

SARS 相关病毒(BJ01 株)的全序列及其比较分析

秦鄂德 祝庆余* 于 曼 范宝昌 常国辉 司炳银 杨保安 彭文明 姜 涛
刘伯华 邓永强 刘 洪 张 雨 王翠娥 李豫川 甘永华 李晓萸 吕富双
谭 刚 曹务春 杨瑞馥

(军事医学科学院微生物学与流行病学研究所, 北京 100071. * 联系人, zhuqy@nic.bmi.ac.cn)

汪 建 李 蔚 徐祖元 李 彦 吴清发 林 伟 陈维军 唐 琳 邓亚军
韩玉军 李昌峰 雷 蒙 李国庆 李文杰 吕 宏 石建萍 童宗中 张 峰
李松岗 刘 斌 刘斯奇 董 伟 王 俊 黄家树 于 军 杨焕明*

(中国科学院北京基因组研究所/生物系统工程国家工程研究中心, 北京 101300. * 联系人, yanghm@genomics.org.cn)

摘要 严重急性呼吸综合征(SARS)是一种新发现的急性传染病. 对一株已确认为 SARS 病原体的冠状病毒(BJ01)进行了全基因组序列测定, 并与其他已知病毒进行了比较分析. 该病毒基因组全长为 29.725 kb, 含 11 个可读框(ORFs), 由 1 个稳定区和 1 个可变区组成, 其中稳定区编码 RNA 依赖性 RNA 聚合酶, 包括两个 ORFs; 可变区含有 4 个蛋白质编码序列(CDSs), 分别编码病毒的 4 个结构蛋白(S, E, M, N 蛋白). 此外, 可变区还包括 5 个非典型的预测蛋白(PUPs). 此病毒基因的排列顺序与其他已知的冠状病毒一致. 通过与已知 RNA 病毒的序列比对, 可以确认该病毒属于冠状病毒科(Coronaviridae). 与 GenBank 中已知的 5 个 SARS 相关病毒全基因组序列的比较分析显示, 已在 30 个核苷酸位置上检出序列差异, 总体突变率为 0.1%, 其中 15 个变异位点在编码区, 可能会导致蛋白质的氨基酸改变(异义突变). S 蛋白中已发现 3 处可能引起该蛋白质物化特征变化的变异, 而该蛋白质可能参与病毒与宿主间的免疫反应. 与病毒包膜形成相关的 M 蛋白中则已发现两处氨基酸变化. 进化分析表明, SARS 病毒可能不是来源于人类, 但没有证据证明是人为制造的. 为了阐明 SARS 相关病毒的病因学以及排除其他可能的 SARS 病原体的存在, 仍需进行更深入的研究.

关键词 严重急性呼吸综合征(SARS) 冠状病毒 基因组 系统发生学 突变

严重急性呼吸综合征(SARS)是一种新发现的急性传染病^[1~5], 它在全球的爆发严重威胁着人类的健康. 根据世界卫生组织(WHO) 2003 年 4 月 29 日的报告, 全球已有 27 个国家和地区总计 5462 人感染 SARS, 导致至少 353 人死亡(<http://www.who.int/csr/sarscountry/en>). 尽管 SARS 病毒是否是惟一病原体还有不同意见, 但目前人们已普遍接受, 引起 SARS 的可能病原体是来自冠状病毒科的变异株^[1~5] (<http://www.nytimes.com/2003/04/24/science/24INFN.html>; <http://www.sciencenews.org/20030329/fob7.asp>).

SARS 相关病毒的全基因组序列可以为鉴定病原体, 探索病因及其发生机制, 阐明传播途径和发病机理, 研发相关的诊断试剂、疫苗和治疗药物等提供必要的基本信息. 近日, 我们完成了一株 SARS 相关病毒的全基因组序列测定, 并与目前已公布的 5 个冠状病毒基因组进行了比较分析.

1 材料与方法

(i) 样品来源. 2003 年 2 月 10 日~3 月 15 日在广州和北京发现 10 例 SARS 疑似病例. 所有患者依据 WHO 诊断标准均确诊为 SARS(<http://www.who.int/csr/sars/guidelines/en/>), 其中两位患者由于呼吸衰竭死亡. 尸检, 取肺部组织标本. 标本用抗生素处理后, 用不同的细胞系(Vero-E6, MDCK, Hep-2, HeLa, BHK-21 和 LLC-MK-2)培养以分离病毒^[6]. 4 株冠状病毒样毒株分离自 Vero-E6 细胞培养物. 病毒制备、反转 PCR 和克隆等均按常规操作.

(ii) 测序、补洞和组装. 采用 MegaBACE 1000 (Amersham)自动测序仪进行测序, 用 Phred (<http://www.phrap.org>)软件识别碱基. 人及其他来源的序列污染去除与拼接用 CrossMatch 和 Phrap 完成(<http://www.phrap.org>). 预组装检出的空洞与低质量区以

ABI 377 测序仪(Applied Biosystems)重新测定 PCR 产物序列, 提高组装质量.

