第 10 卷 第 1 期 生 物 信 息 学 Vol. 10 No. 1
2012 年 03 月 Chinese Journal of Bioinformatics Mar. ,2012

doi:10.3969/j.issn.1672 -5565.2012.01.01

# 起始密码子下游区域 AT 含量优化工具 BestAT 的开发与应用

余劲聪[1],唐龙盘[1],方柏山[1,2*]

(1. 福建省高校工业生物技术重点实验室(华侨大学),福建 厦门,361021;
2. 厦门大学化学化工学院化学工程与生物工程系,福建 厦门,361005)

**摘 要**:基因的表达水平受到起始密码子下游区域 AT 含量的影响,从巨大的序列集中筛选出具有特定 AT 含量和密码子用法特征的同义序列是一个繁琐的工作。本文研发 AT 含量优化工具"BestAT",初步解决了自动获取海量同义序列和充分展示同义序列的密码子用法特性两个关键问题,并且实现了与密码子用法数据库(CUD)的无缝结合,采用了密码子参数的原位标示和 AT 含量曲线等直观方式展示序列特性,为这类实验设计提供有力的支持。

**关键词**: 起始密码子下游区域; AT 含量; 密码子优化; BestAT 软件; 甘油脱氢酶
**中图分类号**:TP31;Q81 **文献标识码**:A **文章编号**:1672 -5565(2012) -01 -001 -04

# Development and application of BestAT tool for optimizing AT content of the initiation codon downstream region

YU Jin-cong[1], TANG Long-pan[1], FANG Bai-shan[1,2*]

(1. *The Key Laboratory for Industrial Biotechnology of Fujian Higher Education (Huaqiao University), Xiamen* 361021, *China*;
2. *Department of Chemical and Biochemical Engineering, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen* 361005, *China*)

**Abstract**: AT content of the initiation codon downstream region affects the expression level of a gene. Picking up synonymous sequences that have specific AT content and codon usage features from tremendous sequence set is a burdensome work. In this paper, the tool named BestAT for AT content optimization had been developed. The tool resolved two key issues, (i) auto-generating mass synonymous sequences and (ii) revealing their codon usage features in details. Moreover, the tool was closely incorporated to the Codon Usage Database (CUD), and supported marking in situ codon parameters and showing visualized AT curve. All of those help to experimental design for AT content optimization powerfully.

**Key words**: initiation codon downstream region; AT content; codon optimization; BestAT software; glycerol dehydrogenase

基因的表达水平受到密码子用法、mRNA 二级结构、tRNA 丰度以及起始密码子下游区域的 AT 含量等多种因素的影响[1],Nishikubo 等[2] 对源自 Thermus thermophilus HB8 的 ndx3 基因的 25 种不同 AT 含量的同义序列(synonymous sequences),分别在 Escherichia coli BL21(DE3)中表达,经比较发现,具有较高 AT 含量的同义序列的表达水平普遍高于较低 AT 含量的表达水平,最大相差达到 10 倍左右,从而提出,通过在密码子下游多采用高 AT 含量的密码子,可能是提高基因表达水平的一般性策略。该策略以调整 AT 含量为目的,但总体而言,仍属于密码子优化(codon optimization)的范畴,因为仅依

收稿日期:2010 -12 -09;修回日期:2011 -04 -07.
基金项目:国家自然科学基金资助项目(21076172,30770059),高等学校博士学科点专项科研基金(20070385001),福建省高校产学合作科技重大项目(2010H6023)。
作者简介:余劲聪,博士研究生,主要从事生物信息学与合成生物学研究,E -mail: yjc@hqu.edu.cn.
* 通讯作者:方柏山,教授,博士生导师,Tel: 0592 -2185869,E -mail: fbs@xmu.edu.cn.

