遗传学报 *Acta Genetica Sinica*, June 2003, 30 (6): 493 ~ 500 ISSN 0379 – 4172

Initial Analysis of Complete Genome Sequences of SARS Coronavirus

CHEN Yun-Jia[1], GAO Ge[1], BAO Yi-Ming[2], Rodrigo LOPEZ[3], WU Jian-Min[1], CAI Tao[1], YE Zhi-Qiang[1], GU Xiao-Cheng[1], LUO Jing-Chu[1,①]

(1. *Collge of Life Sciences, Centre of Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing* 100871, *China*;
2. *National Center for Biotechnology Information, MD* 20892, *USA*;
3. *European Bioinformatics Institute, Hinxton CB*101 *SD, UK*)

Abstract: Multiple sequence alignment among 12 complete SARS coronavirus (SARS-CoV) sequences reveals that the major parts of 29708 b of the genomes have 99.82% identical bases. Forty two nucleotide mismatches were found in addition to the five and six gaps in two genomes. Among them, 28 mismatches result in changes of amino acid in the encoded proteins. Analysis of the changes implies possible effect on the Spike and Membrane protein of the virus, while most of the other changes seem not very significant to alter the structure and function of the proteins. These results have been released on the anti-sars web site maintained by the Centre of Bioinformatics, Peking University (antisars.cbi.pku.edu.cn) and may be of help for further experimental study.

Key words: SARS coronavirus, multiple sequence alignment, sequence analysis, bioinformatics

The outbreak of the Severe Acute Respiratory Syndrome (SARS) starting from southern China early this year has a significant influence on public health. The identification of SARS-CoV as the major causative factor of the SARS disease and the genomic sequencing of the virus makes it possible for bioinformatics study. A total of 16 SARS-CoV genome sequences are available from the nucleic acid database GenBank/EMBL/DDBJ by 20 May 2003. SARS-CoV ZJ01 (AY297028.1) was shown in GenBank after the analysis was performed.

12 complete genomes were retrieved from GenBank (Table 1) with general information such as the name of the sequence, the accession number, the date of deposit, release and latest update. A special code was designated for each sequence, e.g., CA1 for NC_004718.3 from Canada.

A PC/Linux (RedHat 8.0) platform with 2 processors (2.2GHz) and 4GB memory was setup for this study and several bioinformatics tools were installed on the system. ClustalW (ver 1.82) for multiple sequence alignment and PSIPRED (ver 2.3) for protein secondary structure prediction are freely available, while TMHMM (ver 2.0) for transmembrane prediction, PSORT II for protein localisation and HMMER (ver.2.1.1) for constructing hidden Markov model, were kindly provided by the authors.

Multiple sequence alignment of 12 SARS-CoV genomes reveals a high degree of sequence similarity. The major part of the 29708b of all 12 genomes has 99.82% identical bases. Forty two nucleotide mismatches were found in addition to the five and six gaps in two genomes. A table of all mismatched bases was created (Table 2). Information of the positions of the mismatches in the genome, the genetic codons and the proteins they encode and the changes of the amino acids caused by each mismatch are listed in the table. Among all 42 mismatches, 28 cause changes of amino acids in the coded proteins. Protein sequence analysis including secondary structure prediction, transmembrane alpha helix location, signal peptide and nucleic localization signal identification was performed on these proteins to explore the possible alternation of structure, conformation and function which may be affected by the change of the amino acids in special positions. This analysis reveals possible effects on the Spike and Membrane protein of the

收稿日期: 2003 – 05 – 21 修回日期: 2003 – 05 – 27

基金项目: 863 基金 (2001AA231011, 2002AA231061)、国家自然科学基金 (30170232) 资助 [Supported by the "863" Program (2001AA231011, 2002AA231061) and the National Natural Science Foundation of China (30170232)]

作者简介: 陈蕴佳 (1976 –), 男, 汉族, 吉林人, 博士生, 研究方向: 生物信息学

① 通讯作者。罗静初 (1947 –), 男, 上海人, 教授, 博士生导师, 研究方向: 生物信息学。E-mail: luojc@pku.edu.cn; Tel: 86-10-6275-7281

SARS-CoV, while most of other changes seem not very significant. These results have been released on the anti-sars web site maintained by the Centre of Bioinformatics, Peking University (antisars.cbi.pku.edu.cn) and may help biologists for further experimental study. A hypertext version of this paper which contains various materials including sequence data, analysis tools and output results can be found at the above web site.

Acknowledgements: Thanks to Hao Bailin and Zhang Chunting for discussion on genome sequence analysis, and to the email message from Li Wei about the sequencing accuracy information. Thanks to David Lipman, Dennis Benson, Stephen Bryant, Tatiana Tatusova and staff at NCBI viral genome group for their support to the CBI anti-sars web site. Qi Ji and Zheng Nan have provided us with their SARS phylogeny analysis results for online publication.