(iii) 序列注释和比较分析. 核苷酸序列的组成采用 DNA_GC_Calculator 软件分析(<http://www.genome.iastate.edu/ftp/share/DNAgcCal/>). 可读框 (ORFs) 由 ORF Finder (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>)确定. 采用 Blast 比对 nr(非冗余)数据库(<http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>)进行核酸和蛋白质序列的比对分析, 多重序列比对则用 ClustalW1.8 (<ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW>)分析. 蛋白质的糖基化位点用 NetNGlyc1.0 (<http://www.cbs.dtu.dk/services/NetNGlyc>)确定. 蛋白质的物理及化学特征用 Compute PI/MW (http://us.expasy.org/tools/pi_tool.html)预测. 以上所有生物信息学分析均在超级计算机曙光 2000, 曙光 3000 (曙光计算机公司), SUN E10K (SUN Microsystems Inc), SGI Origin 3800 (Silicon Graphics, Inc)和 IBM p690 (IBM Corp)上进行.

2 结果

总共完成了 549 个测序反应, 得到了 207892 nt 的原始数据, 相当于预计的基因组大小 7× 的覆盖率. 经过补洞和改进低质量区域之后, 全长序列为 29.725 kb, 总误差率为 0.0093%. 序列已提交 GenBank(登录号: AY278488), 所有已知序列的克隆可供全世界的合作者使用.

2.1 病毒基因组概况

分析表明, 我们测序的病毒株(BJ01)基因组具有线性正链单链 RNA 病毒的特征, 两侧分别为 5'-甲基

化帽子和 3'-poly(A)⁺尾巴结构. GC 含量为 40.8% (A : U : C : G = 28.5 : 30.7 : 20.0 : 20.8), 在基因组中基本均匀分布(图 1). 两个典型的富 GC 区分别位于 200 ~ 500 和 28200 ~ 28767 nt 处, 接近基因组的两端, 位于编码 RNA 依赖性 RNA 聚合酶和 N(核壳体)蛋白的 ORF 中, 其中 N 蛋白与 PUP5 部分重叠. 基因组中还含有一些重复序列, 可能与单链 RNA 分子的二级结构有关. 5'端序列具有长重复序列形成的典型的二级结构, 如位于 155 ~ 211 nt 的区域与 861 ~ 920 nt 的 60 个核苷酸区域的互补序列高度同源(图 2). 已发现至少 4 个反向重复结构(最小核苷酸数目为 7)(图 3). 约 140 个片段与基因组中可能的回文结构有关 (见补充数据, <http://www.genomics.org.cn/SARS>).

2.2 预测的 ORFs 及其相应蛋白质序列

从此病毒全基因组中共预测出 11 个 ORFs(6 个编码区和 5 个 PUPs), 总长 28.550 kb, 约占全基因组的 96%(表 1, 图 4). 除了 orf 1ab 外, 在 PUP1 和 PUP2, PUP2 和 E 蛋白之间有重叠的 ORFs. 因此, 除了 5' (245 nt)和 3' (356 nt)非编码区(UTRs)外, 共有 6 个基因间区域, 总长 564 nt, 其中 5 个较小(7 ~ 51 nt), 1 个较大(478 nt, 在 PUP4 和 N 蛋白之间). 目前的数据还不足以证明计算模拟出的 PUPs 的可靠性和预测的功能. ORFs 的排列顺序与其他已知 HE(-)冠状病毒完全一致. 密码子使用频度显示 Leu(亮氨酸)的频度最高 (表 2).

编码 RNA 依赖性 RNA 聚合酶的区域(orf 1ab)大约占全基因组的 2/3(246 ~ 21466 nt). 它由两个可读框(orf 1a 和 orf 1b, 246 ~ 13379 和 13379 ~ 21466

