



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria
e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) **PI 0711878-3 A2**

(22) Data de Depósito: 30/05/2007
(43) Data da Publicação: 10/01/2012
(RPI 2140)



(51) *Int.Cl.:*
C12P 21/00
C12N 15/01

(54) **Título:** MÉTODO DE OTIMIZAÇÃO DO CÓDON

(30) **Prioridade Unionista:** 30/05/2006 US 60/809,536,
14/02/2007 US 60/901,687, 14/02/2007 US 60/901,687, 30/05/2006
US 60/809,536

(73) **Titular(es):** Dow Global Technologies Inc.

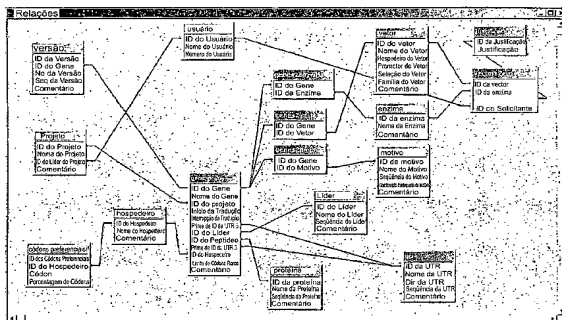
(72) **Inventor(es):** Charles Douglas Hershberger, Steven J.
Stelman, Thomas M. Ramseier

(74) **Procurador(es):** Dannemann, Siemsen, Bigler &
Ipanema Moreira

(86) **Pedido Internacional:** PCT US2007012719 de
30/05/2007

(87) **Publicação Internacional:** WO 2007/142954de
13/12/2007

(57) **Resumo:** MÉTODO DE OTIMIZAÇÃO DO CÓDON. A presente invenção refere-se a uma expressão heteróloga em uma bactéria de *Pseudomonas* hospedeira de uma sequência de polinucleotídeo otimizada codificando uma proteína.





PI0711878-3

Relatório Descritivo da Patente de Invenção para "**MÉTODO DE OTIMIZAÇÃO DO CÓDON**".

REFERÊNCIA CRUZADA A PEDIDOS RELACIONADOS

5 Este pedido reivindica a prioridade ao Pedido Provisório dos Estados Unidos Nº de Série: 60/901.687, depositado em 14 de fevereiro de 2007, e ao Pedido Provisório dos Estados Unidos Nº de Série: 60/809.536, depositado em 30 de maio de 2006, cujas divulgações são incorporadas por referência em suas totalidades.

CAMPO DA INVENÇÃO

10 A presente invenção refere-se, de um modo geral, aos métodos para otimizar os genes para a expressão bacteriana. A invenção adicionalmente se refere a um sistema de banco de dados e ferramentas para a análise dos genes otimizados.

ANTECEDENTES DA INVENÇÃO

15 Diversas bactérias têm sido usadas como células hospedeiras para a preparação de proteínas recombinantes heterólogas. Uma desvantagem significativa de numerosos sistemas bacterianos é o seu uso de códons raros, que é muito diferente dos códons preferenciais nos genes humanos. A presença destes códons raros pode resultar na expressão retardada e redu-
20 zida de genes recombinantes. Em certos aspectos, uma sequência de ácidos nucleicos pode ser modificada para codificar uma variante de polipeptídeo recombinante, onde códons específicos da sequência de ácidos nucleicos tenham sido alterados para códons que sejam favorecidos por um hospedeiro particular e possa resultar em níveis aumentados de expressão (ver,
25 por exemplo, Haas et al., Curr. Biol. 6:315, 1996; Yang et al., Nucleic Acids Res. 24:4592, 1996).

O processo de otimizar a sequência de nucleotídeos que codifica uma proteína expressa de modo heterólogo pode ser uma etapa importante para aperfeiçoar os rendimentos da expressão. As exigências de otimização
30 podem incluir as etapas de aperfeiçoar a capacidade do hospedeiro de produzir a proteína exógena, bem como as etapas de auxiliar o pesquisador em projetar eficientemente construções de expressão. Embora os preços para a

síntese de DNA em escala de gene tenham declinado significativamente nos anos recentes, o investimento na síntese de um gene otimizado para este propósito pode ser caro. Portanto, é importante que uma análise meticulosa seja conduzida para assegurar que todas as exigências de projeto tenham sido adequadamente satisfeitas, antes de prosseguir com a síntese. Além disso, o processo de avaliar os genes sintéticos candidatos e produzir relatórios dos resultados desta análise capazes de serem lidos por seres humanos é um processo que consome tempo.

Embora existam diversas ferramentas para o cálculo de códons preferenciais, estas ferramentas não são geralmente projetadas para relatar o uso do códon em um contexto utilizável. Uma vez que estas ferramentas não comparam um uso calculado com um padrão de referência, requer-se tipicamente uma formatação manual novamente dos dados de saída para distinguir a presença dos códons raros em relação ao sistema de expressão de hospedeiro. A visualização espacial dos códons raros ao longo da sequência de genes traduzida deve também ser efetuada manualmente. Assim, requer-se o treinamento substancial do usuário, incluindo a importação da sequência desejada para o formato correto para cada aplicação.

BREVE SUMÁRIO DA INVENÇÃO

A presente invenção inclui uma sequência de polinucleotídeo sintético que tenha sido otimizada para a expressão heteróloga em uma célula hospedeira bacteriana, tal como o *Pseudomonas fluorescens*.

A presente invenção também proporciona um método de produzir uma proteína recombinante no citoplasma da célula bacteriana, incluindo otimizar uma sequência de polinucleotídeo sintético para a expressão heteróloga em um hospedeiro bacteriano, onde o polinucleotídeo sintético compreende uma sequência de nucleotídeos codificando uma proteína, tal como um antígeno. O método também inclui ligar a sequência de polinucleotídeo sintético otimizada em um vetor de expressão e transformar a bactéria hospedeira com o vetor de expressão. O método adicionalmente inclui cultivar a bactéria hospedeira transformada em um meio de cultura adequado, apropriado para a expressão da proteína, e isolar a proteína. O hospedeiro de

bactéria selecionado pode ser o *Pseudomonas fluorescens*.

As outras modalidades da presente invenção incluem os métodos de otimizar as seqüências de polinucleotídeos sintéticos para a expressão heteróloga em uma célula hospedeira, por identificação e modificação dos códons raros a partir da seqüência de polinucleotídeo sintético que sejam raramente usados no hospedeiro. Além disso, estes métodos podem incluir a identificação e a modificação de seqüências internas putativas de sítios ligadores ribossômicos, bem como a identificação e a modificação de repetições alongadas dos nucleotídeos G ou C a partir da seqüência de polinucleotídeo sintético. Os métodos podem também incluir a identificação e a minimização das estruturas secundárias de mRNA nas regiões de RBS e codificação do gene, bem como a modificação de sítios de restrição de enzimas indesejáveis a partir das seqüências de polinucleotídeos sintéticos.

A presente invenção também proporciona a análise em série e a geração de relatório automáticas de um gene usando um banco de dados e ferramentas para calcular o uso do códon a partir de uma seqüência bruta e graficamente relatar a posição dos códons raros ao longo da seqüência de DNA traduzida. Onde forem projetadas versões candidatas múltiplas de um gene particular, efetua-se uma análise de todas as versões para determinar o melhor candidato para a síntese. Esta comparação, juntamente com uma comparação das versões candidatas com aquela de um códon preferencial de referência, é apresentada em um formato útil capaz de ser lido por ser humano.

BREVE DESCRIÇÃO DAS DIVERSAS VISTAS DOS DESENHOS

A figura 1 ilustra um fluxograma mostrando as etapas que podem ser usadas durante a otimização de uma seqüência de polinucleotídeo sintético;

As figuras 2 e 3 ilustram os perfis de uso dos códons raros mostrando a posição e a distribuição dos códons raros ao longo de uma seqüência de proteína traduzida na cepa de *P. fluorescens* MB214; e

A figura 4 ilustra uma modalidade de um esquema de banco de dados para o banco de dados do gene da presente invenção.

DESCRIÇÃO DETALHADA DA INVENÇÃO

A presente invenção é descrita mais completamente nas partes que se seguem com referência aos desenhos que acompanham, nos quais são mostradas as modalidades preferidas da invenção. Esta invenção pode, entretanto, ser incluída em muitas formas diferentes e não deve ser interpretada como limitada às modalidades descritas neste documento; preferivelmente, estas modalidades são proporcionadas de modo que esta divulgação seja meticulosa e completa, e transmita inteiramente o escopo da invenção para aqueles versados na técnica.

A invenção, de um modo geral, refere-se a um processo para preparar uma proteína recombinante heteróloga em uma célula hospedeira procariótica. O uso do códon da célula hospedeira para os genes da célula hospedeira é determinado. Os códons que ocorrem de modo raro são modificados com os códons que ocorrem frequentemente no ácido nucléico que codifica a proteína recombinante heteróloga na célula hospedeira. A célula hospedeira é então transformada com o ácido nucléico que codifica a proteína recombinante e o ácido nucléico recombinante é expresso.

Conforme usados aqui, os termos "modificar" ou "alterar", ou quaisquer formas deles, significam modificar, alterar, trocar, eliminar, substituir, remover, variar, ou transformar.

A presente invenção também se refere às seqüências de polinucleotídeos sintéticos que codificam uma proteína. As modalidades da presente invenção também proporcionam a expressão heteróloga de um polinucleotídeo sintético em um hospedeiro bacteriano. As outras modalidades incluem uma expressão heteróloga de um polinucleotídeo sintético em *Pseudomonas fluorescens*. As modalidades adicionais da presente invenção também incluem as seqüências de polinucleotídeos otimizadas codificando uma proteína recombinante que podem ser expressas usando um sistema de expressão heteróloga à base de *Pseudomonas fluorescens*. Uma outra modalidade da presente invenção também inclui uma expressão heteróloga de um polinucleotídeo sintético no citoplasma de *Pseudomonas fluorescens*. A modalidade adicional da presente invenção também inclui uma expressão

heteróloga de um polinucleotídeo sintético no periplasma de *Pseudomonas fluorescens*.

