueron más rápidas que la Npu DnaE caracterizada previamente (Zettler J. et al, FEBS Letters 583:909-914 (2009)). En las mismas condiciones de reacción a 25°C, las G1, G8, N1 y I1 fueron respectivamente 31, 6, 9 y 7 veces más rápidas que Npu DnaE, que se

ha caracterizado como una inteína dividida excepcional que tiene la mayor constante de velocidad reportada hasta la fecha (Tabla 4). Los rendimientos de corte y empalme demostraron que, a 5 minutos, las G1, G8, N1 y I1 tienen aproximadamente 90% de formación de producto de corte y empalme.

Tabla 4: Porcentaje de corte y empalme de proteínas y constantes de velocidad de primer orden de la reacción de corte y empalme en trans de proteínas determinados para inteínas divididas.

Inteína	Temp °C	SP (%)	k (s^{-1})	$t_{1/2}$ (s)
G1	25	80-90	$5,7 \times 10^{-2}$	12
G8	25	85-95	$1,7 \times 10^{-2}$	40
N1	25	85-95	$6,2 \times 10^{-2}$	20
I1	25	90-95	$2,0 \times 10^{-2}$	34
DE	25	75-85	$3,4 \times 10^{-3}$	180
DE*	25	75-85	$3,5 \times 10^{-3}$	198
SP=producto de corte y empalme *Zettler J. et al 2009. FEBS Letters 583:909-914				

Los productos de corte y empalme de G1, G8, N1 y I1 se identificaron por espectrometría de masa LC-MS/MS (>90% cobertura de secuencia). Los pesos moleculares determinados fueron consistentes con el valor teórico de 27,3 kDa para todos ellos.

Dadas estas propiedades destacables, este grupo de inteínas divididas naturales parece ser una nueva generación de inteínas divididas ultra-rápidas que pueden usarse para muchas aplicaciones incluyendo ingeniería de proteínas, química celular, ciclación, purificación y otras.

Ejemplo 2: Efecto de la temperatura en la actividad de corte y empalme en trans de Gp41.1

Con el fin de ensayar la versatilidad y robustez de estas inteínas divididas a diferentes temperaturas, Gp41.1 se analizó con más detalle. Se ha mostrado que la actividad de las inteínas se ve afectada por la temperatura. La evidencia de actividad de de corte y empalme de proteínas mediada por inteínas divididas naturales Ssp DnaE y semisintéticas Mtu RecA a baja temperatura tal como 4°C se ha reportado previamente (Martin, D. et al. 2001. Biochemistry, 40:1393-1402 y Lew, B. et al. 1999. Biopolymers (Peptide Science), 51:355-362), pero la actividad a menos de 4°C no se ha documentado previamente, según nuestro conocimiento. Por lo tanto, la actividad de GP41.1 se ensayó a varias temperaturas. Se mezclaron fragmentos N- y C-terminales purificados de Gp41.1 en tampón de corte y empalme a una concentración equimolar de 5µM, y se incubó a 0, 12, 25 y 37°C. La formación del producto de corte y empalme y las velocidades de constante se determinaron, y los resultados se muestran en la Tabla 5.

Sorprendentemente, la inteína Gp41.1 todavía era activa a 0°C. Tenía una $k=5,5 \times 10^{-3} s^{-1}$, y después de 1 hora de reacción bajo dichas condiciones extremas, se formó entre 80-90% del producto de corte y empalme. Cuando la misma reacción se realizó a 12°C, la velocidad de corte y empalme en trans se incrementó hasta una $k=1,3 \times 10^{-2} s^{-1}$ (más rápida que Npu DnaE a 12°C con una $k=2,2 \pm 0.5 \times 10^{-3} s^{-1}$), y después de 1 hora de reacción, se formó entre 85-90% del producto de corte y empalme. A 25°C, la inteína dividida Gp41.1 presentó una $k=5,2 \times 10^{-2} s^{-1}$ (también más rápida que Npu DnaE a 25°C, $k=3,5 \pm 0,2 \times 10^{-3} s^{-1}$) con 90-95% de formación del producto de corte y empalme a los 30 minutos.

También se ha mostrado previamente que muchas inteínas divididas, incluyendo Ssp DnaE y Mtu RecA, presentan rendimientos reducidos y formación incrementada de subproductos de hidrólisis a temperaturas mayores (es decir, 37°C). Otras inteínas, sin embargo, tales como Npu DnaE, son más activas a 37°C. La inteína dividida Gp41.1 presentó su mayor velocidad de actividad a 37°C: $k=1,1 \times 10^{-1} s^{-1}$ (más rápida de nuevo que Npu DnaE a 37°C, $k=1,1 \pm 0,2 \times 10^{-2} s^{-1}$). Notablemente, 90-95% del producto de corte y empalme se había formado en 5 minutos. Estos resultados indican que G1 tiene un rango amplio (de 0 a 37°C) de tolerancia a la temperatura.

Tabla 5: Efecto de la temperatura sobre la actividad de corte y empalme en trans de G 1.

Temp (°C)	k (s^{-1})	$t_{1/2}$ (s)	SP (%)
0	5,5 $x10^{-1}$	126	80-90
12	1,8 $x10^{-2}$	39	80-90
25	5,7 $x10^{-2}$	12	90-95
16,8	1,4 $x10^{-1}$	5	90-95
45	1,8 $x10^{-1}$	4	85-95
50	1,2 $x10^{-1}$	6	85-95
55	8,3 $x10^{-2}$	8	65-75

De forma interesante, a todas las temperaturas analizadas, la actividad de corte y empalme en trans de G1 fue más rápida que la inteína dividida de alto rendimiento Npu DnaE. A 12°C y 25°C, G1 fue 6 y 15 veces más rápida que Npu DnaE, e incluso a 37°C, donde Npu DnaE presenta su actividad más rápida, G1 tuvo una actividad 10 veces más rápida.

Ejemplo 3: Efecto del pH y sal caotrópica en la actividad de corte y empalme en trans

Se ha mostrado que muchas inteínas divididas incluyendo las inteínas Ssp DnaE, y Mtu RecA, presentan rendimientos reducidos y formación incrementada de subproductos de hidrólisis a alto pH o en presencia de agentes desnaturalizantes (Zettler et al., 2009. FEBS letters 583: 909-914). Sin embargo, la eficiencia del corte y empalme de G1 era casi independiente del pH entre 6 y 9 (tabla 6). Sólo se observó una disminución de la actividad a valores de pH extremos tales como 4 y 10. Mtu RecA, por el contrario, tiene un rango de corte y empalme óptimo mucho más estrecho de entre pH 6 y 7,5 (Lew B. et al. Biopolymers. 51:355-362 (1999)). La inteína Ssp DnaE presenta una actividad de corte y empalme en trans máxima a pH 7,0, pro cae a pH mayor (Martin D. et al. 2001. Biochemistry. 40:1393-1402).

Tabla 6: Efecto del pH y la presencia de sales caotrópicas sobre la actividad de corte y empalme en trans de G1

pH	Sal caotrópica	SP (%) en 1 hora
6	-	80-85
7	-	80-85
	Urea 4M	40-45
8	-	80-85
9	-	75-80

En algunos casos, la expresión de proteínas de fusión de inteína podría requerir la presencia de algún agente desnaturalizante para incrementar su solubilidad. Por lo tanto, la tolerancia hacia la presencia de 4M urea en la reacción de corte y empalme también se determinó para la inteína dividida Gp41.1. Los resultados presentados en la tabla 4 muestran una actividad de corte y empalme significativa en presencia de una concentración moderada de urea (4M), con casi el 50% de la actividad de corte y empalme en trans después de 1 hora de reacción respecto a las condiciones optimizadas.

Estas características demuestran la versatilidad inesperada y la robustez de la inteína dividida G1.

Ejemplo 4: Efecto de las exteínas en la actividad de corte y empalme en trans

Los aminoácidos de N-exteína (E^N) que flanquean el dominio de N-inteína no participan directamente en la reacción de corte y empalme en trans, pero podrían tener una influencia en la eficiencia de la reacción. Con el fin de ensayar esta posibilidad, los cinco aminoácidos de la E^N de G1 (TRSGY) se delecionaron, y el nuevo fragmento N-terminal generado ($G1^{N(\Delta ext)}$) se incubó con el fragmento C-terminal correspondiente ($G1^C$) en las mismas condiciones descritas en el ejemplo 1. De forma interesante, se observó actividad de corte y empalme en trans en ausencia de E^N (valor de velocidad 1,8E-3 s^{-1} y un rendimiento de 45%), demostrando que E^N de la inteína dividida G1 no es esencial.

Por el contrario, se ha descrito que el primer aminoácido de la C-exteína está implicado directamente en la reacción de corte y empalme en trans mediada por inteínas e inteínas divididas. Con el fin de ensayar el papel de la serina localizada en la primera posición de la N-exteína G 1 se generaron dos nuevos fragmentos C-terminales de G1. Comprendieron un fragmento C-terminal en el que el dominio E^C fue: (i) parcialmente (sólo se mantuvo la serina que flanquea $\mathbf{I}^N$; $G1^{C(s)}$) o (ii) completamente eliminado ($G1^{C(\Delta ext)}$).

La incubación de $G1^{C(s)}$ con el fragmento N-terminal correspondiente, bien que contenía o carecía de E^N ($G1^N$ o $G1^{N(\Delta ext)}$ respectivamente) resultó en una reacción de corte y empalme en trans eficiente. Por el contrario, cuando $G1^N$ o $G1^{N(\Delta ext)}$) se incuban en presencia de $G1^{C(\Delta ext)}$, no se observa reacción de corte y empalme en trans. En esta situación específica, sorprendientemente, se observa actividad de escisión en C y N. A partir de estos resultados, puede concluirse que un residuo de serina cercano a la IN, y preferiblemente que lo flanquea, es necesario para asegurar una reacción de corte y empalme en trans eficiente.

Ejemplo 5: Efecto sobre la auto-escisión C-terminal de la mutación puntual C1A en la IN en las inteínas divididas Gp41.1, Gp41.8, NrdJ1 y IMPDH1

Algunas inteínas muestran una actividad de auto-escisión independiente en el extremo N o extremo C y requieren residuos de aminoácidos distintos. La mutación en Cys1 a Ala (C1A) en IN inhabilita la escisión en el extremo N pero no el extremo C (referido hasta ahora como auto-escisión C-terminal), mientras la mutación en la Asn154 C-terminal a Ala en I^C inhabilita la escisión en el extremo C pero no el extremo N (Mathys, S. et al. Gene 231:1-13 (1999) y Lu et al. J. Chromatography A. 1218:2553-2560 (2011)). Debido a esta interesante propiedad, algunas inteínas mutadas pueden usarse como péptidos auto-escindibles que permiten una liberación controlada de la proteína de interés de proteínas de fusión. Así, dichas inteínas mutadas pueden usarse en lugar de proteasas comerciales costosas.

En todas las inteínas divididas naturales analizadas hasta la fecha, la mutación C1A suprime el corte y empalme de proteínas. La mutación C1A en las inteínas divididas naturales Npu DnaE y Ssp DnaE bloquea el desplazamiento inicial acilo N a S y bloquea el corte y empalme de proteínas, pero también inhibe casi completamente la reacción de escisión C-terminal (Zettler J., et al. 2009. FEBS Letters 583:909-914) y Ssp DnaE (Martin, D. et al. 2001. Biochemistry. 40:1393-1402). Además, se ha reportado que la actividad de auto-escisión C-terminal se inhibe hasta 90% en la inteína Pab PolII natural de longitud completa con mutación puntual en S1A (Xu, M. y Perler, F. EMBO J. 15:5146-5153 (1996)).

Con el fin de ensayar la actividad de auto-escisión C-terminal, se introdujo la mutación Cys1 a Ala (C1A) en todas las inteínas divididas IN ($G1^{N(C1A)}$, $G8^{N(C1A)}$, $N1^{N(C1A)}$, y $I1^{N(C1A)}$). La numeración y representación esquemática de estas construcciones con mutación puntual se presentan en la Tabla 3 y Fig 2A, respectivamente. Los fragmentos F1 ($I^{N(C1A)}$) y F2 (I^C) purificados de homogenados de *E. coli* se mezclaron a concentraciones equimolares de 5-15 uM, y se realizaron experimentos de curso de tiempo a 25°C. Sorprendentemente, a diferencia de las inteínas divididas naturales Npu DnaE y Ssp DnaE, todas las cuatro inteínas divididas ensayadas mostraron auto escisión C-terminal. En todos los casos, se observaron dos nuevas bandas de proteína que correspondieron en tamaño al fragmento esperado F3 (Trx-H6) y al escindido F4 (Int^C) (Fig 2A). Los valores del rendimiento (% CP) y la constante de velocidad de la reacción de escisión C-terminal a 25°C se muestra en la tabla siguiente.

Tabla 7: Rendimiento de la reacción de escisión C-terminal

Inteína	Temp (°C)	k (s^{-1})	$t_{1/2}$ (min)	CP (%)
G1	25	5,95E-04	19,4	85-95
G8	25	9,50E-05	121,6	85-95
N1	25	2,70E-04	43	85-95
I1	25	3,65E-04	31,6	85-95

Inteína	Temp (°C)	k (s^{-1})	$t_{1/2}$ (min)	CP (%)
CP=Producto escindido				

A diferencia de la exteína N-terminal (E^N), la exteína C-terminal (E^C) participa indirectamente en la reacción de auto-escisión C-terminal. Se cree que la E^C proporciona un entorno apropiado a la I^C para asegurar una reacción de auto-escisión C-terminal eficiente (Zettler J. et al, FEBS Letters 583:909-914 (2009); Lu L. et al, J. Chromatography A. 1218:2553-2560 (2011); Nichols N. et al. Biochemistry. 42:5301-5311 (2003); y Appleby et al., JBC 284:6194-6199 (2009)). Este requerimiento puede ser una limitación importante para varias aplicaciones porque la secuencia de exteína permanecerá unida a la proteína de interés después de la reacción de auto-escisión.

Los 5 aminoácidos de la secuencia de exteína (E^C) que flanquean las inteínas divididas C-terminales se eliminaron. Las construcciones correspondientes ($G1^{C(\Delta ext)}$, $G8^{C(\Delta ext)}$, $N1^{C(\Delta ext)}$, y $I1^{C(\Delta ext)}$, véase la tabla 1) presentaron una unión directa entre la I^C y el gen Trx. Para propósitos de clonación, se mantuvo el sitio de escisión KpnI, pero la presencia de los aminoácidos extra GT no afecta el análisis del papel de las exteínas porque no comparten homología con la secuencia de exteína flanqueante y pueden considerarse como parte de la proteína Trx.

Los fragmentos F1 (gpD-$IN^{(C1A)}$) y F2 (I_C-Trx) correspondientes a Gp41.1 (G1), Gp41.8(G8), NrdJ1(N1), y IMPDH1(I1) se purificaron de homogenados de E. *coli* y se mezclaron a concentraciones equimolares de 5-15 µM. Se realizaron experimentos de curso de tiempo a 25°C. Sorprendentemente, todas las cuatro inteínas divididas naturales mostraron auto-escisión C-terminal, y se observaron dos bandas con una movilidad consistente con F3 (Trx) y F4 (I^C) después de 3 horas de incubación (Fig 2B). Se determinó el rendimiento (% CP) y la constante de velocidad de la reacción de escisión C-terminal a 25°C de G1 y N1. Una comparación de las Tablas 4 y 5 demuestra que el rendimiento es muy alto y es independiente de la presencia de los cinco aminoácidos de la E^C. La ausencia del fragmento E^C resulta en una reducción de la constante de velocidad, pero esta disminución en la velocidad de la reacción de la inteína dividida puede superarse incrementando la temperatura de reacción hasta 37 ó 45°C (Tabla 8). Esta observación inesperada sugiere que estas inteínas pueden funcionar muy eficientemente a altas temperaturas, incluso si estas proteínas no provienen de microorganismos termofílicos.

La secuenciación de proteínas de Edman de los fragmentos F3 liberados en la reacción de escisión C-terminal demostró que los primeros aminoácidos en el fragmento F3 fueron GT. Esto demuestra que la reacción de escisión C-terminal se realizó apropiadamente.

Tabla 8: Rendimiento de la reacción de escisión C-terminal

Inteína	Temp (°C)	k (s')	$t_{1/2}$ (min)	CP (%)
G1	25	9,00E-05	128	85-95
	37	2,4E-04	48	85-95
	45	5,2E-04	9,9	85-95
N1	25	4,00E-05	144	85-95
CP=Producto escindido				

Estos resultados demuestran que, a diferencia de Npu DnaE (Zettler J. et al, FEBS Letters 583:909-914 (2009)), Ssp DnaE (Nichols N. et al., Biochemistry 42:5301-5311 (2003)), y Ssp DnaB (Lu L. et al, J. Chromatography A. 1218:2553-2560 (2011)) Gp41.1 (G1), Gp41.8 (G8), NrdJ1 (N1), y IMPDH1 (I1) son capaces de presentar una escisión C-terminal en ausencia del fragmento flanqueante de C-exteína de 5 aminoácidos (E^C).

