

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 742 199**

51 Int. Cl.:

C12N 9/50

(2006.01)

C12N 9/00

(2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **23.10.2015** **PCT/US2015/057125**

87 Fecha y número de publicación internacional: **12.05.2016** **WO16073228**

96 Fecha de presentación y número de la solicitud europea: **23.10.2015** **E 15790411 (1)**

97 Fecha y número de publicación de la concesión europea: **29.05.2019** **EP 3215614**

54 Título: **Proteínas de fusión de inteína, solubles y métodos para la purificación de las biomoléculas**

30 Prioridad:

03.11.2014 US 201462074494 P

24.08.2015 US 201562209010 P

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
13.02.2020

73 Titular/es:

MERCK PATENT GMBH (100.0%)

Frankfurter Strasse 250

64293 Darmstadt, DE

72 Inventor/es:

ZILLMANN, MARTIN y

ORLANDO, JOE

74 Agente/Representante:

LEHMANN NOVO, María Isabel

ES 2 742 199 T3

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

DESCRIPCIÓN

Proteínas de fusión de inteína, solubles y métodos para la purificación de las biomoléculas.

Solicitudes relacionadas

- 5 Esta solicitud reivindica el beneficio de la Solicitud Provisional de los EE. UU. N.º 62/074,494, presentada el 3 de noviembre de 2014, y la Solicitud Provisional de los EE. UU. N.º 62/209,010, presentada el 24 de agosto de 2015. Todas las enseñanzas de las solicitudes anteriores se incorporan en la presente como referencia.

Campo de la invención

- 10 La presente invención se refiere a proteínas de fusión que comprenden un polipéptido de N-inteína y una contrapartida de solubilización de N-inteína, así como matrices de cromatografía de afinidad que comprenden dichas proteínas de fusión, según se describe en las reivindicaciones.

Antecedentes de la invención

- 15 Los métodos de purificación de proteínas que implican etiquetar una proteína de interés con una etiqueta de afinidad se usan ampliamente en el contexto de los laboratorios, para aplicaciones de investigación y desarrollo, pero han demostrado ser poco prácticos para operaciones de fabricación a gran escala. En la industria del bioprocesamiento, solo se utilizan etiquetas de afinidad escindibles para garantizar que el producto final no contenga la etiqueta, que debe eliminarse durante la producción, por lo general, utilizando una proteasa específica del sitio. La eliminación de la etiqueta de afinidad requiere pasos de proceso adicionales, que aumentan sustancialmente el costo y el tiempo, en especial, a escala industrial. Además, la escisión ineficiente y fuera del sitio conduce a la contaminación del producto proteico final, con proteínas que retienen la etiqueta y los fragmentos de proteínas truncadas, respectivamente, lo que no es aceptable en aplicaciones de bioprocesamiento.

Por consiguiente, existe la necesidad de desarrollar reactivos de cromatografía de afinidad y métodos mejorados que permitan la purificación de proteínas a gran escala, en condiciones industriales.

Compendio de la invención

- 25 Las inteínas constituyen una clase de enzimas autocatalíticas que contienen actividades tanto de proteasa como de ligasa. Una clase de inteínas, llamada "inteínas divididas", comprende dos medias inteínas complementarias, denominadas N-inteína y C-inteína, que se asocian de una manera selectiva y extremadamente estrecha para formar una enzima inteína activa (Shah NH, *et. al*, *J. Amer. Chem. Soc.* 135: 18673-18681; Dassa B., *et al.*, *Nucl. Acids Res.*, 37: 2560-2573 (2009)).

- 30 El uso de inteínas, incluidas las inteínas divididas, en procesos de purificación de proteínas a gran escala ya se ha descrito en el estado de la técnica (véase, por ejemplo, el documento de patente número WO 2013/045632). El uso de inteínas divididas para la separación cromatográfica de las proteínas de interés de mezclas sin procesar también se ha descrito con anterioridad (véase, por ejemplo, la publicación china N.º CN101884910; Guan D., *et al.*, *Biotech. Bioeng.* 110: 2471-2481 (2013); Lu W., *et al.*, *J. Chrom. A*, 1218: 2553-2560 (2011)).

- 35 Sin embargo, el uso de inteínas en procesos de purificación de proteínas a gran escala se ve obstaculizado por su escasa solubilidad cuando se expresan en sistemas de expresión comunes, tales como *E. coli*. Además, no se ha descrito una matriz de cromatografía que incluya un ligando de afinidad basado en inteína, que se una covalentemente a un soporte sólido, que es crítico para procesos eficientes de purificación de proteínas a escala industrial.

- 40 La presente invención proporciona proteínas de fusión solubles que comprenden un polipéptido de N-inteína, capaz de formar un complejo de inteína activa, mediante la asociación con una segunda proteína de fusión que comprende un polipéptido de C-inteína. Las proteínas de fusión que comprenden un polipéptido de N-inteína se pueden unir covalentemente a un soporte sólido, para producir una matriz de cromatografía de afinidad, que es adecuada para aplicaciones de bioprocesamiento a gran escala.

- 45 Por consiguiente, en una realización, la presente invención se refiere a una proteína de fusión, que comprende un polipéptido de N-inteína y una contrapartida de solubilización de N-inteína, unidos por un enlace peptídico. En un aspecto particular de esta realización, la contrapartida de solubilización de N-inteína tiene un peso molecular menor que 15 kDa aproximadamente, un valor de índice alifático menor que 60 aproximadamente y un valor de gran promedio de hidropatía menor que -1, y mejora (por ejemplo, aumenta y/o promueve) la solubilidad del polipéptido de N-inteína. En un aspecto adicional de esta realización, la contrapartida de solubilización de la N-inteína comprende la SEQ ID NO: 15. En otro aspecto más de esta realización, el polipéptido de N-inteína es la N-inteína GP41-1 (SEQ ID NO: 1) o una variante de la misma.

En otra realización, la invención se refiere a una matriz de cromatografía de afinidad que comprende una proteína de fusión, que comprende un polipéptido de N-inteína y una contrapartida de solubilización de N-inteína, unida a un soporte sólido. En un aspecto particular de esta realización, el soporte sólido es una resina de cromatografía que

incluye una base de poliviniléter hidrófilo.

Las proteínas de fusión de N-inteína de la presente invención tienen una solubilidad y una actividad catalítica mejoradas y resultan de utilidad como reactivos para realizar la purificación de proteínas a gran escala (por ejemplo, cromatografía de afinidad) y procesos de modificación (por ejemplo, escisión peptídica y reacciones de ligamiento) cuando se asocian con la correspondiente C-inteína.

Breve descripción de los dibujos

El archivo de la patente o solicitud contiene al menos un dibujo ejecutado en color. La Oficina proporcionará copias de esta publicación de patente o solicitud de patente con dibujos en color, previa solicitud y pago de la tarifa necesaria.

La figura 1 es un diagrama esquemático, que representa un método de purificación por afinidad ejemplar de la invención. El método emplea una matriz de cromatografía de afinidad ejemplar de la invención, que comprende una proteína de fusión que tiene un polipéptido de N-inteína fusionado a una contrapartida de solubilización de N-inteína que está unido a un soporte sólido (superficie). Una segunda proteína de fusión que comprende una C-inteína, que es complementaria a la N-inteína en la matriz de cromatografía de afinidad se fusiona con la proteína diana a purificar (proteína de interés) y a cualquier otro elemento requerido para la expresión, tales como las señales de secreción. La figura 1A muestra los diversos componentes antes de la unión de la proteína de fusión C-inteína a la matriz de cromatografía de afinidad de la N-inteína. La figura 1B muestra la proteína de fusión C-inteína unida a la matriz de cromatografía de afinidad de N-inteína, en condiciones apropiadas (por ejemplo, pH, sal, oxidación/reducción) para la asociación de las inteínas. La figura 1C muestra los componentes después de que las exteínas N y C se han escindido de sus proteínas de fusión respectivas, en condiciones apropiadas para la actividad catalítica del complejo de inteínas.

La figura 2A es una gráfica que representa el efecto de la polaridad de fusión sobre la actividad catalítica (tasa de escisión) para tres contrapartidas de solubilización candidatas de la N-inteína (46, 206, 246; véase la tabla 2) que se fusionaron con el término N (SOLP-NINT) o con el término C (NINT-SOLP) de la N-inteína GP41-1.

La figura 2B es un gráfico que representa el efecto de la polaridad de fusión en la expresión de proteínas en la *E. coli* para tres contrapartidas de solubilización candidatas de N-inteína (46, 206, 246) que se fusionaron con el término N (SOLP-NINT) o el término C (NINT-SOLP) de la N-inteína GP41-1.

La figura 3 es un gráfico que representa las tasas de escisión del sustrato y los títulos de expresión solubles para siete contrapartidas de solubilización candidatas (46, 206, 246, 51, 138, 342, 368) que se fusionaron al extremo C-terminal de la N-inteína GP41-1.

La figura 4 es un gráfico que representa la correlación entre las propiedades físicas calculadas de las contrapartidas de solubilización candidatas y la expresión total (título) o soluble (título soluble) en la *E. coli* de las proteínas de fusión que contienen la contrapartida de solubilización fusionada al término C de la N-inteína. mw: peso molecular; pI: punto isoeléctrico; IA: índice alifático; GRAVY: gran promedio de hidropatía [todos ellos por sus siglas en inglés].

La figura 5A es un gráfico que muestra las frecuencias con las que se encuentran aminoácidos particulares en aproximadamente cien homólogos de la GP41-1 en el residuo correspondiente a la cisteína, en la posición 65 de la inteína GP41-1.

La figura 5B es un gráfico que muestra las frecuencias con las que se encuentran los aminoácidos particulares en aproximadamente cien homólogos de GP41-1 en el residuo correspondiente a la cisteína, en la posición 89 de la inteína GP41-1.

La figura 6 es un injerto que representa las actividades catalíticas (tasas de escisión) de las proteínas de fusión de la contrapartida de solubilización 138 con la N-inteína GP41-1 de tipo salvaje, que contiene dos residuos de cisteína de naturales, ubicados centralmente, en las posiciones 65 y 89 o variantes de la N-inteína GP41-1 que contiene sustituciones de aminoácidos para uno o ambos de los residuos de cisteína, en las posiciones 65 y 89.

La figura 7 representa una estructura de solución de RMN para la contrapartida de solubilización 138 (estructura 1RYK del banco de datos de proteínas). La proteína contiene cuatro dominios de hélice alfa, es globular, tiene una larga bobina no estructurada, que forma la conexión al terminal carboxi de la N-inteína (región en el círculo; no se muestra la N-inteína). Las regiones de bucle GKL y GYQ indicadas por el resaltado amarillo se dirigieron para las inserciones de los residuos de la cisteína para crear las nuevas versiones (GCKL (SEQ ID NO: 61), GCYQ (SEQ ID NO: 62) y GCGCYQ (SEQ ID NO: 63)) de la contrapartida de solubilización 138 (138_GKL22GCKL, 138_GYQ48GCGYQ, y 138_GYQ48GCGYQ).

Descripción detallada de la invención

A continuación se proporciona una descripción de las realizaciones ejemplares de la invención.

I. Definiciones

Para que la presente descripción pueda entenderse con mayor facilidad, primero se definen ciertos términos. A lo largo de la descripción detallada se brindan otras definiciones. A menos que se los defina de manera contraria, todos los términos técnicos y científicos usados en el presente documento tienen el mismo significado que entiende comúnmente un experto en la técnica a la que se refiere esta invención.

5 Las frases “biomolécula de interés” y “molécula diana” se usan indistintamente en el presente documento para referirse a una molécula biológica (por ejemplo, una proteína), un material o un ensamblaje macromolecular, que se debe purificar o eliminar de una mezcla (por ejemplo, una mezcla de proteína cruda). Las biomoléculas de interés
ejemplares incluyen, por ejemplo, péptidos y proteínas recombinantes, incluidos los anticuerpos (por ejemplo,
10 anticuerpos monoclonales), vacunas, virus y otros ensamblajes macromoleculares, tales como nanopartículas y partículas similares a virus que pueden incorporar componentes tanto sintéticos como biomoleculares. A modo de ejemplo, las biomoléculas de interés pueden incluir proteínas y ensamblajes biomoleculares (por ejemplo, producidos por tecnología de ADN recombinante), como por ejemplo, hormonas (por ejemplo, insulina, hormona de crecimiento humana, eritropoyetina, interferones, factor estimulante de colonias de granulocitos, activador de plasminógeno tisular), anticuerpos monoclonales (mAb) y derivados de mAb (por ejemplo, mAb bi-específicos, Fab, scFvs, anticuerpos de tiburones y camélidos), productos terapéuticos derivados de andamios (por ejemplo, DARPin, Affibodies, anticalins), enzimas terapéuticas (por ejemplo, galactosidasa alfa A, alfa-L-iduronidasa, N-acetilgalactosamina-4-sulfatasa, glucocerebrosidasa), toxinas (por ejemplo, botulinum, CRM 197, ricina), vacunas recombinantes (por ejemplo, ántrax, difteria, tétanos, neumonía, virus de la hepatitis B, virus del papiloma humano), partículas similares a virus (por ejemplo, hepatitis B, papiloma humano, influenza, parvovirus, virus Norwalk), así
20 como enzimas industriales (por ejemplo, papaína, bromelina, tripsina, proteinasa K, enzima BENZONASE™, enzima DENERASE™, ureasa, pepsina, etc.) y reactivos de diagnóstico (por ejemplo, glucosa y lactato deshidrogenasa, ADN polimerasas, fosfatasa alcalina, peroxidasa de rábano picante, enzimas de restricción, anticuerpos derivados de hibridomas, etc.). En una realización particular, la molécula diana es un anticuerpo (por ejemplo, un anticuerpo monoclonal) contra una diana terapéutica.

25 El término “proteína de fusión” se refiere a una molécula proteica única, natural, sintética, semisintética o recombinante, que comprende la totalidad o una porción de dos o más polipéptidos heterólogos unidos por enlaces peptídicos.

El término “peptídico”, como se usa en el presente documento, se refiere a péptidos y proteínas de más de dos aminoácidos de longitud, que también pueden incorporar moléculas que no sean de aminoácidos (por ejemplo, cromóforos, fármacos, toxinas, agentes de contraste de imágenes, etc.).
30

El término “polipéptido” se refiere a un polímero de aminoácidos, y no a una longitud específica; por lo tanto, los péptidos, oligopéptidos y proteínas se incluyen dentro de la definición de un polipéptido.

La frase “inteína dividida”, como se usa en el presente documento, se refiere a una proteína, ya sea aislada de la naturaleza o creada mediante tecnología de ADN recombinante, que tiene las siguientes propiedades: (1) la proteína
35 se presenta en dos mitades, que interactúan con alta afinidad y selectividad; (2) las dos mitades deben contener todas las secuencias de inteína requeridas para la actividad catalítica y también pueden contener secuencias peptídicas que no sean inteínas adjuntas; (3) la proteína tiene actividad enzimática solo cuando las dos mitades están estrechamente asociadas y (4) la actividad enzimática es la escisión o ligadura peptídica selectiva del sitio, que sirve para separar las secuencias de inteína de las secuencias peptídicas que no son de inteína o para ligar las
40 secuencias peptídicas que no son de inteína en proteínas lineales o circulares contiguas.

La expresión “inteínas complementarias” se usa en el presente documento para referirse a los porciones [SIC: las porciones] de N-inteína y C-inteína de un par de inteína dividido.

El término “N-inteína”, como se usa en el presente documento, se refiere a un polipéptido de inteína que tiene homología con la porción N-terminal de un único polipéptido de inteína, y que se asocia con una C-inteína complementaria para formar una enzima de inteína activa.
45

El término “C-inteína”, como se usa en el presente documento, se refiere a un polipéptido de inteína que tiene homología con la porción C-terminal de un polipéptido de inteína único, y que se asocia con una C-inteína complementaria para formar una enzima de inteína activa.

El término “exteína”, como se usa en el presente documento, se refiere a las secuencias peptídicas del N y C terminales que se fusionan con las N-inteínas y C-inteínas de la naturaleza y se manipulan (por ejemplo, se escinden o ligan) a través de la acción enzimática de la inteína dividida.
50

El término “ligando”, como se usa en el presente documento, se refiere a una molécula que es capaz de una interacción fuerte y selectiva con otra, especialmente cuando está unida a una superficie, tal como una resina de cromatografía. En algunas realizaciones de esta invención, el ligando puede ser una proteína de fusión N-inteína descrita en el presente documento.
55

La expresión “contrapartida de solubilización”, como se usa en el presente documento, se refiere a una proteína que, cuando se fusiona con una N-inteína, potencia (por ejemplo, aumenta, promueve o mantiene) la cantidad de N-

intestina soluble expresada en la *E. coli*, en relación con la cantidad de N-intestina soluble expresada en ausencia de la contrapartida de solubilización. Por ejemplo, en diversas realizaciones, la expresión de la N-intestina como una proteína de fusión con una contrapartida de solubilización puede aumentar la solubilidad de la N-intestina en al menos aproximadamente el 10 % (por ejemplo, en aproximadamente el 20 %, en aproximadamente el 30 %, en aproximadamente el 40 %, en aproximadamente el 50 %, en aproximadamente el 60 %, en aproximadamente el 70 %, en aproximadamente el 80 %, en aproximadamente el 90 % o más, en relación con la solubilidad de la intestina cuando se expresa sin la contrapartida de solubilización.

En una realización, la contrapartida de solubilización E (SEQ ID NO: 25) se fusiona con una N-intestina y la solubilidad de la proteína de fusión resultante se usa para proporcionar un punto de inicio experimental. Esto es de particular utilidad cuando la N-intestina sola no es soluble o estable.

La frase “molécula parental” o “contraparte de tipo salvaje (wt)” [por sus siglas en inglés] o “proteína wt” o “dominio wt”, como se usa en el presente documento, pretende referirse a una proteína correspondiente (por ejemplo, N-intestina, contrapartida de solubilización de la N-intestina), o un dominio de una proteína, en su forma sustancialmente nativa, que por lo general se usa como control en este documento.

La expresión “identidad de secuencia” significa que dos secuencias de nucleótidos o aminoácidos, cuando están alineadas de manera óptima, como por los programas GAP o BESTFIT que usan pesos de brecha predeterminados, comparten al menos el 70 % de identidad de secuencia o al menos el 80 % de identidad de secuencia o al menos 85 % de identidad de secuencia o al menos 90 % de identidad de secuencia o al menos 95 % de identidad de secuencia o más. Para la comparación de secuencias, normalmente una secuencia actúa como una secuencia de referencia (por ejemplo, secuencia parental), con la que se comparan las secuencias de prueba. Cuando se usa un algoritmo de comparación de secuencias, las secuencias de prueba y referencia se ingresan en una computadora, se designan las coordenadas de las subsecuencias, si es necesario, y se designan los parámetros del programa de algoritmos de secuencia. El algoritmo de comparación de secuencias luego calcula el porcentaje de identidad de secuencia para la o las secuencias de prueba en relación con la secuencia de referencia, en función de los parámetros de programa designados.

La alineación óptima de las secuencias para establecer la comparación se puede llevar a cabo, por ejemplo, mediante el algoritmo de homología local de Smith & Waterman, *Adv. Appl. Math.* 2: 482 (1981), por el algoritmo de alineación de homología de Needleman & Wunsch, *J. Mol. Biol.* 48: 443 (1970), por el método de búsqueda de similitud de Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85: 2444 (1988), mediante implementaciones computarizadas de estos algoritmos (GAP, BESTFIT, FASTA y TFASTA en el paquete de *software* de genética de Wisconsin, Genetics Computer Group, 575 Science Dr., Madison, Wis.), o por inspección visual (véase en general Ausubel *et al.*, *Current Protocols in Molecular Biology*). Un ejemplo de algoritmo que es adecuado para determinar el porcentaje de identidad de secuencia y la similitud de secuencia es el algoritmo BLAST, que se describe en Altschul *et al.*, *J. Mol. Biol.* 215: 403 (1990). El *software* para realizar los análisis de BLAST está disponible públicamente, a través del Centro Nacional de Información Biotecnológica (accesible públicamente a través del servidor de Internet de los Institutos Nacionales de la Salud, NCBI). Normalmente, para realizar la comparación de secuencias pueden usarse los parámetros de programa predeterminados, aunque también el posible utilizar parámetros personalizados. Para las secuencias de aminoácidos, el programa BLASTP emplea como valor predeterminado una longitud de palabra (W) de 3, una expectativa (E) de 10 y la matriz de puntuación BLOSUM62 (ver Henikoff y Henikoff, *Proc. Nat'l. Acad. Sci. USA* 89: 10915 (1989)).

El término “cromatografía”, como se usa en el presente documento, se refiere a una técnica de separación dinámica que separa una molécula diana de interés de otras moléculas presentes en la mezcla y le permite aislarse. Por lo general, en un método de cromatografía, una fase móvil (líquido o gas) transporta una muestra que contiene la molécula diana de interés a través de un medio de fase estacionaria (normalmente sólido). Las diferencias en la partición o afinidad a la fase estacionaria separan las diferentes moléculas, mientras que la fase móvil transporta las distintas moléculas en un momento diferente.

La frase “cromatografía de afinidad”, como se usa en este documento, se refiere a un modo de cromatografía en el que una molécula diana a separar se aísla por su interacción con una molécula (por ejemplo, un ligando de cromatografía de afinidad de acuerdo con esta invención, que comprende una N-intestina y un factor de solubilización de N-intestina) que interactúan específicamente con la molécula diana. En una realización, la cromatografía de afinidad implica la adición de una muestra que contiene una molécula diana (por ejemplo, una inmunoglobulina o una proteína que contiene Fc) a un soporte sólido que lleva un ligando basado en N-intestina, como se describe en este documento.

La expresión “matriz de afinidad” o “matriz de cromatografía de afinidad”, como se usa indistintamente en este documento, se refiere a un soporte cromatográfico sobre el cual se une un ligando de cromatografía de afinidad (por ejemplo, una proteína de fusión N-intestina o un dominio de la misma). El ligando es capaz de unirse a una molécula de interés, a través de la interacción de afinidad (por ejemplo, una proteína de fusión C-intestina complementaria) que se debe purificar o eliminar de una mezcla.

El término “inmunoglobulina”, “Ig” o “anticuerpo” (utilizados como sinónimos en el presente documento) se refiere a

una proteína que tiene una estructura de cadena básica de cuatro polipéptidos, que consiste en dos cadenas pesadas y dos ligeras; dichas cadenas se estabilizan, por ejemplo, por enlaces disulfuro entre cadenas, que tiene la capacidad de unirse específicamente al antígeno. La frase “inmunoglobulina de cadena única” o “anticuerpo de cadena única” (utilizadas indistintamente en el presente documento) se refiere a una proteína que tiene una estructura de cadena de dos polipéptidos, que consiste en una cadena pesada y una cadena ligera, donde dichas cadenas se estabilizan, por ejemplo, por uniones peptídicas de cadenas, que tienen la capacidad de unirse específicamente al antígeno. El término “dominio” se refiere a una región globular de un polipéptido de cadena pesada o ligera, que comprende bucles peptídicos (por ejemplo, que comprenden 3 a 4 bucles peptídicos) estabilizados, por ejemplo, mediante una lámina plisada en β y/o un enlace disulfuro intracadena. Asimismo, los dominios se mencionan aquí como “constantes” o “variables”, en función de la falta relativa de variación de secuencia dentro de los dominios de varios miembros de la clase en el caso de un dominio “constante”, o la variación significativa dentro de los dominios de varios miembros de la clase en el caso de un dominio “variable”. Los “dominios” de anticuerpos o polipéptidos a menudo se denominan indistintamente en la técnica como “regiones” de anticuerpos o polipéptidos. Los dominios “constantes” de las cadenas ligeras de los anticuerpos se denominan indistintamente como “regiones constantes de cadena ligera”, “dominios constantes de cadena ligera”, regiones “CL” o dominios “CL”. Los dominios “constantes” de las cadenas pesadas de anticuerpos se denominan indistintamente como “regiones constantes de cadena pesada”, “dominios constantes de cadena pesada”, regiones “CH” o dominios “CH”. Los dominios “variables” de las cadenas ligeras de anticuerpos se denominan indistintamente como “regiones variables de cadena ligera”, “dominios variables de cadena ligera”, regiones “VL” o dominios “VL”. Los dominios “variables” de las cadenas pesadas de anticuerpos se denominan indistintamente como “regiones variables de cadena pesada”, “dominios variables de cadena pesada”, regiones “VH” o dominios “VH”.