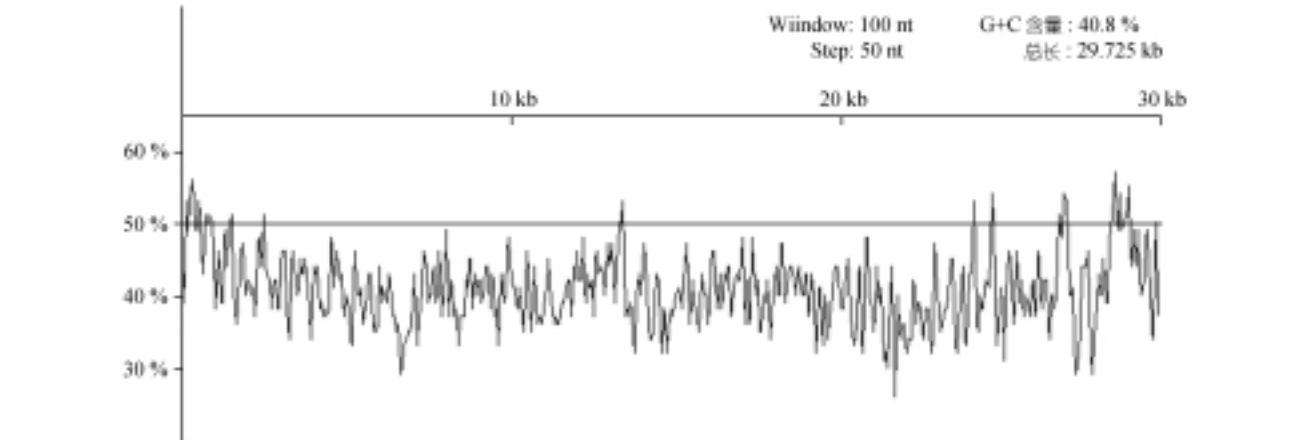


图 1 SARS 相关病毒基因组(BJ01 分离株)的 GC 含量

图 2 在 SARS 相关病毒基因组(BJ01 分离株)中发现的 5'末端重复序列

可能生成两个多肽 (表 1).

S 蛋白(spike protein, 或 E2 糖蛋白前体)是一个弱酸性糖蛋白(pI 5.5), 位置紧邻 RNA 聚合酶(21473 ~ 25240 nt). 它在所有病毒结构蛋白中最大(1255 aa), 分子量约为 139.17 kD. 具有至少 9 个糖基化位置和 1 个疏水性跨膜区.

E 蛋白(小包膜蛋白)位于 26098 ~ 26328 nt, 它可能是此病毒中最小的蛋白(8.36 kD, 76 aa), 在所有病毒结构蛋白中疏水性最强(47.40%), 净(正负)电荷平衡(5.30%). 位于所有其他编码区起始区域附近的保守片段 UCUAAAC 在 E 蛋白中以突变的形式出现于

ORF	位置(nt)	长度(aa)	分子量/kD	等电点(pI)	疏水性(%)	亲水性(%)	电荷(+) (%)	电荷(-)(%)
RNA 聚合酶	246~13379 13379~21466	7073	790.28	6.3	30.80	44.30	11.80	10.50
S 蛋白	21473~25240	1255	139.17	5.5	30.40	44.80	9.10	9.20
E 蛋白	26098~26328	76	8.36	6.0	47.40	32.90	5.30	5.30
M 蛋白	26379~27044	221	25.06	9.3	40.70	36.20	10.90	5.90
N 蛋白	28101~29369	422	46.03	10.1	17.30	54.00	15.40	8.50
PUP1	25249~26073	274	30.90	5.6	34.70	39.10	8.40	8.00
PUP2	25670~26134	154	17.72	11.0	37.00	51.90	19.50	0.60
PUP3	27055~27246	63	7.54	4.7	47.60	42.90	11.10	15.90
PUP4	27254~27622	122	13.94	8.3	33.60	42.60	13.10	8.20
PUP5	28111~28407	98	10.80	4.9	32.70	46.90	9.20	11.2