SARS 冠状病毒全基因组序列初步分析

陈蕴佳[1], 高 歌[1], 鲍一明[2], Rodrigo LOPEZ[3], 吴健民[1], 蔡 涛[1], 叶志强[1], 顾孝诚[1], 罗静初[1,①]

(1. 北京大学生命科学学院, 北京大学生物信息中心, 北京大学蛋白质工程和植物基因工程国家重点实验室, 北京 100871;
2. 美国国家生物技术信息中心, MD 20892, USA; 3. 欧洲生物信息学研究所, Hinxton CB101SD, UK)

摘 要: 对已经完成全序列测定的 12 个 SARS 病毒基因组进行了多序列比对, 发现序列主体部分 29708 b 具有 99.82% 的相同碱基, 除 2 个序列各有 5 个和 6 个碱基的缺失外, 其余部分共有 42 个位点核苷酸碱基的差异, 其中 28 个位点的碱基差异可引起氨基酸残基改变。利用蛋白质二级结构和跨膜螺旋预测以及蛋白质定位等生物信息学工具, 分析了这些产生氨基酸改变部位的蛋白质构像, 推测了可能产生的结构和功能改变, 为进一步生物学实验提供参考。所有分析结果同时在北京大学生物信息中心抗 SARS 网站(antisars.cbi.pku.edu.cn)上发布。

关键词: SARS 冠状病毒; 多序列比对; 蛋白质序列分析; 生物信息学

中图分类号: Q939.47 **文献标识码**: A **文章编号**: 0379-4172(2003)06-0493-08

自 2002 年 11 月在我国广东发现首例重症急性呼吸道疾病(severe acute respiratory syndrome, SARS, 俗称非典型性肺炎)患者以来, 该传染病已在 30 多个国家和地区出现。2003 年 4 月 13 日, 加拿大基因组科学中心率先完成了 SARS 冠状病毒(corona virus)全基因组测序, 截至 2003 年 5 月 20 日, 国际核酸序列数据库 GenBank/EMBL/DDBJ 已发布了 17 个 SARS 病毒基因组序列, 其中 13 个序列的测定已经全部完成, 为深入研究 SARS 病毒的传播、侵染和与宿主的相互作用奠定了分子生物学基础。

自 1977 年 Sanger 首先测定噬菌体 ϕX174 序列全长 5385 b 的基因组序列以来[1], 已经完成全基因组测序的病毒基因组共 1542 个(NCBI 基因组资源网站 2003 年 5 月 24 日数据, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viruses.html)。已测序的病毒基因组中, 单链正链 RNA 病毒为数最多, 共 472 个, 包括 7 个冠状病毒。除 SARS 病毒外, 还有宿主为牛、猪、鼠、禽类等的其他冠状病毒, 以及德国科学家 2000 年 9 月递交的人冠状病毒[2]。

冠状病毒基因组序列全长约 30 kb, 约为人类基因组的 10 万分之一, 大肠杆菌基因组的 15 分之一, 是目前所有单链 RNA 病毒中最大的一种, 具囊膜, 可引起人类和其他动物很多疾病[3]。冠状病毒基因组的最终基因产物共包括 20 多种蛋白质, 分为结构蛋白和非结构蛋白两大类。部分非结构蛋白的功能已经基本确定, 如 RNA 依赖型 RNA 聚合酶(RNA-dependant RNA polymerase, RdRp), 蛋白水解酶等。主要结构蛋白包括核壳蛋白(Nucleocaspid, 即蛋白 N)和 3 ~ 4 种膜蛋白。膜蛋白中 M 蛋白(Membrane protein)含量丰富, 跨膜 3 次, 其 N 端较短, 位于膜外, C 端较长, 位于膜内; S 蛋白(Spike protein)跨膜 1 次, N 端在膜外, C 端在膜内, 是病毒感染宿主细胞的主要成分。E 蛋白(Envelope protein)通常被认为是囊膜的次要结构蛋白, 在鸟类传染性支气管炎病毒(avian infectious bronchitis virus, IBV), 传染性肠胃炎病毒(Transmissible gastroenteritis virus, TGEV)和小鼠肝炎病毒(murine hepatitis virus, MHV)中已得到证实[4]。某些冠状病毒中还可能存在血凝素酯酶

(hemagglutinin-esterase，HE)，但目前对此酶研究不多。冠状病毒 RNA 基因组具有 5′帽子和 3′多聚腺苷酸(PolyA)尾，与普通 mRNA 分子结构相似，和 N 蛋白形成螺旋状核糖核蛋白，并与 M 蛋白、S 蛋白通过出芽方式进入内质网-高尔基体复合物，整合成病毒粒子，最终通过分泌泡到达细胞表面，分泌到胞外[5,6]。