Nos sistemas de expressão heteróloga, as etapas de otimização podem aperfeiçoar a capacidade do hospedeiro de produzir a proteína exógena. A expressão da proteína é governada por um hospedeiro de fatores incluindo aqueles que afetam a transcrição, o processamento do mRNA, e a estabilidade e a iniciação da tradução. As etapas de otimização do polinucleotídeo podem incluir as etapas de melhorar a capacidade do hospedeiro de produzir a proteína exógena, bem como as etapas de auxiliar o pesquisador em projetar eficientemente as construções de expressão. As estratégias de otimização podem incluir, por exemplo, a modificação das regiões de iniciação da tradução, a alteração dos elementos estruturais de mRNA, e o uso de diferentes códons preferenciais. Os parágrafos a seguir discutem os problemas potenciais que podem resultar na expressão heteróloga reduzida da proteína, e as técnicas que podem superar estes problemas.

Uma área que pode resultar na expressão heteróloga reduzida da proteína é uma pausa traducional induzida pelo códon raro. Uma pausa traducional induzida pelo códon raro inclui a presença de códons no polinucleotídeo de interesse que são raramente usados no organismo hospedeiro, que pode ter um efeito negativo sobre a tradução da proteína devido à sua escassez na união de tRNA disponível. Um método de aperfeiçoar a tradução ótima no organismo hospedeiro inclui efetuar a otimização do códon, o que pode resultar nos códons raros do hospedeiro sendo modificados nas seqüências de polinucleotídeos sintéticos.

Uma outra área que pode resultar na expressão heteróloga reduzida da proteína é por iniciação traducional alternada. A iniciação traducional alternada pode incluir uma seqüência de polinucleotídeo sintético inadvertidamente contendo motivos capazes de funcionar como um sítio ligador de ribossomo (RBS). Estes sítios podem resultar na iniciação da tradução de uma proteína truncada a partir de um sítio de gene interno. Um método de reduzir a possibilidade de produzir uma proteína truncada, a qual pode ser difícil de remover durante a purificação, inclui modificar as seqüên-

cias internas putativas do RBS a partir de uma seqüência de polinucleotídeo otimizada.

Uma outra área que pode resultar na expressão heteróloga reduzida da proteína é através do deslizamento da polimerase induzido pela repetição. O deslizamento da polimerase induzido pela repetição envolve as repetições das seqüências de nucleotídeos que tenham sido mostradas causarem o deslizamento ou a falha da DNA polimerase, o que pode resultar nas mutações de deslocamento. Tais repetições podem também causar o deslizamento da RNA polimerase. Em um organismo com uma tendência por altos teores de G+C, pode haver um grau maior de repetições compostas de repetições de nucleotídeos G ou C. Portanto, um método de reduzir a possibilidade de induzir o deslizamento da RNA polimerase inclui alterar as repetições alongadas dos nucleotídeos G ou C.

Uma outra área que pode resultar na expressão heteróloga reduzida da proteína é através das estruturas secundárias interferentes. As estruturas secundárias podem seqüestrar a seqüência do RBS ou o códon de iniciação e têm sido correlacionadas a uma redução na expressão da proteína. As estruturas de laço de haste podem também estar envolvidas na pausa transcricional e na atenuação. Uma seqüência de polinucleotídeo otimizada pode conter estrutura secundária mínima nas regiões do RBS e de codificação do gene da seqüência de nucleotídeos, para permitir a transcrição e a tradução aperfeiçoadas.

Uma outra área que pode efetuar a expressão heteróloga da proteína são os sítios de restrição: Por modificação dos sítios de restrição que poderiam interferir com a subclonagem subsequente das unidades de transcrição nos vetores de expressão hospedeiros, pode ser otimizada uma seqüência de polinucleotídeo.

A otimização de uma seqüência de DNA pode afetar negativa ou positivamente a expressão do gene ou a produção de proteína. Por exemplo, a modificação de um códon menos comum com um códon mais comum pode afetar a meia-vida do mRNA ou alterar a sua estrutura por introdução de uma estrutura secundária que interfere com a tradução da mensagem. Pode,

portanto, ser necessário, em certas situações, alterar a mensagem otimizada.

5 Todo, ou uma parte de um gene pode ser otimizado. Em alguns casos, a modulação desejada da expressão é atingida otimizando-se essencialmente o gene inteiro. Em outros casos, a modulação desejada será atingida otimizando-se parte do, porém não todo o gene.

10 O uso do códon de qualquer seqüência de codificação pode ser ajustado para atingir uma propriedade desejada, por exemplo, altos níveis de expressão em um tipo específico de célula. O ponto de partida para tal otimização pode ser uma seqüência de codificação com 100% de códons comuns, ou uma seqüência de codificação que contenha uma mistura de códons comuns e não-comuns.

15 Duas ou mais seqüências candidatas que difiram em seu uso do códon podem ser geradas e testadas para determinar se elas possuem a propriedade desejada. As seqüências candidatas podem ser avaliadas usando um computador para pesquisar quanto à presença de elementos reguladores, tais como silenciadores e seqüências reforçadoras, e para pesquisar quanto à presença de regiões da seqüência de codificação que poderiam ser convertidas em tais elementos reguladores por uma alteração no uso do códon. Os critérios adicionais podem incluir o enriquecimento para
20 nucleotídeos particulares, por exemplo, A, C, G ou U, os códons preferenciais para um aminoácido particular, ou a presença ou a ausência de estrutura secundária ou terciária de mRNA particular. O ajuste para a seqüência candidata pode ser feito com base em diversos tais critérios.

25 As seqüências candidatas promissoras são construídas e então avaliadas experimentalmente. Os candidatos múltiplos podem ser avaliados independentemente um do outro, ou o processo pode ser iterativo, usando o candidato mais promissor como um novo ponto de partida, ou combinando as regiões de dois ou mais candidatos para produzir um novo híbrido. Podem ser incluídas rodadas adicionais de modificação e avaliação.

30 A modificação do uso do códon de uma seqüência candidata pode resultar na criação ou na destruição de um elemento positivo ou negativo. Em geral, um elemento positivo refere-se a qualquer elemento cuja alte-

ração ou remoção da seqüência candidata possa resultar em uma diminuição na expressão da proteína terapêutica, ou cuja criação possa resultar em um aumento na expressão de uma proteína terapêutica. Por exemplo, um elemento positivo pode incluir uma seqüência reforçadora, um promotor, um elemento promotor a jusante, um sítio de ligação de DNA para um regulador positivo (por exemplo, um ativador transcricional), ou uma seqüência responsável por conferir ou modificar uma estrutura secundária ou terciária de mRNA. Um elemento negativo refere-se a qualquer elemento cuja alteração ou remoção da seqüência candidata possa resultar em um aumento na expressão da proteína terapêutica, ou cuja criação resultaria em uma diminuição na expressão da proteína terapêutica. Um elemento negativo inclui um silenciador, um sítio de ligação de DNA para um regulador negativo (por exemplo, um repressor transcricional), um sítio de pausa transcricional, ou uma seqüência que seja responsável por conferir ou modificar uma estrutura secundária ou terciária de mRNA. Em geral, um elemento negativo surge mais frequentemente do que um elemento positivo. Assim, qualquer alteração no uso do códon que resulte em um aumento na expressão da proteína é mais provável de ter surgido da destruição de um elemento negativo, ao invés da criação de um elemento positivo. Além disso, a alteração da seqüência candidata é mais provável de destruir um elemento positivo do que criar um elemento positivo. Em uma modalidade, uma seqüência candidata é escolhida e modificada de modo a aumentar a produção de uma proteína terapêutica. A seqüência candidata pode ser modificada, por exemplo, alterando-se seqüencialmente os códons ou alterando-se aleatoriamente os códons na seqüência candidata. Uma seqüência candidata modificada é então avaliada por determinação do nível de expressão da proteína terapêutica resultante ou por avaliação de um outro parâmetro, por exemplo, um parâmetro correlacionado ao nível de expressão. Escolhe-se uma seqüência candidata que produza um nível aumentado de uma proteína terapêutica, comparada a uma seqüência candidata inalterada.

Em uma outra abordagem, um ou um grupo de códons pode ser modificado, por exemplo, sem referência à proteína ou estrutura da mensa-

gem, e testado. Alternativamente, um ou mais códons podem ser escolhidos em uma propriedade de nível de mensagem, por exemplo, posição em uma região de teor de GC predeterminado, por exemplo, alto ou baixo, posição em uma região tendo uma estrutura tal como uma seqüência intensificadora ou silenciador, posição em uma região que possa ser modificada para introduzir uma estrutura, tal como uma seqüência intensificadora ou silenciador, posição em uma região tendo, ou predita ter, estrutura secundária ou terciária, por exemplo, emparelhamento dentro da cadeia, emparelhamento entre cadeias, posição em uma região não tendo, ou predita não ter, estrutura secundária ou terciária, p.ex, emparelhamento dentro da cadeia ou entre cadeias. Uma região modificada particular é escolhida se ela produzir o resultado desejado.

Os métodos que geram sistematicamente seqüências candidatas são úteis. Por exemplo, um ou um grupo, por exemplo, um bloco contínuo, de códons, em diversas posições de uma seqüência de ácidos nucleicos sintética, pode ser modificado com códons comuns (ou com códons não comuns, se, por exemplo, a seqüência de partida tiver sido otimizada) e a seqüência resultante avaliada. Os candidatos podem ser gerados otimizando (ou não otimizando) uma dada "janela" de códons na seqüência para gerar um primeiro candidato, e então movendo a janela para uma nova posição na seqüência, e otimizando (ou não otimizando) os códons na nova posição sob a janela para proporcionar um segundo candidato. Os candidatos podem ser avaliados determinando o nível de expressão que eles proporcionam, ou avaliando um outro parâmetro, por exemplo, um parâmetro correlacionado com o nível de expressão. Alguns parâmetros podem ser avaliados por inspeção ou de modo computacional, por exemplo, a sua posse ou privação de teor de GC alto ou baixo; um elemento de seqüência, tal como uma seqüência reforçadora ou um silenciador; estrutura secundária ou terciária, por exemplo, emparelhamento dentro da cadeia ou entre cadeias.

Em certas modalidades, a seqüência de ácidos nucleicos otimizada pode expressar a sua proteína, em um nível que seja pelo menos 110%, 150%, 200%, 500%, 1.000%, 5.000% ou mesmo 10.000% daquele

expresso pela seqüência de ácidos nucléicos que não tenha sido otimizada.

