研究报告
Research Report

茶树中GC2生物决定基因的表现度

Supriyo, Prosenjit
生物技术系, 阿萨姆邦大学, silchar-788011, 阿萨姆邦, 印度
通讯作者, supriyoch_2008@rediffmail.com; 作者

计算分子生物学, 2014 年, 第 3 卷, 第 9 篇
收稿日期: 2014 年 08 月 07 日 接受日期: 2014 年 08 月 07 日 发表日期: 2014 年 08 月 07 日

摘 要 基因表达的有效性受到在基因的编码序列(cds)中使用的密码子的性质的影响。这是由于大多数基因和生物体不均匀使用同义密码子。优先使用某些同义密码子这种现象称为密码子使用偏移 (CUB)。我们分析了在每个密码子位点的标准化的 AT 和 GC 频率。我们观察到基因表达(通过 CAI 测量)和任何密码子位点处的 GC 含量之间的相关性非常弱，除了 GC2 显示与基因表达中度正相关。我们还测量了三个密码子位点的 CAI 和 AT 含量之间的相关性。AT2 与基因表达呈中度负相关。我们进一步观察到 RCBS(基因表达的测量)和 cds 长度之间的强相关性，表明自然选择可能有利于较短基因在更高水平表达。对于该分析，我们最初下载茶树的 350 个编码序列，其中仅发现十个 cds 以起始密码子 ATG 开始，并且长度为三个碱基的精确倍数并且缺乏 N(任何未知碱基)。我们对这十个 cds 的分析显示，在确定基因表达性方面，茶树中同义密码子的第二个位置可能比第三个位置发挥更显着的作用，这从 CUB 和相关分析中可以看出。

关键词 基因表达；相对密码子使用偏移(RCBS)；密码子适应指数(CAI)；密码子使用偏移(CUB)

GC2 Biology Dictates Gene Expressivity in *Camellia sinensis*

Supriyo Chakraborty, Prosenjit Paul
Department of Biotechnology, Assam University, Silchar-788011, Assam, India
Corresponding author, supriyoch_2008@rediffmail.com; Authors

Abstract The effectiveness of the gene expression is influenced by the nature of codons used throughout the coding sequence (cds) of the gene. This is due to the fact that most genes and organisms do not use synonymous codons uniformly. Certain synonymous codons are used preferentially and this phenomenon is called codon usage bias (CUB). We analyzed normalized AT and GC frequency at each codon site. We observed that the correlations between gene expression (measured by CAI) and GC content at any codon site were very weak except GC2s showed moderate positive correlation with gene expression. We also measured the correlations between CAI and AT content at three codon sites. AT2s showed moderate negative correlation with gene expression. We further observed a strong correlation between RCBS (a measure of gene expression) and cds length indicating that natural selection is probably operating in favor of shorter genes to be expressed at higher level. For this analysis, we initially downloaded 350 coding sequences of *Camellia sinensis*, out of which only ten cds were found to begin with the initiator codon ATG, and length as exact multiple of three bases and devoid of N (any unknown base). Our analysis on these ten cds revealed that the second position of synonymous codons in *Camellia sinensis* possibly plays a more prominent role than the third position in determining the gene expressivity as evident from the CUB and the correlation analyses.

Keywords Gene expression; Relative codon usage bias (RCBS); Codon adaptation index (CAI); Codon usage bias (CUB)

基因表达的有效性受到整个基因中所使用密码子的性质影响。由于进化过程，即始终保守的基因在编码序列中几乎没有保持不变。这是由于大多数基因和生物体不使用同义密码子这一事实。优先使用某些同义密码子，称为密码子使用偏好(CUB)的现象。

密码子偏移 ，同义密码子的不等同使用，在物种之间变化很大，在某些情况下，还报道了同一生物体内不同基因之间密码子使用偏移 的显着变化(Bernardi, 1993)。以前的密码子使用分析表明，密码子使用偏移 非常复杂，并且与各种生物因子如基因表达水平相关(Gouy and Gautier, 1982; Sharp and Li, 1986; Sharp et al., 1986; Sharp and Li, 1987)，基因长度(Bains, 1987; Eyre-Walker, 1996)，基因翻译起始信号(Ma, 2002)，蛋白质氨基酸组成(Lobry and Gautier, 1994)，蛋白质结构(D'Onofrio et al., 2002)，tRNA 丰度(Ikemura, 1981, 1982)，突变频率和模式(Sueoka, 1999)和 GC 组成(Sueoka and Kawanishi, 2000)。GC 偏移的影响对密码子偏移 具有主要影响，导致第三个密码子位置的 GC%(也称为 GC3 和 GC 偏移)之间的密切关联(Sueoka, 1988)。由于所有氨基酸(除甲硫氨酸和色氨酸外)在第三位置允许 GC 改变的同义取代，这导致了一种普遍的想法，即使用同义 G/C-末端密码子应当随着 GC 偏好增加而频率增加，而使用 A/T 终止密码子应该减少(Wan et al., 2004)。