Los “anticuerpos” o “inmunoglobulinas” pueden ser monoclonales o policlonales y pueden presentarse en forma monomérica o polimérica, por ejemplo, anticuerpos IgM que existen en forma pentamérica y/o anticuerpos de IgA, que se presentan en forma monomérica, dimerica o multimérica. El término “fragmento” se refiere a una parte o porción de un anticuerpo o a una cadena de anticuerpo que comprende menos residuos de aminoácidos que un anticuerpo intacto o completo o una cadena del anticuerpo. Los fragmentos pueden obtenerse a través del tratamiento químico o enzimático de un anticuerpo intacto o completo o una cadena de anticuerpos. Los fragmentos también se pueden obtener por medios recombinantes. Los fragmentos ejemplares incluyen fragmentos Fab, Fab', F(ab')₂, Fc y/o Fv.

Expresiones tales como “polinucleótido” y “molécula de ácido nucleico”, usadas indistintamente en el presente documento, se refieren a formas poliméricas de nucleótidos de cualquier longitud, ya sean ribonucleótidos o desoxirribonucleótidos. Estos términos incluyen un ADN de cadena simple, doble o triple, ADN genómico, ADNc, ARN, ADN-ARN híbrido, o un polímero que comprende bases de purina y pirimidina, u otras bases de nucleótidos naturales, modificadas química o bioquímicamente, no naturales o derivadas. El esqueleto del polinucleótido puede comprender azúcares y grupos fosfato (como se puede encontrar típicamente en el ARN o ADN), o grupos fosfato o de azúcar modificados o sustituidos. Además, se puede obtener un polinucleótido bicatenario a partir del producto polinucleotídico monocatenario de síntesis química, ya sea sintetizando la cadena complementaria y recociendo las cadenas en condiciones apropiadas, o sintetizando la cadena complementaria *de novo* utilizando una ADN polimerasa con un cebador apropiado. Una molécula de ácido nucleico puede adoptar muchas formas diferentes, por ejemplo, un gen o fragmento de gen, uno o más exones, uno o más intrones, ARNm, ADNc, polinucleótidos recombinantes, polinucleótidos ramificados, plásmidos, vectores, ADN aislado de cualquier secuencia, ARN aislado de cualquier secuencia, sondas de ácidos nucleicos, y cebadores. Un polinucleótido puede comprender nucleótidos modificados, tales como nucleótidos metilados y análogos de nucleótidos, uracilo, otros azúcares y grupos de enlace, tales como fluororibosa y tioato y ramas de nucleótidos. Conforme se usa en este documento, “ADN” o “secuencia de nucleótidos” incluye no solo las bases A, T, C y G, sino que también incluye cualquiera de sus análogos o formas modificadas de estas bases, tales como nucleótidos metilados, modificaciones internucleótidas tales como enlaces no cargados y tioatos, el uso de análogos de azúcar y estructuras de esqueleto modificadas y/o alternativas, tales como poliamidas. En una realización particular, una molécula de ácido nucleico comprende una secuencia de nucleótidos que codifica una proteína de fusión N-inteína o una variante de la misma, como se describe en el presente documento.

II. Proteínas de fusión basadas en inteína

Las inteínas son una clase de enzimas autocatalíticas descubiertas en 1990, que contienen actividades tanto de proteasa como de ligasa, que funcionan en el ciclo de vida natural de estas moléculas. Se ha demostrado que los reactivos de inteína tienen utilidad para la escisión, ligadura y circularización de sustratos peptídicos. En 1998, se descubrió una nueva clase de inteínas denominadas “inteínas divididas”, en las que la enzima se encuentra naturalmente en dos partes, denominadas N-inteína y C-inteína (medias inteínas complementarias). Mientras que las inteínas divididas se han encontrado en una amplia variedad de procariotas inferiores (Zettler J., *et al.*, *FEBS Letters*, 553:909-914 (2009); Dassa B., *et al.*, *Biochemistry*, 46:322-330 (2007); Choi J., *et al.*, *J Mol Biol.* 556: 1093-1106 (2006); Caspi, *et al.*, *Mol Microbiol.* 50: 1569-1577 (2003); Liu X. and Yang J., *J Biol Chem.*, 275:26315-26318 (2003); Wu H., *et al.*, *Proc Natl Acad Sci USA.* 5:9226-9231 (1998)), no se han identificado inteínas divididas en eucariotas (véase la base de datos de inteínas mantenida por New England Biolabs (<http://tools.neb.com/inbase/list.php>)). Recientemente se han caracterizado dos inteínas divididas que son extremadamente rápidas y bastante promiscuas con respecto a las secuencias de inteínas adyacentes. Una clase

es la inteína Npu DnaE (Iwai I., *et al.*, *FEBS Letters* 550: 1853-1858 (2006); Zettler J., *et al.*, *FEBS Letters*, 553:909-914 (2009)) y la otra, las inteínas GP41 divididas, identificadas a partir de datos metagenómicos (Carvajal-Vallejos P., *et al.*, *J. Biol. Chem.* 287: 28686-28696 (2012); Publicación Internacional PCT No. WO2013045632).

Las N- y C-inteínas, con exteínas adjuntas (las dos medias proteínas que se unirán mediante la actividad de la inteína), se asocian de forma extremadamente específica y estrecha a través de múltiples interacciones entre dominios, para formar la enzima inteína activa (Shah NH, *et al.*, *J. Amer. Chem. Soc.* 135: 18673-18681; Dassa B., *et al.*, *Nucl. Acids Res.*, 37: 2560-2573 (2009)). Además de las actividades de ligasa y proteasa presentes en la primera clase de inteínas, las inteínas divididas tienen utilidad en las separaciones de afinidad, debido a la interacción estrecha y selectiva de los dominios de las N- y C-inteínas.

La presente invención se basa, en parte, en el descubrimiento de que la expresión de polipéptidos de inteína como proteínas de fusión con ciertas proteínas heterólogas, referidas aquí como contrapartidas de solubilización, aumenta la solubilidad de la inteína, por lo que la inteína es adecuada como reactivo para la cromatografía de afinidad y otras aplicaciones de purificación y modificación de proteínas que se pueden llevar a la práctica tanto a pequeña como a gran escala. De un modo más específico, la invención proporciona proteínas de fusión altamente solubles que comprenden un polipéptido de N-inteína y una contrapartida de solubilización de N-inteína, capaz de formar un complejo de inteína activa mediante la asociación con un polipéptido de C-inteína complementario.

Por consiguiente, en una realización, la presente invención se refiere a una proteína de fusión que comprende un polipéptido de N-inteína y una contrapartida de solubilización de N-inteína, como se describe en la reivindicación 1. En la técnica, se conoce una variedad de polipéptidos de N-inteína. Las N-inteínas ejemplares incluyen las N-inteínas que se muestran en la tabla 1 y otras descritas en otra parte en este documento. Las N-inteínas descritas en el presente documento, y otras N-inteínas conocidas en la técnica, así como las variantes de dichas N-inteínas que tienen aproximadamente, al menos el 75 % de identidad de secuencia (por ejemplo, aproximadamente, al menos el 80 %, aproximadamente, al menos el 90 %, a aproximadamente, al menos el 95 %, aproximadamente, al menos el 96 %, aproximadamente, al menos el 97 %, aproximadamente, al menos el 98 %, aproximadamente, al menos el 99 % de identidad de secuencia) respecto de una N-inteína de tipo salvaje, pueden incluirse en las proteínas de fusión descritas en el presente documento.

El primer aminoácido en un dominio N-terminal de la inteína suele estar muy conservado y puede ser importante para la reacción de empalme de proteínas. Sin embargo, en algunas realizaciones, el primer aminoácido en un dominio N-terminal de la inteína (por ejemplo, una cisteína, una serina) se puede sustituir con un aminoácido (por ejemplo, un aminoácido distinto de la cisteína o la serina) que evite o reduzca la escisión entre la inteína y un polipéptido heterólogo. En una realización particular, el primer aminoácido en un dominio N-terminal de inteína está sustituido con una alanina.

En una realización particular, las proteínas de fusión N-inteína descritas en este documento comprenden la N-inteína GP41-1 de tipo salvaje (SEQ ID NO: 1 o SEQ ID NO: 29) o una variante de la misma. La variante adecuada de N-inteínas GP41-1 puede tener aproximadamente, al menos el 75 % de identidad de secuencia (por ejemplo, aproximadamente, al menos el 80 %, aproximadamente, al menos el 90 %, aproximadamente, al menos el 95 %, aproximadamente, al menos el 96 %, aproximadamente, al menos el 97 %, aproximadamente, al menos el 98 %, aproximadamente, al menos el 99 % de identidad de secuencia) respecto de la N-inteína GP41-1 de tipo salvaje (SEQ ID NO: 1). Los ejemplos particulares de las variantes de las N-inteínas GP41-1 para su inclusión en las proteínas de fusión de la invención incluyen las variantes de GP41-1 que se asignan a las SEQ ID NO: 2-8 en este documento. En ciertas realizaciones, la variante de N-inteína GP41-1 carece de residuos de cisteína. En una realización particular, uno o más residuos de cisteína que ocurren naturalmente en la N-inteína GP41-1 se eliminan. En otra realización, uno o más residuos de cisteína que ocurren naturalmente en la N-inteína GP41-1 (posiciones 7, 65 y 89 de la SEQ ID NO: 1) están sustituidos con otro residuo de aminoácido (por ejemplo, treonina, lisina o asparagina). En una realización, el residuo de cisteína que se produce naturalmente en la N-inteína GP41-1 en la posición 65 de la SEQ ID NO: 1 está sustituido con otro residuo de aminoácido (por ejemplo, serina, treonina). En una realización particular, el residuo de cisteína en la posición 65 de la SEQ ID NO: 1 está sustituido con treonina. En otra realización más, el residuo de cisteína que se produce naturalmente en la N-inteína GP41-1 en la posición 89 de la SEQ ID NO: 1 está sustituido con otro residuo de aminoácido (por ejemplo, metionina, tirosina). En una realización particular, el residuo de cisteína en la posición 89 de la SEQ ID NO: 1 está sustituido con metionina. En algunas realizaciones, la variante GP41-1 es la variante de N-inteína GP41-1 NINTAA_TM (SEQ ID NO: 6) o la variante de N-inteína GP41-1 NINTAA_TK (SEQ ID NO: 8).

En algunas realizaciones, la variante de N-inteína GP41-1 que carece de algunos o todos los residuos de cisteína es al menos 2 veces, al menos 3 veces, al menos 4 veces, al menos 5 veces, al menos 6 veces, al menos 7 veces, al menos 8 veces, al menos 9 veces o al menos 10 veces más activa que la N-inteína GP41-1 natural en las reacciones de ligadura o escisión. La actividad de inteína, ya sea por escisión o por ligadura, puede analizarse, por lo general, utilizando electroforesis en gel SDS en condiciones reductoras (por ejemplo, Zettler J., Schütz V., Mootz HD, *FEBS Letters* 583: 909-914, 2009; Aranko AS, Züger S, Buchinger E, Iwai H, *PLoS ONE* 4: e5185, 2009). Para ser breves, las reacciones de inteína, por lo general en el transcurso del tiempo, se detienen mediante la adición de

5 un tampón de carga de gel SDS que contenga un agente reductor (por ejemplo, ditioneitol o β-mercaptoetanol), las muestras se hierven para desnaturalizarlas por completo y luego se cargan en un gel de poliacrilamida que contiene SDS junto con marcadores de tamaño de proteína apropiados. Una vez completada la electroforesis, las proteínas de la reacción se separaron de acuerdo con sus pesos moleculares y se pueden visualizar mediante tinción con tintes tradicionales o fluorescentes. Las cantidades de los diversos productos intermedios y productos en función del tiempo se pueden cuantificar mediante densitometría e intensidades en función del tiempo convertidas en tasas enzimáticas (kobs), mediante la aplicación de un programa de ajuste de curvas.

Tabla 1. Ejemplares de las divisiones divididas GP41-1 y sus variantes

SEQ ID NO:	Nombre	Secuencia
1	N-inteína GP41-1	mtrsgyCLDLKTQVQTPQGMKEISNIQVGDVLVS NTGYNEVLNVFPKSKKKSYKITLEDGKEIICSEE HLFPTQTGEMNISGGLKEGMCLYVKEgg
2	Variante de N-inteína GP41-1 NINTΔA_ CC	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLVS NTGYNEVLNVFPKSKKKSYKITLEDGKEIICSEE HLFPTQTGEMNISGGLKEGMCLYVKEgg
3	Variante de N-inteína GP41-1 NINTΔA_ AC	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLVS NTGYNEVLNVFPKSKKKSYKITLEDGKEIASE EHLFPTQTGEMNISGGLKEGMCLYVKEgg
4	Variante de N-inteína GP41-1 NINTΔA_ CK	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLVS NTGYNEVLNVFPKSKKKSYKITLEDGKEIICSEE HLFPTQTGEMNISGGLKEGMKLYVKEgg
5	Variante de N-inteína GP41-1 NINTΔA_ AM	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLVS NTGYNEVLNVFPKSKKKSYKITLEDGKEIASE EHLFPTQTGEMNISGGLKEGMMLYVKEgg
6	Variante de N-inteína GP41-1 NINTΔA_ TM	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLVS NTGYNEVLNVFPKSKKKSYKITLEDGKEITSEE HLFPTQTGEMNISGGLKEGMMLYVKEgg

SEQ ID NO:	Nombre	Secuencia
7	Variante de N-inteína GP41-1 NINTΔA_AK	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLVS NTGYNEVLNVFPKSKKKSYKITLEDGKEIASE EHLFPTQTGEMNISGGLKEGMKLYVKEgg
8	Variante de N-inteína GP41-1 NINTΔA_TK	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLVS NTGYNEVLNVFPKSKKKSYKITLEDGKEIITSEE HLFPTQTGEMNISGGLKEGMKLYVKEgg
9	C-inteína GP41-1 (CINT)	MGKNSMMLKKILKIEELDERELIDIEVSGNHLF YANDILTHN
10	Proteína de fusión de C-inteína-tiorredoxina (CINT_TRX) (la porción de tiorredoxina está subrayada)	MGKNSMMLKKILKIEELDERELIDIEVSGNHLF YANDILTHNMSDKIIHLTDDSFDTDLKADGAI <u>LVDFWAEWCGPCKMIAPILDEIADEYQGKLT</u> <u>AKLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAA</u> <u>TKVGALSKGQLKEFLDANLAHHHHHH</u>

Para las SEQ ID Nos.: 1-8, las secuencias que no son inteínas se indican con texto en minúsculas y las secuencias de inteínas se indican con mayúscula para las SEQ ID NO: 1-8.

Típicamente, los polipéptidos de N-inteína tienen poca solubilidad cuando se expresan en sistemas de expresión comunes, tales como la *E. coli*. La presente invención evita este problema, por ejemplo, expresando la N-inteína como una proteína de fusión con una contrapartida de solubilización de N-inteína, lo que aumenta la solubilidad de la N-inteína (por ejemplo, cuando se expresa en la *E. coli*). Preferiblemente, la contrapartida de solubilización de la N-inteína aumenta la solubilidad del polipéptido de N-inteína, de manera que menos de aproximadamente el 25 % en masa de la proteína de fusión resultante esté presente en los cuerpos de inclusión después de la producción en el sistema de expresión (por ejemplo, *E. coli*). El experto en la técnica puede determinar fácilmente el porcentaje en masa de una proteína expresada que está presente en los cuerpos de inclusión, después de la producción en un sistema de expresión utilizando técnicas y reactivos estándar.

Una persona con los conocimientos comunes en la técnica puede seleccionar fácilmente posibles las contrapartidas de solubilización que pueden aumentar la solubilidad de una N-inteína dada, usando procedimientos conocidos en este ámbito y descritos en el presente documento. Por ejemplo, la probabilidad de generar un producto soluble tras la sobreexpresión en un sistema de expresión (por ejemplo, *E. coli*) se puede calcular utilizando el algoritmo de Wilkinson y Harrison (Wilkinson DL y Harrison RG, *Bio/Technology*, 9: 443, 1991). La predicción de si la proteína contiene una señal de secreción funcional se puede realizar utilizando el algoritmo SignalP 4.1, disponible en el Centro de Análisis de Secuencias Biológicas de la Universidad Técnica de Dinamarca (<http://genome.cbs.dtu.dk/services/SignalP/>). Véanse también los métodos descritos en los ejemplos 1-3 detallados en el presente documento. En última instancia, las contrapartidas de solubilización que proporcionan una mejora óptima de la solubilidad, al tiempo que permiten la actividad catalítica de la inteína máxima deben seleccionarse de las contrapartidas de solubilización candidatas a través de la selección experimental.

Las contrapartidas de solubilización de la N-inteína que tienen ciertas propiedades físicas son particularmente adecuadas para su inclusión en las proteínas de fusión de la invención. Dichas propiedades físicas incluyen un peso

molecular inferior a 15 kDa aproximadamente, un valor de índice alifático (AI) inferior a 60 aproximadamente, y un valor de GRAVY que es inferior a -1. Una persona con los conocimientos comunes en esta técnica puede determinar cada una de estas propiedades para una contrapartida de solubilización dada, usando ensayos y técnicas estándar, por ejemplo, utilizando la herramienta en línea ProtParam (<http://web.expasy.org/tools/protparam/>) que es parte del conjunto de herramientas de bioinformática SwissProt ExPASy.

El gran promedio de hidropaticidad (GRAVY) (Kyte J y Doolittle RF., *J. Mol. Biol.* 157: 105, 1982) de una secuencia de polipéptido lineal se calcula como la suma de los valores de hidropatía de todos los aminoácidos, dividida Por el número de residuos en la secuencia. El aumento de la puntuación positiva indica mayor hidrofobicidad. El cálculo se basa en la escala de Kyte-Doolittle. GRAVY es un método simple para mostrar el carácter hidropático de una proteína.

Alanina	1,8	Leucina	3,8
Arginina	-4,5	Lisina	-3,9
Asparagina	-3,5	Metionina	1,9
Ácido aspártico	-3,5	Fenilalanina	2,8
Cisteína	2,5	Prolina	-1,6
Glutamina	-3,5	Serina	-0,8
Ácido glutámico	-3,5	Treonina	-0,7
Glicina	-0,4	Triptófano	-0,9
Histidina	-3,2	Tirosina	-1,3
Isoleucina	4,5	Valina	4,2

En diversas realizaciones, las proteínas de fusión de N-inteína descritas en el presente documento tienen un valor GRAVY que es menor que -1.

El índice alifático (Ikai, AJ., *J. Biochem.* 88: 1895, 1980) de una proteína se define como el volumen relativo ocupado por las cadenas laterales alifáticas (alanina, valina, isoleucina y leucina). Puede considerarse como un factor positivo para el aumento de la termoestabilidad de las proteínas globulares. El índice alifático de una proteína se calcula de acuerdo con la siguiente fórmula: índice alifático = $X(\text{Ala}) + a * X(\text{Val}) + b * (X(\text{Ile}) + X(\text{Leu}))$. * Los coeficientes a y b son el volumen relativo de la cadena lateral de valina ($a = 2,9$) y de las cadenas laterales de Leu/Ile ($b = 3.9$) a la cadena lateral de alanina. La probabilidad de generar un producto soluble tras la sobreexpresión en la *E. coli* también se puede calcular utilizando el algoritmo de Wilkinson y Harrison (Wilkinson DL y Harrison RG., *Bio/Technology*, **9**: 443, 1991). Otros algoritmos disponibles no necesariamente dan resultados similares. En diversas realizaciones, las proteínas de fusión de N-inteína descritas en este documento tienen un valor de índice alifático (AI) menor que aproximadamente 60, y un valor GRAVY que es menor que -1.

Con preferencia, la contrapartida de solubilización de N-inteína tiene un peso molecular inferior a 15 kDa aproximadamente, un valor de índice alifático inferior a 60 °C aproximadamente.

Los ejemplos de contrapartidas de solubilización de N-inteína particulares se describen en la tabla 2.

