(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2019 年 1 月 31 日 (31.01.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/020054 A1

(51) 国际专利分类号:
G06F 19/20 (2011.01)

(21) 国际申请号: PCT/CN2018/097040

(22) 国际申请日: 2018 年 7 月 25 日 (25.07.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710611752.5 2017年7月25日 (25.07.2017) CN

(71) 申请人: 南京金斯瑞生物科技有限公司(NANJINGJINSIRUI SCIENCE & TECHNOLOGY BIOLOGY CORP.) [CN/CN]; 中国江苏省南京市江宁区科学园雍熙路28号, Jiangsu 211100 (CN)。

(72) 发明人: 樊隆(FAN, Long); 中国江苏省南京市江宁区科学园雍熙路28号, Jiangsu 211100 (CN)。 孙岩(SUN, Yan); 中国江苏省南京市江宁区科学园雍熙路28号, Jiangsu 211100 (CN)。 吴东明(WU, Dongming); 中国江苏省南京市江宁区科学园雍熙路28号, Jiangsu 211100 (CN)。 黄小罗(HUANG, Xiaoluo); 中国江苏省南京市江宁区科学园雍熙路28号, Jiangsu 211100 (CN)。 张丽华(ZHANG, Lihua); 中国江苏省南京市江宁区科学园雍熙路28号, Jiangsu 211100 (CN)。 柳振宇(LIU, Zhenyu); 中国江苏省南京市江宁区科学园雍熙路28号, Jiangsu 211100 (CN)。

(74) 代理人: 北京华睿卓成知识产权代理事务所（普通合伙）(CHENG & PENG INTELLECTUAL PROPERTY LAW OFFICE); 中国北京市东城

WO 2019/020054 A1

(54) **Title:** CODON OPTIMIZATION METHOD BASED ON IMMUNE ALGORITHM

(54) **发明名称:** 一种基于免疫算法的密码子优化方法

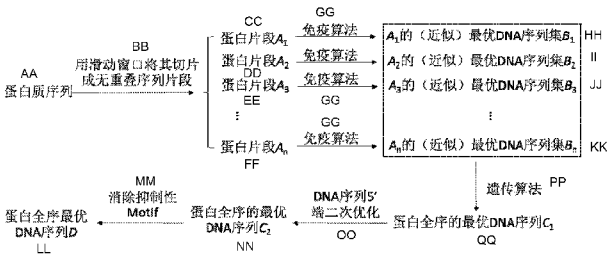

图 1

AA Protein sequence
BB Slice into segments without an overlapping sequence by using a sliding window
CC Protein segment A_1
DD Protein segment A_2
EE Protein segment A_3
FF Protein segment A_n
GG Immune algorithm
HH (Approximately) optimal DNA sequence set B_1 of A_1
II (Approximately) optimal DNA sequence set B_2 of A_2
JJ (Approximately) optimal DNA sequence set B_3 of A_3
KK (Approximately) optimal DNA sequence set B_n of A_n
LL Totally-ordered optimal DNA sequence D of a protein
MM Eliminate inhibition Motif
NN Totally-ordered optimal DNA sequence C2 of the protein
OO Secondary optimization of a 5' end of a DNA sequence
PP Genetic algorithm
QQ Totally-ordered optimal DNA sequence C1 of the protein

(57) **Abstract:** A codon optimization method based on an immune algorithm. The codon optimization method is characterized in that: local multi-target optimization and global multi-target optimization are performed respectively on a protein coding sequence by sequentially using an immune algorithm and a genetic algorithm; the sequence is fine-tuned and optimized by using an exhaustion method, so as to find an optimal expression sequence to the greatest extent. In the present invention, the characteristic of the random global parallel search of the genetic algorithm is kept, and premature convergence is voided to a great extent, and quick convergence

[见续页]

区广渠门内大街90号楼新裕商务大厦B座704, Beijing 100062 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。
— 包括说明书序列表部分(细则5.2(a))。

to an optimal global solution is ensured. In the present invention, by combining the accuracy and efficiency advantages of the immune algorithm and the genetic algorithm for the first time, codon optimization is performed by using a step-based process (local optimization, global optimization and fine-tuned optimization), the high efficiency of the algorithm for performing codon optimization is proved by using an instance test.

(57) 摘要: 一种基于免疫算法的密码子优化方法, 其特征在于先后使用免疫算法和遗传算法分别对蛋白质编码序列进行局部多目标优化和全局多目标优化，再用穷举法对序列进行微调优化，从而最大限度的搜索到最优表达序列。本发明既保留了遗传算法随机全局并行搜索的特点，又在相当大程度上避免未成熟收敛，确保快速收敛于全局最优解。本发明第一次结合利用免疫算法与遗传算法的准确度和效率的优势，通过分步流程(依次分别是局部优化、全局优化、微调优化)进行密码子优化，并通过实例测试证明该算法进行密码子优化的高效性。

一种基于免疫算法的密码子优化方法

技术领域

本发明涉及一种蛋白工程技术，尤其是一种蛋白工程中的密码子优化方法，具体地说是一种基于免疫算法的密码子优化方法。

背景技术

密码子简并性是指在蛋白翻译过程中，一个氨基酸可以由多个不同密码子编码的现象，编码相同氨基酸的不同密码子称为同义密码子，一个长度为200个氨基酸组成的蛋白一般可以由超过10^{20}个不同DNA序列编码。在不同物种中，同义密码子出现的频率并不相同，这种现象称之为密码子的偏好性。密码子优化主要根据宿主表达系统的密码子偏好性等因素，在不改变蛋白氨基酸序列的前提下利用计算机算法从大量DNA编码序列中筛选出能在宿主表达系统中最高效表达蛋白的那条DNA序列。

目前密码子优化过程中常被考虑的影响蛋白表达的主要因素包括宿主细胞密码子偏好性（其常用表征参数有密码子适应度指数 [CAI]、 宿主细胞二联密码子偏好性 [Codon Context]、CBI [Codon Bias Index]、ENC [Effective Number of Codon]、FOP [Frequency of Optimal Codons]、CPP [Codon Preference Parameter]、tAI [tRNA adaptation index])、Hidden Stop Codon数量、GC含量、稀有密码子含量、mRNA抑制性调控模序（motif）数量、mRNA 二级结构（主要包括发夹结构和折叠自由能）、关键密码子和机器学习中数学模型打分、microRNA 结合位点、G4 含量以及蛋白二级结构密码子偏好性(Joshua B. Plotkin & Grzegorz Kudla, *Nature Reviews Genetics,* 2011)。目前可用于密码子优化的软件和算法包括 DNAWorks、Jcat、Synthetic gene designer、GeneDesign 2.0、OPTIMIZER、Eugene、mRNA Optimizer、COOL、D-Tailor、UpGene、GASCO、Codon Harmonization、QPSO、GeMS和ATGME (Evelina Angov, *Biotechnology Journal,* 2011; Nathan Gould et al., Frontiers in Bioengineering and Biotechnology, 2014)。

