(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 119785885 B
(45)授权公告日 2025.08.05

(21)申请号 202411993498.6
(22)申请日 2024.12.31
(65)同一申请的已公布的文献号
申请公布号 CN 119785885 A
(43)申请公布日 2025.04.08
(73)专利权人 合肥综合性国家科学中心大健康研究院
地址 230601 安徽省合肥市经济技术开发区宿松路4090号1号楼202
(72)发明人 崔丰磊 费才溢 徐实
(74)专利代理机构 上海一平知识产权代理有限公司 31266
专利代理师 褚怡霏 徐迅
(51)Int.Cl.
G16B 30/00(2019.01)
G16B 40/00(2019.01)
G16B 50/40(2019.01)
(56)对比文件
CN 116072231 A,2023.05.05
CN 117238374 A,2023.12.15
审查员 程顺超

权利要求书2页 说明书13页
序列表（电子公布） 附图2页

(54)发明名称

一种基于动态规划算法的mRNA序列优化方法

(57)摘要

本发明公开了一种基于动态规划算法的mRNA序列优化方法,涉及mRNA序列优化领域。本发明分别考虑mRNA进入宿主细胞之后,mRNA翻译成蛋白质序列过程中,相关的正(促进)负(阻碍)两方面的因素,采用动态规划算法获得其最优的综合效应,极大提高mRNA最终的蛋白翻译产量,在mRNA疫苗领域具有良好的应用前景。

CN 119785885 B

1.一种mRNA序列优化方法,所述mRNA用于在目标物种或目标细胞中表达,其特征在于,所述方法包括步骤:

S1)提供所述目标物种或目标细胞的编码序列(CDS)中的密码子发生频率,根据密码子发生频率,获得用于替换的密码子集合;

其中,所述的密码子集合不包括稀有密码子,所述稀有密码子为目标物种或目标细胞CDS中每千个密码子中出现次数<10次的密码子;

S2)提供二碱基对RNA内切酶的敏感指数;

S3)基于步骤S1获得的密码子发生频率和步骤S2获得的二碱基对RNA内切酶的敏感指数,计算所述密码子集合中每个密码子的单个密码子评分;

S4)提供待优化的mRNA编码序列或其编码的氨基酸序列,针对所述序列中的每个氨基酸位点,从步骤S1所述的密码子集合中选出对应的所有同义密码子,并基于步骤S3得到的单个密码子评分和相邻密码子之间的连接赋值,利用状态转移方程计算各氨基酸位点上每个同义密码子对应的最大序列评分;

S5)基于步骤S4获得的每个同义密码子对应的最大序列评分,获得全长mRNA编码序列对应的最大序列评分,从而获得最优密码子组合,即为优化后的mRNA编码序列,

其中,所述最优密码子组合是使得全长mRNA编码序列评分最大化的密码子组合。

2.如权利要求1所述的方法,其特征在于,步骤S2中,所述的RNA内切酶选自:RNAase A、RNase T2、RNAase L、或其组合。

3.如权利要求1所述的方法,其特征在于,步骤S3中,单个密码子评分的计算公式如下:

$S(C_j)=D(C_j)+w\cdot f(C_j)$

式中,

$S(C_j)$为单个密码子评分;

$f(C_j)$为密码子C_j的密码子发生频率;

$D(C_j)$为密码子C_j的内部评分;

w为权重系数;

其中,所述内部评分的计算公式如下:

$D(C_j)=pair\_values(C_j[1]C_j[2])+pair\_values(C_j[2]C_j[3])$

式中,

$C_j[1]$、$C_j[2]$、$C_j[3]$分别表示密码子C_j第1、2、3个碱基;

$pair_values(C_j[1]C_j[2])$表示密码子$C_j$第1、2个碱基间的连接赋值;

$pair_values(C_j[2]C_j[3])$表示密码子$C_j$第2、3个碱基间的连接赋值。

4.如权利要求3所述的方法,其特征在于,如$C_j[3]$为G或C,则将内部评分$D(C_j)$乘以1.5倍。

5.如权利要求1所述的方法,其特征在于,步骤S4中,第i位氨基酸的第j个同义密码子对应的最大序列评分DP[i][j]计算公式如下:

$$DP[i][j]=\max_{m\in\{1,2,\ldots,k_{i-1}\}}\left(DP[i-1][m]+V(C_{(i-1)m},C_{ij})+D(C_{ij})+w\cdot f(C_{ij})\right)$$

式中,

DP[i-1][m]表示当第i-1个氨基酸使用其第m个同义密码子时,前i-1个氨基酸能够获

得的最大序列评分；

k_{i-1}表示第i-1个氨基酸对应的同义密码子个数；

$C_{(i-1)m}$表示第i-1个氨基酸对应的第m个同义密码子；

C_{ij}表示第i个氨基酸对应的第j个同义密码子；

$V(C_{(i-1)m},C_{ij})$表示密码子$C_{(i-1)m}$和C_{ij}之间的相邻密码子连接赋值；

$f(C_{ij})$为密码子C_{ij}的密码子发生频率；

$D(C_{ij})$为密码子C_{ij}的内部评分；

w为权重系数。

6.如权利要求5所述的方法，其特征在于，步骤S5中，全长mRNA编码序列对应的最大评分T_{max}通过以下公式得到：

$$\mathcal{T}_{\max} = \max_{j} \mathrm{DP}[n][j]$$

其中，n为全长序列的氨基酸个数。

7.如权利要求5所述的方法，其特征在于，步骤S5中，所述最优密码子组合通过以下公式得到：

$\{C_1,C_2,...,C_n\}$=Backtrack(j^*)；

其中，j*为在全长n个氨基酸中，使得序列评分DP[n][j]最大化的对应密码子，其通过以下公式得到：

$j^*=\arg\max_j DP[n][j]$。

8.如权利要求1所述的方法，其特征在于，所述的方法进一步包括步骤：

S6)从步骤S5获得的mRNA编码序列中，通过同义密码子替换消除负因子序列，所述负因子序列包括内切酶酶切位点和/或低信息熵重复区序列；

所述的低信息熵重复区序列指重复区域超过两个密码子长度的序列片段，其中重复区域是指由多个二碱基或者一碱基重复连接而成的片段。

9.一种用于mRNA序列优化的设备，所述设备包括输入模块，处理模块及输出模块，其中：

所述输入模块用于输入：目标物种和/或目标细胞、以及待优化的mRNA序列或其编码的氨基酸序列；

所述处理模块根据如权利要求1所述的方法进行mRNA序列优化；

所述输出模块用于输出优化后的mRNA序列。

10.一种计算机可读存储介质，其特征在于，其上存储有用于实现权利要求1所述方法的计算机程序。

一种基于动态规划算法的mRNA 序列优化方法

技术领域

[0001] 本发明涉及算法和分子生物学领域。具体地说,本发明涉及一种基于动态规划算法的mRNA序列优化方法。

背景技术

[0002] 相比传统减毒灭活疫苗,mRNA技术具有研发周期短、生产工艺简单、免疫原性强、安全性较高等优势,能广泛应用于传染病疫苗、肿瘤免疫、蛋白替代等多领域。然而,由于mRNA存在相比DNA更易被降解的缺陷,为了使其翻译产生更多的蛋白质,更合适的mRNA序列设计十分必要。

[0003] 现有的市场的常见优化技术中,针对序列考虑的优化特征包括密码子适应指数(CAI)、GC含量、最小自由能(MFE)+二级结构、稀有密码子、低信息熵重复区序列、核糖体载荷等。这些因素在翻译层面可以分为两类:促进或者阻碍。