靠同义密码子替换的方式来达到调整 AT 含量的目的。这种优化策略在其它研究[3,4]中也得到采用并获得了预期的效果。由于密码子高度的简并性，一小段序列可能就存在数量巨大的同义序列，比如“GAA－TTC－CCG－GGG－ATC－CGT－CGA－CCT－GCA－GCC”，虽然仅包含 10 个密码子，但是却存在多达 442368 条的同义序列。从这些海量数据中探讨不同的 AT 含量及具体的密码子用法特性，将是一个极度繁琐的工作。在上述相关研究中，并未给其他研究者提供一个通用的自动化工具。鉴于此，本文在充分分析此类实验辅助设计的核心需求的基础上，从头开发了起始密码子下游区域 AT 含量优化工具“BestAT”，现报道如下。

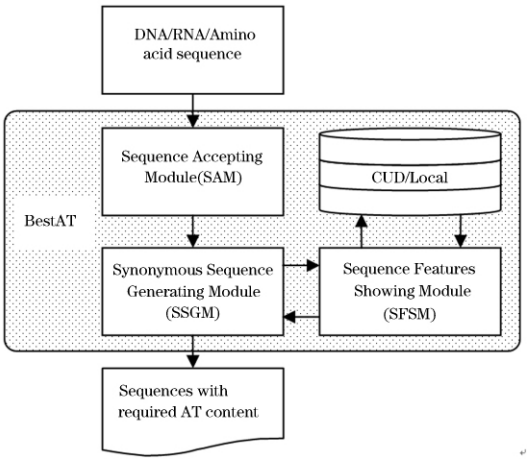



图 1 BestAT 软件的模块组成

Fig. 1 Modular components of the BestAT software

## 1 工具的开发

### 1.1 总体设计

在目的基因起始密码子下游区域序列 AT 含量优化的实验设计中，研究者最关心的两个问题是：(1) 在目的基因的这一区域内，存在多少同义序列，其各自的 AT 含量如何？(2) 相对目标宿主，这些同义序列中的各密码子用法特征如何？因此，为解决这些问题的计算机辅助设计工具，应该具备两个基本功能：①对于给定的序列，从用户处获知该序列的遗传密码类型后，能够自动给出所有的同义序列，并计算对应的 AT 含量，且为了便于浏览序列，可以按照 AT 含量的高低对这些序列依次排列，最终给出一个同义序列列表(同义序列集)；②对于用户指定的目标宿主，能够从一个密码子用法数据源中，自动提取或计算密码子用法参数的值，并标示在指定的同义序列上，以便帮助用户了解该序列的密码子用法特征并作出选择。以满足上述两个核心需求为目的，设置该软件的模块组成为，序列接收模块(sequence accepting module，SAM)、同义序列生成模块(synonymous sequences generating module，SSGM)和序列特征显示模块(sequence features showing module，SFSM)，它们的关系如图 1 所示。一般的流程是，用户输入一段序列文本，经序列接收模块(简称接收模块)转换为软件内部可接受的序列形式。接着，同义序列生成模块(简称生成模块)根据输入序列所属的遗传密码类型，自动生成所有可能的同义序列。同时，当用户选定一个同义序列时，序列特征显示模块(简称特征模块)将即时显示该序列的 AT 含量、密码子参数值等特征，用户可根据这些特征值挑选出需要的序列并输出。

### 1.2 序列接收模块

接收模块用于处理用户输入的序列，当前该模块可接受 DNA、RNA 与蛋白质三种类型的序列，完整覆盖了所有的序列类型，可灵活应对用户的不同需求。为了在同一窗口中接收序列输入，该模块采用了一种“基于特征集的序列类型的自动识别机制”，具体为：如果输入的序列仅包含特征集{A，C，G，T}中的字符，则认为该序列的类型为 DNA，对于 RNA 和蛋白质序列，其特征集分别为{A，C，G，U}、{A，C，D，E，F，G，H，I，K，L，M，N，P，Q，R，S，T，V，W，Y，B，U，X，Z}。对于蛋白质序列的特征集，其中，前 20 个字符为常见的氨基酸单字母表示形式，而后 4 个字符为多义字符，分别代表“Asx，Sec，Xaa，Glx”。假如序列中包含了超过这三个特征集以外的字符，则提示错误。值得注意的是，DNA 与 RNA 特征集显然是蛋白质特征集的子集，考虑到对于一条现实的蛋白质序列，仅出现 A、C、G、T/U 这类残基的可能性较小，因此根据程序的默认设置，这样序列会被自动判定为 DNA 或 RNA 序列，但是对于特殊情况下，用户需要输入这样的蛋白质序列，则可以选择自定义模式，即不采用自动识别模式，而直接定义该序列的类型为蛋白质。综合这两种模式，可以完全解决特征集交叉可能导致的误判问题。在确认序列类型后，如果该序列为：(1) DNA，则直接转入下一模块的处理；(2) RNA，则将序列中的“U”逐一替换为“T”；(3) 蛋白质，则需要用户确认该序列所属的遗传密码类型(事先默认为标准型)，再将该蛋白质序列逆向翻译为简并形式的 DNA 序列。总之，接收模块的功能可以归结为“统一端口、自动识别、分类处理、归一序列”，经这