相较于密码子优化算法中已经被使用的启发式算法（比如粒子群和遗传算法），免疫算法具有其独特优势。免疫算法是基于生物免疫机制提出的一种改进的遗传算法，它将实际求解问题的目标函数对应为抗原，而问题的解对应为抗体。由生物免疫原理可知，生物免疫系统对入侵生命体的抗原通过细胞的分裂和分化作用，自动产生相应的抗体来抵御，这一过程被称为免疫应答。在免疫应答过程中，部分抗体作为记忆细胞保存下来，当同类

抗原再次侵入时，记忆细胞被激活并迅速产生大量抗体，使再次应答比初次应答更快更强烈，体现了免疫系统的记忆功能.抗体与抗原结合后，会通过一系列的反应而破坏抗原.同时，抗体与抗体之间也相互促进和抑制，以维持抗体的多样性及免疫平衡，这种平衡是根据浓度机制进行的，即抗体的浓度越高，则越受抑制；浓度越低，则越受促进，体现了免疫系统的自我调节功能。

发明内容

本发明的目的是针对现有的密码子优化方法存在周期过长，表达准确性较差的问题，发明一种能在有限的时间内有效的完成对密码子优化空间的大规模搜索，即从蛋白编码序列集中筛选出最高效表达的 DNA 序列的基于免疫算法的密码子优化方法。

本发明的技术方案是：

一种基于免疫算法的密码子优化方法,先后使用免疫算法和遗传算法分别对蛋白质编码序列进行局部多目标优化和全局多目标优化，再用穷举法对序列进行微调优化，从而最大限度的搜索到最优表达序列。

具体而言，本发明的方法包括以下三个步骤:；第一步是局部优化，即将蛋白质序列切割成无重叠的序列片段 A_1、A_2 ……A_n，然后利用免疫算法，对每个序列片段完成密码子优化，生成近似最优 DNA 序列集 B_1、B_2……B_n；第二步是全局优化，即利用遗传算法，基于 B_1、B_2……B_n 初始化蛋白质全长的 DNA 编码序列，筛选出蛋白质序列最优 DNA 序列 C_1；第三步是微调优化，包括对编码蛋白质 N 端区域所对应的 DNA 序列 5’端进行穷举法优化，生成 DNA 序列 C_2，并消除表达抑制性模序，最终生成最优表达序列 D。

所述的蛋白质是指由二十个以上的氨基酸组成化合物。在定位上包括分泌蛋白、膜蛋白、胞质蛋白、细胞核内蛋白等；在功能上包含抗体蛋白、调节蛋白、结构蛋白等；在来源上包含同源表达蛋白和异源表达蛋白；在序列上包含天然蛋白和人工改造后的蛋白，完整的蛋白/抗体和截断的部分蛋白/抗体，以及 2 个或 2 个以上蛋白之间、蛋白与肽链之间形成的融合蛋白。本发明中所定义的抗体包括但不限于完整的抗体和 Fab、ScFV、SdAb、嵌合抗体（Chimeric antibody）、双特异性抗体(bispecific antibody)、Fc 融合蛋白等等。

所述的免疫遗传算法采用多目标优化方法对蛋白质片段进行局部优化，群体的初始化基于高表达蛋白编码序列的二联密码子表，直接采用同义密码子对每个基因进行编码；优化过程中通过计算免疫遗传算法的抗体信息熵、抗体群体相似度、抗体浓度和聚合适应度以及更新记忆单元来保证抗体多样性并防止群体退化现象，从而增加算法的全局搜索能力。

所述的遗传算法采用多目标优化方法用来对蛋白质全序进行全局优化，初始化群体基于局部优化的优化后片段随机生成，直接采用每个蛋白质片段的优化序列集对每个基因进行编码。

所述的微调优化利用穷举法对 DNA 序列 5’端的折叠自由能 MFE、Codon Context 和 CAI 进行计算和排序，并根据排序结果选择最佳的蛋白序列 *N*端编码序列。

所述的密码子优化方法至少适用于以下的宿主表达系统：1）哺乳动物表达系统；2）昆虫表达系统；3）酵母表达系统；4）大肠杆菌表达系统；5）枯草芽孢杆菌表达系统；6）植物表达系统和 7）无细胞表达系统。

所述的密码子优化方法至少适用于以下表达载体: 瞬时表达载体和稳定表达载体、病毒表达载体和非病毒表达载体、诱导和非诱导表达载体。

本发明的有益效果是：

免疫算法是一种遗传算法的改进型算法，鉴于免疫算法在优化中防止过早局部收敛的优势，本发明第一次引入免疫算法进行密码子优化进行局部优化，并通过随后的遗传算法进行全局优化，以及最后的微调优化，开发了一种全新的结合了不同算法优势的三步杂合优化算法；更通过下文的实例测试证明该算法进行密码子优化的高效性。

本发明的免疫算法与遗传算法相比，具有如下特点：首先它具有免疫记忆功能，该功能可以加快搜索速度，提高遗传算法的总体搜索能力；其次它具有抗体的多样性保持功能，利用该功能可以提高遗传算法的局部搜索能力；最后它具有自我调节功能，这种功能可用于提高遗传算法的全局搜索能力，避免陷入局部解。所以免疫遗传算法既保留了遗传算法随机全局并行搜索的特点，又在相当大程度上避免未成熟收敛，确保快速收敛于全局最优解。本第一次结合利用免疫算法与遗传算法的准确度和效率的优势，通过分步流程（依次分别是局部优化、全局优化、微调优化）进行密码子优化，并通过实例测试证明该算法进行密码子优化的高效性。

本发明具有速度快，效率高的优点。

附图说明

图 1 是本发明的优化算法流程示意图。

图 2 是本发明的免疫算法流程示意图（即局部优化流程）。

图 3 是本发明的遗传算法流程（即全局优化流程）。

图 4 是本发明的 DNA 序列 5’端优化流程。

图 5 是本发明的测试蛋白基因序列设计示意图。

图 6 是本发明的 pTT 表达载体图谱。

图 7 是本发明的 Western Blotting 结果示意图。

具体实施方式

下面结合附图和实施例对本发明作进一步的说明。

如图 1-7 所示。

一种基于免疫算法的密码子优化方法，它先后使用免疫算法和遗传算法分别对蛋白质编码序列(SEQ ID NO.3 和 SEQ ID NO.4)进行局部多目标优化和全局多目标优化，再用穷举法对序列进行微调优化，从而最大限度的搜索到最优表达序列(SEQ ID NO.5 和 SEQ ID NO.6)，如图 1 所示。其中：

