不同生活习性下原核生物基因组大小与GC含量的关系研究*

林 瀚，黄亚志，张尚宏
（中山大学基因工程教育部重点实验室//生物工程研究中心，广东 广州 510275）

摘 要：对原核生物基因组的研究显示，基因组 GC 含量与基因组大小和氧气偏好性等因素存在着一定的相关性。为了探索基因组大小与 GC 含量的相关性是否受原核生物生活习性的影响，选取了有代表性的 411 种原核生物（包括古细菌与真细菌），分别从最适生长温度、氧气偏好性、运动特性、水生特性和寄生特性等因素进行分析，发现基因组大小与 GC 含量的相关性确实受这些因素的影响，而寄生原核生物中显示出最好的相关性。

关键词：原核生物；基因组大小；GC 含量；生活习性；相关/回归分析
中图分类号：Q754 **文献标志码**：A **文章编号**：0529－6579（2011）03－0090－04

Correlation Between Genome Size and GC Content in Prokaryotes with Different Lifestyles

LIN Han, *HUANG Yazhi*, *ZHANG Shanghong*
(The Key Laboratory of Gene Engineering of Ministry of Education//Biotechnology Research Center, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Researches of prokaryotic genomes have led to the discovery of the correlation of genomic GC content with genome size and aerobics/anaerobics. In order to study the influence of lifestyle of prokaryotes on the correlation between genome size and genomic GC content, we analyzed the effect of optimal growth temperature, aerobics/anaerobics, motility, aquatics, and parasitism, using data from 411 representative prokaryotic species (including archaea and bacteria) and correlation/regression approaches. The results show that the correlation between genome size and genomic GC content is indeed influenced by these factors. Moreover, it was found that the correlation is the most significant in parasitic prokaryotes.

Key words: prokaryote; genome size; GC contents; lifestyle; correlation/regression analysis

基因组 GC 含量（G 与 C 所占的百分比）是基因组组成的标志性指标。早在 20 世纪 50 年代，Lee 等[1]就发现细菌基因组 GC 含量可在 25% ~ 75% 之间。迄今，有两种观点来解释不同生物之间 GC 含量的差异：中性说[2]和选择说[3]。中性说主要强调不同生物之间 GC 含量的差异是由碱基的随机突变和漂移造成，而选择说则认为 GC 含量的差异是环境及生物的生活习性等因素综合作用的结果。

在选择说的模式下，Thiery 等[4]分析了一些脊椎动物的基因组，发现温血脊椎动物的基因组 GC 含量要比冷血脊椎动物的高。Galtier 等[5]和 Hurst 等[6]则研究了原核基因组 GC 含量与生物最适生长温度的相关性，发现总体上相关关系并不明显，而一些 RNA（如 16S rRNA）的 GC 含量却与相应细菌的最适生长温度有较好的相关性。Musto 等[7]进

* 收稿日期：2010－08－05
基金项目：国家自然科学基金资助项目（30270752）；广东省自然科学基金资助项目（031616）；国家 973 计划资助项目（2005CB724600）；国家人才基金（基地）项目（J0730638）
作者简介：林瀚（1985 年生），男，硕士研究生；通讯作者：张尚宏；E-mail：lsszsh@mail.sysu.edu.cn

一步采用按科分组分析以降低其他因素干扰的方法研究这种关系，结果也只发现了一些局部的规律。另一方面，近年的研究显示，原核生物 GC 含量与其基因组大小却有一定的总体相关性[8]。

由于基因组的 GC 含量可能要受到生物生存环境及生活习性的影响，所以从总体上去分析 GC 含量与基因组大小的相关性并不一定能反映两者之间的确切关系。为了找到更精确的规律，本文采用单因素和双因素分组分析的方法研究基因组大小与 GC 含量的关系，探索在不同环境和生活习性下两者的相关性。

1　材料和方法

1.1　研究材料

本研究一共选取了有代表性的 411 种原核生物基因组进行分析，包括 371 种真细菌和 40 种古细菌[9]。它们的全序列以及生存环境的数据均从 NCBI 基因组数据库（http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi）中获得。

1.2　相关分析和回归分析

首先对所选取的原核生物样本的基因组大小与 GC 含量进行回归分析和计算决定系数 R^2（相关系数 r 的平方），验证相关性的存在。然后，将这些原核生物按照最适生长温度、氧气偏好性、运动特性、水生特性和寄生特性的不同进行单因素分组和双因素分组[9]，对每一组原核生物的基因组大小与 GC 含量进行同样的分析。

进一步，从分组的回归分析结果中选出线性相关关系最好的一组，进行二核苷酸和三核苷酸频率（数据源自我们以往的研究[10]）与基因组大小的相关分析，以验证在基因组大小与 GC 含量相关关系存在的情况下，是否同时也有寡聚核苷酸频率上的偏好性。

