第18卷4期 中国病毒学 18(4):335—339
2003年8月 VIROLOGICA SINIGA August 2003

SARS病毒与其他冠状病毒的基因组比较*

孙 啸，谢建明，周士新，谢雪英，陆祖宏

(东南大学生物科学与医学工程系，江苏南京，210096)

Genome Comparison Between *SARS-associated coronavirus* and Other *Coronaviruses*

SUN Xiao, XIE Jian-ming, ZHOU Shi-xin, XIE Xue-ying, LU Zu-hong

(Department of Biological Science & Medical Engineering, Southeast University, Nanjing 210096, China)

Abstract: The genome sequences of *SARS-associated coronavirus* (SARS-CoV) were compared with other *Coronaviruses*. Similar sequences were found by searching biomolecular databases. Sequence analysis showed that SARS-CoV had similar genomic organization and structural proteins while compared with other *Coronaviruses*. SARS-CoV is related to other known *Coronaviruses*. But there were some specific sequences in SARS-CoV genome. The virulence of SARS may derive from variations of ORF1a and S protein and the specificity of some nonstructural proteins. The result of word frequency analysis indicated that SARS-CoV does not closely resemble any of known *Coronaviruses*.

Key words: *SARS-associated coronavirus*(SARS-CoV); *Coronavirus*; Genomics; Sequence Analysis

摘要：本文利用生物信息学方法比较SARS病毒和其他冠状病毒基因组。通过数据库搜索，找出与SARS病毒基因组相似的核酸或蛋白质序列，并对相似序列进行比对，分析它们的共性和差异。结果表明，SARS病毒在基因组的组织上及结构蛋白质方面与现有冠状病毒有比较大的相似性，SARS病毒基因组与冠状病毒基因组相关。但是，SARS病毒基因组还存在一些特异性序列，ORF1a和S蛋白(特别是S1)的变化以及SARS-CoV特异性的非结构蛋白可能是SARS发病机理与传染特性区别于其他冠状病毒的分子基础。在全基因组水平上进行核酸单词出现频率分析，结果表明，SARS病毒远离已知的其他冠状病毒，单独成为一类。

关键词：SARS病毒（SARS-CoV)；冠状病毒，基因组学，序列分析

中图分类号：R511，R373 **文献标识码**：A **文章编号**：1003-5125(2003)04-0335-05

1 引言

公元2003年春季，一种呼吸系统疾病——非典型肺炎在全世界多个国家和地区流行，严重地威胁着人类的健康。世界卫生组织将其称为严重急性呼吸综合征（Severe Acute Respiratory Syndromes, SARS），该病原是冠状病毒的一个变种*SARS-associated coronavirus*(SARS-CoV)[1, 2, 3]。

冠状病毒的基因组全部集中在一条正向的RNA单链，其长度一般在3万个核苷酸左右。靠近基因组5'端有两个很大的开放阅读框（Open Reading Frame, ORF)，编码多聚蛋白(polyprotein)，其序列长度约占整个基因组的三分之二。这两个ORF分别为1a和1b，通过核糖体移码位点相互连接。ORF1a和ORF1b的翻译产物是多聚蛋白前体，由病毒蛋白酶在多个位点切割开。病毒蛋白酶在多聚蛋白基因的表达过程中起着突出的作用[4]。在ORF1b下游，有4到9个基因，它们分别编码冠状病毒的结构蛋白（如S、M、N及E蛋白）和非结构蛋白[5]。SARS-CoV属于冠状病毒。先后有多个SARS-CoV株的基因组被完全或部分测序[6, 7, 8]。从基因组结构来看，SARS-CoV具有典型的冠状病毒基因组的特点。我们以Tor2 SARS-CoV基因组序列（GenBank的登录号为AY274119）作为SARS基

收稿日期：2003-05-19，修回日期：2003-07-06
* 基金项目：国家高技术计划（863）资助项目（2002AA231071)；江苏省自然科学基金资助项目（BK2002057)。
作者简介：孙啸（1962-)，男，博士，主要从事生物信息学研究。Tel: 025-3795174. E-mail: xsun@seu.edu.cn.

