

球孢白僵菌基因组密码子使用模式 及其影响因素分析^①

张瑞芝¹, 陈章宝¹, 陈德敏², 渠 铮², 向 恒²

1. 西南大学 药学院, 重庆 400715; 2. 西南大学 动物科技学院, 重庆 400715

摘要: 球孢白僵菌 *Beauveria bassiana* 是一类成功用于全球生物防治的重要病原真菌, 这是由于其宿主范围广及毒力多样化特性决定的. 利用球孢白僵菌已公布全基因组数据, 开展密码子使用偏好分析, 以理解其基因组特征和宿主适应性. 尽管通过密码子指数 CAI, CBI, ENc 的计算, 认为球孢白僵菌密码子使用偏性较弱. 但根据 RSCU 值鉴定最优密码子, 发现球孢白僵菌基因组强烈偏好使用第三位碱基 G/C 的密码子, 尤其是 C 结尾的密码子. 对密码子第三位碱基进行 PR2 分析, 发现球孢白僵菌密码子偏好性还表现在倾向于使用嘧啶结尾密码子, 即 U 使用频率高于 A, C 使用频率高于 G. 该偏好模式通过相关性分析, 认为核苷酸组成是影响球孢白僵菌密码子偏好性的一个重要因素. 而 ENc 图 and 对应分析则表明除核苷酸组成外, 球孢白僵菌密码子使用偏好还受到基因功能及其他多种因素的影响. 中性绘图分析认为, 球孢白僵菌密码子使用偏好受到外界选择下的定向突变压力作用, 并推测该选择来自于其宿主. 而该定向突变压力也是加快球孢白僵菌自身基因组进化和毒力多样化的一种动力, 使其能适应在 700 多种昆虫宿主中寄生.

关键词: 球孢白僵菌; 密码子使用偏好; 定向突变压力; 核苷酸组成

中图分类号: Q949.32

文献标志码: A

文章编号: 1673-9868(2018)05-0066-09

作为一类知名的昆虫致病菌, 球孢白僵菌 *Beauveria bassiana* 由于其广泛的宿主范围(超过 700 种昆虫^[1])及感染不同昆虫的毒力多样性, 被世界各地广泛用于害虫生物防治^[2], 在农林业生产中展现出了巨大的价值. 其中, 作为全球最大的害虫防治项目, 我国利用球孢白僵菌防治亚洲玉米螟 *Ostrinia furnacalis* 和马尾松毛虫 *Dendrolimus punctatus*^[3-4], 取得了显著成效. 此外, 球孢白僵菌能够产生大量的次生代谢物, 这些代谢物在工业上也具有巨大的应用价值.

球孢白僵菌 ARSEF 2860 的全基因组测序计划由中国科学院上海生命科学研究院和浙江大学的科学家在 2012 年完成^[5]. 通过第二代测序技术, 获得了 76.6 倍覆盖度的基因组测序数据, 并组装成 33.7 Mb 基因组大小的 242 条 scaffolds 数据, 以及 10 364 个蛋白质编码基因被注释. 通过 interproscan 分析, 7 283 个基因被归类为 3 002 个基因家族, 这表明球孢白僵菌存在着普遍的基因家族膨胀现象^[5]. 许多物种特异的毒力基因, 诸如编码一类小型富含半胱氨酸分泌蛋白的 SSCPs 基因, 在球孢白僵菌基因组中被鉴定^[5]. 上述 2 种特征被认为与其宿主范围和致病策略有关. 此外, 球孢白僵菌具有特殊的表达转录因子(TFs)用于调控和激活特定基因, 以适应不同的昆虫宿主^[5].

生物体内, 在遗传物质(DNA 或 RNA)编码蛋白质(protein)的翻译过程中, 遗传密码(genetic code)是一套起着关键作用的规则. 标准遗传密码包含 61 个编码、20 种氨基酸的密码子和 3 个终止密码子. 由于除

^① 收稿日期: 2017-03-23

基金项目: 国家自然科学基金项目(31302036); 中央高校基本科研业务费项目(XDJK2015B024, XDJK2009C149); 西南大学本科生科技创新基金项目(20152802005).

作者简介: 张瑞芝(1982-), 女, 硕士, 实验师, 主要从事病原微生物学研究.

通信作者: 向 恒, 博士, 高级实验师.

蛋氨酸 Methionine 和色氨酸 Tryptophan 外, 其余氨基酸都可以由 2 个其以上的密码子进行编码, 使得密码子具有简并性. 这些简并的同义密码子 synonymous codons 虽然编码同一个氨基酸, 但在不同物种中使用的频率却不尽相同^[6], 由此产生的现象称之为密码子使用偏好性(Codon Usage Bias). 密码子偏好性广泛存在于生物体中, 并受核苷酸组成^[7]、翻译过程^[8]、tRNA 丰度^[9]、基因功能^[10]、mRNA 结构^[11]、基因长度^[12]、蛋白质疏水结构^[13]、环境温度^[14]等多种因素的影响, 尤其是基因突变与自然选择压力之间的平衡决定了密码子的使用偏好^[15]. 因此, 针对遗传密码子使用偏好的分析将能更好地理解不同生物体在分子进化、环境适应、基因组特征等方面的异同, 是基因组学中的一个重要课题.