[0004] 然而,目前采用的方法仍有待改进。例如,密码子适应指数(CAI)通过比较基因中密码子的使用频率与该物种中高表达基因的密码子使用偏好之间的相似度来计算。由CAI的计算过程可得知,它考虑的仅仅是单个密码子的性能,并未考虑序列中密码子的上下文的连接性。GC含量实际上指征了可以被切割的核心关键碱基-尿嘧啶(U)的减少。但GC含量过高会引起局部地区形成低信息熵重复区序列。低信息熵重复区引起体外转录(IVT)步骤中聚合酶的滑窗效应,而无法对合成的mRNA进行质检。因此,优化算法既要实现尿嘧啶的减少,又要防止局部形成低信息熵重复序列。最近有报道表明,mRNA结构以及对应衍生计算的指数(MFE)对于核糖体的阻滞作用非常微小,主要对核糖体起阻滞作用的是密码子对应的tRNA的丰富度。在细胞内环境中,mRNA的三级结构是不稳定的,变化无常的,可能存在被蛋白激酶R(PKR)等识别的降解敏感结构,但是降解的最终步骤都是内切酶的对RNA的切割水解。此外,优化技术还需要考虑运算的时间。一个密码子可能有多个同义密码子可以替换,以100个密码子为例,每个密码子平均三个同义密码子,如果采用所有的同义密码子穷举的算法,则计算量为3^{100},该计算次数是无法达到的。

[0005] 因此,本领域需要开发一种mRNA优化方法,最大化实现mRNA抗降解、提升翻译和蛋白生产效率,并且提高运算速度。

发明内容

[0006] 本发明的目的就是提供一种mRNA优化方法。

[0007] 在本发明的第一方面,提供了一种mRNA序列优化方法,所述mRNA用于在目标物种或目标细胞中表达,所述方法包括步骤:

[0008] S1)提供所述目标物种或目标细胞的编码序列(CDS)中的密码子发生频率,根据密码子发生频率,获得用于替换的密码子集合;

[0009] 其中,所述的密码子集合不包括稀有密码子,所述稀有密码子为目标物种或目标细胞CDS中每千个密码子中出现次数＜10次的密码子;

[0010] S2)提供二碱基对RNA内切酶的敏感指数;

[0011] S3)基于步骤S1获得的密码子发生频率和步骤S2获得的二碱基对RNA内切酶的敏感指数,计算所述密码子集合中每个密码子的单个密码子评分;

[0012] S4)提供待优化的mRNA编码序列或其编码的氨基酸序列,针对所述序列中的每个氨基酸位点,从步骤S1所述的密码子集合中选出对应的所有同义密码子,并基于步骤S3得到的单个密码子评分和相邻密码子之间的连接赋值,利用状态转移方程计算各氨基酸位点上每个同义密码子对应的最大序列评分;

[0013] S5)基于步骤S4获得的每个同义密码子对应的最大序列评分,获得全长mRNA编码序列对应的最大序列评分,从而获得最优密码子组合,即为优化后的mRNA编码序列,

[0014] 其中,所述最优密码子组合是使得全长mRNA编码序列评分最大化的密码子组合。

[0015] 在另一优选例中,步骤S1中,所述目标物种CDS序列为目标物种中所有能编码蛋白的基因;且目标细胞CDS序列为目标细胞类型中表达量前1000的能够编码蛋白的基因。

[0016] 在另一优选例中,步骤S1中,所述密码子发生频率为某一密码子在目标物种或目标细胞CDS的每千个密码子中出现的次数。

[0017] 在另一优选例中,步骤S1中,所述密码子集合不包含起始密码子AUG和终止密码子UGA。

[0018] 在另一优选例中,步骤S2中,所述的RNA内切酶选自:RNAase A、RNase T2、RNAase L、或其组合。

[0019] 在另一优选例中,步骤S2中,所述二碱基对RNA内切酶的敏感指数表示每两个碱基之间的磷酸二酯键对RNA内切酶的敏感性,敏感性越高,敏感指数越低。

[0020] 在另一优选例中,二碱基对RNA内切酶的敏感指数范围为[-10,10],各二碱基组合的敏感指数为:AA为3、AU为-8、AC为6、AG为6、UA为-10、UU为-10、UC为0、UG为0、CA为5、CU为6、CC为10、CG为10、GA为5、GU为-10、GC为10、GG为10。

[0021] 在另一优选例中,步骤S3中,二碱基对RNA内切酶的敏感指数与密码子发生频次通过权重参数w相加,权重参数w表示两种成分对于最终翻译产量的贡献程度。

[0022] 在另一优选例中,步骤S3中,单个密码子评分的计算公式如下:

[0023] $S(C_j) = D(C_j) + w \cdot f(C_j)$

[0024] 式中,

[0025] $S(C_j)$为单个密码子评分;

[0026] $f(C_j)$为密码子C_j的密码子发生频率;

[0027] $D(C_j)$为密码子C_j的内部评分;

[0028] w为权重系数;

[0029] 其中,所述内部评分的计算公式如下:

[0030] $D(C_j) = pair\_values(C_j[1]C_j[2]) + pair\_values(C_j[2]C_j[3])$

[0031] 式中,

[0032] $C_j[1]$、$C_j[2]$、$C_j[3]$分别表示密码子Cj第1、2、3个碱基;

[0033] $pair_values(C_j[1]C_j[2])$表示密码子Cj第1、2个碱基间的连接赋值;

[0034] $pair_values(C_j[2]C_j[3])$表示密码子Cj第2、3个碱基间的连接赋值。

[0035] 在另一优选例中,所述的连接赋值与二碱基对RNA内切酶的敏感指数成正相关。

[0036] 在另一优选例中,所述的权重系数w范围为0.1-1,较佳地,w选自0.1、0.2、0.3、0.4、.05、0.6、0.7、0.8、0.9或1。

[0037] 在另一优选例中,如$C_j[3]$为G或C,则将内部评分$D(C_j)$乘以1.5倍。

[0038] 在另一优选例中,步骤S4中,设定序列的初始评分为0。

[0039] 在另一优选例中,步骤S4中,相邻密码子的连接赋值$V(C_j,C_k)$计算公式如下:

[0040] $V(C_j,C_k)$=pair_values($C_j[3]C_k[1]$)

[0041] 式中,

[0042] C_j、C_k分别表示mRNA序列的第j和k个密码子,其中k=j+1;

[0043] pair_values($C_j[3]C_k[1]$)表示密码子C_j第3个碱基和密码子C_k第1个碱基间的连接赋值。

[0044] 在另一优选例中,步骤S4中,对于第i位氨基酸的第j个同义密码子,以第i-1位氨基酸可取的同义密码子为变量,计算其能够获得的最大序列评分DP[i][j]。

[0045] 在另一优选例中,步骤S4中,第i位氨基酸的第j个同义密码子对应的最大序列评分DP[i][j]计算公式如下:

[0046] $$\mathrm{DP}[i][j]=\max_{m\in\{1,2,\ldots,k_{i-1}\}}\left(\mathrm{DP}[i-1][m]+V(C_{(i-1)m},C_{ij})+D(C_{ij})+w\cdot f(C_{ij})\right)$$

[0047] 式中,

[0048] DP[i-1][m]表示当第i-1个氨基酸使用其第m个同义密码子时,前i-1个氨基酸能够获得的最大序列评分;

[0049] k_{i-1}表示第i-1个氨基酸对应的同义密码子个数;

[0050] $C_{(i-1)m}$表示第i-1个氨基酸对应的第m个同义密码子;