因组的蓝本，利用生物信息学方法对它进行详细分析，通过生物分子序列数据库搜索和序列比对，研究 SARS-CoV 基因组与其他已知冠状病毒基因组的同源性，研究 SARS-CoV 的结构蛋白与其他冠状病毒的差异，同时利用我们提出的基因组核酸单词出现频率分析方法，分析与 SARS-CoV 相关的一组冠状病毒的进化关系。

2 SARS-CoV 基因组序列分析

利用美国生物信息研究所(NCBI)提供的 Blast 系列程序[9]搜索生物分子序列数据库，这些程序包括 Blastn、Blastp、Blastx 等。试图通过数据库搜索，发现与 Tor2 SARS-CoV 同源的序列，包括核酸序列和蛋白质序列。

取 Tor2 SARS-CoV 全基因组序列，用程序 Blastn 搜索核酸序列数据库，返回结果大部分是已知冠状病毒（如鼠科肝炎病毒）的小片段，其中最长的相似片段约 100nt。在某个已知的冠状病毒基因组中往往有多个显著相似的片段，例如，在鼠科肝炎病毒基因组中，各个显著相似片段长度的总和约为 300nt。将 SARS 病毒与已知冠状病毒基因组序列用程序 Blast2 进行两两比较，发现相似部分主要集中在 13kb 到 21kb 之间，即 SARS-CoV 的 ORF1b 区域，而其他区域中显著相似的序列片段极少。

从核酸序列数据库搜索和序列比较结果来看，只能说明 SARS 与冠状病毒有关，但是，由于没有发现较长的显著相似片段，故不能深入说明 SARS-CoV 与已知冠状病毒的关系。一般来说，蛋白质序列比核酸序列保守，通过蛋白质序列的比较，可能会发现种属之间更多的关系。根据 Tor2 SARS-CoV 基因组的注释，取 ORF1ab 翻译的多聚蛋白序列，其长度为 7073 个氨基酸，利用程序 Blastp 搜索蛋白质序列数据库，搜索结果见图 1。

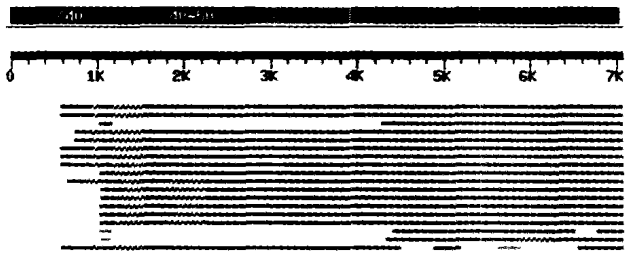

图 1 Tor2 SARS-CoV 多聚蛋白序列数据库搜索结果

Fig.1 Database search result of Tor2 SARS-CoV polyprotein

Each line represents a similar sequence comes from *Murine hepatitis virus, Bovine coronavirus, Porcine epidemic diarrhea virus, Transmissible gastroenteritis virus, Human coronavirus 229E or Avian infectious bronchitis virus.*

图 1 中带有刻度的横线代表 SARS-CoV 的蛋白序列，其下方的一系列横线表示在数据库中找到的与 SARS-CoV 多聚蛋白显著相似的蛋白质序列。搜索结果表明，SARS-CoV 的多聚蛋白序列与多个已知冠状病毒存在着大片段的显著相似，相似片段长度基本接近 SARS-CoV 多聚蛋白全序列的长度。对于 Tor2 SARS-CoV 基因组的其他蛋白质，我们也进行了序列比较，发现了一些显著的相似序列，结果将在下一节讨论。