一、免疫算法（即局部优化，流程见图 2）。

该步骤的优化变量个数 ***L*** 为 2，即对每个片段优化 Codon Context 和 CAI 这两个特征（具体描述见下文），属于多目标优化。假设免疫系统由 ***N*** 个抗体组成（即群体规模为 ***N***），每个抗体基因长度为 ***M***（等同于蛋白质序列的氨基酸个数 ***M***），直接采用同义密码子对每个基因进行编码。

（1）根据不同宿主表达系统的基础数据集（即高表达蛋白的编码序列）计算密码子频率表和二联密码子频率表，供生成序列和计算 codon context 和 CAI 使用。

（2）初次应答时，初始抗体根据二联密码子频率产生。具体以蛋白质序列 $a_1a_2\cdots a_m$ 为例，假设 a_1 的同义密码子是 c_{11} 和 c_{12}，a_2 的同义密码子是 c_{21}、c_{22} 和 c_{23}。首个氨基酸 a_1 的密码子根据密码子频率表中 c_{11} 和 c_{12} 的频率选取。二联氨基酸 a_1a_2 对应的二联密码子为 $c_{11}c_{21}$、$c_{11}c_{22}$、$c_{11}c_{23}$、$c_{12}c_{21}$、$c_{12}c_{22}$ 和 $c_{12}c_{23}$，其中二联同义密码子有两组，包括[$c_{11}c_{21}$、$c_{11}c_{22}$、$c_{11}c_{23}$]和[$c_{12}c_{21}$、$c_{12}c_{22}$、$c_{12}c_{23}$]。假设 a_1 选取的密码子为 C_{11}，则氨基酸 a_2 的密码子根据 $c_{11}c_{21}$、$c_{11}c_{22}$ 和 $c_{11}c_{23}$ 的频率从 c_{21}、c_{22} 和 c_{23} 中选择一个。如果 a_1 选取的密码子是 C_{12}，则根据 $c_{12}c_{21}$、$c_{12}c_{22}$ 和 $c_{12}c_{23}$ 的频率选择氨基酸 a_2 的密码子 c_{21}、c_{22} 和 c_{23} 中的一个。简言之，除第一个氨基酸直接根据密码子频率表选取密码子以外，其他氨基酸的密码子的选取都与它的上一个氨基酸的密码子的选取有关，并由它们的二联同义密码子的频率决定。

（3）非初次应答时，群体由父代个体和记忆单元中存储的 ***K*** 个抗体组成，记忆单元抗体记录有优化历史中出现过的 ***K*** 个最佳抗体，其中适应度低的抗体在优化过程中逐步被更高适应度的个体替代。

（4）计算抗体的适应度 ***F***（包括 $F_{[codon\ Context]}$ 和 $F_{[CAI]}$），根据多目标优化选择 N 个子代个体并对新群体完成交叉和变异操作。这里的变异是随机突变密码子。

（5）计算抗体群体相似度 ***S***

本发明利用 Shannon 的平均信息熵 ***H(N)*** 来度量群体相似度 ***S***。

首先 P_{ij} 为同义密码子 ***i*** 出现在氨基酸 ***j*** 上的概率，即：

$$P_{\mathrm{ij}} = \frac{\mathrm{N_{ij}}}{\mathrm{N}},$$

其中 N_{ij} 为群体所有个体的第 ***j*** 个氨基酸位置上其同义密码子 ***i*** 出现的总个数。则 ***Hj(N)*** 为第 ***j*** 个基因（即蛋白序列的第 ***j*** 个氨基酸）的信息熵，定义为：

$$\mathrm{H_j}(N) = -\sum_{i=1}^{N} P_{ij} \log_2 P_{ij}$$

整个群体的平均信息熵为：

$$H(N) = -\frac{1}{M}\sum_{j=1}^{M} H_j(N)$$

群体相似度 ***S*** 的定义为：

$$S = \frac{1}{1+H(N)}$$

（6）随着优化的进行，群体中抗体的相似度不断提高，为了避免抗体的同质性，提高抗体的多样度，从而提高全局搜索能力，防止未成熟收敛，当群体相似度 ***S*** 大于阈值 S_0 时，模仿免疫系统细胞的新陈代谢功能，产生 ***P*** 个新抗体，生成过程同上述（2），使抗体总数达到 ***P+N***。如果群体相似度 ***S*** 小于阈值 S_0 则群体继续直接进入下一代进化，并更新记忆单元。

（7）当 $S>S_0$ 时，对抗体群体 ***P+N*** 计算抗体浓度和聚合适应度。其中抗体浓度是指每个抗体在群体中与其相似抗体所占的百分比，即：

$$C_i = \frac{A_i}{N-1}$$

其中 ***Ai*** 指与抗体 ***i*** 相似度大于相似度常数 λ 的抗体个数。λ 指两个个体比较时在 ***M*** 个密码子中相同的密码子的个数。

聚合适应度 ***F'*** 是依据抗体浓度对抗体适应度 ***F*** 进行修正后的值，即：

$$F_i^{'} = \alpha\frac{F_i}{\sum_i^N F_i} + (1-\alpha)\frac{A_i}{\sum_i^N A_i}(0<\alpha<1)$$

根据聚合适应度选取子代群体，更新记忆单元，并进入下一轮优化，由于我们同时考虑了codon context 和 CAI 两个序列特征，所以 $F'_{[codon\ context]}$基于 $F_{[codon\ context]}$计算，$F'_{[CAI]}$基于 $F_{[CAI]}$计算。如果达到终止代数则停止进化，并输出单个蛋白片段的优化序列集。

二、遗传算法（即全局优化，流程见图 3）。

基于免疫算法优化生成的所有蛋白片段的优化序列集，随机生成初始化群体 N，根据遗传算法的流程，完成适应度计算、子代群体的选取、交叉、变异和记忆体更新，到达终止代数则停止进化，并输出蛋白全序的最优 DNA 编码序列，整个流程属于多目标优化。优化过程中我们直接采用每个蛋白质片段的优化序列集对每个基因进行编码。

三、微调优化。

微调优化包括两步，首先是对 DNA 5’端进行优化，然后消除表达抑制性模序。其中 DNA 5’端的优化过程如图四，使用穷举法列举出蛋白 N 端氨基酸序列（8-15 个氨基酸）所有可能的 DNA 编码序列，并计算它们的 codon context 和 CAI，然后将蛋白序列起始密码子上游的载体序列 50bp （默认值为 50bp，长度可选范围 0～50 bp）与其依序连接，并利用 mfold 软件计算连接后的序列的折叠自由能（minimum free energy，MFE）。根据折叠自由能（值越大越好）、codon context（值越大越好）和 CAI（值越大越好）对信号肽的编码序列进行排序，选择出最佳 5’端序列。