[0051] C_{ij}表示第i个氨基酸对应的第j个同义密码子;

[0052] $V(C_{(i-1)m},C_{ij})$表示密码子$C_{(i-1)m}$和C_{ij}之间的相邻密码子连接赋值;

[0053] $f(C_{ij})$为密码子C_{ij}的密码子发生频率;

[0054] $D(C_{ij})$为密码子C_{ij}的内部评分;

[0055] w为权重系数。

[0056] 在另一优选例中,步骤S5中,从步骤S4得到的序列评分中选出全长mRNA编码序列对应的最大评分,并回溯该评分对应的各个密码子,从而得到最优密码子组合。

[0057] 在另一优选例中,全长mRNA编码序列对应的最大评分T_{max}通过以下公式得到:

[0058] $$\mathcal{T}_{\max}=\max_j \mathrm{DP}[n][j]$$

[0059] 其中,n为全长序列的氨基酸个数。

[0060] 在另一优选例中,步骤S5中,所述最优密码子组合通过以下公式得到:

[0061] $\{C_1,C_2,\ldots,C_n\}$=Backtrack(j^*);

[0062] 其中,j*为在全长n个氨基酸中,使得序列评分DP[n][j]最大化的对应密码子,其通过以下公式得到:

[0063] $j^*=\arg\max_j$DP[n][j]。

[0064] 在另一优选例中,所述的方法进一步包括步骤:

[0065] S6)从步骤S5获得的mRNA编码序列中,通过同义密码子替换消除负因子序列,所述负因子序列包括内切酶酶切位点和/或低信息熵重复区序列。

[0066] 在另一优选例中,所述的内切酶酶切位点为sapI酶切位点。

[0067] 在另一优选例中,所述的低信息熵重复区序列指重复区域超过两个密码子长度的序列片段,其中重复区域是指由多个二碱基或者一碱基重复连接而成的片段。

[0068] 在另一优选例中,所述消除采用穷举法进行,并且对所有可能的同义密码子替换检验是否引入新的负因子序列。

[0069] 在另一优选例中,从已消除负因子序列的mRNA编码序列中挑选出序列评分最大化的mRNA编码序列,从而得到优化后的mRNA编码序列;所述序列评分通过步骤S4所述方法计算。

[0070] 在另一优选例中,步骤S4进一步包括:

[0071] 对待优化的mRNA编码序列或其编码的氨基酸序列进行预处理。

[0072] 在另一优选例中,对氨基酸序列进行序列预处理包括:

[0073] 1)判定序列中的所有字母是否属于氨基酸的简称;

[0074] 2)判定序列的第一个字母是否为甲硫氨酸-M,如不是则将M放置在首位。

[0075] 在另一优选例中,对mRNA内部的CDS序列进行序列预处理包括选自下组的一个或多个步骤:

[0076] 1)将序列中的胸腺嘧啶(T)替换为尿嘧啶(U);

[0077] 2)判定序列长度是否是3的倍数;

[0078] 3)判定碱基组成单元是否为A、U、C和G;

[0079] 4)将序列的起始密码子更换为AUG,终止密码子更换为UGA。

[0080] 在另一优选例中,所述的预处理进一步包括判定CDS序列的长度,如CDS序列长度≥2000nt,选择同一基因编码的更短蛋白序列的mRNA作为输入序列。

[0081] 在另一优选例中,所述方法进一步包括:将预处理前CDS序列翻译的氨基酸序列与优化后的mRNA编码序列翻译的氨基酸序列进行比较,从而验证优化后序列的正确性。

[0082] 在本发明的第二方面,提供了一种用于mRNA序列优化的设备,所述设备包括输入模块,处理模块及输出模块,其中:

[0083] 所述输入模块用于输入:目标物种和/或目标细胞、以及待优化的mRNA序列或其编码的氨基酸序列;

[0084] 所述处理模块根据如本发明第一方面所述的方法进行mRNA序列优化;

[0085] 所述输出模块用于输出优化后的mRNA序列。

[0086] 在本发明的第三方面,提供了一种计算机设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,所述处理器执行所述程序时实现如本发明第一方面所述的方法。

[0087] 在本发明的第四方面,提供了一种计算机可读存储介质,其上存储有用于实现如本发明第一方面所述方法的计算机程序。

[0088] 在本发明的第五方面,提供了一种mRNA疫苗,所述mRNA的编码序列经过如本发明第一方面所述的方法优化。

[0089] 应理解,在本发明范围内中,本发明的上述各技术特征和在下文(如实施例)中具体描述的各技术特征之间都可以互相组合,从而构成新的或优选的技术方案。限于篇幅,在此不再一一累述。

附图说明

[0090] 下列附图用于说明本发明的具体实施方案,而不用于限定由权利要求书所界定的本发明范围。

[0091] 图1显示了本发明系统示意图。

[0092] 图2显示了本发明优化方法流程图。

[0093] 图3显示了mRNA结构示意图。

具体实施方式

[0094] 本发明人经过广泛而深入的研究,首次开发了一种基于动态规划算法的mRNA序列优化方法。本发明的方法找到了mRNA抗降解与翻译速度之间的最优平衡点、优化后的mRNA无重复低信息熵序列片段存在,且序列内无sapI酶切位点。相对现有的密码子适应指数,本发明综合考虑了单个密码子内部的连接性以及相邻密码子之间的连接性,从而提供了针对全长序列的密码子优化方法。同时,本发明采用动态规划算法找到最优的密码子组合,相对于传统的穷举法节省了大量算力,提升了运算速度。在此基础上,完成了本发明。

[0095] 序列优化影响因素

[0096] 1.密码子频率

[0097] 在生物体中,DNA转录生成mRNA序列,mRNA序列的翻译区(coding sequence,CDS)是由密码子顺序连接组成。每个密码子由三个碱基组成的,每一个密码子对应着唯一的一个氨基酸,一个氨基酸可能有多个密码子对应,对应于同一个氨基酸的多个密码子称为同义密码子。组成mRNA翻译区(CDS序列)的单个密码子的相对适应度(W)或者每千个密码子中出现的频率表征了各个密码子的使用频率,对应细胞内的转运RNA(tRNA)的丰度。各密码子对应的使用频率具有物种特异性。在蛋白质肽链的合成过程中,mRNA上的密码子与tRNA分子在核糖体作用下结合,将密码子对应的氨基酸次第连接合成蛋白质肽链。不同的转运RNA在细胞内的丰度不一样,从而影响氨基酸结合到多肽上的效率。

[0098] 2.密码子第三位碱基

[0099] 密码子第三位为G或者C(GC3)时,与tRNA上的反密码子碱基之间形成的氢键数量为3,相对AT之间的2个氢键,更为稳定,有助于提高密码子与反密码子的特异性结合,从而确保蛋白质合成的准确性。另外,GC3与序列整体GC含量均能预示更长的mRNA的半衰期。

[0100] 3.核糖核酸序列(RNA)内切酶

[0101] RNA的结构组成如图3所示,其中的CDS区位于整个序列的内部中段,而不是两端。则与之相关的降解机制主要是通过核糖核酸(RNA)内切酶对磷酸二酯键的打开过程而发生。据调查,相关的RNA内切酶包括RNase A、RNase T1、RNase T2和RNase L。RNAase T1存在于真菌与细菌中,RNase A存在于脊椎动物中,RNase T2存在于除了古细菌以外的所有生物中,RNase L(核糖核酸酶L)在脊椎动物中高度保守,包括哺乳动物、鸟类、爬行动物、两栖动物和鱼类。因此,对于mRNA疫苗或者相关产品的用途,需要关注RNase A、RNase T2、RNase L。这三种内切酶的切割的序列特征如下:RNase A主要切割嘧啶(C-胞嘧啶,U-尿嘧啶)之后的磷酸二酯键;RNase T2主要切割位于GU,AU之间的磷酸二酯键;RNase L主要切割UU和UA。这三种RNA内切酶对其他的碱基残基的连接也同样可以切割,但是效率不高。

