不同国家和地区 SARS-CoV-2 序列的密码子使用偏好性及聚类分析

刘建忠，郑智捷

（云南大学云南省软件工程重点实验室，昆明 650091）

摘要：目的：研究 8 个国家和地区 SARS-CoV-2 序列的密码子使用偏好性和演化关系。方法：分别求出 8 个国家和地区 SARS-CoV-2 序列中的最大相似序列组的基准 SARS-CoV-2 序列，用 CodonW 软件分析基准 SARS-CoV-2 序列的密码子使用偏好，用 MEGA-X 软件分析各国家和地区基准 SARS-CoV-2 序列之间演化关系。结果：基准 SARS-CoV-2 序列的 G+C 含量均低于 A+U 含量，其中加拿大最低，中国次之；8 个国家和地区的共有使用偏好性密码子有 19 个，不包括中国的有 7 个，不包括加拿大的有 1 个，不包括中国和加拿大的有 1 个，这总计 28 个密码子中有 25 个以 A、U 结尾；中国有特有偏好性密码子 7 个，其中 6 个以 C、G 结尾，加拿大有 1 个，以 U 结尾；3 个发育树将基准 SARS-CoV-2 序列分成两大类，美国、印度、智利和比利时为一类，澳大利亚、加拿大、英格兰和中国为另一类，在大类中进一步划分时存在不确定性。结论：SARS-CoV-2 为了适应不同宿主在不断地发生突变，与宿主发生了协同进化，呈现出多源传染、多路演化态势。

关键词：生物工程；SARS-CoV-2；基准 SARS-CoV-2 序列；密码子使用偏好性；聚类分析

中图分类号：[Q812]　　**文献标识码**：A

文章编号：1674-2850(2021)03-0314-10　　**开放科学（资源服务）标识码（OSID）**：

Codon usage bias and cluster analysis of SARS-CoV-2 sequences in different countries and territories

LIU Jianzhong, ZHENG Jeffrey

(*Key Laboratory for Software Engineering of Yunnan Province*, *Yunnan University*, *Kunming* 650091, *China*)

Abstract: Objective: The codon usage bias and evolution of the SARS-CoV-2 sequences in eight different countries and territories were studied. Methods: The datum SARS-CoV-2 sequences of the largest similar sequence group among SARS-CoV-2 sequences in eight different countries and territories were calculated, the codon usage bias of the datum SARS-CoV-2 sequences was analyzed using CodonW package, the evolutionary relationships between the datum SARS-CoV-2 sequences were analyzed using MEGA-X. Results: G+C contents of the SARS-CoV-2 sequences were all lower than A+U contents, and the lowest measure was observed in Canada, followed by China. There were 19 commonly owned preference codons in eight different countries and territories, and 7 in seven different countries and territories (excluding China), 1 in seven different countries and territories (excluding Canada), 1 in six different countries and territories (excluding China and Canada). 25 of these 28 codons ended with A and U. There were 7 proprietary preference codons in China, 6 of which ended with C or G, and 1 in Canada that ended with U. The datum SARS-CoV-2 sequences were divided into two major categories by three phylogenetic trees, with the datum SARS-CoV-2 sequences in the USA, India, Chile and Belgium as one category and the datum SARS-CoV-2 sequences in Australia, Canada, England and China as the other category. In further

基金项目：国家自然科学基金（62041213）

作者简介：刘建忠（1955—），男，研究员，主要研究方向：反演集合理论及其应用. E-mail: liujianz6655@126.com

classification, there was uncertainty. Conclusion: In order to adapt to different hosts, SARS-CoV-2 was constantly mutating and co-evolving with the host, showing a trend of multi-source infection and multi-channel evolution.

Key words: biotechnology; SARS-CoV-2; datum SARS-CoV-2 sequence; codon usage bias; cluster analysis

0 引言

密码子使用偏好性被认为是一种演化的结果，密码子使用偏好性可揭示病毒的演化关系，提高我们对病毒 DNA 表达调控的认识，并有助于疫苗的设计[1]。为了解不同国家和地区 SARS-CoV-2 之间的特性、差别和关系，本文对 8 个国家和地区：智利、比利时、澳大利亚、加拿大、英格兰、印度、美国和中国的一些 SARS-CoV-2 序列，进行了密码子使用偏好性分析，并且分析了它们之间的演化关系。

每个国家或地区的 SARS-CoV-2 序列都很多，众多的 SARS-CoV-2 序列之间存在着差异，有些差异值还很大。因此，需要首先解决的问题是，如何找出一个国家或地区的样本序列，即在一个国家或地区众多的 SARS-CoV-2 序列中，寻找出一条最具有代表性的样本 SARS-CoV-2 序列。