Conforme ilustrado pela figura 1, o processo de otimização pode começar pela identificação da seqüência de aminoácidos desejada a ser expressa de modo heterólogo pelo hospedeiro. A partir da seqüência de aminoácidos, uma seqüência de polinucleotídeo ou DNA candidata pode ser projetada. Durante o projeto da seqüência de DNA sintética, a freqüência do uso do códon pode ser comparada com o uso do códon do organismo de expressão hospedeiro e os códons raros do hospedeiro podem ser modificados na seqüência sintética. Adicionalmente, a seqüência de DNA candidata sintética pode ser modificada para remover os sítios de restrição de enzimas indesejáveis e adicionar ou alterar quaisquer seqüências de sinais, ligadores ou regiões não traduzidas desejados. A seqüência de DNA sintética pode ser analisada quanto à presença de estrutura secundária que possa interferir com o processo de tradução, tal como as repetições de G/C e as estruturas de laço. Antes da seqüência de DNA candidata ser sintetizada, o projeto de seqüência otimizada pode ser checado para verificar que a seqüência codifica corretamente a seqüência de aminoácidos desejada. Finalmente, a seqüência de DNA candidata pode ser sintetizada usando práticas de síntese de DNA, tais como aquelas conhecidas na técnica.

Em uma outra modalidade da invenção, o uso do códon geral em um organismo hospedeiro, tal como *Pseudomonas fluorescens*, pode ser utilizado para otimizar a expressão da seqüência de polinucleotídeo heteróloga. Podem ser avaliadas a porcentagem e a distribuição dos códons que raramente seriam considerados como preferidos para um aminoácido particular no sistema de expressão de hospedeiro. Os valores de 5% e 10% de uso podem ser usados como valores de corte para a determinação dos códons raros. Por exemplo, os códons listados na TABELA 1 têm uma ocorrência calculada de menos do que 5% no genoma de *Pseudomonas fluorescens* MB214 e seriam geralmente evitados em um gene otimizado, expresso em um hospedeiro de *Pseudomonas fluorescens*.

TABELA 1

Aminoácido(s)	Códon(s) Usado(s)	% de Ocorrência
G Gly	GGA	3,26
I Ile	ATA	3,05
L Leu	CTA	1,78
	CTT	4,57
	TTA	1,89
R Arg	AGA	1,39
	AGG	2,72
	CGA	4,99
S Ser	TCT	4,18

Uma variedade de células hospedeiras pode ser usada para a expressão de um produto de gene heterólogo. A célula hospedeira pode ser selecionada a partir de uma população apropriada de células de *E. coli* ou células de *Pseudomonas*. As pseudomônadas e as bactérias intimamente relacionadas, como usadas neste documento, são co-extensivas com o grupo definido aqui como "Subgrupo de Proteobactérias Gram(-) 1". O "Subgrupo de Proteobactérias Gram(-) 1" é mais especificamente definido como o grupo de Proteobactérias pertencente às famílias e/ou aos gêneros descritos como incidindo naquela "Parte" taxonômica chamada "Bastões e Cocos Aeróbicos Gram-Negativos" por R. E. Buchanan e N. E. Gibbons (eds.), *Bergey's Manual of Determinative Bacteriology*, pp. 217-289 (8ª ed., 1974) (The Williams & Wilkins Co., Baltimore, Md., EUA) (nas partes que se seguem, "Bergey (1974)"). A célula hospedeira pode ser selecionada a partir do Subgrupo de Proteobactérias Gram-Negativas 18, que é definido como o grupo de todas as subespécies, variedades, cepas, e outras unidades subespeciais da espécie *Pseudomonas fluorescens*, incluindo aquelas pertencentes, por exemplo, aos que se segue (com o ATCC ou outros números de depósitos da(s) cepa(s) ilustrativa(s) mostrados em parêntese): *P. fluorescens* biótipo A, também chamado biovar 1 ou biovar I (ATCC 13525); *P. fluorescens* biótipo B, também chamado biovar 2 ou biovar II (ATCC 17816); *P. fluorescens* biótipo C, também chamado biovar 3 ou biovar III (ATCC 17400); *P. fluorescens* biótipo F, também chamado biovar 4 ou biovar IV (ATCC 12983); *P.*

fluorescens biótipo G, também chamado biovar 5 ou biovar V (ATCC 17518); *P. fluorescens* biovar VI; *P. fluorescens* Pf0-1; *P. fluorescens* Pf-5 (ATCC BAA-477); *P. fluorescens* SBW25; e *P. fluorescens* subsp. *cellulosa* (NCIMB 10462).

5 A célula hospedeira pode ser selecionada a partir do Subgrupo de Proteobactérias Gram-Negativas 19, que é definido como o grupo de todas as cepas de *P. fluorescens* biótipo A, incluindo a cepa de *P. fluorescens* MB101, e seus derivados.

10 Em uma modalidade, a célula hospedeira pode ser quaisquer das Proteobactérias da ordem Pseudomonadales. Em uma modalidade particular, a célula hospedeira pode ser quaisquer das Proteobactérias da família Pseudomonadaceae. Em uma modalidade particular, a célula hospedeira pode ser selecionada a partir de um ou mais dos que seguem: Subgrupo de Proteobactérias Gram-negativas 1, 2, 3, 5, 7, 12, 15, 17, 18 ou 19.

15 As cepas adicionais de *P. fluorescens* que podem ser usadas na presente invenção incluem *P. fluorescens* Migula e *P. fluorescens* Loitokitok, tendo as seguintes designações de ATCC: [NCIB 8286]; NRRL B-1244; cepa NCIB 8865 COI; cepa NCIB 8866 CO2; 1291 [ATCC 17458; IFO 15837; NCIB 8917; LA; NRRL B-1864; pirrolidina; PW2 [ICMP 3966; NCPPB 967; NRRL B-899]; 13475; NCTC 10038; NRRL B-1603 [6; IFO 15840]; 52-IC; CCEB 488-A [BU 140]; CCEB 553 [DEM 15/47]; IAM 1008 [AHH-27]; IAM 1055 [AHH-23]; 1 [IFO 15842]; 12 [ATCC 25323; NIH 11; den Dooren de Jong 216]; 18 [IFO 15833; WRRL P-7]; 93 [TR-10]; 108[52-22; IFO 15832]; 143 [IFO 15836; PL]; 149 [2-40-40; IFO 15838]; 182 [IFO 3081; PJ 73]; 184
20 [IFO 15830]; 185[W2 L-1]; 186 [IFO 15829; PJ 79]; 187 [NCPPB 263]; 188 [NCPPB 316]; 189 [PJ227; 1208]; 191 [IFO 15834; PJ 236; 22/1]; 194 [Klinge R-60; PJ 253]; 196 [PJ 288]; 197 [PJ 290]; 198[PJ 302]; 201 [PJ 368]; 202 [PJ 372]; 203 [PJ 376]; 204 [IFO 15835; PJ 682]; 205[PJ686]; 206 [PJ 692]; 207 [PJ 693]; 208 [PJ 722]; 212 [PJ 832]; 215 [PJ 849]; 216 [PJ885]; 267 [B-
25 9]; 271 [B-1612]; 401 [C71A; IFO 15831; PJ 187]; NRRL B-3178 [4; IFO 15841]; KY8521; 3081; 30-21; [IFO 3081]; N; PYR; PW; D946-B83 [BU 2183; FERM-P 3328]; P-2563 [FERM-P 2894; IFO 13658]; IAM-1126 [43F]; M-1;
30

A506 [A5-06]; A505[A5-05-1]; A526 [A5-26]; B69; 72; NRRL B4290; PMW6 [NCIB 11615]; SC 12936; A1 [IFO 15839]; F 1847 [CDC-EB]; F 1848 [CDC 93]; NCIB 10586; P17; F-12; AmMS 257; PRA25; 6133D02; 6519E01; Ni; SC15208; BNL-WVC; NCTC 2583 [NCIB 8194]; H13; 1013 [ATCC 11251];
 5 CCEB 295]; IFO 3903; 1062; ou Pf-5.

A transformação das células hospedeiras de *Pseudomonas* com o(s) vetor(es) pode ser efetuada usando qualquer metodologia de transformação conhecida na técnica, e as células hospedeiras bacterianas podem ser transformadas como células intactas ou como protoplastos (isto é, inclu-
 10 indo os citoplastos). As metodologias de transformação incluem as metodologias de poração, por exemplo, eletroporação, fusão de protoplasto, conjugação bacteriana, e tratamento com cátion divalente, por exemplo, tratamento com cloreto de cálcio ou tratamento com $\text{CaCl}/\text{Mg}^{2+}$, ou outros métodos bastante conhecidos na técnica. Ver, por exemplo, Morrison, *J. Bact.*,
 15 132:349-351 (1977); Clark-Curtiss e Curtiss, *Methods in Enzymology*, 101:347-362 (Wu et al., eds, 1983), Sambrook et al., *Molecular Cloning, A Laboratory Manual* (2^a ed. 1989); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); e *Current Protocols in Molecular Biology* (Ausubel et al., eds., 1994)).

Conforme usado neste documento, o termo "fermentação" inclui tanto as modalidades nas quais é empregada a fermentação literal, quanto as modalidades nas quais são empregados outros modos de cultura não fermentativos. A fermentação pode ser efetuada em qualquer escala. Nas modalidades da presente invenção, o meio de fermentação pode ser sele-
 25 cionado a partir dentre meios ricos, meios mínimos, e meios de sais mine-rais; um meio rico pode também ser usado. Em uma outra modalidade, um meio mínimo ou um meio de sais minerais é selecionado. Ainda em uma ou- tra modalidade, um meio mínimo é selecionado. Em mais uma outra modali-
 30 são geralmente usados.