[0102] 动态规划算法

[0103] 本发明基于动态规划算法进行mRNA序列的优化,相比传统的穷举法,节省算力并提升了运算速度。本发明动态规划算法使用状态转移方程实现,具体包括:

[0104] 1)状态定义:每个状态代表当前构建到某一特定位置(即某一特定氨基酸或密码子)的部分序列及其累计评分;

[0105] 2)状态转移:根据同义密码子的选择及其对应的评分,更新累计评分并记录最优路径;

[0106] 其中,初始状态为序列的起始位置,累计评分为零。

[0107] 最优序列筛选:在完成动态规划过程后,依据累计评分对所有可能生成的序列进行排序,选取评分最高的为最优解。

[0108] 序列优化方法

[0109] 本发明提供了一种基于动态规划算法的mRNA序列优化方法,通过动态规划算法找到抗降解与翻译速度之间的最优平衡点。本发明的方法包括如下步骤:

[0110] S1:计算密码子的发生频次。对于已知在某些特定细胞中翻译表达的mRNA疫苗,收集目标细胞的RNAseq数据,计算表达谱,获得其表达水平前1000的能够翻译生成蛋白的基因。如果没有目标细胞相关RNAseq数据或具体表达细胞未知,可以收集目标物种的所有的CDS序列。对每一个密码子,计算在所有的CDS序列中每千个密码子中出现的次数。如果某个密码子的出现的次数低于10,则为稀有密码子,该稀有密码子不作为同义替换候选者。特别地,起始密码子使用AUG,终止密码子使用UGA。

[0111] S2:构建二碱基对核糖核酸(RNA)内切酶的敏感指数。本发明中,序列的抗降解性通过二碱基对核糖核酸(RNA)内切酶的敏感指数进行定义。RNA内切酶具备不同的切割特性,例如RNAaseA在嘧啶(C、U)及RNAase T2在GU、AU等敏感连接处高效切割,RNAase L则倾向高效切割UU、UA序列。根据这些酶在细胞内反应通路中的上下游关系,切割效率,分子相似度等因素,本发明对16种二核苷酸组合进行了内切酶敏感指数定义。在一种实施方式中,自定义赋值范围为[-10,10],AA为3、AU为-8、AC为6、AG为6、UA为-10、UU为-10、UC为0、UG为0、CA为5、CU为6、CC为10、CG为10、GA为5、GU为-10、GC为10、GG为10。

[0112] S3:序列基础预处理。如果需要处理的初始序列是氨基酸序列则验证所有字母是否属于氨基酸的简称(["A","R","N","D","C","Q","E","G","H","I","L","K","M","F","P","S","T","W","Y","V"])。第一个字母如果不是甲硫氨酸-M,则在第一位加上M。如果处理的是信使核糖核酸(mRNA)内部的翻译序列(coding sequence,CDS),将序列中的胸腺嘧啶(T)替换为尿嘧啶(U),验证其长度是否是3的倍数且碱基组成单元是否为["A","U","C","G"]。起始密码子更换为AUG,终止密码子更换为UGA。若待优化的mRNA的翻译序列(CDS)过长(>=2knt)时,即使经过优化设计,序列内的不可避免的负面因子的积累可能导致其在细胞内降解,失去设计价值。因此在存在更短蛋白序列的情况下,需要从uniprot中选取同一基因编码的更短的蛋白序列的mRNA进行设计。

[0113] S4:密码子打分。为了翻译产生更多的蛋白,mRNA序列需要两方面的属性:1)抵抗降解,2)核糖体的高速翻译;S2步骤中已经定义了抵抗降解的属性,S1中计算了核糖体高速翻译的主要属性——密码子的发生频次。整个序列由单个密码子顺序连接而成,因此计算每个密码子内部的S1+S2步骤的两种属性的加和对于后续的动态规划算法中作为状态转移的点十分必要。另外,密码子第三位为G或C时,密码子与反密码子之间的结合能更强,有利

于翻译效率的提升。因此,若密码子第三位为G或C,则对应连接赋值乘以1.5倍。综上所述,每个密码子的内部评分$D(C_j)$计算如下:

[0114] $D(C_j)=pair\_values(C_j[1]C_j[2])+pair\_values(C_j[2]C_j[3])$

[0115] 式中,

[0116] $C_j[1]$、$C_j[2]$、$C_j[3]$分别表示密码子Cj第1、2、3个碱基;

[0117] $pair_values(C_j[1]C_j[2])$表示密码子Cj第1、2个碱基间的连接赋值;

[0118] $pair_values(C_j[2]C_j[3])$表示密码子Cj第2、3个碱基间的连接赋值。

[0119] 连接赋值表示碱基之间的连接强度,其可以通过上述的二碱基对内切酶的敏感指数计算得到。

[0120] 若$C_j[3]\in\{G,C\}$则

[0121] $D(C_j)=1.5\times D(C_j)$

[0122] 密码子最终评分公式:

[0123] $S(C_j)=D(C_j)+w\cdot f(C_j)$

[0124] 符号定义:

[0125] $S(C_j)$为单个密码子评分;

[0126] $f(C_j)$为密码子C_j的密码子发生频率;

[0127] $D(C_j)$为密码子C_j的内部评分;

[0128] w为权重系数。

[0129] 其中,权重系数w用于平衡抗降解能力与密码子使用频率的重要性。如翻译效率对最终产量的影响较大,则权重系数较大,如降解率影响较大,则权重系数较小。权重系数w范围为0.1-1,较佳地,w选自0.1、0.2、0.3、0.4、.05、0.6、0.7、0.8、0.9或1。

[0130] S5:动态规划算法计算最优序列。整条序列由单个密码子顺序连接组成,除了起始和终止密码子,每个密码子连接上游和下游两个密码子,因此有两个连接赋值。在起始密码子(AUG)和终止密码子(UGA)已经固定的前提下,为避免重复计算,针对每个密码子只计算上游的连接。因此,除了每个密码子的内部评分$D(C_j)$外,密码子之间的连接赋值$V(C_i,C_{i+1})$,也是影响整体评分的重要因素。如使用穷举法计算每个同义密码子,然后选择最优路径,计算量将非常大。因此,本发明使用动态规划算法来优化序列。动态规划通过状态转移方程将这两者综合考虑,确保每一步的选择不仅优化当前密码子的评分,还兼顾与前一个密码子的连接赋值,从而实现整体评分的最大化。具体来说,包括以下步骤:

[0131] 1)计算相邻密码子的连接赋值:

[0132] $V(C_j,C_k)=pair\_values(C_j[3]C_k[1])$

[0133] 式中,

[0134] C_j、C_k分别表示mRNA序列的第j和k个密码子,其中k=j+1;

[0135] $pair_values(C_j[3]C_k[1])$表示密码子$C_j$第3个碱基和密码子$C_k$第1个碱基间的连接赋值。

[0136] 2)动态规划的状态转移:

[0137] $$DP[i][j]=\max_{m\in\{1,2,\ldots,k_{i-1}\}}\left(DP[i-1][m]+V(C_{(i-1)m},C_{ij})+D(C_{ij})+w\cdot f(C_{ij})\right)$$