四、上述流程相关细节

（1） 基础数据集及二联密码子表生成

基础数据集是指不同宿主表达系统中高表达蛋白及其所对应的 DNA 编码序列。二联密码子表是指基础数据集的所有二联密码子相对适应度（计算方法见下文）。

（2）codon context 和 CAI 的计算流程

a) 密码子相对适应度 w_{ij}:

$$w_{ij} = \frac{x_{ij}}{x_{i\max}}$$

其中 x_{ij}表示基础数据集中第 i 种氨基酸的第 j 个同义密码子的出现个数，x_{imax}表示基础数据集中第 i 种氨基酸使用频率最高的同义密码子出现的个数。

b) 目标序列的密码子适应指数（Codon Adaptation Index，CAI）:

$$CAI = \left(\prod_{k=1}^{L} w_k \right)^{\frac{1}{L}}$$

其中 ***L*** 指目标序列（即蛋白质序列或片段）的氨基酸个数，w_k为每个氨基酸密码子使用的密码子对应的基础数据集的密码子相对适应度。CAI 的值介于 0 到 1 之间。优化过程中我们尽量提高编码 DNA 的 CAI 的值。

c) 二联密码子相对适应度 p_k：

$$p_k = \frac{\alpha_{CC}^{k}}{\alpha_{AA}^{j(k)}}, \forall k \subseteq \{1,2,\mathrm{K},3721\}$$

其中二联密码子有 3721 种（$61 \times 61 = 3721$，不考虑终止密码子），α_{CC}^{k}表示第 ***k*** 种二联密码子在蛋白序列基础数据集或目标序列（即蛋白质序列或片段）中出现的个数，$\alpha_{AA}^{j(k)}$ 表示表示该二联密码子对应的二联氨基酸出现的个数。

d) 目标序列的二联密码子适应指数（Codon Context，CC）：

$$CC = 1 - \frac{\sum_{k=1}^{3721} \left| p_0^k - p_1^k \right|}{3721}$$

其中 p_0^k 表示目标序列的第 ***k*** 种二联密码子的相对适应度，p_1^k 表示基础数据集的第 ***k*** 种二联密码子的相对适应度。CC 的值介于 0 到 1 之间。优化过程中我们尽量提高编码 DNA 的 CC 的值。

（3）免疫算法和遗传算法的多目标优化过程中子代群体选择可使用 NSGA2 和 SPEA2 算法（默认使用 NSGA2），交叉使用两点交叉。

以下通过一个实例进一步说明本发明的优点：

测试例使用的宿主表达系统是 CHO 细胞系，一共优化测序了两个蛋白质（相关信息见表一）。JNK3 蛋白序列如 SEQ ID NO.1 所示，GFP 蛋白序列如 SEQ ID NO.2 所示；优化前 JNK3 蛋白和 GFP 蛋白编码序列分别如 SEQ ID NO.3 和 SEQ ID NO.4 所示，优化后 JNK3 蛋白和 GFP 蛋白编码序列分别如 SEQ ID NO.5 和 SEQ ID NO.6 所示。

表一：优化测试蛋白序列信息

蛋白质	GenBank 登录号 （野生型）	标签（tag）	标签位置

JNK3	U34820.1	Flag tag	C 末端
GFP	AY174111.1	Flag tag	C 末端

按照图 5 所示，合成编码测试蛋白的基因片段，并通过 EcoR I 和 Hind III 酶切位点将其分别克隆到 pTT5 表达载体（购买自 NRC， 质粒图谱如图 6 所示）。

CHO 3E7 细胞瞬转表达步骤：

1、 将处于对数生长期的 CHO 3E7 悬浮细胞用新鲜的 FreeStyle CHO 培养基稀释到 $5X10^5$ 个细胞/mL， 每个 125mL 三角摇瓶中接种 30mL 细胞悬液。

2、 将细胞在 37℃ 5% CO_2 条件下进行悬浮培养。

3、 当细胞密度达到 $1-1.2X10^6$ 个/mL 时，通过 PEI 转染试剂将克隆有目的基因的质粒载体按照 1ug/ml 的用量分别转染 CHO 3E7 细胞。

4、 转染 48 小时后，将培养基经 1500 转/min 离心，收获细胞。样品可于-80℃冰箱内保存。

Western Blot 实验步骤：

利用抗 Flag tag 抗体，通过 Western Blotting 检测细胞裂解液中目标蛋白的表达量，beta-actin 蛋白作为内参，每个质粒的表达实验重复三次，Western Blotting 结果见图 7。详细步骤如下。

1、 使用细胞裂解液裂解 CHO 细胞，对蛋白浓度进行测定。

2、 向蛋白溶液中加入 5X SDS-PAGE 蛋白上样缓冲液，沸水浴加热 10 分钟。

3、 用微量移液器将蛋白样品加入 SDS-PAGE 胶加样孔内，每孔上样 20 ul。

4、 使用 140V 恒压电泳 60 分钟，溴酚蓝到达胶的底端处附近即可停止电泳。

5、 转膜电压为 100 V，低温转膜时间为 60 分钟。

6、 转膜完毕后把蛋白膜放置到预先准备好的洗涤液中，漂洗 1-2 分钟洗去膜上的转膜液。

7、 摇床上缓慢摇动室温封闭 45 分钟。

8、 加入稀释好的一抗,室温缓慢摇动孵育一小时。

9、 加入洗涤液,在摇床上缓慢摇动洗涤 5 分钟，共洗涤 3 次。

10、加入稀释好的二抗,室温缓慢摇动孵育一小时。

11、加入洗涤液,在摇床上缓慢摇动洗涤 5 分钟，共洗涤 3 次。

12、化学发光检测。

13、使用 Image J 软件对 Western Blotting 结果图片进行定量分析。

表二：优化前后蛋白相对表达量（经 Western Blotting 检测）

GFP（相对表达量±标准差 ）	**JNK3**（相对表达量±标准差）

优化后	22.06 ± 1.78	8.01 ± 0.21
野生型	1.19 ± 0.16	1.09 ± 0.10
比率	18.37 ± 2.90	7.42 ± 0.58

*相对表达量：蛋白表达量除以野生型序列三次重复实验中表达量的最小值

由表二可见，JNK3 和 GFP 蛋白经过本专利的三步杂合密码子优化后，表达量分别较野生型序列提升 7.42±0.58 倍和 18.37±2.90 倍，充分证明新算法的高效性。在公司的实际生产中，我们也比较测试了该算法与其他算法对多个蛋白的优化效果，同样证明该算法更加稳定高效。