一系列的处理后，用户输入的序列转换为软件内部可识别的形式。

### 1.3　同义序列生成模块

生成模块需要两个参数，一是前一模块传递下来的序列，二是该序列所属的遗传密码类型。前者每个位置上的密码子，都将成为决定该位置需要采用哪个同义密码子组的标识，可以描述为“特征密码子－同义密码子组”。后者指定了每种同义密码子组内，共涉及哪些具体的密码子，可以描述为“同义密码子组（特征密码子）－密码子”。基于上述两个参数，在一般情况下，只要采用通常的递归算法，即可遍历所有的同义序列。但是从 1.2 小节可知，在某些情况下，前一模块传递来的序列中可能包含多义字符，此时特定位置上的密码子并非具体的某个密码子，而是“多义密码子”，面对这种情况，需要进行一个预处理，即用多义密码子代表的任一具体的密码子，替代该位置的多义密码子，从而可符合一般情况下的处理模式。在递归得到每条同义序列时，同时计算该序列的 AT 含量，并在最终的同义序列集中，根据 AT 含量的高低对集内各成员进行排序，从而便于用户对同义序列列表的浏览。

### 1.4　序列特征显示模块

特征模块为用户提供除 AT 含量以外的其它序列或密码子特征的信息。特征模块支持从本地和远程两种数据源中提取密码子用法（codon usage table，CUT）数据，并计算相关密码子参数的值。对于本地数据源，支持本地 CUT 的导入，且软件本身已自带一些常用物种的 CUT 数据。对于远程数据源，支持从密码子用法数据库[5]（codon usage database，CUD，http://www.kazusa.or.jp/codon）下载需要的 CUT 数据。当前 CUD 可为研究者提供多达 35799 个不同物种的 CUT 数据，通常每个 CUT 中包含“Amino acid”、“Codon”、“Number”、“/1000”和“Fraction”五项主要数据。BestAT 与 CUD 的无缝结合，极大增强了本软件对 CUT 的支持性，且在上述五项数据的基础上，可再自动计算相对同义密码子用法[6]（relative synonymous codon usage，RSCU）、相对适应度[7]（relative adaptiveness，RA）等其它参数的值。用户可选择任何一种参数，在序列上“原位标示”该参数的值，从而便于观察该序列中每个密码子相对宿主的使用情况，比如是否属于偏好密码子或者稀有密码子等。此外，该模块也提供 AT 含量的可视化方式，即 AT 含量曲线。用户可参照这些指标的情况，依据需要或经验，选择特定的同义序列进行实际的实验操作。