[0138] 其中,DP[i][j]表示前i个氨基酸序列中,当第i个氨基酸使用其第j个同义密码子

时所能获得的最大序列评分；

[0139] DP[i-1][m]表示当第i-1个氨基酸使用其第m个同义密码子时，前i-1个氨基酸对应的最大序列评分；

[0140] K_{i-1}表示第i-1个氨基酸对应的同义密码子个数；

[0141] $C_{(i-1)m}$表示第i-1个氨基酸对应的第m个同义密码子；

[0142] C_{ij}表示第i个氨基酸对应的第j个同义密码子；

[0143] $V(C_{(i-1)m},C_{ij})$表示密码子$C_{(i-1)m}$和C_{ij}之间的相邻密码子连接赋值；

[0144] $f(C_{ij})$为密码子C_{ij}的密码子发生频率；

[0145] $D(C_{ij})$为密码子C_{ij}的内部评分。

[0146] 3)计算最优得分：

[0147] $$\mathcal{T}_{\max} = \max_{j} \mathrm{DP}[n][j]$$

[0148] 其中，n为全长序列的氨基酸个数。

[0149] 4)最优密码子序列回溯

[0150] $\{C_1,C_2,\ldots,C_n\}=\mathrm{Backtrack}(j^*)$

[0151] 其中，$j^*=\arg\max_j \mathrm{DP}[n][j]$

[0152] 通过定义明确的评分体系和动态规划模型，本方法能够在考虑密码子抗内切降解能力、密码子使用频率以及密码子间连接效应的基础上，系统性地选择最优密码子序列。

[0153] S6:消除负面因子片段序列。本发明的算法还包括消除影响体外转录(IVT)生产的小片段序列，包括：

[0154] 1)酶切位点。本发明的算法需要消除可能受到体外转录(IVT)所用细胞中核酸酶影响的酶切位点。在一种实施方式中，所述酶切位点包括sapI酶切位点，其序列特征如下：

[0155] 5’...GCTC TTC(N)1▼...3’

[0156] 3’...CGAGAAG(N)4▲...5’

[0157] 当该特征序列片段存在时，会导致模版DNA片段被切开而最终降解。使用同义密码子替换来消除酶切位点片段时，由于涉及的密码子少，可以通过穷举法列出所有可能的方法，并使得最终选取的同义替换方案的整体序列按照前述打分方式得分最高。

[0158] 2)低信息熵重复区序列。如果序列中出现多个重复碱基，mRNA合成以及之后产物测序验证均需要聚合酶。而聚合酶在重复碱基片段保真度较差。由于核糖体的覆盖范围为2个密码子，因此若重复区域超过两个密码子长度则需要进行去除。“重复区域”是指由多个二碱基或者一碱基重复连接而成的片段。类似地，通过穷举法选择打分最高的同义密码子进行替换。

[0159] 本发明的方法还可以进一步包括：将输入的CDS序列转换为氨基酸序列待与最终的优化后的序列对应的氨基酸序列进行比较，核验最终优化后序列的正确性。

[0160] 本发明的主要优点包括：

[0161] 1)本发明选取了对mRNA翻译为蛋白质的过程影响较明显的促进因素和阻碍因素，最终选择密码子发生频次、密码子第三位碱基、相邻的上下游密码子连接性、以及酶切位点和低信息熵重复区序列消除作为主要的序列优化点。本发明优化后的mRNA序列U碱基含量与高频密码子的使用达到最优平衡状态，无重复低信息熵序列片段存在，且序列内无sapI

酶切位点。

[0162] 2)相对现有的密码子适应指数,本发明综合考虑了单个密码子内部的连接性以及相邻密码子之间的连接性,从而提供了针对全长序列的密码子优化方法。

[0163] 3)本发明采用动态规划算法找到最优的密码子组合,相对于传统的穷举法节省了大量算力,提升了运算速度。

[0164] 下面结合具体实施例,进一步阐述本发明。应理解,这些实施例仅用于说明本发明而不用于限制本发明的范围。下列实施例中未注明具体条件的实验方法,通常按照常规条件,例如Sambrook等人,分子克隆:实验室手册(New York:Cold Spring Harbor Laboratory Press,1989)中所述的条件,或按照制造厂商所建议的条件。除非另外说明,否则百分比和份数是重量百分比和重量份数。

[0165] 实施例1猪瘟蛋白I215L mRNA序列优化

[0166] 将猪瘟蛋白I215L转染到HEK293T中的进行翻译表达。HEK293T为人的肾源细胞,因此宿主为人。

[0167] I215L序列如下:

[0168] MVSRFLIAEYRHLIENPSENFKISVNENNITEWDVILRGPPDTLYEGGLFKAKVAFPPEYPYAPPKLTFTSEMWHPNIYPDGRLCISILHGDNAEEQGMTWSPAQKIDTILLSVISLLNEPNPDSPANVDAAKSYRKYVYKEDLESYPMEVKKTV

[0169] KKSLDECSPEDIEYFKNAASNVPPIPSDAYEDECEEMEDDTYILTYDDDEEEE

[0170] DEEMDDE(SEQ ID NO:1)

[0171] 设计分为以下步骤:

[0172] S1:计算密码子的发生频次。收集人的所有CDS序列,对每一个密码子,计算在所有的CDS序列中每千个密码子中出现的次数,如下表1所示。表格中稀有密码子不作为序列优化时的同义替换候选者。表中不包含起始及终止密码子。起始密码子使用AUG,终止密码子使用UGA。

[0173] 表1密码子发生频次

[0174]

密码子	频率	密码子	频率	密码子	频率	密码子	频率
GCC	26	CGG	11	GAC	24	CAG	36
GCU	19	CGC	9	GAU	24	CAA	14
GCA	17	CGA	6	UGC	11	GAG	40
GCG	6	CGU	5	UGU	10	GAA	34
AGA	13	AAU	18	CAC	15	GGC	20
AGG	12	AAC	18	CAU	12	GGA	17
AUC	18	GGG	15	CUG	36	AGC	20
AUU	16	GGU	11	CUC	18	UCC	17
AUA	8	UUC	17	CUU	14	UCU	17
CUG	36	UUU	17	UUG	13	UCA	14
CUC	18	CCU	20	CUA	7	AGU	14
CUU	14	CCC	19	AAG	32	UCG	4
UUG	13	CCA	19	AAA	27	GUG	26

UUA	9	CCG	6	ACC	18	GUC	13
AAG	32	UGG	11	ACA	17	GUU	12
AAA	27	UAC	13	ACU	14	UAU	12
UGG	11	UAU	12	ACG	6	UAC	13
GUC	13	GUG	26	GUA	8	GUU	12

[0175] S2:构建二碱基对核糖核酸(RNA)内切酶的敏感指数。该指数主要是基于在真核生物中已知的RNA内切酶包括RNAase A和RNase T2家族,以及RNAase L切割特性而自定义构建。RNAase A在嘧啶(C、U)及RNAase T2在GU、AU等敏感连接处高效切割,RNAase L则倾向高效切割UU、UA序列。其余连接类型也同样受到这些内切酶的切割,但是效率不如上述特殊位点。

[0176] 二核苷酸组合共有16种,根据这些酶在细胞内反应通路中的上下游关系,切割效率,分子相似度等因素,此处自定义赋值范围为[-10,10],如AA为3、AU为-8、AC为6、AG为6、UA为-10、UU为-10、UC为0、UG为0、CA为5、CU为6、CC为10、CG为10、GA为5、GU为-10、GC为10、GG为10。

[0177] 表2二碱基对RNA内切酶的敏感指数

[0178]