En los experimentos descritos anteriormente, se ha demostrado que la escisión en C puede realizarse en ausencia de la E^C. No obstante, no se analizó la influencia que tiene la naturaleza de los primeros aminoácidos después de la I^C sobre la eficiencia de la escisión en C. Para aplicaciones de la escisión en C, se prefiere que la proteína liberada de la proteína de fusión escindida no contenga ningún aminoácido extra en su extremo N-terminal. Con el fin de determinar si la inteína dividida G1 era capaz de producir una escisión "limpia" independientemente del primer aminoácido de la proteína de interés, se realizó una nueva batería de construcciones. En estas construcciones, el primer aminoácido de la proteína Trx se mutó a todas las variantes naturales de aminoácidos, y se clonó

directamente en el dominio C-terminal de G1. Sorprendentemente, todas las variantes mostraron una actividad de escisión en C significativa.

El rendimiento, constante de velocidad y t(1/2) determinados para cada construcción se resumen en la Tabla 9.

Tabla 9: Resumen del rendimiento y parámetros cinéticos (k y t(1/2)) de la escisión en C de varias construcciones de G1. * Corresponde al primer residuo después de la I^C. ** Serina es el aminoácido natural encontrado en G1

	Escisión C-Terminal		
	Rendimiento	k	t(1/2)
Residuo*	(%)	s-1	min
Met	85,63	8,70E-04	13,28
Cys	83,62	6,83E-04	16,91
Thr	87,51	5,73E-04	20,15
Arg	87,32	5,10E-04	22,65
Lys	86,71	5,10E-04	22,65
Ser**	89,40	4,83E-04	23,90
Gln	88,85	4,70E-04	24,58
His	71,13	4,47E-04	25,86
Ala	75,82	3,57E-04	32,39
Tyr	91,34	3,30E-04	35,01
Phe	72,60	3,07E-04	37,67
Asn	71,20	2,87E-04	40,30
Trp	94,12	2,80E-04	41,26
Val	86,68	2,73E-04	42,27
Leu	69,32	2,73E-04	42,27
Asp	68,61	2,47E-04	46,83
Ile	88,37	1,73E-04	66,65
Gly	75,18	1,37E-04	84,53
Glu	96,18	1,27E-04	91,20
Pro	57,82	6,00E-05	192,54

Ejemplo 6: Auto-escisión N-terminal con las inteínas divididas Gp41.1, Gp41.8, NrdJ1 y IMPDH1.

Se ha descrito para varias inteínas e inteínas divididas que el bloqueo de la escisión C-terminal mediante la mutación de la Asn justo aguas arriba de la exteína todavía permite que ocurra la escisión en N. Esta mutación se introdujo en el dominio I^C de las inteínas divididas Gp41.1, Gp41.8, NrdJ1 y IMPDH1 ($G1^{C(N\rightarrow A)}$, $G8^{C(N\rightarrow A)}$, $N1^{C(N\rightarrow A)}$ y $I1^{C(N\rightarrow A)}$ respectivamente) reemplazando la Asn justo aguas arriba de la exteína por una Ala. Como se ha descrito previamente, estas proteínas de fusión se produjeron en *E coli,* se purificaron y se incubaron posteriormente con cantidades equivalentes de los equivalentes correspondientes $G1^N$, $G8^N$, $N1^N$ y $I1^N$, esencialmente como se ha descrito previamente. Sorprendentemente, la reacción de auto-escisión N-terminal fue altamente ineficiente, como puede concluirse por la observación de que sólo uno de los dos productos esperados de auto-escisión N-terminal se observó claramente en geles de SDS-PAGE (I^N-H6). El segundo producto esperado correspondiente al ST-gpD-E^N también se observó, pero como una banda muy débil demostrando que la reacción de auto-escisión N-terminal fue muy ineficiente. Además, se observó un producto intermedio que correspondía probablemente al fragmento C-terminal ($G1^{C(N\rightarrow A)}$, $G8^{C(N\rightarrow A)}$, $N1^{C(N\rightarrow A)}$ o $I1^{C(N\rightarrow A)}$) unido al ST-gpD-E^N como un subproducto principal.

Con el fin de incrementar la eficiencia de la escisión en N, se realizó una segunda ronda de mutaciones puntuales en todos los mutantes únicos ensayados previamente ($G1^{C(N\rightarrow A)}$, $G8^{C(N\rightarrow A)}$, $N1^{C(N\rightarrow A)}$ y $I1^{C(N\rightarrow A)}$). El primer residuo de Ser de la E^C se mutó a Ala. Los nuevos dobles mutantes generados ($G1^{C(NS\rightarrow A)}$, $G8^{C(NS\rightarrow A)}$, $N1^{C(NS\rightarrow A)}$ y $I1^{C(NS\rightarrow A)}$) se expresaron en *E coli,* y las proteínas purificadas se incubaron a 25°C con una concentración equivalente del equivalente correspondiente ($G1^N$, $G8^N$, $N1^N$ y $I1^N$, respectivamente). En este caso, el análisis por SDS-PAGE mostró que la auto-escisión N-terminal ocurría muy eficientemente. En conclusión, la mutación del residuo de Ser en la primera posición de la E^C combinada con la mutación en el último residuo de Asn de la I^C permite una escisión N-terminal eficiente. Por ejemplo, la incubación del $G1^{C(N/S\rightarrow A)}$ purificado con el $G1^N$, resultó en una velocidad de escisión N-terminal de 5,7 E-4 s^{-1} y un rendimiento de 70%.

Todas las publicaciones, patentes, solicitudes de patentes, sitios de internet, y números de acceso/secuencias de bases de datos (incluyendo tanto secuencias de polinucleótidos como de polipéptidos) citadas en la presente memoria se incorporan por ésta por referencia en su totalidad para todos los propósitos en el mismo grado que si se indicara específicamente e individualmente que cada publicación, solicitud de patente, sitio de internet, o número de acceso/secuencia de bases de datos individual estuviera así incorporada por referencia.

Debe apreciarse que se pretende que la sección de Descripción Detallada, y no las secciones de Compendio y Resumen, se usen para interpretar las reivindicaciones. Las secciones de Compendio y Resumen pueden mostrar una o más pero no todas las realizaciones ejemplares de la presente invención según se contempla por el o los inventores y, así, no se pretende que limiten la presente invención y las reivindicaciones adjuntas de ninguna manera.

La descripción anterior de las realizaciones específicas revelará tan completamente la naturaleza general de la invención que otros pueden, aplicando el conocimiento en la experiencia en la técnica, modificar y/o adaptar fácilmente para varias aplicaciones dichas realizaciones específicas, sin experimentación excesiva, sin alejarse del concepto general de la presente invención. Por lo tanto, se pretende que dichas adaptaciones y modificaciones estén en el significado y rango de equivalentes de las realizaciones descritas, tomando como base la enseñanza y guía presentadas en la presente memoria. Debe entenderse que la fraseología o terminología de la presente memoria es para el propósito de descripción y no de limitación, de manera que esa terminología o fraseología de la presente especificación debe interpretarse por el experto en la técnica a la luz de las enseñanzas y guía.

La amplitud y alcance de la presente invención no deben estar limitados por ninguna de las realizaciones ejemplares descritas anteriormente, sino que debe definirse sólo según las reivindicaciones siguientes y sus equivalentes.

Listado de secuencias

<110> ERA BIOTECH, S.A.

<120> INTEÍNAS DIVIDIDAS Y USOS DE ÉSTAS

<130> P7749PC00

<150> US 61/540101
<151> 28-09-2011

<150> EP12171848
<151> 13-06-2012

<160> 108

<170> PatentIn versión 3.5

<210> 1
<211> 678
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del fragmento N de GP-41.1

<400> 1

```
ccatggccag ttggagccac ccgcagttcg aaaaagcgag caaagaaacc tttacccatt      60

accagccgca gggcaacagt gacccggctc ataccgcaac cgcgcccggc ggattgagtg     120

cgaaagcgcc tgcaatgacc ccgctgatgc tggacacctc cagccgtaag ctggttgcgt     180

gggatggcac caccgacggt gctgccgttg gcattcttgc ggttgctgct gaccagacca     240

gcaccacgct gacgttctac aagtccggca cgttccgtta tgaggatgtg ctctggccgg     300

aggctgccag cgacgagacg aaaaaacgga ccgcgtttgc cggaacggca atcagcatcg     360

ttggatccac ccgtagcggt tattgcctgg acctgaaaac ccaggtgcag accccgcagg     420

gcatgaagga gattagcaac attcaggtgg gcgacctggt tctgagcaac accggctata     480

atgaggtgct gaacgtgttc ccgaagagca aaaagaagag ctacaagatc acgctggagg     540

acggcaagga aatcatttgc agcgaagaac atctgtttcc gacccagacc ggcgaaatga     600

atattagcgg tggcctgaaa gaaggcatgt gcctgtatgt gaaagagggc ggtcaccacc     660

atcatcacca ctaagctt                                                  678
```

<210> 2
<211> 223
<212> PRT
<213> Secuencia Artificial

<220>
<223> Proteína del fragmento N de GP-41.1

<400> 2

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
```

```
1               5                   10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Thr Arg Ser Gly Tyr Cys
        115                 120                 125

Leu Asp Leu Lys Thr Gln Val Gln Thr Pro Gln Gly Met Lys Glu Ile
    130                 135                 140

Ser Asn Ile Gln Val Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr Asn
145                 150                 155                 160

Glu Val Leu Asn Val Phe Pro Lys Ser Lys Lys Lys Ser Tyr Lys Ile
                165                 170                 175

Thr Leu Glu Asp Gly Lys Glu Ile Ile Cys Ser Glu Glu His Leu Phe
            180                 185                 190

Pro Thr Gln Thr Gly Glu Met Asn Ile Ser Gly Gly Leu Lys Glu Gly
        195                 200                 205

Met Cys Leu Tyr Val Lys Glu Gly Gly His His His His His His
    210                 215                 220
```

<210> 3
<211> 88
<212> PRT
<213> Secuencia Artificial

<220>
<223> GP 41.1 (InteínaN)

<400> 3

```
Cys Leu Asp Leu Lys Thr Gln Val Gln Thr Pro Gln Gly Met Lys Glu
1               5                   10                  15

Ile Ser Asn Ile Gln Val Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr
            20                  25                  30

Asn Glu Val Leu Asn Val Phe Pro Lys Ser Lys Lys Lys Ser Tyr Lys
        35                  40                  45

Ile Thr Leu Glu Asp Gly Lys Glu Ile Ile Cys Ser Glu Glu His Leu
    50                  55                  60

Phe Pro Thr Gln Thr Gly Glu Met Asn Ile Ser Gly Gly Leu Lys Glu
65                  70                  75                  80

Gly Met Cys Leu Tyr Val Lys Glu
                85
```

<210> 4
<211> 5
<212> PRT
<213> Secuencia Artificial

<220>
<223> GP 41.1 (ExteínaN)

<400> 4

```
Thr Arg Ser Gly Tyr
1               5
```

<210> 5
<211> 520
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del fragmento C de GP41.1

<400> 5

```
catatgggca aaaacagcat gatgctgaag aagatcctga agatcgagga gctggacgag       60

cgcgagctga ttgatatcga agtgagcggc aaccacctgt tctacgccaa tgacattctg      120

acgcataata gcagcagcga tgtgggtacc ggatctgata aaattattca tctgactgat      180

gattcttttg atactgatgt acttaaggca gatggtgcaa tcctggttga tttctgggca      240

cactggtgcg gtccgtgcaa aatgatcgct ccgattctgg atgaaatcgc tgacgaatat      300

cagggcaaac tgaccgttgc aaaactgaac atcgatcaca acccgggcac tgcgccgaaa      360

tatggcatcc gtggtatccc gactctgctg ctgttcaaaa acggtgaagt ggcggcaacc      420

aaagtgggtg cactgtctaa aggtcagttg aaagagttcc tcgacgctaa cctggccggc      480

tctgaattca gatctcatca ccatcaccat cactaagctt                            520
```

<210> 6
<211> 170
<212> PRT
<213> Secuencia Artificial

<220>
<223> Proteína del fragmento C de GP41.1

<400> 6

```
Met Gly Lys Asn Ser Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu
1               5                  10                  15

Leu Asp Glu Arg Glu Leu Ile Asp Ile Glu Val Ser Gly Asn His Leu
            20                  25                  30

Phe Tyr Ala Asn Asp Ile Leu Thr His Asn Ser Ser Ser Asp Val Gly
        35                  40                  45

Thr Gly Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr
    50                  55                  60

Asp Val Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His
65                  70                  75                  80

Trp Cys Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala
                85                  90                  95

Asp Glu Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His
            100                 105                 110

Asn Pro Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu
        115                 120                 125

Leu Leu Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu
    130                 135                 140

Ser Lys Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser
145                 150                 155                 160

Glu Phe Arg Ser His His His His His His
                165                 170
```

<210> 7
<211> 37
<212> PRT
<213> Secuencia Artificial

<220>
<223> GP 41.1 (InteínaC)

<400> 7

```
Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu Leu Asp Glu Arg Glu
1               5                  10                  15

Leu Ile Asp Ile Glu Val Ser Gly Asn His Leu Phe Tyr Ala Asn Asp
            20                  25                  30

Ile Leu Thr His Asn
        35
```

<210> 8

<211> 5
<212> PRT
<213> Secuencia Artificial

<220>
<223> GP 41.1 (ExteínaC)

<400> 8

Ser Ser Ser Asp Val
1 5

<210> 9
<211> 5
<212> PRT
<213> Secuencia Artificial

<220>
<223> Potenciador de E coli

<400> 9

Met Gly Lys Asn Ser
1 5

<210> 10
<211> 681
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del fragmento N de GP 41.8

<400> 10

ccatggccag ttggagccac ccgcagttcg aaaaagcgag caaagaaacc tttacccatt 60

accagccgca gggcaacagt gacccggctc ataccgcaac cgcgcccggc ggattgagtg 120

cgaaagcgcc tgcaatgacc ccgctgatgc tggacacctc cagccgtaag ctggttgcgt 180

gggatggcac caccgacggt gctgccgttg gcattcttgc ggttgctgct gaccagacca 240

gcaccacgct gacgttctac aagtccggca cgttccgtta tgaggatgtg ctctggccgg 300

aggctgccag cgacgagacg aaaaaacgga ccgcgtttgc cggaacggca atcagcatcg 360

ttggatccag ccaactgaat cgttgcctga gcctggatac gatggttgtg accaatggca 420

aagcgattga gattcgtgat gtgaaagtgg gcgattggct ggaaagcgaa tgtggcccgg 480

tgcaggtgac cgaagtgctg ccgattatca agcagccggt gtttgaaatt gtgctgaaga 540

gcggcaaaaa gatccgtgtg agcgcgaatc ataaattccc gaccaaagat ggcctgaaaa 600

ccatcaatag cggtctgaaa gttggcgact tcctgcgtag ccgtgcgaaa ggcggccatc 660

atcaccacca tcactaagct t 681

<210> 11
<211> 224
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del fragmento N de GP 41.8

<400> 11

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1               5                   10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Ser Gln Leu Asn Arg Cys
        115                 120                 125

Leu Ser Leu Asp Thr Met Val Val Thr Asn Gly Lys Ala Ile Glu Ile
    130                 135                 140

Arg Asp Val Lys Val Gly Asp Trp Leu Glu Ser Glu Cys Gly Pro Val
145                 150                 155                 160

Gln Val Thr Glu Val Leu Pro Ile Ile Lys Gln Pro Val Phe Glu Ile
                165                 170                 175

Val Leu Lys Ser Gly Lys Lys Ile Arg Val Ser Ala Asn His Lys Phe
            180                 185                 190

Pro Thr Lys Asp Gly Leu Lys Thr Ile Asn Ser Gly Leu Lys Val Gly
        195                 200                 205

Asp Phe Leu Arg Ser Arg Ala Lys Gly Gly His His His His His His
    210                 215                 220
```

<210> 12
<211> 89
<212> PRT
<213> Secuencia Artificial

<220>
<223> GP41.8 (Inteína-N)

<400> 12

```
Cys Leu Ser Leu Asp Thr Met Val Val Thr Asn Gly Lys Ala Ile Glu
1               5                  10                  15

Ile Arg Asp Val Lys Val Gly Asp Trp Leu Glu Ser Glu Cys Gly Pro
            20                  25                  30

Val Gln Val Thr Glu Val Leu Pro Ile Ile Lys Gln Pro Val Phe Glu
        35                  40                  45

Ile Val Leu Lys Ser Gly Lys Lys Ile Arg Val Ser Ala Asn His Lys
    50                  55                  60

Phe Pro Thr Lys Asp Gly Leu Lys Thr Ile Asn Ser Gly Leu Lys Val
65                  70                  75                  80

Gly Asp Phe Leu Arg Ser Arg Ala Lys
                85
```

<210> 13
<211> 5
<212> PRT
<213> Secuencia Artificial

<220>
<223> GP41.8 (Exteína-N)

<400> 13

```
Ser Gln Leu Asn Arg
1               5
```

<210> 14
<211> 529
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del extremo C de GP41.8