茶(*Camellia sinensis*)是世界上最受欢迎的饮料之一，归因于其多样化的品种和品质、味道和促进作用，也对人体健康有益处。它由于有迷人的香气、愉快的味道和许多药用的好处吸引了民众许多注意，从公元前 3000 已经在社会上开始消费(Kliman and Bernal, 2005)。 在茶叶中也发现了许多次生代谢物，例如多酚、生物碱(例如咖啡因)、维生素(A，B1，B2，E，C)、多糖、挥发油和矿物质(Lin et al., 2003)。

尽管它在遗传学的几个领域中具有根本的重要性，但是长期以来一直在努力测量 CUB。测序技术的进步提供了不同的生物丰富的基因组数据。随着许多生物体的全基因组测序的出现，CUB 的研究再次引起学者关注。在本文中，我们提出了通过分析密码子适应指数(CAI)研究茶树 CUB、相对密码子使用偏性(RCBS)、优化密码子频率(Fop)、相对同义密码子的使用价值(Fop)、有效密码子数目(ENc)、GC 含量、GC 偏移和 AT 偏移。

1 结果

1.1 总密码子使用分析

由于全基因组序列不可用于茶树，在本研究中仅使用 10 个基因。表 1 显示了所选择的基因及其登录号以及总体 RCBS、CAI、GC%、GC1s、GC2s 和 GC3s 值。发现茶树的编码序列富含 A 和/或 T。但是在铜绿假单胞菌 中，以 G 和/或 C 结束的密码子在整个编码区中明显占优势(Gupta and Ghosh, 2001)。然而，总体密码子使用值可能掩盖了基因中密码子使用偏移 的一些异质性，这可能叠加在这个生物的极端基因组组成上。

1.2 密码子使用变异

在第三同义密码子位置(GC3s)处由基因使用的密码子的有效数目(Nc)和(G + C)百分比用于研究茶树中的基因之间的密码子使用变异。Wright(1990)建议 Nc 对 GC3s 的图可以有效地用于探索基因之间的密码子使用变异。Wright(1990)证明，如果基因的密码子使用偏差具有除基因组 GC 组成以外的某些影响，则基因的实际分布与未选择下的预期分布的比较可以是指示性的。

图 1 显示了茶树中不同基因的 Nc 分布。Nc 的平均值和标准偏差值分别为 15.2 和 0.42637，表明基因中密码子使用偏差有很大的变化。基因中密码子使用偏差的变化从第二个同义密码子位置的(G + C)分布进一步证实，如图 2 所示。这些结果表明，除了组成约束，其他趋势可能影响总体密码子使用茶树中的基因变异。

1.3 RCBS 和 CAI 值之间的关系

每个基因已经演变出适应密码子使用模式的基因表达水平，并且 RCBS 值>0.5 和 CAI 值>0.5 表现出有利的密码子使用。因此，我们选择这两个指数作为基于书面证据的有效表达度量。CAI 和 RCBS 基因的表达水平已经显示。从我们的分析，我们已经发现十个基因中的六个具有 RCBS 和 CAI 值，每个大于 0.5，表明这六个茶树中的基因可以作为高度表达的基因。

当我们将它们绘制在图上时，RCBS 和 CAI 显示类似的模式(图 3)。我们进一步分析了编码区的长度和基因的表达水平之间的关系。与以前的其他研究(Ikemura, 1981; Ikemura, 1982; Moriyama and Powell, 1998)一致，我们的数据支持较小尺寸的高度表达的基因。我们观察到 RCBS 随编码的蛋白质的长度减少。在 RCBS 和蛋白质长度之间观察到显着的负相关。在图 4 中，我们将 RCBS 作为基因长度的函数。