一般来说，一个国家或地区的最大相似 SARS-CoV-2 序列组相对可靠。为此，首先在一个国家或地区的众多序列中，寻找出最大相似 SARS-CoV-2 序列组，然后再求出该国家或地区最大相似 SARS-CoV-2 序列组的基准 SARS-CoV-2 序列[2]，理论上可以假设认为这个最大相似 SARS-CoV-2 序列组中的所有 SARS-CoV-2 序列都是由这条初始的基准 SARS-CoV-2 序列演化而来。因此，这个基准 SARS-CoV-2 序列可代表该国家或地区的样本。在此基础上用 CodonW 软件分析该国家或地区的基准 SARS-CoV-2 序列的密码子使用偏好性，从而得出各国家和地区之间 SARS-CoV-2 序列密码子使用偏好性的差别。最后用 MEGA-X 软件分析各国家和地区基准 SARS-CoV-2 序列之间的演化关系。

1 材料与方法

1.1 数据来源

本文研究所涉及的数据均来自 https://www.gisaid.org/，共下载了 8 个国家和地区（2020 年 4 月至 5 月发布）的 SARS-CoV-2 序列：智利（130 条）、比利时（300 条）、澳大利亚（500 条）、加拿大（120 条）、英格兰（500 条）、印度（100 条）、美国（500 条）和中国（490 条）。

1.2 数据分析方法

1.2.1 提取基准 SARS-CoV-2 序列方法

分别在每个国家或地区的众多 SARS-CoV-2 序列中，用 MEGA-X 软件寻找其最大相似 SARS-CoV-2 序列组，得到各国家或地区的最大相似 SARS-CoV-2 序列组结果：智利（25 条）、比利时（100 条）、澳大利亚（74 条）、加拿大（48 条）、英格兰（458 条）、印度（63 条）、美国（181 条）、中国（198 条）。

假设各国家和地区的最大相似 SARS-CoV-2 序列组都是由一条初始的基准 SARS-CoV-2 序列[2]演变而来。为此，用滤波的方式求出各国家和地区最大相似 SARS-CoV-2 序列组的基准 SARS-CoV-2 序列。具体方法如下：

1）将 n 条相似 SARS-CoV-2 序列 Seq$i$（i=1,2,⋯,n）按序号（位点）排列对齐；

2）对某一序号 j，每条 SARS-CoV-2 序列对应值有：Seq$i$[j]（i=1,2,⋯,n）；

3）设基准 SARS-CoV-2 序列 DatumSeq[*j*]等于 Seq*i*[*j*]（*i*=1,2,⋯,*n*）中最多的值。

例如，*n*=5 时，对于 *j* 位，若 Seq1[*j*]=Seq2[*j*]=Seq3[*j*]=*T*，Seq4[*j*]=*A*，Seq5[*j*]=*C*，则 *T* 值最多，故 DatumSeq[*j*]=*T*.

本文主要研究各国家和地区的基准 SARS-CoV-2 序列之间的关系。

1.2.2 使用偏好性分析与聚类分析方法

用 CodonW 软件分析基准 SARS-CoV-2 序列的密码子使用偏好性，用 MEGA-X 软件聚类分析各国家和地区的基准 SARS-CoV-2 序列，利用生成的系统发育树分析它们之间的演化关系。

2 基准 SARS-CoV-2 序列密码子使用偏好性分析与聚类分析

2.1 各国家和地区基准 SARS-CoV-2 序列的密码子使用偏好性分析

本研究采用密码子使用偏好性参数和相对同义密码子使用度（relative synonymous codon usage，RSCU）衡量密码子使用偏好性。

密码子使用偏好性参数有：T3s、C3s、A3s、G3s（同义密码子在第 3 位上相应碱基的出现频率），CAI（密码子适应性指数），CBI（密码子偏爱指数），Fop（最优密码子的使用频率），Nc（密码子有效数），GC3s（密码子第 3 位的 G+C 含量），GC（基因的 G+C 含量），L-sym（同义氨基酸数），L-aa（氨基酸数），Gravy（平均亲水性值），Aromo（芳香性值）.