<400> 14

```
catatgtgcg agatcttcga gaacgagatc gactgggatg aaatcgcgag cattgagtat      60

gtgggcgttg aggagaccat tgacatcaac gtgacgaacg accgcctgtt cttcgcaaac     120

ggcattctga cccataatag cgcggtggaa gagggtaccg gatctgataa aattattcat     180

ctgactgatg attcttttga tactgatgta cttaaggcag atggtgcaat cctggttgat     240

ttctgggcac actggtgcgg tccgtgcaaa atgatcgctc cgattctgga tgaaatcgct     300

gacgaatatc agggcaaact gaccgttgca aaactgaaca tcgatcacaa cccgggcact     360

gcgccgaaat atggcatccg tggtatcccg actctgctgc tgttcaaaaa cggtgaagtg     420

gcggcaacca aagtgggtgc actgtctaaa ggtcagttga aagagttcct cgacgctaac     480

ctggccggct ctgaattcag atctcatcac catcaccatc actaagctt                 529
```

<210> 15
<211> 173
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA GP41.8

<400> 15

Met Cys Glu Ile Phe Glu Asn Glu Ile Asp Trp Asp Glu Ile Ala Ser
1 5 10 15

Ile Glu Tyr Val Gly Val Glu Glu Thr Ile Asp Ile Asn Val Thr Asn
20 25 30

Asp Arg Leu Phe Phe Ala Asn Gly Ile Leu Thr His Asn Ser Ala Val
35 40 45

Glu Glu Gly Thr Gly Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser
50 55 60

Phe Asp Thr Asp Val Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe
65 70 75 80

Trp Ala His Trp Cys Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp
85 90 95

Glu Ile Ala Asp Glu Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn
100 105 110

Ile Asp His Asn Pro Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile
115 120 125

Pro Thr Leu Leu Leu Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val
130 135 140

Gly Ala Leu Ser Lys Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu
145 150 155 160

Ala Gly Ser Glu Phe Arg Ser His His His His His His
165 170

<210> 16
<211> 45
<212> PRT
<213> Secuencia Artificial

<220>
<223> GP41.8 (InteínaC)

<400> 16

Met Cys Glu Ile Phe Glu Asn Glu Ile Asp Trp Asp Glu Ile Ala Ser
1 5 10 15

Ile Glu Tyr Val Gly Val Glu Glu Thr Ile Asp Ile Asn Val Thr Asn
20 25 30

Asp Arg Leu Phe Phe Ala Asn Gly Ile Leu Thr His Asn
35 40 45

<210> 17
<211> 5

<212> PRT
<213> Secuencia Artificial

<220>
<223> GP41.8 (ExteínaC)

<400> 17

Ser Ala Val Glu Glu
1 5

<210> 18
<211> 729
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del extremo N de NrdJ1

<400> 18

ccatggccag ttggagccac ccgcagttcg aaaaagcgag caaagaaacc tttacccatt 60

accagccgca gggcaacagt gacccggctc ataccgcaac cgcgcccggc ggattgagtg 120

cgaaagcgcc tgcaatgacc ccgctgatgc tggacacctc cagccgtaag ctggttgcgt 180

gggatggcac caccgacggt gctgccgttg gcattcttgc ggttgctgct gaccagacca 240

gcaccacgct gacgttctac aagtccggca cgttccgtta tgaggatgtg ctctggccgg 300

aggctgccag cgacgagacg aaaaaacgga ccgcgtttgc cggaacggca atcagcatcg 360

ttggatccgg caccaatccg tgttgcctgg tgggcagcag cgagatcatc acccgtaact 420

acggcaaaac cacgatcaaa gaggtggttg agatcttcga caacgacaag aatatccagg 480

tgctggcgtt caacacccac acggacaata tcgaatgggc cccaattaaa gcggcgcaac 540

tgacccgtcc aaacgcagag ctggtggaac tggaaattaa caccctgcat ggcgtgaaaa 600

ccatccgttg caccccggat catccagtgt ataccaaaaa tcgtgactat gtgcgcgccg 660

atgagctgac cgatgatgat gaactggtgg tggcgattgg cggccatcac caccatcacc 720

actaagctt 729

<210> 19
<211> 240
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo N NrdJ1

<400> 19

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1               5                  10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Gly Thr Asn Pro Cys Cys
        115                 120                 125

Leu Val Gly Ser Ser Glu Ile Ile Thr Arg Asn Tyr Gly Lys Thr Thr
    130                 135                 140

Ile Lys Glu Val Val Glu Ile Phe Asp Asn Asp Lys Asn Ile Gln Val
145                 150                 155                 160

Leu Ala Phe Asn Thr His Thr Asp Asn Ile Glu Trp Ala Pro Ile Lys
                165                 170                 175

Ala Ala Gln Leu Thr Arg Pro Asn Ala Glu Leu Val Glu Leu Glu Ile
            180                 185                 190

Asn Thr Leu His Gly Val Lys Thr Ile Arg Cys Thr Pro Asp His Pro
        195                 200                 205

Val Tyr Thr Lys Asn Arg Asp Tyr Val Arg Ala Asp Glu Leu Thr Asp
    210                 215                 220

Asp Asp Glu Leu Val Val Ala Ile Gly Gly His His His His His His
225                 230                 235                 240
```

<210> 20
<211> 105
<212> PRT
<213> Secuencia Artificial

<220>
<223> NrdJ1 (InteínaN)

<400> 20

```
Cys Leu Val Gly Ser Ser Glu Ile Ile Thr Arg Asn Tyr Gly Lys Thr
1               5                   10                  15

Thr Ile Lys Glu Val Val Glu Ile Phe Asp Asn Asp Lys Asn Ile Gln
            20                  25                  30

Val Leu Ala Phe Asn Thr His Thr Asp Asn Ile Glu Trp Ala Pro Ile
        35                  40                  45

Lys Ala Ala Gln Leu Thr Arg Pro Asn Ala Glu Leu Val Glu Leu Glu
    50                  55                  60

Ile Asn Thr Leu His Gly Val Lys Thr Ile Arg Cys Thr Pro Asp His
65                  70                  75                  80

Pro Val Tyr Thr Lys Asn Arg Asp Tyr Val Arg Ala Asp Glu Leu Thr
                85                  90                  95

Asp Asp Asp Glu Leu Val Val Ala Ile
            100                 105
```

<210> 21
<211> 5
<212> PRT
<213> Secuencia Artificial

<220>
<223> NrdJ1 (ExteínaN)

<400> 21

```
Gly Thr Asn Pro Cys
1               5
```

<210> 22
<211> 514
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del extremo C de NrdJ1

<400> 22

```
catatggaag cgaagaccta catcggtaaa ctgaagagcc gcaagattgt tagcaacgag       60

gacacctacg atatccagac cagcacgcat aatttctttg cgaacgacat cctggtgcac      120

aacagcgaaa ttgtgctggg taccggatct gataaaatta ttcatctgac tgatgattct      180

tttgatactg atgtacttaa ggcagatggt gcaatcctgg ttgatttctg ggcacactgg      240

tgcggtccgt gcaaaatgat cgctccgatt ctggatgaaa tcgctgacga atatcagggc      300

aaactgaccg ttgcaaaact gaacatcgat cacaacccgg gcactgcgcc gaaatatggc      360

atccgtggta tcccgactct gctgctgttc aaaaacggtg aagtggcggc aaccaaagtg      420

ggtgcactgt ctaaaggtca gttgaaagag ttcctcgacg ctaacctggc cggctctgaa      480

ttcagatctc atcaccatca ccatcactaa gctt                                  514
```

<210> 23

<211> 168
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo C de NrdJ1

<400> 23

```
Met Glu Ala Lys Thr Tyr Ile Gly Lys Leu Lys Ser Arg Lys Ile Val
1               5                   10                  15

Ser Asn Glu Asp Thr Tyr Asp Ile Gln Thr Ser Thr His Asn Phe Phe
            20                  25                  30

Ala Asn Asp Ile Leu Val His Asn Ser Glu Ile Val Leu Gly Thr Gly
        35                  40                  45

Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val
    50                  55                  60

Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys
65                  70                  75                  80

Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu
                85                  90                  95

Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro
            100                 105                 110

Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu
        115                 120                 125

Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys
    130                 135                 140

Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe
145                 150                 155                 160

Arg Ser His His His His His His
                165
```

<210> 24
<211> 40
<212> PRT
<213> Secuencia Artificial

<220>
<223> NrdJ1 (Inteínac)

<400> 24

```
Met Glu Ala Lys Thr Tyr Ile Gly Lys Leu Lys Ser Arg Lys Ile Val
1               5                   10                  15

Ser Asn Glu Asp Thr Tyr Asp Ile Gln Thr Ser Thr His Asn Phe Phe
```

20 25 30

Ala Asn Asp Ile Leu Val His Asn
35 40

<210> 25
<211> 5
<212> PRT
<213> Secuencia Artificial

<220>
<223> NrdJ1 (Exteínac)

<400> 25

Ser Glu Ile Val Leu
1 5

<210> 26
<211> 681
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del extremo N de DNA-E

<400> 26

ccatggccag ttggagccac ccgcagttcg aaaaagcgag caaagaaacc tttacccatt 60

accagccgca gggcaacagt gacccggctc ataccgcaac cgcgcccggc ggattgagtg 120

cgaaagcgcc tgcaatgacc ccgctgatgc tggacacctc cagccgtaag ctggttgcgt 180

gggatggcac caccgacggt gctgccgttg gcattcttgc ggttgctgct gaccagacca 240

gcaccacgct gacgttctac aagtccggca cgttccgtta tgaggatgtg ctctggccgg 300

aggctgccag cgacgagacg aaaaaacgga ccgcgtttgc cggaacggca atcagcatcg 360

ttggatcctg tttaagctat gaaacggaaa tattgacagt agaatatgga ttattaccga 420

ttggtaaaat tgtagaaaag cgcatcgaat gtactgttta tagcgttgat aataatggaa 480

atatttatac acaacctgta gcacaatggc acgatcgcgg agaacaagag gtgtttgagt 540

attgtttgga agatggttca ttgattcggg caacaaaaga ccataagttt atgactgttg 600

atggtcaaat gttgccaatt gatgaaatat ttgaacgtga attggatttg atgcgggttg 660

ataatttgcc gaattaagct t 681

<210> 27
<211> 224
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo N de DNA-E

<400> 27

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1               5                  10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Cys Leu Ser Tyr Glu Thr
        115                 120                 125

Glu Ile Leu Thr Val Glu Tyr Gly Leu Leu Pro Ile Gly Lys Ile Val
    130                 135                 140

Glu Lys Arg Ile Glu Cys Thr Val Tyr Ser Val Asp Asn Asn Gly Asn
145                 150                 155                 160

Ile Tyr Thr Gln Pro Val Ala Gln Trp His Asp Arg Gly Glu Gln Glu
                165                 170                 175

Val Phe Glu Tyr Cys Leu Glu Asp Gly Ser Leu Ile Arg Ala Thr Lys
            180                 185                 190

Asp His Lys Phe Met Thr Val Asp Gly Gln Met Leu Pro Ile Asp Glu
        195                 200                 205

Ile Phe Glu Arg Glu Leu Asp Leu Met Arg Val Asp Asn Leu Pro Asn
    210                 215                 220
```

<210> 28
<211> 102
<212> PRT
<213> Secuencia Artificial

<220>
<223> DNA-E (Inteínan)

<400> 28

```
Cys Leu Ser Tyr Glu Thr Glu Ile Leu Thr Val Glu Tyr Gly Leu Leu
1               5                   10                  15

Pro Ile Gly Lys Ile Val Glu Lys Arg Ile Glu Cys Thr Val Tyr Ser
            20                  25                  30

Val Asp Asn Asn Gly Asn Ile Tyr Thr Gln Pro Val Ala Gln Trp His
        35                  40                  45

Asp Arg Gly Glu Gln Glu Val Phe Glu Tyr Cys Leu Glu Asp Gly Ser
    50                  55                  60

Leu Ile Arg Ala Thr Lys Asp His Lys Phe Met Thr Val Asp Gly Gln
65                  70                  75                  80

Met Leu Pro Ile Asp Glu Ile Phe Glu Arg Glu Leu Asp Leu Met Arg
                85                  90                  95

Val Asp Asn Leu Pro Asn
            100
```

<210> 29
<211> 496
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del extremo C de DNA-E

<400> 29

```
catatgatca aaatagccac acgtaaatat ttaggcaaac aaaatgtcta tgacattgga       60

gttgagcgcg accataattt tgcactcaaa aatggcttca tagcttctaa ttgtttcaat      120

ggtaccggat ctgataaaat tattcatctg actgatgatt cttttgatac tgatgtactt      180

aaggcagatg gtgcaatcct ggttgatttc tgggcacact ggtgcggtcc gtgcaaaatg      240

atcgctccga ttctggatga aatcgctgac gaatatcagg gcaaactgac cgttgcaaaa      300

ctgaacatcg atcacaaccc gggcactgcg ccgaaatatg gcatccgtgg tatcccgact      360

ctgctgctgt tcaaaaacgg tgaagtggcg gcaaccaaag tgggtgcact gtctaaaggt      420

cagttgaaag agttcctcga cgctaacctg gccggctctg aattcagatc tcatcaccat      480

caccatcact aagctt                                                      496
```

<210> 30
<211> 162
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo C de DNA-E

<400> 30

```
Met Ile Lys Ile Ala Thr Arg Lys Tyr Leu Gly Lys Gln Asn Val Tyr
1               5                   10                  15

Asp Ile Gly Val Glu Arg Asp His Asn Phe Ala Leu Lys Asn Gly Phe
            20                  25                  30

Ile Ala Ser Asn Cys Phe Asn Gly Thr Gly Ser Asp Lys Ile Ile His
        35                  40                  45

Leu Thr Asp Asp Ser Phe Asp Thr Asp Val Leu Lys Ala Asp Gly Ala
    50                  55                  60

Ile Leu Val Asp Phe Trp Ala His Trp Cys Gly Pro Cys Lys Met Ile
65                  70                  75                  80

Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu Tyr Gln Gly Lys Leu Thr
                85                  90                  95

Val Ala Lys Leu Asn Ile Asp His Asn Pro Gly Thr Ala Pro Lys Tyr
            100                 105                 110

Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu Phe Lys Asn Gly Glu Val
        115                 120                 125

Ala Ala Thr Lys Val Gly Ala Leu Ser Lys Gly Gln Leu Lys Glu Phe
    130                 135                 140

Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe Arg Ser His His His His
145                 150                 155                 160

His His
```

<210> 31
<211> 36
<212> PRT
<213> Secuencia Artificial

<220>
<223> DNA-E (InteínaC)

<400> 31

```
Met Ile Lys Ile Ala Thr Arg Lys Tyr Leu Gly Lys Gln Asn Val Tyr
1               5                   10                  15

Asp Ile Gly Val Glu Arg Asp His Asn Phe Ala Leu Lys Asn Gly Phe
            20                  25                  30

Ile Ala Ser Asn
        35
```

<210> 32
<211> 717
<212> ADN
<213> Secuencia Artificial

<220>

<223> ADN del extremo N de IMPDH

<400> 32

ccatggccag ttggagccac ccgcagttcg aaaaagcgag caaagaaacc tttacccatt 60

accagccgca gggcaacagt gacccggctc ataccgcaac cgcgcccggc ggattgagtg 120

cgaaagcgcc tgcaatgacc ccgctgatgc tggacacctc cagccgtaag ctggttgcgt 180

gggatggcac caccgacggt gctgccgttg gcattcttgc ggttgctgct gaccagacca 240

gcaccacgct gacgttctac aagtccggca cgttccgtta tgaggatgtg ctctggccgg 300

aggctgccag cgacgagacg aaaaaacgga ccgcgtttgc cggaacggca atcagcatcg 360

ttggatccgg cattggcggt ggctgctttg tgccgggcac cctggtgaac acggaaaacg 420

gcctgaagaa aatcgaggaa attaaggtgg gcgacaaggt gttcagccat accggcaaac 480

tgcaggaagt tgtggacacg ctgatctttg accgcgacga agaaatcatc agcattaacg 540

gcatcgactg cacgaaaaac cacgagttct acgtgatcga caaggagaac gcgaaccgtg 600

tgaacgaaga caatatccat ctgttcgcgc gttgggttca cgcggaggag ctggacatga 660

aaaaacatct gctgattgag ctggaaggcg gccatcatca ccaccaccac taagctt 717

<210> 33
<211> 236
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo N de IMPDH

<400> 33

Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1 5 10 15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
20 25 30

```
Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Gly Ile Gly Gly Gly Cys
        115                 120                 125

Phe Val Pro Gly Thr Leu Val Asn Thr Glu Asn Gly Leu Lys Lys Ile
    130                 135                 140

Glu Glu Ile Lys Val Gly Asp Lys Val Phe Ser His Thr Gly Lys Leu
145                 150                 155                 160

Gln Glu Val Val Asp Thr Leu Ile Phe Asp Arg Asp Glu Glu Ile Ile
                165                 170                 175

Ser Ile Asn Gly Ile Asp Cys Thr Lys Asn His Glu Phe Tyr Val Ile
            180                 185                 190

Asp Lys Glu Asn Ala Asn Arg Val Asn Glu Asp Asn Ile His Leu Phe
        195                 200                 205

Ala Arg Trp Val His Ala Glu Glu Leu Asp Met Lys Lys His Leu Leu
    210                 215                 220

Ile Glu Leu Glu Gly Gly His His His His His His
225                 230                 235
```