为了进一步从基因组水平分析 SARS-CoV 与目前已知冠状病毒的关系，我们利用程序 Blastx，根据 Tor2 SARS-CoV 全基因组的核酸序列搜索蛋白质序列数据库，将基因组序列先按照不同的阅读框翻译成蛋白质序列，然后用翻译的序列搜索蛋白质序列数据库，搜索结果见图 2。从图 2 可以看出，通过数据库搜索找到许多与 SARS-CoV 基因有联系的冠状病毒基因组（见图 2 中的标注），其中显著相似的蛋白质序列分别是多聚蛋白（图中 0-21k）与 S 蛋白（图中标注），这两个蛋白质的编码序列占整个基因组长度的 5/6。另外，S 蛋白编码区域后面其他结构蛋白基因也存在着显著相似序列，但是由于这些序列相对比较短，用整个基因组搜索没有发现它们。取基因组的最后 5000nt 的片段去搜索数据库，则可以发现它们。BLASTx 的搜索结果说明，SARS-CoV 在基因组的组织上及结构蛋白质方面与现有冠状病毒有比较大的相似性，SARS-CoV 基因组与其他冠状病毒基因组相关。

然而，对照比较图 1 和图 2，可以发现，在 SARS-CoV 基因组中存在与其他冠状病毒截然不同的区域，这些区域在图 2 中以虚线箭头标注。取这些区域的共同部分，得到 1 个非相似区域，其位置处于基因组的 4137-4870，代表 SARS-CoV 基因组的一个特异性片段。根据 SARS-CoV 基因组核酸序列与其他冠状病毒蛋白质序列的比对结果，详细分析这个区域，可以发现：SARS-CoV 在该区域中核酸序列长度为 1kb 左右，但是，其他冠状病毒蛋白质序列对应此区域的长度比较短，这说明非相似区域可能是 SARS-CoV 基因组中插入的一段序列。

综合前面的序列比较结果、多聚蛋白对应的各个多肽比较结果以及下一节对结构蛋白的比较结果，我们可以得到 SARS-CoV 基因组同源序列的分布全貌。如图 3 所示，代表 SARS-CoV 基因组序列的坐标轴上方的横线显示了同源序列。其中，NSP1..NSP13 为多聚蛋白分解后对应的各个多肽。

图 3 从整个基因组显示了 SARS-CoV 与现有冠状病毒显著相似性，除基因组两端外，同源序列几乎覆盖了整个基因组。因此，SARS-CoV 来自于现有冠状病毒。