表 1 RCBS, CAI, CDS 长度, GC 含量分析和茶树中基因的登录号

Table 1 RCBS, CAI, CDS length, GC content analysis and accession number for *Camellia sinensis* genes

Sl. no	Gene name	Accession number	CDS length (bp)	CAI	RCBS	GC content (%)			
						GC	GC1	GC2	GC3
1	Acetyl CoA carboxylase	DQ366599	1800	0.6351	0.0413	47.6	55.0	42.3	45.3
2	Polyphenol oxidase (PPO)	FJ656220	1800	0.6118	0.0408	49.1	52.3	40.5	54.3
3	pRB mRNA for retinoblastoma related protein	AB247284	3078	0.5455	0.0252	42.9	47.9	43.9	37.1
4	cycD3-2 mRNA for cyclin D3-2	AB247283	1119	0.3605	0.0637	43.3	50.4	34.3	45.0
5	cycD3 mRNA for cyclin D3-1	AB247282	1116	0.5152	0.0629	46.7	53.0	38.2	48.9
6	cycb mRNA for cyclin B	AB247280	1323	0.4688	0.0541	45.4	53.7	39.5	43.1
7	*Stearoyl acyl carrier protein desaturase*	KC242133	1191	0.3339	0.0599	45.3	54.2	37.8	44.1
8	Cultivar Longjing43 glycerol-3-phosphate acyltransferase	KC920896	1353	0.4935	0.0529	46.4	52.5	42.1	44.6
9	Omega-3 fatty acid desaturase (FAD8)	KC847167	1359	0.5694	0.0536	46.7	53.0	42.2	44.8
10	AMP deaminase	KC700025	2571	0.5943	0.0298	44.2	52.5	37.5	42.6

GC 含量的理想百分比范围在 30%至 70%之间(图 5 和图 6)。在该范围之外的任何峰对转录和翻译效率有不利影响。在十个基因中，发现基因多酚氧化酶(PPO)具有在 32%至 75%的理想范围之外的 GC 含量。这些结果表明，存在于基因的 CDS 序列中的 GC 含量影响该特定基因的表达性。通过将 GCI 值对 GC 含量作图，再次证实了基因中 GC 含量的差异，并且发现由于它们的 GC 含量，两个基因显示与基因表达性呈负相关性(图 7)。

1.4 GC 和 AT 含量与基因表达模式的关系

我们分析了在每个密码子位点的标准化的 AT 和 GC 频率。我们观察到通过 CAI 测量的基因表达和在任何密码子位点的 GC 含量之间的相关性非常弱(rGC1 = 0.069，rGC2 = 0.604 和 rGC3 = 0.186)(图 8)。因此，与其他相矛盾的是，第三个密码子位置的 GC 含量是茶树中基因表达的非常差的预测因子(Sharp and Lloyad, 1993; Gerton et al., 2000; Marin et al., 2003)。但在第二个密码子位置，GC 含量显示与基因表达中度正相关。由于茶树的编码序列富含 AT，我们还分析了每个密码子位点的 AT 频率，发现第二个密码子位置显示与基因表达的中度负相关。

2 讨论

简而言之，我们已经提出了基因的表达测量，设计了从相对密码子偏好和密码子适应指数来预测基因表达水平。基于基因表达和密码子组成是强相关的假设，已经定义了密码子适应指数以提供对基因中密码子偏好程度的直观有意义的测量。我们概述了一个简单的方法来评估基因中密码子偏移 指数的强度，作为其可能的表达水平的指导，并结合茶树基因分析进行说明。

本研究的目的是：(a)分析 CAI、RCBS、GC 偏差、GC 含量、茶树的基因的密码子的相对位置。(b)将上述参数与基因表达模式相关联。根据我们在本研究中提到的目标，我们从茶树中选择十个基因进行 CUB 分析。使用由我们开发的 perl 程序来检索具有正确的初始和终止密码子的准确编码序列。为了使取样误差最小化，我们仅取得大于或等于 1000bp 的那些编码序列。通过使用由我们开发的基于 PERL 的程序计算所有上述用于 CUB 分析的参数。