O meio de sais minerais consiste em sais minerais e uma fonte de carbono, tal como, por exemplo, glicose, sacarose, ou glicerol. Os exem-

plos de meios de sais minerais incluem, por exemplo, o meio M9, o meio de *Pseudomonas* (ATCC 179), o meio de Davis e Mingioli (ver, BD Davis e ES Mingioli (1950) em *J. Bact.* 60:17-28). Os sais minerais usados para preparar os meios de sais minerais incluem aqueles selecionados a partir dentre, por exemplo, fosfatos de potássio, sulfato ou cloreto de amônio, sulfato ou cloreto de magnésio, e minerais residuais, tais como cloreto de cálcio, borato, e sulfatos de ferro, cobre, manganês, e zinco. Nenhuma fonte de nitrogênio orgânico, tal como peptona, triptona, aminoácidos, ou um extrato de levedura, é incluído em um meio de sais minerais. Em vez disso, uma fonte de nitrogênio inorgânico é usada e esta pode ser selecionada a partir dentre, por exemplo, sais de amônio, amônia aquosa, e amônia gasosa. Um meio de sais minerais pode conter glicose como a fonte de carbono. Em comparação com os meios de sais minerais, os meios mínimos podem também conter sais minerais e uma fonte de carbono, porém podem ser suplementados com, por exemplo, baixos níveis de aminoácidos, vitaminas, peptonas, ou outros ingredientes, embora estes sejam adicionados em níveis muito mínimos.

Em uma modalidade, os meios podem ser preparados usando os diversos componentes listados abaixo. Os componentes podem ser adicionados na seguinte ordem: primeiro o $(\text{NH}_4)\text{HPO}_4$, o KH_2PO_4 e ácido cítrico podem ser dissolvidos em aproximadamente 30 litros de água destilada; então, uma solução de microelementos pode ser adicionada, seguida pela adição de um agente antiespuma, tal como o Ucolub N 115. Então, após a esterilização a quente (tal como em aproximadamente 121 graus C), as soluções estéreis de glicose MgSO_4 e tiamina-HCl podem ser adicionadas. O controle do pH em aproximadamente 6,8 pode ser atingido usando amônia aquosa. A água destilada estéril pode então ser adicionada para ajustar o volume inicial para 371 menos o estoque de glicerol (123 mL). As substâncias químicas estão comercialmente disponíveis a partir de diversos fornecedores, tais como a Merck. Este meio pode permitir um cultivo em densidade de células alto (HCDC) para o crescimento da espécie de *Pseudomonas* e bactérias relacionadas. O HCDC pode iniciar como um processo em bate-

lada, o qual é seguido por um cultivo de batelada alimentada de duas fases. Após o crescimento ilimitado na parte da batelada, o crescimento pode ser controlado em uma taxa específica reduzida por um período de 3 tempos de duplicação, nos quais a concentração de biomassa pode ser aumentada diversas vezes. Os detalhes adicionais de tais procedimentos de cultivo são descritos em Riesenber, D.; Schulz, V.; Knorre, W. A.; Pohl, H. D.; Korz, D.; Sanders, E. A.; Ross, A.; Deckwer, W. D. (1991) "High cell density cultivation of *Escherichia coli*, at controlled specific growth rate" *J Biotechnol*: 20(1) 17-27.

5 TABELA-US- 00005 TABELA 5 Composição do meio Concentração inicial do componente KH_2PO_4 13,3 g l^{-1} $(\text{NH}_4)_2\text{HPO}_4$ 4,0 g l^{-1} Ácido cítrico 1,7 g l^{-1} $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ 1,2 g l^{-1} Solução de metais residuais 10 ml l^{-1} Tiamina HCl 4,5 mg l^{-1} Glicose- H_2O 27,3 g l^{-1} Antiespuma Ucolub N115 0,1 ml l^{-1} Solução de alimentação $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ 19,7 g l^{-1} Glicose- H_2O 770 g l^{-1} NH_3 23 g l^{-1} Solução de metais residuais 6 g l^{-1} citrato de Fe(III) 1,5 g l^{-1} $\text{MnCl}_2 \cdot 4\text{H}_2\text{O}$ 0,8 g l^{-1}

10 $\text{ZnCH}_2\text{COO l}_2 \cdot 2\text{H}_2\text{O}$ 0,3 g l^{-1} H_3BO_3 0,25 g l^{-1} $\text{Na}_2\text{MoO}_4 \cdot 2\text{H}_2\text{O}$ 0,25 g l^{-1} $\text{CoCl}_2 \cdot 6\text{H}_2\text{O}$ 0,15 g l^{-1} $\text{CuCl}_2 \cdot 2\text{H}_2\text{O}$ 0,84 g l^{-1} sal de Na_2 de ácido etileno diaminotetracético $2\text{H}_2\text{O}$ (Titriplex III, Merck).

As seqüências descritas neste pedido podem ser homólogas (terem identidade similar). As proteínas e/ou as seqüências de proteínas são "homólogas" quando elas forem derivadas, natural ou artificialmente, de uma proteína ou seqüência de proteínas ancestral comum. Similarmente, os ácidos nucleicos e/ou as seqüências de ácidos nucleicos são homólogas quando eles forem derivados, natural ou artificialmente, de um ácido nucleico ou seqüência de ácidos nucleicos ancestral comum. Por exemplo, qualquer ácido nucleico de ocorrência natural pode ser modificado por qualquer método de mutagênese disponível para incluir um ou mais códons seletores. Quando expresso, este ácido nucleico mutagenizado codifica um polipeptídeo compreendendo um ou mais aminoácidos não naturais. O processo de mutação pode, obviamente, alterar de modo adicional um ou mais códons padrões, desse modo alterando um ou mais aminoácidos padrões na proteína mutante resultante também. A homologia é geralmente deduzida a partir da similaridade das seqüências entre dois ou mais ácidos nucleicos ou proteínas (ou

20

25

30

suas seqüências). A porcentagem exata de similaridade entre as seqüências, que é útil em estabelecer a homologia, varia com o ácido nucléico e a proteína em questão, porém usa-se rotineiramente tão pouco quanto 25% de similaridade das seqüências para estabelecer a homologia. Podem também
5 ser usados níveis mais elevados de similaridade das seqüências, por exemplo, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 96%, 97%, 98% ou 99% ou mais, para estabelecer a homologia. Os métodos para determinar as porcentagens de similaridades das seqüências (por exemplo, BLASTP e BLASTN usando parâmetros preestabelecidos) são descritos neste documento e estão, em geral, disponíveis.
10

Os polipeptídeos podem compreender uma seqüência de sinal (ou líder) na extremidade N terminal da proteína, a qual orienta de modo contraducional ou pós-traducional a transferência da proteína. O polipeptídeo pode também ser conjugado a um ligador ou outra seqüência para a facilidade da síntese, purificação ou identificação do polipeptídeo (por exemplo, poli-His), ou para aumentar a ligação do polipeptídeo a um suporte sólido.
15

Quando comparando as seqüências de polipeptídeos, duas seqüências são ditas serem "idênticas" se a seqüência de aminoácidos nas duas seqüências for igual quando alinhada para a correspondência máxima, como descrito abaixo. As comparações entre as duas seqüências são tipicamente efetuadas comparando-se a seqüência sobre uma janela de comparação para identificar e comparar as regiões locais da similaridade das seqüências. Uma "janela de comparação", como usada neste documento, refere-se a um segmento de pelo menos cerca de 20 posições contíguas,
20 normalmente 30 a cerca de 75, 40 a cerca de 50, em que uma seqüência pode ser comparada a uma seqüência de referência do mesmo número de posições contíguas após as duas seqüências serem alinhadas de modo ótimo.
25

O alinhamento ótimo das seqüências para comparação pode ser
30 conduzido usando o programa Megalign no software integrado Lasergene do software bioinformatics (DNASTAR, Inc., Madison, Wis.), usando parâmetros preestabelecidos. Este programa inclui diversos esquemas de alinhamento

- descritos nas seguintes referências: Dayhoff, M. O. (1978) A model of evolutionary change in proteins - Matrices for detecting distant relationships. Em Dayhoff, M. O. (ed.) Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington D.C. Vol. 5, Supl. 3, págs. 345
- 5 358; Hein J. (1990) Unified Approach to Alignment and Phylogenies págs. 626 645 Methods in Enzymology vol. 183, Academic Press, Inc., San Diego, Calif.; Higgins, D. G. e Sharp, P. M. (1989) CABIOS 5:151 153; Myers, E. W. e Muller W. (1988) CABIOS 4:11 17; Robinson, E. D. (1971) Comb. Theor 11:105; Santou, N. Nes, M. (1987) Mol. Biol. Evol. 4:406 425; Sneath, P. H.
- 10 A. e Sokal, R. R. (1973) Numerical Taxonomy -- the Principles and Practice of Numerical Taxonomy, Freeman Press, San Francisco, Calif.; Wilbur, W. J. e Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. EUA* 80:726 730.

Alternativamente, o alinhamento ótimo das seqüências para comparação pode ser conduzido pelo algoritmo de identidade local de Smith

15 e Waterman (1981) *Add. APL. Math* 2:482, pelo algoritmo de alinhamento de identidade de Needleman e Wunsch (1970) *J. Mol. Biol.* 48:443, pela busca por métodos de similaridade de Pearson e Lipman (1988) *Proc. Natl. Acad. Sci. EUA* 85: 2444, por implementações computadorizadas destes algoritmos (GAP, BESTFIT, BLAST, FASTA, e TFASTA no Pacote de Softwares

20 Wisconsin Genetics, Genetics Computer Group (GCG), 575 Science Dr., Madison, Wis.), ou por inspeção.

Um exemplo de algoritmos que podem ser adequados para determinar a porcentagem de identidade das seqüências e similaridade das seqüências são os algoritmos BLAST e BLAST 2.0, que são descritos em

25 Altschul et al. (1977) *Nucl. Acids Res.* 25:3389 3402 e Altschul et al. (1990) *J. Mol. Biol.* 215:403 410, respectivamente. O BLAST e o BLAST 2.0 podem ser usados, por exemplo, com os parâmetros descritos neste documento, para determinar a porcentagem de identidade das seqüências para os polinucleotídeos e os polipeptídeos da invenção. O software para efetuar as análises por BLAST está publicamente disponível através do National Center

30 for Biotechnology Information. Para as seqüências de aminoácidos, uma matriz de pontuação pode ser usada para calcular a pontuação cumulativa. A

extensão dos acertos de palavras em cada direção é interrompida quando: a pontuação cumulativa de alinhamento diminui pela quantidade X a partir de seu valor máximo atingido; a pontuação cumulativa vai para zero ou abaixo devido ao acúmulo de um ou mais alinhamentos de resíduos de pontuação negativa; ou o fim de uma ou outra seqüência é atingido. Os parâmetros do algoritmo BLAST W, T e X determinam a sensibilidade e a velocidade do alinhamento.