<210> 34
<211> 101
<212> PRT
<213> Secuencia Artificial

<220>
<223> IMPDH (Inteínan)

<400> 34

```
Cys Phe Val Pro Gly Thr Leu Val Asn Thr Glu Asn Gly Leu Lys Lys
```

1 5 10 15

Ile Glu Glu Ile Lys Val Gly Asp Lys Val Phe Ser His Thr Gly Lys
20 25 30

Leu Gln Glu Val Val Asp Thr Leu Ile Phe Asp Arg Asp Glu Glu Ile
35 40 45

Ile Ser Ile Asn Gly Ile Asp Cys Thr Lys Asn His Glu Phe Tyr Val
50 55 60

Ile Asp Lys Glu Asn Ala Asn Arg Val Asn Glu Asp Asn Ile His Leu
65 70 75 80

Phe Ala Arg Trp Val His Ala Glu Glu Leu Asp Met Lys Lys His Leu
85 90 95

Leu Ile Glu Leu Glu
100

<210> 35
<211> 5
<212> PRT
<213> Secuencia Artificial

<220>
<223> IMPDH (Exteínan)

<400> 35

Gly Ile Gly Gly Gly
1 5

<210> 36
<211> 514
<212> ADN
<213> Secuencia Artificial

<220>
<223> ADN del extremo C de IMPDH

<400> 36

catatgaagt tcaagctgaa ggagatcacg agcatcgaga ccaagcacta caagggcaag 60

gtgcacgatc tgaccgtgaa tcaggaccac agctataacg tgcgcggcac cgtggtgcat 120

aatagcattt gcagcaccgg taccggatct gataaaatta ttcatctgac tgatgattct 180

tttgatactg atgtacttaa ggcagatggt gcaatcctgg ttgatttctg ggcacactgg 240

tgcggtccgt gcaaaatgat cgctccgatt ctggatgaaa tcgctgacga atatcagggc 300

aaactgaccg ttgcaaaact gaacatcgat cacaacccgg gcactgcgcc gaaatatggc 360

atccgtggta tcccgactct gctgctgttc aaaaacggtg aagtggcggc aaccaaagtg 420

ggtgcactgt ctaaaggtca gttgaaagag ttcctcgacg ctaacctggc cggctctgaa 480

ttcagatctc atcaccatca ccatcactaa gctt 514

<210> 37
<211> 168

<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo C de IMPDH

<400> 37

```
Met Lys Phe Lys Leu Lys Glu Ile Thr Ser Ile Glu Thr Lys His Tyr
1               5                  10                  15

Lys Gly Lys Val His Asp Leu Thr Val Asn Gln Asp His Ser Tyr Asn
            20                  25                  30

Val Arg Gly Thr Val Val His Asn Ser Ile Cys Ser Thr Gly Thr Gly
        35                  40                  45

Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val
    50                  55                  60

Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys
65                  70                  75                  80

Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu
                85                  90                  95

Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro
            100                 105                 110

Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu
        115                 120                 125

Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys
    130                 135                 140

Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe
145                 150                 155                 160

Arg Ser His His His His His His
                165
```

<210> 38
<211> 40
<212> PRT
<213> Secuencia Artificial

<220>
<223> IMPDH (InteínaC)

<400> 38

Met Lys Phe Lys Leu Lys Glu Ile Thr Ser Ile Glu Thr Lys His Tyr
1 5 10 15

Lys Gly Lys Val His Asp Leu Thr Val Asn Gln Asp His Ser Tyr Asn
20 25 30

Val Arg Gly Thr Val Val His Asn
35 40

<210> 39
<211> 5
<212> PRT
<213> Secuencia Artificial

<220>
<223> IMPDH (ExteínaC)

<400> 39

Ser Ile Cys Ser Thr
1 5

<210> 40
<211> 8
<212> PRT
<213> Secuencia Artificial

<220>
<223> Etiqueta Strep

<400> 40

Trp Ser His Pro Gln Phe Glu Lys
1 5

<210> 41
<211> 107
<212> PRT
<213> Secuencia Artificial

<220>
<223> gpD

<400> 41

Lys Glu Thr Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala
1 5 10 15

His Thr Ala Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met
20 25 30

Thr Pro Leu Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp
35 40 45

Gly Thr Thr Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp
50 55 60

Gln Thr Ser Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr
65 70 75 80

Glu Asp Val Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg
85 90 95

Thr Ala Phe Ala Gly Thr Ala Ile Ser Ile Val
100 105

<210> 42
<211> 6
<212> PRT
<213> Secuencia Artificial

<220>
<223> H6

<400> 42

His His His His His His
1 5

<210> 43
<211> 111
<212> PRT
<213> Secuencia Artificial

<220>
<223> Trx

<400> 43

Gly Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp
1 5 10 15

Val Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp
20 25 30

Cys Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp
35 40 45

Glu Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn
50 55 60

Pro Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu
65 70 75 80

Leu Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser
85 90 95

Lys Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser
100 105 110

<210> 44
<211> 13
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia ChsXcplhXTXXG comprendida en la caja N1

<220>
<221> VARIANTE
<222> (2).. (2)
<223> /nota = "Xaa es un aminoácido hidrofóbico"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /nota = "Xaa es un aminoácido pequeño"

<220>
<221> VARIANTE
<222> (4)..(4)
<223> /nota = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /nota = "Xaa es un aminoácido cargado"

<220>
<221> VARIANTE
<222> (6).. (6)
<223> /nota = "Xaa es un aminoácido polar"

<220>
<221> VARIANTE
<222> (7) .. (7)
<223> /nota = "Xaa es un aminoácido grande"

<220>
<221> VARIANTE
<222> (8) .. (8)
<223> /nota = "Xaa es un aminoácido hidrofóbico"

<220>
<221> VARIANTE
<222> (9) .. (9)
<223> /nota = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (11)..(12)
<223> /nota = "Xaa es cualquier aminoácido"

<400> 44

Cys Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Thr Xaa Xaa Gly
1 5 10

<210> 45
<211> 13
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia comprendida en el dominio N-terminal de inteína

<220>
<221> VARIANTE
<222> (1) .. (1)
<223> /reemplazar = "Cys"

<220>
<221> VARIANTE
<222> (2).. (2)
<223> /reemplazar = "Leu"
/reemplazar = "Phe"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (3) .. (3)
<223> /reemplazar = "Ser" /reemplazar = "Thr"
/reemplazar = "Val"
/reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (4) .. (4)
<223> /reemplazar = "Leu" /reemplazar = "Pro"
/reemplazar = "Gly"
/reemplazar = "Tyr"

<220>
<221> VARIANTE
<222> (5) .. (5)
<223> /reemplazar = "Asp" /reemplazar = "Glu"
/reemplazar = "Lys"
/reemplazar = "Gly"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Thr"
/reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (7) .. (7)
<223> /reemplazar = "Glu"
/reemplazar = "Gln"
/reemplazar = "Leu"

<220>
<221> VARIANTE
<222> (7) .. (7)
<223> /reemplazar = "Met"
/reemplazar = "Lys"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Ile"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (9) .. (9)
<223> /reemplazar = "Leu" /reemplazar = "Gln"
/reemplazar = "Val"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (9) .. (9)
<223> /reemplazar = "Lys"
/reemplazar = "Asp"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (10)..(10)
<223> /reemplazar = "Thr"
/reemplazar = "Ile"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (11) .. (11)
<223> /reemplazar = "Val" /reemplazar = "Pro"
/reemplazar = "Gln"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "Glu"
/reemplazar = "Lys"
/reemplazar = "Leu"

<220>
<221> VARIANTE
<222> (12)..(12)
<223> /reemplazar = "Glu" /reemplazar = "Gln"
/reemplazar = "Gly"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (12) .. (12)
<223> /reemplazar = "Tyr"
/reemplazar = "Ile"
/reemplazar = "Glu"

<220>
<221> VARIANTE
<222> (13)..(13)
<223> /reemplazar = "Tyr" /reemplazar = "Gly"
/reemplazar = "Lys"
/reemplazar = "Pro"

<220>
<221> VARIANTE
<222> (13)..(13)
<223> /reemplazar = "Asp"

<400> 45

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5 10

<210> 46
<211> 13
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia comprendida en el dominio N-terminal de inteína

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Cys"

<220>
<221> VARIANTE
<222> (2)..(2)
<223> /reemplazar = "Leu"
/reemplazar = "Phe"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /reemplazar = "Ser" /reemplazar = "Thr"
/reemplazar = "Val"
/reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (4)..(4)
<223> /reemplazar = "Leu"
/reemplazar = "Pro"
/reemplazar = "Gly"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /reemplazar = "Asp"
/reemplazar = "Lys"
/reemplazar = "Gly"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Thr"
/reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (7)..(7)
<223> /reemplazar = "Gln" /reemplazar = "Leu,"
/reemplazar = "Met"
/reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (7)..(7)
<223> /reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Ile"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Gln" /reemplazar = "Val"
/reemplazar = "Asn"
/reemplazar = "Lys"

<220>

<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Asp"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (10)..(10)
<223> /reemplazar = "Thr"
/reemplazar = "Ile"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "Pro" /reemplazar = "Gln"
/reemplazar = "Asn"
/reemplazar = "Glu"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "Lys"
/reemplazar = "Leu"

<220>
<221> VARIANTE
<222> (12)..(12)
<223> /reemplazar = "Glu" /reemplazar = "Gln"
/reemplazar = "Gly"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (12)..(12)
<223> /reemplazar = "Tyr"
/reemplazar = "Ile"
/reemplazar = "Glu"

<220>
<221> VARIANTE
<222> (13)..(13)
<223> /reemplazar = "Gly" /reemplazar = "Lys"
/reemplazar = "Pro"
/reemplazar = "Asp"

<400> 46

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5 10

<210> 47
<211> 15
<212> PRT
<213> Secuencia Artificial

<220>
<223> Sequence GXXhXhTXaHXhhTX comprendida en la caja N3

<220>
<221> VARIANTE
<222> (2)..(3)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE

<222> (4) .. (4)
<223> /note = "Xaa es un aminoácido hidrofóbico"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /note = "Xaa es un aminoácido hidrofóbico"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /note = "Xaa es un aminoácido ácido"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (12)..(13)
<223> /note = "Xaa es un aminoácido hidrofóbico"

<220>
<221> VARIANTE
<222> (15)..(15)
<223> /note = "Xaa es cualquier aminoácido"

<400> 47

Gly Xaa Xaa Xaa Xaa Xaa Thr Xaa Xaa His Xaa Xaa Xaa Thr Xaa
1 5 10 15

<210> 48
<211> 15
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia comprendida en el dominio N-terminal de inteína

<220>
<221> VARIANTE
<222> (1) .. (1)
<223> /reemplazar = "Gly"
/reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (2) .. (2)
<223> /reemplazar = "Ser" /reemplazar = "Lys"
/reemplazar = "Gln"
/reemplazar = "Asn"

<220>
<221> VARIANTE

<222> (2)..(2)
<223> /reemplazar = "Phe"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /reemplazar = "Leu" /reemplazar = "Glu"
/reemplazar = "Lys"
/reemplazar = "Arg"

<220>
<221> VARIANTE
<222> (4)..(4)
<223> /reemplazar = "Ile"
/reemplazar = "Leu"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /reemplazar = "Arg" /reemplazar = "Ile"
/reemplazar = "Val"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Ala" /reemplazar = "Cys"
/reemplazar = "Val"
/reemplazar = "Glu"

<220>
<221> VARIANTE
<222> (7)..(7)
<223> /reemplazar = "Thr"
/reemplazar = "Ser"
/reemplazar = "Asp"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Lys" /reemplazar = "Glu"
/reemplazar = "Ala"
/reemplazar = "Pro"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Asp" /reemplazar = "Glu"
/reemplazar = "Asn"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (10)..(10)
<223> /reemplazar = "His"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "Lys" /reemplazar = "Leu"

/reemplazar = "Gln"
/reemplazar = "Met"

<220>
<221> VARIANTE
<222> (12)..(12)
<223> /reemplazar = "Phe"
/reemplazar = "Val"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (13)..(13)
<223> /reemplazar = "Met" /reemplazar = "Pro"
/reemplazar = "Phe"
/reemplazar = "Tyr"

<220>
<221> VARIANTE
<222> (13)..(13)
<223> /reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (14)..(14)
<223> /reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (15)..(15)
<223> /reemplazar = "Val" /reemplazar = "Gln"
/reemplazar = "Lys"
/reemplazar = "Leu"

<400> 48

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5 10 15

<210> 49
<211> 15
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia comprendida en el dominio N-terminal de inteína

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Gly"
/reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (2)..(2)
<223> /reemplazar = "Lys" /reemplazar = "Gln"
/reemplazar = "Asn"
/reemplazar = "Phe"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /reemplazar = "Glu"
/reemplazar = "Lys"
/reemplazar = "Arg"

<220>
<221> VARIANTE
<222> (4)..(4)
<223> /reemplazar = "Ile"
/reemplazar = "Leu"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /reemplazar = "Arg" /reemplazar = "Ile"
/reemplazar = "Val"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (6).. (6)
<223> /reemplazar = "Cys"
/reemplazar = "Val"
/reemplazar = "Glu"

<220>
<221> VARIANTE
<222> (7)..(7)
<223> /reemplazar = "Thr"
/reemplazar = "Ser"
/reemplazar = "Asp"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Glu" /reemplazar = "Ala"
/reemplazar = "Pro"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Asp" /reemplazar = "Glu"
/reemplazar = "Asn"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (10)..(10)
<223> /reemplazar = "His"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "Lys" /reemplazar = "Leu"
/reemplazar = "Gln"
/reemplazar = "Met"

<220>
<221> VARIANTE
<222> (12)..(12)
<223> /reemplazar = "Phe"
/reemplazar = "Val"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (13)..(13)
<223> /reemplazar = "Pro" /reemplazar = "Phe"

/reemplazar = "Tyr"
/reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (14)..(14)
<223> /reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (15)..(15)
<223> /reemplazar = "Gln"
/reemplazar = "Lys"
/reemplazar = "Leu"

<400> 49

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5 10 15

<210> 50
<211> 14
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia XhhDIpVXXpHXFX comprendida en la caja C1

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (2)..(3)
<223> /note = "Xaa es un aminoácido hidrofóbico"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /note = "Xaa es un aminoácido polar"

<220>
<221> VARIANTE
<222> (8)..(9)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (10)..(10)
<223> /note = "Xaa es un aminoácido polar"

<220>
<221> VARIANTE
<222> (12)..(12)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (14)..(14)
<223> /note = "Xaa es cualquier aminoácido"

<400> 50

Xaa Xaa Xaa Asp Ile Xaa Val Xaa Xaa Xaa His Xaa Phe Xaa
1 5 10

<210> 51
<211> 14
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia comprendida en el dominio C-terminal de inteína

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Asn" /reemplazar = "Glu"
/reemplazar = "Leu"
/reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Gln" /reemplazar = "Asp"
/reemplazar = "Pro"
/reemplazar = "Arg"

<220>
<221> VARIANTE
<222> (2)..(2)
<223> /reemplazar = "Val"
/reemplazar = "Leu"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /reemplazar = "Tyr" /reemplazar = "Ile"
/reemplazar = "Val"
/reemplazar = "His"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /reemplazar = "Phe"

<220>
<221> VARIANTE
<222> (4)..(4)
<223> /reemplazar = "Asp"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /reemplazar = "Ile"
/reemplazar = "Leu"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Gly" /reemplazar = "Glu"
/reemplazar = "Thr"
/reemplazar = "Gln"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (7)..(7)
<223> /reemplazar = "Val"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Glu" /reemplazar = "Ser"
/reemplazar = "Thr"
/reemplazar = "Asp"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Asn"
/reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Arg" /reemplazar = "Gly"
/reemplazar = "Asp"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Gln"
/reemplazar = "Ser"
/reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (10).. (10)
<223> /reemplazar = "Asp" /reemplazar = "Glu"
/reemplazar = "Asn"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (10)..(10)
<223> /reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "His" /reemplazar = "Arg"
/reemplazar = "Ser"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (12).. (12)
<223> /reemplazar = "Asn" /reemplazar = "Leu"
/reemplazar = "Ser"
/reemplazar = "Ile"

<220>

<221> VARIANTE
<222> (12)..(12)
<223> /reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (13)..(13)
<223> /reemplazar = "Phe" /reemplazar = "Tyr"
/reemplazar = "Leu"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (14)..(14)
<223> /reemplazar = "Ala" /reemplazar = "Tyr"
/reemplazar = "Phe"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (14)..(14)
<223> /reemplazar = "Cys"
/reemplazar = "Ser"

<400> 51

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5 10

<210> 52
<211> 14
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia comprendida en el dominio C-terminal de inteína

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Glu" /reemplazar = "Leu"
/reemplazar = "Lys"
/reemplazar = "Gln"

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Asp"
/reemplazar = "Pro"
/reemplazar = "Arg"