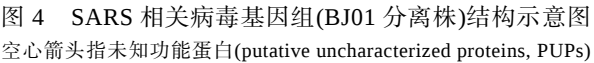


表 2 SARS 相关病毒基因组(BJ01 分离株)可读框(ORF)的密码子使用频率

RNA 聚合酶				S 蛋白				E 蛋白				M 蛋白				N 蛋白			
Leu (L)	675	9.50%		Thr (T)	100	8.00%		Val (V)	14	18.40%		Leu (L)	31	14.00%		Gly (G)	45	10.70%	
Val (V)	579	8.20%		Leu (L)	99	7.90%		Leu (L)	14	18.40%		Ala (A)	19	8.60%		Ser (S)	35	8.30%	
Ala (A)	511	7.20%		Ser (S)	96	7.60%		Ser (S)	7	9.20%		Ile (I)	18	8.10%		Gln (Q)	34	8.10%	
Thr (T)	495	7.00%		Val (V)	91	7.30%		Thr (T)	5	6.60%		Val (V)	16	7.20%		Ala (A)	34	8.10%	
Ser (S)	458	6.50%		Ala (A)	84	6.70%		Asn (N)	5	6.60%		Gly (G)	15	6.80%		Thr (T)	33	7.80%	
Lys (K)	415	5.90%		Phe (F)	83	6.60%		Tyr (Y)	4	5.30%		Arg (R)	15	6.80%		Pro (P)	31	7.30%	
Gly (G)	419	5.90%		Asn (N)	81	6.50%		Phe (F)	4	5.30%		Thr (T)	13	5.90%		Arg (R)	31	7.30%	
Asp (D)	395	5.60%		Gly (G)	78	6.20%		Ala (A)	4	5.30%		Asn (N)	13	5.90%		Lys (K)	29	6.90%	
Asn (N)	366	5.20%		Ile (I)	77	6.10%		Ile (I)	3	3.90%		Ser (S)	12	5.40%		Leu (L)	26	6.20%	
Glu (E)	348	4.90%		Asp (D)	74	5.90%		Glu (E)	3	3.90%		Phe (F)	11	5.00%		Asn (N)	25	5.90%	
Ile (I)	343	4.80%		Lys (K)	60	4.80%		Cys (C)	3	3.90%		Tyr (Y)	9	4.10%		Asp (D)	22	5.20%	
Phe (F)	331	4.70%		Pro (P)	57	4.50%		Pro (P)	2	2.60%		Trp (W)	7	3.20%		Glu (E)	14	3.30%	
Tyr (Y)	324	4.60%		Gln (Q)	55	4.40%		Lys (K)	2	2.60%		Met (M)	7	3.20%		Phe (F)	13	3.10%	
Pro (P)	274	3.90%		Tyr (Y)	54	4.30%		Gly (G)	2	2.60%		Glu (E)	7	3.20%		Val (V)	11	2.60%	
Arg (R)	259	3.70%		Glu (E)	42	3.30%		Arg (R)	2	2.60%		Lys (K)	6	2.70%		Tyr (Y)	11	2.60%	
Gln (Q)	234	3.30%		Cys (C)	39	3.10%		Met (M)	1	1.30%		Asp (D)	6	2.70%		Ile (I)	11	2.60%	
Cys (C)	233	3.30%		Arg (R)	39	3.10%		Asp (D)	1	1.30%		Pro (P)	5	2.30%		Met (M)	7	1.70%	
Met (M)	177	2.50%		Met (M)	20	1.60%		Trp (W)	0	0.00%		Gln (Q)	5	2.30%		Trp (W)	5	1.20%	
His (H)	160	2.30%		His (H)	15	1.20%		His (H)	0	0.00%		His (H)	3	1.40%		His (H)	5	1.20%	
Trp (W)	77	1.10%		Trp (W)	11	0.90%		Gln (Q)	0	0.00%		Cys (C)	3	1.40%		Cys (C)	0	0.00%	
总计	7073	100%			1255	100%			76	100%			221	100%			422	100%	

基因上游 200 碱基处.

M 蛋白(膜蛋白, 基质蛋白, 或 E1 膜糖蛋白)为中
等大小蛋白(25.06 kD, 221 aa), 等电点较高(pI 9.3),
位于 26379 ~ 27044 nt.

N 蛋白(核壳体蛋白)的大小仅次于 S 蛋白(46.03
kD, 422 aa), 位于 3'端(28101 ~ 29369 nt), 它是一个
典型的碱性蛋白, 等电点高达 10.1, 在所有预测蛋白
中亲水性最强(54.00%)而疏水性最弱(17.30%).

2.3 冠状病毒基因组的变异和系统发生学

根据发现的变异位置的不平均分布, 我们可将
病毒全基因组划分为两个区域. RNA 聚合酶的 5'稳
定区突变率相对较低(0.09%, 19/21, 220). 3'可变区含
有 4 个编码区编码病毒结构(S, E, M, N 蛋白)基因和 5
个 PUPs, 突变率较高(0.21%, 11/5, 447) (表 3). 通过
对 5 个 SARS 冠状病毒分离株基因组序列的比较分
析, 我们发现了 30 个变异位置, 累计突变率约为
0.1%. 变异中很大一部分(63.3%, 19/30)为转换. C-T

类型占 68.4%(13/19), A-G 占 31.6%(6/19). 其中有 9
个已在两条不同的序列中找到.

约一半的变异(15/30)在蛋白编码区中可能为异
义突变, 在 PUPs 中也有 2 个变异(表 4). RNA 聚合酶
中发现的 19 个变异位置中, 约一半(10)可导致氨基
酸变化. 这一比率在 S 蛋白(3/4)和 M 蛋白(2/2)中更
高, 可能导致物化特征的改变, 尚不知是否与免疫反
应相关的膜蛋白高级结构和功能有关. 未发现任何
缺失或插入突变.