在分析茶树的编码序列后，发现基因富含 AT。但是在铜绿假单胞菌的情况下，显然在整个编码区中以 G 和/或 C 结尾的密码子占主导。我们还通过分析密码子的有效数(Nc)来预测密码子使用的异质性。Nc 的平均值和标准偏差分别为 15.2 和 0.42637，说明茶树中的基因密码子使用偏差存在较大差异。基因之间密码子使用偏差的变化从第三同义密码子位置处的(G + C)的分布进一步证实。这些结果表明，除了组成限制，其他趋势可能会影响茶树中基因的总体密码子使用变异。

每个基因已经演变出适应密码子使用模式的

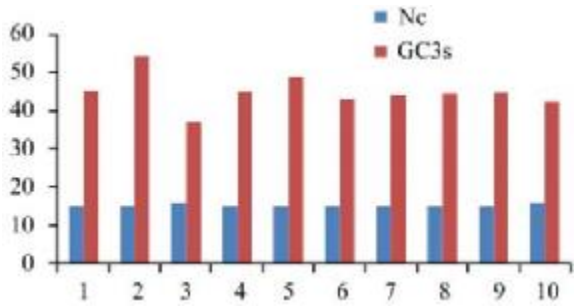

图 1 茶树基因的 Nc 分布

Figure 1 Nc distribution of *Camellia sinensis* genes

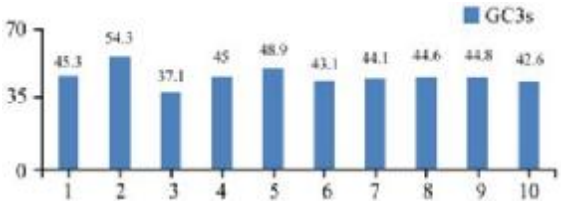

图 2 茶树基因的 GC3s 分布

Figure 2 GC3s distribution of *Camellia sinensis* genes

基因表达水平，并且 RCBS 值>0.5 和 CAI 值>0.5 表现出有利的密码子使用。我们计算了这些基因的 CAI 和 RCBS 值，发现茶树中 10 个基因中有 6 个是高度表达的基因。我们还分析了 GC 含量对密码子相对位置的分布;结果显示，除了基因 PPO，所有其他基因具有理想的 GC 百分比。

我们分析了在每个密码子位点的标准化的 AT 和 GC 频率。我们观察到通过 CAI 测量的基因表达和在任何密码子位点的 GC 含量之间的相关性非常弱。GC2 与基因表达呈中度正相关(0.604)。我们还测量了在任何密码子位点的 CAI 和 AT 含量之间的相关性。AT2 与基因表达显示中度负相关(-0.604)。

此外，我们的分析进一步揭示了在确定基因表达水平上茶树中同义密码子的第二个位置比第三个位置发挥更显著的作用，如由 CHAI 和 GC2s 之间的正相关系数(0.064)与具有 GC1 和 GC3 的 CAI 的相关系数(0.069 和 0.187)相比所揭示的现象。这与以下事实矛盾：大肠杆菌中密码子的第三位置在确定基因表达中起主要作用，尽管茶树和大肠杆菌都是富含 AT 的。与具有 AT1 和 AT3 的 CAI 的相关系数(-0.069 和-0.172)相比，CAI 和 AT2s 之间的最高负相关性(-0.064)进一步证实了这一点。这可能是由于本研究用了少量的编码序列，并且只有具有高 CAI 和 RCBS 的基因。