RSCU 是指某个密码子与其无偏好使用时频率之间的比值，可以衡量密码子的使用偏好程度。若 RSCU 值为 1，表明此时密码子无使用偏好性，若 RSCU 值大于 1，说明该密码子存在使用偏好性，使用频率较高，大于 1.6 表示该密码子是强使用偏好性密码子，小于 1 则相反[3]。

本文利用 CodonW 软件（https://sourceforge.net/projects/codonw/）分别分析各国家和地区基准 SARS-CoV-2 序列的密码子使用偏好性参数和 RSCU，结果如表 1、表 2 所示。

表 1 密码子使用偏好性相关参数值

Tab. 1 Codon usage bias related parameter values

国家或地区	T3s	C3s	A3s	G3s	CAI	CBI	Fop	Nc	GC3s	GC	L-sym	L-aa	Gravy	Aromo
印度	0.430 9	0.268 0	0.395 3	0.178 3	0.205	−0.042	0.400	50.67	0.350	0.396	8 801	9 180	−0.081 111	0.148 257
美国	0.430 9	0.268 0	0.395 4	0.178 0	0.205	−0.042	0.400	50.68	0.350	0.397	8 801	9 180	−0.082 462	0.148 257
比利时	0.430 8	0.267 9	0.395 8	0.178 2	0.206	−0.042	0.400	50.64	0.350	0.396	8 779	9 157	−0.081 817	0.148 411
智利	0.430 7	0.268 1	0.395 5	0.178 2	0.206	−0.042	0.400	50.67	0.350	0.396	8 797	9 175	−0.082 191	0.148 338
英格兰	0.430 8	0.267 9	0.395 8	0.178 2	0.206	−0.042	0.400	50.64	0.350	0.396	8 779	9 157	−0.081 850	0.148 411
澳大利亚	0.431 7	0.267 7	0.398 8	0.174 9	0.206	−0.043	0.399	50.42	0.347	0.396	8 661	9 038	−0.085 694	0.149 480
加拿大	0.458 4	0.198 7	0.381 5	0.213 4	0.206	−0.032	0.391	49.86	0.319	0.385	9 292	9 702	0.152 999	0.099 567
中国	0.339 3	0.239 0	0.385 6	0.300 1	0.165	−0.063	0.352	53.20	0.414	0.394	8 695	9 184	0.415 570	0.106 490

统计结果表明，基准 SARS-CoV-2 序列的 G+C 含量均低于 A+U 含量，其中加拿大的含量最低，中国次之。

从统计结果看，智利、美国、印度、比利时、英格兰 5 个国家和地区的密码子使用偏好性参数（T3s，C3s，A3s，G3s，CAI，CBI，Fop，Nc，GC3s，L-sym，L-aa，Gravy，Aromo）很相近，其次是与澳大利亚相近，与中国的差别最大。

表 2 和表 3 表明，8 个国家和地区的基准 SARS-CoV-2 序列的共有使用偏好性密码子有 19 个：UUU、

UUA、UUG、AUU、UCU、UCA、CCU、CCA、ACU、ACA、GCU、GCA、CAA、AAA、UGU、AGU、AGA、AGG 和 GGU，并且 AGA 是强使用偏好性密码子。其中，以 U 结尾 9 个，以 A 结尾 8 个，以 G 结尾 2 个。