碱基对	赋值	碱基对	赋值	碱基对	赋值	碱基对	赋值
AA	3	CA	5	GA	5	TA	-10
AC	6	CC	10	GC	10	TC	0
AG	6	CG	10	GG	10	TG	0
AT	-8	CT	6	GT	-10	TT	-10

[0179] S3:序列基础预处理。如果需要处理的初始序列是氨基酸序列则验证所有字母是否属于氨基酸的简称(["A","R","N","D","C","Q","E","G","H","I","L","K","M","F","P","S","T","W","Y","V"])。第一个字母如果不是甲硫氨酸-M,则将M放置在首位。如果处理的是信使核糖核酸(mRNA)内部的翻译序列(coding sequence,CDS),将序列中的胸腺嘧啶(T)替换为尿嘧啶(U),验证其长度是否是3的倍数且碱基组成单元是否为["A","U","C","G"]。起始密码子更换为AUG,终止密码子更换为UGA。为了核验最终优化后序列的正确性,将输入的CDS序列转换为氨基酸序列待与最终的优化后的序列对应的氨基酸序列进行比较。若待优化的mRNA的翻译序列(CDS)过长(>=2knt)时,即使经过优化设计,序列内的不可避免的负面因子的积累可能导致其在细胞内降解,失去设计价值,因此在如果存在更短蛋白序列的情况下,需要从uniprot中选取同一基因编码的更短的蛋白序列的mRNA进行设计。

[0180] I215L长度仅为215个AA,且每个氨基酸字母对应正确,长度没有超过2knt,符合要求。

[0181] S4:密码子打分。为了翻译产生更多的蛋白,mRNA序列需要两方面的属性:1)抵抗降解,2)核糖体的高速翻译;S2步骤中已经定义了抵抗降解的属性,S1中计算了核糖体高速翻译的主要属性-密码子的发生频次。整个序列由单个密码子顺序连接而成,因此计算每个密码子内部的S1+S2步骤的两种属性的加和对于后续的动态规划算法中作为状态转移的点十分必要。另外,密码子第三位为G或C时,密码子与反密码子之间的结合能更强,有利于翻译效率的提升。因此,若密码子第三位为G或C,则对应连接赋值乘以1.5倍。

[0182] 表3密码子内部评分

[0183]

密码子	评分	密码子	评分	密码子	评分	密码子	评分
AAG	16.7	AAA	8.7	AAU	-3.2	AAC	15.3
ACC	25.8	ACA	12.7	ACU	13.4	ACG	24.6
AGA	12.3	AGG	25.2	CGG	31.1	UCC	16.7
UCU	7.7	UCA	6.4	AGC	26	AUG	-9.9
AUC	-10.2	AUU	-16.4	AGU	-2.6	CAA	9.4
CAC	18	CAU	-1.8	CAG	20.1	UUG	-13.7
CCU	18	CCC	31.9	CCA	16.9	GAU	-0.6
CUG	12.6	CUC	10.8	CUU	-2.6	GCU	17.9
GAG	20.5	GAA	11.4	GAC	18.9	GGA	16.7
GCA	16.7	GGU	1.1	GCC	32.6	GUC	-13.7
GGG	31.5	UGU	-9	GGC	32	UAU	-16.8
GUU	-18.8	UAC	-4.7	GUG	-12.4	UUC	-13.3
UGC	16.1	UGG	16.1	UUU	-18.3		

[0184] 其中罕见密码子已经去除。

[0185] S5:动态规划算法计算最优序列。整条序列由单个密码子顺序连接组成,除了起始和终止密码子,每个密码子连接上游和下游两个密码子,也就是有两个连接赋值。在起始密码子(AUG)和终止密码子(UGA)已经固定的前提下,为避免重复计算,针对每个密码子只计算上游的连接。因此,除了每个密码子的内部评分D(Cj)外,密码子之间的连接赋值V(Ci,Ci+1),也是影响整体评分的重要因素。使用状态转移方程实现整体评分的最大化。

[0186] 根据上述算法得出的结果序列如下:

[0187] AUGGUGAGCCGGUUCCUGAUCGCCGAGUACCGGCACCUGAUCGAGAACCCCAGCGAGAACUUCAAGAUCAGCGUGAACGAGAACAACAUCACCGAGUGGGACGUGAUCCUGCGGGGCCCCCCCCGACACCCUCUACGAGGGCGGCCUCUUCAAGGCCAAGGUGGCCUUCCCCCCCCGAGUACCCCUACGCCCCCCCCCAAGCUGACCUUCACCAGCGAGAUGUGGCACCCCAACAUCUACCCCGACGGCCGGCUCUGCAUCAGCAUCCUGCACGGCGACAACGCCGAGGAGCAGGGCAUGACCUGGAGCCCCGCCCAGAAGAUCGACACCAUCCUGCUGAGCGUGAUCAGCCUGCUGAACGAGCCCAACCCCGACAGCCCCGCCAACGUGGACGCCGCCAAGAGCUACCGGAAGUACGUCUACAAGGAGGACCUGGAGAGCUACCCCAUGGAGGUGAAGAAGACCGUGAAGAAGAGCCUGGACGAGUGCAGCCCCGAGGACAUCGAGUACUUCAAGAACGCCGCCAGCAACGUGCCCCCCAUCCCCAGCGACGCCUACGAGGACGAGUGCGAGGAGAUGGAGGACGACACCUACAUCCUGACCUACGACGACGACGAGGAGGAGGAGGACGAGGAGAUGGACGACGAGUGA(SEQ ID NO:2)

[0188] S6:消除负面因子片段序列。主要是影响体外转录(IVT)生产的小片段序列需要消除。1)sapI酶切位点。2)低信息熵重复区序列。SEQ ID NO:2中下划线示出了低信息熵重复区序列片段,SEQ ID NO:3中下划线示出了优化后的对应序列片段。

[0189] AUGGUGAGCCGGUUCCUGAUCGCCGAGUACCGGCACCUGAUCGAGAACCCCAGCGAGAACUUCAAGAUCAGCGUGAACGAGAACAACAUCACCGAGUGGGACGUGAUCCUGCGGGGCCCACCCGACACCCUCUACGAGGGCGGCCUCUUCAAGGCCAAGGUGGCCUUCCCACCCGAGUACCCCUACGCCCCACCCAAGCUGACCUUCACCAGCGAGAUGUGGCACCCCAACAUCUACCCCGACGGCCGGCUCUGCAUCAGCAUCCUGCACGGCGACAACGCCGAGGAGCAGGGCAUGACCUGGAGCCCCGCCCAGAAGAUCGACACCAUCCUGCUGAGCGUGAUCAGCCUGCUGAACGAGCCCAACCCCGAC

AGCCCCGCCAACGUGGACGCCGCCAAGAGCUACCGGAAGUACGUCUACAAGGAGGACCUGGAGAGCUACCCCAUGG
AGGUGAAGAAGACCGUGAAGAAGAGCCUGGACGAGUGCAGCCCCGAGGACAUCGAGUACUUCAAGAACGCCGCCAG
CAACGUGCCCCCAAUCCCCAGCGACGCCUACGAGGACGAGUGCGAGGAGAUGGAGGACGACACCUACAUCCUGACC
UACGACGACGACGAGGAGGAGGAGGACGAGGAGAUGGACGACGAGUGA(SEQ ID NO:3)