## 2　工具的应用

甘油脱氢酶（glycerol dehydrogenase，EC 1.1.1.6）是参与甘油代谢反应的一种氧化还原酶，之前本实验室[8,9]已从克雷伯杆菌（Klebsiella pneumoniae）中扩增出甘油脱氢酶基因 gldA，并采用 pET 表达载体转入 E. coli BL21（DE3），成功实现了该基因的异源表达。以 Nishikubo 等[2]的实验设计作为参考，将 dhaT 的第 2 至 6 位密码子区域的序列，输入 BestAT 进行分析。结果可知，该区域共存在 1728 条同义序列，如图 2 所示，根据 AT 含量的不同，可将同义序列集分为 8 个子集，同义序列的数量在各子集中的分布呈正态式，最低的 AT 含量为 26.7%，最高为 80.0%，原始序列（野生型）处于中等水平为 53.3%。作为前期探索，初步确定具有最高 AT 含量的同义序列作为实验对象，基于对序列中每个具体密码子使用频率的考虑，如表 1 所示，因为 H1 中 ACA 比 H2 中 ACT 的使用频率更高，所以最终确定的实验对象为 H1（优化型）。将野生型和优化型对应的完整基因序列与表达载体 pET－32a（＋）连接，并分别转入 E. coli BL21（DE3）中诱导表达，一定时间后收集菌体超声破碎，提取粗酶液，参考文献[10,11]的方法测定酶活力。结果可知，优化型的酶活力为 191.3 U/mL，而相同条件下野生型的酶活力为 48.3 U/mL，仅是采用 AT 含量优化的方法，就将酶活力提高了 3 倍多[12]。

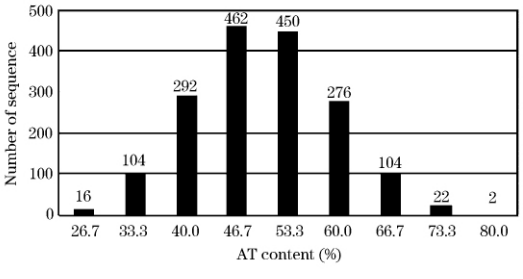



图 2　dhaT 基因第 2 到 6 位密码子区域的同义序列在不同 AT 含量中的分布

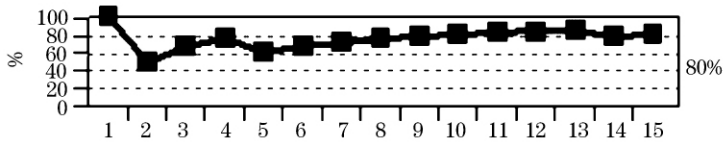
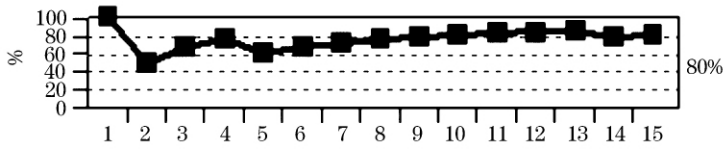
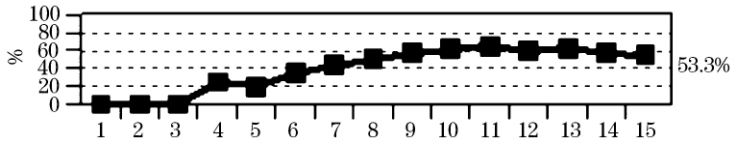
Fig. 2　Distribution of synonymous sequences between the second and sixth codon for dhaT gene by AT content

## 3　总结与展望

多项研究[1-4]表明，起始密码子下游序列的 AT 含量对基因表达水平有明显的影响，通过 PCR 引物法或基因的从头合成，可方便地在这一区域引入突变，获取具有特定 AT 含量的同义序列，但是如何从

表 1 dhaT 基因第 2 到 6 位密码子区域的同义序列及其 AT 含量特征

Table 1 Synonymous sequences between the second and sixth codon for dhaT gene and their AT content features

| Sequence (2 – 6) & RA values | AT curve & AT content |
|---|---|
| AGA – ACA – TAT – TTA – AGA$^{H1}$<br>44.7 – 80.0 – 100 – 46.6 – 44.7 |  |
| AGA – ACT – TAT – TTA – AGA$^{H2}$<br>44.7 – 69.1 – 100 – 46.6 – 44.7 |  |
| CGC – ACT – TAT – TTG – AGG$^{WT}$<br>88.1 – 69.1 – 100 – 34.5 – 24.9 |  |

Notes: H1, high AT content type 1; H2, high AT content type 2; WT, wild type.