表 2 基准 SARS-CoV-2 序列的密码子 RSCU 值

Tab. 2 Codon RSCU values of datum SARS-CoV-2 sequences

氨基酸	密码子	智利	美国	印度	比利时	英格兰	加拿大	澳大利亚	中国
Phe	UUU***	1.29	1.29	1.29	1.29	1.29	1.41	1.29	1.27
	UUC	0.71	0.71	0.71	0.71	0.71	0.59	0.71	0.73
Leu	UUA***	1.58	1.57	1.58	1.58	1.58	1.36	1.58	1.32
	UUG***	1.21	1.21	1.20	1.21	1.21	1.33	1.22	1.32
	CUU**	1.40	1.41	1.41	1.39	1.39	1.25	1.40	0.99
	CUC	0.53	0.53	0.53	0.53	0.53	0.45	0.53	0.42
	CUA	0.74	0.74	0.74	0.74	0.74	0.89	0.74	0.97
	CUG	0.54	0.54	0.54	0.54	0.54	0.72	0.54	0.97
Ile	AUU***	1.43	1.43	1.43	1.43	1.43	1.47	1.45	1.49
	AUC	0.77	0.77	0.77	0.77	0.77	0.56	0.74	0.62
	AUA	0.80	0.80	0.80	0.80	0.80	0.96	0.81	0.89
Val	GUU**	1.70	1.70	1.70	1.70	1.70	1.62	1.71	0.93
	GUC	0.74	0.74	0.74	0.74	0.74	0.53	0.74	0.40
	GUA*	0.88	0.88	0.88	0.88	0.88	0.91	0.88	1.03
	GUG*	0.68	0.68	0.68	0.68	0.68	0.94	0.66	1.65
Ser	UCU***	1.41	1.41	1.41	1.40	1.40	1.84	1.40	1.15
	UCC	0.56	0.56	0.56	0.56	0.56	0.48	0.56	0.66
	UCA***	1.28	1.28	1.28	1.28	1.28	1.60	1.31	1.60
	UCG	0.25	0.25	0.25	0.25	0.25	0.33	0.23	0.34
Pro	CCU***	1.46	1.47	1.46	1.47	1.47	1.74	1.49	1.39
	CCC	0.58	0.58	0.58	0.55	0.55	0.37	0.56	0.66
	CCA***	1.58	1.58	1.58	1.59	1.59	1.66	1.60	1.53
	CCG	0.38	0.38	0.38	0.39	0.39	0.23	0.36	0.42
Thr	ACU***	1.30	1.30	1.30	1.30	1.30	1.61	1.30	1.05
	ACC	0.88	0.88	0.88	0.88	0.88	0.56	0.89	0.81
	ACA***	1.46	1.46	1.46	1.47	1.46	1.60	1.48	1.75
	ACG	0.35	0.35	0.35	0.35	0.36	0.23	0.33	0.40
Ala	GCU***	1.80	1.80	1.80	1.80	1.80	1.99	1.80	1.34
	GCC	0.73	0.73	0.73	0.73	0.73	0.55	0.72	0.71
	GCA***	1.24	1.24	1.24	1.24	1.24	1.19	1.24	1.51
	GCG	0.23	0.23	0.23	0.23	0.23	0.28	0.24	0.45
Tyr	UAU**	1.10	1.10	1.10	1.10	1.10	1.08	1.10	0.79
	UAC*	0.90	0.90	0.90	0.90	0.90	0.92	0.90	1.21
His	CAU**	1.04	1.04	1.04	1.04	1.04	1.11	1.04	0.96
	CAC*	0.96	0.96	0.96	0.96	0.96	0.89	0.96	1.04
Gln	CAA***	1.43	1.43	1.43	1.43	1.43	1.20	1.44	1.10
	CAG	0.57	0.57	0.57	0.57	0.57	0.80	0.56	0.90

续表

氨基酸	密码子	智利	美国	印度	比利时	英格兰	加拿大	澳大利亚	中国
Asn	AAU**	1.10	1.10	1.10	1.10	1.10	1.25	1.10	0.93
	AAC*	0.90	0.90	0.90	0.90	0.90	0.75	0.90	1.07
Lys	AAA***	1.45	1.45	1.45	1.45	1.45	1.19	1.45	1.05
	AAG	0.55	0.55	0.55	0.55	0.55	0.81	0.55	0.95
Asp	GAU**	1.15	1.15	1.15	1.15	1.15	1.20	1.15	0.88
	GAC*	0.85	0.85	0.85	0.85	0.85	0.80	0.85	1.12
Glu	GAA**	1.38	1.38	1.38	1.38	1.38	1.37	1.39	0.99
	GAG*	0.62	0.62	0.62	0.62	0.62	0.63	0.61	1.01
Cys	UGU***	1.16	1.16	1.16	1.16	1.16	1.42	1.15	1.18
	UGC	0.84	0.84	0.84	0.84	0.84	0.58	0.85	0.82
Arg	CGU*	0.57	0.58	0.57	0.57	0.57	1.14	0.57	0.65
	CGC	0.43	0.43	0.43	0.43	0.43	0.49	0.44	0.36
	CGA	0.34	0.33	0.33	0.34	0.34	0.48	0.34	0.47
	CGG	0.36	0.36	0.36	0.36	0.36	0.34	0.34	0.30
Ser	AGU***	1.46	1.46	1.46	1.46	1.46	1.29	1.46	1.35
	AGC**	1.04	1.04	1.04	1.05	1.05	0.46	1.04	0.91
Arg	AGA***	2.98	2.97	2.98	2.99	2.99	2.22	3.02	2.57
	AGG***	1.32	1.32	1.33	1.31	1.31	1.33	1.28	1.65
Gly	GGU***	1.51	1.51	1.51	1.51	1.51	2.09	1.51	1.04
	GGC	0.85	0.85	0.85	0.85	0.86	0.77	0.86	0.85
	GGA**	1.13	1.14	1.13	1.13	1.13	0.85	1.14	1.23
	GGG	0.51	0.50	0.51	0.51	0.50	0.29	0.49	0.88

注：终止密码子及 Trp 和 Met（RSCU 值为 1）未列入表中；*某个国家或地区特有使用偏好性密码子，**某几个国家或地区共有使用偏好性密码子，….这个所对应的国家或地区特有的、或这几个所对应的国家和地区共有的使用偏好性密码子的 RSCU 值；*** 8 个国家和地区共有使用偏好性密码子，___这 8 个所对应的国家和地区共有使用偏好性密码子的 RSCU 值