考虑到球孢白僵菌巨大的经济价值及其宿主范围广的特性, 本文将通过生物信息学方法分析球孢白僵菌基因组数据, 确定其密码子使用模式, 明确其密码子偏好性, 并描述影响其密码子使用偏好的主要因素, 从而了解其基因组特征, 为球孢白僵菌的真菌-昆虫互作研究奠定科学基础.

1 材料和方法

1.1 基因组及编码基因序列数据

本文采用的球孢白僵菌 ARSEF 2860 基因组数据(序列号 NZ_ADAH00000000)及 10 364 个蛋白质编码基因序列, 均下载自 GenBank 数据库(<http://www.ncbi.nlm.nih.gov/genbank>).

1.2 核苷酸组成及中性绘图和 PR2 分析

运用自编 PERL 脚本(可从网址 https://github.com/hxiang1019/calc_GC_content.git 下载), 计算上述 10 364 个球孢白僵菌编码基因及其密码子第一、第二、第三位碱基上的 GC 含量, 分别记为 GC_{cds} , P_1 , P_2 , P_3 . 将各基因 P_1 和 P_2 的平均值 P_{12} 作为纵坐标, 以 P_3 作为横坐标, 进行中性绘图分析(Neutrality analysis), 以确定密码子 3 位置间核苷酸组成相关性, 进而检测突变选择平衡是否显著影响密码子使用偏好^[16]. 中性图中每个点代表一个编码基因, 当该基因的 P_{12} 与 P_3 显著相关时, 表明其变异更多是受到自身遗传漂变的非定向突变压力作用, 而很少受到或不受到外界选择下定向突变压力的作用, 因而成中性, 该点应位于对角线上; 而如果相关性不显著, 该点应位于水平线上, 表明其变异受到很大程度的定向突变压力作用. 中性绘图分析回归线越呈水平状态, 其斜率越接近于 0, 表明 P_{12} 与 P_3 越不相关, 密码子使用偏好就更多地受定向突变压力影响; 反之, 斜率越接近于 1, 表明 P_{12} 与 P_3 相关, 密码子偏好就更多地受遗传漂变作用^[17].

此外, 本研究也计算了 10 364 个基因在密码子第三位上的核苷酸组成, 记为 A_3 , U_3 , C_3 , G_3 , 进而以 AU 偏量 $[A_3/(A_3+U_3)]$ 和 GC 偏量 $[G_3/(G_3+C_3)]$ 作为横纵坐标, 进行 PR2 分析(Parity rule 2 analysis), 以检测密码子在嘌呤和嘧啶使用上的偏好^[18]. 在 PR2 分析中, 每个点代表 1 个编码基因, 若该基因第三位密码子使用 A 多于 U, 则位于图中上半部分; 若使用 G 多于 C, 则位于图中右半部分. 而若某一基因的密码子偏好性仅受到核苷酸组成这 1 个因素的影响, 那该基因的密码子第三位碱基上 G_3 和 C_3 及 A_3 和 U_3 的使用频率应该一致, 该基因应处于中间位置^[18].

1.3 密码子指数计算及 ENc 绘图分析

运用 CodonW 软件(John Peden, <http://www.molbiol.ox.ac.uk/cu>, 版本号 1.4.2), 计算 10 364 个球孢白僵菌编码基因的各项密码子指数, 如密码子适应指数(Codon Adaptation Index, CAI)、密码子偏好指数(Codon Bias Index, CBI)、最优密码子使用频率(Frequency of optimal codon, Fop)、有效密码子数(Effective Number of Codon, ENc)、同义密码子第三位 GC 含量(GC_{3s})、氨基酸疏水性指数(*Gravy*)、氨基酸芳香性指数(*Aromo*)等. 其中, 密码子适应指数 CAI 一般用于反映基因的表达情况, 其数值介于 0 到 1 间, 该值越大表明基因的偏好性越强, 表达水平也越高^[19]. 密码子偏好指数 CBI 也能衡量基因的偏好性和表达程度, 并能反映出基因中最优密码子的组成情况, 其值介于 0 到 1 间时, 数值越大基因偏好性越强; 而当 CBI 值小于 0 时, 则表示最优密码子使用频率比平均使用频率还要少, 表现出弱偏好性^[20]. 有效密码子数 ENc 是分析密码子使用偏好最具参考意义的指数, 其值通常与 CAI 和 CBI 相关, ENc 值越接近 20, 说明基因偏性越强, 表达程度越高; 越接近 61 则刚好相反^[21].