与许多其他的 RNA 病毒相同, 5 个 SARS 相关病
毒基因组中, RNA 依赖性 RNA 聚合酶高度保守, 检
测出的突变率只有 0.09%, 这在已发现突变的 5 个编
码区中是最低的(表 3). 最保守的一段位于 SARS 相
关病毒的 14573 ~ 16894 nt 处, 与牛冠状病毒中的相
应区域 (AF391542, AF220295)和鼠肝炎病毒的相应
区域(AF208067, AF029248, AF208066, AF207902,
AF201929) 的同一性约为 70% (640 ~ 643/914 aa).
在所有 17 个已知冠状病毒中检测到 1 个异常保守的

表 3 5 株 SARS 相关病毒基因组的突变统计

ORF	长度(nt)	置换数	置换比(%)
RNA 聚合酶	21220	19	0.09
S 蛋白	3767	4	0.11
M 蛋白	665	2	0.30
PUP1	824	3	0.36
PUP3	191	1	0.52
Non-ORF		1	
总计	29725	30	0.10

表 4 5 株 SARS 相关病毒基因组在碱基和氨基酸水平的突变情况

ORF	核苷酸位置 (BJ01)	氨基酸位置	BJ01 ^{a)}	TOR2 ^{a)}	US ^{a)}	CUHK ^{a)}	HKU ^{a)}	相对频率 ^{b)}
RNA 聚合酶	2582	779	T (Val)	C (Val)	T (Val)	T (Val)	C (Val)	C(2)/T(3)
	7727	2494	G (Pro)	G (Pro)	G (Pro)	T (Pro)	G (Pro)	T(1)/G(4)
	7900	2552	C (Ala)	C (Ala)	T (Val)	C (Ala)	C (Ala)	T(1)/C(4)
	7911	2556	G (Asp)	G (Asp)	G (Asp)	G (Asp)	A (Asn)	A(1)/G(4)
	8368	2708	G (Ser)	G (Ser)	G (Ser)	G (Ser)	C (Thr)	C(1)/G(4)
	8398	2718	G (Arg)	G (Arg)	G (Arg)	G (Arg)	C (Thr)	C(1)/G(4)
	8553	2770	T (Leu)	G (Val)	G (Val)	G (Val)	G (Val)	T(1)/G(4)
	9385	3047	C (Ala)	T (Val)	T (Val)	C (Ala)	T (Val)	C(2)/T(3)
	9460	3072	T (Val)	T (Val)	T (Val)	C (Ala)	T (Val)	C(1)/T(4)
	9835	3197	T (Val)	C (Ala)	C (Ala)	C (Ala)	C (Ala)	T(1)/C(4)
	10568	3441	C (Thr)	A (Thr)	A (Thr)	A (Thr)	A (Thr)	C(1)/A(4)
	16603	5453	C (Ala)	C (Ala)	T (Ala)	C (Ala)	C (Ala)	T(1)/C(4)
	17545	5767	G (Glu)	T (Asp)	T (Asp)	G (Glu)	T (Asp)	G(2)/T(3)
	17827	5861	C (Arg)	C (Arg)	C (Arg)	T (Arg)	C (Arg)	T(1)/C(4)
	18046	5934	G (Lys)	G (Lys)	G (Lys)	G (Lys)	A (Lys)	A(1)/G(4)
	19045	6267	A (Glu)	A (Glu)	G (Glu)	G (Glu)	A (Glu)	G(2)/A(3)
S 蛋白	19819	6525	G (Val)	A (Val)	A (Val)	A (Val)	A (Val)	G(1)/A(4)
	20577	6778	A (Gln)	G (Arg)	A (Gln)	A (Gln)	A (Gln)	G(1)/A(4)
	20891	6883	G (Asp)	T (Tyr)	G (Asp)	G (Asp)	G (Asp)	T(1)/G(4)
	21702	77	A (Asp)	G (Gly)	G (Gly)	A (Asp)	G (Gly)	A(2)/G(3)
	22203	244	C (Thr)	T (Ile)	T (Ile)	C (Thr)	T (Ile)	C(2)/T(3)
PUP1	23201	577	T (Ser)	G (Ala)	T (Ser)	T (Ser)	T (Ser)	G(1)/T(4)
	24853	1127	T (Leu)	T (Leu)	C (Leu)	T (Leu)	T (Leu)	C(1)/T(4)
	25550	101	T (Met)	T (Met)	T (Met)	T (Met)	A (Lys)	A(1)/T(4)
M 蛋白	25654	136	C (Gln)	A (Lys)	A (Lys)	A (Lys)	A (Lys)	C(1)/A(4)
	26031	261	C (Pro)	A (Pro)	A (Pro)	A (Pro)	A (Pro)	C(1)/A(4)
PUP3	26581	68	C (Ala)	C (Ala)	C (Ala)	C (Ala)	T (Val)	T(1)/C(4)
	26838	154	T (Ser)	T (Ser)	C (Pro)	T (Ser)	T (Ser)	C(1)/T(4)
Non-ORF	27224	57	T (Leu)	C (Pro)	C (Pro)	C (Pro)	C (Pro)	T(1)/C(4)
碱基(氨基酸)突变数 ^{c)}	27808		C	T	T	C	T	C(2)/T(3)

a) BJ01/AY278488.2, US/AY278741.1, CUHK/AY278554.1, HKU/AY278491.2, TOR2/NC_004718.1. b) 在该位点处通过比较基因组序列得出的突变数目. c) 通过比较 BJ01 分离株的碱基和氨基酸序列得出的突变数目

GC 含量丰富区域, 其位置在基因组 3'端(29571 ~ 29602, 32 nt, cgaggccacgcggagtacgatcgagggtacag), 在 N 蛋白的编码区外, 但它的生物学功能不详.