本文利用生物信息学手段，对已经完成全序列测定的 SARS 冠状病毒基因组序列进行初步分析，旨在为进一步分析和生物学实验提供一些参考。

1　材料和方法

1.1　SARS 冠状病毒基因组序列来源

SARS 冠状病毒全基因组序列取自美国国家生物技术信息中心(National Center for Biotechnology Information，NCBI)的核酸序列数据库 GenBank。16 个 SARS 冠状病毒基因组序列中，12 个为完整基因组，表 1 列出 13 个完整基因组的序列，以递交日期为序(由浙江疾病控制中心和军事医学科学院测定的冠状病毒全基因组序列 ZJ01 于 2003 年 5 月 19 日发布，本文未作分析)。序列长度范围为 29705 ~ 29751 b，相差为 46 b。其中某些序列 3 ′端有多聚腺苷酸(PolyA)加尾信号，5′端有长度不等的多余碱基，可能与测序过程有关。其中香港中文大学递交的 CUHK-W1 和 CUHK-Su10 均为 29736 b，新加坡基因组研究中心递交的 SIN2500、SIN2679 和 SIN2774 等 3 个序列均为 29711 b。其中 TOR2 等 6 个序列已有关于编码蛋白质的初步注释信息，主要蛋白注释内容基本相同。

表 1　12 个完成全基因组测序的 SARS 冠状病毒

Table 1　12 SARS coronaviruses with complete genome sequence

代　码	序列名称	数据库代码	序列长度	递交日期	首次发布日期	最近更新日期
Code	Name	Accession	Length	Submit	First Release	Latest update
CA1	TOR2	NC _ 004718.3	29751b	13-Apr-2003	14-Apr-2003	16-May-2003
BJ1	BJ01	AY278488.2	29725b	17-Apr-2003	21-Apr-2003	01-May-2003
US1	Urbani	AY278741.1	29727b	17-Apr-2003	21-Apr-2003	21-Apr-2003
HK1	HKU-39849	AY278491.2	29742b	17-Apr-2003	18-Apr-2003	18-Apr-2003
HK2	CUHK-W1	AY278554.2	29736b	17-Apr-2003	18-Apr-2003	14-May-2003
HK3	CUHK-Su10	AY282752.1	29736b	24-Apr-2003	07-May-2003	07-May-2003
SG1	SIN2500	AY283794.1	29711b	27-Apr-2003	09-May-2003	09-May-2003
SG2	SIN2677	AY283795.1	29705b	27-Apr-2003	09-May-2003	09-May-2003
SG3	SIN2679	AY283796.1	29711b	27-Apr-2003	09-May-2003	09-May-2003
SG4	SIN2748	AY283797.1	29706b	27-Apr-2003	09-May-2003	09-May-2003
SG5	SIN2774	AY283798.1	29711b	27-Apr-2003	09-May-2003	09-May-2003
TW1	TW1	AY291451.1	29729b	06-May-2003	14-May-2003	14-May-2003

CA1：加拿大基因组科学中心(Genome Sciences Centre，Canada)；US1：美国疾病控制中心(Centers for Disease Control and Prevention，USA)；HK1：香港大学(The University of Hong Kong)；HK2/HK3：香港中文大学 (Chinese University of Hong Kong)；BJ1：军事医学科学院/华大基因组中心(Academy of Military Medical Sciences/Beijing Genomics Institute)；SG1-5：新加坡基因组研究所(Genome Institute of Singapore)；TW1：中国台湾大学，National Taiwan University。

1.2　分析平台和工具

本文采用基于双 CPU 的 PC 服务器作冠状病毒基因组序列分析平台，主频 2.2 GHz，内存 4 GB，操作系统为 Linux RedHat8.0 版。多序列比对采用 ClustalW 1.82 版[7]，跨膜螺旋预测采用 TMHMM 2.0 版[8]，蛋白质二级结构预测采用 PSIPRED 2.3 版[9]，隐马模型构建程序选用 HMMER2.1.1 版[10]，蛋白质亚细胞定位预测采用 PSORT II 程序[11]。上述分析程序均安装在本地服务器上，其中 ClustalW、PSIPRED 和 HMMER 可从文献提供的网站上下载，TMHMM 和 PSORT 由作者特别提供。Pfam 数据库 7.8 版[12] 从网上下载。

1.3　分析方法

1.3.1　多序列比对

对上述 12 个冠状病毒完整基因组序列按 FASTA 格式存放为一个输入文件，并对注释行作适当编辑。按序列来源统一命名，如 BJ1 表示来自北京的第 1 个序列、SG1 表示来自新加坡的第 1 个序列等；删除‘SARS complete genome’等冗余信息，增加基因组长度和更新日期。用 ClustalW 对上述 12 个基因组进行多序列比对，分析初步比对结果后发现，所有 12 个序列具有较高的相似性，除 SG2(SIN2677)在近

3′端有连续 6 个碱基缺失，SG4(SIN2748)在近 3′端有相临 5 个碱基缺失外，其他部位没有任何多碱基或单碱基插入和缺失。根据以上初步结果，对 12 个序列进行适当编辑，去除 5′端多余碱基和 3′端多聚腺苷酸，保留序列主体部分 29708 个碱基。对编辑后的 12 个序列重新进行多序列比对，找出对比对结果中 42 个差异位点的碱基和密码子改变，及其引起编码蛋白相应氨基酸残基的改变，所得结果如表 2 所示，表中碱基差异引起氨基酸改变的蛋白质的名称和功能如表 3 所示，这些蛋白质在基因组上对应的位置如图 1 所示。