本发明未涉及部分均与现有技术相同或可采用现有技术加以实现。

权 利 要 求 书

1、一种基于免疫算法的密码子优化方法，其特征在于先后使用免疫算法和遗传算法分别对蛋白质编码序列进行局部多目标优化和全局多目标优化，再用穷举法对序列进行微调优化，从而最大限度的搜索到最优表达序列。

2、根据权利要求 1 所述的优化方法，其特征是它包括以下三个步骤：；第一步是局部优化，即将蛋白质序列切割成无重叠的序列片段 A_1、A_2 ……A_n，然后利用免疫算法，对每个序列片段完成密码子优化，生成近似最优 DNA 序列集 B_1、B_2……B_n；第二步是全局优化，即利用遗传算法，基于 B_1、B_2……B_n初始化蛋白质全长的 DNA 编码序列，筛选出蛋白质序列最优 DNA 序列 C_1；第三步是微调优化，包括对编码蛋白质 N 端区域所对应的 DNA 序列 5’端进行穷举法优化，生成 DNA 序列 C_2，并消除表达抑制性模序，最终生成最优表达序列 D。

3、根据权利要求 1 或 2 所述的优化方法，其特征是所述的蛋白质是指由二十个以上的氨基酸组成化合物。在定位上包括分泌蛋白、膜蛋白、胞质蛋白、细胞核内蛋白等；在功能上包含抗体蛋白、调节蛋白、结构蛋白等；在来源上包含同源表达蛋白和异源表达蛋白；在序列上包含天然蛋白和人工改造后的蛋白，完整的蛋白/抗体和截断的部分蛋白/抗体，以及 2 个或 2 个以上蛋白之间、蛋白与肽链之间形成的融合蛋白。本发明中所定义的抗体包括但不限于完整的抗体和 Fab、ScFV、SdAb、嵌合抗体、双特异性抗体、Fc 融合蛋白等等。

4、根据权利要求 1 或 2 所述的优化方法，其特征是所述的免疫遗传算法采用多目标优化方法对蛋白质片段进行局部优化，群体的初始化基于高表达蛋白编码序列的二联密码子表，直接采用同义密码子对每个基因进行编码；优化过程中通过计算免疫遗传算法的抗体信息熵、抗体群体相似度、抗体浓度和聚合适应度以及更新记忆单元来保证抗体多样性并防止群体退化现象，从而增加算法的全局搜索能力。

5、根据权利要求 1 或 2 所述的优化方法，其特征是所述的遗传算法采用多目标优化方法用来对蛋白质全序进行全局优化，初始化群体基于局部优化的优化后片段随机生成，直接采用每个蛋白质片段的优化序列集对每个基因进行编码。

6、根据权利要求 1 或 2 所述的优化方法，其特征是所述的微调优化利用穷举法对 DNA 序列 5’端的折叠自由能 MFE、Codon Context 和 CAI 进行计算和排序，并根据排序结果选择最佳的蛋白序列 N端编码序列。

7、根据权利要求 1 或 2 所述的优化方法，其特征是所述的密码子优化方法至少适用于以下的宿主表达系统：1）哺乳动物表达系统；2）昆虫表达系统；3）酵母表达系统；4）大肠杆菌表达系统；5）枯草芽孢杆菌表达系统；6）植物表达系统和 7）无细胞表达系统。

8、根据权利要求 1 或 2 所述的优化方法，其特征是所述的密码子优化方法至少适用于以下表达载体: 瞬时表达载体和稳定表达载体、病毒表达载体和非病毒表达载体、诱导和非诱导表达载体。

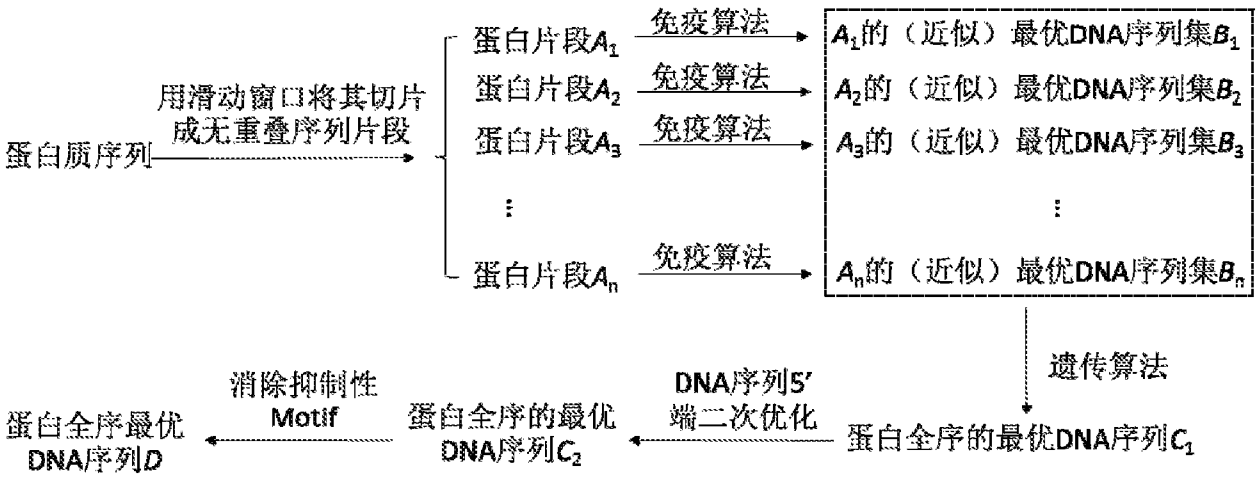
蛋白质序列
用滑动窗口将其切片成无重叠序列片段
蛋白片段A_1
蛋白片段A_2
蛋白片段A_3
蛋白片段A_n
免疫算法
免疫算法
免疫算法
免疫算法
A_1的（近似）最优DNA序列集B_1
A_2的（近似）最优DNA序列集B_2
A_3的（近似）最优DNA序列集B_3
A_n的（近似）最优DNA序列集B_n
遗传算法
蛋白全序的最优DNA序列C_1
DNA序列5′端二次优化
蛋白全序的最优DNA序列C_2
消除抑制性Motif
蛋白全序最优DNA序列D