<220>
<221> VARIANTE
<222> (2)..(2)
<223> /reemplazar = "Val"
/reemplazar = "Leu"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /reemplazar = "Tyr" /reemplazar = "Ile"
/reemplazar = "Val"
/reemplazar = "His"

<220>

<221> VARIANTE
<222> (3)..(3)
<223> /reemplazar = "Phe"

<220>
<221> VARIANTE
<222> (4)..(4)
<223> /reemplazar = "Asp"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /reemplazar = "Ile"
/reemplazar = "Leu"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Gly" /reemplazar = "Glu"
/reemplazar = "Thr"
/reemplazar = "Gln"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (7)..(7)
<223> /reemplazar = "Val"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Glu" /reemplazar = "Ser"
/reemplazar = "Thr"
/reemplazar = "Asp"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Asn"
/reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Gly" /reemplazar = "Asp"
/reemplazar = "Asn"
/reemplazar = "Gln"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Ser"
/reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (10)..(10)
<223> /reemplazar = "Asp" /reemplazar = "Glu"
/reemplazar = "Asn"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (10)..(10)
<223> /reemplazar = "Lys"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "His" /reemplazar = "Arg"
/reemplazar = "Ser"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (11)..(11)
<223> /reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (12)..(12)
<223> /reemplazar = "Asn" /reemplazar = "Leu"
/reemplazar = "Ser"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (12)..(12)
<223> /reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (13)..(13)
<223> /reemplazar = "Phe" /reemplazar = "Tyr"
/reemplazar = "Leu"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (14)..(14)
<223> /reemplazar = "Ala" /reemplazar = "Tyr"
/reemplazar = "Phe"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (14)..(14)
<223> /reemplazar = "Cys"
/reemplazar = "Ser"

<400> 52

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5 10

<210> 53
<211> 9
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia hNXIhXHNn comprendida en la caja C2

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /note = "Xaa es un aminoácido hidrofóbico"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /note = "Xaa es un aminoácido hidrofóbico"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /note = "Xaa es cualquier aminoácido"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /note = "Xaa es un aminoácido nucleofílico"

<400> 53

Xaa Asn Xaa Ile Xaa Xaa His Asn Xaa
1 5

<210> 54
<211> 9
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia comprendida en el dominio C-terminal de inteína

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Leu" /reemplazar = "Ala"
/reemplazar = "Val"
/reemplazar = "Ile"

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Cys"

<220>
<221> VARIANTE
<222> (2).. (2)
<223> /reemplazar = "Asn"
/reemplazar = "Arg"

<220>
<221> VARIANTE
<222> (3)..(3)
<223> /reemplazar = "Gly" /reemplazar = "Asp"
/reemplazar = "Ala"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (4).. (4)
<223> /reemplazar = "Ile"
/reemplazar = "Phe"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /reemplazar = "Leu"
/reemplazar = "Ile"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Val" /reemplazar = "Ile"
/reemplazar = "Thr"
/reemplazar = "Ala"

<220>
<221> VARIANTE
<222> (7)..(7)
<223> /reemplazar = "His"
/reemplazar = "Ser"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Ser"
/reemplazar = "Thr"
/reemplazar = "Cys"

<400> 54

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5

<210> 55
<211> 9
<212> PRT
<213> Secuencia Artificial

<220>
<223> Secuencia comprendida en el dominio C-terminal de inteína

<220>
<221> VARIANTE
<222> (1)..(1)
<223> /reemplazar = "Ala" /reemplazar = "Val"
/reemplazar = "Ile"
/reemplazar = "Cys"

<220>
<221> CARACTERÍSTICA_MISC
<222> (1)..(8)
<223> /nota = "secuencia de inteína"

<220>
<221> VARIANTE
<222> (2)..(2)
<223> /reemplazar = "Asn"
/reemplazar = "Arg"

<220>
<221> VARIANTE
<222> (3)..(3)

<223> /reemplazar = "Gly" /reemplazar = "Asp"
/reemplazar = "Ala"
/reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (4)..(4)
<223> /reemplazar = "Ile"
/reemplazar = "Phe"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (5)..(5)
<223> /reemplazar = "Leu"
/reemplazar = "Val"

<220>
<221> VARIANTE
<222> (6)..(6)
<223> /reemplazar = "Val"
/reemplazar = "Ile"
/reemplazar = "Thr"

<220>
<221> VARIANTE
<222> (7)..(7)
<223> /reemplazar = "His"

<220>
<221> VARIANTE
<222> (8)..(8)
<223> /reemplazar = "Asn"

<220>
<221> VARIANTE
<222> (9)..(9)
<223> /reemplazar = "Ser"
/reemplazar = "Thr"
/reemplazar = "Cys"

<220>
<221> CARACTERÍSTICA_MISC
<222> (9)..(9)
<223> /note = "primer aminoácido de la exteína"

<400> 55

Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa Xaa
1 5

<210> 56
<211> 223
<212> PRT
<213> Secuencia Artificial

<220>
<223> Proteína del fragmento N de GP-41.1 C1A

<400> 56

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1               5                  10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Thr Arg Ser Gly Tyr Ala
        115                 120                 125

Leu Asp Leu Lys Thr Gln Val Gln Thr Pro Gln Gly Met Lys Glu Ile
    130                 135                 140

Ser Asn Ile Gln Val Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr Asn
145                 150                 155                 160

Glu Val Leu Asn Val Phe Pro Lys Ser Lys Lys Lys Ser Tyr Lys Ile
                165                 170                 175

Thr Leu Glu Asp Gly Lys Glu Ile Ile Cys Ser Glu Glu His Leu Phe
            180                 185                 190

Pro Thr Gln Thr Gly Glu Met Asn Ile Ser Gly Gly Leu Lys Glu Gly
        195                 200                 205
Met Cys Leu Tyr Val Lys Glu Gly Gly His His His His His His
    210                 215                 220
```

<210> 57
<211> 224
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del fragmento N de GP 41.8 C1A

<400> 57

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1               5                  10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Ser Gln Leu Asn Arg Ala
        115                 120                 125

Leu Ser Leu Asp Thr Met Val Val Thr Asn Gly Lys Ala Ile Glu Ile
    130                 135                 140

Arg Asp Val Lys Val Gly Asp Trp Leu Glu Ser Glu Cys Gly Pro Val
145                 150                 155                 160

Gln Val Thr Glu Val Leu Pro Ile Ile Lys Gln Pro Val Phe Glu Ile
                165                 170                 175

Val Leu Lys Ser Gly Lys Lys Ile Arg Val Ser Ala Asn His Lys Phe
            180                 185                 190

Pro Thr Lys Asp Gly Leu Lys Thr Ile Asn Ser Gly Leu Lys Val Gly
        195                 200                 205

Asp Phe Leu Arg Ser Arg Ala Lys Gly Gly His His His His His His
    210                 215                 220
```

<210> 58
<211> 240
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo N de NrdJ1 C1A

<400> 58

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1               5                  10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Gly Thr Asn Pro Cys Ala
        115                 120                 125

Leu Val Gly Ser Ser Glu Ile Ile Thr Arg Asn Tyr Gly Lys Thr Thr
    130                 135                 140

Ile Lys Glu Val Val Glu Ile Phe Asp Asn Asp Lys Asn Ile Gln Val
145                 150                 155                 160

Leu Ala Phe Asn Thr His Thr Asp Asn Ile Glu Trp Ala Pro Ile Lys
                165                 170                 175

Ala Ala Gln Leu Thr Arg Pro Asn Ala Glu Leu Val Glu Leu Glu Ile
            180                 185                 190
Asn Thr Leu His Gly Val Lys Thr Ile Arg Cys Thr Pro Asp His Pro
        195                 200                 205

Val Tyr Thr Lys Asn Arg Asp Tyr Val Arg Ala Asp Glu Leu Thr Asp
    210                 215                 220

Asp Asp Glu Leu Val Val Ala Ile Gly Gly His His His His His His
225                 230                 235                 240
```

<210> 59
<211> 236
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo N de IMPDH C1A

<400> 59

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1               5                  10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
        35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
    50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
65                  70                  75                  80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
                85                  90                  95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
            100                 105                 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Gly Ile Gly Gly Gly Ala
        115                 120                 125

Phe Val Pro Gly Thr Leu Val Asn Thr Glu Asn Gly Leu Lys Lys Ile
    130                 135                 140

Glu Glu Ile Lys Val Gly Asp Lys Val Phe Ser His Thr Gly Lys Leu
145                 150                 155                 160
Gln Glu Val Val Asp Thr Leu Ile Phe Asp Arg Asp Glu Glu Ile Ile
                165                 170                 175

Ser Ile Asn Gly Ile Asp Cys Thr Lys Asn His Glu Phe Tyr Val Ile
            180                 185                 190

Asp Lys Glu Asn Ala Asn Arg Val Asn Glu Asp Asn Ile His Leu Phe
        195                 200                 205

Ala Arg Trp Val His Ala Glu Glu Leu Asp Met Lys Lys His Leu Leu
    210                 215                 220

Ile Glu Leu Glu Gly Gly His His His His His His
225                 230                 235
```

<210> 60
<211> 165
<212> PRT
<213> Secuencia Artificial

<220>
<223> Proteína del fragmento C de GP41.1 deltaext

<400> 60

```
Met Gly Lys Asn Ser Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu
1               5                  10                  15

Leu Asp Glu Arg Glu Leu Ile Asp Ile Glu Val Ser Gly Asn His Leu
            20                  25                  30

Phe Tyr Ala Asn Asp Ile Leu Thr His Asn Gly Thr Gly Ser Asp Lys
        35                  40                  45

Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val Leu Lys Ala
    50                  55                  60

Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys Gly Pro Cys
65                  70                  75                  80

Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu Tyr Gln Gly
                85                  90                  95

Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro Gly Thr Ala
            100                 105                 110

Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu Phe Lys Asn
        115                 120                 125
Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys Gly Gln Leu
    130                 135                 140

Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe Arg Ser His
145                 150                 155                 160

His His His His His
                165
```

<210> 61
<211> 168
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo C de GP41.8 deltaext

<400> 61

```
Met Cys Glu Ile Phe Glu Asn Glu Ile Asp Trp Asp Glu Ile Ala Ser
1               5                  10                  15

Ile Glu Tyr Val Gly Val Glu Glu Thr Ile Asp Ile Asn Val Thr Asn
            20                  25                  30

Asp Arg Leu Phe Phe Ala Asn Gly Ile Leu Thr His Asn Gly Thr Gly
        35                  40                  45

Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val
    50                  55                  60

Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys
65                  70                  75                  80

Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu
                85                  90                  95

Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro
            100                 105                 110

Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu
        115                 120                 125

Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys
    130                 135                 140

Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe
145                 150                 155                 160

Arg Ser His His His His His His
                165
```

<210> 62
<211> 163
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo C de NrdJ1 deltaext

<400> 62

```
Met Glu Ala Lys Thr Tyr Ile Gly Lys Leu Lys Ser Arg Lys Ile Val
1               5                  10                  15

Ser Asn Glu Asp Thr Tyr Asp Ile Gln Thr Ser Thr His Asn Phe Phe
            20                  25                  30

Ala Asn Asp Ile Leu Val His Asn Gly Thr Gly Ser Asp Lys Ile Ile
        35                  40                  45

His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val Leu Lys Ala Asp Gly
    50                  55                  60

Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys Gly Pro Cys Lys Met
65                  70                  75                  80

Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu Tyr Gln Gly Lys Leu
                85                  90                  95

Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro Gly Thr Ala Pro Lys
            100                 105                 110

Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu Phe Lys Asn Gly Glu
        115                 120                 125

Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys Gly Gln Leu Lys Glu
    130                 135                 140

Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe Arg Ser His His His
145                 150                 155                 160

His His His
```

<210> 63
<211> 163
<212> PRT
<213> Secuencia Artificial

<220>
<223> PROTEÍNA del extremo C de IMPDH deltaext

<400> 63

```
Met Lys Phe Lys Leu Lys Glu Ile Thr Ser Ile Glu Thr Lys His Tyr
1               5                  10                  15

Lys Gly Lys Val His Asp Leu Thr Val Asn Gln Asp His Ser Tyr Asn
            20                  25                  30

Val Arg Gly Thr Val Val His Asn Gly Thr Gly Ser Asp Lys Ile Ile
        35                  40                  45

His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val Leu Lys Ala Asp Gly
    50                  55                  60

Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys Gly Pro Cys Lys Met
65                  70                  75                  80

Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu Tyr Gln Gly Lys Leu
                85                  90                  95

Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro Gly Thr Ala Pro Lys
            100                 105                 110

Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu Phe Lys Asn Gly Glu
        115                 120                 125

Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys Gly Gln Leu Lys Glu
    130                 135                 140

Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe Arg Ser His His His
145                 150                 155                 160

His His His
```

<210> 64
<211> 106
<212> PRT
<213> Secuencia Artificial

<220>
<223> Región N-terminal de la inteína NrdA2

<400> 64

```
Cys Leu Thr Gly Asp Ala Lys Ile Asp Val Leu Ile Asp Asn Ile Pro
1               5                  10                  15

Ile Ser Gln Ile Ser Leu Glu Glu Val Val Asn Leu Phe Asn Glu Gly
```

20 25 30

Lys Glu Ile Tyr Val Leu Ser Tyr Asn Ile Asp Thr Lys Glu Val Glu
35 40 45

Tyr Lys Glu Ile Ser Asp Ala Gly Leu Ile Ser Glu Ser Ala Glu Val
50 55 60

Leu Glu Ile Ile Asp Glu Glu Thr Gly Gln Lys Ile Val Cys Thr Pro
65 70 75 80

Asp His Lys Val Tyr Thr Leu Asn Arg Gly Tyr Val Ser Ala Lys Asp
85 90 95

Leu Lys Glu Asp Asp Glu Leu Val Phe Ser
100 105

<210> 65
<211> 34
<212> PRT
<213> Secuencia Artificial

<220>
<223> Región C-terminal de la inteína NrdA2

<400> 65

Met Gly Leu Lys Ile Ile Lys Arg Glu Ser Lys Glu Pro Val Phe Asp
1 5 10 15

Ile Thr Val Lys Asp Asn Ser Asn Phe Phe Ala Asn Asn Ile Leu Val
20 25 30

His Asn

<210> 66
<211> 166
<212> PRT
<213> Secuencia Artificial

<220>
<223> G1C(S)

<400> 66

Met Gly Lys Asn Ser Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu
1 5 10 15

Leu Asp Glu Arg Glu Leu Ile Asp Ile Glu Val Ser Gly Asn His Leu
20 25 30

```
Phe Tyr Ala Asn Asp Ile Leu Thr His Asn Ser Gly Thr Gly Ser Asp
         35                  40                  45

Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val Leu Lys
     50                  55                  60

Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys Gly Pro
65                  70                  75                  80

Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu Tyr Gln
                85                  90                  95

Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro Gly Thr
            100                 105                 110

Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu Phe Lys
         115                 120                 125

Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys Gly Gln
     130                 135                 140

Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe Arg Ser
145                 150                 155                 160

His His His His His His
                165
```

<210> 67
<211> 218
<212> PRT
<213> Artificial

<220>
<223> G1N(deltaext)

<400> 67

```
Met Ala Ser Trp Ser His Pro Gln Phe Glu Lys Ala Ser Lys Glu Thr
1               5                   10                  15

Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala His Thr Ala
            20                  25                  30

Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met Thr Pro Leu
         35                  40                  45

Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp Gly Thr Thr
     50                  55                  60

Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp Gln Thr Ser
```

65 70 75 80

Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr Glu Asp Val
85 90 95

Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg Thr Ala Phe
100 105 110

Ala Gly Thr Ala Ile Ser Ile Val Gly Ser Cys Leu Asp Leu Lys Thr
115 120 125

Gln Val Gln Thr Pro Gln Gly Met Lys Glu Ile Ser Asn Ile Gln Val
130 135 140

Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr Asn Glu Val Leu Asn Val
145 150 155 160

Phe Pro Lys Ser Lys Lys Lys Ser Tyr Lys Ile Thr Leu Glu Asp Gly
165 170 175

Lys Glu Ile Ile Cys Ser Glu Glu His Leu Phe Pro Thr Gln Thr Gly
180 185 190

Glu Met Asn Ile Ser Gly Gly Leu Lys Glu Gly Met Cys Leu Tyr Val
195 200 205

Lys Glu Gly Gly His His His His His His
210 215

<210> 68
<211> 170
<212> PRT
<213> Artificial

<220>
<223> Proteína del fragmento C de GP41.1 N a A

<400> 68

Met Gly Lys Asn Ser Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu
1 5 10 15

Leu Asp Glu Arg Glu Leu Ile Asp Ile Glu Val Ser Gly Asn His Leu
20 25 30

Phe Tyr Ala Asn Asp Ile Leu Thr His Ala Ser Ser Ser Asp Val Gly
35 40 45

Thr Gly Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr
50 55 60

```
Asp Val Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His
65                  70                  75                  80

Trp Cys Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala
                85                  90                  95

Asp Glu Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His
            100                 105                 110

Asn Pro Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu
        115                 120                 125

Leu Leu Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu
    130                 135                 140

Ser Lys Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser
145                 150                 155                 160

Glu Phe Arg Ser His His His His His His
                165                 170
```