表 2 SARS 冠状病毒基因组序列差异

Table 2 Mismatches among 12 genome sequences of SARS coronavirus*

编号	位点	次数	碱基变化 Mismatch												密码子	蛋白(位点)	氨基酸
No	Site	Freq	CA1	BJ1	US1	HK1	HK2	HK3	SG1	SG2	SG3	SG4	SG5	TW1	Codon	Protein(site)	AA
1	1476	1	a	a	a	a	a	a	a	G	a	a	a	a	AGA-AGG	P65(225)	R-R
2	2601	1	t	t	t	C	t	t	t	t	t	t	t	t	GTT-GTC	P65(600)	V-V
3	3165	1	a	a	a	a	a	a	a	a	a	a	a	G	TCA-TCG	Nsp1(149)	S-S
4	7746	1	g	g	g	g	T	g	g	g	g	g	g	g	CCG-CCT	Nsp1(1676)	P-P
5	7919	1	c	c	T	c	c	c	c	c	c	c	c	c	GCT-GTT	Nsp1(1734)	A-V
6	7930	1	g	g	g	A	g	g	g	g	g	g	g	g	GAC-AAC	Nsp1(1738)	D-N
7	8387	1	g	g	g	C	g	g	g	g	g	g	g	g	AGT-ACT	Nsp(1890)	S-T
8	8417	1	g	g	g	C	g	g	g	g	g	g	g	g	AGA-ACA	Nsp1(1900)	R-T
9	8572	1	g	T	g	g	g	g	g	g	g	g	g	g	GTA-TTA	Nsp1(1952)	V-L
10	9404	2	t	C	t	t	C	t	t	t	t	t	t	t	GTT-GCT	Nsp1(2229)	V-A
11	9479	1	t	t	t	t	C	t	t	t	t	t	t	t	GTA-GCA	Nsp1(2254)	V-A
12	9854	1	c	T	c	c	c	c	c	c	c	c	c	c	GCC-GTC	Nsp1(2379)	A-V
13	10587	1	a	C	a	a	a	a	a	a	a	a	a	a	ACA-ACC	3cl(201)	T-T
14	13494	1	g	g	g	A	g	g	g	g	g	g	g	g	GTT-AGT	RdRp(42)	V-S
15	13495	1	t	t	t	G	t	t	t	t	t	t	t	t	GTT-AGT	RdRp(42)	V-S
16	16622	1	c	c	T	c	c	c	c	c	c	c	c	c	GCC-GCT	Nsp10(152)	A-A
17	17564	2	t	G	t	t	G	t	t	t	t	t	t	t	GAT-GAG	Nsp10(466)	D-E
18	17846	2	c	c	c	c	T	T	c	c	c	c	c	c	CGC-CGT	Nsp10(560)	R-R
19	18065	1	g	g	g	A	g	g	g	g	g	g	g	g	AAG-AAA	Nsp11(32)	K-K
20	18282	1	c	c	c	c	c	c	c	c	A	c	c	c	CTA-ATA	Nsp11(105)	L-I
21	18965	1	t	t	t	t	t	t	t	t	t	t	A	t	ATT-ATA	Nsp11(332)	I-I
22	19064	2	a	a	G	a	G	a	a	a	a	a	a	a	GAA-GAG	Nsp11(365)	E-E
23	19084	4	c	c	c	c	c	c	T	T	c	T	T	c	ACA-ATA	Nsp11(372)	T-I
24	19838	1	a	G	a	a	a	a	a	a	a	a	a	a	GTA-GTG	Nsp12(96)	V-V
25	21721	2	g	A	g	g	A	g	g	g	g	g	g	g	GGC-GAC	Spike(77)	G-D
26	22222	2	t	C	t	t	C	t	t	t	t	t	t	t	ATT-ACT	Spike(244)	I-T
27	23174	1	c	c	c	c	c	c	c	c	T	c	c	c	TCC-TCT	Spike(561)	S-S
28	23220	1	G	t	t	t	t	t	t	t	t	t	t	t	TCT-GCT	Spike(577)	S-A
29	23792	1	c	c	c	c	c	c	c	c	c	c	T	c	GTC-GTT	Spike(767)	V-V
30	24872	1	t	t	C	t	t	t	t	t	t	t	t	t	CTT-CTC	Spike(1127)	L-L
31	25298	1	A	g	g	g	g	g	g	g	g	g	g	g	GGA-AGA	SARS3a(11)	G-R
32	25569	1	t	t	t	A	t	t	t	t	t	t	t	t	ATG-AAG	SARS3a(101)	M-K
33	25673	1	a	C	a	a	a	a	a	a	a	a	a	a	AAG-CAG	SARS3a(136)	K-Q
34	26050	1	a	C	a	a	a	a	a	a	a	a	a	a	CCA-CCC	SARS3a(261)	P-P
			a		a	a	a	a	a	a	a	a	a	a	CAA-CCA	SARS3b(121)	Q-P
35	26428	1	g	g	g	g	g	g	A	g	g	g	g	g	GAG-AAG	Membrane(11)	E-K
36	26477	1	t	t	t	t	t	G	t	t	t	t	t	t	TTC-TGC	Membrane(27)	F-C
37	26600	1	c	c	c	T	c	c	c	c	c	c	c	c	GCT-GTT	Membrane(68)	A-V
38	26857	1	t	t	C	t	t	t	t	t	t	t	t	t	TCC-CCC	Membrane(154)	S-P
39	27111	1	a	a	a	a	a	a	a	G	a	a	a	a	GAG-GGG	SARS6(13)	E-G
40	27243	1	c	T	c	c	c	c	c	c	c	c	c	c	CCT-CTT	SARS6(57)	P-L
41	27827	2	t	C	t	t	C	t	t	t	t	t	t	t	TGC-CGC	SARS8a(17)	C-R
42	28696	1	g	g	g	g	g	T	g	g	g	g	g	g	GGT-TGT	Nucleocapsid(193)	G-C
总数		52	2	12	5	9	9	3	2	3	2	1	3	1			