2　结　果

2.1　原核生物基因组大小与 GC 含量的总体相关性

所分析的原核生物基因组大小大部分都在 1～6 Mb 范围内，而 GC 含量则一般在 20%～75% 之间（图 1）。回归分析显示，基因组大小与 GC 含量总体上存在着具统计学意义的正相关（有关参数见图 1）。

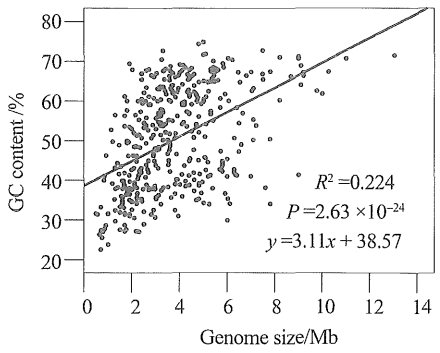

图 1　原核生物基因组大小与基因组 GC 含量的回归分析（样本数 n=411）

Fig. 1　Regression analysis of genome size versus genomic GC content of prokaryotes (sample size n=411)

(R^2: coefficient of determination; x: genome size; y: GC content)

2.2　不同组别原核生物基因组大小与 GC 含量的回归分析

从图 2a 可以看出，中温原核生物基因组大小与 GC 含量呈现较显著的正相关关系；而嗜热原核生物的这种关系则较差，且它们的基因组较小（图 2b）。好氧原核生物也显示出一定的正相关关系（图 2c）；厌氧原核生物却没有什么明显的规律（图 2d）；兼性厌氧原核生物则有较好的正相关关系（图 2e）。不运动原核生物的正相关关系比较显著（图 2f），而运动原核生物的则没那么明显（图 2g）。水生原核生物基因组大小与 GC 含量没有明显的关系（图 2h）；寄生原核生物则呈现明显的正相关关系（图 2i）；兼性寄生原核生物也呈现一定的关系（图 2j）。

从以上组别的原核生物基因组大小与 GC 含量的相关系数看，寄生原核生物那一组的最大，其回归直线斜率也最大。因此，寄生生活习性对维持或增强基因组大小与基因组 GC 含量的相关性有较大的作用。

2.3　双因素分组情况下基因组大小与 GC 含量的相关分析

从表 1 可以看出，除了厌氧寄生这一组外，其余和寄生有关的组别的基因组大小与 GC 含量正相关关系都非常好，r 值都在 0.65 以上，这说明寄生的生活方式可能存在某些因素能够维持或促进基因组大小与 GC 含量呈现较好的相关关系。另一方面，所有和厌氧有关的组别的 r 值都在 0.30 以下，P 值也比较大，说明这些组别的相关关系都比较差。这可能是厌氧的生活方式存在某些阻碍基因组大小与 GC 含量呈正相关关系的因素，且这些因素的作用比寄生的正向作用还要强。此外，除厌氧不运动组和水生不运动组外，其余和不运动相关的组别的基因组大小与 GC 含量正相关关系都比较好。