Tabla 2. Contrapartidas ejemplares de solubilización de N-inteína

SEQ ID NO:	Nombre	GID	Secuencia
11	Contrapartida de solubilización 46: profago de Qin; proteína inducida por choque frío	170081219	MREYPNGEKTHLTVMAAGFPSL TGDHKVIYVAADRHTVSEEILE AAIRLLS

SEQ ID NO:	Nombre	GID	Secuencia
12	Contrapartida de solubilización 206: proteína hipotética ECDH10B_1576 [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	170081120	MSHLDEVIARVDAAIEESVIAH MNEILLALSDDAELSREDRYTQ QQRLRTAIAHHGRKHKEDMEA RHEQLTKGGTIL
13	Contrapartida de solubilización 246: proteína hipotética ECDH10B_1388 [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	170080950	MNKETQPIDRETLLEANKIIRE HEDTLAGIEATGVTQRNGVLVF TGDYFLDEQGLPTAKSTAVFNM FKHLAHLVSEKYHLVD
14	Contrapartida de solubilización 51: proteína hipotética ECDH10B_0422 [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	170080051	MSLENAPDDVKLAVDLIVLLEE NQIPASTVLRALDIVKRDEYKKL TRDDEAEK
15	Contrapartida de solubilización 138: proteína putativa de respuesta al estrés [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	170083502	MNKDEAGGNWKQFKGKVKEQ WGKLTDDDMTHIEGKRDQLVG KIQERYGYQKDQAEKEVVDWE TRNEYRW
16	Contrapartida de solubilización 138 GKL22GCKL: proteína putativa de respuesta al estrés [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	NA	MNKDEAGGNWKQFKGKVKEQ WGCKLTDDDMTHIEGKRDQLV GKIQERYGYQKDQAEKEVVDW ETRNEYRW
17	Contrapartida de solubilización 138 GYQ48GCGY: proteína putativa de respuesta al estrés [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	NA	MNKDEAGGNWKQFKGKVKEQ WGKLTDDDMTHIEGKRDQLVG KIQERYGCGYQKDQAEKEVVDW ETRNEYRW
18	Contrapartida de solubilización 138 GYQ48GCGY: proteína putativa de respuesta al estrés [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	NA	

SEQ ID NO:	Nombre	GID	Secuencia
			MNKDEAGGNWKQFKGKVKEQ WGKLTDDDMTIIIEGKRDQLVG KIQERYGCGYQKDQAEKEVVD WETRNEYRW
19	Contrapartida de solubilización 342: proteína hipotética ECDH10B_2487 [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	170081941	MIAEFESRILALIDGMVDHASDD ELFASGYLRGHLTLAIAELES GDHSAQAVHTTVS QSLEKAIGAG ELSPRDQALVTDMWENLFQQA SQQ
20	Contrapartida de solubilización 368: proteína putativa de modulación sigma(54) [<i>Escherichia coli</i> cepa K-12, subcepa DH10B]	170082737	MQLNITGNNVEITEALREFVTA KFAKLEQYFDRINQVYVVLKVE
			KVTHTSDATLHVNGGEIHASAE GQDMYAAIDGLIDKLARQLTKH KDKLKQH
21	Contrapartida de solubilización A: componente EspA del sistema de secreción tipo 3	9626250	MDTSNATSVVNVSASSSTSTIYD LGNMSKDEVVKLFEELGVFQA AILMFSYMYQAQSNLSIAKFAD MNEASKASTTAQKMANLVDAK IADVQSSTDKNNAKLPQDVID YINDPRNDISVTGISDLSGDLSA GDLQTVKAAISAKANNLTTVVN NSQLEIQQMSNTLNLLTSARSD VQSLQYRTISAISLGK
22	Contrapartida de solubilización B: proteasa fh8 de <i>Fasciola hepatica</i> [duela del hígado] (HiTag)	147611	MPSVEVEKLLHVLDRNGDGKV SAEELKAFADDSKYPLDSNLIK AFIKEHDKNKDGKLDLKVLSIL SS
23	Contrapartida de solubilización C: proteasa fh8 de <i>Fasciola hepatica</i> (HiTag)	387618410	MPSVEVEKLLH

SEQ ID NO:	Nombre	GID	Secuencia
24	Contrapartida de solubilización D: glutatión S-transferasa	251787291	MGQLIDGVWHDTWYDTKSTGG KFQRSASAFRNWLTADGAPGPT GKGGFAAEKDRYHLYVSLACP WAHRTLIMRKLKGLEPFISVSV VNPLMLENGWTFDDSFPGATG DTLYQHEFLYQLYLHADPHYSG RVTVPVLWDKKNHTIVSNESAE IIRMFNTAFDALGAKAGDYYP ALQPKIDELNGWIYDTVNNGVY KAGFATSQQAYDEAVAKVFESL ARLEQILGQHRYLTGNQLTEADI RLWTTLVRFDPVYVTHFKCDK HRISDYLNLYGFLRDIYQMPGIA ETVNFHDHIRNHYFRSHKTINPTG IISIGPWQDLDEPHGRDVRFG
25	Contrapartida de solubilización E: proteína de estabilización de cabeza de ADN [fago lambda de la enterobacteria]	410480759	MASWSHPQFEKASKETFTHYQP QGNSDPAHTATAPGGLSAKAPA MTPLMLDTSSRKLVAWDGTTD GAAVGILAVAADQTSTLTFYK SGTFRYEDVLWPEAASDETKKR TAFAGTAISIV
26	Contrapartida de solubilización F: proteína que se une a la maltosa	218465276	MKIKTGARILALSALTMMFSA SALAKIEEGKLVIWINGDKGYN GLAEVGKKFEKDTGIKVTVEHP DKLEEKFPQVAATGDGPDIIFW AHDRFGGYAQSGLLAEITPDKA FQDKLYPFTWDAVRYNGKLIAY PIAVEALSLIYNKDLLPNPPKTW EEIPALDKELKAKGKSALMFNL QEPYFTWPLIAADGGYAFKYEN
			GKYDIKDVGVNDNAGAKAGLTF LVDLIKNKHMNADTDYSIAEAA FNKGETAMTINGPWAWSNIDTS KVNYGVTVLPTFKGQPSKPFVG VLSAGINAASPNKELAKEFLEN YLLTDEGLEAVNKDKPLGAVAL KSYEEELAKDPRIAATMENAQK GEIMPNIQMSAFWYAVRTAVI NAASGRQTVDEALKDAQTRITK

SEQ ID NO:	Nombre	GID	Secuencia
27	Contrapartida de solubilización G: tiorredoxina	218465276	MSDKIIHLTDDSFDTDVLKADG AILVDFWAEWCGPCKMIAPILD EIADEYQGKLTVAKLNIDQNPG TAPKYGIRGIPTLLLFKNGEVAA TKVGALSKGQLKEFLDANLA
28	Contrapartida de solubilización H: factor de terminación de transcripción NusA	387509083	MNKEILAVVEAVSNEKALPREK IFEALESALATATKKKYEQIDV RVQIDRKSGDFDTFRRWL VVDE VTQPTKEITLEAARYEDES LNLG DYVEDQIESVTFDRITTQTAKQV IVQKVREAERAMVVDQFREHE GEITGVVKKVNRDNISLDLGNN AEAVILREDMLPRENFRPGDRV RGVLYSVRPEARQAQLFVTRSK PEMLIELFRIEVP EIGEEVIEIKAA ARDPGSRAKIAVKTNDKRIDPV GACVGMRGARVQAVSTELGGE RIDIVLWDDNPAQFVINAMAPA DVASIVVDEDKHTMDIAVEAGN LAQAIGRNGQNVR LASQLSGWE LNVMTVDDLQAKHQAEAHAAI DTFTKYLDIDEDFATVLVEEGFS TLEELAYVPMKELLEIEGLDEPT VEALRERAKNALATIAQAQEEES LGDNKPADDLLNLEGVDRDLAF KLAARGVCTLEDLAEQGIDDLA DIEGLTDEKAGALIMAARNICW FGDEA

“GID” se refiere a la ID de GenBank (<http://www.ncbi.nlm.nih.gov/genbank>). Las contrapartidas de solubilización A-H son contrapartidas de solubilización conocidas, muchas de los cuales se han incorporado a sistemas de fusión disponibles en el mercado para aumentar el rendimiento y la solubilidad de las proteínas recombinantes producidas en la *E. coli*. Los aminoácidos que han sido sustituidos o insertados artificialmente en la secuencia principal se resaltan en negrita y están subrayados.

En una realización particular, la contrapartida de solubilización de N-inteína es, o comprende, toda o una parte de la contrapartida de solubilización 138 (SEQ ID NO: 15), o una variante de la misma (por ejemplo, la contrapartida de solubilización 138 GKL22GCKL (SEQ ID NO: 16); contrapartida de solubilización 138 GYQ48GCGY (SEQ ID NO: 17); contrapartida de solubilización 138 GYQ48GCGY (SEQ ID NO: 18)).

Los métodos para preparar proteínas de fusión, o quiméricas, son bien conocidos en la técnica, incluidas, entre otras, técnicas estándar de ADN recombinante. Por ejemplo, los fragmentos de ADN que codifican diferentes secuencias de proteínas (por ejemplo, una N-inteína y una contrapartida de solubilización de N-inteína; una C-inteína y una molécula diana) se unen entre sí en el marco, de acuerdo con las técnicas convencionales. En otra realización, el gen de fusión puede sintetizarse mediante técnicas convencionales, que incluyen sintetizadores de

ADN automatizados. Alternativamente, la amplificación por PCR de los fragmentos de ácido nucleico puede llevarse a cabo utilizando cebadores de anclaje que dan lugar a salientes complementarios entre dos fragmentos de ácido nucleico consecutivos, que posteriormente se pueden recocer y volver a amplificar para generar una secuencia de ácido nucleico quimérica (véase Ausubel *et al.*, *Current Protocols in Molecular Biology*, 1992). Además, hay muchos vectores de expresión disponibles comercialmente que ya codifican un resto de fusión (por ejemplo, un resto GST, un resto Fc).

Con preferencia, la proteína de fusión se expresa a partir de un ácido nucleico codificante en células huésped procarióticas o eucariotas transfectadas o establemente transfectadas o estables o en organismos no humanos. Las células hospedadoras comunes o los organismos no humanos para la expresión de proteínas recombinantes incluyen, por ejemplo, *Escherichia coli*, *Corynebacterium glutamicum*, *Pseudomonas fluorescens*, *Lactococcus lactis*, *Pichia pastoris*, *Saccharomyces cerevisiae*, *Zea mize*, *Nicotinia tabacum*, *Daucus carota*, células SF9, células CHO (por ejemplo, células CHO DG44, células CHO DXB11), células NS0, células HEK 293 y animales no humanos completos, como vacas y cabras. En una realización, la proteína de fusión N-inteína se expresa en *E. coli*. Luego, la proteína de fusión N-inteína expresada se puede purificar, separándola de las proteínas celulares contaminantes utilizando métodos convencionales de separación y cromatografía, como la clarificación por filtración profunda, la purificación por anión y la cromatografía de intercambio catiónico y la concentración por ultrafiltración.

La proteína heteróloga (por ejemplo, la contrapartida de solubilización de N-inteína, molécula diana) puede fusionarse a cualquier extremo del polipéptido de inteína. En una realización, una contrapartida de solubilización de N-inteína se une al extremo N-terminal de un polipéptido de N-inteína. En otra realización, una contrapartida de solubilización de N-inteína se une al término C de un polipéptido de N-inteína.

En algunas realizaciones, el polipéptido de inteína (por ejemplo, N-inteína, C-inteína) y la proteína heteróloga (por ejemplo, la contrapartida de solubilización de N-inteína, molécula diana) se unen directamente a través de un enlace peptídico. En otras realizaciones, la proteína de fusión incluye una molécula espaciadora o enlazadora entre el polipéptido de inteína (por ejemplo, N-inteína, C-inteína) y la proteína heteróloga (por ejemplo, la contrapartida de solubilización de N-inteína, molécula diana). Las moléculas espaciadoras/enlazadoras adecuadas son conocidas en la técnica.

En las proteínas de fusión descritas en el presente documento, el dominio N-terminal de la inteína se puede fusionar directamente (por ejemplo, a través de un enlace peptídico) o indirectamente (por ejemplo, a través de una secuencia de aminoácidos de enlace) a un polipéptido heterólogo. Por lo tanto, en algunas realizaciones, un polipéptido heterólogo se fusiona directa o indirectamente al extremo N de un dominio N-terminal de la inteína. En ciertas realizaciones, 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, He, Gly, Glu y Pro.

En algunas realizaciones, la proteína de fusión comprende un enlace entre el polipéptido heterólogo y la secuencia de inteína. Por ejemplo, la proteína de fusión puede comprender un conector entre el término C de la proteína heteróloga y el término N del dominio N-terminal de la inteína. El enlace puede tener, por ejemplo, una longitud de entre alrededor de 1 y aproximadamente 10 aminoácidos. En algunas realizaciones, el enlace puede tener una longitud de entre alrededor de 1 y aproximadamente 5 aminoácidos. Por ejemplo, el enlace puede contener 1, 2, 3, 4 o 5 aminoácidos. En algunas realizaciones, el último aminoácido del enlace que contacta con el polipéptido heterólogo y el extremo N del dominio N-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 y Pro.

En algunas realizaciones, el enlace puede comprender una secuencia de exteína. En algunas realizaciones, el enlace puede comprender una secuencia de exteína nativa. En algunas realizaciones, la exteína comprende una secuencia seleccionada del grupo que consiste en las SEC ID NOS: 4, 8, 13, 17, 21, 25, 35 y 39 del documento de patente WO201345632. En algunas realizaciones, un enlace que comprende aminoácidos de una exteína comprende, por ejemplo, el primero (es decir, N-terminal) de aproximadamente 1 a aproximadamente 5 aminoácidos de una secuencia seleccionada del grupo que consiste en las SEQ ID NO: 4, 8, 13, 17, 21, 25, 35 y 39. En algunas realizaciones, el enlace comprende aproximadamente 1, 2, 3, 4 o 5 aminoácidos de una secuencia seleccionada del grupo que consiste en las SEQ ID NO: 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 en la naturaleza (por ejemplo, una N-inteína GP41-1 y una C-inteína GP41-1). En otras realizaciones, una proteína de fusión comprende un dominio de inteína y un dominio de exteína, que no se encuentra junto con ese dominio de inteína particular en la naturaleza, también denominado en este documento un "dominio de exteína heterólogo". A modo de ejemplo, una proteína de fusión puede comprender un dominio de inteína GP41-1 y un dominio de exteína IMPDH.

Las proteínas de fusión de la invención pueden incluir, además, opcionalmente uno o más rótulos detectables. Los rótulos adecuados para su uso de acuerdo con la presente invención son conocidos en la técnica y, por lo general, incluyen cualquier molécula que, por su naturaleza química, y ya sea por medios directos o indirectos, proporcione una señal identificable que permita la detección de una proteína. Así, por ejemplo, las proteínas de fusión se pueden rotular de una manera convencional, como con moléculas reporteras específicas, fluoróforos, materiales radiactivos o enzimas (por ejemplo, peroxidasas, fosfatasa). En una realización particular, las proteínas de fusión incluyen uno o más colorantes fluorescentes como rótulos detectables. Los métodos estándar para modificar una proteína para

incluir un rótulo detectable son conocidos en la técnica.

En diversas realizaciones, la invención se refiere, además, a un ácido nucleico aislado, que comprende una secuencia de nucleótidos que codifica una proteína de fusión de la invención, un vector de expresión que comprende dichos ácidos nucleicos y una célula huésped que lleva dichos vectores de expresión.

5 III. Matrices de cromatografía de afinidad que comprenden proteínas de fusión N-inteína

Las proteínas de fusión descritas en el presente documento que contienen polipéptidos de N-inteína y contrapartidas de solubilización de N-inteína tienen utilidad, entre otras cosas, como ligandos para aplicaciones de cromatografía de afinidad. Por consiguiente, la presente invención, en ciertas realizaciones, proporciona matrices de cromatografía de afinidad que comprenden una proteína de fusión, la cual comprende un polipéptido de N-inteína y una contrapartida de solubilización de N-inteína unida a un soporte sólido.

En una realización particular, el soporte sólido es una resina de cromatografía. En una cierta realización, la resina de cromatografía incluye una base hidrófila de poliviniléter. Las resinas de cromatografía adecuadas que tienen una base hidrófila de poliviniléter incluyen, aunque no taxativamente, resinas ESHMUNO® (EMD Millipore Corporation).

En otra realización, la resina de cromatografía es un medio polimérico basado en metacrilato sintético (por ejemplo, cuentas con un tamaño de partícula en el intervalo de 20-40 µm aproximadamente o de alrededor de 40-90 µm). En algunas realizaciones, la resina de cromatografía tiene funcionalidad de ácido carboxílico. Las resinas de cromatografía adecuadas que tienen funcionalidad de ácido carboxílico incluyen, aunque no taxativamente, resinas COO de FRACTOGEL® (EMD Millipore Corporation).

Otros soportes sólidos adecuados para las matrices de cromatografía de afinidad de la invención pueden incluir, por ejemplo, vidrio de poros controlados, sílice, óxido de circonio, óxido de titanio, agarosa, polimetacrilato, poliácrlato, poliacrilamida, alcohol polivinílico y poliestireno, así como sus derivados (por ejemplo, sus aleaciones).

Un material poroso utilizado como soporte sólido puede comprender un compuesto hidrófilo, un compuesto hidrófobo, un compuesto oleófilo, un compuesto oleófilo o cualquier combinación de los mismos. El material poroso puede comprender un polímero o un copolímero. Los ejemplos de materiales porosos adecuados incluyen, aunque no taxativamente, poliéter sulfona, poliamida, por ejemplo, nailon, polisacáridos, tales como, por ejemplo, agarosa y celulosa, poliácrlato, polimetacrilato, poliacrilamida, polimetacrilamida, politetrafluoroetileno, polisulfona, poliéster, fluoruro de polivinilideno, polipropileno, polietileno, alcohol polivinílico, policarbonato, polímero de un fluorocarbono, por ejemplo, poli(tetrafluoroetileno-co-perfluoro(alquil vinil éter)), vidrio, sílice, circonia, óxido de titanio, cerámica, metal y sus aleaciones.

El material poroso puede comprender una molécula orgánica o inorgánica o una combinación de moléculas orgánicas e inorgánicas y puede comprender uno o más grupos funcionales, por ejemplo, un grupo hidroxilo, un grupo tiol, un grupo amino, un grupo carbonilo o un grupo ácido carboxílico, adecuado para reaccionar, por ejemplo, formando enlaces covalentes para una modificación química adicional, con el fin de unirse covalentemente a una proteína. En otra realización, el material poroso puede no poseer un grupo funcional, pero puede estar recubierto con una capa de material que lleve grupos funcionales, tales como un grupo hidroxilo, un grupo tiol, un grupo aminoácido, un grupo carbonilo o un grupo ácido carboxílico.

En algunas realizaciones, se usa una matriz de separación por afinidad convencional, por ejemplo, de naturaleza orgánica y basada en polímeros que exponen una superficie hidrófila a los medios acuosos usados, por ejemplo, exponen grupos hidroxilo (-OH), carboxi (-COOH), carbonilo (-CHO, o RCO-R'), carboxamido (-CONH₂, posiblemente en formas N-sustituidas), amino (-NH₂, posiblemente en forma sustituida), oligo- o polietilenoxi en sus superficies externas y, si están presentes, también en sus superficies internas. En una realización, los polímeros pueden basarse, por ejemplo, en polisacáridos, tales como dextrano, almidón, celulosa, pululano, agarosa, etc., que se han reticulado ventajosamente, por ejemplo con biseóxidos, epihalohidrinatos, bromuro de alilo, éter alilglicídico hidrocarburos inferiores sustituidos con 1,2,3-trihalo, para proporcionar una porosidad y rigidez adecuadas. En otra realización, el soporte sólido comprende cuentas de agarosa porosas. Los diversos soportes utilizados en la presente invención pueden prepararse fácilmente, de acuerdo con métodos estándar conocidos en la técnica, tales como, por ejemplo, la gelificación en suspensión inversa descrita, por ejemplo, en Hjerten, *Biochim Biophys Acta* 79 (2), 393-398 (1964). Alternativamente, las matrices base pueden ser productos disponibles comercialmente, como los medios FastFlow SEPHAROSE™ (GE Healthcare, Uppsala, Suecia). En algunas realizaciones, particularmente ventajosas para separaciones a gran escala, el soporte está adaptado para aumentar su rigidez y, por lo tanto, hace que la matriz sea más adecuada para altas velocidades de flujo.

De un modo alternativo, el soporte sólido se puede basar en polímeros sintéticos, como alcohol polivinílico, polihidroxialquil-acrilatos, polihidroxialquil-metacrilatos, poliacrilamidas, polimetacrilamidas, etc. En el caso de los polímeros hidrófobos, tales como las matrices basadas en bencenos sustituidos con divinilo y monovinilo, la superficie de la matriz a menudo se hidrofaliza, para exponer grupos hidrofílicos, como se los definió con anterioridad, a un líquido acuoso circundante. Dichos polímeros pueden producirse fácilmente de acuerdo con métodos estándar, véase, por ejemplo, Arshady, *Chimica e L'Industria* 70 (9), 70-75 (1988). De un modo alternativo, se puede usar un producto disponible comercialmente, como SOURCE™ (GE Healthcare, Uppsala, Suecia) y la

resina POROS (Applied Biosystems, Foster City, CA).

En otras realizaciones más, el soporte sólido comprende un soporte de naturaleza inorgánica, por ejemplo, sílice, óxido de circonio, óxido de titanio y sus aleaciones. La superficie de las matrices inorgánicas a menudo se modifica para incluir grupos reactivos adecuados. Los ejemplos incluyen CM Zirconia (CIPHERGEN-BioSeptra (Cergy-pontoise, Francia)) y soportes CPG® (Millipore Corporation).

En algunas realizaciones, el soporte sólido puede basarse, por ejemplo, en zirconia, óxido de titanio o sílice en forma de vidrio de poro controlado, que puede modificarse para que contenga grupos reactivos y/o admita el empapado cáustico, para acoplarse a los ligandos.

Los formatos de soporte sólido ejemplares incluyen, aunque no taxativamente, una cuenta (esférica o irregular), una fibra hueca, una fibra sólida, una almohadilla, un gel, una membrana, un casete, una columna, un chip, un portaobjetos, un plato o un monolito.

Con respecto al formato de una matriz, en una realización, está en forma de un monolito poroso. En una realización alternativa, la matriz está en forma de cuentas o partículas que puede ser porosas o no porosas. Las matrices en forma de cuentas o partículas se pueden usar como lecho compacto o en forma suspendida. Las formas suspendidas incluyen aquellas conocidas como lechos expandidos y suspensiones puras, en las cuales las partículas o cuentas pueden moverse con libertad. En el caso de monolitos, lechos compactados y lechos expandidos, el procedimiento de separación comúnmente sigue la cromatografía convencional con un gradiente de concentración. En el caso de una suspensión pura, se utilizará el modo discontinuo. Además, se puede usar un soporte sólido en formas tales como una superficie, un chip, un capilar o un filtro.

La matriz también podría estar en forma de membrana en un cartucho. La membrana podría estar en formato de lámina plana, espiral o fibra hueca.

En ciertas realizaciones, el soporte sólido puede ser un soporte soluble, por ejemplo, un polímero soluble o un polímero soluble en agua. Los soportes solubles ejemplares incluyen, aunque no taxativamente, un biopolímero tal como, por ejemplo, una proteína o un ácido nucleico. El polímero también puede ser un polímero soluble sintético, tal como, por ejemplo, incluso aunque no en forma taxativa, un polímero que contenga grupos cargados negativamente (carboxílico o sulfónico), grupos cargados positivamente (amina cuaternaria, amina terciaria, grupos secundarios o primarios), grupos hidrófobos (grupos fenilo o butilo), grupos hidrófilos (grupos hidroxilo o amino) o una combinación de los anteriores. Se pueden encontrar polímeros solubles sintéticos ejemplares en la publicación internacional PCT n.º WO2008091740 y en la publicación estadounidense n.º US20080255027, cuyas enseñanzas completas se incorporan aquí como referencia.

En algunas realizaciones, el soporte sólido puede incluir una molécula de avidina (por ejemplo, estreptavidina) y la proteína de fusión N-inteína puede comprender una etiqueta de biotina (por ejemplo, una molécula de biotina unida covalentemente a la contrapartida de solubilización en la proteína de fusión), de tal manera que la unión de la proteína de fusión al soporte sólido se logre a través de la interacción de las moléculas de avidina y biotina.

Las proteínas de fusión de N-inteína de la invención se pueden unir al soporte sólido en un solo sitio en la proteína de fusión (unión de punto único) o en más de un sitio en la proteína de fusión (unión multipunto). Con preferencia, el polipéptido de N-inteína en la proteína de fusión está orientado aleado del soporte sólido, cuando la proteína de fusión está unida al soporte sólido. Por ejemplo, se pueden colocar grupos de aminoácidos reactivos únicos (por ejemplo, residuos de cisteína) en la contrapartida de solubilización, en ubicaciones que son distales a la región activa del dominio N-inteína, para garantizar que la N-inteína se aleje del soporte sólido.

Con preferencia, el o los sitios (por ejemplo, grupos de aminoácidos reactivos únicos) en la proteína de fusión que participan en la unión al soporte sólido se encuentran exclusivamente en la contrapartida de solubilización de N-inteína. Por consiguiente, para lograr esto, puede ser necesario modificar el polipéptido de N-inteína para eliminar el aminoácido que proporciona el sitio reactivo único (por ejemplo, cisteína), por ejemplo, mediante la eliminación o sustitución de dichos aminoácidos, dondequiera que se encuentren en el N-inteína. Los métodos para eliminar o sustituir aminoácidos en una proteína son bien conocidos en la técnica.

La proteína de fusión de N-inteína inmovilizada puede ser adecuada para separaciones cromatográficas de columna o de múltiples pocillos o puede ser paramagnética, de modo que pueda capturarse desde la solución mediante la aplicación de un campo magnético.

Se puede usar cualquier técnica adecuada para unir una proteína de fusión descrita en este documento a un soporte, por ejemplo, un soporte sólido que incluye aquellos que son de amplio conocimiento en la técnica y descritos en el presente documento. Por ejemplo, en algunas realizaciones, la proteína de fusión puede unirse a un soporte mediante técnicas de acoplamiento convencionales que utilizan, por ejemplo, grupos tiol, amino y/o carboxi presentes en la proteína de fusión. Por ejemplo, los bisepóxidos, epiclorhidrina, CNBr, N-hidroxisuccinimida (NHS), etc., son reactivos de acoplamiento muy conocidos. En algunas realizaciones, se introduce un espaciador entre el soporte y la proteína de fusión, lo cual mejora la disponibilidad de la proteína de fusión y facilita el acoplamiento químico de la proteína de fusión al soporte.