* 表中第 1 列为差异碱基编号(No) 第 2 为变化位点(Site)，以 CA1(NC_004718.3)为参考 第 3 列为该位点上碱基变化次数(Freq) 碱基变化(Mismatch)栏中列出 12 条序列 42 个位点碱基变化，序列代码见表 1 密码子(Codon)栏列出密码子改变 蛋白质(位点)栏(Protein(site))给出碱基变化所在蛋白质及变化位点 氨基酸(AA)栏列出氨基酸残基改变，用单字符表示。

表 3　SARS 冠状病毒基因组序列差异引起氨基酸残基改变的蛋白质

Table 3　Proteins with amino acid change deduced from mismatches of SARS corona virus genome sequences

蛋白名称	说　明	残基改变数
Nsp1	非结构蛋白 1，可能对应于其它冠状病毒似木瓜蛋白酶(PLP-1 或 PLP-2)	8
RdRp	RNA 依赖性 RNA 聚合酶	1
Nsp10	非结构蛋白 10，功能不详，含与金属离子结合的结构域和 NTP 酶/螺旋酶结构域	1
Nsp11	非结构蛋白 11，功能不详	2
Spike	即 S 蛋白，负责病毒侵染	3
SARS3a	第 3 个基因第 1 个阅读框所编码蛋白，功能未知	3
SARS3b	第 3 个基因第 2 个阅读框所编码蛋白，功能未知	1
Membrane	即 M 蛋白，对病毒粒子的形成起关键作用	4
SARS6	第 6 个基因所编码蛋白，功能未知	2
SARS8a	第 8 个基因第 1 个阅读框所编码蛋白，功能未知	1
Nucleocapsid	核壳蛋白，即 N 蛋白	1

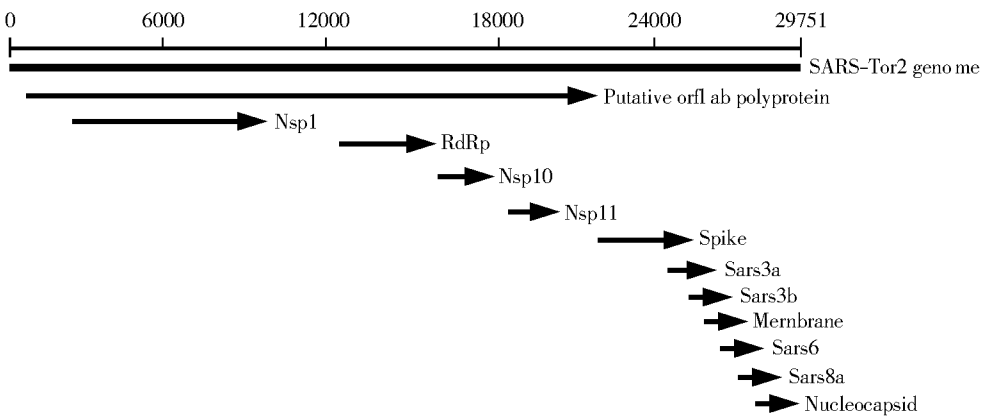

图 1　基因组序列差异引起氨基酸残基改变的蛋白质在基因组上位置

Fig.1　Positions of proteins with amino acid changes deduced from mismatches of genome sequences of SARS coronaviras

1.3.2　编码蛋白质特性分析

对表 2 所示具有氨基酸改变的 11 个蛋白质进行初步分析，预测其可能的二级结构和跨膜螺旋，分析氨基酸残基改变的位点处二级结构构像变化，并推测这些变化对蛋白质的空间结构构像及活性的影响。