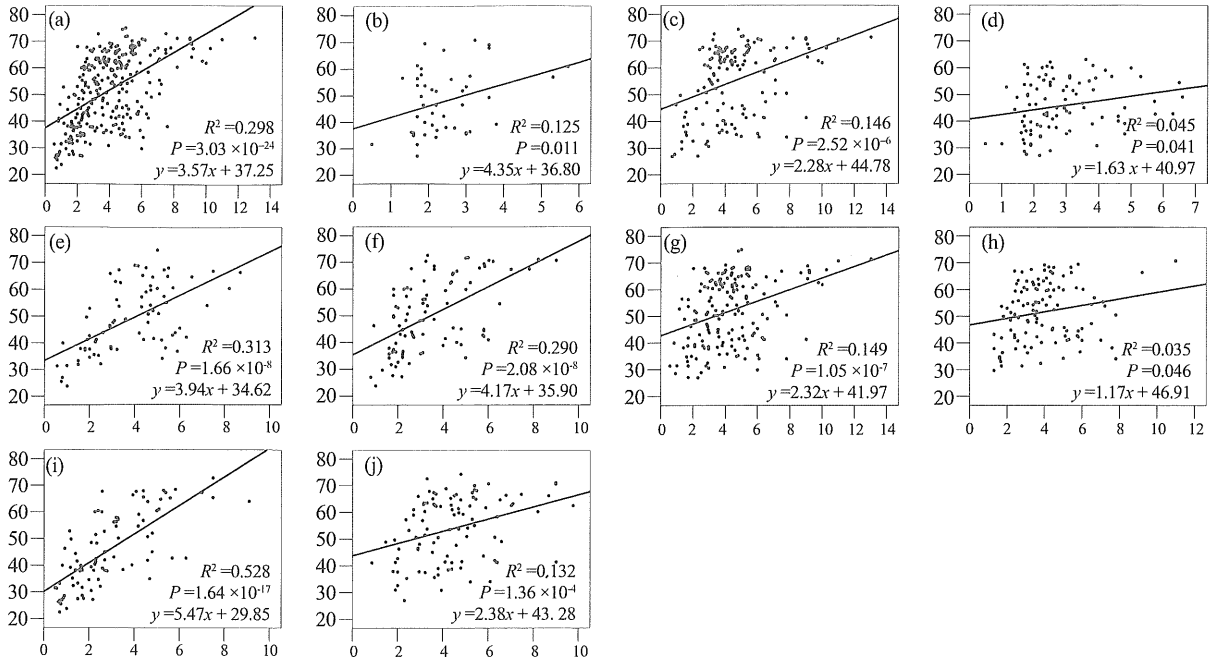

图2 不同组别原核生物基因组大小与基因组GC含量的回归分析

Fig. 2 Regression analysis of genome size versus genomic GC content in different prokaryote groups

[Abscissa: genome size (Mb); ordinate: GC content (%). For other explanations see legend of Fig. 1 (a): mesophilic group, n=295; (b): hyperthermophilic group, n=51; (c): aerobic group, n=146; (d): anaerobic group, n=97; (e): facultative group, n=87; (f): immotile group, n=94; (g): motile group, n=181; (h): aquatic group, n=113; (i): host-associated group, n=99; (j): multiple group, n=106]

表1 双重生活习性因素分组情况下基因组大小与GC含量的相关分析

Table 1 Correlation between genome size and genomic GC content in various groups classified by combinations of lifestyles

Group	r	P	Group	r	P
Aerobic-Aquatic	0.159	0.326	Facultative-Host-associated	0.810	2.60×10^{-5}
Aerobic-Host-associated	0.658	2.59×10^{-4}	Facultative-Multiple	0.503	1.28×10^{-5}
Aerobic-Multiple	0.220	0.161	Facultative-Motile	0.391	9.58×10^{-3}
Aerobic-Immotile	0.515	4.22×10^{-3}	Aquatic-Immotile	0.279	0.248
Aerobic-Motile	0.334	2.79×10^{-3}	Aquatic-Motile	0.300	0.064
Anaerobic-Aquatic	0.182	0.429	Host-associated-Immotile	0.648	6.12×10^{-4}
Anaerobic-Host-associated	0.282	0.172	Host-associated-Motile	0.665	1.54×10^{-4}
Anaerobic-Immotile	0.259	0.222	Multiple-Immotile	0.576	0.012
Anaerobic-Motile	0.234	0.249	Multiple-Motile	0.189	0.168

2.4 寄生原核生物基因组寡聚核苷酸频率与基因组大小的相关/回归分析

2.4.1 二核苷酸频率与基因组大小的相关/回归分析 从表2可以看出，全由强核苷酸（C或G）或全由弱核苷酸（A或T）组成的二核苷酸的频率（%）与基因组大小（Mb）的相关系数和回归系数的绝对值都明显比其他二核苷酸（一强一弱核苷酸组成）的要大。这说明随着基因组大小增大，对能增加GC含量的二核苷酸的偏好性明显增加，对能减少GC含量的二核苷酸的偏好性明显减小。此外，相关系数和回归系数的值还显示出二核苷酸的链对称特征，即寡聚核苷酸的频率与其反向互补序列的频率很相近[10-11]。

2.4.2 三核苷酸频率与基因组大小的相关/回归分析 除CCC和GGG外，GC含量为100%或0%的三核苷酸频率与基因组大小的相关系数绝对值都在0.65以上，回归系数绝对值都在0.45以上，呈明显的相关关系（表3）。因此，随着基因组增大，对GC含量高于AT含量的三核苷酸的偏好性也增加，而对GC含量低于AT含量的三核苷酸的偏好

性则减小。三核苷酸的链对称特征同样可从相关系数和回归系数的值显示出来。

表 2　寄生原核生物基因组二核苷酸频率与基因组大小的相关/回归分析

Table 2　Correlation/regression analysis of genomic dinucleotide frequency versus genome size in parasitic prokaryotes

Dinucleotide	r	Regression	Dinucleotide	r	Regression
AA	-0.708**	-1.628**	GA	0.522**	0.237**
AC	0.234*	0.080*	GC	0.701**	1.432**
AG	-0.207	-0.077	GG	0.683**	0.932**
AT	-0.664**	-1.049**	GT	0.211	0.073
CA	0.084	0.034	TA	-0.690**	-1.317**
CC	0.693**	0.946**	TC	0.529**	0.238**
CG	0.737**	1.796**	TG	0.058	0.024
CT	-0.210	-0.079	TT	-0.708**	-1.643**