La unión de una proteína de fusión N-inteína a un soporte sólido se puede lograr de muchas maneras diferentes, la mayoría de las cuales son bien conocidas en la técnica, así como las descritas en el presente documento. Véase, por ejemplo, Hermanson *et al.*, *Immobilized Affinity Ligand Techniques*, Academic Press, pp. 51-136 (1992). Por ejemplo, los ligandos de proteínas se pueden acoplar a un soporte sólido a través de grupos activos ya sea sobre la superficie del soporte sólido o bien al ligando de las proteínas, como, por ejemplo, grupo hidroxilo, tiol, epóxido, amino, carbonilo, epóxido o ácido carboxílico. La unión se puede lograr usando químicas conocidas, que incluyen, entre otros, el uso de bromuro de cianógeno (CNBr), éster de N-hidroxi succinimida, activación de epoxi (bisoxirano) y aminación reductora.

En una realización particular, se usa una resina de cromatografía (por ejemplo, cuentas) que tiene grupos ácido carboxílico (-COOH) o amino (-NH₂). En una realización adicional, la resina de cromatografía también tiene grupos hidroxilo (-OH) y/u otro grupo funcional que se puede convertir en -COOH o -NH₂ o -OH.

En algunas realizaciones, el acoplamiento de proteínas dirigido por tiol se puede usar para unir la proteína de fusión N-inteína de la invención a un soporte sólido. El acoplamiento de proteínas dirigido a tiol se ha descrito en la bibliografía. Véase, por ejemplo, Ljungquist, et al., *Eur. J. Biochem.* Vol. 186, pp. 558-561 (1989). Se sabe que las maleimidas reaccionan selectivamente con grupos tiol, a pH 7,0-7,5. Con un pH > 8, también pueden reaccionar con grupos de aminas y, además, tienden a hidrolizarse (Greg T. Hermanson, *Bioconjugation Techniques*, Academic Press, 2008; Ian Johnson, Michelle T.Z. Spence, *Molecular Probes Handbook, A Guide to Fluorescent Probes and Labeling Technologies*, 2010). Por debajo del pH 8, las yodoacetamidas también son altamente selectivas hacia los grupos tiol (Greg T. Hermanson, *Bioconjugation Techniques*, Academic Press, 2008; Ian Johnson, Michelle T.Z. Spence, *Molecular Probes Handbook, A Guide to Fluorescent Probes and Labeling Technologies*, 2010). Sin embargo, las yodoacetamidas son intrínsecamente inestables en la luz y la mayoría de los enlaces disponibles comercialmente no son solubles en agua y/o son muy caros. Dado que la selectividad de la yodoacetamida hacia los grupos tiol no es superior a la maleimida, la maleimida suele ser la mejor opción para la fabricación a gran escala.

En algunas realizaciones, el ligando de N-inteína se puede acoplar a AMP o FG-COO activado con yodoacetamida a través de un único grupo sulfhidrilo disponible en el dominio de solubilización. La densidad del ligando de la resina derivada se puede calcular midiendo el agotamiento de una proteína de fusión que contenga la C-inteína de la solución. Hasta la fecha, se ha logrado una densidad de ligando de N-inteína no optimizada de 1 g/litro de FG-COO.

Muchas proteínas también se han acoplado con éxito a resinas activadas con epoxi, tales como FRACTOGEL® Epoxy. El epóxido reacciona con grupos amino primarios, grupos hidroxilo y sulfhidrilo y produce matrices de afinidad muy estables (PV Kuznetsov 1993. *Pharmaceutical Chemistry Journal* 27: 439-52).

En algunas realizaciones, los ligandos de proteínas pueden acoplarse a un soporte sólido a través de un enlace intermedio. El enlace puede comprender al menos un grupo funcional acoplado a un resto de enlace. El resto de enlace puede comprender cualquier molécula capaz de ser acoplada a un grupo funcional. Por ejemplo, el resto de enlace puede incluir un grupo alquilo, alquenilo o alquinilo, cualquiera de ellos. El resto de enlace puede comprender una cadena de carbono, que varía de 1 a 30 átomos de carbono. En algunas realizaciones, el enlace puede comprender más de 30 átomos de carbono. El resto de enlace puede comprender al menos un heteroátomo, tal como nitrógeno, oxígeno y azufre. El resto de unión puede comprender una cadena ramificada, una cadena no ramificada o una cadena cíclica. El resto de enlace puede estar sustituido con dos o más grupos funcionales.

La elección de las condiciones apropiadas de tampón para acoplar un ligando de proteína a un soporte sólido está dentro de la capacidad del experto en la técnica. Los tampones adecuados incluyen cualquier tampón que no contenga amina, tal como los tampones de carbonato, bicarbonato, sulfato, fosfato y acetato. Cuando se utiliza la química asociativa, la concentración de sal del tampón dependerá del grupo asociativo utilizado. Por ejemplo, la concentración de sal puede estar en el rango de 5 nM-100 mM. Cuando se usa una especie cargada, la concentración de sal puede ser de al menos 5 nM pero de menos de 0,1 M, al menos 5 nM pero menos de 0,01 M, al menos 5 nM pero menos de 0,001M. En ciertas realizaciones, la concentración de sal puede ser de 0,01 M. Cuando se usa una especie hidrófoba, por lo general, es conveniente una alta concentración de sal. Por lo tanto, la concentración de sal puede ser mayor que 0,001 M, mayor que 0,01 M, o mayor que 0,1 M.

En algunas realizaciones, cuando se usa química asociativa, la reacción se realiza a una temperatura que varía de 0 °C a 99 °C. En ciertas realizaciones, el método de reacción se pone en práctica a una temperatura inferior a 60 °C, inferior a 40 °C, inferior a 20 °C, o inferior a 10 °C. En algunas realizaciones, el método de la invención se pone en práctica a una temperatura de 4 °C aproximadamente. En otras realizaciones, el método de la invención se pone en práctica a una temperatura de 20 °C.

En otras realizaciones, la proteína de fusión N-inteína se puede combinar con varios modificadores (membranas, superficies poliméricas, fluorescentes u otros rótulos de detección) en combinación con productos químicos de reticulación o condensación apropiados para formar un aducto covalente que incluya a proteína de fusión N-inteína y el modificador.

IV. Métodos para usar las proteínas de fusión basadas en inteínas de la invención

Las proteínas de fusión descritas en el presente documento que contienen polipéptidos de N-inteína y contrapartidas

de solubilización de N-inteína, y las matrices de cromatografía de afinidad que comprenden dichas proteínas de fusión, tienen utilidad, entre otras cosas, en los métodos de purificación por afinidad, métodos de selección de complejos de inteína dividida activa adecuadas para uso en métodos de purificación por afinidad y métodos de escisión y ligadura peptídicos, como se describe más adelante en este documento.

- 5 Por consiguiente, la presente descripción, en ciertas realizaciones, se refiere a un método de purificación por afinidad de una molécula diana en una muestra. En un aspecto de esta realización, el método comprende: a) proporcionar una muestra que contiene una primera proteína de fusión, que comprende un polipéptido de C-inteína unido a una molécula diana mediante un enlace peptídico; b) poner en contacto la muestra con una matriz de cromatografía de afinidad que comprende una segunda proteína de fusión, en donde la segunda proteína de fusión
10 comprende un polipéptido de N-inteína unido por un enlace peptídico a una contrapartida de solubilización de N-inteína que promueve la solubilidad del polipéptido de N-inteína, en condiciones en las que el polipéptido de C-inteína en la primera proteína de fusión se une selectivamente al polipéptido de N-inteína en la segunda proteína de fusión, para formar un complejo de inteína que es inactivo; c) lavar la matriz de cromatografía de afinidad que contiene el complejo de inteína inactivo, para eliminar los contaminantes no unidos; d) exponer el complejo de inteína a condiciones en las que el complejo de inteína es activo y escinde la molécula diana del polipéptido de C-inteína y e) recuperar la molécula diana escindida.

La proteína de fusión que comprende un polipéptido de N-inteína unido a una contrapartida de solubilización de N-inteína puede ser cualquiera de las proteínas de fusión de N-inteína descritas en otra parte en el presente documento.

- 20 La muestra que contiene la proteína de fusión que comprende un polipéptido de C-inteína unida a una molécula diana puede ser cualquier muestra adecuada (por ejemplo, muestra biológica). En una realización, la muestra es una preparación o mezcla de proteína sin procesar (por ejemplo, un extracto celular).

- La molécula diana puede ser cualquier biomolécula de interés. A modo de ejemplo, las biomoléculas de interés pueden incluir proteínas y ensamblajes biomoleculares (por ejemplo, producidos por tecnología de ADN recombinante), como, por ejemplo, hormonas (por ejemplo, insulina, hormona de crecimiento humana, eritropoyetina, interferones, factor estimulante de colonias de granulocitos, activador de plasminógeno tisular), anticuerpos monoclonales (mAb) y derivados de mAb (por ejemplo, mAb bi-específicos, Fab, scFv, anticuerpos de tiburones y camélidos), productos terapéuticos derivados de andamios (por ejemplo, DARPin, Affibodies, anticalins), enzimas terapéuticas (por ejemplo, galactosidasa alfa A, alfa-L-iduronidasa, N-acetilgalactosamina-4-sulfatasa, glucocerebrosidasa), toxinas (por ejemplo, botulinum, CRM 197, ricina), vacunas recombinantes (por ejemplo, ántrax, difteria, tétanos, neumonía, virus de la hepatitis B, virus del papiloma humano), partículas similares a virus (por ejemplo, hepatitis B, papiloma humano, influenza, parvovirus, virus Norwalk), así como enzimas industriales (por ejemplo, papaína, bromelina, tripsina, proteinasa K, enzima BENZONASE™, enzima DENERASE™, ureasa, pepsina, etc.) y reactivos de diagnóstico (por ejemplo, glucosa y lactato deshidrogenasa, ADN polimerasas, fosfatasa alcalina, peroxidasa de rábano picante, enzimas de restricción, anticuerpos derivados de hibridomas, etc.). En una realización particular, la molécula diana es un anticuerpo (por ejemplo, un anticuerpo monoclonal) a una diana terapéutica.

- Dependiendo de la inteína particular utilizada (por ejemplo, especie *Synechocystis* (Ssp) DnaB, *Nostoc punctiforme* (Npu) DnaE, GP41-1), las condiciones de carga, lavado, escisión y elución difieren significativamente. No obstante, los expertos en la técnica pueden determinar fácilmente las condiciones apropiadas de carga, lavado, escisión y elución para una inteína particular. Las condiciones adecuadas (por ejemplo, concentración de agentes caotrópicos y reductores/oxidantes, iones metálicos (por ejemplo, cinc, calcio, estroncio, magnesio, manganeso), agentes excluyentes de volumen (por ejemplo PEG, PVP, dextrano), detergentes, sales, temperatura y pH) para inteínas particulares incluyen, aunque no taxativamente, las condiciones descritas a continuación. Para la inteína GP41-1 en particular, se sabe que la actividad no se ve afectada relativamente por un pH variable en el rango de 6-10 (Carvajal-Vallejos P., *et al.*, *J. Biol. Chem.* 287: 28686-28696 (2012)).

- Un experto en la técnica puede determinar las condiciones en las cuales el polipéptido de C-inteína en la primera proteína de fusión se une selectivamente al polipéptido de N-inteína en la segunda proteína de fusión para formar un complejo de inteína catalíticamente inactivo, para una inteína dada. En general, para el proceso a escala industrial, el cambio de temperatura durante una separación cromatográfica se considera poco práctico, ya que el cambio resultaría en un largo paso de equilibrio para garantizar que la columna y la temperatura del empaque sean uniformes en todo momento. Las condiciones de unión ejemplares incluyen: a) una temperatura variable en el rango de aproximadamente 4-25 °C, y un tampón que comprenda Tris/HCl 50 mM, NaCl 300 mM, EDTA 1 mM, glicerol al 10 % (v/v), DTT 2 mM, pH = 7 (por ejemplo, para GP41-1, véase Carvajal-Vallejos P., *et al.*, *J. Biol. Chem.* 287: 28686-28696 (2012)); b) una temperatura variable en el rango de 4-25 °C aproximadamente y un tampón que comprenda NaAc 50 mM, NaCl 0,5 M, pH = 5 (por ejemplo, para la inteína DnaB, véase Lu W., *et al.*, *J. Chrom. A*, 1218: 2553-2560 (2011)) y c) una temperatura variable en el rango de 4-25 °C aproximadamente y un tampón que comprenda NaCl 0,5 M, Tris-HCl 10 mM, cloruro de zinc 0,5 mM, pH = 8 (por ejemplo, para Npu DnaE, véase Guan D., *et al.*, *Biotech. Bioeng.* 110: 2471-2481 (2013)).

- 60 De un modo similar, las condiciones que promueven la actividad catalítica del complejo de inteína pueden variar

dependiendo de las inteínas utilizadas y pueden ser determinadas por un experto en la técnica. Las condiciones ejemplares para promover la actividad de la inteína catalítica incluyen: a) un tampón que comprenda Tris-HCl 50 mM, pH = 7,0, NaCl 300 mM, EDTA 1 mM; b) un tampón, que comprenda L-arginina 0,3 M, EDTA 5 mM, un tampón fosfato 50 mM, pH = 6,5 y c) un tampón que comprenda NaCl 0,5 M, Tris-HCl 10 mM, DTT 50 mM, pH = 8,0.

5 En un aspecto de esta realización, el método puede comprender, además, limpiar, regenerar y/o almacenar las matrices de cromatografía de afinidad de la invención. Típicamente, una matriz de cromatografía de afinidad puede limpiarse en condiciones alcalinas o ácidas, dependiendo de la composición de la matriz. Un experto en la técnica puede determinar las condiciones adecuadas para limpiar, regenerar, restaurar y/o almacenar una matriz de afinidad.

10 Un método de purificación de afinidad ejemplar se proporciona en la figura 1 y el ejemplo 10 descritos en el presente documento.

En otra realización más, la descripción se refiere a un método de selección de un complejo de inteína catalíticamente activo, que es adecuado para uso en purificación por afinidad. En un aspecto de esta realización, el método comprende: a) poner en contacto una primera proteína de fusión que comprende un polipéptido de C-inteína unido a una molécula diana mediante un enlace peptídico, con una segunda proteína de fusión que comprende un polipéptido de N-inteína unido a una Contrapartida de solubilización de N-inteína por un enlace peptídico, en condiciones en las que el polipéptido de C-inteína en la primera proteína de fusión se une selectivamente al polipéptido de N-inteína en la segunda proteína de fusión, para formar un complejo de inteína y b) determinar si la molécula diana se escinde del polipéptido de C-inteína en condiciones que promueven la actividad de inteína, en donde la presencia de la molécula diana escindida es indicativa de un complejo de inteína catalíticamente activo.

Las N-inteínas y las C-inteínas empleadas en el método de esta realización pueden ser cualquier par de inteínas divididas complementarias, tales como, por ejemplo, los pares de inteína divididos descritos en este documento (por ejemplo, la N-inteína GP41-1 y las C-inteínas).

Las condiciones en las cuales el polipéptido de C-inteína en la primera proteína de fusión se une selectivamente al polipéptido de N-inteína en la segunda proteína de fusión para formar un complejo de inteína catalíticamente inactivo pueden variar, dependiendo de las inteínas utilizadas y se pueden determinar apelando a los conocimientos comunes en la técnica. Las condiciones de unión ejemplares incluyen: a) una temperatura variable en el rango de 4-25 °C aproximadamente, y un tampón que comprenda Tris-HCl 100 mM, NaCl 25 mM, cloruro de zinc 0,1 mM, pH = 9; b) una temperatura comprendida en el intervalo de 4-25 °C aproximadamente y un tampón que comprenda NaAc 50 mM, NaCl 0,5 M, pH = 5 y c) una temperatura variable en el rango de 4-25 °C aproximadamente y un tampón que comprenda NaCl 0,5 M, Tris-HCl 10 mM, pH = 8.

La molécula diana puede ser cualquier molécula diana adecuada, que incluye, entre otras, cualquiera de las moléculas diana descritas en el presente documento.

Ejemplos

35 Ejemplo 1. Selección de la caracterización de las contrapartidas de solubilización candidatos para la N-inteína GP41-1

Utilizando el conjunto de las más de 4000+ proteínas de *Escherichia* conocidas como punto de partida, se seleccionaron siete contrapartidas de solubilización (véase la tabla 2, SEQ ID NOs: 11-15, 19, 20) para las pruebas, aplicando los siguientes criterios:

- 40 (1) las proteínas seleccionadas carecían de residuos de cisteína;
- (2) las proteínas seleccionadas se predijeron vía simulación computacional para ser solubles cuando se sobreexpresan en la *E. coli*;
- (3) las proteínas seleccionadas tenían un peso molecular inferior a 11 kDa;
- (4) las proteínas seleccionadas se predijeron como no secretables, ya fuera vía simulación computacional o por conocimiento concreto;
- 45 (5) cuando se disponía de información sobre las interacciones de las proteínas, las proteínas seleccionadas eran monoméricas más que multiméricas;
- (6) cuando se disponía de información sobre la función de las proteínas, las proteínas seleccionadas no tenían un carácter regulador o tóxico, lo que significaba que no participaban en el control de las principales vías celulares o que podían causar la muerte de las *E. coli* que las sobreexpresaba (por ejemplo, nucleasas, polimerasas, etc.) y
- 50 (7) se favorecieron aquellas proteínas que tenían estructuras cristalográficas por rayos X o RMN conocidas.

La tabla 3 proporciona las propiedades físicas (peso molecular (mw), pH isoelectrico (pI), probabilidad de expresión

soluble en *E. coli*, si se predice que la proteína se secretará en *E. coli*, el gran promedio de la hidrofobicidad (GRAVY), y el índice alifático (AI) para las inteínas y las contrapartidas de solubilización evaluadas en este estudio, que se calcularon utilizando algoritmos disponibles públicamente. Todos los parámetros físicos, con la excepción de la probabilidad de solubilidad tras la sobreexpresión en la *E. coli* y la predicción de probabilidad de secreción, se calcularon utilizando la herramienta en línea ProtParam (<http://web.expasy.org/tools/protparam/>) que es parte del conjunto de herramientas de bioinformática SwissProt ExPASy. El peso molecular se expresa en daltons. El pI para cada proteína es el valor de pH en el que la proteína no tiene carga neta. El punto isoeléctrico (pI) es un pH en el que la carga neta de la proteína es cero. En el caso de las proteínas, el punto isoeléctrico depende principalmente de siete aminoácidos cargados: glutamato (grupo δ -carboxilo), aspartato (grupo β -carboxilo), cisteína (grupo tiol), tirosina (grupo fenol), histidina (cadenas laterales de imidazol), lisina (grupo ϵ -amonio) y arginina (grupo guanidinio). Adicionalmente, se debe tener en cuenta la carga de los grupos terminales de proteínas (NH₂ i COOH). Cada uno de ellos tiene su constante de disociación ácida única, conocida como pK. Además, la carga neta de la proteína está en estrecha relación con el pH de la solución (tampón). Teniendo en cuenta esto, podemos usar la ecuación de Henderson-Hasselbach para calcular la carga de proteína en cierto pH:

Aminoácido	NH ₂	COOH	C	D	E	H	K	R	Y
pKa (wikipedia)	8,2	3,65	8,18	3,9	4,07	6,04	10,54	12,48	10,46

El promedio general de hidropaticidad (GRAVY) (Kyte J y Doolittle RF., J. Mol. Biol. 157: 105, 1982) de una secuencia polipeptídica lineal se calcula como la suma de los valores de hidropatía de todos los aminoácidos, dividida por el número de residuos en la secuencia. El aumento de la puntuación positiva indica mayor hidrofobicidad. El cálculo se basa en la escala de Kyte-Doolittle. GRAVY es un método simple para mostrar el carácter hidrofóbico de una proteína.

Alanina	1,8	Leucina	3,8
Arginina	-4,5	Lisina	-3,9
Asparagina	-3,5	Metionina	1,9
Ácido aspártico	-3,5	Fenilalanina	2,8
Cisteína	2,5	Prolina	-1,6
Glutamina	-3,5	Serina	-0,8
Ácido glutámico	-3,5	Treonina	-0,7
Glicina	-0,4	Triptófano	-0,9
Histidina	-3,2	Tirosina	-1,3
Isoleucina	4,5	Valina	4,2

El índice alifático (Ikai, AJ., J. Biochem. 88: 1895, 1980) de una proteína se define como el volumen relativo ocupado por las cadenas laterales alifáticas (alanina, valina, isoleucina y leucina). Puede considerarse como un factor positivo para el aumento de la termoestabilidad de las proteínas globulares. El índice alifático de una proteína se calcula de acuerdo con la siguiente fórmula: Índice alifático = X (Ala) + a * X (Val) + b * (X (Ile) + X (Leu)). * Los coeficientes a y b son el volumen relativo de la cadena lateral de valina (a = 2,9) y de las cadenas laterales de Leu/Ile (b = 3,9) a la cadena lateral de alanina. La probabilidad de generar un producto soluble tras la sobreexpresión en *E. coli* también se puede calcular utilizando el algoritmo de Wilkinson y Harrison (Wilkinson DL y Harrison RG., Bio/Technology, 9: 443, 1991). Otros algoritmos disponibles no dan necesariamente resultados similares.

La predicción de si la proteína contiene una señal de secreción funcional se realizó utilizando el algoritmo SignalP 4.1, disponible en el Centro para el Análisis de Secuencias Biológicas de la Universidad Técnica de Dinamarca (<http://genome.cbs.dtu.dk/services/SignalP/>).