2　结果与讨论

2.1　多序列比对

12 个 SARS 冠状病毒基因组多序比对结果表明，除 5′和 3′端外，序列主体部分 29708 b 具有很高的相似性，相同碱基 29654 个，占 99.82%；SG2 和 SG4 两序列各有 6 个和 5 个碱基的缺失，SG4 有一个位点未知。除此之外的差异部位共 42 处，其中一个位点有连续 2 个碱基的差异，而 2 个序列在同一位点出现差异的有 7 处，4 个序列在同一位点出现差异的有 1 处。按万分之一测序误差计(李蔚，电子邮件通讯)，12 个基因组测序误差引起的差异总数应该不会超过 36 处，2 个或 2 个以上在同一位点序列出现碱基差异的可能性极小。可以认为，12 个来自不同地区、不同时间的 SARS 冠状病毒基因组可能产生了点突变。十分有趣的是，香港中文大学递交的 2 个序列，长度均为 29736 b，有 10 个位点的差异，而长度均为 29711 b 的 SG1、SG3 和 SG5 却只有 4 个位点差异(表 2)。

比较已有 6 条基因组序列的注释信息，发现主要蛋白的注释基本相同，我们以加拿大基因组科学中心 2003 年 5 月 16 日第 3 版的 CA1 (NC_004718.3)序列及注释信息为参照，分析了 42 个位点碱基改变引起的密码子和相应的氨基酸残基改

变,发现有 1 个碱基位于 2 个读码框(SARS3a 和 SARS3b)内,但其中仅有一个碱基变化可引起氨基酸残基改变(SARS3b,Q-P);此外,另有 14 个位点碱基的改变不引起氨基酸残基改变,即密码子不同,但所编码氨基酸相同;而其余 27 个位点则可引起氨基酸改变,此中 7 个为疏水残基 A-V、V-A、L-I 之间的改变,3 个为相似残基 D-N、S-T、D-E 之间的改变,另外 17 个则为非相似残基之间的改变,如 V-S、G-D、G-R、P-L 等。

2.2 编码蛋白质氨基酸残基改变分析

假定上述 12 个序列的测定完全正确,对 28 个位点氨基酸残基改变在不同编码蛋白中的分布和可能引起的结构功能变化,分别讨论如下。

2.2.1 结构蛋白

出现氨基酸改变的结构蛋白包括功能已知的 S、M、E 和 N 等蛋白,以及功能未知的 SARS3a、SARS3b、SARS6 和 SARS8a 蛋白。对这些蛋白分析如下:

S 蛋白(即 Spike 蛋白)是负责病毒侵染的主要部分[13]。S 蛋白的差异可导致病毒宿主的更迭,并可使宿主产生感冒、腹膜炎、肠胃炎等多种疾病[14~16]。S 蛋白同时也是影响冠状病毒侵染宿主程度的主要成分[17]。在多数冠状病毒中,S 蛋白存在两种形式,一种为独立的单链蛋白,另一种为两个大小相似的蛋白剪切产物,称 S1 和 S2。S1 是与宿主细胞受体相互作用的主要部分[18],S2 则是参与细胞融合的主要部分。HMMER-PFAM 分析表明,SARS 病毒 S 蛋白 75~609 区与冠状病毒 S 蛋白的 S1 域相似,641~1247 区与 S2 域相似,787~1221 区与病毒融合糖蛋白 F0 相似。SARS 病毒 S 蛋白中 3 个突变位点 G-D,I-T 和 S-A 均位于 S1 区。考虑到氨基酸性质的变化,以及 S1 在侵染宿主细胞中的作用,这些位点的改变有可能影响 S 蛋白与宿主细胞受体之间相互作用。

M 蛋白是病毒粒子的主要结构蛋白之一,对病毒粒子的形成起关键作用。TMHMM 分析表明,此蛋白有 3 个跨膜区。4 个氨基酸改变的位点中,第 154 位(S-P)位于病毒粒子内部,可能会影响到与 N 蛋白的相互作用,以及病毒粒子的形成。其余 3 个位点中,第 11 位位点变化位于病毒粒子外部,可能影响与宿主受体的作用,但 M 蛋白与宿主受体结合的实验证据尚无文献报道。第 27(F-C)和 68(A-V)位变化位点位于跨膜区,推测不会影响跨膜区的构像和性质。

N 蛋白(即核壳蛋白)与病毒基因组结合,形成螺旋状结构,并与 M 蛋白相互作用,是主要的病毒结构蛋白之一。对 31 个冠状病毒 N 蛋白的多序列比对表明,SARS 病毒有两个序列保守区。蛋白质亚细胞定位程序 PSORT 分析表明,N 蛋白中存在一些核定位信号,但改变的位点(G-C)不在这些区域内,因此推测这个变化不会对这些保守区产生直接影响。

SARS3a 蛋白全长 274 个残基,共有 G-R、M-K、K-Q 等 3 个位点改变。TMHMM 预测表明,此蛋白包括 3 个跨膜区,可能是病毒膜上结构蛋白之一,但 3 个变化位点均分布在跨膜区外。SARS3b 蛋白全长 154 个残基,PSORT 和 TMHMM 预测表明此蛋白不含有跨膜区、信号肽,却含有核定位信号;但仅有的一个突变位点(Q-P)并不在核定位信号区。SARS6 蛋白全长 63 个残基,有两个位点改变(E-G 和 P-L)。PSIPRED 二级结构预测结果表明,第一个位点位于 α-螺旋区;第二个位点位于回环区。SARS8a 蛋白很短,全长只有 39 个残基,PSIPRED 二级结构预测表明,所改变的第 17 位(C-R)位于 α-螺旋区。上述位点的改变,是否会影响这些蛋白质的功能,需要直接的实验证据。