此外, 我们还以 ENc 值为纵坐标、 GC_{3s} 值为横坐标, 绘制 ENc 图(ENc-plot), 用于探测各基因的密码子使用是否只受其核苷酸组成压力的影响, 还是有其他因素在起作用^[21]. 在 ENc 图中, 曲线 $ENc = 2 +$

$GC_{3s} + 29/[GC_{3s}^2 + (1 - GC_{3s})^2]$ 表示密码子偏好性仅由核苷酸组成所决定的标准情况. 若某基因沿着或靠近标准曲线分布, 表明该基因的密码子偏好性仅仅受 GC 含量的影响; 若远离标准曲线, 则表明该基因还受到其他诸如选择压力等因素的影响^[21].

1.4 RSCU 计算及对应分析

我们使用 CodonW 软件, 计算了球孢白僵菌基因组的相对同义密码子使用值(relative synonymous codon usage, RSCU). 作为密码子偏好性研究的一个重要指数, RSCU 值是密码子实际观测值与理论观测值的比值. 当 RSCU 为 1 时, 表示实际观测值与理论值相同, 即各同义密码子使用频率相同, 密码子使用无偏好. 当 RSCU 大于 1 时, 表示该密码子使用频率高于理论值; 反之则使用频率低于理论值.

在 RSCU 的基础上, 我们将球孢白僵菌编码基因中 CAI 值居于最高、最低的 5% 样本取出生成 2 个数据集, 分别作为高、低表达组, 再利用卡方检验(chi-squared test)比较 2 组对应的 RSCU 值, 若某密码子在高表达组的使用频率显著地($p\text{-value} < 0.05$)高于低表达组, 则确定其为最优密码子(optimal codon)^[22], 并将 RSCU 值大于 1 的密码子认定为偏好密码子(preferred codon), 而 RSCU 值小于 0.10 的密码子则认定为使用频率低的稀有密码子(rare codon).

基于球孢白僵菌各编码基因的 RSCU 值, 利用 CodonW 软件进行对应分析(Correspondence analysis, COA)^[23], 并比较其基因组内部的变异情况, 进而分析影响密码子偏好的因素. 对应分析是一种多元统计方法, 用于调查基因组内各基因间 RSCU 值的差异^[23]. 通过多维空间, 对应分析不仅能展示各基因的分布, 还能反映影响密码子偏好性各种因素间的主次关系, 从而探测影响密码子偏好的因素是否多元化^[13].

文中相关性分析、方差分析、显著性检验等数据统计分析及绘图由 SPSS 18.0 和 Microsoft Excel 2010 完成.

2 结 果

2.1 密码子指数分析

球孢白僵菌基因组的多项密码子特征, 包括基因组 GC 含量(GC_{genome})、编码基因 GC 含量(GC_{cds})、密码子第一、二、三位 GC 含量(P_1, P_2, P_3)、以及密码子适应指数(CAI)、密码子偏好指数(CBI)、最优密码子使用频率(Fop)、有效密码子数(ENc)等被计算(表 1). 其中, 球孢白僵菌 CAI 平均值(0.079)和 CBI 平均值(0.018)接近于 0, ENc 平均值(48.82)较大, 表明球孢白僵菌密码子使用偏好较弱, 基因表达水平偏低.

核苷酸组成是影响密码子使用偏好的 1 个重要因素, 球孢白僵菌的编码基因 GC 含量 GC_{cds} (57.5%) 比其基因组 GC 含量 GC_{genome} (51.5%) 更高(表 1), 表明编码区比非编码的序列更为稳定. 相关性分析显示球孢白僵菌的 GC_{cds} 不仅与其 P_1 ($r = 0.623, p\text{-value} < 0.01$) 和 P_2 ($r = 0.480, p\text{-value} < 0.01$) 显著相关, 还与 P_3 ($r = 0.869, p\text{-value} < 0.01$) 显著相关, 表明球孢白僵菌编码基因 GC 含量受到 3 个密码子位置的影响. 有效密码子数 ENc 和密码子 3 个位置间存在相关性, P_1 ($r = 0.343, p\text{-value} < 0.01$), P_2 ($r = 0.130, 0.01 < p\text{-value} < 0.05$), P_3 ($r = 0.812, p\text{-value} < 0.01$), 表明核苷酸组成可能是影响球孢白僵菌密码子使用偏好的因素.