图 1

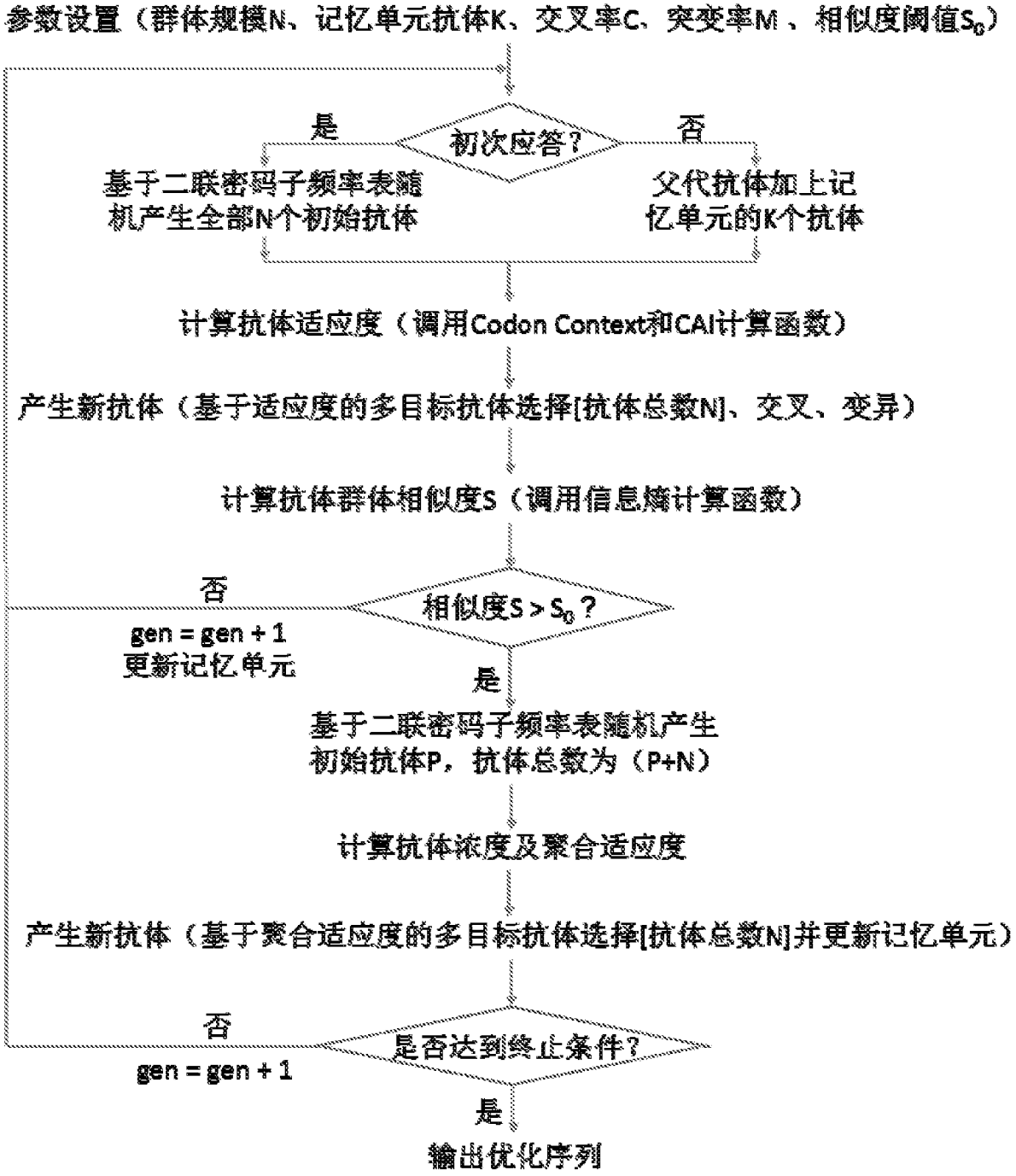
参数设置（群体规模N、记忆单元抗体K、交叉率C、突变率M 、相似度阈值S_0）
初次应答？
是
否
基于二联密码子频率表随机产生全部N个初始抗体
父代抗体加上记忆单元的K个抗体
计算抗体适应度（调用Codon Context和CAI计算函数）
产生新抗体（基于适应度的多目标抗体选择[抗体总数N]、交叉、变异）
计算抗体群体相似度S（调用信息熵计算函数）
相似度S > S_0？
否
gen = gen + 1
更新记忆单元
是
基于二联密码子频率表随机产生初始抗体P，抗体总数为（P+N）
计算抗体浓度及聚合适应度
产生新抗体（基于聚合适应度的多目标抗体选择[抗体总数N]并更新记忆单元）
是否达到终止条件？
否
gen = gen + 1
是
输出优化序列