<210> 69
<211> 173
<212> PRT
<213> Artificial

<220>
<223> Fragmento C-terminal de GP41-8 N a A

<400> 69

```
Met Cys Glu Ile Phe Glu Asn Glu Ile Asp Trp Asp Glu Ile Ala Ser
1               5                   10                  15

Ile Glu Tyr Val Gly Val Glu Glu Thr Ile Asp Ile Asn Val Thr Asn
            20                  25                  30

Asp Arg Leu Phe Phe Ala Asn Gly Ile Leu Thr His Ala Ser Ala Val
        35                  40                  45

Glu Glu Gly Thr Gly Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser
    50                  55                  60

Phe Asp Thr Asp Val Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe
65                  70                  75                  80

Trp Ala His Trp Cys Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp
                85                  90                  95
```

```
Glu Ile Ala Asp Glu Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn
            100                 105                 110

Ile Asp His Asn Pro Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile
        115                 120                 125

Pro Thr Leu Leu Leu Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val
    130                 135                 140

Gly Ala Leu Ser Lys Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu
145                 150                 155                 160

Ala Gly Ser Glu Phe Arg Ser His His His His His His
                165                 170
```

<210> 70
<211> 168
<212> PRT
<213> Artificial

<220>
<223> Fragmento C-terminal de NrdJ1 N a A

<400> 70

```
Met Glu Ala Lys Thr Tyr Ile Gly Lys Leu Lys Ser Arg Lys Ile Val
1               5                   10                  15

Ser Asn Glu Asp Thr Tyr Asp Ile Gln Thr Ser Thr His Asn Phe Phe
            20                  25                  30

Ala Asn Asp Ile Leu Val His Ala Ser Glu Ile Val Leu Gly Thr Gly
        35                  40                  45

Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val
    50                  55                  60

Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys
65                  70                  75                  80

Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu
                85                  90                  95

Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro
            100                 105                 110

Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu
        115                 120                 125

Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys
    130                 135                 140

Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe
145                 150                 155                 160

Arg Ser His His His His His His
                165
```

<210> 71
<211> 168
<212> PRT
<213> Artificial

<220>
<223> Fragmento C-terminal de IMPDH1 N a A

<400> 71

```
Met Lys Phe Lys Leu Lys Glu Ile Thr Ser Ile Glu Thr Lys His Tyr
1               5                  10                  15

Lys Gly Lys Val His Asp Leu Thr Val Asn Gln Asp His Ser Tyr Asn
            20                  25                  30

Val Arg Gly Thr Val Val His Ala Ser Ile Cys Ser Thr Gly Thr Gly
        35                  40                  45

Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val
    50                  55                  60

Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys
65                  70                  75                  80

Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu
                85                  90                  95

Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro
            100                 105                 110

Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu
        115                 120                 125

Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys
    130                 135                 140

Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe
145                 150                 155                 160

Arg Ser His His His His His His
                165
```

<210> 72
<211> 170
<212> PRT
<213> Artificial

<220>
<223> Fragmento C-terminal de GP41.1 N/S a A

<400> 72

```
Met Gly Lys Asn Ser Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu
1               5                  10                  15

Leu Asp Glu Arg Glu Leu Ile Asp Ile Glu Val Ser Gly Asn His Leu
            20                  25                  30

Phe Tyr Ala Asn Asp Ile Leu Thr His Ala Ala Ser Ser Asp Val Gly
        35                  40                  45

Thr Gly Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr
    50                  55                  60

Asp Val Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His
65                  70                  75                  80

Trp Cys Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala
                85                  90                  95

Asp Glu Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His
            100                 105                 110

Asn Pro Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu
        115                 120                 125

Leu Leu Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu
    130                 135                 140

Ser Lys Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser
145                 150                 155                 160

Glu Phe Arg Ser His His His His His His
                165                 170
```

<210> 73
<211> 173
<212> PRT
<213> Artificial

<220>
<223> Fragmento C-terminal de GP41.8 N/S a A

<400> 73

```
Met Cys Glu Ile Phe Glu Asn Glu Ile Asp Trp Asp Glu Ile Ala Ser
1               5                  10                  15

Ile Glu Tyr Val Gly Val Glu Glu Thr Ile Asp Ile Asn Val Thr Asn
            20                  25                  30

Asp Arg Leu Phe Phe Ala Asn Gly Ile Leu Thr His Ala Ala Ala Val
        35                  40                  45

Glu Glu Gly Thr Gly Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser
    50                  55                  60

Phe Asp Thr Asp Val Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe
65                  70                  75                  80

Trp Ala His Trp Cys Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp
                85                  90                  95

Glu Ile Ala Asp Glu Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn
            100                 105                 110

Ile Asp His Asn Pro Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile
        115                 120                 125

Pro Thr Leu Leu Leu Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val
    130                 135                 140

Gly Ala Leu Ser Lys Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu
145                 150                 155                 160

Ala Gly Ser Glu Phe Arg Ser His His His His His His
                165                 170
```

<210> 74
<211> 168
<212> PRT
<213> Artificial

<220>
<223> Fragmento C-terminal de NrdJ1 N/S a A

<400> 74

```
Met Glu Ala Lys Thr Tyr Ile Gly Lys Leu Lys Ser Arg Lys Ile Val
1               5                  10                  15

Ser Asn Glu Asp Thr Tyr Asp Ile Gln Thr Ser Thr His Asn Phe Phe
```

20 25 30

Ala Asn Asp Ile Leu Val His Ala Ala Glu Ile Val Leu Gly Thr Gly
35 40 45

Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val
50 55 60

Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys
65 70 75 80

Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu
85 90 95

Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro
100 105 110

Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu
115 120 125

Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys
130 135 140

Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe
145 150 155 160

Arg Ser His His His His His His
165

<210> 75
<211> 168
<212> PRT
<213> Artificial

<220>
<223> Fragmento C-terminal de IMPDH1 N/S a A

<400> 75

Met Lys Phe Lys Leu Lys Glu Ile Thr Ser Ile Glu Thr Lys His Tyr
1 5 10 15

Lys Gly Lys Val His Asp Leu Thr Val Asn Gln Asp His Ser Tyr Asn
20 25 30

Val Arg Gly Thr Val Val His Ala Ala Ile Cys Ser Thr Gly Thr Gly
35 40 45

Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp Val
50 55 60

```
Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp Cys
65                  70                  75                  80

Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp Glu
                85                  90                  95

Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn Pro
            100                 105                 110

Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu Leu
        115                 120                 125

Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser Lys
    130                 135                 140

Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser Glu Phe
145                 150                 155                 160

Arg Ser His His His His His His
                165
```

<210> 76
<211> 109
<212> PRT
<213> bacteriófago lambda

<400> 76

```
Lys Glu Thr Phe Thr His Tyr Gln Pro Gln Gly Asn Ser Asp Pro Ala
1               5                   10                  15

His Thr Ala Thr Ala Pro Gly Gly Leu Ser Ala Lys Ala Pro Ala Met
            20                  25                  30

Thr Pro Leu Met Leu Asp Thr Ser Ser Arg Lys Leu Val Ala Trp Asp
        35                  40                  45

Gly Thr Thr Asp Gly Ala Ala Val Gly Ile Leu Ala Val Ala Ala Asp
    50                  55                  60

Gln Thr Ser Thr Thr Leu Thr Phe Tyr Lys Ser Gly Thr Phe Arg Tyr
65                  70                  75                  80

Glu Asp Val Leu Trp Pro Glu Ala Ala Ser Asp Glu Thr Lys Lys Arg
                85                  90                  95

Thr Ala Phe Ala Gly Thr Ala Ile Ser Ile Val Gly Ser
            100                 105
```

<210> 77
<211> 111
<212> PRT
<213> Escherichia coli

<400> 77

```
Gly Ser Asp Lys Ile Ile His Leu Thr Asp Asp Ser Phe Asp Thr Asp
1               5                   10                  15

Val Leu Lys Ala Asp Gly Ala Ile Leu Val Asp Phe Trp Ala His Trp
            20                  25                  30

Cys Gly Pro Cys Lys Met Ile Ala Pro Ile Leu Asp Glu Ile Ala Asp
        35                  40                  45

Glu Tyr Gln Gly Lys Leu Thr Val Ala Lys Leu Asn Ile Asp His Asn
    50                  55                  60

Pro Gly Thr Ala Pro Lys Tyr Gly Ile Arg Gly Ile Pro Thr Leu Leu
65                  70                  75                  80

Leu Phe Lys Asn Gly Glu Val Ala Ala Thr Lys Val Gly Ala Leu Ser
                85                  90                  95

Lys Gly Gln Leu Lys Glu Phe Leu Asp Ala Asn Leu Ala Gly Ser
            100                 105                 110
```

<210> 78
<211> 9
<212> PRT
<213> Artificial

<220>
<223> Caja C1 de la región C-terminal de la inteína GP41-1

<400> 78

```
Ala Asn Asp Ile Leu Thr His Asn Ser
1               5
```

<210> 79
<211> 88
<212> PRT
<213> Artificial

<220>
<223> N-inteína gp41-1

<400> 79

```
Cys Leu Asp Leu Lys Thr Gln Val Gln Thr Pro Gln Gly Met Lys Glu
1               5                   10                  15
```

Ile Ser Asn Ile Gln Val Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr
20 25 30

Asn Glu Val Leu Asn Val Phe Pro Lys Ser Lys Lys Lys Ser Tyr Lys
35 40 45

Ile Thr Leu Glu Asp Gly Lys Glu Ile Ile Cys Ser Glu Glu His Leu
50 55 60

Phe Pro Thr Gln Thr Gly Glu Met Asn Ile Ser Gly Gly Leu Lys Glu
65 70 75 80

Gly Met Cys Leu Tyr Val Lys Glu
85

<210> 80
<211> 27
<212> PRT
<213> Artificial

<220>
<223> N-inteína gp41-2

<400> 80

Cys Leu Asp Leu Lys Thr Gln Val Gln Thr Gln Gln Gly Leu Lys Asp
1 5 10 15

Ile Ser Asn Ile Gln Val Gly Asp Leu Val Leu
20 25

<210> 81
<211> 46
<212> PRT
<213> Artificial

<220>
<223> N-inteína gp41-3

<400> 81

Cys Leu Asp Leu Lys Thr Gln Val Gln Thr Pro Gln Gly Met Lys Glu
1 5 10 15

Ile Ser Asn Ile Gln Val Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr
20 25 30

Asn Glu Val Leu Asn Val Phe Pro Lys Ser Lys Lys Lys Ser
35 40 45

<210> 82
<211> 88
<212> PRT
<213> Artificial

<220>
<223> N-inteína gp41-4

<400> 82

```
Cys Leu Asp Leu Lys Thr Gln Val Gln Thr Pro Gln Gly Met Lys Glu
1               5                  10                  15

Ile Ser Asn Ile Gln Val Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr
            20                  25                  30

Asn Glu Val Leu Asn Val Phe Pro Lys Ser Lys Lys Lys Ser Tyr Lys
        35                  40                  45

Ile Thr Leu Glu Asp Gly Lys Glu Ile Ile Cys Ser Glu Glu His Leu
    50                  55                  60

Phe Pro Thr Gln Thr Gly Glu Met Asn Ile Ser Gly Gly Leu Lys Glu
65                  70                  75                  80

Gly Met Cys Leu Tyr Val Lys Glu
                85
```

<210> 83
<211> 88
<212> PRT
<213> Artificial

<220>
<223> N-inteína gp41-5

<400> 83

```
Cys Leu Asp Leu Lys Thr Gln Val Gln Thr Pro Gln Gly Met Lys Glu
1               5                  10                  15

Ile Ser Asn Ile Gln Val Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr
            20                  25                  30

Asn Glu Val Leu Asn Val Phe Pro Lys Ser Lys Lys Lys Ser Tyr Lys
        35                  40                  45

Ile Thr Leu Glu Asp Gly Lys Glu Ile Ile Cys Ser Glu Glu His Leu
    50                  55                  60

Phe Pro Thr Gln Thr Gly Glu Met Asn Ile Ser Gly Gly Leu Lys Glu
65                  70                  75                  80

Gly Met Cys Leu Tyr Val Lys Glu
                85
```

<210> 84
<211> 43
<212> PRT
<213> Artificial

<220>
<223> N-inteína gp41-6

<400> 84

```
Ser Tyr Lys Ile Thr Leu Glu Asp Gly Lys Glu Ile Ile Cys Ser Glu
1               5                   10                  15

Glu His Leu Phe Pro Thr Gln Asn Gly Glu Val Asn Ile Lys Gly Gly
            20                  25                  30

Leu Lys Glu Gly Met Cys Leu Tyr Val Lys Glu
        35                  40
```

<210> 85
<211> 88
<212> PRT
<213> Artificial

<220>
<223> N-inteína gp41-7

<400> 85

```
Cys Leu Asp Leu Lys Thr Gln Val Gln Thr Pro Gln Gly Met Lys Glu
1               5                   10                  15

Leu Ser Asn Ile Gln Val Gly Asp Leu Val Leu Ser Asn Thr Gly Tyr
            20                  25                  30

Asn Gln Val Leu Asn Val Phe Pro Lys Ser Lys Lys Lys Ser Tyr Lys
        35                  40                  45

Ile Thr Leu Glu Asp Gly Lys Glu Ile Ile Cys Ser Glu Glu His Leu
    50                  55                  60

Phe Pro Thr Gln Asn Gly Glu Val Asn Ile Lys Gly Gly Leu Lys Glu
65                  70                  75                  80

Gly Met Cys Leu Tyr Val Lys Glu
                85
```

<210> 86
<211> 89
<212> PRT
<213> Artificial

<220>
<223> N-inteína gp41-8

<400> 86

```
Cys Leu Ser Leu Asp Thr Met Val Val Thr Asn Gly Lys Ala Ile Glu
1               5                  10                  15

Ile Arg Asp Val Lys Val Gly Asp Trp Leu Glu Ser Glu Cys Gly Pro
            20                  25                  30

Val Gln Val Thr Glu Val Leu Pro Ile Ile Lys Gln Pro Val Phe Glu
        35                  40                  45

Ile Val Leu Lys Ser Gly Lys Lys Ile Arg Val Ser Ala Asn His Lys
    50                  55                  60

Phe Pro Thr Lys Asp Gly Leu Lys Thr Ile Asn Ser Gly Leu Lys Val
65                  70                  75                  80

Gly Asp Phe Leu Arg Ser Arg Ala Lys
                85
```

<210> 87
<211> 101
<212> PRT
<213> Artificial

<220>
<223> N-inteína IMPDH-1

<400> 87

```
Cys Phe Val Pro Gly Thr Leu Val Asn Thr Glu Asn Gly Leu Lys Lys
1               5                  10                  15

Ile Glu Glu Ile Lys Val Gly Asp Lys Val Phe Ser His Thr Gly Lys
            20                  25                  30

Leu Gln Glu Val Val Asp Thr Leu Ile Phe Asp Arg Asp Glu Glu Ile
        35                  40                  45

Ile Ser Ile Asn Gly Ile Asp Cys Thr Lys Asn His Glu Phe Tyr Val
    50                  55                  60

Ile Asp Lys Glu Asn Ala Asn Arg Val Asn Glu Asp Asn Ile His Leu
65                  70                  75                  80

Phe Ala Arg Trp Val His Ala Glu Glu Leu Asp Met Lys Lys His Leu
                85                  90                  95

Leu Ile Glu Leu Glu
            100
```