cds 的组成偏差在形成密码子使用中起关键作

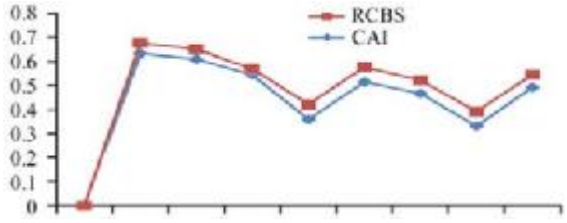

图 3 RBCS 和 CAI 值之间的关系

Figure 3 The relationship between RCBS and CAI values

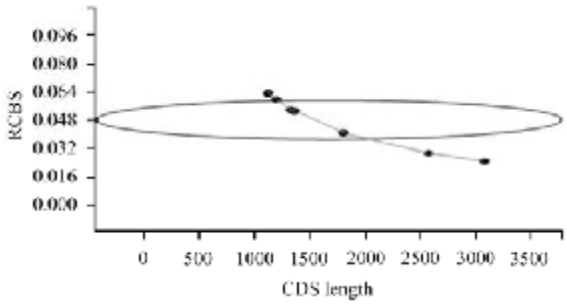

图 4 CARBS 和蛋白质长度之间的关系

Figure 4 Relationship between RCBS and protein length

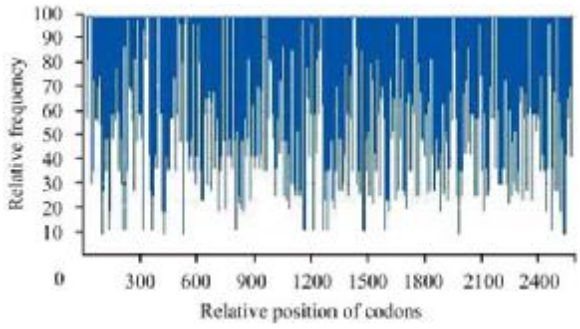

图 5 基因 AMP 脱氨酶 CDS 长度的密码子使用频率分布

Figure 5 The distribution of codon usage frequency along the length of the CDS for the gene AMP deaminese

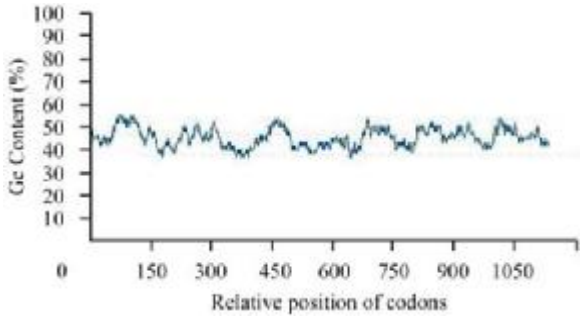

图 6 硬脂酰基载体蛋白去饱和酶基因的 GC 含量的百分比范围

Figure 6 The percentage range of GC content for the Stearoyl acyl carrier protein desaturase gene

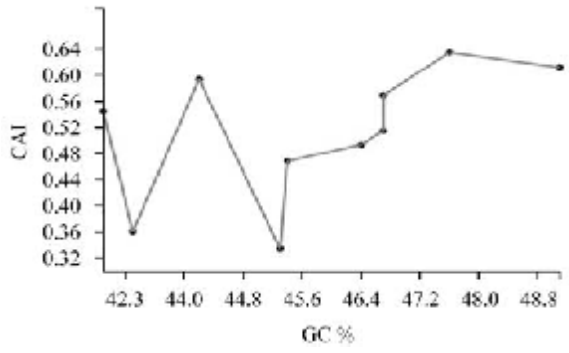

图 7 茶树基因中 GC 含量 CAI 图

Figure 7 CAI plotted against the GC content for *Camellia sinensis* genes

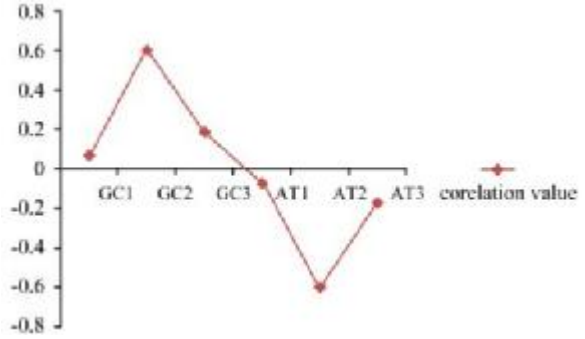

图 8 不同密码子位置的 CIA 和 GC / AT 含量之间的相关性

Figure 8 Correlation between CAI and GC/AT content at different codon positions

用。GC 含量对密码子使用偏差具有主要影响，致使在第三密码子位置的 GC%之间的密切关联，也称为 GC3 生物学。由于所有氨基酸(甲硫氨酸和色氨酸除外)允许在密码子第三位置进行 GC 改变的同义取代，这提出了即使用同义 G / C-末端密码子可以增加基因的表达能力，而使用的 A / T 终止密码子可以降低基因表达的水平的普遍观点。对于该分析，我们最初下载茶树的 350 个编码序列，其中仅发现十个 cds 以起始密码子 ATG 开始，并且长度为三个碱基的精确倍数并且缺乏 N(任何未知碱基)。从 cds 的 CUB 分析和在三个密码子位点的 GC / AT 含量与 CAI 值之间的相关分析显而易见，结果表明，在决定基因表现力方面，茶树中同义密码子的第二位置可能起到比第三位密码子更突出的作用。