巨大的同义序列集中挑选出特定的优化型,是这类实验设计的一个关键。本文研发的 AT 含量优化工具“BestAT”,初步解决了自动获取海量同义序列和充分展示同义序列的密码子用法特性两个问题,并且具有三个特点:(1) 与 CUD 的无缝结合,从 CUD 自动获取密码子用法数据,但同时也支持本地输入;(2) 直观的原位标示,在每个密码子上原位标示密码子参数值;(3) AT 含量的可视化,给出 AT 含量曲线。因此,BestAT 有望成为这类实验设计的一个必不可少的辅助工具。众所周知,基因的表达还受到其它因素的影响,如 mRNA 的二级结构、tRNA 丰度、表达载体、宿主、培养基等等,尤其是与起始密码子下游区域同样存在紧密关系的 downstream box 元件[13,14],对表达水平也有极大的影响。

## 参考文献(References):


[1] Hatfield G W, Roth D A. Optimizing scaleup yield for protein production: Computationally Optimized DNA Assembly (CODA) and Translation Engineering [J]. Biotechnol Annu Rev, 2007, 13: 27 – 42.

[2] Nishikubo T, Nakagawa N, Kuramitsu S, Masui R. Improved heterologous gene expression in Escherichia coli by optimization of the AT – content of codons immediately downstream of the initiation codon [J]. J Biotechnol, 2005, 120(4): 341 – 346.

[3] Wakamatsu T, Nakagawa N, Kuramitsu S, Masui R. Structural basis for different substrate specificities of two ADP – ribose pyrophosphatases from Thermus thermophilus HB8 [J]. J Bacteriol, 2008, 190(3): 1108 – 1117.

[4] Maertens B, Spriestersbach A, von Groll U, Roth U, Kubicek J, Gerrits M, Graf M, Liss M, Daubert D, Wagner R, Schäfer F. Gene optimization mechanisms: a multi – gene study reveals a high success rate of full – length human proteins expressed in Escherichia coli [J]. Protein Sci, 2010, 19(7): 1312 – 1326.

[5] Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000 [J]. Nucleic Acids Res, 2000, 28(1): 292.

[6] Sharp P M, Li W H. The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications [J]. Nucleic Acids Res, 1987, 15(3): 1281 – 1295.

[7] Fuhrmann M, Hausherr A, Ferbitz L, Schödl T, Heitzer M, Hegemann P. Monitoring dynamic expression of nuclear genes in Chlamydomonas reinhardtii by using a synthetic luciferase reporter gene [J]. Plant Mol Biol, 2004, 55(6): 869 – 881.

[8] 张婷婷, 方柏山, 王耿, 王飞飞. 克雷伯杆菌甘油脱氢酶基因的克隆表达与纯化 [J]. 生物工程学报, 2008, 24(3): 495 – 499.

[9] 李梓君, 方柏山, 杨仲丽, 刘嘉. 应用易错 PCR 定向进化甘油脱氢酶 [J]. 华侨大学学报(自然科学版), 2010, 31(6): 661 – 666.

[10] Malaoui H, Marczak R. Separation and characterization of the 1, 3 – propanediol and glycerol dehydrogenase activities from Clostridium butyricum E5 wild – type and mutant D [J]. J Appl Microbiol., 2001, 90(6): 1006 – 1014.

[11] Katrlík J, Mastihuba V, Voštiar I, šefčovičová J, štefuca V, Gemeiner P. Amperometric biosensors based on two different enzyme systems and their use for glycerol determination in samples from biotechnological fermentation process [J]. Analytica Chimica Acta, 2006, 566(1): 11 – 18.

[12] 唐龙盘, 余劲聪, 戴丹凤, 方柏山. 甘油脱氢酶基因在大肠杆菌中的密码子优化表达 [J]. 微生物学报, 2011, 51(4): 76 – 81.

[13] Sprengart M L, Fuchs E, Porter A G. The downstream box: an efficient and independent translation initiation signal in Escherichia coli [J]. EMBO J, 1996, 15(3): 665 – 674.

[14] Etchegaray J P, Inouye M. Translational enhancement by an element downstream of the initiation codon in Escherichia coli [J]. J. Biol. Chem., 1999, 274(15): 10079 – 10085.