表 1 球孢白僵菌基因组密码子特征

特 征	数 值	特 征	数 值
基因组大小	33.4 Mbp	密码子第一、二位 GC 含量 P_{12}	0.536 ± 0.040
编码基因数	10 364	密码子第三位 GC 含量 P_3	0.667 ± 0.111
基因组 GC 含量 GC_{genome}	0.515	密码子适应指数 CAI	0.079 ± 0.030
编码基因 GC 含量 GC_{cds}	0.575 ± 0.050	密码子偏好指数 CBI	0.018 ± 0.078
编码基因长度	$1\,487.55 \pm 1\,119.39$ bp	最优密码子使用频率 Fop	0.387 ± 0.059
密码子第一位 GC 含量 P_1	0.615 ± 0.049	有效密码子数 ENc	48.82 ± 8.25
密码子第二位 GC 含量 P_2	0.458 ± 0.058		

2.2 ENc 绘图分析

为了检验球孢白僵菌密码子使用偏好是否仅受核苷酸组成的影响, 课题组绘制了 ENc 图(图 1). 在图 1 中, 除少部分基因沿着标准曲线分布外, 大部分球孢白僵菌基因都位于标准曲线下方, 表明核苷酸组成不

是影响球孢白僵菌密码子偏好性的唯一因素,还存在其他因素的作用. 某些基因 ENc 值等于 61, 位于图 1 中最顶端, 表明它们编码氨基酸的所有 61 个密码子都在使用, 这些基因无密码子偏好性. 此外, 由于绝大多数球孢白僵菌基因的同义密码子第三位 GC 含量(GC_{3s})超过了 50%, 在 ENc 图中表现出了基因组右倾的情况(图 1).

2.3 RSCU 分析

为了调查球孢白僵菌基因组的密码子使用情况, 课题组计算了球孢白僵菌的相对同义密码子使用值 $RSCU$ (表 2).

表 2 球孢白僵菌基因组相对同义密码子使用表

氨基酸	密码子	相对同义密码子使用值	氨基酸	密码子	相对同义密码子使用值
Phe	UUU	0.97	Ser	UCU	0.50
	UUC	<u>1.03</u>		UCC	<u>2.02*</u>
Leu	UUA	0.02—		UCA	0.17
	UUG	0.22		UCG	<u>1.46*</u>
	CUU	0.38		AGU	0.12
	CUC	<u>3.55*</u>		AGC	<u>1.71*</u>
	CUA	0.13	Pro	CCU	0.44
	CUG	<u>1.70*</u>		CCC	<u>2.40*</u>
	AUU	<u>1.08*</u>		CCA	0.15
Ile	AUC	<u>1.86*</u>		CCG	<u>1.01</u>
	AUA	0.06—	Thr	ACU	0.36
Met	AUG	1.00		ACC	<u>2.22*</u>
Val	GUU	0.36		ACA	0.22
	GUC	<u>2.83*</u>		ACG	<u>1.21*</u>
	GUA	0.08—		GAA	0.32
	GUG	0.74		GAG	<u>1.68*</u>
Tyr	UAU	0.21	Cys	UGU	0.14
	UAC	<u>1.79*</u>		UGC	<u>1.86*</u>
STOP	UAA	<u>1.74</u>	Arg	CGU	0.57
	UAG	0.74		CGC	<u>4.47*</u>
	UGA	0.52		CGA	0.19
His	CAU	0.23		CGG	0.44
	CAC	<u>1.77*</u>		AGA	0.14
Gln	CAA	0.32		AGG	0.18
	CAG	<u>1.68*</u>	Gly	GGU	0.49
Asn	AAU	0.27		GGC	<u>3.20*</u>
	AAC	<u>1.73*</u>		GGA	0.14
Lys	AAA	0.20	Ala	GGG	0.17
	AAG	<u>1.80*</u>		GCU	0.46
Asp	GAU	0.29		GCC	<u>2.70*</u>
	GAC	<u>1.71*</u>		GCA	0.13
Trp	UGG	1.00		GCG	0.71

注: 最优密码子以符号 * ($p\text{-value}<0.01$)表示, 稀有密码子以符号—($RSCU<0.10$)表示, 偏好密码子($RSCU>1$)以下划线表示.

调查 $RSCU$ 大于 1 的使用频率相对高的偏好密码子(表 2, 下划线表示), 发现球孢白僵菌基因组强烈偏好核苷酸 G/C 结尾密码子(表 3, 25 个偏好密码子中有 23 个都是 G/C 结尾的), 尤其是 C 结尾的(表 3, 25 个偏好密码子中有 16 个是 C 结尾的). 卡方鉴定出的最优密码子(表 2, 符号 * 表示)也表现出同样的情况, 22 个最优密码子中有 21 个都是 G/C 结尾的、15 个是 C 结尾的(表 3), 其中 CUC, GUC, UCC, CCC, ACC, CGC, GGC, GCC 这 8 个 $RSCU$ 大于 2 的最优密码子全是 C 结尾的. 此外, 3 个稀有密码子 UUA, AUA, GUA 也被鉴定(表 2, 符号—表示), 它们以 A 结尾.

球孢白僵菌基因组 G/C 结尾密码子的偏好性表现与其基因组 GC 含量高(51.5%)有关, 尤其与第三位密码子 GC 含量 P_3 非常高(66.7%)(表 1)密切相关, 也表明了核苷酸组成确实是影响球孢白僵菌密码子使用偏好的一个重要因素.