Em uma abordagem, a "porcentagem de identidade das seqüências" é determinada comparando-se duas seqüências otimamente alinhadas sobre uma janela de comparação de pelo menos 20 posições, onde a porção da seqüência de polipeptídeos na janela de comparação pode compreender adições ou remoções (isto é, lacunas) de 20 por cento ou menos, normalmente 5 a 15 por cento, ou 10 a 12 por cento, em comparação com as seqüências de referência (que não compreendem adições ou remoções) para o alinhamento ótimo das duas seqüências. A porcentagem é calculada determinando-se o número de posições nas quais ocorre o resíduo de aminoácido idêntico em ambas as seqüências para produzir o número de posições correlacionadas, dividindo-se o número de posições correlacionadas pelo número total de posições na seqüência de referência (isto é, o tamanho da janela) e multiplicando-se os resultados por 100 para produzir a porcentagem de identidade das seqüências.

Dentro de outras modalidades ilustrativas, as seqüências otimizadas de códons podem incluir um polipeptídeo, o qual pode ser um polipeptídeo de fusão que compreende os polipeptídeos múltiplos como descritos neste documento, ou que compreende pelo menos um polipeptídeo como descrito neste documento e uma seqüência não relacionada, tal como uma proteína de tumor conhecida. Um par de fusão pode, por exemplo, auxiliar em proporcionar epítomos auxiliares T (um par de fusão imunológico), preferivelmente epítomos auxiliares T reconhecidos por seres humanos, ou pode auxiliar em expressar a proteína (uma seqüência reforçadora da expressão) em rendimentos maiores do que a proteína recombinante nativa. Certos pares de fusão preferidos são pares de fusão tanto imunológicos

quanto reforçadores da expressão. Outros pares de fusão podem ser selecionados de modo a aumentar a solubilidade do polipeptídeo ou capacitar o polipeptídeo de ser alvejado para os compartimentos intracelulares desejados. Ainda adicionalmente, os pares de fusão incluem as marcas de afinidade, que facilitam a purificação do polipeptídeo.

Os polipeptídeos de fusão podem geralmente ser preparados usando técnicas padrões, incluindo a conjugação química. De preferência, um polipeptídeo de fusão é expresso como um polipeptídeo recombinante, permitindo a produção de níveis aumentados, em relação a um polipeptídeo não fundido, em um sistema de expressão. Resumidamente, as seqüências de ácidos nucléicos codificando os componentes de polipeptídeos podem ser formadas separadamente, e ligadas em um vetor de expressão apropriado. A extremidade de 3' da seqüência de DNA codificando um componente de polipeptídeo é ligada, com ou sem um ligador de peptídeo, à extremidade de 5' de uma seqüência de DNA codificando o segundo componente de polipeptídeo de modo que os quadros de leitura das seqüências estejam em fase. Isto permite a tradução para um polipeptídeo de fusão individual que conserva a atividade biológica de ambos os polipeptídeos componentes.

Uma seqüência ligadora de peptídeo pode ser empregada para separar o primeiro e o segundo componentes de polipeptídeos por uma distância suficiente para assegurar que cada polipeptídeo duplique em suas estruturas secundárias e terciárias. Tal seqüência ligadora de peptídeo é incorporada no polipeptídeo de fusão usando práticas padrões bastante conhecidas na técnica. As seqüências ligadoras de peptídeos adequadas podem ser escolhidas com base nos seguintes fatores: (1) sua capacidade de adotar uma conformação estendida flexível; (2) sua incapacidade de adotar uma estrutura secundária que poderia interagir com os epítomos funcionais sobre o primeiro e o segundo polipeptídeos; e (3) a ausência de resíduos hidrofóbicos ou carregados que poderiam reagir com os epítomos funcionais sobre o polipeptídeo. As seqüências ligadoras de peptídeos preferidas contêm resíduos Gly, Asn e Ser. Outros aminoácidos neutros próximos, tais como Thr e Ala, podem também ser usados na seqüência ligadora. As seqüên-

cias de aminoácidos que podem ser empregadas utilmente como ligadoras incluem aquelas divulgadas em Maratea et al., *Gene* 40:39 46, 1985; Murphy et al., *Proc. Natl. Acad. Sci. EUA* 83:8258 8262, 1986; Pat. U.S. Nº 4.935.233 e Pat. U.S. Nº 4.751.180. A seqüência ligadora pode geralmente ser de 1 a cerca de 50 aminoácidos de comprimento. As seqüências ligadoras não são requeridas quando o primeiro e o segundo polipeptídeos tiverem regiões de aminoácidos N terminais não essenciais que podem ser usadas para separar os domínios funcionais e impedir a interferência estérica.

As seqüências de DNA ligadas são operavelmente ligadas a elementos reguladores transcricionais ou traducionais adequados. Os elementos reguladores, responsáveis pela expressão do DNA, estão localizados somente 5' em relação à seqüência de DNA codificando os primeiros polipeptídeos. Similarmente, os códons de interrupção requeridos para finalizar os sinais de terminação de tradução e transcrição estão somente presentes 3' em relação à seqüência de DNA codificando o segundo polipeptídeo.

A presente invenção também proporciona a análise em série e a geração de relatório automáticas de um gene usando um banco de dados e ferramentas para calcular o uso do códon a partir de uma seqüência bruta e graficamente relatar a posição dos códons raros ao longo da seqüência de DNA traduzida. Diversas novas ferramentas foram desenvolvidas para auxiliar neste processo, onde a análise e a geração de relatório são completadas automaticamente, reduzindo o tempo requerido gasto por um pesquisador.

Nos estágios iniciais do esboço do projeto, uma seqüência de codificação de proteína pode ser avaliada para determinar se é conveniente a otimização de todo ou parte do gene. Embora não haja critério absoluto em fazer esta determinação, uma estratégia envolve a avaliação da porcentagem e a distribuição de códons que seriam considerados raramente preferidos para um aminoácido particular no sistema de expressão de hospedeiro. Os valores de 5% e 10% de uso são comumente usados como valores de corte para a determinação dos códons raros. Por exemplo, os códons listados na Tabela 1 têm uma ocorrência calculada de menos do que 5% no genoma de MB214, e seriam preferencialmente evitados em um gene otimiz-

do a ser expresso neste hospedeiro. Para verificar se um gene de interesse poderia ser expresso de modo heterólogo sem a otimização, pode-se determinar que porcentagem de códons raros existe naquele gene e se eles residem em posições que poderiam ter um efeito deletério sobre a expressão (isto é, próximos à extremidade de 5' do gene ou concentrados juntos nos agrupamentos).

Para lidar com estas questões, a ferramenta da presente invenção é projetada para calcular o uso do códon a partir de uma sequência de ORF bruta e graficamente relatar a posição dos códons raros ao longo de uma sequência de DNA traduzida. Adicionalmente, uma tabela de cores codificadas pode ser apresentada para comparar o uso do códon do gene apresentado com aquele dos códons preferenciais de referência de MB214. Para permitir a portabilidade, remover a dependência sobre qualquer pacote de bioinformática de suporte particular e proporcionar a facilidade de uso, a nova ferramenta pode ser escrita como um programa CGI inteiramente na linguagem de programação Perl, e ser acessível como um formulário via um navegador.

Em uso, uma sequência de nucleotídeos não formatada é colocada no formulário e apresentada, e os relatórios formatados são retornados. Os resultados da amostra são mostrados nas Figuras 2 e 3, e na Tabela 2.

Tabela 2

AA	Códon	Uso MB214	Uso do Gene	Diferença	AA	Códon	Uso MB214	Uso do Gene	Diferença
A	GCA	9,82	18,18	-8,36	N	AAC	75,63	63,64	11,99
A	GCC	50,65	9,09	41,56	N	AAT	24,37	36,36	-11,99
A	GCG	30,30	0,00	30,30	P	CCA	12,52	66,67	-54,15
A	GCT	9,23	72,73	-63,50	P	CCC	26,76	0,00	26,76
C	TGC	79,33	0,00	79,33	P	CCG	48,94	0,00	48,94
C	TGT	20,67	100,00	-79,33	P	CCT	11,78	33,33	-21,55
D	GAC	67,10	41,67	25,43	Q	CAA	34,05	66,67	-32,62
D	GAT	32,90	58,33	-25,43	Q	CAG	65,95	33,33	32,62
E	GAA	54,89	27,27	27,62	R	AGA	1,39	27,27	-25,88
E	GAG	45,11	72,73	-27,62	R	AGG	2,72	9,09	-6,37
F	TTC	68,10	62,50	5,60	R	CGA	4,99	0,00	4,99
F	TTT	31,90	37,50	-5,60	R	CGC	55,83	0,00	55,83

AA	Códon	Uso MB214	Uso do Gene	Dife- rença	AA	Códon	Uso MB214	Uso do Gene	Dife- rença
G	GGA	3,26	28,57	-25,31	R	CGG	13,81	18,18	-4,37
G	GGC	63,53	0,00	63,53	R	CGT	21,25	45,45	-24,20
G	GGG	13,92	28,57	-14,65	S	AGC	40,26	12,50	27,76
G	GGT	19,29	42,86	-23,57	S	AGT	9,69	37,50	-27,81
H	CAC	63,13	50,00	13,13	S	TCA	5,19	25,00	-19,81
H	CAT	36,87	50,00	-13,13	S	TCC	18,92	12,50	6,42
I	ATA	3,05	18,18	-15,13	S	TCG	21,75	0,00	21,75
I	ATC	73,52	18,18	55,34	S	TCT	4,18	12,50	-8,32
I	ATT	23,44	63,64	-40,20	T	ACA	5,95	18,18	-12,23
K	AAA	36,51	9,09	27,42	T	ACC	65,39	45,45	19,94
K	AAG	63,49	90,91	-27,42	T	ACG	19,72	0,00	19,72
L	CTA	1,78	0,00	1,78	T	ACT	8,94	36,36	-27,42
L	CTC	14,12	13,64	0,48	V	GTA	8,97	0,00	8,97
L	CTG	57,10	22,73	34,37	V	GTC	26,59	8,33	18,26
L	CTT	4,57	31,82	-27,25	V	GTG	56,02	41,67	14,35
L	TTA	1,89	4,55	-2,66	V	GTT	8,43	50,00	-41,57
L	TTG	20,55	27,27	-6,72	W	TGG	100,00	100,00	0,00
M	ATG	100,00	100,00	0,00	Y	TAC	69,23	80,00	-10,77
					Y	TAT	30,77	20,00	10,77

A Tabela 2 representa uma tabela da freqüência de códons, listando para cada par de aminoácido/códon: i) a porcentagem de freqüência do códon no MB214, ii) a porcentagem de freqüência do códon no gene analisado, e iii) a porcentagem de diferença entre o uso no gene analisado versus o MB214. O destaque indica o uso do códon no MB214 de menos do que 10%. O destaque de valores "0,00" na coluna de Uso do Gene indica um códon raro que não é usado na seqüência analisada.