2.2.2 非结构蛋白

非结构蛋白中产生氨基酸改变的包括非结构蛋白 1(Non-structural protein 1,Nsp1)、RNA 依赖型 RNA 聚合酶(RdRp)、非结构蛋白 10(Nsp10)和非结构蛋白 11(Nsp11)。

Nsp1 全长 2422 个氨基酸残基,共有 8 个残基改变,均位于 C 端区,其中 2 个位于可能的跨膜区中,分别为 V-L 和 V-A,均为疏水性残基之间的变化,不会影响跨膜螺旋构像。其余改变部位中 3 个为疏水性残基之间的变化,即 A-V、V-L 和 V-A;1 个为天冬氨酸和天冬酰胺之间的改变(D-N),位于回环区;这 4 个位点的改变估计不会产生很大影响。仅有 1 个位于螺旋区内(S-T),1 个位于 β 折叠(R-T)。前者为性质相似的丝氨酸和苏氨酸,影响不大;后者为碱性残基精氨酸和亲水性残基苏氨酸之间的改变,影响也不会很大。

RdRp 是病毒基因组复制和子基因组 mRNA 合成必须的聚合酶,RdRp 的突变将影响病毒复制和转录。SARS 冠状病毒 RdRp 全长 932 个残基,所改变

的位点位于第 42 位(V-S)。图 2 为 SARS 病毒和其他 6 种冠状病毒 RdRp 序列比对结果。从图 2 中可以看到 SARS-CoV 第 42 位不是保守位点,可以允许 V、A、S、R 等不同残基。

Nsp10 蛋白含有金属离子结合结构域和 NTP 酶/螺旋酶结构域,因此很有可能是病毒复制、转录的重要辅助酶。此蛋白中改变位点位于第 466 位(D-E),均为酸性氨基酸,推测不会影响 Nsp10 蛋白结构。Nsp11 蛋白全长 527 个残基,与已知蛋白没有任何相似性,含两个改变位点 L-I 和 T-I,其中 L 和 I 性质相似,但 T(苏氨酸)和 I(异亮氨酸)性质不同,是否会对 Nsp11 蛋白构像和功能产生影响,需要进一步实验验证。

图 2　RdRp 蛋白 N 端多序列比对结果

Fig.2　Multiple sequence alignment of N terminal of RdRp among 7 coronaviruses

通过上面的分析可以看出,12 个 SARS 冠状病毒基因组序列的 42 处变化,对多数蛋白产物不会产生大的结构功能改变,可能产生改变的主要是 S 蛋白的 S1 部分和 M 蛋白。尽管这些变化仅为个别残基改变,但有可能导致宿主种类、侵染部位、侵染强度的改变。目前,尽管已经有了 SARS 冠状病毒基因组序列分析[21]和蛋白酶 3C-like 空间结构模型[22]等报道,国内也已开始发表相关文章[23,24]。关于冠状病毒蛋白产物结构功能尚无足够的实验证据,以上分析仅为生物学实验提供一些参考信息。

本文网络版包括所有原始数据和分析结果的链接,其中表 2 和图 2 为彩色版,可通过北京大学生物信息中心 SARS 相关生物信息学专业网站(antisars.cbi.pku.edu.cn)进入。该网站同时刊登国内外有关 SARS 研究的最新消息、文献和研究结果,提供基因组序列、蛋白质结构等数据查询和下载,以及国外主要 SARS 研究生物信息学网站的链接。该网站已在美国国家生物技术信息中心(NCBI)和欧洲生物信息学研究所(EBI)的 SARS 专门网页建立了链接[http://www.ncbi.nlm.nih.gov/genomes/SARS/SARS.html,http://www.ebi.ac.uk/2can/disease/SARS.html]。

致　谢：感谢郝柏林院士和张春霆院士关于 SARS 基因组序列分析的讨论和修改意见,感谢李蔚博士关于测序精度的电子邮件,感谢美国国家生物技术信息中心主任 Lipman 院士、GenBank 数据库主任 Benson 博士、蛋白质结构组主任 Bryant 博士、基因组 Tatusova 博士以及病毒基因组全体工作人员对北京大学生物信息中心抗 SARS 网站的支持,感谢戚继和郑楠同学为该网站提供了 SARS 病毒网络版论文。

参考文献(References):

[1] Sanger F, Air G M, Barrell B G, Brown N L, Coulson A R, Fiddes C A, Hutchison C A, Slocombe P M, Smith M. Nucliotide sequence of bacteriophage phi X174 DNA. *Nature*, 1977, 265: 687 ~ 95.

[2] Thiel V, Herold J, Schelle B and Siddell S G. Infectious RNA transcribed in vitro from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. *J Gen Virol*, 2001, 82 (Pt6): 1273 ~ 1281.

[3] Lai M M, Cavanagh D. The molecular biology of coronaviruses. *Adv Virus Res*, 1997, 48: 1 ~ 100.