表 3　基准 SARS-CoV-2 序列的使用偏好性密码子及其对应的氨基酸

Tab. 3　Bias codons and their corresponding amino acids of datum SARS-CoV-2 sequences

氨基酸	密码子	智利	美国	印度	比利时	英格兰	加拿大	澳大利亚	中国
Phe	UUU***	△	△	△	△	△	△	△	△
Leu	UUA***	△	△	△	△	△	△	△	△
	UUG***	△	△	△	△	△	△	△	△
	CUU**	O	O	O	O	O	O	O	
Ile	AUU***	△	△	△	△	△	△	△	△
Val	GUU**	O	O	O	O	O	O	O	
	GUA*								+
	GUG*								+
Ser	UCU***	△	△	△	△	△	△	△	△
	UCA***	△	△	△	△	△	△	△	△
Pro	CCU***	△	△	△	△	△	△	△	△
	CCA***	△	△	△	△	△	△	△	△
Thr	ACU***	△	△	△	△	△	△	△	△
	ACA***	△	△	△	△	△	△	△	△

续表

氨基酸	密码子	智利	美国	印度	比利时	英格兰	加拿大	澳大利亚	中国
Ala	GCU***	△	△	△	△	△	△	△	△
	GCA***	△	△	△	△	△	△	△	△
Tyr	UAU**	O	O	O	O	O	O	O	
	UAC*								+
His	CAU**	O	O	O	O	O	O	O	
	CAC*								+
Gln	CAA***	△	△	△	△	△	△	△	△
Asn	AAU**	O	O	O	O	O	O	O	
	AAC*								+
Lys	AAA***	△	△	△	△	△	△	△	△
Asp	GAU**	O	O	O	O	O	O	O	
	GAC*								+
Glu	GAA**	O	O	O	O	O	O	O	
	GAG*								+
Cys	UGU***	△	△	△	△	△	△	△	△
Arg	CGU*						+		
Ser	AGU***	△	△	△	△	△	△	△	△
	AGC**	O	O	O	O	O		O	
Arg	AGA***	△	△	△	△	△	△	△	△
	AGG***	△	△	△	△	△	△	△	△
Gly	GGU***	△	△	△	△	△	△	△	△
	GGA**	O	O	O	O	O		O	O

注：终止密码子及 Trp 和 Met（RSCU 值为 1）未列入表中；*某国家或地区特有使用偏好性密码子，+所对应的国家或地区；**某几个国家或地区共有使用偏好性密码子，O 所对应的这几个国家或地区；*** 8 个国家和地区共有使用偏好性密码子，△所对应的这 8 个国家和地区

其次，除中国外的 7 个国家和地区有 7 个共有使用偏好性密码子（如表 3 所示）：CUU、GUU、UAU、CAU、AAU、GAU 和 GAA（以 U 结尾 6 个，以 A 结尾 1 个），其中 GUU 是强使用偏好性密码子。

而中国有 7 个特有使用偏好性密码子（如表 3 所示）：GUA、GUG、UAC、CAC、AAC、GAC 和 GAG（以 C 结尾 4 个，以 G 结尾 2 个，以 A 结尾 1 个），其中 GUG 是强使用偏好性密码子。