Tabla 3. Parámetros físicos medidos para las inteínas y las contrapartidas de solubilización de N-inteína utilizadas en este estudio

PROTEÍNA	PESO MOLECULAR (D)	PI	PROBABILIDAD DE SOLUBILIDAD EN LA <i>E. COLI</i>	SECRETADA	GRAVY	ÍNDICE ALIFÁTICO
CONTRAPARTIDA DE SOLUBILIZACIÓN 46	5639,4	5,8	74	NO	-0,086	95,7
CONTRAPARTIDA DE SOLUBILIZACIÓN 206	8799,8	5,6	97	NO	-0,636	93,9
CONTRAPARTIDA DE SOLUBILIZACIÓN 246	9385,6	5,4	84	NO	-0,38	89,3
CONTRAPARTIDA DE SOLUBILIZACIÓN 51	6024,8	4,5	98	NO	-0,391	117,7
CONTRAPARTIDA DE SOLUBILIZACIÓN 138	8325,2	5,4	69	NO	-1,526	48,0
CONTRAPARTIDA DE SOLUBILIZACIÓN 138_ GKL22GCKL	8428,3	5,4	60	NO	-1,469	47,3
CONTRAPARTIDA DE SOLUBILIZACIÓN 138_ GYQ48GCYQ	8428,3	5,4	60	NO	-1,469	47,3
CONTRAPARTIDA DE SOLUBILIZACIÓN 138_ GYQ48GCGYQ	8485,4	5,4	60	NO	-1,454	46,6
CONTRAPARTIDA DE SOLUBILIZACIÓN 342	10000	4,3	98	NO	-0,148	93,4
CONTRAPARTIDA DE SOLUBILIZACIÓN 368	10750,2	6,5	72	NO	-0,398	95,5
N-INTEÍNA GP41-1, VARIANTE NINTΔA_ CC	10589,1	5,8	21	NO	-0,44	80,1
C-INTEÍNA GP41-1 (CINT)	4916,7	5,1	86	NO	-0,312	111,4
FUSIÓN DE C-INTEÍNA GP41-1 CON TIORREDOXINA (CINT_ TRX)	17528,1	5,5	78	NO	-0,189	101,35
CONTRAPARTIDA DE SOLUBILIZACIÓN A	20504,9	4,6	34	NO	-0,14	91,5
CONTRAPARTIDA DE SOLUBILIZACIÓN B	7597,6	5,6	51	NO	-0,606	96,0
CONTRAPARTIDA DE SOLUBILIZACIÓN C	1281,5	5,4	94	NO	0,127	123,6
CONTRAPARTIDA DE SOLUBILIZACIÓN D	37389,1	6,3	42	NO	-0,388	77,9

PROTEÍNA	PESO MOLECULAR (D)	PI	PROBABILIDAD DE SOLUBILIDAD EN LA <i>E. COLI</i>	SECRETADA	GRAVY	ÍNDICE ALIFÁTICO
CONTRAPARTIDA DE SOLUBILIZACIÓN E	12740,2	5,8	41	NO	-0,299	60,4
CONTRAPARTIDA DE SOLUBILIZACIÓN F	43387,6	5,5	40	NO	-0,251	85,4
CONTRAPARTIDA DE SOLUBILIZACIÓN G	12629,4	5,7	74	NO	-0,143	97,6
CONTRAPARTIDA DE SOLUBILIZACIÓN H	54870,9	4,5	95	NO	-0,278	98,0

Ejemplo 2. Creación de construcciones de expresión de proteínas de *E. coli*

Las construcciones de plásmidos que llevan la secuencia de codificación para las posibles contrapartidas de solubilización 46, 206 y 246 se fusionaron con la secuencia de codificación para NINTΔA_CC, ya fuera a través del aminoácido amino o carboxi terminal de NINTΔA_CC y se insertaron en una versión de pJ414 de DNA2.0. Estas construcciones se transformaron en células BL21 DE3 de *E. coli* competentes, utilizando métodos convencionales y se aislaron colonias resistentes a la ampicilina. La producción de proteínas del tamaño esperado se confirmó mediante electroforesis de poliacrilamida SDS (SDS PAGE).

Los transformantes de cada una de las 6 construcciones se cultivaron en 2 ml de LB, que contenía 100 µg de ampicilina/ml (LB + Amp) de una reserva de glicerol de células BL21 DE3 de *E. coli* transformadas con la construcción correspondiente. Se dejó que este preinóculo creciera durante la noche a 37 °C y 250 rpm y se usó para inocular 200 ml de LB + Amp (inóculo al 1 %). El cultivo se incubó a 37 °C y 250 rpm, a una OD₆₀₀ de entre 0,5 - 0,6. La expresión de proteínas se indujo mediante la adición de 0,4 mM de IPTG. La temperatura se redujo a 30 °C y el cultivo se incubó a esta temperatura y 250 rpm durante 5 horas. Después de ese período, las células se recolectaron mediante centrifugación (4500 g, 25 min, 4 °C), el sobrenadante se descartó y el gránulo celular se mantuvo a -80 °C para una purificación adicional de la proteína.

La región codificante para la proteína de sustrato de prueba CINT_TRX se clonó en pSABAD92A (número de acceso al GenBank HM070247) y se transformó en células BL21 DE3 competentes. Los transformantes exitosos se aislaron en caldo Luria más 50 µg/ml de carbenicilina (LB + C). La producción de proteínas del tamaño esperado se confirmó utilizando SDS PAGE. Las reservas de glicerol de cada uno de los tres clones/construcción BL21 se almacenaron a -80 °C.

Se usó una pequeña cantidad de reservas de glicerol BL21 congelado para inocular un cultivo de 5 ml en LB + C a 37 °C, 250 rpm. Al día siguiente, se usaron 0,1 ml del cultivo desarrollado durante toda la noche para inocular 10 ml de LB + C y este cultivo se desarrolló a 37 °C, 250 rpm a una OD₆₀₀ de 0,6 a 0,9. Los cultivos se indujeron con arabinosa al 0,02 % a 28 °C, 250 rpm durante 5 horas. Después de la inducción, las células se recolectaron mediante centrifugación (4500 g, 25 min, 4 °C), el sobrenadante se descartó y el gránulo celular se mantuvo a -80 °C para una purificación adicional de la proteína.

Ejemplo 3. Determinación de la relación soluble frente a insoluble y de la cantidad total de proteínas expresadas

Para determinar los rendimientos de expresión y la relación soluble:insoluble para cada construcción, se centrifugaron alícuotas de los cultivos desarrollados correspondientes a biomásas equivalentes, cultivados como se indicó con anterioridad, a 5000 g, 15 min, 4 °C. Después de descartar el sobrenadante del cultivo, las células se resuspendieron en 200 µl de un tampón de solubilización que consistía en Tris 50 mM, pH 8, NaCl 300 mM, Triton X-100 al 0,5 %. Las células se rompieron por sonicación (10 ráfagas x 3, Branson 250 Sonifier, con tiempo entre cada serie para permitir el enfriamiento de la muestra). Para separar la fracción soluble de la insoluble, las muestras se centrifugaron a 16.000 g y 4 °C durante 10 min. Las fracciones solubles se retiraron en un tubo separado, mientras que las fracciones insolubles se resuspendieron en 200 µl del mismo tampón de solubilización por sonicación (utilizando los mismos parámetros que en la sonicación anterior).

Las proteínas recombinantes en lisados celulares se cuantificaron después de gel de SDS PAGE, luego de la tinción con Coomassie, usando análisis densitométrico de la utilización de una curva de BSA cuantificada como referencia. Se cargaron tres volúmenes de muestra diferentes por clon junto con la curva estándar de BSA (6 puntos de 0,2 a 1,2 µg). Las intensidades de las bandas de proteínas se determinaron mediante densitometría utilizando el

software "Quantity One" (Biorad). Se aplica un factor de corrección para cada proteína considerando la relación de peso molecular BSA/proteína recombinante.

Las concentraciones para proteínas purificadas se determinaron utilizando los coeficientes de extinción calculados y su absorbancia se calculó a 280 nm.

5 Ejemplo 4. Purificación de proteínas expresadas

Para purificar la proteína de fusión de C-inteína INT_TRX utilizada como sustrato de escisión en su totalidad, las células de *E. coli* que expresaban la proteína se resuspendieron en un tampón que contenía Tris-HCl 50 mM, pH = 8,0, NaCl 300 mM, 0,5X CellLytic B (Sigma -Aldrich) e imidazol 20 mM. Las células se sonicaron en hielo durante 20 minutos, con un ciclo de actividad pulsada del 30 % (Branson 250 Sonifier) y se centrifugaron durante 30 minutos a 34500 g a 4°C. La fusión de C-inteína soluble se purificó a partir del sobrenadante en las columnas His-Trap HP (GE Healthcare), siguiendo las instrucciones del fabricante. Las fracciones eluidas que contenían las proteínas de fusión C-inteína purificadas se agruparon, se dializaron contra el tampón de escisión (Tris-HCl 50 mM, pH = 7,0, NaCl 300 mM, EDTA 1 mM, glicerol al 10 % (v/v); CB) en presencia de DTT 2 mM y se almacenaron en alícuotas a -80 °C.

Las fusiones de N-inteína se purificaron en condiciones nativas, a partir de células de *E. coli* que albergan las construcciones de expresión. Los gránulos celulares se resuspendieron en un tampón que contenía Tris-HCl 100 mM, pH = 8,0, NaCl 150 mM y EDTA 1 mM. Las células se sometieron a sonicación en hielo durante 20 minutos, con un ciclo de actividad pulsada del 30 % (Branson 250 Sonifier) y se centrifugaron durante 30 minutos a 34500 g a 4 °C. La fracción soluble de la fusión de N-inteína se purificó por cromatografía en columnas His-Trap HP, como se describió con anterioridad. Las fracciones eluidas que contenían la proteína purificada se agruparon, se dializaron contra CB, en presencia de DTT 2 mM y se almacenaron en alícuotas a -80 °C.

20 Ejemplo 5. Determinación de la cinética de escisión para proteínas expresadas

Las reacciones *in vitro* se realizaron como se describió previamente (Carvajal-Vallejos P., *et al.*, *J. Biol. Chem.* 287: 28686, 2012). En resumen, las proteínas de fusión N y C purificadas se preincubaron brevemente por separado, en las condiciones de prueba correspondientes. La reacción de escisión se inició mezclando proteínas de fusión N-inteína y C-inteína complementarias en un tampón de escisión, a concentraciones equimolares de 5 µM y se incubaron a 25 °C y 37 °C. Para los experimentos relacionados con esta invención, la contrapartida de escisión siempre fue CINT_TRX. Las alícuotas se eliminaron a intervalos específicos y la reacción se detuvo mediante la adición de un tampón SDS PAGE, que contenía un 8 % de SDS (p/v) y un 20 % de β-mercaptoetanol (v/v), seguido de 5 minutos de ebullición. Los productos de las reacciones se cuantificaron mediante SDS PAGE (4-12 % de geles de Bis-Tris de Novex, Invitrogen, Carlsbad, EE. UU.), seguido de la tinción con azul brillante de Coomassie (Sigma). Las intensidades relativas de las bandas de proteínas se determinaron densitométricamente, utilizando el programa de Quantity One (BioRad). Los diferentes productos de escisión se normalizaron de acuerdo con su peso molecular correspondiente. El porcentaje de escisión de proteínas se calculó a partir de la relación de los productos escindidos y el precursor marcado con inteína CINT_TRX. Las tasas constantes (kobs) se determinaron utilizando el software GraFit (Erithacus, Surrey, Reino Unido), ajustando los datos a la ecuación $P = P_0 (1 - e^{-kt})$, donde P es la cantidad de producto de fusión de C-inteína escindido formado en el momento t; P₀ es la cantidad máxima de producto escindido que se puede obtener (rendimiento); e es la constante de Euler y k es la tasa observada. Todas las reacciones se trataron como procesos irreversibles, de estado preestabilizado y de primer orden, bajo el supuesto de que, después de la asociación rápida de los dos fragmentos de inteína complementarios, la escisión de la proteína de fusión C-inteína se produce como una reacción monomolecular.

Ejemplo 6. Determinación de la ubicación y propiedades óptimas para las contrapartidas de solubilización NINT

Se crearon fusiones NINTΔA_CC para potenciales contrapartidas de solubilización en ambas posibles orientaciones (es decir, se fusionaron con el término N o C de NINTΔA_CC) para todas las contrapartidas de solubilización. Las seis construcciones resultantes se expresaron en la *E. coli*, y la proteína producida se analizó con respecto a la cantidad total producida y la solubilidad, como se describió con anterioridad. Además, se purificó la proteína de cada construcción y se caracterizó la velocidad de escisión, utilizando CINT_TRX purificado como sustrato. Los resultados de este análisis se muestran en las figuras. 2A y 2B. Mientras que la fusión de la contrapartida de solubilización al término N de NINTΔA_CC produce mayores cantidades de proteína en la *E. coli* para todas las construcciones probadas que la fusión de la contrapartida de solubilización al término C, se observa lo contrario de esta tendencia cuando se miden las tasas de escisión. El trabajo publicado utilizando un sistema diferente de inteínas divididas mientras estos estudios estaban en curso demostró una relación similar entre la ubicación de la contrapartida de solubilización y la N-inteína y la actividad de N-inteína que se explicó al referirse a información estructural conocida que indica una mayor probabilidad de interferencia estérica entre los dominios de inteína, cuando la fusión se realiza en la polaridad opuesta (Guan D, Ramírez M, Chen Z., *Biotechnol Bioeng.* 110: 2471, 2013).

Este trabajo se amplió para incluir las contrapartidas de solubilización adicionales 51, 138, 342 y 368 (véanse las tablas 2 y 3), que eran distintas de las contrapartidas de solubilización caracterizadas previamente, en términos de tamaño y punto isoeléctrico (pI). Todas estas se fusionaron al término carboxilo de NINTΔA_CC, como se mostró con anterioridad con las contrapartidas de solubilización 46, 206 y 246, para producir fusiones que tenían la actividad

catalítica más alta. Estas construcciones se expresaron en la *E. coli*, se purificaron y se analizaron para determinar la tasa de escisión, como se describió con anterioridad. Los resultados de estos análisis se presentan en la figura 3. Mientras que la contrapartida de solubilización 246 claramente tiene la mayor actividad, el mejor compromiso entre la actividad catalítica y la expresión soluble se observó para la contrapartida de solubilización 138.

- 5 Para entender las propiedades de la contrapartida de solubilización 138 que la hacen efectiva para la solubilización de N-inteína durante la expresión en la *E. coli*, los parámetros de proteína calculados para cada una de las solubilizaciones candidatas en la tabla 3 se correlacionaron con el título soluble en la figura 4. Si bien ninguno de estos parámetros se correlacionó fuertemente con la expresión general, tanto los valores de AI como de GRAVY mostraron una correlación negativa con el título soluble.

10 Ejemplo 7. Selección de aminoácidos para el reemplazo de residuos de cisteína en NINTΔA_CC

La N-inteína GP41-1 aislada de fuentes naturales contiene tres residuos de cisteína, pero uno había sido reemplazado previamente para obtener NINTΔA_CC, la construcción madre para esta invención. Los dos residuos de cisteína restantes contenidos en NINTΔA_CC se seleccionaron para su reemplazo, de modo que se pudiera introducir un único residuo de cisteína reactivo en el dominio de solubilización, para su posterior inmovilización u otra modificación.

- 15 Para identificar los aminoácidos que pudieran sustituirse por los dos residuos de cisteína en NINTA_CC y aún producir una proteína de inteína estable y funcional, se realizaron varios análisis filogenéticos, en los que se alinearon secuencias de proteínas y variantes de aminoácidos naturales en las posiciones 65 y 89 en la SEQ ID NO: 1 fueron examinados. Sería de esperar que el reemplazo de las cisteínas internas que se presentan naturalmente en GP41-1 con otros aminoácidos que se encuentran en estas posiciones en inteínas similares produzca proteínas variantes de GP41-1 funcionales y/o estables, ya que la selección natural ha permitido que estas variantes persistan en la naturaleza. Cuando dicho análisis se realiza con N-inteínas de la clase de inteína GP41 (1, 2, 3, 4, 5, 6; Dassa B., *et al.*, *Nucl. Acids Res.*, 37: 2560-2573 (2009)), se encuentra que los dos residuos de cisteína en las posiciones 65 y 89 en la SEQ ID NO: 1 están altamente conservados, lo que sugiere que la sustitución de estas cisteínas afectaría adversamente la actividad y/o la estabilidad de la inteína GP41-1. Sin embargo, si el análisis se amplía para incluir proteínas ligeramente más divergentes, que pueden tener o no función de inteína, se identifican muchas proteínas que tienen homología con el gen *phoH* del regulón de fosfato de la *E. coli*. Se obtuvieron aproximadamente cien homólogos de GenBank, usando la herramienta de búsqueda BLAST y se alinearon con el algoritmo CLUSTAL, aplicando la herramienta gratuita, BioEdit (Hall TA., *Nucl. Acids. Symp. Ser.* 41:95, 1999). Los resultados de este análisis se muestran en la figura 5, donde la numeración de la posición se basa en NINTΔA_CC (SEQ ID NO: 2). A partir de este análisis, queda claro que la treonina y la alanina se producen con frecuencia en la posición 65 y que la lisina, la metionina y las asparaginas se producen con frecuencia en la posición 89, lo que indica que la sustitución de las cisteínas naturales con estos aminoácidos naturales debería producir una proteína estable.

Ejemplo 8. Selección de las variantes de aminoácidos de NINTΔA_CC para propiedades óptimas

- 35 Se crearon, expresaron, purificaron y caracterizaron construcciones basadas en NINTΔA_CC (SEQ ID NO: 2), que contenían las sustituciones de aminoácidos que se muestran en la tabla 4, como se describió con anterioridad.

- 40 Las mediciones de la tasa de escisión para las proteínas de fusión de N-inteína preparadas a partir de cada construcción se detallan en la figura 6. La NINTΔA_CC madre (+ cnt) se muestra a la izquierda para establecer una comparación. Los aminoácidos en las posiciones 65 y 89 se muestran en la parte inferior de la figura. Un residuo de treonina en la posición 65 produce fusiones de N-inteína que tienen significativamente más actividad que la progenitora. De las construcciones probadas, la fusión de N-inteína que tiene una treonina en la posición 65 y una metionina en la posición 89 proporcionó una construcción con una velocidad catalítica aproximadamente tres veces mayor que la construcción original.

Tabla 4. Variantes de N-inteína GP41-1

Variante de N-inteína GP41-1	Secuencia
NINTΔA_CC (SEQ ID NO:2)	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLSNTGYNEVLNVFPKSKKK SYKITLEDGKEII <u>C</u> SEEHLPQTQTGEMNISGGLKEGM <u>C</u> LYVKEgg
NINTΔA_AC (SEQ ID NO:3)	

Variante de N-inteína GP41-1	Secuencia
	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLSNTGYNEVLNVFPKSKKK SYKITLEDGKEII <u>A</u> SEEHLFPTQTGEMNISGGLKEGM <u>C</u> LYVKEgg
NINTΔA_ CK (SEQ ID NO:4)	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLSNTGYNEVLNVFPKSKKK SYKITLEDGKEII <u>C</u> SEEHLFPTQTGEMNISGGLKEGM <u>K</u> LYVKEgg
NINTΔA_ AM (SEQ ID NO:5)	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLSNTGYNEVLNVFPKSKKK SYKITLEDGKEII <u>A</u> SEEHLFPTQTGEMNISGGLKEGM <u>M</u> LYVKEgg
NINTΔA_ TM (SEQ ID NO:6)	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLSNTGYNEVLNVFPKSKKK SYKITLEDGKEII <u>T</u> SEEHLFPTQTGEMNISGGLKEGM <u>M</u> LYVKEgg
NINTΔA_ AK (SEQ ID NO:7)	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLSNTGYNEVLNVFPKSKKK SYKITLEDGKEII <u>A</u> SEEHLFPTQTGEMNISGGLKEGM <u>K</u> LYVKEgg
NINTΔA_ TK (SEQ ID NO:8)	mtrsgyALDLKTQVQTPQGMKEISNIQVGDVLSNTGYNEVLNVFPKSKKK SYKITLEDGKEII <u>T</u> SEEHLFPTQTGEMNISGGLKEGM <u>K</u> LYVKEgg
Para las SEQ ID Nos.: 2-8, las secuencias que no son inteínas e indican en minúscula y las secuencias de inteínas se indican en mayúsculas.	

Ejemplo 9. Estrategia para la introducción de residuos únicos de cisteína en la contrapartida de solubilización 138

Para permitir la modificación química de la proteína de fusión N-inteína sin disminuir su actividad catalítica, el sitio de la probable modificación debería estar lo más alejado posible del sitio activo de la N-inteína. En ausencia de información estructural para la contrapartida de solubilización 138, un enfoque razonable sería realizar un análisis filogenético, como se describió con anterioridad para la N-inteína GP41-1 (véase el ejemplo 7), determinar las regiones de la proteína que muestran una alta variabilidad, modificar estas por la inserción de una cisteína, y luego probar todas las construcciones resultantes. Sin embargo, hay una estructura de solución de RMN disponible para la contrapartida de solubilización 138 (estructura 1RYK de banco de datos de proteínas), que se muestra en la figura 7. La proteína contiene cuatro dominios de hélice alfa, es globular, tiene una larga bobina no estructurada, que forma la conexión al terminal carboxi de la N-inteína (región en el círculo; no se muestra la N-inteína). Las regiones de bucle GKL y GYQ indicadas por el resaltado amarillo se dirigieron para las inserciones de los residuos de la cisteína para crear las nuevas versiones (GCKL (SEQ ID NO: 61), GCYQ (SEQ ID NO: 62) y GCGYQ (SEQ ID NO: 63)) de la contrapartida de solubilización 138 (138_GKL22GCKL (SEQ ID NO: 16), 138_GYQ48GCRYQ (SEQ ID NO: 17) y 138_GYQ48GCGYQ (SEQ ID NO: 18)).

Ejemplo 10. Acoplamiento de la proteína de fusión N-inteína (ligando) a la resina de cromatografía

Una proteína de fusión soluble que contiene la contrapartida de solubilización 138_GYQ48GCGYQ (SEQ ID NO: 18) fusionada al término carboxilo de la variante de GP41-1 NINTΔA_TM (SEQ ID NO: 6) se expresa a partir de un ácido nucleico codificador en la *E. coli* y posteriormente se separa de proteínas celulares contaminantes utilizando métodos de separación convencionales.

- 5 La proteína de fusión N-inteína purificada se acopla luego a una resina de cromatografía FRACTOGEL® o ESHMUNO® (EMD Millipore Corporation), a través del único sitio de cisteína reactiva en el dominio asociado de solubilización de la proteína de fusión utilizando técnicas estándar.

En la preparación para la activación, 5 ml de resina húmeda de FRACTOGEL® COO (FG-COO) se lavan una vez con agua desionizada y tres veces con ácido 2-(N-morfolino)etansulfónico 150 mM, a pH 6,5 (tampón MES), en un embudo Buechner y se transfieren a una botella de vidrio Schott. Se disuelven 0,1035 g de 1-etil-3-(3-dimetilamino-propil)carbodiimida (EDC) en 3 ml de tampón MES y se agregan a FG-COO. La mezcla se incuba durante 2 minutos a temperatura ambiente. Se agrega una solución de 0,1372 g de ácido trifluoroacético de N-(3-aminopropil)maleimida (APM), en 4 ml de tampón MES y la mezcla se mantiene a temperatura ambiente durante la noche con agitación. El pH se mantuvo a 6,5, a través de la titulación con NaOH 1M. Para el almacenamiento, la resina activada se resuspende en NaCl 150 mM, que contiene etanol al 20 % y se almacena en un refrigerador. Para el análisis de la funcionalización, se prepara una solución al 50 % v/v de resina activada en tampón fosfato 100 mM que contiene NaCl 150 mM, a pH = 7,2 (tampón PO). Se mezclan 0,5 ml de FG-COO activado con 1 ml de solución de clorhidrato de cisteína 204 uM, en tampón PO y se incuba durante una hora. Una muestra de FG-COO que no se ha activado con AMP se procesa en paralelo, como control negativo. La resina se lava luego extensivamente con tampón PO, NaCl 0,5 M y se resuspende en NaCl 0,5 M, para el análisis de grupos sulfhidrilo libres usando reactivo de Ellmann (ácido 5,5'-ditio-bis- (ácido 2-nitrobenzoico)) y métodos conocidos. Con este análisis se determina una densidad de ligando de hasta 400 µmol por gramo de resina seca.

Ejemplo 11. Purificación por afinidad de una tiorredoxina usando proteínas de fusión de inteína

La resina que contiene la proteína de fusión N-inteína inmovilizada, preparada de acuerdo con el ejemplo 10 de este documento se empaqueta en una columna de cromatografía estándar, y una mezcla de proteína sin procesar, que contiene la proteína de fusión CINT_TRX (SEQ ID NO: 10), que incluye la tiorredoxina de la molécula diana fusionada al término carboxi de la C-inteína GP41-1, se agrega a la columna que contiene la proteína de fusión N-inteína inmovilizada, a una temperatura variable en el rango de 4-25 °C, y usando un tampón de carga que contiene Tris-HCl 100 mM, NaCl 25 mM, cloruro de zinc 0,1 mM, pH = 9, para permitir una fuerte interacción entre los dominios de la N-inteína y C-inteína GP41-1, sin permitir la catálisis de la inteína.