As Figuras 2 e 3 ilustram os resultados dos perfis de uso dos códons raros, mostrando a posição e a distribuição dos códons raros ao longo de uma seqüência de proteína traduzida. Os códons destacados são representados com menos do que 5% e 10% de freqüência na cepa de *P. fluorescens* MB214 nas Figuras 2 e 3, respectivamente. A porcentagem global e o número absoluto de códons incidindo abaixo de 5% ou 10% de uso são também indicados seguindo a seqüência traduzida nas Figuras 2 e 3, respectivamente.

O banco de dados e as ferramentas para análise dos genes oti-

mizados são também proporcionados. Assim que um gene tiver sido analisado e uma determinação feita que a síntese de uma versão otimizada do gene é garantida, uma ou mais versões sintéticas do gene podem ser projetadas. Cada um dos candidatos de projeto do gene resultantes pode ser
5 analisado antes da síntese, para assegurar o cumprimento com todos os critérios de projeto. Para guardar a trilha dos genes apresentados, os critérios de projeto associados, e as versões de candidatos sintéticos resultantes a serem analisadas, um banco de dados relacional é proporcionado para armazenar esta informação.

10 Para funcionar com o código Perl existente em um ambiente Linux, em uma modalidade particular da invenção, o PostgreSQL foi selecionado como o banco de dados relacional. Os dados podem ser introduzidos no, e extraídos do, banco de dados criado usando, por exemplo, o módulo DBI de Perl. O esquema do banco de dados pode ser projetado para permitir
15 a flexibilidade em selecionar elementos a serem incluídos na unidade de transcrição sintética (por exemplo, seqüência de proteína, seqüência líder, e UTR's). Os vetores de expressão e os hospedeiros podem ser definidos para assegurar a compatibilidade do gene sintético com os múltiplos sítios de clonagem dos vetores e os códons preferenciais dos hospedeiros. Os motivos
20 que devem ser evitados na seqüência final podem também ser definidos, e as versões sintéticas candidatas para cada gene podem ser armazenadas. Uma modalidade representativa do esquema do banco de dados para o banco de dados do gene é ilustrada na figura 4, com os nomes de arquivo no banco de dados real representados em letra minúscula.

25 Para facilitar a entrada de dados no banco de dados sem requerer experiência em SQL, em uma modalidade particular da invenção, uma interface com o usuário foi desenvolvida consistindo em formulários HTML gerados por CGI. A interface com o usuário pode também proporcionar uma camada de checagem de erros para assegurar que todos os valores introduzidos sejam válidos.
30

A introdução de um novo gene requer um formulário HTML gerado por CDI completado e o aperto de um botão ENVIAR. Os valores po-

dem ser introduzidos no formulário livremente nas caixas de texto ou selecionados a partir de menus predefinidos ativar e caixa de seleção. Estes menus podem ser construídos automaticamente a partir dos valores atualmente disponíveis no banco de dados. Novos valores podem ser adicionados para cada menu clicando-se em um respectivo hiperlink "Adicionar", o
 5 que gera um novo formulário HTML específico para aquela entrada de dados. Se forem detectados erros no envio, o usuário pode ser retornado para o formulário e apresentado com mensagens descrevendo as correções necessárias que devem ser feitas. Todos os valores anteriormente introduzidos
 10 podem ser conservados sobre o formulário de modo que somente os valores relacionados ao erro possam ser modificados ou introduzidos novamente.

Após introduzir um novo gene, uma cotação pode ser requerida de um fornecedor externo para o projeto e a síntese do gene candidato/unidade de transcrição. O processo pode ser iniciado introduzindo informação na página do site da internet do fornecedor. Para facilitar este processo e impedir erros de entrada de dados, uma ferramenta pode ser proporcionada que permita a preparação dos dados necessários diretamente do banco de dados para o formato requerido. Esta ferramenta pode permitir que um usuário gere a informação requerida para uma cotação selecionando um
 15 nome do gene a partir de um menu ativar automaticamente gerado de todos os genes disponíveis no banco de dados, na hora que a página foi carregada. Assim que um gene for selecionado, o pressionamento de um botão ENVIAR gera um formulário com três campos que pode ser colocado diretamente no formulário de requisição de cotação do fornecedor. Um hiperlink para esta
 20 página pode também ser proporcionado.

Devido à redundância no código genético, existem diversas seqüências de codificação diferentes que podem ser geradas para um candidato de gene sintético. Os fornecedores tipicamente proporcionarão múltiplas versões sintéticas candidatas para cada gene, para permitir que um pesquisador selecione a versão que mais proximamente seja compatível com os
 30 critérios de projeto requeridos. Estas seqüências podem ser adicionadas ao banco de dados e associadas com o envio do respectivo gene usando a re-

de. O nome do gene pode então ser selecionado a partir de um menu ativar automaticamente gerado, e um número da versão, a seqüência, e quaisquer comentários descritivos podem ser introduzidos. Assim que enviado, o pipeline de análise automatizado pode ser corrido para determinar qual das versões enviadas no banco de dados é a melhor para a síntese.

Um programa (por exemplo, um programa Perl) pode ser incluído para automatizar o processo de avaliação de cada versão sintética candidata, para assegurar o cumprimento com os critérios de projeto como enviados para o banco de dados. Cada versão de gene sintético pode ser extraída do banco de dados, juntamente com as especificações de projeto relevantes, e corrida através de uma série de análises. Esta análise pode incluir uma ou mais das que seguem:

1) O GCG (disponível do Accelrys Software, Inc., San Diego, CA) CODONFREQUENCY pode ser corrido para determinar o uso do códon da versão sintética. Os arquivos de saída são analisados e a presença de quaisquer códons raros, definida por uma porcentagem do valor de corte armazenado no banco de dados para cada gene, pode ser detectada;

2) O GCG MAPSORT pode ser corrido para determinar a presença de quaisquer enzimas de restrição indesejadas que possam interferir com a subclonagem futura. A lista de enzimas de restrição avaliadas pode ser extraída do banco de dados através de relações entre enzimas, vetores de expressão, e genes. Os arquivos de saída podem ser analisados para detectar a presença de qualquer sítio de restrição a partir da lista de enzimas;

3) O GCG FINDPATTERNS pode ser corrido para detectar a presença de quaisquer motivos de seqüência que devam ser evitados na versão sintética. Cada padrão pode ser definido no banco de dados, juntamente com o número de combinações inadequadas toleradas para aquele padrão específico. Os arquivos de saída podem ser analisados para detectar a presença de quaisquer dos motivos de seqüência deletérios definidos;

4) Um programa (por exemplo, um programa Perl) pode ser corrido para detectar a resistência de quaisquer estruturas de laço de haste

presentes. O programa pode sequencialmente correr o GCG STEMLOOP para encontrar posições de laços de hastes putativas na seqüência, extrair as coordenadas destes laços, e então correr as coordenadas dos laços através do GCG MFOLD para determinar a energia livre da estrutura de laço. Os resultados de saída podem ser classificados por energia livre e os dados para os cinco laços mais fortes podem ser extraídos. Adicionalmente, a energia livre do laço mais forte pode ser relatada para propósitos comparativos; e

5) o GCG BESTFIT pode ser corrido para comparar as traduções de peptídeos das seqüências de DNA nativas e sintéticas, para assegurar que nenhuma mutação tenha sido introduzida por erro. As seqüências traduzidas podem ser geradas por GCG TRANSLATE. Os resultados de saída podem ser analisados e relatados.

Um relatório pode ser gerado no formato HTML para visualização ou impressão em um navegador da Web ou no Microsoft Word. O relatório pode incluir um relatório de resumo dos resultados das análises na forma de tabelas. Por exemplo, conforme ilustrado na Tabela 3, uma coluna pode ser proporcionada para cada versão sintética e uma linha para cada análise.

TABELA 3

Critérios	v1	v2	v3
Códons Raros			
≥5 G's ou C's			
Seqüência de SD de Gene Interno			
Estrutura de laço de haste de gene interno mais forte			
Sítios de restrição únicos			
A proteína codificada pelo gene sintético é idêntica à seqüência de proteína original			

Neste modo, um pesquisador pode comparar os resultados para cada versão e selecionar a versão mais adequada para a síntese. Se a análise indicar que nenhuma das versões atende os critérios de projeto, podem

ser requeridas versões adicionais e a análise pode ser corrida novamente até que uma versão adequada seja obtida. O relatório pode também incluir os dados brutos a partir de cada análise, para propósitos de documentação. Os dados para cada versão de gene podem ser organizados por análise efetuada e as partes relevantes dos dados de saída podem ser destacadas para a facilidade de leitura.

A presente invenção é explicada em maior detalhe nos Exemplos que se seguem. Estes exemplos são pretendidos como ilustrativos da invenção e não são para serem considerados como limitação dela.

10 EXEMPLOS

EXEMPLO 1

Projeto de Gene Sintético a partir de *P. fluorescens*

Uma região do DNA contendo uma sequência de Shine-Dalgarno ótima e um sítio de enzima de restrição *SpeI* único foi adicionada a montante da sequência de codificação. Uma região do DNA contendo três códons de interrupção e um sítio de enzima de restrição *XhoI* único foi adicionada a jusante da sequência de codificação. Todos os códons raros ocorrendo no *Pfenex* ORFome com menos do que 5% de uso do códon foram modificados para evitar o bloqueio ribossômico. Todos os sítios ligadores ribossômicos de genes internos que eram iguais ao padrão *aggaggt_n5-10dtg* com duas combinações inadequadas ou menos foram modificados para evitar os produtos de proteínas truncadas. As extensões de cinco ou mais nucleotídeos C, ou cinco ou mais G foram eliminadas para evitar o deslizamento da RNA polimerase. As estruturas de laços de hastes de genes internos fortes, especialmente aquelas cobrindo o sítio ligador ribossômico, foram modificadas. O gene sintético foi sintetizado por DNA2.0, Inc. (Menlo Park, CA).