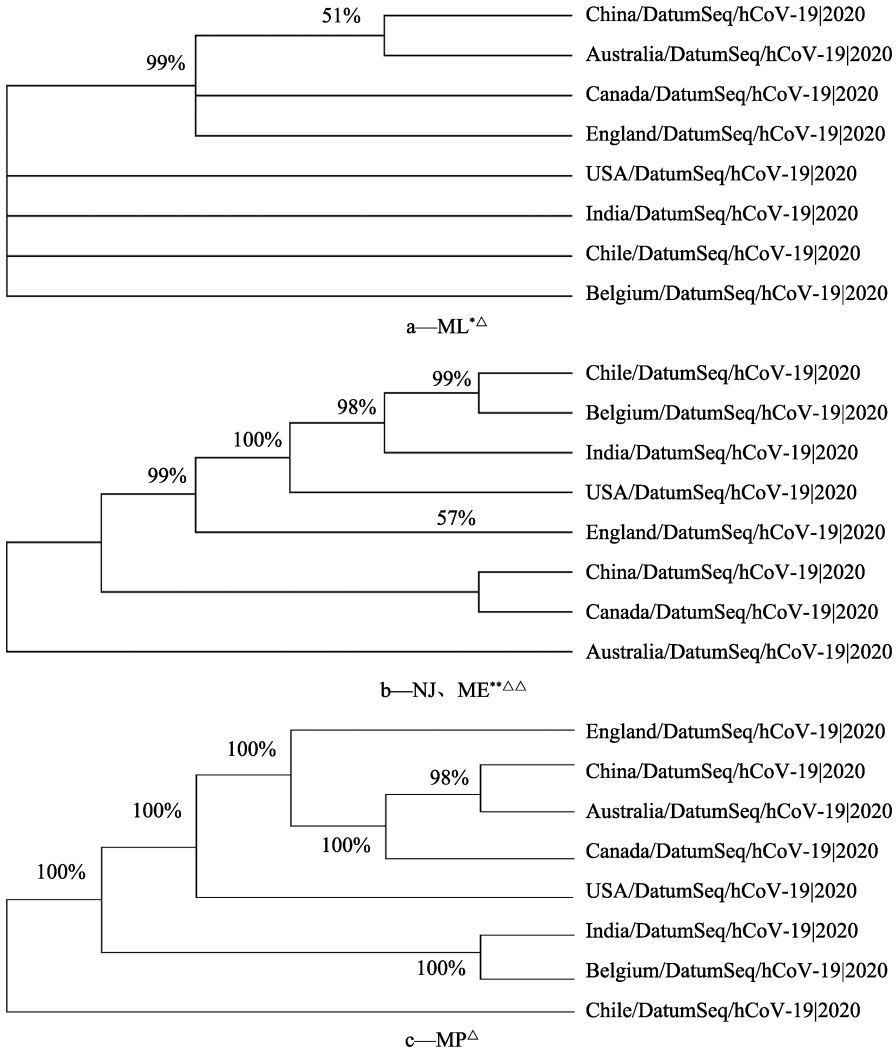
除加拿大外的 7 个国家和地区有 1 个共有使用偏好性密码子 GGA. 加拿大有 1 个特有使用偏好性密码子 CGU. 除中国和加拿大外的其他 6 个国家和地区有 1 个共有使用偏好性密码子 AGC.

2.2 各国家和地区基准 SARS-CoV-2 序列的聚类分析

本文用 MEGA-X（https://www.megasoftware.net/）软件的最大似然法（maximum likelihood，ML）、最大简约法（maximum parsimony，MP）、邻接法（neighbor-joining，NJ）和最小进化法（minimum evolution，ME）4 种方法对各国家和地区基准 SARS-CoV-2 序列之间演化关系进行系统发育树分析，共得出 3 种系统发育树（ME 和 NJ 得到相同的系统发育树）。

由图 1a 可以看出，ML 法生成的系统发育树将 8 个国家和地区的基准 SARS-CoV-2 序列分成亲缘关系不同的两类：美国、印度、智利和比利时为第一类；澳大利亚、加拿大、英格兰和中国为第二类。在

第二类中，中国与澳大利亚亲缘关系较近。图 1b 为 NJ 法和 ME 法生成的相同的系统发育树，图 1c 为 MP 法生成的系统发育树。从拓扑结构看，它们与图 1a 是等价的，均将 8 个国家或地区的基准 SARS-CoV-2 序列分成亲缘关系不同的两类。区别只是发育树的细致程度不同，它们将两类中的 8 个国家和地区之间的关系做了更加细致的描述。



注：自展值均为 2 000；*模型/方法为 GTR，**模型/方法为 *p*-距离；△空缺/丢失数据处理为使用所有位点，△△空缺/丢失数据处理为完全删除，其余参数设置默认

图 1 系统发育树

Fig. 1 Phylogenetic trees

3 结果与分析

3.1 基准 SARS-CoV-2 序列密码子使用偏好性分析

8 个国家和地区的基准 SARS-CoV-2 序列的 G+C 含量均低于 A+U 含量，其中加拿大的含量最低，中国的次之。智利、美国、印度、比利时、英格兰 5 个国家和地区的密码子使用偏好性相近，与中国的差别大。

8 个国家和地区的基准 SARS-CoV-2 序列中有 19 个共有使用偏好性密码子，但每个国家或地区的密

码子使用偏好性强度不一定相同。以强使用偏好性（RSCU 值大于 1.6）密码子 AGA 的 RSCU 值为例（如表 2 所示），中国是 2.57，澳大利亚是 3.02，而加拿大是 2.22. 再如密码子 GCU，中国是弱偏好性（1.34），其他国家和地区都是强偏好性（≥1.8）。亲缘关系与密码子使用偏好性有关，亲缘关系较近，基因的密码子使用偏好性较相似。