根据 GenBank 中公布的 17 个不同的冠状病毒总共 64 株的 RNA 聚合酶的核苷酸/氨基酸序列分析, 我们构建了一个系统发生树(图 5). 比较基因组分析包括鸟类中的 4 种病毒(禽传染性支气管炎病毒, 禽传染性喉气管炎病毒, 火鸡冠状病毒和海雀冠状病毒), 啮齿类动物中的 2 种(鼠肝炎病毒和鼠冠状病毒), 家畜或宠物中的 7 种(马冠状病毒, 犬冠状病毒, 猫冠状病毒, 猪流行性痢疾病毒, 猪红血球凝集性脑脊髓炎病毒, 猪遗传性肠胃炎病毒和牛冠状病毒), 以及人类的 4 种(人冠状病毒 229E, 人冠状病毒 OC43, 人肠冠状病毒 4408, SARS 相关冠状病毒). 除了 5' RNA 聚合酶稳定区, 其他编码区或 PUPs 在 SARS 相关病毒和所有目前测序的基因组之间没有发现明显的同源性.

3 讨论

3.1 SARS 相关病毒的分类

在 SARS 相关冠状病毒被确定之前, 关于 SARS 的病原体有很多假设. 干咳、无并发白细胞增高、抗菌素治疗无效等因素排除了 SARS 产生于细菌、衣原体、支原体感染的可能性^[3]. 这种疾病通过呼吸方式

快速传播则表明了病毒感染的可能性. 广泛的非典型症状显示可能为 RNA 病毒感染. 这可能源自于它比 DNA 病毒高得多的复制错误率, 而 DNA 病毒则表现出典型的、相对统一的症状. 基于上述假设, 尽管不能排除其他的病原体的存在, 我们还是把注意力集中在 RNA 病毒上.

最近, 862 个 RNA 病毒的全序列或部分序列已在 GenBank 中列出, 包括 80 个类反转录病毒(retroid)、219 个双链 RNA 病毒、100 个单链 RNA 负链病毒和 463 个单链 RNA 正链病毒 (<http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viruses.html>). 与 GenBank 中所有已知序列进行的基因组水平上的比对说明将它归为冠状病毒家族是合理的 (见补充数据<http://www.genomics.org.cn/SARS>). 这一分类与电子显微镜下的形态表征一致, 即所有冠状病毒家族成员都具有冠状外表特征 (因此命名“冠状”)^[8]. 需要强调的是, 这一相似性主要是源自于 RNA 依赖性 RNA 聚合酶高度保守上, 而与其抗原性、毒性或 SARS 相关病毒的适应性无直接关系.

分类和进化研究对于阐明 SARS 的发病机理、病原起源和进化以及追踪其来源具有很重要的意义, 这有助于预测未来可能再一次出现的致死性传染病. 进化树同时给出一个令专家和大众困惑了很久的问题的答案, 即 SARS 病毒应该是从动物起源的. 那么,

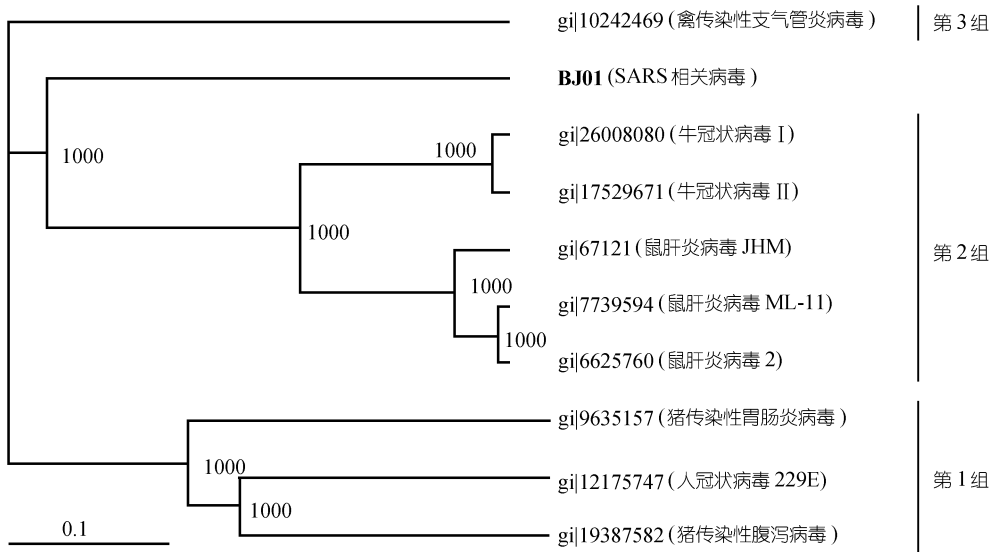


图 5 基于全长 RNA 聚合酶氨基酸序列的 SARS 相关病毒基因组的进化树

根据 RNA 聚合酶序列的类似性推测,病原可能是牛源或鼠源.其他家畜和宠物,由于与人有亲密的接触,成为病原体的可能性也存在.