3 材料与方法

3.1 数据收集

从 NCBI(www.ncbi.nlm.nih.gov)下载 Camellia sinensis 的编码序列(cds)。为了使抽样误差最小化，我们只取得大于或等于 1000bp 并具有正确的初始和终止密码子并缺少 N(任何未知碱基)的 cds。使用我们开发的 perl 中的程序来检索精确的编码序列。最后，选择 10 个序列用于 CUB 分析。

3.2 模型

使用相对密码子使用偏移 和密码子适应指数来研究基因之间的总体密码子使用变化。RCBS 是在随机密码子使用的假设下，观察到的密码子的频率与期望频率的差异，其中基础组合物在序列中的三个位点存在偏移。RCBS 是指示基因中每个密码子的 RCB 的影响的基因的总得分。RCB 反映了基因表达的水平。基因的表达量度由 RCBS 表示(Hertog et al., 1993)。RCBS 值接近 0 表示密码子缺乏偏移，因此可用于比较不同组的基因。

基因表达水平与基因的密码子使用差异有关，该基因密码子在三个密码子位点处偏向核苷酸成分。令 f(x，y，z)为基因的密码子三联体(x，y，z)的归一化密码子频率。然后，基因中密码子三联体(x，y，z)的相对密码子偏移(RCB)定义为：

$$dxyz=\frac{f(x,y,z)-f1(x)f2(y)f3(z)}{f1(x)f2(y)f3(z)}$$

其中，f1(x)是在第一密码子位置的 x 的归一化频率，f2(y)是在第二密码子位置的 y 的归一化频率，f3(z)是基因的第三个密码子位置处的 z 的归一化频率。频率 f1，f2，f3 已经从基因的密码子样品集合导出，并且频率的归一化在密码子的基因长度上进行，以试图补偿 RCB 与密码子总数的预期增加，以这样的方式定量基因的密码子偏好的程度，使得可以在基因组内和基因组之间进行比较。如前所述，d_{xyz} 包含比其他更多的定量信息，因为它考虑了密码子使用以及基本组成偏差。基因的表达量度为：

$$RCBS = \left(\prod_{i=1}^{L}(1+d^{i}_{xyz})\right)^{1/L} - 1$$

其中，d^{i}_{xyz} 是基因的第 i 个密码子的使用差异。L 是基因中的密码子数。

基因表现力再次通过计算参数密码子适应指数来测定(Sharp and Li, 1986)。它基本上测量从给定基因到参考基因的氨基酸密码子用途的距离。CAI 将翻译最佳密码子定义为在高度表达的基因中频繁出现的那些密码子。

$$CAI(L(g)) = \exp\left(\frac{\sum_{c=1}^{L} \log w_{c(i)}}{L}\right) = \left(\prod_{i=1}^{L} w_{c(i)}\right)^{1/L}$$

其中，*L* 是基因 g 的长度，$w_c(1)$是密码子 c 在参考基因(不是 g)中的相对适应性。相对适应性定义为：

$$w_c = \frac{f_c}{\max(f_s)}, s \in \{c_a\},$$

其中，fc 是密码子 c 的频率，密码子 c 是基因 g 中的 l^{th} 密码子。a 是由 *c* 编码的氨基酸，$\{C_a\}$是编码氨基酸 a 的同义密码子集合。某些密码子将在基因中出现多次。因此，我们可以重写方程式来对密码子而不是长度求和，并且使用计数而不是频率。这使得对实际基因的依赖更清楚。更常见的公式是：

$$CAI(o(g)) = \exp\left(\frac{1}{O_{tot}} \sum_{c \in C} o_c \log w_c\right) = \left(\prod_{c \in C} o_c \log w_c\right)^{o_{tot}}$$

有效密码子数(Nc)是序列中使用的不同密码子的总数(Wright, 1990)。Nc 的值范围从 20(其中每个氨基酸只使用一个密码子)至 61(对于标准遗传密码)，其中所有可能的同义密码子以相同的频率使用。Nc 测量偏向使用较小的密码子子集，远离等同使用同义密码子。例如，如上所述，高度表达的基因选择使用较少的密码子。Nc 的基本思想类似于来自群体遗传学的接合性概念，其涉及来自两个生物体的基因的相似性。