<210> 88
<211> 133
<212> PRT
<213> Artificial

<220>
<223> N-inteína NrdA-1

<400> 88

Cys Val Ala Gly Asp Thr Lys Ile Lys Ile Lys Tyr Pro Glu Ser Val
1 5 10 15

Gly Asp Gln Tyr Gly Thr Trp Tyr Trp Asn Val Leu Glu Lys Glu Ile
20 25 30

Gln Ile Glu Asp Leu Glu Asp Tyr Ile Ile Met Arg Glu Cys Glu Ile
35 40 45

Tyr Asp Ser Asn Ala Pro Gln Ile Glu Val Leu Ser Tyr Asn Ile Glu
50 55 60

Thr Gly Glu Gln Glu Trp Lys Pro Ile Thr Ala Phe Ala Gln Thr Ser
65 70 75 80

Pro Lys Ala Lys Val Met Lys Ile Thr Asp Glu Glu Ser Gly Lys Ser
85 90 95

Ile Val Val Thr Pro Glu His Gln Val Phe Thr Lys Asn Arg Gly Tyr
100 105 110

Val Met Ala Lys Asp Leu Ile Glu Thr Asp Glu Pro Ile Ile Val Asn
115 120 125

Lys Asp Met Asn Phe
130

<210> 89
<211> 106
<212> PRT
<213> Artificial

<220>
<223> N-inteína NrdA-2

<400> 89

Cys Leu Thr Gly Asp Ala Lys Ile Asp Val Leu Ile Asp Asn Ile Pro

1 5 10 15

Ile Ser Gln Ile Ser Leu Glu Glu Val Val Asn Leu Phe Asn Glu Gly
20 25 30

Lys Glu Ile Tyr Val Leu Ser Tyr Asn Ile Asp Thr Lys Glu Val Glu
35 40 45

Tyr Lys Glu Ile Ser Asp Ala Gly Leu Ile Ser Glu Ser Ala Glu Val
50 55 60

Leu Glu Ile Ile Asp Glu Glu Thr Gly Gln Lys Ile Val Cys Thr Pro
65 70 75 80

Asp His Lys Val Tyr Thr Leu Asn Arg Gly Tyr Val Ser Ala Lys Asp
85 90 95

Leu Lys Glu Asp Asp Glu Leu Val Phe Ser
100 105

<210> 90
<211> 105
<212> PRT
<213> Artificial

<220>
<223> N-inteína NrdA-4

<400> 90

Cys Leu Ala Gly Asp Thr Thr Val Thr Val Leu Glu Gly Asp Ile Val
1 5 10 15

Phe Glu Met Thr Leu Glu Asn Leu Val Ser Leu Tyr Lys Asn Val Phe
20 25 30

Ser Val Ser Val Leu Ser Phe Asn Pro Glu Thr Gln Lys Gln Glu Phe
35 40 45

Lys Pro Val Thr Asn Ala Ala Leu Met Asn Pro Glu Ser Lys Val Leu
50 55 60

Lys Ile Thr Asp Ser Asp Thr Gly Lys Ser Ile Val Cys Thr Pro Asp
65 70 75 80

His Lys Val Phe Thr Lys Asn Arg Gly Tyr Val Ile Ala Ser Glu Leu
85 90 95

Asn Ala Glu Asp Ile Leu Glu Ile Lys
100 105

<210> 91
<211> 65
<212> PRT
<213> Artificial

<220>
<223> N-inteína NrdA-5

<400> 91

```
His Thr Glu Thr Val Arg Arg Val Gly Thr Ile Thr Ala Phe Ala Gln
1               5                  10                  15

Thr Ser Pro Lys Ser Lys Val Met Lys Ile Thr Asp Glu Glu Ser Gly
            20                  25                  30

Asn Ser Ile Val Val Thr Pro Glu His Lys Val Phe Thr Lys Asn Arg
        35                  40                  45

Gly Tyr Val Met Ala Lys Asn Leu Val Glu Thr Asp Glu Leu Val Ile
    50                  55                  60

Asn
65
```

<210> 92
<211> 49
<212> PRT
<213> Artificial

<220>
<223> N-inteína NrdA-6

<400> 92

```
Tyr Val Cys Ser Arg Asp Asp Thr Thr Gly Phe Lys Leu Ile Cys Thr
1               5                  10                  15

Pro Asp His Met Ile Tyr Thr Lys Asn Arg Gly Tyr Ile Met Ala Lys
            20                  25                  30

Tyr Leu Lys Glu Asp Asp Glu Leu Leu Ile Asn Glu Ile His Leu Pro
        35                  40                  45

Thr
```

<210> 93
<211> 105
<212> PRT
<213> Artificial

<220>
<223> N-inteína NrdJ-1

<400> 93

```
Cys Leu Val Gly Ser Ser Glu Ile Ile Thr Arg Asn Tyr Gly Lys Thr
1               5                   10                  15

Thr Ile Lys Glu Val Val Glu Ile Phe Asp Asn Asp Lys Asn Ile Gln
            20                  25                  30

Val Leu Ala Phe Asn Thr His Thr Asp Asn Ile Glu Trp Ala Pro Ile
        35                  40                  45

Lys Ala Ala Gln Leu Thr Arg Pro Asn Ala Glu Leu Val Glu Leu Glu
    50                  55                  60

Ile Asp Thr Leu His Gly Val Lys Thr Ile Arg Cys Thr Pro Asp His
65                  70                  75                  80

Pro Val Tyr Thr Lys Asn Arg Gly Tyr Val Arg Ala Asp Glu Leu Thr
                85                  90                  95

Asp Asp Asp Glu Leu Val Val Ala Ile
            100                 105
```

<210> 94
<211> 105
<212> PRT
<213> Artificial

<220>
<223> N-inteína NrdJ-2

<400> 94

```
Cys Leu Val Gly Ser Ser Glu Ile Ile Thr Arg Asn Tyr Gly Lys Thr
1               5                   10                  15

Thr Ile Lys Glu Val Val Glu Ile Phe Asp Asn Asp Lys Asn Ile Gln
            20                  25                  30

Val Leu Ala Phe Asn Thr His Thr Asp Asn Ile Glu Trp Ala Pro Ile
        35                  40                  45

Lys Ala Ala Gln Leu Thr Arg Pro Asn Ala Glu Leu Val Glu Leu Glu
    50                  55                  60

Ile Asn Thr Leu His Gly Val Lys Thr Ile Arg Cys Thr Pro Asp His
65                  70                  75                  80

Pro Val Tyr Thr Lys Asn Arg Asp Tyr Val Arg Ala Asp Glu Leu Thr
                85                  90                  95

Asp Asp Asp Glu Leu Val Val Ala Ile
            100                 105
```

<210> 95
<211> 38
<212> PRT
<213> Artificial

<220>
<223> C-inteína gp41-1

<400> 95

Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu Leu Asp Glu Arg Glu
1 5 10 15

Leu Ile Asp Ile Glu Val Ser Gly Asn His Leu Phe Tyr Ala Asn Asp
20 25 30

Ile Leu Thr His Asn Ser
35

<210> 96
<211> 38
<212> PRT
<213> Artificial

<220>
<223> C-inteína gp41-2

<400> 96

Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu Leu Asp Glu Arg Glu
1 5 10 15

Leu Ile Asp Ile Glu Val Ser Gly Asn His Leu Phe Tyr Ala Asn Ala
20 25 30

Ile Leu Thr His Asn Ser
35

<210> 97
<211> 26
<212> PRT
<213> Artificial

<220>
<223> C-inteína gp41-7

<400> 97

Met Met Leu Lys Lys Ile Leu Lys Ile Glu Glu Leu Asp Glu Arg Glu
1 5 10 15

Leu Ile Asp Ile Glu Val Ser Gly Asn His
20 25

<210> 98
<211> 46
<212> PRT
<213> Artificial

<220>
<223> C-inteína gp41-8

<400> 98

Met Cys Glu Ile Phe Glu Asn Glu Ile Asp Trp Asp Glu Ile Ala Ser
1 5 10 15

Ile Glu Tyr Val Gly Val Glu Glu Thr Ile Asp Ile Asn Val Thr Asn
20 25 30

Asp Arg Leu Phe Phe Ala Asn Gly Ile Leu Thr His Asn Ser
35 40 45

<210> 99
<211> 47
<212> PRT
<213> Artificial

<220>
<223> C-inteína gp41-9

<400> 99

Met Ile Met Lys Asn Arg Glu Arg Phe Ile Thr Glu Lys Ile Leu Asn
1 5 10 15

Ile Glu Glu Ile Asp Asp Asp Leu Thr Val Asp Ile Gly Met Asp Asn
20 25 30

Glu Asp His Tyr Phe Val Ala Asn Asp Ile Leu Thr His Asn Thr
35 40 45

<210> 100
<211> 41
<212> PRT
<213> Artificial

<220>
<223> C-inteína IMPDH-1

<400> 100

Met Lys Phe Lys Leu Lys Glu Ile Thr Ser Ile Glu Thr Lys His Tyr
1 5 10 15

Lys Gly Lys Val His Asp Leu Thr Val Asn Gln Asp His Ser Tyr Asn
20 25 30

Val Arg Gly Thr Val Val His Asn Ser
35 40

<210> 101
<211> 43
<212> PRT
<213> Artificial

<220>
<223> C-inteína IMPDH-2

<400> 101

Met Lys Phe Thr Leu Glu Pro Ile Thr Lys Ile Asp Ser Tyr Glu Val
1 5 10 15

Thr Ala Glu Pro Val Tyr Asp Ile Glu Val Glu Asn Asp His Ser Phe
20 25 30

Cys Val Glu Asn Gly Phe Val Val His Asn Ser
35 40

<210> 102
<211> 41
<212> PRT
<213> Artificial

<220>
<223> C-inteína IMPDH-3

<400> 102

Met Lys Phe Lys Leu Val Glu Ile Thr Ser Lys Glu Thr Phe Asn Tyr
1 5 10 15

Ser Gly Gln Val His Asp Leu Thr Val Glu Asp Asp His Ser Tyr Ser
20 25 30

Ile Asn Asn Ile Val Val His Asn Ser
35 40

<210> 103
<211> 35
<212> PRT
<213> Artificial

<220>
<223> C-inteína NrdA-2

<400> 103

Met Gly Leu Lys Ile Ile Lys Arg Glu Ser Lys Glu Pro Val Phe Asp
1 5 10 15

Ile Thr Val Lys Asp Asn Ser Asn Phe Phe Ala Asn Asn Ile Leu Val
20 25 30

His Asn Cys
35

<210> 104
<211> 34
<212> PRT
<213> Artificial

<220>
<223> NrdA-3

<400> 104

Met Leu Lys Ile Glu Tyr Leu Glu Glu Glu Ile Pro Val Tyr Asp Ile
1 5 10 15

Thr Val Glu Glu Thr His Asn Phe Phe Ala Asn Asp Ile Leu Ile His
20 25 30

Asn Cys

<210> 105
<211> 28
<212> PRT
<213> Artificial

<220>
<223> C-inteína NrdA-5

<400> 105

Met Leu Lys Ile Glu Tyr Leu Glu Glu Glu Ile Pro Val Tyr Asp Ile
1 5 10 15

Thr Val Glu Gly Thr His Asn Leu Ala Tyr Ser Leu
20 25

<210> 106
<211> 33
<212> PRT
<213> Artificial

<220>
<223> C-inteína NrdA-6

<400> 106

Met Gly Ile Lys Ile Arg Lys Leu Glu Gln Asn Arg Val Tyr Asp Ile
1 5 10 15

Lys Val Glu Lys Ile Ile Ile Phe Cys Asn Asn Ile Leu Val His Asn
20 25 30

Cys

<210> 107
<211> 34
<212> PRT
<213> Artificial

<220>
<223> C-inteína NrdA-7

<400> 107

Met Leu Lys Ile Glu Tyr Leu Glu Glu Glu Ile Pro Val Tyr Asp Ile
1 5 10 15

Thr Val Glu Lys Thr Asn Asn Phe Phe Ala Asn Asp Ile Leu Val His
20 25 30

Asn Cys

<210> 108

<211> 41
<212> PRT
<213> Artificial

<220>
<223> C-inteína NrdJ-1

<400> 108

```
Met Glu Ala Lys Thr Tyr Ile Gly Lys Leu Lys Ser Arg Lys Ile Val
1               5                   10                  15

Ser Asn Glu Asp Thr Tyr Asp Ile Gln Thr Ser Thr His Asn Phe Phe
            20                  25                  30

Ala Asn Asp Ile Leu Val His Asn Ser
        35                  40
```

REIVINDICACIONES

1. Una proteína de fusión que comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y (ii) un polipéptido heterólogo, en la que el polipéptido heterólogo es C-terminal respecto al dominio de inteína.

2. La proteína de fusión de la reivindicación 1 en la que el último aminoácido del dominio de inteína es glutamina o asparagina.

3. La proteína de fusión de la reivindicación 2 en la que el primer aminoácido del péptido heterólogo es serina, cisteína o treonina.

4. La proteína de fusión de la reivindicación 1 en la que el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina y en la que el primer aminoácido del polipéptido heterólogo es un aminoácido distinto de serina, cisteína o treonina.

5. Una proteína de fusión que comprende (i) un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y (ii) un polipéptido heterólogo, en la que el polipéptido heterólogo es N-terminal respecto al dominio de inteína.

6. La proteína de fusión de la reivindicación 5, en la que el primer aminoácido del dominio de inteína es una serina o cisteína.

7. La proteína de fusión de las reivindicaciones 5 ó 6 en la que el primer aminoácido del dominio de inteína es un aminoácido distinto de serina o cisteína.

8. Una proteína de fusión que comprende un primer dominio de inteína, un segundo dominio de inteína, y un polipéptido heterólogo, en la que el polipéptido heterólogo es N-terminal respecto al primer dominio de inteína, y en la que el polipéptido heterólogo es C-terminal respecto al segundo dominio de inteína y en la que

(a) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:3 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:7;

(b) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:12 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:16;

(c) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:20 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:24; o

(d) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:34 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:38; o

(e) el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:64 y el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:65.

9. La proteína de fusión de la reivindicación 8 en la que el polipéptido heterólogo y el segundo dominio de inteína están conectados por un enlace peptídico o por un conector y en la que el primer aminoácido del polipéptido heterólogo o el primer aminoácido del conector es serina, cisteína, o treonina.

10. Una composición o kit de partes que comprende un primer componente y un segundo componente en la que

(i) el primer componente es la proteína de fusión de una cualquiera de las reivindicaciones 1-3 y

(ii) el segundo componente se selecciona del grupo que consiste en la proteína de fusión de la reivindicación 7 y un dominio N-terminal de inteína;

en la que

a. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:7 y el dominio de inteína de la proteína de fusión de la reivindicación 7 o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:3;

b. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:16 y el dominio de inteína de la proteína de fusión de la reivindicación 7 o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:12;

c. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:24 y el dominio de inteína de la proteína de fusión de la reivindicación 7 o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:20;

d. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:38 y el dominio de inteína de la proteína de fusión de la reivindicación 7 o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:34 o

e. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:65 y el dominio de inteína de la proteína de fusión de la reivindicación 7 o el dominio N-terminal de inteína es al menos 75% idéntico a SEQ ID NO:64.

11. Una composición o kit de partes que comprende un primer componente y un segundo componente en la que

(i) el primer componente es la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 y

(ii) el segundo componente se selecciona del grupo que consiste en las proteínas de fusión de una cualquiera de las reivindicaciones 4 y un dominio C-terminal de inteína; en la que

a. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:3 y el dominio de inteína de la proteína de fusión de la reivindicación 4 o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:7;

b. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:12 y el dominio de inteína de la proteína de fusión de la reivindicación 4 o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:16;

c. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:20 y el dominio de inteína de la proteína de fusión de la reivindicación 4 o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:24;

d. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:34 y el dominio de inteína de la proteína de fusión de la reivindicación 4 o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:38 o

e. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:64 y el dominio de inteína de la proteína de fusión de la reivindicación 4 o el dominio C-terminal de inteína es al menos 75% idéntico a SEQ ID NO:65.

12. Una composición o kit de partes que comprende la proteína de fusión de una cualquiera de las reivindicaciones 1-3 y la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 en la que

a. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:7 y el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:3;

b. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:16 y el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:12;

c. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:24 y el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:20;

d. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:38 y el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:34; o

e. el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 1-3 es al menos 75% idéntico a SEQ ID NO:65 y el dominio de inteína de la proteína de fusión de una cualquiera de las reivindicaciones 5 ó 6 es al menos 75% idéntico a SEQ ID NO:64.

13. Un método seleccionado del grupo que consiste en:

(i) un método para escindir un polipéptido heterólogo de un dominio de inteína en el que el polipéptido heterólogo está unido al extremo C del dominio de inteína, que comprende incubar la composición de la reivindicación 10 o poner en asociación los componentes del kit de partes de la reivindicación 10 bajo condiciones que permitan la escisión de proteínas mediada por inteínas,

(ii) un método para escindir un polipéptido heterólogo de un dominio de inteína en el que el polipéptido heterólogo está unido al extremo N del dominio de inteína, que comprende incubar la composición de la reivindicación 11 o poner en asociación los componentes del kit de partes de la reivindicación 11 bajo condiciones que permitan la escisión de proteínas mediada por inteínas,

(iii) un método para unir covalentemente el extremo N de un primer polipéptido al extremo C de un segundo polipéptido que comprende incubar la composición de la reivindicación 12 o poner en asociación los componentes del kit de partes de la reivindicación 12 bajo condiciones que permitan el corte y empalme de la inteína en el que dicho polipéptido es el polipéptido heterólogo que forma parte de la proteína de fusión según cualquiera de las reivindicaciones 1 a 3 y dicho segundo polipéptido es el polipéptido heterólogo que forma parte de la proteína de fusión según cualquiera de las reivindicaciones 5 ó 6.

(iv) un método para ciclar un polipéptido heterólogo que comprende incubar la proteína de fusión de una cualquiera de las reivindicaciones 8 ó 9 bajo condiciones que permitan el corte y empalme de inteína en el que el polipéptido heterólogo es el polipéptido heterólogo que forma parte de la proteína de fusión de una cualquiera de las reivindicaciones 8 ó 9.

14. Un vector que comprende un polinucleótido que codifica un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65 y al menos un sitio de clonación aguas abajo de dicho polinucleótido que permite la clonación de un polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el dominio de inteína y el polipéptido codificado por el polinucleótido de interés.

15. Un vector según la reivindicación 14 en el que el último aminoácido del dominio de inteína es glutamina o asparagina.

16. Un vector según la reivindicación 15 en el que el penúltimo aminoácido del dominio de inteína es histidina.

17. Un vector según la reivindicación 14 en el que el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina.

18. Un vector según la reivindicación 17 en el que el último aminoácido del dominio de inteína es un aminoácido distinto de asparagina o glutamina, en el que el vector comprende además un polinucleótido que codifica un polipéptido que forma un péptido conector entre el dominio de inteína y el polipéptido codificado por el péptido heterólogo y en el que el primer aminoácido de dicho conector es un aminoácido distinto de serina, cisteína o treonina.