SARS 病毒是否为人造的问题也得到了普遍关注,它是否是恐怖分子有目的制造的,或是否来自于某个生物实验室不小心泄露.根据我们的序列分析和基因组知识,这些可能性很小.首先,这完全不是简单地将已知的天然基因剪切、重整或以其他已知方式重组的一个产物,因为其序列除 RNA 聚合酶外,均出现高度的序列变异.人工制造这个病毒中惟一可以运用的科学知识是其保守的基因顺序,好似直接由冠状病毒模拟出来.第二,如果它是人为的,那么它或者是未参照任何最新的技术而被随意合成出来的人造品,要么是基于目前还完全未知的技术.

各种各样的毒株以及典型的基因组结构在进化中都发生了高比率的突变,这些突变可以解释其较快的适应性和致命的毒性.可能的机制如 RNA 重组或重新配对,可大大加快病毒通过获得生长优势和增强毒性而进化的速度.现在就对有关病毒的起源下结论还为时过早,进一步的证据将来自于更深入的研究.

3.2 SARS 相关病毒的系统发生学分析

我们对 5 个 SARS 冠状病毒的全基因组序列进行了比较分析,这些病毒分别从加拿大、美国和中国(香港,北京)的病人体内得到的.据此构建了一个 SARS 相关冠状病毒的系统发生树(图 6).

系统发生树显示, BJ01 株与香港分离到的一株 (gi|30027610/CUHK)最接近.香港分离到的另一株

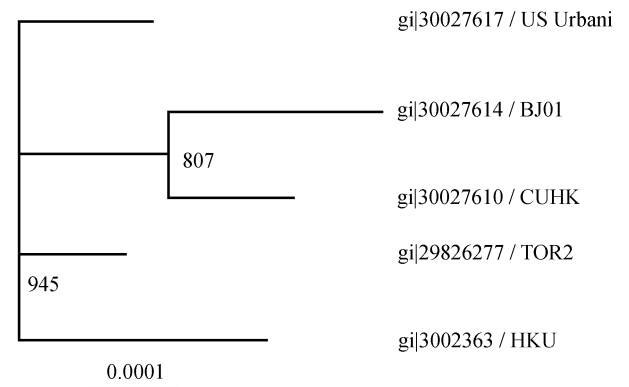


图 6 基于全基因组序列的 5 株 SARS 相关病毒基因组的进化树

(gi|30023963/HKU) 则更接近多伦多的病毒株 (gi|29826277/TOR2). 这可能是由于提取出毒株的那位多伦多患者曾到香港旅行过^[1]. 美国发现的毒株与其他毒株都相距甚远. 如果考虑所有这些都发生在这么短的一段时间内,我们可以想见病毒的突变率极高.当然,这也与 RNA 复制的高错误率相一致^[8]. 进一步的结论只有在获取了更多的分子流行病学数据,依据不同地区分离到的更多的病毒毒株序列,才有可能更清晰地阐明 SARS 在全球传播的途径和传播中可能发生的突变.

3.3 序列变异及临床意义

即使鉴别到的突变中有一大部分(61.3%)位于编码 RNA 聚合酶的序列中(表 3),考虑到该酶分子量比较大,在病毒蛋白的编码区中相对突变率并不高.因此, RNA 聚合酶可以用于不同毒株的基因型鉴别,需要说明的是这与相应的病毒免疫原性并无直接关系.更重要的是, RNA 聚合酶可以支持不依赖于宿主细胞的病毒转录^[9]. 因此,可以以它为靶体进行药物设计或筛选抑制 SARS 相关病毒的复制,原理是基于它只存在于 RNA 病毒中,因此这种对 RNA 依赖性 RNA 复制的抑制不会影响人体细胞的正常功能.治疗艾滋病的“鸡尾酒疗法”就是这样一个成功的例子^[10].

S 蛋白在已知的冠状病毒中的突变率为 0.11%,与全基因组的突变率(0.10%)相近.令人惊讶的是,4 个我们已经鉴别的突变中有 3 个是异义的,其位于测出的氨基酸的疏水性/亲水性明显变化的区域(表 4). 我们已经知道冠状病毒是先通过 S 蛋白与细胞受体结合,以使病毒进入宿主细胞^[11]. 所以我们推测 S 蛋白的突变应该与 SARS 病毒逃逸免疫监测的能力相关.而且它可能是引发体液免疫反应和细胞毒 T 淋巴细胞^[11]的主要抗原.因此, S 蛋白作为潜在的可选抗原,除了其在诊断方面的应用外,还可用于研制疫苗和抑制病毒的中和抗体.