EXEMPLO 2

Projeto de Gene Sintético a partir de *P. fluorescens*

30 Os aminoácidos da metionina 21 até a glutamina 520 foram incluídos no produto de proteína expresso final. Todos os códons raros ocorrendo no *Pfenex* ORFome com menos do que 5% de uso do códon foram

modificados para evitar o bloqueio ribossômico. Todos os sítios ligadores ribossômicos de genes internos que eram iguais ao padrão $\text{aggaggt}_{5-10}\text{dtg}$ com duas combinações inadequadas ou menos foram modificados para evitar os produtos de proteínas truncadas. As extensões de cinco ou mais nucleotídeos C, ou cinco ou mais G foram eliminadas para evitar o deslizamento da RNA polimerase. As estruturas de laços de hastes de genes internos fortes, especialmente aquelas cobrindo o sítio ligador ribossômico, foram modificadas. Uma sequência de DNA codificando o líder de secreção periplasmático *pbp* de 24 aminoácidos foi fundida à extremidade de 5' da sequência otimizada. Uma região do DNA contendo uma sequência de Shine-Dalgarno ótima e um sítio de enzima de restrição *SpeI* único foi adicionada a montante da sequência de codificação. Uma região do DNA contendo três códon de interrupção e um sítio de enzima de restrição *XhoI* único foi adicionada a jusante da sequência de codificação. O gene sintético foi sintetizado por DNA2.0, Inc.

A presente invenção não é para ser limitada no escopo pelas modalidades específicas descritas neste documento. De fato, diversas modificações da invenção, além daquelas descritas aqui, tornar-se-ão aparentes para aqueles versados na técnica a partir da descrição precedente. Tais modificações são pretendidas incidir no escopo das reivindicações em anexo.

REIVINDICAÇÕES

1. Método de produzir uma proteína recombinante compreendendo:

5 otimizar uma seqüência de polinucleotídeo sintético para a expressão heteróloga em uma bactéria de *Pseudomonas fluorescens* hospedeira, em que o polinucleotídeo sintético compreende uma seqüência de nucleotídeos codificando uma proteína;

ligar a seqüência de polinucleotídeo sintético otimizada a um vetor de expressão;

10 transformar a bactéria de *Pseudomonas fluorescens* hospedeira com o vetor de expressão;

cultivar a bactéria de *Pseudomonas fluorescens* hospedeira transformada em um meio de cultura adequado, apropriado para a expressão da proteína; e

15 isolar a proteína.

2. Método de acordo com a reivindicação 1, em que a otimização da seqüência de polinucleotídeo sintético para a expressão heteróloga na bactéria de *Pseudomonas fluorescens* hospedeira adicionalmente compreende identificar e modificar os códons raros a partir da seqüência de polinucleotídeo sintético que sejam raramente usados na bactéria de *Pseudomonas fluorescens* hospedeira.

25 3. Método de acordo com a reivindicação 2, em que a otimização da seqüência de polinucleotídeo sintético para a expressão heteróloga na bactéria de *Pseudomonas fluorescens* hospedeira adicionalmente compreende identificar e modificar as seqüências internas putativas de sítios ligados ribossômicos a partir da seqüência de polinucleotídeo sintético.

30 4. Método de acordo com a reivindicação 2, em que a otimização da seqüência de polinucleotídeo sintético para a expressão heteróloga na bactéria de *Pseudomonas fluorescens* hospedeira adicionalmente compreende identificar e modificar as repetições alongadas dos nucleotídeos G ou C a partir da seqüência de polinucleotídeo sintético.

5. Método de acordo com a reivindicação 2, em que a otimização

da seqüência de polinucleotídeo sintético para a expressão heteróloga na bactéria de *Pseudomonas fluorescens* hospedeira adicionalmente compreende identificar e minimizar a estrutura secundária de mRNA nas regiões de RBS e codificação do gene da seqüência de polinucleotídeo sintético.

5 6. Método de acordo com a reivindicação 2, em que a otimização da seqüência de polinucleotídeo sintético para a expressão heteróloga na bactéria de *Pseudomonas fluorescens* hospedeira adicionalmente compreende identificar e modificar os sítios de restrição de enzimas indesejáveis a partir da seqüência de polinucleotídeo sintético.

10 7. Método de acordo com a reivindicação 2, em que a identificação e a modificação dos códons raros compreendem identificar e modificar os códons tendo uma ocorrência de menos do que 10% no genoma bacteriano de *Pseudomonas fluorescens*.

15 8. Método de acordo com a reivindicação 2, em que a identificação e a modificação dos códons raros compreendem identificar e modificar os códons tendo uma ocorrência de menos do que 5% no genoma bacteriano de *Pseudomonas fluorescens*.

20 9. Método de acordo com a reivindicação 1, em que a otimização da seqüência de polinucleotídeo sintético para a expressão heteróloga adicionalmente compreende identificar e modificar os códons a partir da seqüência de polinucleotídeo sintético para aumentar a expressão.

 10. Método de acordo com a reivindicação 2, em que a modificação dos códons raros compreende substituir os códons raros por códons de ocorrência freqüente.

25 11. Método de produzir uma proteína recombinante compreendendo:

 identificar e modificar os códons raros a partir da seqüência de polinucleotídeo sintético que sejam raramente usados na bactéria de *Pseudomonas* hospedeira;

30 identificar e modificar as seqüências internas putativas de sítios ligadores ribossômicos a partir da seqüência de polinucleotídeo sintético;

 identificar e modificar as repetições alongadas dos nucleotídeos

G ou C a partir da sequência de polinucleotídeo sintético;

identificar e minimizar a estrutura secundária de mRNA nas regiões de RBS e codificação do gene da sequência de polinucleotídeo sintético;

identificar e modificar os sítios de restrição de enzimas indesejáveis a partir da sequência de polinucleotídeo sintético para formar uma sequência de polinucleotídeo sintético otimizada;

ligar a sequência de polinucleotídeo sintético otimizada a um vetor de expressão;

transformar a bactéria de *Pseudomonas* hospedeira com o vetor de expressão;

cultivar a bactéria de *Pseudomonas* hospedeira transformada em um meio de cultura adequado, apropriado para a expressão da proteína; e

isolar a proteína.

12. Método de acordo com a reivindicação 11, em que a bactéria de *Pseudomonas* hospedeira é o *Pseudomonas fluorescens*.

13. Método de acordo com a reivindicação 11, em que a bactéria de *Pseudomonas* hospedeira é a cepa de *Pseudomonas fluorescens* MB101.

14. Método de acordo com a reivindicação 12, em que a identificação e a modificação dos códons raros compreendem identificar e modificar os códons tendo uma ocorrência de menos do que 10% no genoma bacteriano de *Pseudomonas fluorescens*.

15. Método de acordo com a reivindicação 12, em que a identificação e a modificação dos códons raros compreendem identificar e modificar os códons tendo uma ocorrência de menos do que 5% no genoma bacteriano de *Pseudomonas fluorescens*.

16. Método de analisar os genes otimizados, compreendendo: proporcionar um banco de dados de otimização de genes para a bactéria de *Pseudomonas fluorescens*;

introduzir os dados dos genes no banco de dados;

identificar os vetores de expressão ou os hospedeiros;

emitir uma requisição de síntese de um gene candidato ou uni-

dade de transcrição;

adicionar as seqüências de genes otimizados no banco de dados;

5 avaliar uma ou mais versões sintéticas do(s) gene(s) candidato(s) sintetizado(s) para assegurar o cumprimento com a requisição de síntese; e

analisar a uma ou mais versões sintéticas do(s) gene(s) candidato(s).

10 17. Método de acordo com a reivindicação 16, adicionalmente compreendendo gerar um relatório dos resultados da análise da uma ou mais versões sintéticas do(s) gene(s) candidato(s).

18. Método de acordo com a reivindicação 16, em que a análise da uma ou mais versões sintéticas do(s) gene(s) candidato(s) compreende a análise do(s) gene(s) candidato(s) por inspeção ou computacionalmente.

15 19. Método de acordo com a reivindicação 16, em que a análise da uma ou mais versões sintéticas do(s) gene(s) candidato(s) compreende a análise do nível de expressão proporcionado pelo(s) gene(s) candidato(s).

20 20. Método de acordo com a reivindicação 16, em que a análise da uma ou mais versões sintéticas do(s) gene(s) candidato(s) compreende a análise da sua posse ou privação de teor de GC alto ou baixo, de um elemento de seqüência, ou da estrutura do(s) gene(s) candidato(s).