表 3 球孢白僵菌基因组最优密码子、偏好密码子($RSCU>1$)、稀有密码子的第三位碱基统计数

密码子第三位碱基	最优密码子	偏好密码子	稀有密码子
A/U/G/C	22	25	3
A/U	1	2	3
G/C	21	23	0
A	0	1	3
U	1	1	0
G	6	7	0
C	15	16	0

2.4 PR2 绘图分析

既然球孢白僵菌的最优密码子和偏好密码子使用 C 结尾远多于 G 结尾(表 3), 为此, 利用 $PR2$ 图探讨密码子在嘌呤(A、G)和嘧啶(U、C)使用上的偏好性(图 2). 如图 2 所示, 大多数基因位于图像的左下方. 计算球孢白僵菌的 GC 偏量 $[G_3/(G_3+C_3)]$ 和 AU 偏量 $[A_3/(A_3+U_3)]$ 分别为 0.436 和 0.439, 表明密码子第三位偏好 U 高于 A, C 高于 G, 即嘧啶高于嘌呤.

球孢白僵菌密码子中嘌呤和嘧啶的使用频率具有不对称性, 表明除核苷酸组成外还有其他因素影响其密码子使用偏好, 这与 ENc 图(图 1)的结果相呼应. 此外, 个别基因位于 $PR2$ 图的横坐标轴上(图 2), 它们被认为不使用 A 结尾密码子.

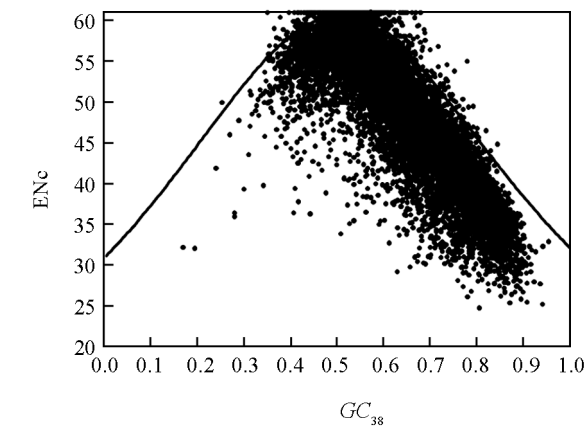


图 1 球孢白僵菌基因组的 ENc 图

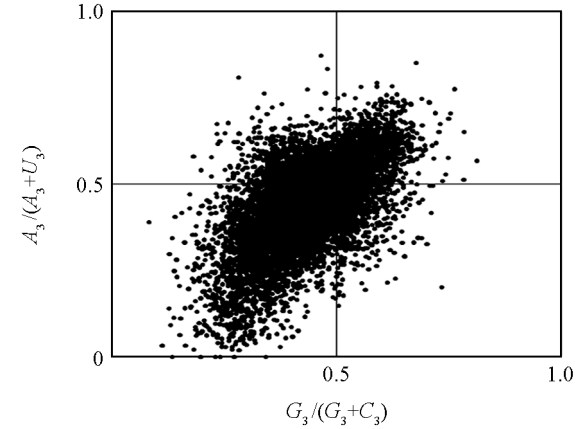


图 2 球孢白僵菌基因组的 $PR2$ 图

图中每一个点代表一个基因, 标准曲线公式为 $ENc = 2 + GC_{3s} + 29/[GC_{3s}^2 + (1 - GC_{3s})^2]$, 大部分基因不位于标准曲线上, 表明球孢白僵菌基因组密码子除受核苷酸组成影响外, 还有其他多种因素起作用.

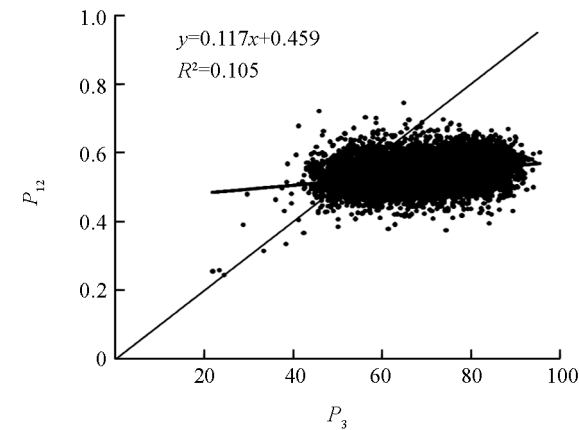
图中每一个点代表一个基因, 以 GC 偏量 $[G_3/(G_3+C_3)]$ 为横坐标, 以 AU 偏量 $[A_3/(A_3+U_3)]$ 为纵坐标. 球孢白僵菌基因组的 GC 偏量和 AU 偏量的平均值分别为 0.436 和 0.439.