La columna cargada se lava luego para eliminar los contaminantes no unidos y débilmente unidos, utilizando un tampón de lavado que contiene detergente (por ejemplo, Triton X100, ND40) o sal (por ejemplo, acetato, fosfato, cloruro, sales de sulfato de sodio, amonio o potasio).

La escisión y elución de la porción de tiorredoxina de la proteína de fusión C-inteína se realiza mediante la adición de un tampón de escisión (Tris-HCl 50 mM, pH = 7,0, NaCl 300 mM, EDTA 1 mM). La tiorredoxina escindida se recupera luego en el eluato.

Tabla 5. Inteínas divididas ejemplares y sus secuencias.

Inteína	SEQ ID NO:	Secuencia
Dominio N-terminal de GP41.1	29	<u>CLDLKTQVQT PQGMKEISNI QVGDVLVSNT</u> GYNEVLNVFP KSKKKSYKIT LEDGKEIICS EEHLFPTQTG EMNISGGLKE GMCLYVKE
Dominio N-terminal de GP41.8	30	<u>CLSLDTMVVT NGKAIEIRDV KVGDWLESEC</u> GPVQVTEVLP IIKQPVFIEIV LKSGKKIRVS ANHKFPTKDG LKTINSGLKV GDFLRSRA
Dominio N-terminal de NrdJ1	31	

Inteína	SEQ ID NO:	Secuencia
		<u>CLVGSSEIIT</u> RNYGKTTIKE VVEIFDNDKN IQVLAFNTHT DNIEWAPIKA AQLTRPNAEL VELEINTLHG VKTIRCTPDH PVYTKNRDYY RADELTDDE LVVAI
Dominio N-terminal de IMPDH1	32	<u>CFVPGTLVNT</u> ENGLKKIEEI KVGDKVFSHT GKLQEVVDTL IFDRDEEIS INGIDCTKNH EFYVIDKENA NRVNEDNIHL FARWVHAEEL DMKKHLLIEL E
Dominio N-terminal de NrdA-2	33	<u>CLTGDAKIDV</u> LIDNIPISQI SLEEVVNLFN EGKEIYVLSY NIDTKEVEYK EISDAGLISE SAEVLEIIDE ETGQKIVCTP DHKVYTLNRG YVSAKDLKED DELVFS
Dominio N-terminal de Npu DNA-E (código de acceso del Genbank ZP_00111398)	34	<u>CLSYETEILT</u> VEYGLLPYGK IVEKRIECTV YSVDNNGNIY TQPVAQWHDR GEQEVFEYCL EDGSLIRATK DHKFMTVDGQ MLPIDEIFER ELDLMRVDNL PN
Dominio N-terminal de Ssp DNA-B (código de acceso del Genbank Q55418)	35	<u>CISGDSLISL</u> ASTGKRVSIGK DLLDEKDFEI WAINEQTMKL ESAKVSFVFC TGKKLVYILK TRLGRTIKAT ANHRFLTIDG WKRLDELSLK EHIALPRKLE SSSLQ
Dominio C-terminal de GP41.1	9	MMLKKILKIE ELDERELIDI EVSGNHLFYA <u>NDILTHN</u>
Dominio C-terminal de GP41.8	36	MCEIFENEID WDEIASIEYV GVEETIDINV TNDRLFF <u>ANG ILTHN</u>
Dominio C-terminal de NrdJ1	37	

Inteína	SEQ ID NO:	Secuencia
		MEAKTYIGKL KSRKIVSNED TYDIQTSTHN FF <u>ANDILVHN</u>
Dominio C-terminal de IMPDH1	38	MKFKLKEITS IETKHYKGKV HDLTVNQDHS YN <u>VRGTVVHN</u>
Dominio C-terminal de NrdA-2	39	MGLKIIKRES KEPVFDITVK DNSNFF <u>FANNI</u> <u>LVHN</u>
Dominio C-terminal de Npu DNA-E (código de acceso del Genbank ZP_ 00111398)	40	MIKIATRKYL GKQNVYDIGV ERDHNFALK <u>N</u> <u>GFIASN</u>
Dominio C-terminal de Ssp DNA-B (código de acceso del Genbank Q55418)	41	SPEIEKLSQS DIYWDSIVSI TETGVVEEVF DLTVPGPHNF V <u>ANDIIVHN</u>

Las secuencias subrayadas corresponden a las cajas N1 de los dominios N-terminales de la inteína. Las secuencias subrayadas dobles corresponden a las cajas C1 de los dominios C-terminales de la inteína (por ejemplo, que carecen del primer aminoácido de la exteína).

5 Secuencias adicionales de N-inteína ejemplares

gp 41-2

CLDLKTQVQTQQGLKDISNIQVGDLVL (SEQ ID NO:42)

gp 41-3

CLDLKTQVQTQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKS (SEQ ID NO:43)

10

gp 41-4

CLDLKTQVQTQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKS**SYKITLEDGKEII**

CSEEHLFPTQTGEMNISGGLKEGMCLYVKE (SEQ ID NO:44)

gp 41-5

CLDLKTQVQTQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKS**SYKITLEDGKEII**

CSEEHLFPTQTGEMNISGGLKEGMCLYVKE (SEQ ID NO:45)

gp 41-6

15

SYKITLEDGKEIICSEEHLFPTQNGEVNIKGLKEGMCLYVKE (SEQ ID NO:46)

gp 41-7

MMLKKILKIEELDERELIDIEVSGNH (SEQ ID NO:47)

NrdA-1

CVAGDTKIKIKYPESVGDQYGTWYWNVLEKEIQIEDLEDYIIMRECEIYDSNAPQIEV
LSYNIETGEQEWKPITAFATSPKAKVMKITDEESGKSIVVTPEHQVFTKNRGYVMA
KDLIETDEPIIVNKDMNF (SEQ ID NO:48)

NrdA-4

CLAGDTTTVTVLEGDIVFEMTLENLVSLYKNVFSVSVLSFNPETQKQEFKPVNAALM
NPESKVLKITDSDTGKSIVCTPDHKVFTKNRGYVIASELNAEDILEIK (SEQ ID NO:49)

5 NrdA-5

HTETVRRVGTITAFATSPKSKVMKITDEESGNSIVVTPEHKVFTKNRGYVMAKNLV
ETDELVIN (SEQ IDNO:50)

NrdA-6

YVCSRDDTTGFKLICTPDHMIYTKNRGYIMAKYLKEDDELLINEIHLPT (SEQ ID NO:51)

NrdJ-1

10 CLVGSSEIITRNYGKTTIKEVVEIFDNDKNIQVLA FNTHTDNIEWAPIKAAQLTRPNAE
LVELEIDTLHG VKTIRCTPDHPVYTKNRGYVRADELTDDEL VVAI (SEQ ID NO:52)

NrdJ-2

CLVGSSEIITRNYGKTTIKEVVEIFDNDKNIQVLA FNTHTDNIEWAPIKAAQLTRPNAE
LVELEINTLHG VKTIRCTPDHPVYTKNRD YVRADELTDDEL VVAI (SEQ ID NO:53)

Secuencias adicionales de CN-inteína ejemplares

gp 41-9

15 MIMKNRERFITEKILNIEEIDDDLTVDIGMDNEDHYFVANDILTHNT (SEQ ID NO:54)

IMPDH-2

MKFTLEPITKIDSYEVTAEPVYDIEVENDHSFCV eNGFVVHNS (SEQ ID NO:55)

IMPDH-3

MKFKLVEITSKETFNYSQG-VHDLTVEDDHSYSI-NNIVVHNS (SEQ ID NO:56),

20 NrdA-3

MLKIEYLEEEIPVYDITVEETHNFFANDILIHNC (SEQ ID NO:57),

NrdA-5

MLKIEYLEEEIPVYDITVEGTHNLAYSL (SEQ ID NO:58),

NrdA-6

25 MGIKIRKLEQNRVYDIKVEKIIIFCANNILVHNC (SEQ ID NO:59), and

NrdJ-1

MEAKTYIGKLKSRKIVSNEDTYDIQTSTHNFFANDILVHNS (SEQ ID NO:60).

Las enseñanzas relevantes de todas las patentes, solicitudes publicadas y referencias citadas en este documento se incorporan como referencia en su totalidad.

30 A menos que se indique lo contrario, todos los números que expresan cantidades de ingredientes, condiciones de expresión, condiciones de tratamiento, etc., utilizados en la memoria descriptiva, incluidas las reivindicaciones, deben entenderse como modificados en todos los casos por expresiones tales como "aproximadamente/alrededor

de". Por consiguiente, a menos que se indique lo contrario, los parámetros numéricos son aproximaciones y pueden variar dependiendo de las propiedades deseadas que se pretenden obtener mediante la presente invención. Salvo que se indique lo contrario, la frase "al menos" que precede a una serie de elementos debe entenderse como que se refiere a cada elemento de la serie. Los expertos en la materia reconocerán, o podrán determinar recurriendo a la experimentación que no exceda la de la práctica rutinaria, muchos equivalentes a las realizaciones específicas de la invención descritas en este documento. Dichos equivalentes quedan incluidos en las siguientes reivindicaciones.

5

Aunque esta invención se ha mostrado y descrito particularmente con referencias a realizaciones de ejemplo de la misma, los expertos en la materia entenderán que se pueden realizar diversos cambios en la forma y detalles sin apartarse del alcance de la invención abarcada por las reivindicaciones adjuntas.

10

LISTA DE SECUENCIAS

- <110> Merck Patent GmbH
- <120> Proteínas de fusión de inteína, solubles y métodos para la purificación de las biomoléculas
- 5 <130> N410412EP
- <140> EP 15790411,1
- <141> 2015-10-23
- <150> US 62/074,494
- <151> 2014-11-03
- 10 <150> US 62/209,010
- <151> 2015-08-24
- <160> 63
- <170> FastSEQ para Windows, Versión 4.0
- <210> 1
- 15 <211> 96
- <212> PRT
- <213> Secuencia artificial
- <220>
- <223> N-inteína GP41-1 con secuencias que no son de inteínas a los flancos
- 20 <400> 1
- | | | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Thr | Arg | Ser | Gly | Tyr | Cys | Leu | Asp | Leu | Lys | Thr | Gln | Val | Gln | Thr |
| 1 | | | | 5 | | | | | 10 | | | | | 15 | |
| Pro | Gln | Gly | Met | Lys | Glu | Ile | Ser | Asn | Ile | Gln | Val | Gly | Asp | Leu | Val |
| | | | 20 | | | | | 25 | | | | | 30 | | |
| Leu | Ser | Asn | Thr | Gly | Tyr | Asn | Glu | Val | Leu | Asn | Val | Phe | Pro | Lys | Ser |
| | | 35 | | | | | 40 | | | | | 45 | | | |
| Lys | Lys | Lys | Ser | Tyr | Lys | Ile | Thr | Leu | Glu | Asp | Gly | Lys | Glu | Ile | Ile |
| | 50 | | | | | 55 | | | | | 60 | | | | |
| Cys | Ser | Glu | Glu | His | Leu | Phe | Pro | Thr | Gln | Thr | Gly | Glu | Met | Asn | Ile |
| 65 | | | | | 70 | | | | | 75 | | | | 80 | |
| Ser | Gly | Gly | Leu | Lys | Glu | Gly | Met | Cys | Leu | Tyr | Val | Lys | Glu | Gly | Gly |
| | | | | 85 | | | | | 90 | | | | | 95 | |
- <210> 2
- <211> 96
- <212> PRT
- 25 <213> Secuencia artificial
- <220>
- <223> Variante de N-inteína GP41-1 con secuencias que no son de inteínas a los flancos
- <400> 2

Met	Thr	Arg	Ser	Gly	Tyr	Ala	Leu	Asp	Leu	Lys	Thr	Gln	Val	Gln	Thr
1				5					10					15	
Pro	Gln	Gly	Met	Lys	Glu	Ile	Ser	Asn	Ile	Gln	Val	Gly	Asp	Leu	Val
			20					25					30		
Leu	Ser	Asn	Thr	Gly	Tyr	Asn	Glu	Val	Leu	Asn	Val	Phe	Pro	Lys	Ser
		35					40					45			

Lys	Lys	Lys	Ser	Tyr	Lys	Ile	Thr	Leu	Glu	Asp	Gly	Lys	Glu	Ile	Ile
	50					55					60				
Cys	Ser	Glu	Glu	His	Leu	Phe	Pro	Thr	Gln	Thr	Gly	Glu	Met	Asn	Ile
65					70					75				80	
Ser	Gly	Gly	Leu	Lys	Glu	Gly	Met	Cys	Leu	Tyr	Val	Lys	Glu	Gly	Gly
				85					90					95	

<210> 3

5 <211> 96

<212> PRT

<213> Secuencia artificial

<220>

10 <223> Variante de N-inteína GP41-1 con secuencias que no son de inteínas a los flancos

<400> 3

Met	Thr	Arg	Ser	Gly	Tyr	Ala	Leu	Asp	Leu	Lys	Thr	Gln	Val	Gln	Thr
1				5					10					15	
Pro	Gln	Gly	Met	Lys	Glu	Ile	Ser	Asn	Ile	Gln	Val	Gly	Asp	Leu	Val
			20					25					30		
Leu	Ser	Asn	Thr	Gly	Tyr	Asn	Glu	Val	Leu	Asn	Val	Phe	Pro	Lys	Ser
		35					40					45			
Lys	Lys	Lys	Ser	Tyr	Lys	Ile	Thr	Leu	Glu	Asp	Gly	Lys	Glu	Ile	Ile
	50					55					60				
Ala	Ser	Glu	Glu	His	Leu	Phe	Pro	Thr	Gln	Thr	Gly	Glu	Met	Asn	Ile
65					70					75				80	
Ser	Gly	Gly	Leu	Lys	Glu	Gly	Met	Cys	Leu	Tyr	Val	Lys	Glu	Gly	Gly
				85					90					95	

<210> 4

<211> 96

15 <212> PRT

<213> Secuencia artificial

<220>

<223> Variante de N-inteína GP41-1 con secuencias que no son de inteínas a los flancos

20 <400> 4

Met	Thr	Arg	Ser	Gly	Tyr	Ala	Leu	Asp	Leu	Lys	Thr	Gln	Val	Gln	Thr
1				5					10					15	
Pro	Gln	Gly	Met	Lys	Glu	Ile	Ser	Asn	Ile	Gln	Val	Gly	Asp	Leu	Val
			20					25					30		
Leu	Ser	Asn	Thr	Gly	Tyr	Asn	Glu	Val	Leu	Asn	Val	Phe	Pro	Lys	Ser
		35					40					45			
Lys	Lys	Lys	Ser	Tyr	Lys	Ile	Thr	Leu	Glu	Asp	Gly	Lys	Glu	Ile	Ile
	50					55					60				
Cys	Ser	Glu	Glu	His	Leu	Phe	Pro	Thr	Gln	Thr	Gly	Glu	Met	Asn	Ile
65					70					75				80	
Ser	Gly	Gly	Leu	Lys	Glu	Gly	Met	Lys	Leu	Tyr	Val	Lys	Glu	Gly	Gly
				85					90					95	

<210> 5

<211> 96

<212> PRT

5 <213> Secuencia artificial

<220>

<223> Variante de N-inteína GP41-1 con secuencias que no son de inteínas a los flancos

<400> 5

Met	Thr	Arg	Ser	Gly	Tyr	Ala	Leu	Asp	Leu	Lys	Thr	Gln	Val	Gln	Thr
1				5					10					15	
Pro	Gln	Gly	Met	Lys	Glu	Ile	Ser	Asn	Ile	Gln	Val	Gly	Asp	Leu	Val
			20					25					30		
Leu	Ser	Asn	Thr	Gly	Tyr	Asn	Glu	Val	Leu	Asn	Val	Phe	Pro	Lys	Ser
		35					40					45			
Lys	Lys	Lys	Ser	Tyr	Lys	Ile	Thr	Leu	Glu	Asp	Gly	Lys	Glu	Ile	Ile
	50					55					60				
Ala	Ser	Glu	Glu	His	Leu	Phe	Pro	Thr	Gln	Thr	Gly	Glu	Met	Asn	Ile
65					70					75				80	
Ser	Gly	Gly	Leu	Lys	Glu	Gly	Met	Met	Leu	Tyr	Val	Lys	Glu	Gly	Gly
				85					90					95	

<210> 6

<211> 96

<212> PRT

<213> Secuencia artificial

15 <220>

<223> Variante de N-inteína GP41-1 con secuencias que no son de inteínas a los flancos

<400> 6

Met	Thr	Arg	Ser	Gly	Tyr	Ala	Leu	Asp	Leu	Lys	Thr	Gln	Val	Gln	Thr
1				5					10					15	
Pro	Gln	Gly	Met	Lys	Glu	Ile	Ser	Asn	Ile	Gln	Val	Gly	Asp	Leu	Val
			20					25					30		
Leu	Ser	Asn	Thr	Gly	Tyr	Asn	Glu	Val	Leu	Asn	Val	Phe	Pro	Lys	Ser
		35					40					45			
Lys	Lys	Lys	Ser	Tyr	Lys	Ile	Thr	Leu	Glu	Asp	Gly	Lys	Glu	Ile	Ile
	50					55					60				
Thr	Ser	Glu	Glu	His	Leu	Phe	Pro	Thr	Gln	Thr	Gly	Glu	Met	Asn	Ile
65					70					75				80	
Ser	Gly	Gly	Leu	Lys	Glu	Gly	Met	Met	Leu	Tyr	Val	Lys	Glu	Gly	Gly
				85					90					95	

<210> 7

20

<211> 96

<212> PRT

<213> Secuencia artificial

<220>

5 <223> Variante de N-inteína GP41-1 con secuencias que no son de inteínas a los flancos

<400> 7

Met	Thr	Arg	Ser	Gly	Tyr	Ala	Leu	Asp	Leu	Lys	Thr	Gln	Val	Gln	Thr
1				5					10					15	
Pro	Gln	Gly	Met	Lys	Glu	Ile	Ser	Asn	Ile	Gln	Val	Gly	Asp	Leu	Val
			20					25					30		
Leu	Ser	Asn	Thr	Gly	Tyr	Asn	Glu	Val	Leu	Asn	Val	Phe	Pro	Lys	Ser
		35					40					45			
Lys	Lys	Lys	Ser	Tyr	Lys	Ile	Thr	Leu	Glu	Asp	Gly	Lys	Glu	Ile	Ile
	50					55					60				
Ala	Ser	Glu	Glu	His	Leu	Phe	Pro	Thr	Gln	Thr	Gly	Glu	Met	Asn	Ile
65					70					75				80	
Ser	Gly	Gly	Leu	Lys	Glu	Gly	Met	Lys	Leu	Tyr	Val	Lys	Glu	Gly	Gly
				85					90					95	

<210> 8

10 <211> 96

<212> PRT

<213> Secuencia artificial

<220>

15 <223> Variante de N-inteína GP41-1 con secuencias que no son de inteínas a los flancos

<400> 8

Met	Thr	Arg	Ser	Gly	Tyr	Ala	Leu	Asp	Leu	Lys	Thr	Gln	Val	Gln	Thr
1				5					10					15	
Pro	Gln	Gly	Met	Lys	Glu	Ile	Ser	Asn	Ile	Gln	Val	Gly	Asp	Leu	Val
			20					25					30		
Leu	Ser	Asn	Thr	Gly	Tyr	Asn	Glu	Val	Leu	Asn	Val	Phe	Pro	Lys	Ser
		35					40					45			
Lys	Lys	Lys	Ser	Tyr	Lys	Ile	Thr	Leu	Glu	Asp	Gly	Lys	Glu	Ile	Ile
	50					55					60				
Thr	Ser	Glu	Glu	His	Leu	Phe	Pro	Thr	Gln	Thr	Gly	Glu	Met	Asn	Ile
65					70					75				80	
Ser	Gly	Gly	Leu	Lys	Glu	Gly	Met	Lys	Leu	Tyr	Val	Lys	Glu	Gly	Gly
				85					90					95	

<210> 9

<211> 42

20 <212> PRT

<213> cianófago

<400> 9

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						
		35					40								

<210> 10

<211> 157

<212> PRT

<213> Secuencia artificial

5 <220>

<223> GP41-1 C-inteína-proteína de fusión de tiorredoxinas

<400> 10

```

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 Met 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 Glu 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 Gln 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 His His His His His
145      150      155

```

10

<210> 11

<211> 51

<212> PRT

<213> *E. coli*

15 <400> 11

```

Met Arg Glu Tyr Pro Asn Gly Glu Lys Thr His Leu Thr Val Met Ala
 1      5      10      15
Ala Gly Phe Pro Ser Leu Thr Gly Asp His Lys Val Ile Tyr Val Ala
 20      25      30
Ala Asp Arg His Val Thr Ser Glu Glu Ile Leu Glu Ala Ala Ile Arg
 35      40      45
Leu Leu Ser
 50

```

<210> 12

<211> 77

<212> PRT

20 <213> *E. coli*

<400> 12

ES 2 742 199 T3

Met Ser His Leu Asp Glu Val Ile Ala Arg Val Asp Ala Ala Ile Glu
 1 5 10 15
 Glu Ser Val Ile Ala His Met Asn Glu Leu Leu Ile Ala Leu Ser Asp
 20 25 30
 Asp Ala Glu Leu Ser Arg Glu Asp Arg Tyr Thr Gln Gln Gln Arg Leu
 35 40 45
 Arg Thr Ala Ile Ala His His Gly Arg Lys His Lys Glu Asp Met Glu
 50 55 60
 Ala Arg His Glu Gln Leu Thr Lys Gly Gly Thr Ile Leu
 65 70 75

<210> 13

<211> 83

<212> PRT

5 <213> *E. coli*

<400> 13

Met Asn Lys Glu Thr Gln Pro Ile Asp Arg Glu Thr Leu Leu Lys Glu
 1 5 10 15
 Ala Asn Lys Ile Ile Arg Glu His Glu Asp Thr Leu Ala Gly Ile Glu
 20 25 30
 Ala Thr Gly Val Thr Gln Arg Asn Gly Val Leu Val Phe Thr Gly Asp
 35 40 45
 Tyr Phe Leu Asp Glu Gln Gly Leu Pro Thr Ala Lys Ser Thr Ala Val
 50 55 60
 Phe Asn Met Phe Lys His Leu Ala His Val Leu Ser Glu Lys Tyr His
 65 70 75 80
 Leu Val Asp

<210> 14

<211> 53

10 <212> PRT

<213> *E. coli*

<400> 14

Met Ser Leu Glu Asn Ala Pro Asp Asp Val Lys Leu Ala Val Asp Leu
 1 5 10 15
 Ile Val Leu Leu Glu Glu Asn Gln Ile Pro Ala Ser Thr Val Leu Arg
 20 25 30
 Ala Leu Asp Ile Val Lys Arg Asp Tyr Glu Lys Lys Leu Thr Arg Asp
 35 40 45
 Asp Glu Ala Glu Lys
 50

<210> 15

15 <211> 69

<212> PRT

<213> *E. coli*

<400> 15

ES 2 742 199 T3

Met Asn Lys Asp Glu Ala Gly Gly Asn Trp Lys Gln Phe Lys Gly Lys
 1 5 10 15
 Val Lys Glu Gln Trp Gly Lys Leu Thr Asp Asp Asp Met Thr Ile Ile
 20 25 30
 Glu Gly Lys Arg Asp Gln Leu Val Gly Lys Ile Gln Glu Arg Tyr Gly
 35 40 45
 Tyr Gln Lys Asp Gln Ala Glu Lys Glu Val Val Asp Trp Glu Thr Arg
 50 55 60
 Asn Glu Tyr Arg Trp
 65