[0190] 实施例2猪瘟蛋白pB602L的mRNA序列优化

[0191] 将猪瘟蛋白pB602L转染到HEK293T中的进行翻译表达。HEK293T为人的肾源细胞。宿主同样为人。

[0192] pB602L序列如下：

[0193] MAEFNIDELLKNVLEDPSTEISEETLKQLYQRTNPYKQFKNDSRVAFCSFTNLREQYIRRLIMTSFIG
YVFKALQEWMPSYSKPTHTTKTLLSELITLVDTLKQETNDVPSESVVNTILSIADSCKTQTQKSKEAKTTIDSFLR
EHFVFDPNLHAQSAYTCASTCADTNVDTCASTCASTCASTCASTCASTCASTCASTGASTCADTNVDTCASTCADT
NVDTCASTCADTNVDTCASTCADTNVNTCASMCADTNVDTCASTCANTCASTEYTDLADPERIPLHIMQKTLNVPN
ELQADIDAITQTPQGYRAAAHILQNIELHQSIKHMLENPRAFKPILFNTKITRYLSQHIPPQDTFYKWNYYIEDNY
EELRAATESIYPEKPDLEFAFIIYDVVDSSNQQKVDEFYYKYKDQIFSEVSSIQLGNWTLLGSFKANRERYNYFNQ
NNEIIKRILDRHEEDLKIGKEILRNTIYHKKAKNIQETGPDAPGLSIYNSTFHTDSGIKGLLSFKELKNLEKASGN
IKKAREYDFIDDCEEKIKQLLSKENLTPDEESELIKTKKQLNNALEMLNVPDDTIRVDMWVNNNNKLEKEILYTKA
EL(SEQ ID NO:4)

[0194] 该序列优化步骤中,S1、S2、S4与前述步骤一致。

[0195] S3运行结果:pB602L长度为602个AA,且每个氨基酸字母对应正确,对应的核苷酸的长度为1806nt,没有超过2knt。序列符合要求。

[0196] 通过S5动态优化获得序列如下：

[0197] AUGGCCGAGUUCAACAUCGACGAGCUGCUGAAGAACGUGCUGGAGGACCCCAGCACCGAGAUCAGCGA
GGAGACCCUGAAGCAGCUCUACCAGCGGACCAACCCCUACAAGCAGUUCAAGAACGACAGCCGGGUGGCCUUCUGC
AGCUUCACCAACCUGCGGGAGCAGUACAUCCGGCGGCUGAUCAUGACCAGCUUCAUCGGCUACGUCUUCAAGGCCC
UGCAGGAGUGGAUGCCCAGCUACAGCAAGCCCACCCACACCACCAAGACCCUGCUGAGCGAGCUGAUCACCCUGGU
GGACACCCUGAAGCAGGAGACCAACGACGUGCCCAGCGAGAGCGUGGUGAACACCAUCCUGAGCAUCGCCGACAGC
UGCAAGACCCAGACCCAGAAGAGCAAGGAGGCCAAGACCACCAUCGACAGCUUCCUGCGGGAGCACUUCGUCUUCG
ACCCCAACCUGCACGCCCAGAGCGCCUACACCUGCGCCAGCACCUGCGCCGACACCAACGUGGACACCUGCGCCAG
CACCUGCGCCAGCACCUGCGCCAGCACCUGCGCCAGCACCUGCGCCAGCACCUGCGCCAGCACCUGCGCCAGCACC
GGCGCCAGCACCUGCGCCGACACCAACGUGGACACCUGCGCCAGCACCUGCGCCGACACCAACGUGGACACCUGCG
CCAGCACCUGCGCCGACACCAACGUGGACACCUGCGCCAGCACCUGCGCCGACACCAACGUGAACACCUGCGCCAG
CAUGUGCGCCGACACCAACGUGGACACCUGCGCCAGCACCUGCGCCAACACCUGCGCCAGCACCGAGUACACCGAC
CUGGCCGACCCCGAGCGGAUCCCCCUGCACAUCAUGCAGAAGACCCUGAACGUGCCCAACGAGCUGCAGGCCGACA
UCGACGCCAUCACCCAGACCCCCCAGGGCUACCGGGCCGCCGCCCACAUCCUGCAGAACAUCGAGCUGCACCAGAG
CAUCAAGCACAUGCUGGAGAACCCCCGGGCCUUCAAGCCCAUCCUCUUCAACACCAAGAUCACCCGGUACCUGAGC
CAGCACAUCCCACCCCAGGACACCUUCUACAAGUGGAACUACUACAUCGAGGACAACUACGAGGAGCUGCGGGCCG
CCACCGAGAGCAUCUACCCCGAGAAGCCCGACCUGGAGUUCGCCUUCAUCAUCUACGACGUGGUGGACAGCAGCAA
CCAGCAGAAGGUGGACGAGUUCUACUACAAGUACAAGGACCAGAUCUUCAGCGAGGUGAGCAGCAUCCAGCUGGGC
AACUGGACCCUGCUGGGCAGCUUCAAGGCCAACCGGGAGCGGUACAACUACUUCAACCAGAACAACGAGAUCAUCA

AGCGGAUCCUGGACCGGCACGAGGAGGACCUGAAGAUCGGCAAGGAGAUCCUGCGGAACACCAUCUACCACAAGAA
GGCCAAGAACAUCCAGGAGACCGGCCCCGACGCCCCCGGCCUGAGCAUCUACAACAGCACCUUCCACACCGACAGC
GGCAUCAAGGGCCUGCUGAGCUUCAAGGAGCUGAAGAACCUGGAGAAGGCCAGCGGCAACAUCAAGAAGGCCCGGG
AGUACGACUUCAUCGACGACUGCGAGGAGAAGAUCAAGCAGCUGCUGAGCAAGGAGAACCUGACCCCCGACGAGGA
GAGCGAGCUGAUCAAGACCAAGAAGCAGCUGAACAACGCCCUGGAGAUGCUGAACGUGCCCGACGACACCAUCCGG
GUGGACAUGUGGGUGAACAACAACAACAAGCUGGAGAAGGAGAUCCUCUACACCAAGGCCGAGCUCUGA(SEQ ID
NO:5)

[0198] S6:消除负面因子片段序列。主要是影响体外转录(IVT)生产的小片段序列需要消除。1)sapI酶切位点。2)低信息熵重复区序列。优化后,这两种负面因子片段序列并不存在。

[0199] 在本发明提及的所有文献都在本申请中引用作为参考,就如同每一篇文献被单独引用作为参考那样。此外应理解,在阅读了本发明的上述讲授内容之后,本领域技术人员可以对本发明作各种改动或修改,这些等价形式同样落于本申请所附权利要求书所限定的范围。