2.5 中性绘图分析

既然球孢白僵菌密码子偏好性不只受到核苷酸组成的作用, 因此通过中性绘图分析(图 3), 检测突变选择压力是否为影响密码子使用偏好的一个重要因素. 如中性分析图所示(图 3), 球孢白僵菌基因组的斜率为 0.117, 趋近于 0, 其 P_{12} 与 P_3 相关系数为 0.105, 双尾检验未达到显著水平($p\text{-value}>0.05$), 表明 P_{12} 与 P_3 不相关, 认为球孢白僵菌密码子偏好性受到外界自然选择下的定向突变压力作用多于自身遗传漂变下的非定向压力作用.

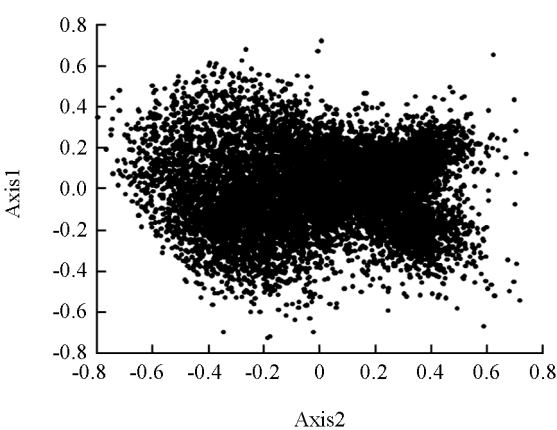
2.6 对应分析

除上述核苷酸组成和突变选择压力外, 课题组利用对应分析检测球孢白僵菌密码子偏好性是否还受到其他因素的作用, 以及这些因素的主次关系. 除对应分析图所展示的第一、二轴外(图 4), 球孢白僵菌的前四轴分别占据了总变异度的 25.13%, 11.46%, 4.10%, 3.03%, 表明球孢白僵菌密码子偏好性以第一轴(约占 1/4)和第二轴(约占 1/10)为主. 因此, 将前 4 轴与各项密码子指数进行相关性分析(表 4), 发现第一轴与代表核苷酸组成的 GC_{cds} ($r=-0.666$, $p\text{-values}<0.01$) 和 GC_{3s} ($r=-0.915$, $p\text{-values}<0.01$), 以及代表基因功能的氨基酸疏水性指数 Gravy($r=-0.245$, $0.01<p\text{-values}<0.05$) 和芳香性指数 Aromo($r=-0.203$, $0.01<p\text{-values}<0.05$) 均显著相关, 表明核苷酸组成和基因功能是影响球孢白僵菌密码子偏好性的最主要因素. 同时, 第二轴与 CAI , CBI , Fop , ENc 等密码子指数显著相关($p\text{-values}<0.01$), 表明也存在着其他因素对球孢白僵菌密码子使用偏好的影响. 这与上述分析结果一致.



图中每一个点代表一个基因, 以密码子第三位 GC 含量 P_3 为横坐标, 以密码子第一、二位 GC 含量 P_{12} 为纵坐标.

图 3 球孢白僵菌基因组的中性分析图



图中每一个点代表一个基因, 以第二轴(Axis 2)为横坐标, 以第一轴(Axis 1)为纵坐标.

图 4 球孢白僵菌基因组的对应分析图

表 4 球孢白僵菌基因组密码子指数与对应分析 4 个轴的相关性分析

	第一轴	第二轴	第三轴	第四轴
编码基因 GC 含量 GC_{cds}	-0.666 *	-0.359 *	0.070	0.109
同义密码子第三位 GC 含量 GC_{3s}	-0.915 *	-0.354 *	-0.078	0.004
编码基因长度	0.152	0.062	0.086	-0.052
密码子适应指数 CAI	-0.099	0.760 *	-0.277 @	-0.144
密码子偏好指数 CBI	-0.036	0.689 *	-0.375 *	-0.071
最优密码子使用频率 Fop	-0.153	0.668 *	-0.307 *	-0.121
有效密码子数 ENc	0.081	0.848 *	0.060	-0.053
氨基酸疏水性指数	-0.245 @	-0.086	-0.136	-0.041
氨基酸芳香性指数	-0.203 @	0.065	0.017	-0.011

注: 符号 * ($p\text{-value}<0.01$) 和 @ ($0.01<p\text{-value}<0.05$) 表示相关性显著水平.

3 讨 论

密码子使用偏好是各物种在长期进化过程中形成的特征,是基因组及进化模式的特有反映^[24],其相关研究已在诸多生物体中被报道.利用已公布的球孢白僵菌基因组及 10 364 条编码基因序列,课题组开展了球孢白僵菌密码子偏好性研究.通过密码子指数 CAI , CBI , ENc 的计算,发现球孢白僵菌密码子使用偏性较弱,基因表达水平也较低.尽管如此,根据 $RSCU$ 值鉴定最优密码子,发现球孢白僵菌基因组使用 G/C 结尾密码子的频率远远大于 A/U 结尾的,几乎没有 A/U 结尾最优密码子的存在(表 2 和表 3).球孢白僵菌这种 G/C 偏好模式与其基因组的高 G/C 含量一致,尤其与其第三位密码子 G/C 含量(P_3)高达 66.66% 有关(表 1).大部分物种都表现出这种模式,即基因组 G/C 含量高的物种倾向于使用 G/C 丰富的密码子,而 A/U 高的偏好使用 A/U 丰富的密码子^[25].因此,影响球孢白僵菌密码子偏好性的一个重要因素就是其核苷酸组成,而球孢白僵菌对应分析(图 4)及相关性分析(表 4)结果也支持了该结论.