M 蛋白中发现的两个突变都是 C-T 转换的,而且是异义的.由于它相对较小,其突变率(0.30%, 2/665 nt)在病毒结构蛋白中是最高的,且明显高于基因组的总突变率(表 3). 通过比较更多的冠状病毒基因组序列,将可能进一步阐明突变规律并解释所导致的可能的结构和功能的改变.预测的 M 蛋白结构与它的功能相一致,它横跨包膜的双分子脂膜,并

且与包膜的形成和膜转运有关,同时可能参与免疫反应. M 蛋白可作为诊断和治疗用的另一候选蛋白.

在两个内部蛋白 E 蛋白和 N 蛋白中还没有发现突变位点. 一个解释是因为它们对于病毒自身的存活都很重要. E 蛋白是病毒结构蛋白中最小,作为一个位于病毒体中的磷蛋白它参与包膜的组装过程. 考虑到它在进化中的保守性, E 蛋白可以成为药物设计的一个潜在靶点,但是需要对其进行更多的深入研究. 对其蛋白的物理/化学特性的预测结果显示, N 蛋白是一个与基因组 RNA 相关的高碱性蛋白,可形成螺旋核壳体. 它也可作为药物设计与治疗的靶点.

总之,对基因组和遗传学的认识可以让我们深入探索 SARS 相关病毒的发病机理和毒性. 不同毒株的更多序列数据可以拓宽我们对病毒的病因学及其进化的认识,而且对于 SARS 病毒的诊断、预防及治疗具有重要意义. 在开发诊断试剂、预防和治疗用的抗体和疫苗,以及研究免疫反应方面, S 蛋白,也许同时还有 M 蛋白,将是首要考虑的对象. 但是若考虑开发药物, RNA 聚合酶将是一个更有效的靶体,可直接抑制病毒 RNA 复制,通过 N 蛋白可影响病毒核壳体的结构, E 蛋白则可干扰包膜的形成和病毒的装配. M 蛋白和 E 蛋白可作为抑制免疫信号传递路径的靶体,而对信号传递的阻断对于病毒破坏免疫系统十分重要. 结合基因组学和蛋白质组学,对 PUPs 的表达以及未知功能的研究将会取得突破. 更加广泛地研究,掌握更加丰富的数据资源,将使我们逐步了解 SARS 冠状病毒的致病机理. 它如何以这样一个微小的基因组巧妙而有效地利用人体资源,成功地越过人体防御而迅速扩增并给人类带来巨大危害的奥秘必将展现在我们面前.

致谢 感谢科学技术部、中国科学院、国家自然科学基金委员会的经费支持. 同样感谢中国疾病预防控制中心、北京协和医院、北京市政府和杭州市政府的支持合作.

参 考 文 献

1 Poutanen S M, Low D E, Henry B, et al. Identification of severe acute respiratory syndrome in Canada. *N Engl J Med*, 2003, 348: 1995~2005

2 Lee N, Hui D, Wu A, et al. A major outbreak of severe acute respiratory syndrome in Hong Kong. *N Engl J Med*, 2003, 348: 1986~1994

3 Peiris J S, Lai S T, Poon L L, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet*, 2003, 361: 1319 ~ 1325

4 Ksiazek T G, Erdman D, Goldsmith C S, et al. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med*, 2003, 348: 1953 ~ 1966

5 Drosten C, Gunther S, Preiser W, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med*, 2003, 348: 1967 ~ 1976

6 祝庆余, 秦鄂德, 王翠娥, 等. 非典型肺炎病例标本中新型冠状病毒的分离与鉴定. *中国生物工程杂志*, 2003, 23: 106~108

7 Alam S L, Atkins J F, Gesteland R F. Programmed ribosomal frameshifting: much ado about knotting! *Proc Natl Acad Sci USA*, 1999, 96: 14177 ~ 14179

8 Regenmortel V, edi. *Virus Taxonomy*. 7th ed. Academic Press, 2001

9 Uchil PD, Satchidanandam V. Characterization of RNA synthesis, replication mechanism, and *in vitro* RNA-dependent RNA polymerase activity of Japanese encephalitis virus. *Virology*, 2003, 307: 358 ~ 371

10 Henkel J. Attacking AIDS with a “cocktail” therapy? *FDA Consum*, 1999, 33: 12 ~ 17

11 Popova R, Zhang X. The spike but not the hemagglutinin/esterase protein of bovine coronavirus is necessary and sufficient for viral infection. *Virology*, 2002, 294: 222 ~ 236

(2003-04-29 收稿, 2003-05-19 收修改稿)