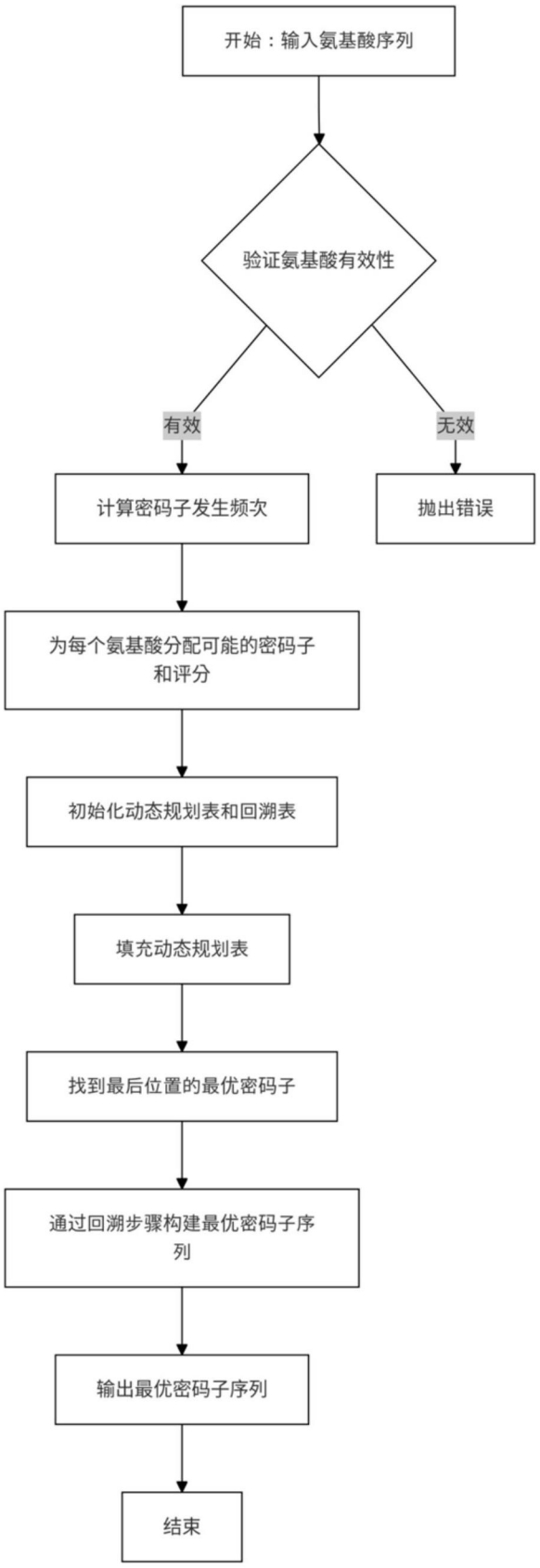
开始：输入氨基酸序列
验证氨基酸有效性
有效
无效
计算密码子发生频次
抛出错误
为每个氨基酸分配可能的密码子和评分
初始化动态规划表和回溯表
填充动态规划表
找到最后位置的最优密码子
通过回溯步骤构建最优密码子序列
输出最优密码子序列
结束

图1

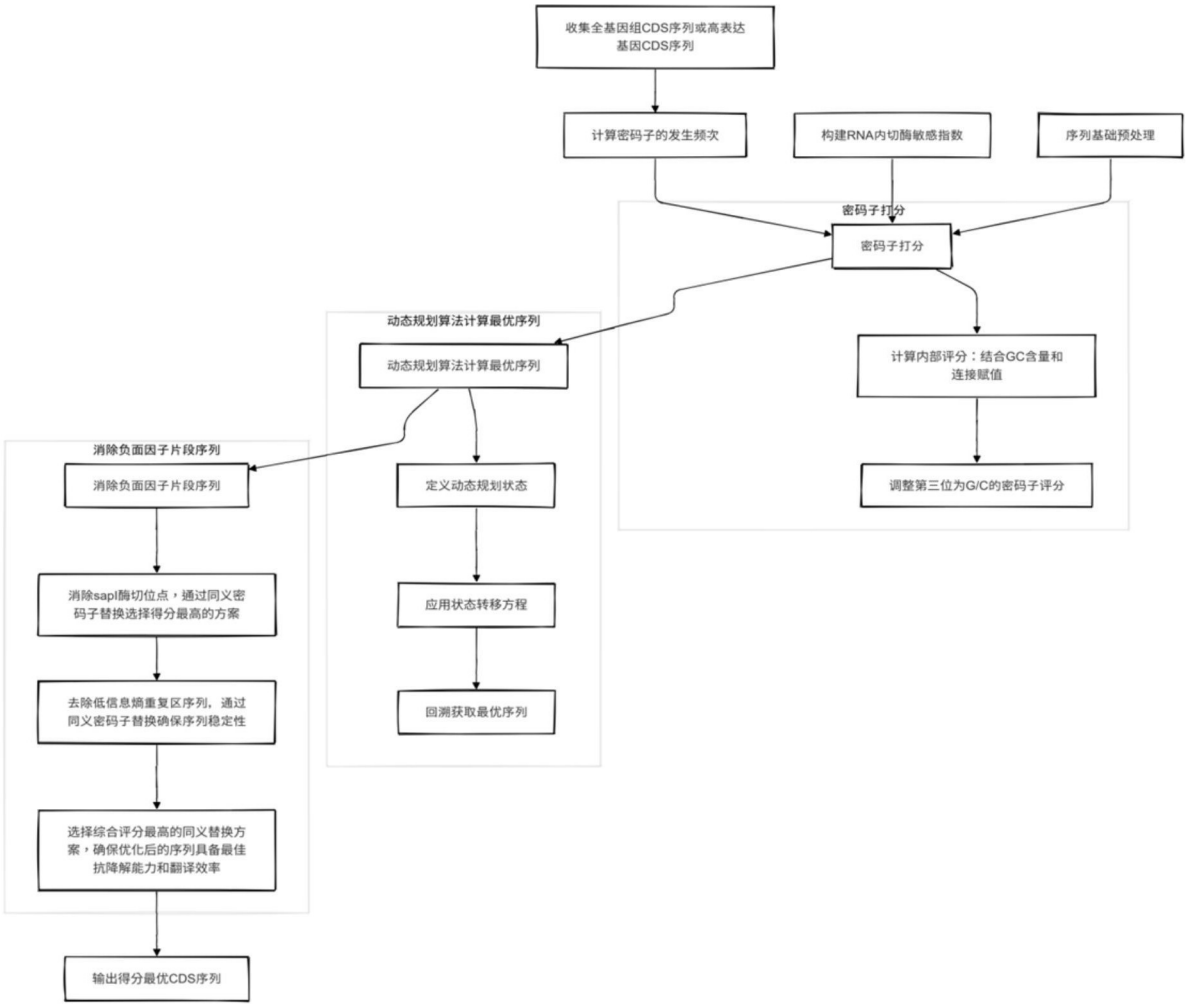
收集全基因组CDS序列或高表达基因CDS序列
计算密码子的发生频次
构建RNA内切酶敏感指数
序列基础预处理
密码子打分
密码子打分
动态规划算法计算最优序列
动态规划算法计算最优序列
计算内部评分：结合GC含量和连接赋值
消除负面因子片段序列
消除负面因子片段序列
定义动态规划状态
调整第三位为G/C的密码子评分
消除sapI酶切位点，通过同义密码子替换选择得分最高的方案
应用状态转移方程
去除低信息熵重复区序列，通过同义密码子替换确保序列稳定性
回溯获取最优序列
选择综合评分最高的同义替换方案，确保优化后的序列具备最佳抗降解能力和翻译效率
输出得分最优CDS序列

图2

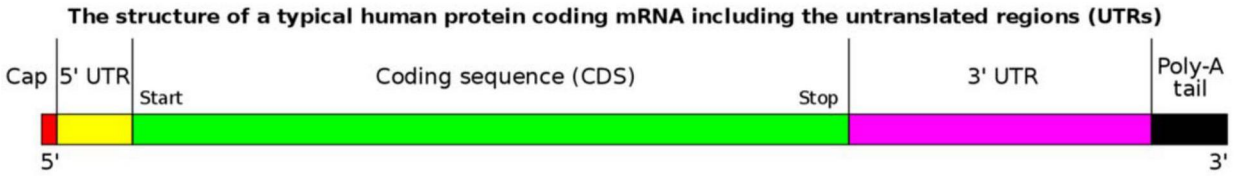
The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5' UTR
Start
Coding sequence (CDS)
Stop
3' UTR
Poly-A tail
5'
3'

图3