对密码子第三位的碱基进行 $PR2$ 分析,发现球孢白僵菌的偏好性还表现在倾向于使用嘧啶结尾的密码子,即 U 使用频率高于 A, C 使用频率高于 G(图 2).因此,尽管核苷酸组成被认定为影响球孢白僵菌密码子的一个重要因素,但其密码子在嘌呤和嘧啶上的不对称性表明了还存在着其他因素的作用^[18].而 ENc 图(图 1)和对应分析(表 4)的结果也支持了该结论.

由于 Necsculea 等^[26]认为突变选择压力是引起密码子不对称性的最大因素,为此课题组对球孢白僵菌基因组进行了中性绘图分析,结果表明球孢白僵菌密码子受到外界自然选择下的定向突变压力作用(图 3),这与一类营寄生生活的真菌是一样的,它们的密码子也受到强烈的定向突变压力作用^[27].因此,认为球孢白僵菌密码子受到的定向突变压力可能与其寄生生活有关,压力来自于其宿主.

为了维持基因稳定,大多数的序列突变都是同义突变.然而,在外界选择的定向突变压力下, P_{12} 与 P_3 不相关、非同义突变增多、基因变化加快、毒力基因多样化,致使球孢白僵菌可以更快地调节自己,以适应其宿主范围广泛的寄生生活方式.表 4 中第一轴与代表基因功能的氨基酸疏水性指数 Gravy 和芳香性指数 Aromo 的相关性,揭示了影响球孢白僵菌密码子使用偏好的另一个主要因素是基因功能,这从一个侧面印证了定向突变压力下的基因功能变化确实是其基因组的一个重要特征.此外,球孢白僵菌基因组中存在着大量的转座元件^[5],可能也是在定向突变压力下加快自身基因组进化和遗传变异的一种方式,以适应其在 700 多种昆虫宿主中的寄生.当然,更可靠的结论需要通过更多的关于球孢白僵菌及近源物种基因组的诸如 Ka/Ks 计算、正选择检验、系统进化树比较等大量分子进化研究来完成.

参考文献:

- [1] GOETTEL M S, POPRAWski T J, VANDENBERG J D, et al. Safety to Nontarget Invertebrates of Fungal Biocontrol Agents [J]. Safety of Microbial Insecticides, 1989, 564(2): 209—231.
- [2] Thomas M B, Read A . Can Fungal Biopesticides Control Malaria [J]. Nature Reviews Microbiology, 2007, 5(5): 377—83.
- [3] 季香云, 杨长举. 白僵菌的致病性与应用 [J]. 中国生物防治学报, 2003, 19(2): 82—85.
- [4] 李春香, 张淑红. 白僵菌对害虫致病性的研究进展 [J]. 唐山师范学院学报, 2005, 27(5): 40—43.
- [5] XIAO G, YING S H, ZHENG P, et al. Genomic Perspectives on the Evolution of Fungal Entomopathogenicity in *Beauveria bassiana* [J]. Scientific Reports, 2012, 2(7): 483—483.