<210> 16

<211> 70

<212> PRT

5 <213> *E. coli*

<400> 16

Met Asn Lys Asp Glu Ala Gly Gly Asn Trp Lys Gln Phe Lys Gly Lys
 1 5 10 15
 Val Lys Glu Gln Trp Gly Cys Lys Leu Thr Asp Asp Asp Met Thr Ile
 20 25 30
 Ile Glu Gly Lys Arg Asp Gln Leu Val Gly Lys Ile Gln Glu Arg Tyr
 35 40 45
 Gly Tyr Gln Lys Asp Gln Ala Glu Lys Glu Val Val Asp Trp Glu Thr
 50 55 60
 Arg Asn Glu Tyr Arg Trp
 65 70

<210> 17

<211> 70

10 <212> PRT

<213> *E. coli*

<400> 17

Met Asn Lys Asp Glu Ala Gly Gly Asn Trp Lys Gln Phe Lys Gly Lys
 1 5 10 15
 Val Lys Glu Gln Trp Gly Lys Leu Thr Asp Asp Asp Met Thr Ile Ile
 20 25 30
 Glu Gly Lys Arg Asp Gln Leu Val Gly Lys Ile Gln Glu Arg Tyr Gly
 35 40 45

Cys Tyr Gln Lys Asp Gln Ala Glu Lys Glu Val Val Asp Trp Glu Thr
 50 55 60
 Arg Asn Glu Tyr Arg Trp
 65 70

15 <210> 18

<211> 71

<212> PRT

<213> *E. coli*

20 <400> 18

Met Asn Lys Asp Glu Ala Gly Gly Asn Trp Lys Gln Phe Lys Gly Lys
 1 5 10 15
 Val Lys Glu Gln Trp Gly Lys Leu Thr Asp Asp Asp Met Thr Ile Ile
 20 25 30
 Glu Gly Lys Arg Asp Gln Leu Val Gly Lys Ile Gln Glu Arg Tyr Gly
 35 40 45
 Cys Gly Tyr Gln Lys Asp Gln Ala Glu Lys Glu Val Val Asp Trp Glu
 50 55 60
 Thr Arg Asn Glu Tyr Arg Trp
 65 70

<210> 19

<211> 92

<212> PRT

5 <213> *E. coli*

<400> 19

Met Ile Ala Glu Phe Glu Ser Arg Ile Leu Ala Leu Ile Asp Gly Met
 1 5 10 15
 Val Asp His Ala Ser Asp Asp Glu Leu Phe Ala Ser Gly Tyr Leu Arg
 20 25 30
 Gly His Leu Thr Leu Ala Ile Ala Glu Leu Glu Ser Gly Asp Asp His
 35 40 45
 Ser Ala Gln Ala Val His Thr Thr Val Ser Gln Ser Leu Glu Lys Ala
 50 55 60
 Ile Gly Ala Gly Glu Leu Ser Pro Arg Asp Gln Ala Leu Val Thr Asp
 65 70 75 80
 Met Trp Glu Asn Leu Phe Gln Gln Ala Ser Gln Gln
 85 90

<210> 20

<211> 95

10 <212> PRT

<213> *E. coli*

<400> 20

Met Gln Leu Asn Ile Thr Gly Asn Asn Val Glu Ile Thr Glu Ala Leu
 1 5 10 15
 Arg Glu Phe Val Thr Ala Lys Phe Ala Lys Leu Glu Gln Tyr Phe Asp
 20 25 30
 Arg Ile Asn Gln Val Tyr Val Val Leu Lys Val Glu Lys Val Thr His
 35 40 45
 Thr Ser Asp Ala Thr Leu His Val Asn Gly Gly Glu Ile His Ala Ser
 50 55 60
 Ala Glu Gly Gln Asp Met Tyr Ala Ala Ile Asp Gly Leu Ile Asp Lys
 65 70 75 80
 Leu Ala Arg Gln Leu Thr Lys His Lys Asp Lys Leu Lys Gln His
 85 90 95

<210> 21

15 <211> 192

<212> PRT

<213> *E. coli*

<400> 21

Met Asp Thr Ser Asn Ala Thr Ser Val Val Asn Val Ser Ala Ser Ser
 1 5 10 15
 Ser Thr Ser Thr Ile Tyr Asp Leu Gly Asn Met Ser Lys Asp Glu Val
 20 25 30
 Val Lys Leu Phe Glu Glu Leu Gly Val Phe Gln Ala Ala Ile Leu Met
 35 40 45
 Phe Ser Tyr Met Tyr Gln Ala Gln Ser Asn Leu Ser Ile Ala Lys Phe
 50 55 60
 Ala Asp Met Asn Glu Ala Ser Lys Ala Ser Thr Thr Ala Gln Lys Met
 65 70 75 80
 Ala Asn Leu Val Asp Ala Lys Ile Ala Asp Val Gln Ser Ser Thr Asp
 85 90 95
 Lys Asn Ala Lys Ala Lys Leu Pro Gln Asp Val Ile Asp Tyr Ile Asn
 100 105 110
 Asp Pro Arg Asn Asp Ile Ser Val Thr Gly Ile Ser Asp Leu Ser Gly
 115 120 125
 Asp Leu Ser Ala Gly Asp Leu Gln Thr Val Lys Ala Ala Ile Ser Ala
 130 135 140
 Lys Ala Asn Asn Leu Thr Thr Val Val Asn Asn Ser Gln Leu Glu Ile
 145 150 155 160
 Gln Gln Met Ser Asn Thr Leu Asn Leu Leu Thr Ser Ala Arg Ser Asp
 165 170 175
 Val Gln Ser Leu Gln Tyr Arg Thr Ile Ser Ala Ile Ser Leu Gly Lys
 180 185 190

<210> 22

<211> 68

<212> PRT

5 <213> *Fasciola hepatica*

<400> 22

Met Pro Ser Val Glu Val Glu Lys Leu Leu His Val Leu Asp Arg Asn
 1 5 10 15
 Gly Asp Gly Lys Val Ser Ala Glu Glu Leu Lys Ala Phe Ala Asp Asp
 20 25 30
 Ser Lys Tyr Pro Leu Asp Ser Asn Lys Ile Lys Ala Phe Ile Lys Glu
 35 40 45
 His Asp Lys Asn Lys Asp Gly Lys Leu Asp Leu Lys Glu Leu Val Ser
 50 55 60
 Ile Leu Ser Ser
 65

<210> 23

<211> 11

10 <212> PRT

<213> *Fasciola hepatica*

<400> 23

Met Pro Ser Val Glu Val Glu Lys Leu Leu His
 1 5 10

<210> 24

15 <211> 328

<212> PRT

<213> *E. coli*

<400> 24

```

Met Gly Gln Leu Ile Asp Gly Val Trp His Asp Thr Trp Tyr Asp Thr
 1      5      10      15
Lys Ser Thr Gly Gly Lys Phe Gln Arg Ser Ala Ser Ala Phe Arg Asn
 20      25      30
Trp Leu Thr Ala Asp Gly Ala Pro Gly Pro Thr Gly Lys Gly Gly Phe
 35      40      45
Ala Ala Glu Lys Asp Arg Tyr His Leu Tyr Val Ser Leu Ala Cys Pro
 50      55      60
Trp Ala His Arg Thr Leu Ile Met Arg Lys Leu Lys Gly Leu Glu Pro
 65      70      75      80
Phe Ile Ser Val Ser Val Val Asn Pro Leu Met Leu Glu Asn Gly Trp
 85      90      95
Thr Phe Asp Asp Ser Phe Pro Gly Ala Thr Gly Asp Thr Leu Tyr Gln
 100      105      110
His Glu Phe Leu Tyr Gln Leu Tyr Leu His Ala Asp Pro His Tyr Ser
 115      120      125
Gly Arg Val Thr Val Pro Val Leu Trp Asp Lys Lys Asn His Thr Ile
 130      135      140
Val Ser Asn Glu Ser Ala Glu Ile Ile Arg Met Phe Asn Thr Ala Phe
 145      150      155      160
Asp Ala Leu Gly Ala Lys Ala Gly Asp Tyr Tyr Pro Pro Ala Leu Gln
 165      170      175
Pro Lys Ile Asp Glu Leu Asn Gly Trp Ile Tyr Asp Thr Val Asn Asn
 180      185      190
Gly Val Tyr Lys Ala Gly Phe Ala Thr Ser Gln Gln Ala Tyr Asp Glu
 195      200      205
Ala Val Ala Lys Val Phe Glu Ser Leu Ala Arg Leu Glu Gln Ile Leu
 210      215      220
Gly Gln His Arg Tyr Leu Thr Gly Asn Gln Leu Thr Glu Ala Asp Ile
 225      230      235      240
Arg Leu Trp Thr Thr Leu Val Arg Phe Asp Pro Val Tyr Val Thr His
 245      250      255
Phe Lys Cys Asp Lys His Arg Ile Ser Asp Tyr Leu Asn Leu Tyr Gly
 260      265      270
Phe Leu Arg Asp Ile Tyr Gln Met Pro Gly Ile Ala Glu Thr Val Asn
 275      280      285
Phe Asp His Ile Arg Asn His Tyr Phe Arg Ser His Lys Thr Ile Asn
 290      295      300
Pro Thr Gly Ile Ile Ser Ile Gly Pro Trp Gln Asp Leu Asp Glu Pro
 305      310      315      320
His Gly Arg Asp Val Arg Phe Gly
 325

```

<210> 25

<211> 120

<212> PRT

5 <213> Fago lambda de la enterobacteria

<400> 25

```

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

```

ES 2 742 199 T3

```

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
      115          120

```

<210> 26

<211> 396

<212> PRT

5

<213> *E. coli*

<400> 26

```

Met Lys Ile Lys Thr Gly Ala Arg Ile Leu Ala Leu Ser Ala Leu Thr
 1          5          10          15
Thr Met Met Phe Ser Ala Ser Ala Leu Ala Lys Ile Glu Glu Gly Lys
      20          25          30
Leu Val Ile Trp Ile Asn Gly Asp Lys Gly Tyr Asn Gly Leu Ala Glu
      35          40          45
Val Gly Lys Lys Phe Glu Lys Asp Thr Gly Ile Lys Val Thr Val Glu
      50          55          60
His Pro Asp Lys Leu Glu Glu Lys Phe Pro Gln Val Ala Ala Thr Gly
      65          70          75          80
Asp Gly Pro Asp Ile Ile Phe Trp Ala His Asp Arg Phe Gly Gly Tyr
      85          90          95
Ala Gln Ser Gly Leu Leu Ala Glu Ile Thr Pro Asp Lys Ala Phe Gln
      100          105          110
Asp Lys Leu Tyr Pro Phe Thr Trp Asp Ala Val Arg Tyr Asn Gly Lys
      115          120          125
Leu Ile Ala Tyr Pro Ile Ala Val Glu Ala Leu Ser Leu Ile Tyr Asn
      130          135          140
Lys Asp Leu Leu Pro Asn Pro Pro Lys Thr Trp Glu Glu Ile Pro Ala
      145          150          155          160
Leu Asp Lys Glu Leu Lys Ala Lys Gly Lys Ser Ala Leu Met Phe Asn
      165          170          175          180
Leu Gln Glu Pro Tyr Phe Thr Trp Pro Leu Ile Ala Ala Asp Gly Gly
      180          185          190          195
Tyr Ala Phe Lys Tyr Glu Asn Gly Lys Tyr Asp Ile Lys Asp Val Gly
      195          200          205          210
Val Asp Asn Ala Gly Ala Lys Ala Gly Leu Thr Phe Leu Val Asp Leu
      210          215          220          225
Ile Lys Asn Lys His Met Asn Ala Asp Thr Asp Tyr Ser Ile Ala Glu
      225          230          235          240
Ala Ala Phe Asn Lys Gly Glu Thr Ala Met Thr Ile Asn Gly Pro Trp
      245          250          255          260
Ala Trp Ser Asn Ile Asp Thr Ser Lys Val Asn Tyr Gly Val Thr Val
      260          265          270          275
Leu Pro Thr Phe Lys Gly Gln Pro Ser Lys Pro Phe Val Gly Val Leu
      275          280          285          290
Ser Ala Gly Ile Asn Ala Ala Ser Pro Asn Lys Glu Leu Ala Lys Glu
      290          295          300          305
Phe Leu Glu Asn Tyr Leu Leu Thr Asp Glu Gly Leu Glu Ala Val Asn
      305          310          315          320
Lys Asp Lys Pro Leu Gly Ala Val Ala Leu Lys Ser Tyr Glu Glu Glu
      325          330          335          340
Leu Ala Lys Asp Pro Arg Ile Ala Ala Thr Met Glu Asn Ala Gln Lys
      340          345          350          355
Gly Glu Ile Met Pro Asn Ile Pro Gln Met Ser Ala Phe Trp Tyr Ala
      355          360          365          370
Val Arg Thr Ala Val Ile Asn Ala Ala Ser Gly Arg Gln Thr Val Asp
      370          375          380

```

```

Glu Ala Leu Lys Asp Ala Gln Thr Arg Ile Thr Lys
385          390          395

```

<210> 27

<211> 109

<212> PRT

<213> *E. coli*

5

<400> 27

```

Met 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 Glu 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 Gln 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
          100          105

```

<210> 28

<211> 495

<212> PRT

<213> *E. coli*

10

<400> 28

```

Met Asn Lys Glu Ile Leu Ala Val Val Glu Ala Val Ser Asn Glu Lys
 1          5          10          15
Ala Leu Pro Arg Glu Lys Ile Phe Glu Ala Leu Glu Ser Ala Leu Ala
          20          25          30
Thr Ala Thr Lys Lys Lys Tyr Glu Gln Glu Ile Asp Val Arg Val Gln
 35          40          45
Ile Asp Arg Lys Ser Gly Asp Phe Asp Thr Phe Arg Arg Trp Leu Val
 50          55          60
Val Asp Glu Val Thr Gln Pro Thr Lys Glu Ile Thr Leu Glu Ala Ala
 65          70          75          80
Arg Tyr Glu Asp Glu Ser Leu Asn Leu Gly Asp Tyr Val Glu Asp Gln
          85          90          95
Ile Glu Ser Val Thr Phe Asp Arg Ile Thr Thr Gln Thr Ala Lys Gln
          100          105          110
Val Ile Val Gln Lys Val Arg Glu Ala Glu Arg Ala Met Val Val Asp
          115          120          125
Gln Phe Arg Glu His Glu Gly Glu Ile Ile Thr Gly Val Val Lys Lys
          130          135          140
Val Asn Arg Asp Asn Ile Ser Leu Asp Leu Gly Asn Asn Ala Glu Ala
          145          150          155          160
Val Ile Leu Arg Glu Asp Met Leu Pro Arg Glu Asn Phe Arg Pro Gly
          165          170          175
Asp Arg Val Arg Gly Val Leu Tyr Ser Val Arg Pro Glu Ala Arg Gly
          180          185          190
Ala Gln Leu Phe Val Thr Arg Ser Lys Pro Glu Met Leu Ile Glu Leu
          195          200          205
Phe Arg Ile Glu Val Pro Glu Ile Gly Glu Glu Val Ile Glu Ile Lys
          210          215          220
Ala Ala Ala Arg Asp Pro Gly Ser Arg Ala Lys Ile Ala Val Lys Thr
          225          230          235          240
Asn Asp Lys Arg Ile Asp Pro Val Gly Ala Cys Val Gly Met Arg Gly

```

245 250 255
 Ala Arg Val Gln Ala Val Ser Thr Glu Leu Gly Gly Glu Arg Ile Asp
 260 265 270
 Ile Val Leu Trp Asp Asp Asn Pro Ala Gln Phe Val Ile Asn Ala Met
 275 280 285
 Ala Pro Ala Asp Val Ala Ser Ile Val Val Asp Glu Asp Lys His Thr
 290 295 300
 Met Asp Ile Ala Val Glu Ala Gly Asn Leu Ala Gln Ala Ile Gly Arg
 305 310 315
 Asn Gly Gln Asn Val Arg Leu Ala Ser Gln Leu Ser Gly Trp Glu Leu
 325 330 335
 Asn Val Met Thr Val Asp Asp Leu Gln Ala Lys His Gln Ala Glu Ala
 340 345 350
 His Ala Ala Ile Asp Thr Phe Thr Lys Tyr Leu Asp Ile Asp Glu Asp
 355 360 365
 Phe Ala Thr Val Leu Val Glu Glu Gly Phe Ser Thr Leu Glu Glu Leu
 370 375 380
 Ala Tyr Val Pro Met Lys Glu Leu Leu Glu Ile Glu Gly Leu Asp Glu
 385 390 395 400
 Pro Thr Val Glu Ala Leu Arg Glu Arg Ala Lys Asn Ala Leu Ala Thr
 405 410 415
 Ile Ala Gln Ala Gln Glu Glu Ser Leu Gly Asp Asn Lys Pro Ala Asp
 420 425 430
 Asp Leu Leu Asn Leu Glu Gly Val Asp Arg Asp Leu Ala Phe Lys Leu
 435 440 445
 Ala Ala Arg Gly Val Cys Thr Leu Glu Asp Leu Ala Glu Gln Gly Ile
 450 455 460
 Asp Asp Leu Ala Asp Ile Glu Gly Leu Thr Asp Glu Lys Ala Gly Ala
 465 470 475 480
 Leu Ile Met Ala Ala Arg Asn Ile Cys Trp Phe Gly Asp Glu Ala
 485 490 495

<210> 29

<211> 88

<212> PRT

5 <213> N-inteína GP41-1 (cianófago)

<400> 29

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 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> 30

<211> 88

10 <212> PRT

<213> Desconocido

<220>

<223> Dominio N-terminal de GP41.8

<400> 30

15 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
85

```

<210> 31

<211> 105

5 <212> PRT

<213> Desconocido

<220>

<223> Dominio N-terminal de NrdJ1

<400> 31

```

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

```

10 <210> 32

<211> 101

<212> PRT

<213> Desconocido

15 <220>

<223> Dominio N-terminal de IMPDH1

<400> 32

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> 33

<211> 106

<212> PRT

5 <213> Desconocido

<220>

<223> Dominio N-terminal de NrdA-2

<400> 33

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

10 <210> 34

<211> 102

<212> PRT

<213> *Nostoc punctiforme*

<400> 34

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

15 <210> 35

<211> 105

<212> PRT

<213> *Synechocystis*

<400> 35

Cys Ile Ser Gly Asp Ser Leu Ile Ser Leu Ala Ser Thr Gly Lys Arg
1 5 10 15
Val Ser Ile Lys Asp Leu Leu Asp Glu Lys Asp Phe Glu Ile Trp Ala
20 25 30
Ile Asn Glu Gln Thr Met Lys Leu Glu Ser Ala Lys Val Ser Arg Val
35 40 45
Phe Cys Thr Gly Lys Lys Leu Val Tyr Ile Leu Lys Thr Arg Leu Gly
50 55 60
Arg Thr Ile Lys Ala Thr Ala Asn His Arg Phe Leu Thr Ile Asp Gly
65 70 75 80

Trp Lys Arg Leu Asp Glu Leu Ser Leu Lys Glu His Ile Ala Leu Pro
85 90 95
Arg Lys Leu Glu Ser Ser Ser Leu Gln
100 105

<210> 36

<211> 45

<212> PRT

<213> Desconocido

<220>

<223> Dominio C-terminal de GP41.8

<400> 36

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> 37

<211> 40

<212> PRT

<213> Desconocido

<220>

<223> Dominio C-terminal de NrdJ1

<400> 37

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> 38

<211> 40

<212> PRT

<213> Desconocido

<220>

5 <223> Dominio C-terminal de IMPDH1

<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> 34

10 <212> PRT

<213> Desconocido

<220>

<223> Dominio C-terminal de NrdA-2

<400> 39

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														

15 <210> 40

<211> 36

<212> PRT

<213> *Nostoc punctiforme*

20 <400> 40

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> 41

<211> 48

<212> PRT

25 <213> *Synechocystis*

<400> 41

Ser Pro Glu Ile Glu Lys Leu Ser Gln Ser Asp Ile Tyr Trp Asp Ser
 1 5 10 15
 Ile Val Ser Ile Thr Glu Thr Gly Val Glu Glu Val Phe Asp Leu Thr
 20 25 30
 Val Pro Gly Pro His Asn Phe Val Ala Asn Asp Ile Ile Val His Asn
 35 40 45

<210> 42

<211> 27

<212> PRT

5 <213> Desconocido

<220>

<223> Secuencia de N-inteína GP41-2

<400> 42

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

10 <210> 43

<211> 46

<212> PRT

<213> Desconocido

<220>

15 <223> Secuencia de N-inteína GP41-3

<400> 43

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> 44

<211> 88

20 <212> PRT

<213> Desconocido

<220>

<223> Secuencia de N-inteína GP41-4

<400> 44

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 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> 45

<211> 88

<212> PRT

5 <213> Desconocido

<220>

<223> Secuencia de N-inteína GP41-5

<400> 45

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 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

10 <210> 46

<211> 43

<212> PRT

<213> Desconocido

<220>

15 <223> Secuencia de N-inteína GP41-6

<400> 46

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> 47

<211> 26

20 <212> PRT

<213> Desconocido

<220>

<223> Secuencia de N-inteína GP41-7

<400> 47

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> 48

5 <211> 133

<212> PRT

<213> Desconocido

<220>

<223> Secuencia de N-inteína NrdA-1

10 <400> 48

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> 49

<211> 105

<212> PRT

15 <213> Desconocido

<220>

<223> Secuencia de N-inteína NrdA-4

<400> 49

Cys	Leu	Ala	Gly	Asp	Thr	Thr	Val	Thr	Val	Leu	Glu	Gly	Asp	Ile	Val
1				5					10					15	

20

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> 50

<211> 65

<212> PRT

5 <213> Desconocido

<220>

<223> Secuencia de N-inteína NrdA-5

<400> 50

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

10 <210> 51

<211> 49

<212> PRT

<213> Desconocido

<220>

15 <223> Secuencia de N-inteína NrdA-6

<400> 51

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> 52

<211> 105

20 <212> PRT

<213> Desconocido

<220>

<223> Secuencia de N-inteína NrdJ-1

<400> 52

```

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> 53

5 <211> 105

<212> PRT

<213> Desconocido

<220>

<223> Secuencia de N-inteína NrdJ-2

10 <400> 53

```

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> 54

<211> 47

<212> PRT

15 <213> Desconocido

<220>

<223> Secuencia de C-inteína GP41-9

<400> 54