[4] Siddell S G The small-membrane protein, p. 181 ~ 189. In S. G. Siddell (ed.). The Coronaviridae. Plenum Press, New York, N. Y.

[5] Klumperman J, Locker J K, Meijer A, Horzinek M C, Geuze H J, Rottier P J. Coronavirus M proteins accumulate in the Golgi complex beyond the site of virion budding. *J Virol*, 1994, 68: 6523 ~ 6534.

[6] Krijnse-Locker J, Ericsson M, Rottier P J, Griffiths G.. Characterization of the budding compartment of mouse hepatitis virus: evidence that transport from the RER to the Golgi complex requires only one vesicular transport step. *J Cell Biol*, 1994, 124: 55 ~ 70.

[7] Thompson J D, Higgins D G, Gibson T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res*, 1994, 22: 4673 ~ 4680.

[8] Sonnhammer E L, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 1998, 6: 175 ~ 182.

[9] McGuffin L J, Bryson K, Jones D T. The PSIPRED Protein Structure Prediction Server. *Bioinformatics*, 2000, 16(4): 404 ~ 405.

[10] Eddy S R. Multiple alignment using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol*, 1995, 3: 114 ~ 120.

[11] Nakai K and Horton P. PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem Sci*, 1999, 24: 34 ~ 35.

[12] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S R, Griffiths-Jones S, Howe K L, Marshall M, Sonnhammer E L. The Pfam Protein Families Database. *Nucl Acids Res*, 2002, 30: 276 ~ 280.

[13] Krueger D K, Kelly S M, Lewicki D N, Ruffolo R, Gallagher T M.. Variations in Disparate Regions of the Murine Coronavirus Spike Protein Impact the Initiation of Membrane Fusion. *J Virol*, 2001, 75: 2792 ~ 2802.

[14] Das Sarma J, Fu L, Tsai J C, Weiss S R, Lavi E. Demyelination determinants map to the spike glycoprotein gene of coronavirus mouse hepatitis virus. *J Virol*, 2000, 74: 9206 ~ 9213.

[15] Leparc-Goffart I, Hingley S T, Chua M M, Phillips J, Lavi E, Weiss S R. Targeted recombination within the spike gene of murine coronavirus mouse hepatitis virus-A59: Q159 is a determinant of hepatotropism. *J Virol*, 1998, 72: 9628 ~ 9636.

[16] Sanchez C M, Izeta A, Sanchez-Morgado J M, Alonso S, Sola I, Balasch M, Plana-Duran J, Enjuanes L. Targeted recombination demonstrates that the spike gene of transmissible gastroenteritis coronavirus is a determinant of its enteric tropism and virulence. *J Virol*, 1999, 73: 7607 ~ 7618.

[17] Navas S, Seo S H, Chua M M, Sarma J D, Lavi E, Hingley S T, Weiss S R. Murine Coronavirus Spike Protein Determines the Ability of the Virus To Replicate in the Liver and Cause Hepatitis. *J Virol*, 2001, 75: 2452 ~ 2457.

[18] Kubo H, Yamada Y K, Taguchi F. Localization of neutralizing epitopes and the receptor-binding site within the amino-terminal 330 amino acids of the murine coronavirus spike protein. *J Virol*, 1994, 68: 5403 ~ 5410.

[19] Klumperman J, Locker J K, Meijer A, Horzinek M C, Geuze H J, Rottier P J. Coronavirus M proteins accumulate in the Golgi complex beyond the site of virion budding. *J Virol*, 1994, 8: 6523 ~ 6534.

[20] Krijnse-Locker J, Ericsson M, Rottier P J, Griffiths G. Characterization of the budding compartment of mouse hepatitis virus: evidence that transport from the RER to the Golgi complex requires only one vesicular transport step. *J Cell Biol*, 1994, 124: 55 ~ 70.

[21] Ruan Y J, Wei C L, Ee L A, Vega V B, Thoreau H, Yun S T S, Chia J M, Ng P, Chiu K P, Lim L, Tao Z, Peng C K, Ean L O L, Lee N M, Sin L Y, Ng L F P, Chee R E, Stanton L W, Long P M, Liu E T. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *The Lancet*? Published online 9 May 2003.

[22] Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus Main Proteinase (3CL^{pro}) Structure: Basis for Design of Anti-SARS Drugs. *Science*, Published online 13 May 2003.

[23] QIN E 'de, ZHU Qingyu, YA Man, *et al*. A complete sequence and comparative analysis of SARS-associated virus (Isolate BJ01). *Chinese Science Bulletin*, 2003, 48: 941 ~ 948.

[24] ZHANG Wen-Guang, LI Jin-Quan, ZHOU Huan-Min. Genomic characterization of SARS coronavirus: A novel member of coronavirus. *Acta Genetica Sinica*, 2003, 30(6): 501 ~ 508.
张文广, 李金泉, 周欢敏. 冠状病毒的新成员——SARS 的基因组特性. 遗传学报, 2003, 30(6): 501 ~ 508.

(责任编辑 李绍武)