19. Un vector que comprende un polinucleótido que codifica un dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y al menos un sitio de clonación aguas arriba de dicho polinucleótido que permite la clonación de un polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el polipéptido codificado por el polinucleótido de interés y el dominio de inteína.

20. El vector según la reivindicación 19 en el que el primer aminoácido del dominio de inteína es serina o cisteína.

21. El vector según la reivindicación 19 en el que el primer aminoácido del dominio de inteína es un aminoácido distinto de serina o cisteína.

22. Un vector que comprende un polinucleótido que codifica un primer dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65, al menos un sitio de clonación aguas abajo de dicho polinucleótido que permite la clonación de un polinucleótido de interés, y un polinucleótido aguas abajo del sitio de clonación, que codifica un segundo dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64, de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende el polipéptido codificado por el polinucleótido de interés y el primer y segundo dominios de inteína en el que

a. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:7, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:3;

b. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:16, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:12;

c. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:24, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:20;

d. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:38, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:34 o

e. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:65, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:64.

23. El vector según la reivindicación 22 en el que el último aminoácido del dominio de inteína es glutamina o asparagina, en el que el penúltimo aminoácido del dominio de inteína es histidina y/o en el que el primer aminoácido del segundo dominio de inteína es serina o cisteína.

24. Un vector que comprende:

(i) un polinucleótido que codifica un primer dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 7, 16, 24, 38 y 65,

(ii) un primer sitio de clonación aguas abajo de dicho polinucleótido que codifica un primer dominio de inteína,

(iii) un polinucleótido que codifica un segundo dominio de inteína al menos 75% idéntico a una secuencia seleccionada del grupo que consiste en SEQ ID NOs: 3, 12, 20, 34 y 64 y

(iv) un segundo sitio de clonación aguas arriba de dicho polinucleótido que codifica un segundo dominio de inteína,

en el que el primer sitio de clonación permite la clonación de un primer polinucleótido de interés y el segundo sitio de clonación permite la clonación de un segundo polinucleótido de interés de manera que se forma un polinucleótido que codifica una proteína de fusión que comprende, en dicho orden, el polipéptido codificado por el segundo polinucleótido de interés, el segundo dominio de inteína, el primer dominio de inteína y el polipéptido codificado por el segundo polinucleótido de interés y en el que

a. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:7, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:3;

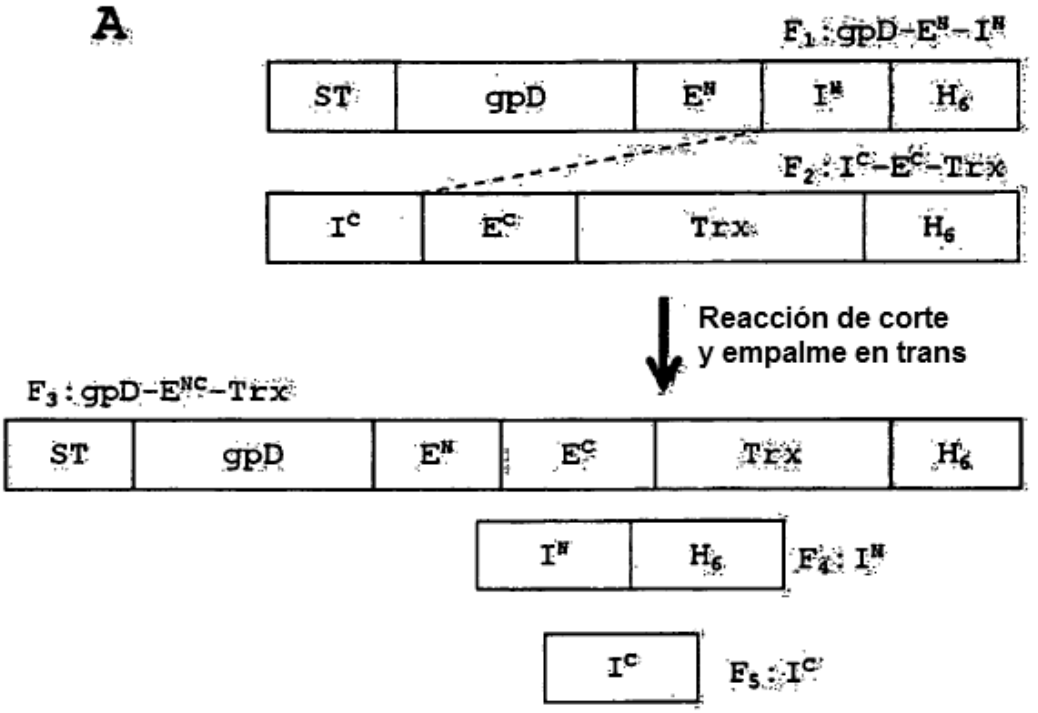
b. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:16, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:12;

c. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:24, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:20;

d. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:38, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:34 o

e. si el primer dominio de inteína es al menos 75% idéntico a SEQ ID NO:65, entonces el segundo dominio de inteína es al menos 75% idéntico a SEQ ID NO:64.

25. Un vector según la reivindicación 24 en el que el primer aminoácido del segundo dominio de inteína es cisteína o serina, en el que el último aminoácido del primer dominio de inteína es glutamina o asparagina, en el que el penúltimo aminoácido del primer dominio de inteína es histidina y/o en el que el primer aminoácido del segundo polipéptido de interés es cisteína, serina o treonina.


A
F1 : gpD-EN-IN
ST
gpD
EN
IN
H6
F2 : IC-EC-Trx
IC
EC
Trx
H6
Reacción de corte
y empalme en trans
F3 : gpD-ENC-Trx
ST
gpD
EN
EC
Trx
H6
IN
H6
F4 : IN
IC
F5 : IC

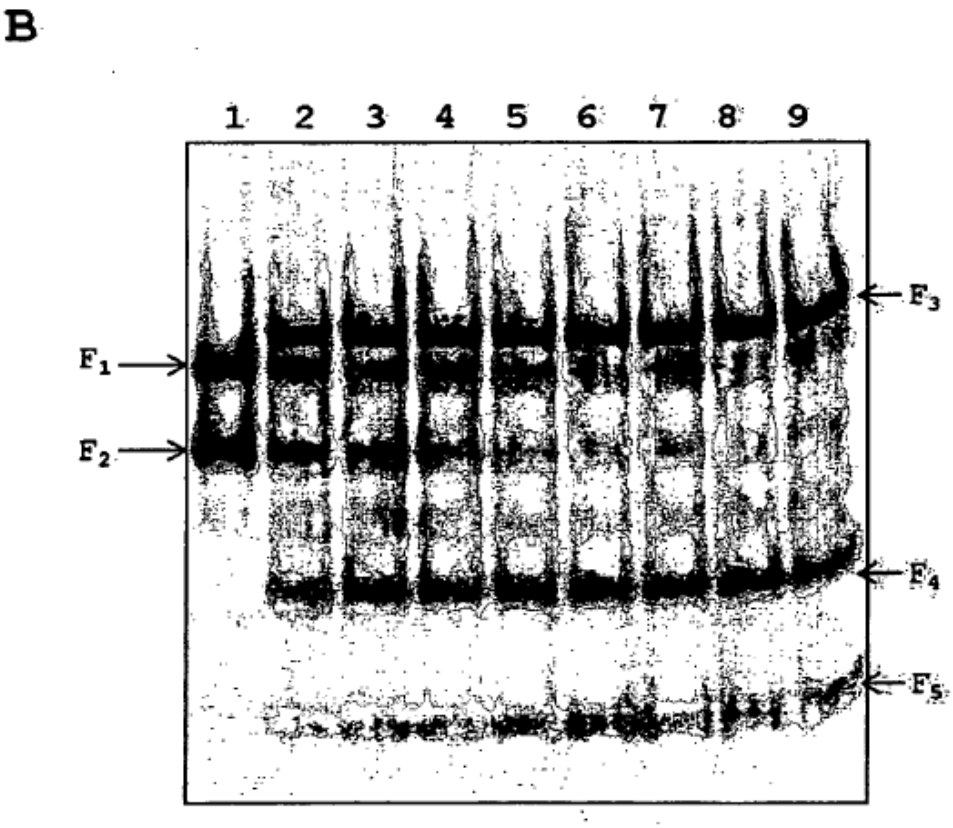

B
1 2 3 4 5 6 7 8 9
F1
F2
F3
F4
F5

FIG. 1

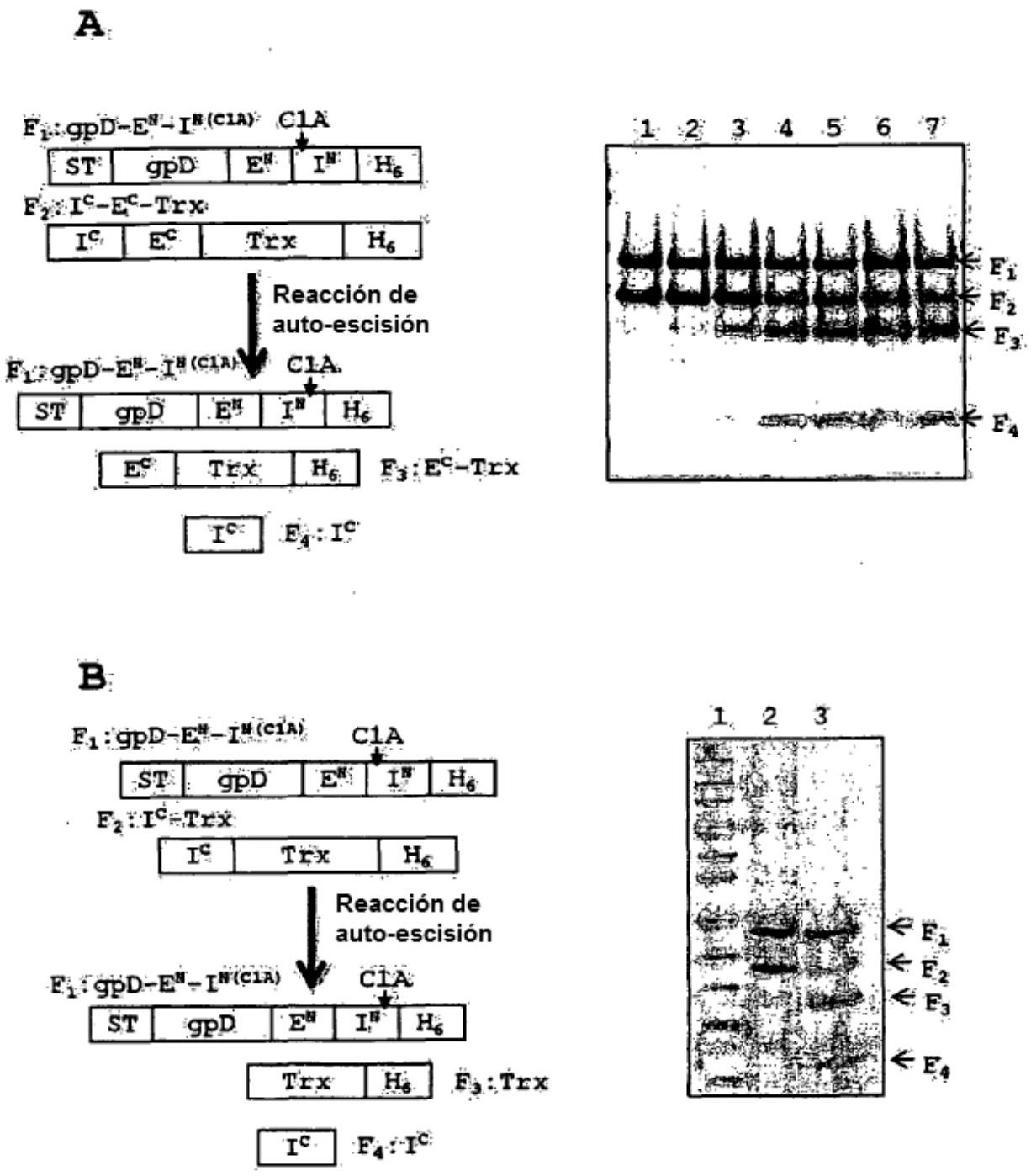

A
F1: gpD-EN-IN(C1A)
C1A
ST
gpD
EN
IN
H6
F2: IC-EC-Trx
IC
EC
Trx
H6
Reacción de auto-escisión
F1: gpD-EN-IN(C1A)
C1A
ST
gpD
EN
IN
H6
EC
Trx
H6
F3: EC-Trx
IC
F4: IC
1 2 3 4 5 6 7
F1
F2
F3
F4
B
F1: gpD-EN-IN(C1A)
C1A
ST
gpD
EN
IN
H6
F2: IC-Trx
IC
Trx
H6
Reacción de auto-escisión
F1: gpD-EN-IN(C1A)
C1A
ST
gpD
EN
IN
H6
Trx
H6
F3: Trx
IC
F4: IC
1 2 3
F1
F2
F3
F4

FIG. 2

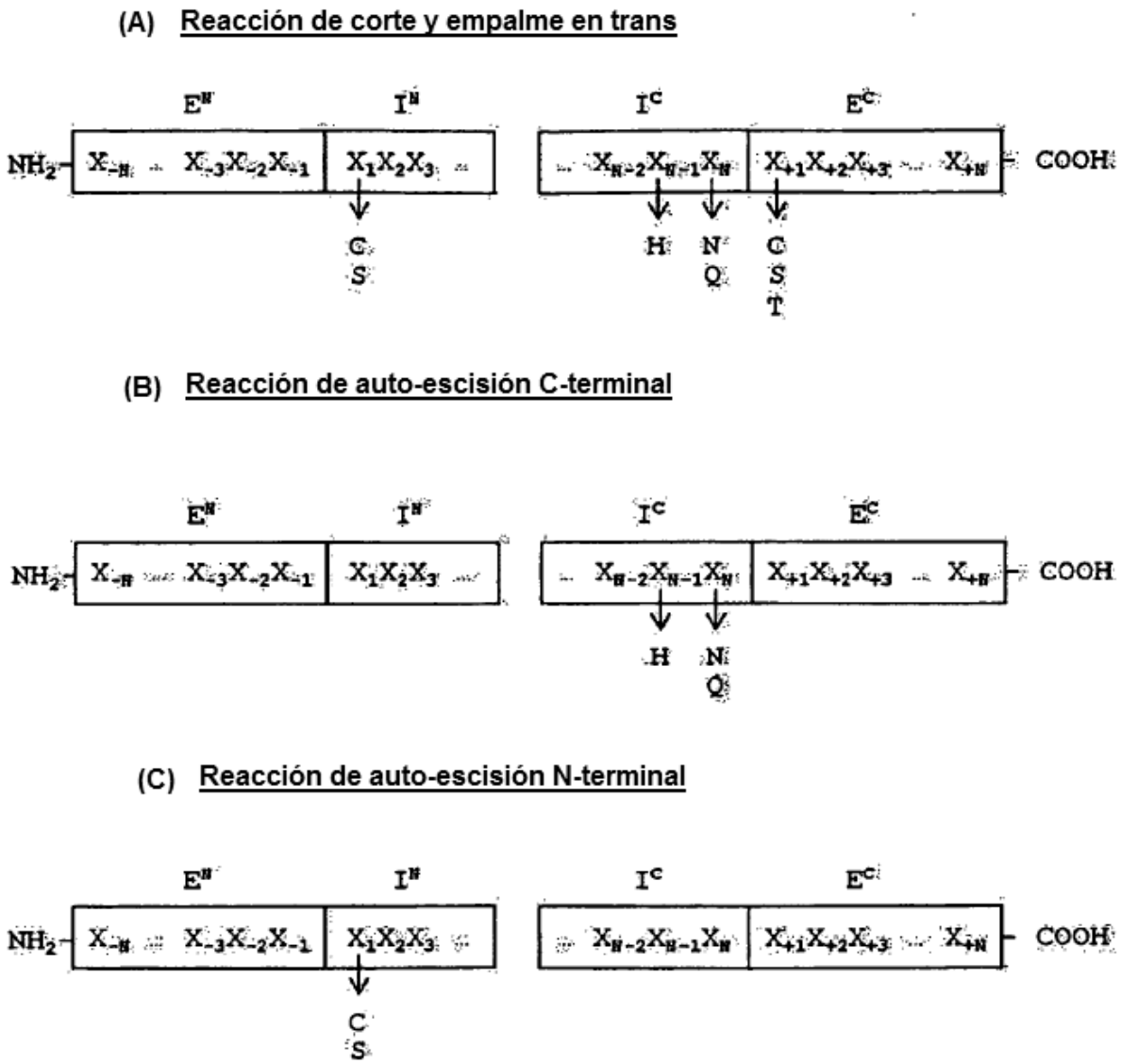

(A) Reacción de corte y empalme en trans
E^N
I^N
I^C
E^C
NH2
COOH
C
S
H
N
Q
C
S
T
(B) Reacción de auto-escisión C-terminal
E^N
I^N
I^C
E^C
NH2
COOH
H
N
Q
(C) Reacción de auto-escisión N-terminal
E^N
I^N
I^C
E^C
NH2
COOH
C
S

FIG. 3