8 个国家和地区的 19 个共有使用偏好性密码子中以 U 结尾 9 个，A 结尾 8 个，G 结尾 2 个；除中国外的 7 个国家和地区有 7 个共有使用偏好性密码子，其中以 U 结尾 6 个，以 A 结尾 1 个。这些共有使用偏好性密码子几乎均以 A 或 U 结尾，这与流感病毒、以往的 SARS 等相关病毒的使用偏好性密码子以 A、U 结尾居多[4]的结果相似。

中国有 7 个特有使用偏好性密码子，加拿大有 1 个特有使用偏好性密码子，除中国和加拿大外的其他 6 个国家和地区有 1 个共有使用偏好性密码子，说明这 6 个国家和地区之间亲缘关系更相近。

3.2 基准 SARS-CoV-2 序列聚类分析

MEGA-X 软件的 4 种系统发育树生成方法生成的 3 个发育树（如图 1 所示）之间存在着拓扑结构的同构性，因此可以认为上述发育树具有统计学上的稳定性。

3 个发育树都将 8 个国家和地区的基准 SARS-CoV-2 序列分成两大类，但在大类中进一步分成亲缘关系不同的小类时，存在着不确定性。

例如，在图 1a 和图 1c 中，中国与澳大利亚亲缘关系更接近，而在图 1b 中中国与加拿大亲缘关系更接近；图 1b 中，比利时与智利亲缘关系更接近，而在图 1c 中比利时与印度亲缘关系更接近。由此可见，关于中国、澳大利亚和加拿大三者之间的亲缘关系，统计计算方法略有不同就可能会得出不同结果。同理，比利时、印度和智利三者之间的亲缘关系也是统计计算方法略有不同就可能会得出不同结果。

有研究发现基于密码子使用偏好性的聚类结果不一定能完全、真实地反映基因组之间的系统发育关系[5]，这提示基因组的分类不仅与碱基序列组成有关，还可能与各种水平上的综合结果以及遗传关系等有关。

3.3 基准 SARS-CoV-2 序列演化的暗示

研究显示，基准 SARS-CoV-2 序列密码子偏爱以 A、U 结尾，这与流感病毒、以往的 SARS 等相关病毒相似。追溯历史，流感病毒从 1918 年至 2009 年近一个世纪的演变，其基因的 G+C 和 GC3s 值始终小于 50%，表明流感病毒比较偏爱使用 A、U 和以其结尾的密码子。而以往的 SARS 等相关病毒的研究资料显示也是主要偏爱以 A 和 T 结尾的密码子。

密码子使用偏好性是在生物演化过程中逐步形成的，突变和翻译选择被认为是影响密码子使用偏好的主要原因[6~7]。有研究认为，在 AU(T)含量极高的基因组中，突变压力是主要影响其同义密码子使用的因素[8~9]。原因是 GC 键相对于 AU(T)键间多了一个氢键，因此 GC 键断裂所需的能量要远大于 AU(T)键，AU(T)相较于 GC 而言更易断裂而产生变动[10]。

目前普遍认为，病毒在同一种宿主间传播的时间足够长以后，为了更好地利用宿主内的资源和提高自身的生存能力，病毒的密码子使用偏好性会与宿主的密码子使用偏好性趋同，即病毒编码氨基酸时会倾向于使用该宿主编码同种氨基酸所使用的密码子[11]。病毒这种与宿主协同进化的功能，使得病毒基因能针对不同的宿主基因环境，呈现出不同的演化态势。

由此认为，在本文基准 SARS-CoV-2 序列的分析中，中国与澳大利亚、加拿大亲缘关系的不确定性，

比利时与智利、印度亲缘关系的不确定性，这种小类分化的不确定性应该与流感病毒、以往的 SARS 等相关病毒一样，都是由它们的 AU(T)键变化导致的。这说明 SARS-CoV-2 为了适应不同宿主在不断地发生突变，呈现出多路演化态势。图 1b、图 1c 呈现出 8 个国家和地区的基准 SARS-CoV-2 序列为适应不同宿主而发生的多路演化态势。

另一项研究工作发现，不同国家的 SARS-CoV-2 具有不同的拓扑熵；甚至在一些国家内，不同地区的拓扑熵也可能不同（见文献[12]中图 12）。这与本文的 SARS-CoV-2 具有多路演化态势的结论是一致的。近来相继报道的英国变异新冠病毒（B.1.17）、南非变异新冠病毒（B.1.351）和美国加利福尼亚变异新冠病毒（CAL.20C）也印证了这一点。