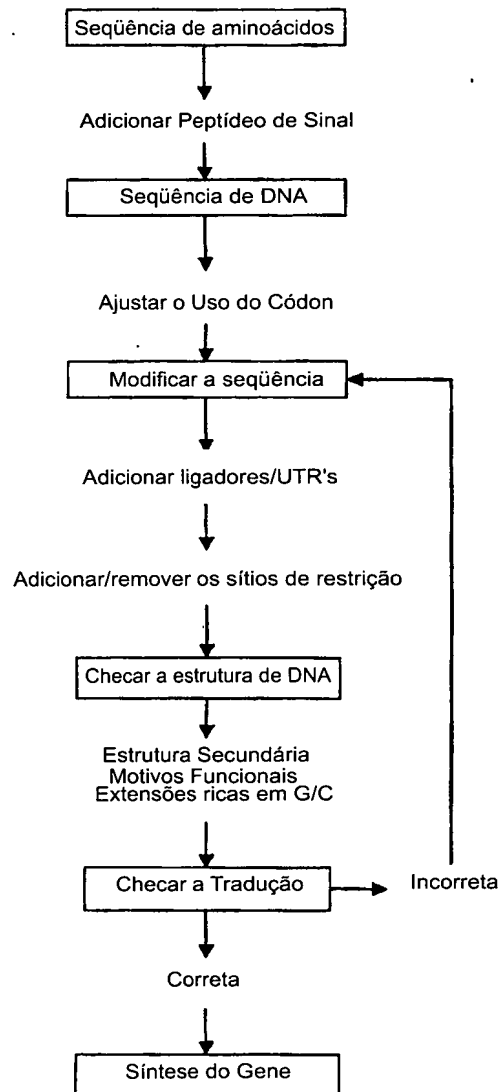


FIG. 1

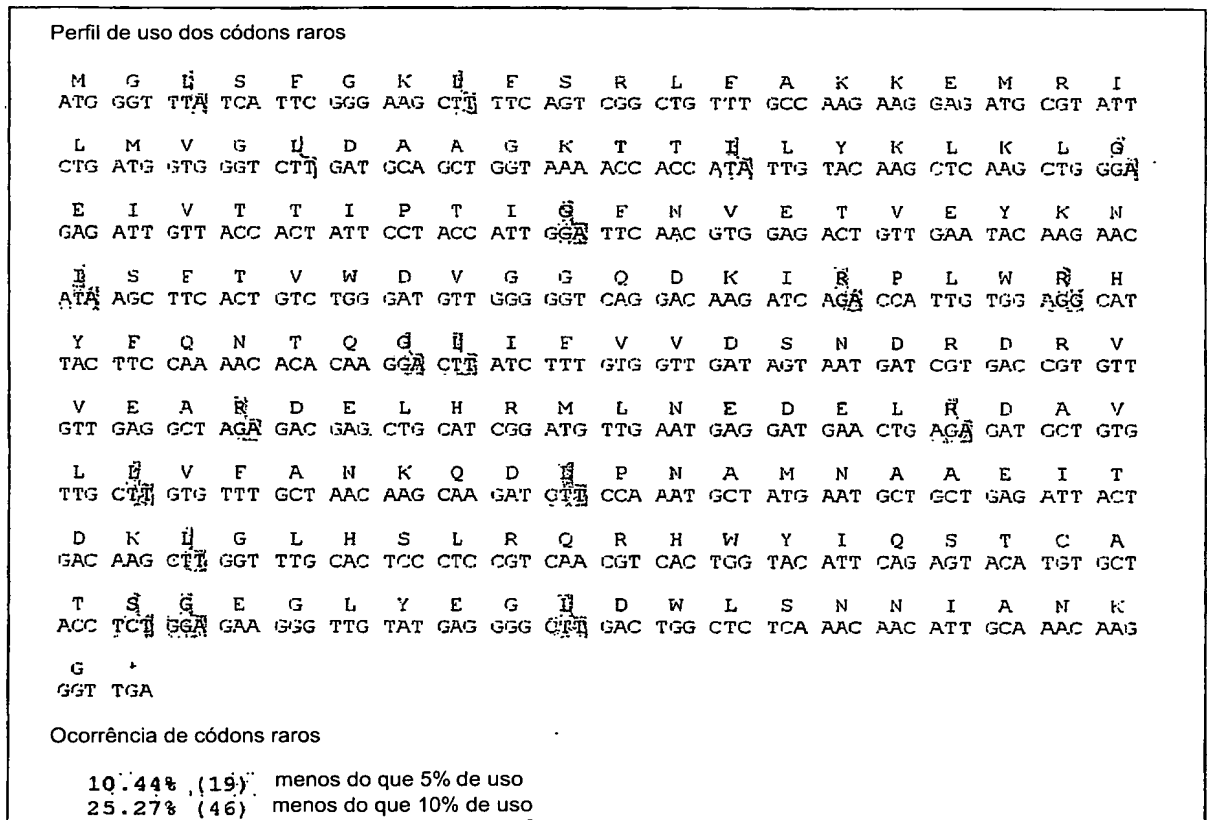


FIG. 2

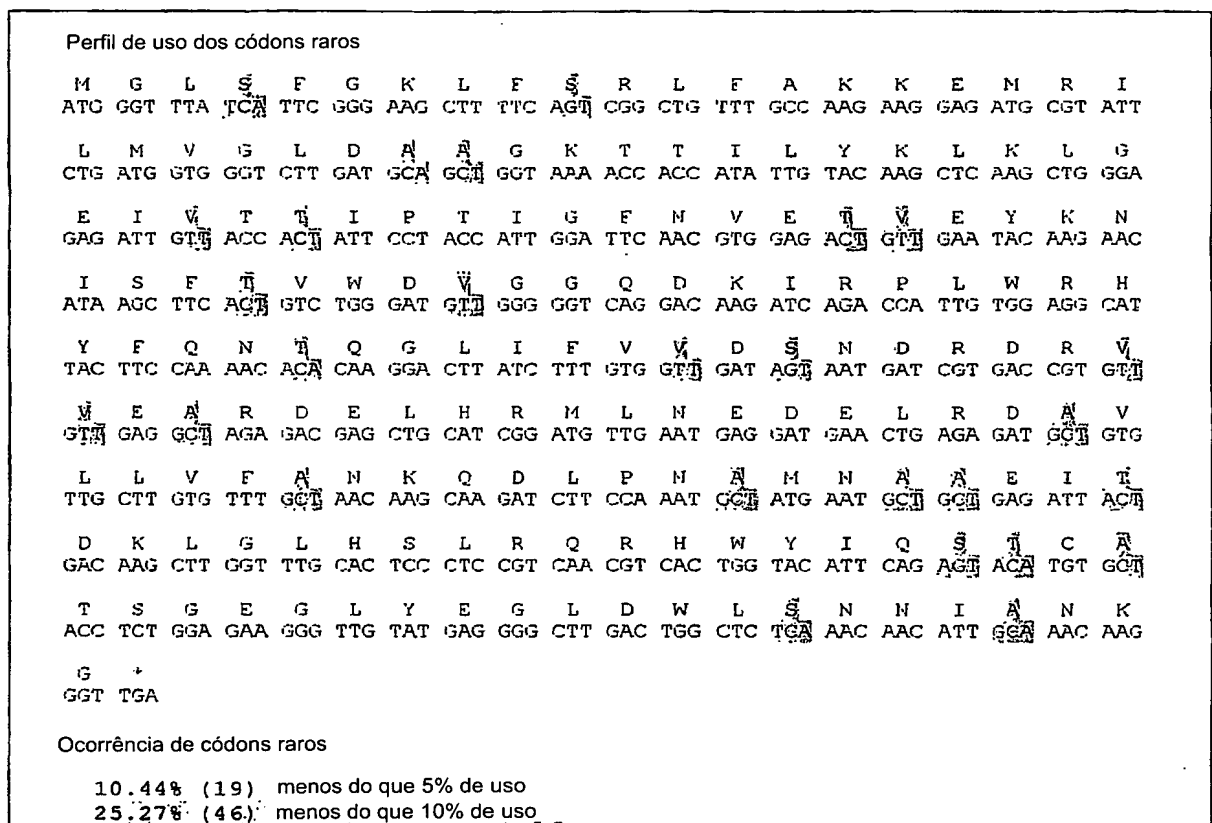


FIG. 3

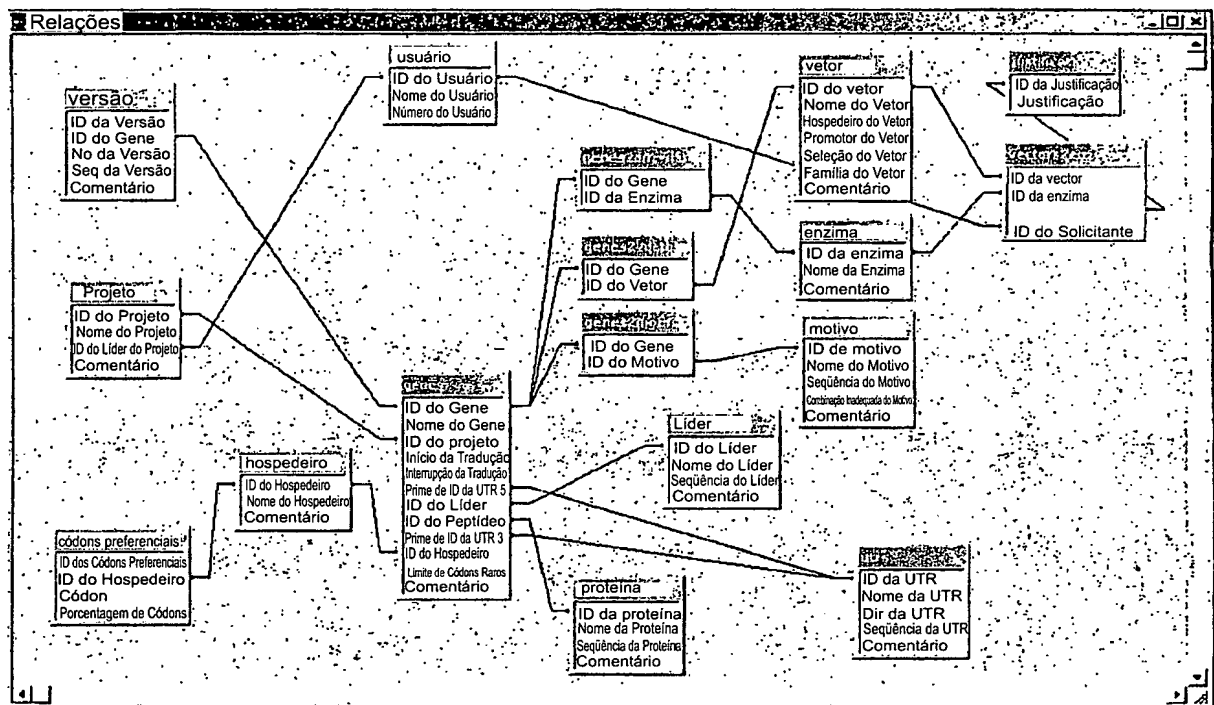


FIG. 4

RESUMO

Patente de Invenção: "MÉTODO DE OTIMIZAÇÃO DO CÓDON".

A presente invenção refere-se a uma expressão heteróloga em uma bactéria de *Pseudomonas* hospedeira de uma sequência de polinucleo-
5 tideo otimizada codificando uma proteína.