图 2

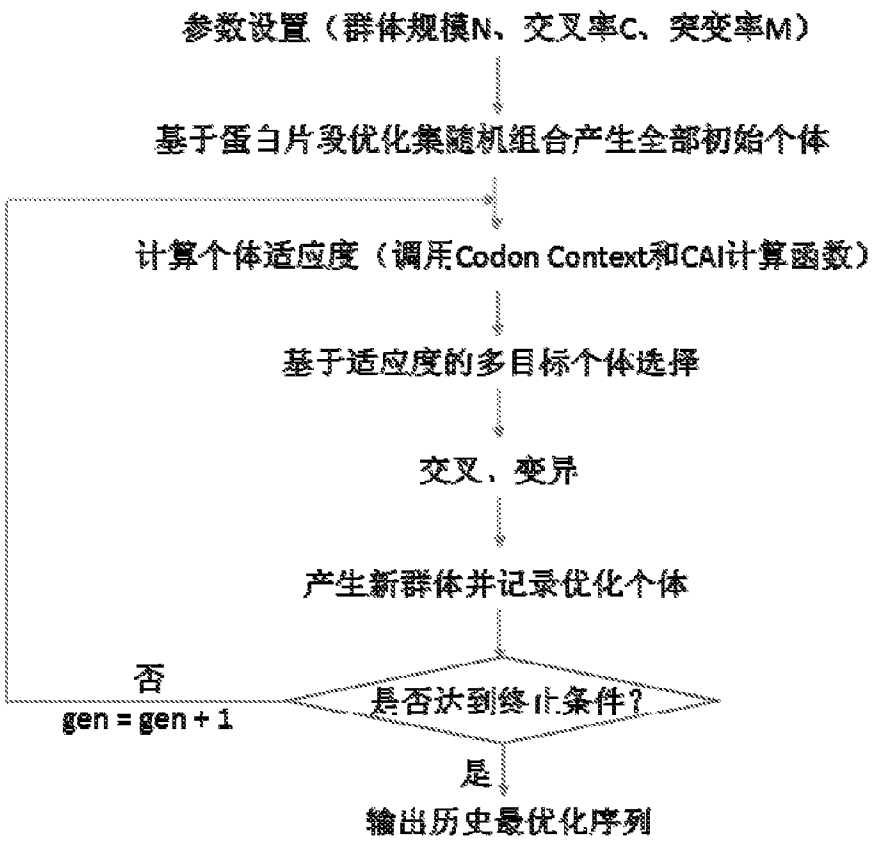

图 3

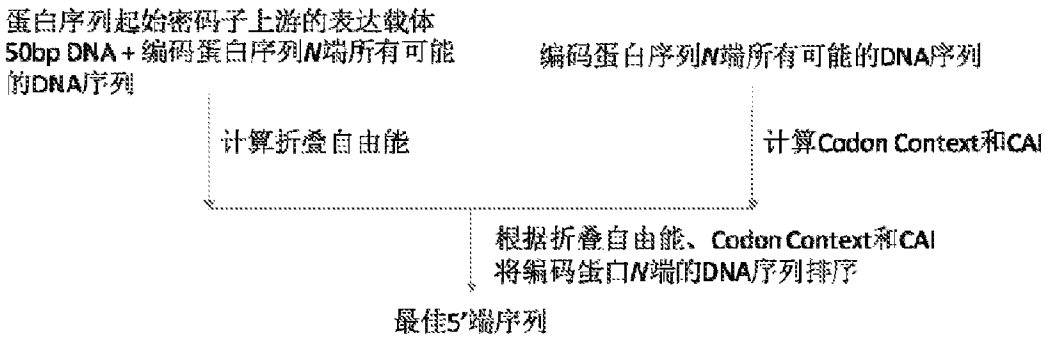

图 4

JNK3

	1～1242 bp	1243~1266 bp
优化前:	U34820.1	Flag Tag
优化后:	优化后的 DNA 序列	Flag Tag

GFP

	1~3048 bp	3049~3072 bp
优化前:	AY174111.1	Flag Tag
优化后:	优化后的 DNA 序列	Flag Tag

图 5

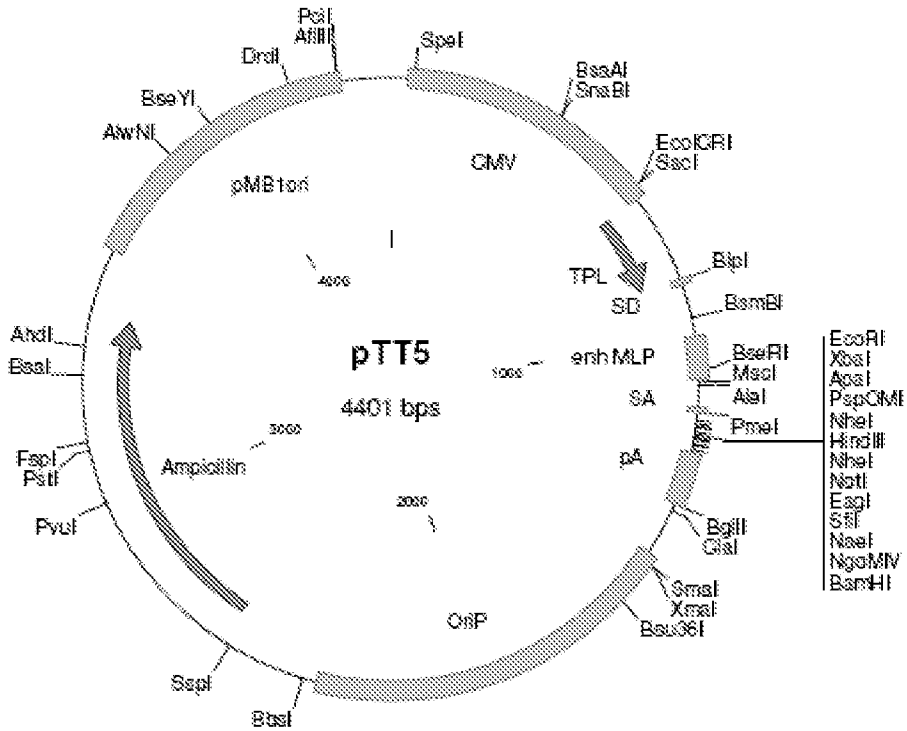
pTT5
4401 bps
CMV
enh MLP
TPL
SD
SA
pA
pMB1 ori
Ampicillin
OriP
1000
2000
3000
4000
PstI
AflIII
SpeI
DraI
BsaAI
SnaBI
BseYI
AlwNI
EcoICRI
SacI
BlpI
BsmBI
AhdI
BsaI
FspI
PstI
PvuI
SspI
BssSI
BseRI
MeeI
AleI
PmeI
EcoRI
XbaI
AccI
PspOMI
NheI
HindIII
NheI
NotI
EagI
SfiI
NaeI
NgoMIV
BamHI
BglII
ClaI
SmaI
XmaI
Bsu36I

图 6

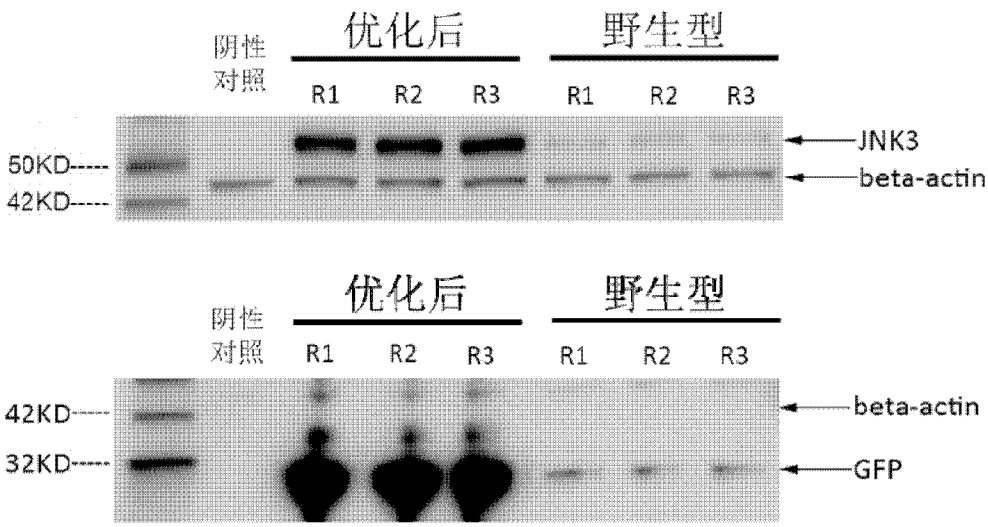
阴性
对照
优化后
野生型
R1
R2
R3
R1
R2
R3
50KD-----
42KD-----
JNK3
beta-actin
阴性
对照
优化后
野生型
R1
R2
R3
R1
R2
R3
42KD-----
32KD-----
beta-actin
GFP

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/097040

A. CLASSIFICATION OF SUBJECT MATTER

G06F 19/20(2011.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F19/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNTXT, CNKI, VEN, SIPOABS, USTXT, WOTXT, EPTXT, SPRINGER, GOOGLE SCHOLAR: 密码子, 基码, 免疫, 遗传, 算法, 优化, 局部, 全局, 穷举, codon, immune, immunity, genetic, algorithm, optimize, local, global, exhaustion