Regression: regression coefficient; *: $P < 0.05$; **: $P < 0.01$

表 3　寄生原核生物基因组中 GC 含量 100% 或 0% 的三核苷酸频率与基因组大小的相关/回归分析

Table 3　Correlation/regression analysis of genomic trinucleotide (GC content 100% or 0%) frequency versus genome size in parasitic prokaryotes

Trinucleotide	r	Regression	Trinucleotide	r	Regression
CCC	0.582**	0.190**	AAA	-0.688**	-0.790**
CCG	0.737**	0.561**	AAT	-0.676**	-0.547**
CGC	0.704**	0.661**	ATA	-0.650**	-0.450**
CGG	0.735**	0.557**	ATT	-0.676**	-0.550**
GCC	0.716**	0.541**	TAA	-0.679**	-0.546**
GCG	0.702**	0.657**	TAT	-0.651**	-0.451**
GGC	0.712**	0.537**	TTA	-0.681**	-0.547**
GGG	0.561**	0.184**	TTT	-0.685**	-0.796**

Regression: regression coefficient; *: $P < 0.05$; **: $P < 0.01$

3　讨　论

本研究分析了各种类型原核生物的基因组。嗜热原核生物基因组大小与 GC 含量的相关性比较差，说明高温可能对其有较大的负面影响，原因可能是基因组的增大受到生存温度的制约。氧气的偏好性同样对这种相关性有影响，好氧原核生物的相关性不如兼性厌氧的，这可能是由于好氧的生活习性导致基因组 GC 含量偏高造成[12]；厌氧原核生物中较差的关系则可能是由伴随厌氧生活习性的多种因素导致。

寄生原核生物的基因组大小与 GC 含量有着最为明显的正相关关系；在对其基因组二、三核苷酸频率的分析中，同样显示出这种规律。寄生原核生物具有如此好的相关关系可能是由于某些与寄生生存方式相关的因素具促进作用而造成。另一方面，也可能是这种相关关系在生物基因组起源时就存在，寄生的生活方式导致寄生原核生物与“世”隔绝而少受外界环境因素影响，使这种相关关系维持得比较好。因此，原始生物基因组中是否就存在基因组大小与 GC 含量这种相关性值得进一步探索。总体而言，GC 含量作为基因组的基本指标，它与普遍存在于基因组序列中的链对称一起，蕴含着基因组起源与进化的重要信息[10-11]。

参考文献：

[1] LEE K Y, WAHL R, BARBU E. Contenu en bases puriques et pyrimidiques des acides desoxyribonucleiques des bacteries [J]. Ann Inst Pasteur, 1956, 91: 212-224.

[2] SUEOKA N. On the genetic basis of variation and heterogeneity of DNA base composition [J]. Proc Natl Acad Sci USA, 1962, 48: 582-592.

[3] BERNARDI G. Compositional constraints and genome evolution [J]. J Mol Evol, 1986, 24: 1-11.

[4] THIERY J P, MACAYA G, BERNARDI G. An analysis of eukaryotic genomes by density gradient centrifugation [J]. J Mol Biol, 1976, 108: 219-235.

[5] GALTIER N, LOBRY J R. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes [J]. J Mol Evol, 1997, 44: 632-636.

[6] HURST L D, MERCHANT A R. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes [J]. Proc R Soc Lond B Biol Sci, 2001, 268: 493-497.

[7] MUSTO H, NAYA H, ZAVALA A. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes [J]. FEBS Letters, 2004, 573: 73-77.

[8] MUSTO H, NAYA H, ZAVALA A. Genomic GC level, optimal growth temperature, and genome size in prokaryotes [J]. Biochem Biophys Res Commun, 2006, 347: 1-3.

[9] 林瀚. 原核生物基因组大小和 GC 含量相关性研究 [D]. 广州：中山大学, 2008: 25-35.

[10] ZHANG S H, HUANG Y Z. Characteristics of oligonucleotide frequencies across genomes: conservation versus variation, strand symmetry, and evolutionary implications[J/OL]. Nature Proceedings, 2008. [http://hdl.handle.net/10101/npre.2008.2146.1].

[11] ZHANG S H, HUANG Y Z. Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA [J]. Bioinformatics, 2010, 26: 478-485.

[12] NAYA H H, ROMERO A, ZAVALA B. Aerobics increases the genomic guanine plus cytosine content (GC%) in prokaryotes [J]. J Mol Biol, 2002, 55: 260-264.