由图 1 可以看出，美国、印度、比利时和智利共同影响着其他 4 个国家和地区的基准 SARS-CoV-2 序列，呈现出一种处于多源传染的多路演化态势。

4 结论

研究发现，8 个国家和地区基准 SARS-CoV-2 序列的 G+C 含量均低于 A+U 含量，其中加拿大含量最低，中国次之。

智利、美国、印度、比利时、英格兰 5 个国家和地区的密码子使用偏好性参数很相近，其次是与澳大利亚相近，与中国的差别最大。

8 个国家和地区有 19 个共有使用偏好性密码子，除中国外的 7 个国家和地区有 7 个共有使用偏好性密码子，除加拿大外的 7 个国家和地区有 1 个共有使用偏好性密码子，除中国和加拿大外的 6 个国家和地区有 1 个共有使用偏好性密码子，在这总计 28 个使用偏好性密码子中有 25 个以 A、U 结尾。此外，中国有 7 个特有使用偏好性密码子，其中 6 个以 C、G 结尾；加拿大有 1 个特有使用偏好性密码子，以 U 结尾。

本文中生成的 3 个发育树将 8 个国家和地区的基准 SARS-CoV-2 序列分成亲缘关系不同的两大类：美国、印度、智利和比利时为第一类；澳大利亚、加拿大、英格兰和中国为第二类。但在大类中进一步划分亲缘关系不同的小类时，存在着不确定性，即统计计算方法略有不同就可能会得出不同结果。研究认为，这种小类分化的不确定性应该与流感病毒、以往的 SARS 等相关病毒一样，都是由于它们的 AU(T)键易断裂而产生的变动导致的。这说明 SARS-CoV-2 为了适应不同宿主在不断地发生突变，与宿主发生了协同进化，呈现出多源传染、多路演化态势。

[参考文献] (References)

[1] NASRULLAH I, BUTT A M, TAHIR S, et al. Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution[J]. BMC Evol. Biol., 2015, 15(1), DOI: 10.11861512862-015-0456-4.

[2] LIU J Z, ZHENG J. Observing variations of differences on COVID-19 in different regions extracting type and mutation information[EB/OL]. Beijing: Sciencepaper Online. http://www.paper.edu.cn/releasepaper/content/202005-51.

[3] 徐利娟，钟金城，陈智华，等. 流感病毒基因的密码子偏好性及聚类分析[J]. 生物信息学，2010，8（2）：175-179，186.
XU L J, ZHONG J C, CHEN Z H, et al. Cluster analysis and codon usage bias studies on genes of influenza virus[J]. China Journal of Bioinformatics, 2010, 8(2): 175-179, 186. (in Chinese)

[4] GU W J, ZHOU T, MA J M, et al. Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales[J]. Virus Research, 2004, 101(2): 155-161.

[5] CHRISTIANSON M L．Codon usage patterns distort phylogenies from or of DNA sequences[J]. American Journal of Botany, 2005, 92(8): 1221-1233.

[6] KARLIN S, MRAZEK J. What drives codon choices in human genes[J]. Journal of Molecular Biology, 1996, 262(4): 459-472.

[7] LESNIK T, SOLOMOVICI J, DEANA A, et a1. Ribosome traffic in *E. coli* and regulation of gene expression[J]. Journal of Theoretical Biology, 2000, 202(2): 175-185.

[8] ZHAO S, ZHANG Q, CHEN Z H, et a1. The factors shaping synonymous codon usage in the genome of *Burkholderia mallei*[J]. Journal of Genetics and Genomics, 2007, 34(4): 362-372.

[9] ZHONG J C, LI Y M, ZHAO S, et a1. Mutation pressure shapes codon usage in the GC-Rich genome of foot-and-mouth disease virus[J]. Virus Genes, 2007, 35(3): 767-776.

[10] 王文斌，于欢，邱相坡. 黄芩叶绿体基因组重复序列及密码子偏好性分析[J]. 分子植物育种，2018，16（8）：2445-2452. WANG W B, YU H, QIU X P. Analysis of repeat sequence and codon bias of chloroplast genome in scutellaria baicalensis[J]. Molecular Plant Breeding, 2018, 16(8): 2445-2452. (in Chinese)

[11] VERVOORT E B, van RAVESTEIN A, van PEIJ N N, et al. Optimizing heterologous expression in *Dictyostelium*: importance of 5’ codon adaptation[J]. Nucleic Acids Research, 2000, 28(10): 2069-2074.

[12] ZHENG J, LIU J Z. A visual framework of meta genomic analysis on variations of whole SARS-CoV-2 sequences[J]. Journal of EC Neurology, 2021(2): 49-70.

（责任编辑：段桃）