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	郑建刚 等 (ZHENG, Jiangang et al.). "基于信息熵的DNA免疫遗传算法 (DNA-Immune-Genetic Algorithm Based on Information Entropy)" *计算机仿真 (Computer Simulation),* Vol. 23, No. (06), 30 June 2006 (2006-06-30), pages 163-165	1, 3-8
A	CN 106951726 A (SUZHOU GENEWIZ BIOTECHNOLOGY CO., LTD.) 14 July 2017 (2017-07-14) entire document	1-8
A	CN 101490262 A (DSM IP ASSETS BV) 22 July 2009 (2009-07-22) entire document	1-8
A	傅平 等 (FU, Ping et al.). "基于信息熵的免疫遗传算法聚类分析 (Clustering Analysis of Immune-genetic Algorithm Based on Information Entropy)" *计算机工程 (COMPUTER ENGINEERING),* Vol. 34, No. (6), 31 March 2008 (2008-03-31), pages 227-228 and 232	1-8
A	SANDHU, K.S. et al. "GASCO: Genetic Algorithm Simulation for Codon Optimization" *In Silico Biology,* Vol. 8, No. (2), 13 April 2008 (2008-04-13), pages 187-191	1-8

☐ Further documents are listed in the continuation of Box C.

☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 October 2018	**31 October 2018**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/097040

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
   a. ☑ forming part of the international application as filed:
      ☑ in the form of an Annex C/ST.25 text file.
      ☐ on paper or in the form of an image file.
   b. ☐ furnished together with the international application under PCT Rule 13*ter*.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
   c. ☐ furnished subsequent to the international filing date for the purposes of international search only:
      ☐ in the form of an Annex C/ST.25 text file (Rule 13*ter*.1(a)).
      ☐ on paper or in the form of an image file (Rule 13*ter*.1(b) and Administrative Instructions, Section 713).

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/097040

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106951726	A	14 July 2017	None			
CN	101490262	A	22 July 2009	US	2014377800	A1	25 December 2014
				JP	2009540845	A	26 November 2009
				US	2009286280	A1	19 November 2009
				EA	200900096	A1	30 June 2009
				BR	PI0713795	A2	06 November 2012
				EA	015925	B1	30 December 2011
				EP	2035561	A1	18 March 2009
				EP	2423315	B1	07 January 2015
				CN	101490262	B	26 September 2012
				WO	2008000632	A1	03 January 2008
				US	8812247	B2	19 August 2014
				AU	2007263880	A1	03 January 2008
				ES	2534282	T3	21 April 2015
				JP	5250850	B2	31 July 2013
				BR	PI0713795	B1	20 March 2018
				DK	2423315	T3	13 April 2015
				EP	2423315	A1	29 February 2012
				CA	2657975	A1	03 January 2008

国际检索报告

国际申请号
PCT/CN2018/097040

A. 主题的分类

G06F 19/20(2011.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F19/-

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNABS,CNTXT,CNKI,VEN,SIPOABS,USTXT,WOTXT,EPTXT,SPRINGER,GOOGLE SCHOLAR:密码子，基码，免疫，遗传，算法，优化，局部，全局，穷举，codon，immune，immunity，genetic，algorithm，optimize，local，global，exhaustion

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	郑建刚 等. "基于信息熵的DNA免疫遗传算法" 计算机仿真，第23卷，第06期，2006年 6月 30日 (2006-06-30)， 第163-165页	1，3-8
A	CN 106951726 A (苏州金唯智生物科技有限公司) 2017年 7月 14日 (2017-07-14) 全文	1-8
A	CN 101490262 A (帝斯曼知识产权资产管理有限公司) 2009年 7月 22日 (2009-07-22) 全文	1-8
A	傅平 等. "基于信息熵的免疫遗传算法聚类分析" 计算机工程，第34卷，第6期，2008年 3月 31日 (2008-03-31)， 第227-228、232页	1-8
A	Kuljeet Singh Sandhu 等. "GASCO: Genetic Algorithm Simulation for Codon Optimization" In Silico Biology，第8卷，第2期，2008年 4月 13日 (2008-04-13)， 第187-191页	1-8

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2018年 10月 24日	2018年 10月 31日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 陈学元 电话号码 62411980

表 PCT/ISA/210 (第2页) (2015年1月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/097040

检索报告引用的专利文件			公布日（年/月/日）	同族专利			公布日（年/月/日）
CN	106951726	A	2017年 7月 14日	无			
CN	101490262	A	2009年 7月 22日	US	2014377800	A1	2014年 12月 25日
				JP	2009540845	A	2009年 11月 26日
				US	2009286280	A1	2009年 11月 19日
				EA	200900096	A1	2009年 6月 30日
				BR	PI0713795	A2	2012年 11月 6日
				EA	015925	B1	2011年 12月 30日
				EP	2035561	A1	2009年 3月 18日
				EP	2423315	B1	2015年 1月 7日
				CN	101490262	B	2012年 9月 26日
				WO	2008000632	A1	2008年 1月 3日
				US	8812247	B2	2014年 8月 19日
				AU	2007263880	A1	2008年 1月 3日
				ES	2534282	T3	2015年 4月 21日
				JP	5250850	B2	2013年 7月 31日
				BR	PI0713795	B1	2018年 3月 20日
				DK	2423315	T3	2015年 4月 13日
				EP	2423315	A1	2012年 2月 29日
				CA	2657975	A1	2008年 1月 3日

国际检索报告

国际申请号
PCT/CN2018/097040

第I栏　核苷酸和/或氨基酸序列(续第1页第1.c项)

1. 关于国际申请中所公开的任何核苷酸和/或氨基酸序列,国际检索是基于下列序列表进行的：

a. ☑ 作为国际申请的一部分提交的：
   - ☑ 附件C/ST.25文本文件形式
   - ☐ 纸件或图形文件形式

b. ☐ 根据细则13之三.1(a)仅为国际检索目的以附件C/ST.25文本文件形式与国际申请同时提交的：

c. ☐ 仅为国际检索目的在国际申请日之后提交的：
   - ☐ 附件C/ST.25文本文件形式（细则13之三.1(a)）
   - ☐ 纸件或图形文件形式（细则13之三.1(b)和行政规程第713段）

2. ☐ 另外，在提交/提供了多个版本或副本的序列表的情况下，提供了关于随后提交的或附加的副本中的信息与申请时提交的作为申请一部分的序列表的信息相同或未超出申请时提交的申请中的信息范围（如适用）的所需声明。

3. 补充意见：