在密码子使用的环境中，多个同义密码子以类似于多个等位基因的方式进行处理。氨基酸 Za 的纯合性测量相似性程度，并基于相对密码子频率计算：

$$Z_a = \frac{O_a \sum_{c \in Ca} \int \frac{2}{ac} - 1}{O_a - 1}$$

氨基酸的有效密码子的数目是纯合性的倒数：

$$Na = Za^{-1}$$

The value of Na ranges from 1 to the number of synonymous codons ka (the codon degeneracy). With equal codon usage, homozygosity is minimal and the value of Na is the number of synonymous codons. The overall number of effective codons for a gene (Nc) is a sum of average homozygosities Za for different redundancy classes k (in set K of all redundancy classes):

Na 的值的范围从 1 到同义密码子 ka 的数目(密码子简并性)。使用相等的密码子，纯合性是最小的，Na 的值是同义密码子的数目。基因的有效密码子的总数(Nc)是不同冗余类 k(所有冗余类别的集合 K)的平均纯合子 Za 的总和：

$$Nc = \sum_{k=K} n_k \overline{N}_{a-k}$$

每个冗余类别：

$$\overline{N}_a = \frac{1}{n_k} \sum_{a \in K_k} N_a$$

当密码子使用模式比预期更均匀时，可以发生 Nc>61，在这种情况下将其重新调整为 61.如果没有观察到氨基酸或非常罕见，则将该值替换为同一冗余类别中氨基酸的平均纯合性。如果 Ile 氨基酸缺失(具有三个同义密码子的冗余类中的唯一成员)，则从其他冗余类别的平均纯合性估计相应的 Z。

例如，在异亮氨酸的情况下：

$$Z_{k=3} = \frac{1}{3}\left(\left(\frac{2}{\overline{Z}_{k=3}} - 1\right)^{-1} - \left(\frac{2}{3\overline{Z}_{k=4}} - \frac{1}{3}\right)^{-1} + \left(\frac{2}{5\overline{Z}_{k=6}} - \frac{3}{5}\right)^{-1}\right)$$

当基因的氨基酸之间存在大的差异时，可以使用所有单个氨基酸的 Nc 之和，而不是采用每个冗余类别的平均值的总和：

$$Nc = \sum_{a \in A} N_a$$

GC3s 是(G + C)的频率，A3s，T3s，G3s 和 C3s 是 A，T，G 和 C 在密码子的同义第三位置的分布(Gupta and Ghosh, 2001)。GC 偏移和 AT 偏移分别定义为 DNA 序列的(G-C)与(G + C)和(A-T)与(A + T)的比率(Wright, 1990)。

3.3 分析

所有上述参数通过使用我们开发的 PERL 程序计算。此后我们测量了所有上述参数与茶树的基因表达性之间的相关性。

作者贡献

S.C 对本研究进行构思，并进行软件分析。P.P. 负责分析数据集并撰写初稿，整理数据和表格。所有作者阅读并同意了最终稿件。

致谢

我们感谢在这项研究 Assam University, Silchar, Assam, India 提供必要的设施。我们真诚地感谢塞森博士，主任和其他工作人员，以及阿萨姆大学计算机中心和 Silchar 的帮助，他们为这项研究工作提供互联网接入的支持。

参考文献

Bains W., 1987, Codon distribution in vertebrate genes may be used to predict gene length, J Mol. Biol., 197(3): 379-388
http://dx.doi.org/10.1016/0022-2836(87)90551-1

Bernardi G., 1993, The vertebrate genome: isochores and evolution, Mol. Biol. Evol., 10: 186-204

D'Onofrio G., Ghosh T.C., and Bernardi G., 2002, The base composition of the genes is correlated with the secondary structures of the encoded proteins, Gene, 300(1-2): 179-187
http://dx.doi.org/10.1016/S0378-1119(02)01045-4

Eyre-Walker A., 1996, Synonymous codon bias is related to gene length in *Escherichia coli:* selection for translational accuracy? Mol Biol Evol., 13(6): 864-872
http://dx.doi.org/10.1093/oxfordjournals.molbev.a025646

Gerton J.L., DeRisi J., Shroff, R., Lichten M., Brown P.O., and Petes T.D., 2000, Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*, Proc. Natl. acad. Sci. USA, 97(21), 11383-11390
http://dx.doi.org/10.1073/pnas.97.21.11383