- [6] GRANTHAM R, GAUTIER C, GOUY M. Codon Frequencies in 119 Individual Genes Confirm Consistent Choices of Degenerate Bases According to Genome Type [J]. *Nucleic Acids Res*, 1980, 8(9): 1893—1912.
- [7] OSAWA S, OHAMA T, YAMAO F, et al. Directional Mutation Pressure and Transfer RNA in Choice of the Third Nucleotide of Synonymous Two-Codon Sets [J]. *Proc Natl Acad Sci USA*, 1988, 85(4): 1124—1128.
- [8] SHARP P M, LI W H. Codon Usage in Regulatory Genes in *Escherichia coli* Does Not Reflect Selection for ‘rare’ Codons [J]. *Nucleic Acids Res*, 1986, 14(19): 7737—7749.
- [9] IKEMURA T. 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]. *J Mol Biol*, 1981, 151(3): 389—409.
- [10] CHIAPELLO H, AI E. Codon Usage and Gene Function are Related in Sequences of *Arabidopsis thaliana* [J]. *Gene*, 1998, 209(1): 1—38.
- [11] OREŠIĆ M, Shalloway D. Specific Correlations Between Relative Synonymous Codon Usage and Protein Secondary Structure [J]. *J Mol Biol*, 1998, 281(1): 31—48.
- [12] MORIYAMA E N, POWELL J R. Gene Length and Codon Usage Bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli* [J]. *Nucleic Acids Res*, 1998, 26(13): 3188—3193.
- [13] ROMERO H, ZAVALA A, MUSTO H. Codon Usage in Chlamydia Trachomatis is the Result of Strand-Specific Mutational Biases and a Complex Pattern of Selective Forces [J]. *Nucleic Acids Res*, 2000, 28(10): 2084—2090.
- [14] SAU K, DEB A. Temperature Influences Synonymous Codon and Amino Acid Usage Biases in the Phages Infecting Extremely Thermophilic Prokaryotes [J]. *In Silico Biology*, 2009, 9(1): 1—9.
- [15] MORTON B R. The Role of Context-Dependent Mutations in Generating Compositional and Codon Usage Bias in Grass Chloroplast DNA [J]. *J Mol Evol*, 2003, 56(5): 616—629.
- [16] SUEOKA N. Directional Mutation Pressure and Neutral Molecular Evolution [J]. *Proc Natl Acad Sci USA*, 1988, 85(8): 2653—2657.
- [17] SUEOKA N, Kawanishi Y. DNA G+C Content of the Third Codon Position and Codon Usage Biases of Human Genes [J]. *Gene*, 2000, 261(1): 53—62.
- [18] SUEOKA N. Translation-Coupled Violation of Parity Rule 2 in Human Genes is Not the Cause of Heterogeneity of the DNA G+C Content of Third Codon Position [J]. *Gene*, 1999, 238(1): 53—58.
- [19] 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.
- [20] MORTON B R. Chloroplast DNA Codon Use: Evidence for Selection at the psb A Locus Based on tRNA Availability [J]. *J Mol Evol*, 1993, 37(3): 273—280.
- [21] WRIGHT F. The ‘Effective Number of Codons’ Used in a Gene [J]. *Gene*, 1990, 87(1): 23—29.
- [22] LIU Q. Analysis of Codon Usage Pattern in the Radioresistant Bacterium *Deinococcus radiodurans* [J]. *Biosystems*, 2006, 85(2): 99—106.
- [23] GREENACRE M J. Theory and Applications of Correspondence Analysis [J]. London: Academic Press, 1984.
- [24] SHARP P M, COWE E, HIGGINS D G, et al. Codon Usage Patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a Review of the Considerable Within-Species Diversity [J]. *Nucleic Acids Res*, 1988, 16(17): 8207—8211.
- [25] HERSHBERG R, PETROV D A. General Rules for Optimal Codon Choice [J]. *PLoS Genetics*, 2009, 5(7): 50—55.
- [26] NECSULEA A, LOBRY J R. A New Method for Assessing the Effect of Replication on DNA Base Composition Asym-

- metry [J]. Molecular Biology & Evolution, 2007, 24(10): 2169–2179.
- [27] XIANG H, ZHANG R Z, BUTLER R R, et al. Comparative Analysis of Codon Usage Bias Patterns in Microsporidian Genomes [J]. Plos One, 2015, 10(6): e0129223.

Codon Usage Bias (CUB) Pattern and Its Influencing Factors for the Whole Genome of *Beauveria bassiana*

ZHANG Rui-zhi¹, CHEN Zhang-bao¹, CHEN De-min²,
QU Zheng², XIANG Heng²

1. School of Pharmaceutical Sciences, Southwest University, Chongqing 400715, China;

2. School of Animal Science and Technology, Southwest University, Chongqing 400715, China

Abstract: An extremely wide host range of *Beauveria bassiana* makes this fungus one of the most successful commercial fungal entomopathogens for worldwide biological pest control. In a study reported herein, the codon usage bias (CUB) for the whole genome of *B. bassiana* was analyzed based on the data available so as to understand its genomic features and host adaptation. Although the presence of a weak CUB of *B. bassiana* genome was revealed based on an analysis of codon indices (CAI, CBI and ENc), nearly all of the optimal and the preferred codons in *B. bassiana* genome were found of G/C at the third codon position, especially C. *Parity* Rule 2 (PR2) plot analysis of the third codon position of the codon showed that pyrimidines (U and C) were used more frequently than purines (A and G). Correlation analysis of this CUB pattern indicated that nucleotide composition and gene function were two main factors accounting for *B. bassiana* CUB, and the ENc-plot and the correspondence analysis suggested that in addition to nucleotide composition, *B. bassiana* CUB was influenced by gene function and a variety of other factors as well. Neutrality analysis revealed that the directed mutational pressure under natural selection rather than the non-directed mutational pressure under random genetic drift played a more active role in the *B. bassiana* genome, suggesting that it might be a key motivation of adapting to over 700 species of insect hosts for *B. bassiana*.

Key words: *Beauveria bassiana*; codon usage bias; directed mutational pressure; nucleotide composition

责任编辑 夏 娟