```

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

```

20 <210> 55

<211> 42

<212> PRT

<213> Desconocido

<220>

5 <223> Secuencia de C-inteína IMPDH-2

<400> 55

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	Asn	Gly	Phe	Val	Val	His	Asn	Ser						
		35					40								

<210> 56

<211> 41

10 <212> PRT

<213> Desconocido

<220>

<223> Secuencia de C-inteína IMPDH-3

<400> 56

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								

15 <210> 57

<211> 34

<212> PRT

<213> Desconocido

20 <220>

<223> Secuencia de C-inteína NrdA-3

<400> 57

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> 58

25 <211> 28

<212> PRT

<213> Desconocido

<220>

<223> Secuencia de C-inteína NrdA-5

<400> 58

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> 59

5 <211> 33

<212> PRT

<213> Desconocido

<220>

<223> Secuencia de C-inteína NrdA-6

10 <400> 59

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> 60

<211> 41

<212> PRT

15 <213> Desconocido

<220>

<223> Secuencia de C-inteína NrdJ-1

<400> 60

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								

20 <210> 61

<211> 4

<212> PRT

<213> Secuencia artificial

<220>

25 <223> Región de bucle de la *E. coli*

<400> 61

Gly	Cys	Lys	Leu
1			

<210> 62

<211> 4
 <212> PRT
 <213> Secuencia artificial
 <220>
 5 <223> Región de bucle de la *E. coli*
 <400> 62
 Gly Cys Tyr Gln
 1
 <210> 63
 <211> 5
 10 <212> PRT
 <213> Secuencia artificial
 <220>
 <223> Región de bucle de la *E. coli*
 <400> 63
 Gly Cys Gly Tyr Gln
 15 1 5

REIVINDICACIONES

1. Una proteína de fusión que comprende un polipéptido de N-inteína y una contrapartida de solubilización de N-inteína unidos por un enlace peptídico, en el que la contrapartida de solubilización de N-inteína tiene un peso molecular inferior a 15 kDa, un valor de índice alifático menor que 60 y un valor de gran promedio de hidropatía menor que -1, y que aumenta la solubilidad del polipéptido de N-inteína, en comparación con el polipéptido de N-inteína expresado en ausencia de la contrapartida de solubilización.
2. La proteína de fusión según la reivindicación 1, en la que:
 - a) el polipéptido de N-inteína es la N-inteína GP41-1 (SEQ ID NO: 1), o una variante que comprende una secuencia de aminoácidos que tiene al menos el 90 % de identidad de la GP41-1 y, opcionalmente, en donde la variante de N-inteína GP41-1 tiene un aminoácido que no es un residuo de cisteína en las posiciones 7, 65 y 89 de la SEQ ID NO: 1 y, opcionalmente, en donde la variante de N-inteína GP41-1 tiene una alanina en la posición 7, un residuo de treonina o una alanina residuo en la posición 65 y un residuo de metionina, un residuo de lisina o un residuo de asparagina en la posición 89 de la SEQ ID NO: 1 o
 - b) el polipéptido de N-inteína comprende una secuencia seleccionada de las SEQ ID NO: 1-8 y 29-56.
3. La proteína de fusión según una cualquiera de las reivindicaciones 1-2, en donde el polipéptido de N-inteína carece de un aminoácido que se usa para unir la proteína de fusión a un soporte sólido y, opcionalmente, en donde el aminoácido usado para la unión de la proteína de fusión a un soporte sólido es la cisteína.
4. La proteína de fusión según una cualquiera de las reivindicaciones 1-3, en la que:
 - a) la contrapartida de solubilización de la N-inteína se une al extremo N-terminal del polipéptido de la N-inteína o
 - b) la contrapartida de solubilización de la N-inteína se une al extremo C-terminal del polipéptido de la N-inteína.
5. La proteína de fusión según cualquiera de las reivindicaciones precedentes, en la que la contrapartida de solubilización de N-inteína es la contrapartida de solubilización de la N-inteína 138 (SEQ ID NO: 15), o en la que la variante de la contrapartida de solubilización de la N-inteína 138 comprende la SEQ ID NO: 16, SEQ ID NO: 17, o SEQ ID NO: 18.
6. La proteína de fusión según cualquiera de las reivindicaciones precedentes, en donde la proteína de fusión se produce en la *E. coli*, en condiciones en las que menos del 25 % en masa de la proteína de fusión producida está presente en los cuerpos de inclusión.
7. La proteína de fusión según cualquiera de las reivindicaciones anteriores, en la que la proteína de fusión se modifica para incluir un marcador detectable y, opcionalmente, en el que el marcador detectable es un tinte fluorescente.
8. La proteína de fusión según cualquiera de las reivindicaciones precedentes, en donde la proteína de fusión se expresa en *Escherichia coli*, *Corynebacterium glutamicum*, *Pseudomonas fluorescens*, *Lactococcus lactis*, *Pichia pastoris*, *Saccharomyces cerevisiae*, *Zea mays*, *Nicotinia tabacum*, *Daucus carota*, células SF9, células CHO, células NS0 o células HEK 293.
9. Un ácido nucleico que codifica la proteína de fusión según las reivindicaciones 1-8 y, opcionalmente, en el que:
 - a) el ácido nucleico se incluye en un vector de expresión o
 - b) el ácido nucleico se incluye en un plásmido.
10. Una célula hospedadora que comprende el ácido nucleico según la reivindicación 9 y, opcionalmente, en donde la célula hospedadora es *E. coli*.
11. Una matriz de cromatografía de afinidad que comprende la proteína de fusión según una cualquiera de las reivindicaciones 1-8 unida a un soporte sólido, opcionalmente en la que:
 - a) el soporte sólido es una resina de cromatografía y, opcionalmente, en la que la resina de cromatografía incluye una base hidrófila de poliviniléter o
 - b) el soporte sólido es una cuenta, una fibra hueca, una fibra sólida, una almohadilla, un gel, una membrana, un casete, una columna, un chip, una placa, un plato o un monolito y, opcionalmente, en donde el soporte sólido es una cuenta magnética.
12. La matriz de cromatografía de afinidad según la reivindicación 11, en la que el soporte sólido comprende vidrio de poro controlado, sílice, óxido de circonio, óxido de titanio, agarosa, polimetacrilato, poliácido, poliácidoamida, alcohol polivinílico, poliestireno o derivados de los mismos.

13. La matriz de cromatografía de afinidad según una cualquiera de las reivindicaciones 11-12, en la que la matriz comprende, además, una molécula espaciadora entre la proteína de fusión y el soporte sólido.

14. La matriz de cromatografía de afinidad según una cualquiera de las reivindicaciones 11-13, en la que:

5 a) la proteína de fusión está unida al soporte sólido en un sitio único en la contrapartida de solubilización de la N-inteína o

b) la proteína de fusión se une al soporte sólido en más de un sitio en la contrapartida de solubilización de la N-inteína.

10 15. La matriz de cromatografía de afinidad según una cualquiera de las reivindicaciones 1-14, en la que el polipéptido de la N-inteína en la proteína de fusión permanece activo cuando la proteína de fusión está unida al soporte sólido y, opcionalmente, en donde el polipéptido de N-inteína en la fusión la proteína está orientado de modo que se aleje del soporte sólido cuando la proteína de fusión se une al soporte sólido.

FIG. 1

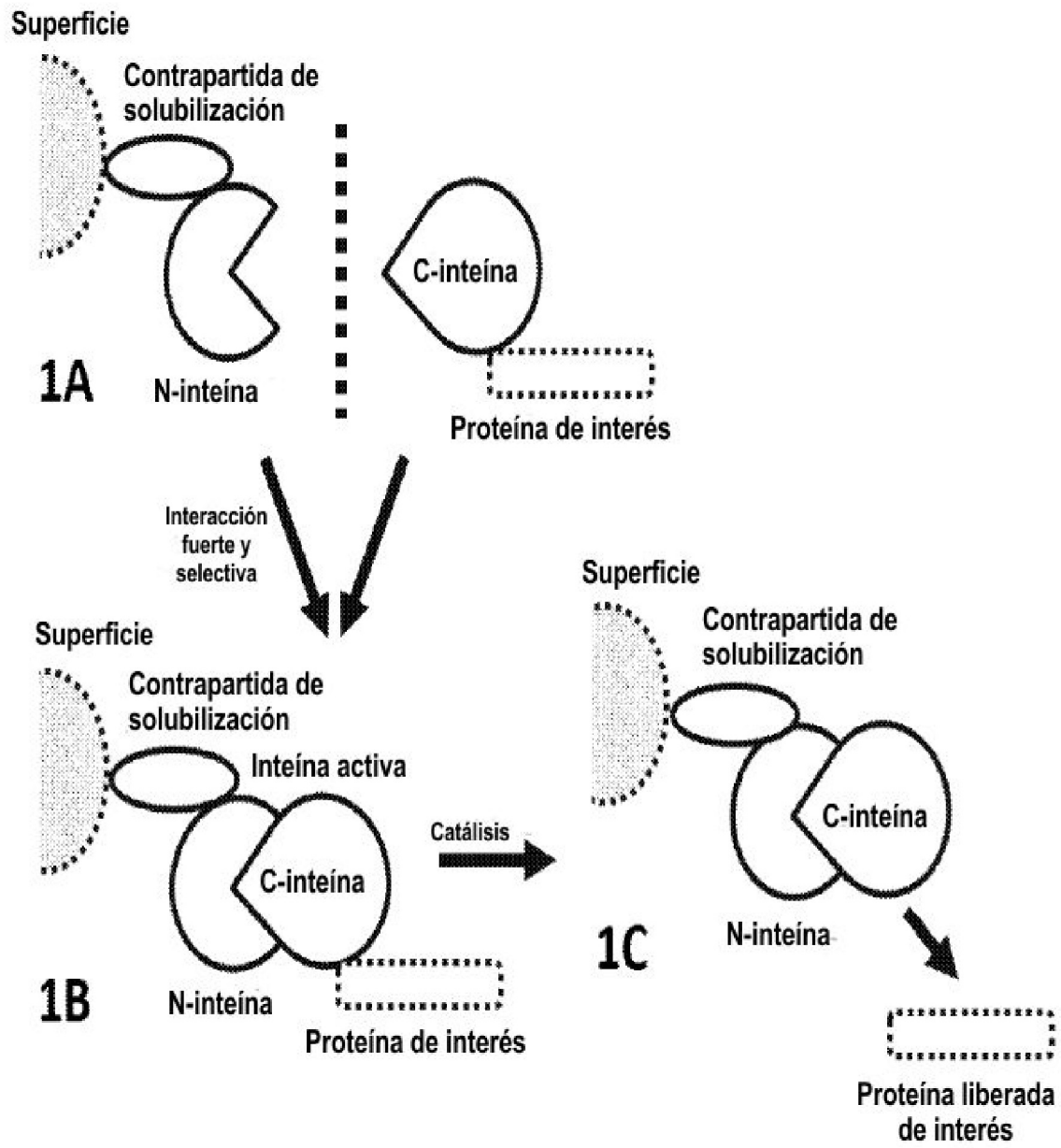


FIG. 2

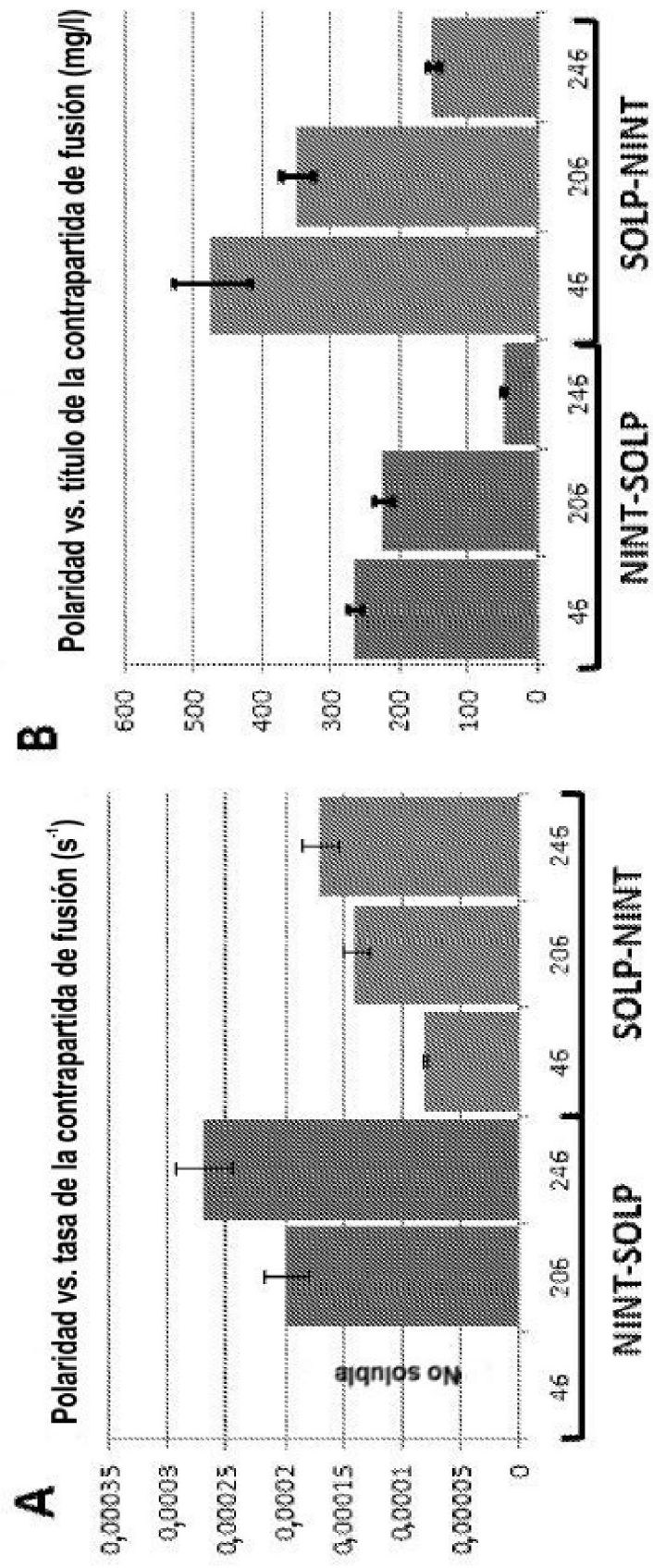


FIG. 3

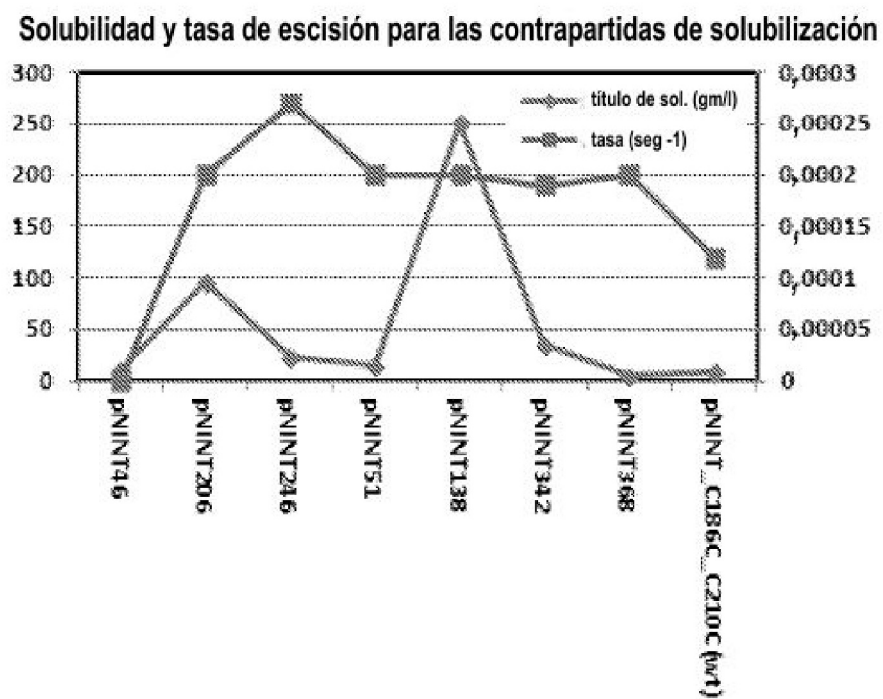
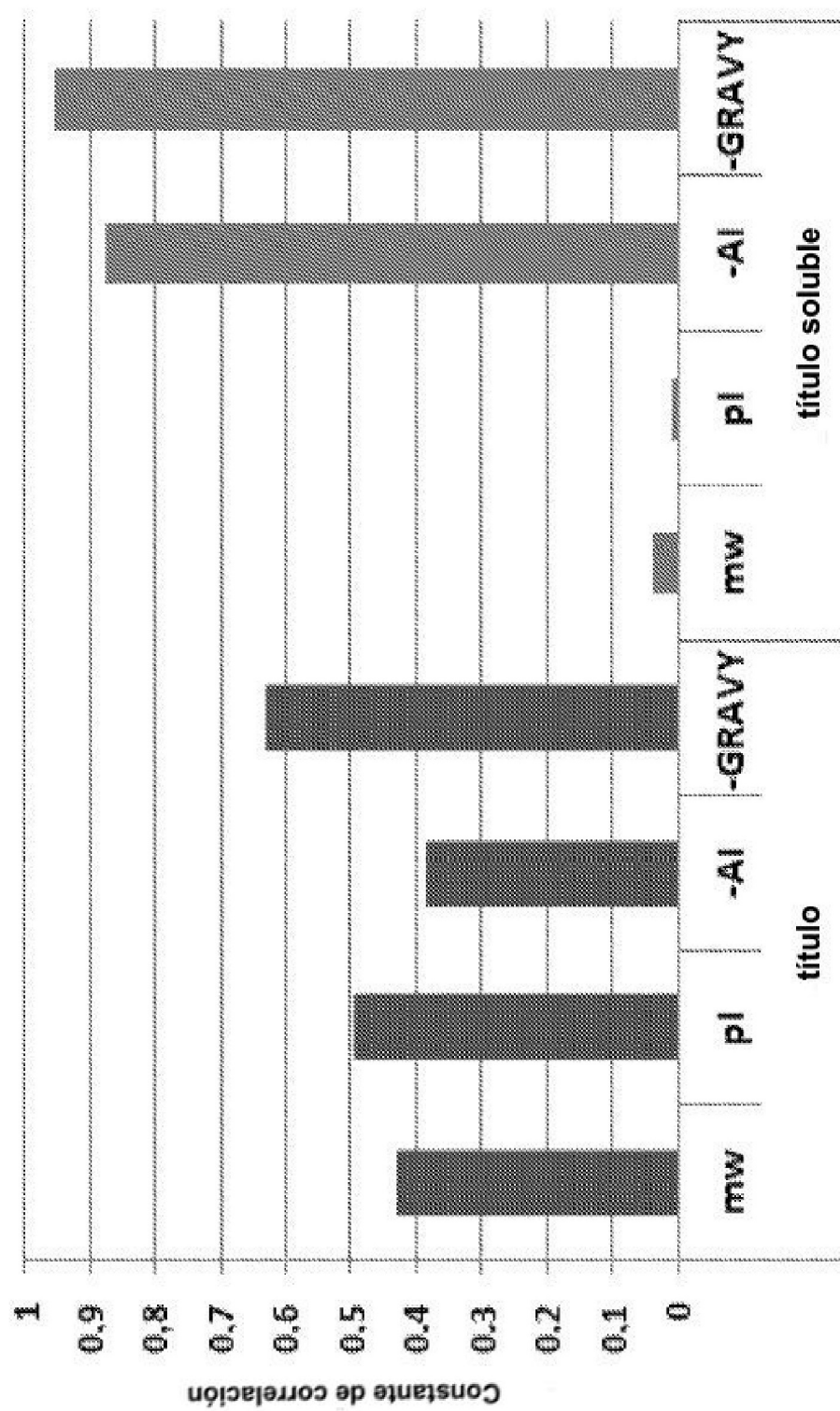


FIG. 4



FIGS. 5A-5B

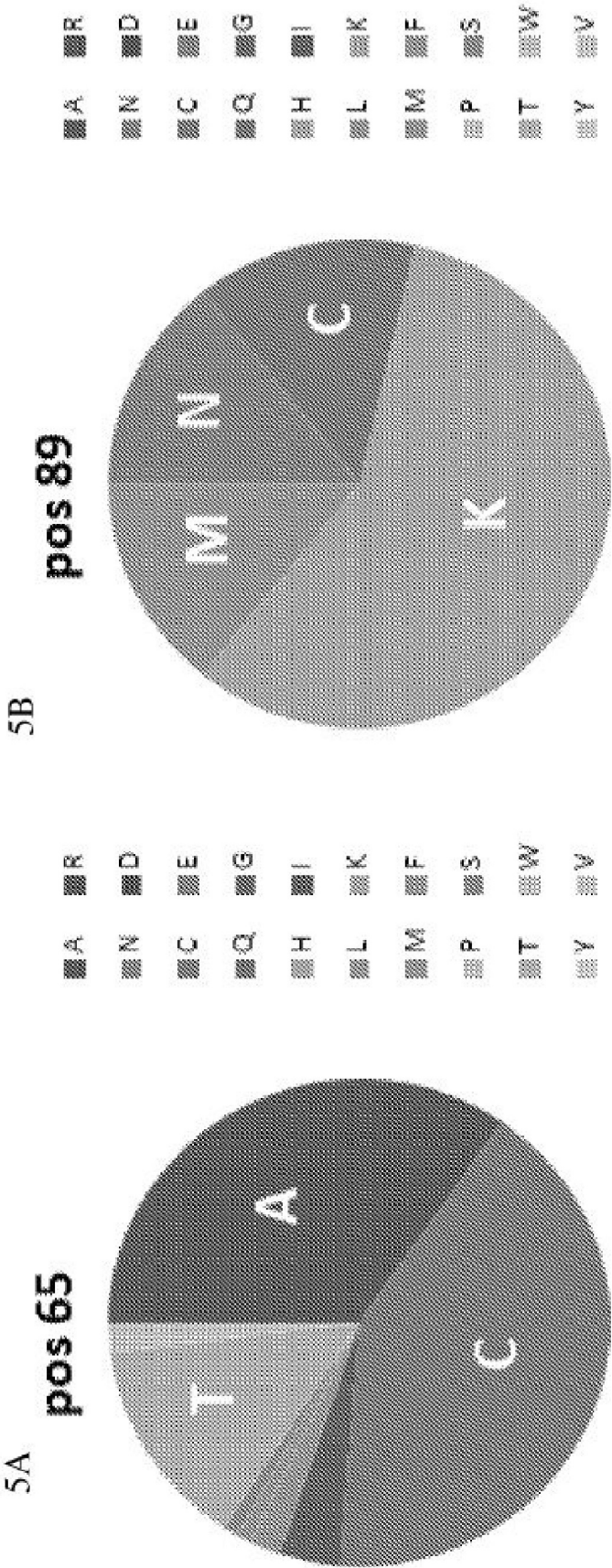


FIG. 6

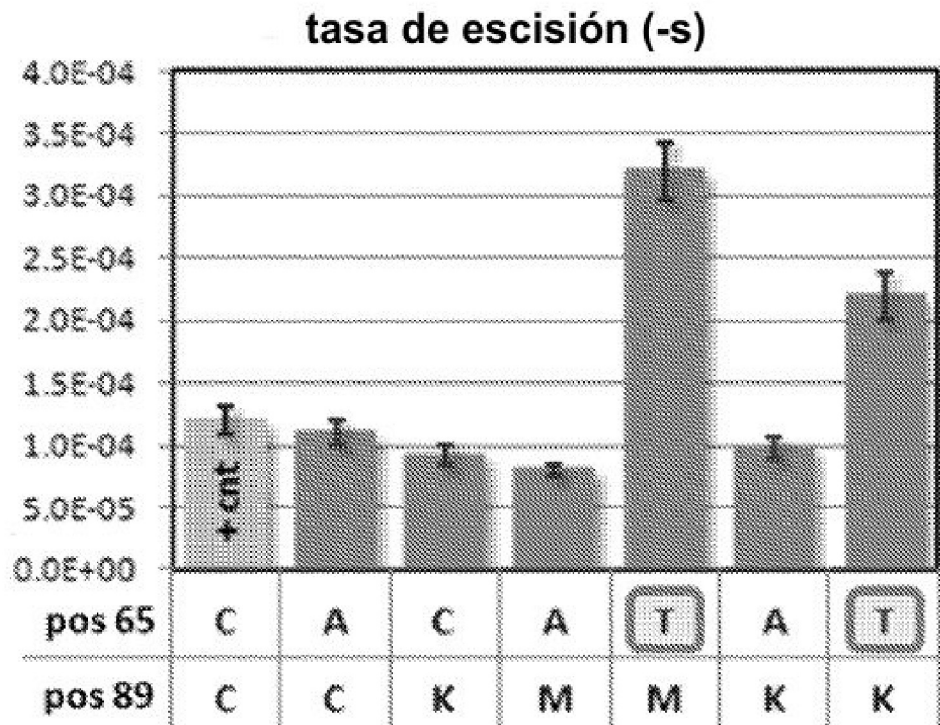


FIG. 7