Gouy M., and Gautier C., 1982, Codon usage in bacteria: correlation with gene expressivity, Nucleic Acids Res., 10: 7055-7074
http://dx.doi.org/10.1093/nar/10.22.7055

Gupta S.K., and Ghosh T.C., 2001, Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*, Gene, 273: 63-70
http://dx.doi.org/10.1016/S0378-1119(01)00576-5

Hertog H.G., Hollman P.C., Katan M.B., and Kromhout D., 1993, Intake of potentially anticarcinogenic flavonoids and their determinations in adults in the Netherlands, Nutr. Cancer, 20(1): 21-29
http://dx.doi.org/10.1080/01635589309514267

Ikemura T., 1981, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, J Mol. Biol., 151(3): 389-409
http://dx.doi.org/10.1016/0022-2836(81)90003-6

Ikemura T., 1982, Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs, J Mol. Biol., 158(4): 573-597
http://dx.doi.org/10.1016/0022-2836(82)90250-9

Kliman R.M., and Bernal C.A., 2005, Unusual usage of AGG and TTG codons in humans and their viruses, Gene, 352: 92-99
http://dx.doi.org/10.1016/j.gene.2005.04.001

Lobry J.R., and Gautier C., 1994, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, Nucleic Acids Res., 22(15): 3174-3180
http://dx.doi.org/10.1093/nar/22.15.3174

Ma J., Campbell A., and Karlin S., 2002, Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures, J Bacteriol., 184(20): 5733-5745
http://dx.doi.org/10.1128/JB.184.20.5733-5745.2002

Marin A., Gallardo M., Kato Y., Shirahige K., Gutiérrez G., Ohta K., and Aguilera A., 2003, Relationship between G+C content, ORF length and mRNA concentration in *Saccharomyces cerevisiae*, Yeast, 20(8): 703-711
http://dx.doi.org/10.1002/yea.992

Moriyama E.N., and Powell J.R., 1998, Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*, Nucleic Acids Res.,

26(13): 3188-3193
http://dx.doi.org/10.1093/nar/26.13.3188

Roymondal U., Das S., and Sahoo S., 2009, Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome, DNA Res., 16(1): 13-30
http://dx.doi.org/10.1093/dnares/dsn029

Sharp P.M., and Li W.H., 1986, An evolutionary perspective on synonymous codon usage in unicellular organisms, J Mol. Evol., 24(1-2): 28-38
http://dx.doi.org/10.1007/BF02099948

Sharp P.M., and Li W.H., 1987, The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res., 15(3): 1281-1295
http://dx.doi.org/10.1093/nar/15.3.1281

Sharp P.M., and Lloyad A.T., 1993, Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure, Nucleic Acids Res, 21(2): 179-183
http://dx.doi.org/10.1093/nar/21.2.179

Sharp P.M., Tuohy T.M., and Mosurski K.R., 1986, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes, Nucleic Acids Res., 14: 5125-5143
http://dx.doi.org/10.1093/nar/14.13.5125

Sueoka N., 1988, Directional mutation pressure and neutral molecular evolution, Proc. Natl. Acad. Sci., 85(8): 2653-2657
http://dx.doi.org/10.1073/pnas.85.8.2653

Sueoka N., 1999, Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C, J Mol. Evol., 49(1): 49-62
http://dx.doi.org/10.1007/PL00006534

Sueoka N., and Kawanishi Y., 2000, DNA G+C content of the third codon position and codon usage biases of human genes**,** Gene, 261(1): 53-62
http://dx.doi.org/10.1016/S0378-1119(00)00480-7

Wan X.F., Xu D., Kleinhofs A., and Zhou J., 2004, Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes, BMC Evolutionary Biology, 4: 19
http://dx.doi.org/10.1186/1471-2148-4-19

Wright F., 1990, The ‘effective number of codons’ used in a gene, Gene, 87(1): 23-29
http://dx.doi.org/10.1016/0378-1119(90)90491-9

Y.S. Lin, Y.J. Tasi, J.S. Tsay, and J.K. Lin, 2003, Factors affecting the levels of tea polyphenols and caffeine in tea leaves, J. Agric. Food Chem., 51(7): 1864-1873
http://dx.doi.org/10.1021/jf021066b