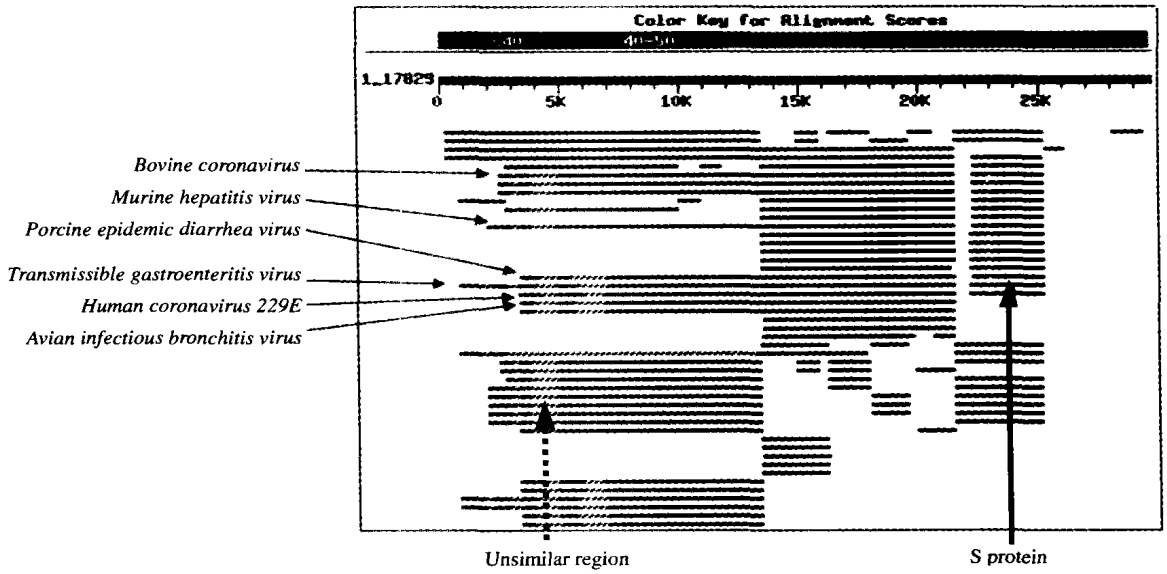

图 2 Tor2 SARS-CoV 全基因组 Blastx 搜索结果

Fig.2 Blastx search result of whole genome of Tor2 SARS-CoV

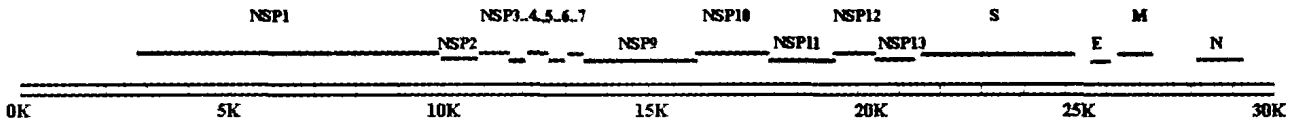

图 3 Tor2 SARS-CoV 全基因组同源序列分布

Fig.3 Outline of homologous sequences of Tor2 SARS-CoV genome

3 SARS-CoV 蛋白质同源性分析

多聚蛋白及结构蛋白在病毒进入细胞、病毒粒子形成和释放过程中起着重要的作用，这几个蛋白与 SARS-CoV 的传染性、毒性密切相关。根据 Tor2 SARS-CoV 基因组的注释，分别取得这些蛋白质的序列，然后利用程序 BLASTp 搜索蛋白质数据库，并分析所得到的序列比对结果。

ORF1a 中存在多个复制酶的亚单位，可能会较大的改变宿主细胞膜的结构[4]，而 ORF1b 与病毒的 RNA 合成密切相关[10]。虽然数据库序列比较结果表明这两段都与已知的冠状病毒的存在相似蛋白质序列，但是 ORF1b 相似程度更大，而 ORF1a 相似程度比较小。前面对基因组核酸序列的两两比较结果也说明 ORF1b 比较保守，而 ORF1a 变化较大。因此，SARS-CoV ORF1a 的变化可能导致复制酶与宿主细胞膜的结合性能发生变化，相对于其他冠状病毒可能会更大程度地改变宿主细胞膜的结构，由此产生较强的毒性。

S 蛋白主要有两个结构域，即 S1 和 S2。S1 可以自主地与细胞受体结合，而 S2 调节病毒与细胞膜的融合[11,12]。通过数据库搜索，发现 S 蛋白与鼠肝炎病毒最相似，整个蛋白质序列的相同部分达到 30%。同时发现 S 蛋白序列的后半部分与已知的冠状病毒 S2 保守序列（Corona_S2）高度相似(图 4A)，匹配的序列比对部分达到 90.7%。仔细观察针对 S 蛋白的数据库搜索结果还可以发现，S 蛋白后半部分序列与已知冠状病毒 S 蛋白的序列比对得较好，而前半部分比对得较差。说明 S 蛋白的 S1 部分变化较大。由于 S 蛋白(特别是 S1)与宿主细胞上的受体结合，该蛋白的变化将直接导致病毒的抗原性的变化，因此 SARS-CoV S 蛋白前半部分的变化可以解释两

个问题，即为什么人类目前对 SARS-CoV 的免疫力很低？为什么 SARS-CoV 具有比较高的毒性？

膜蛋白 M 与已知冠状病毒 M 蛋白保守序列 Corona_M 高度相似，匹配的序列比对部分达到 92.9%(图 4 B)。将 SARS-CoV 的 M 蛋白与其他冠状病毒的 M 蛋白进行比较，可以发现 M 蛋白与其他冠状病毒的相似部分可以达到 40%以上，其中，最大相似蛋白为猪血凝性脑脊髓炎病毒的 M 蛋白，相同的部分达到 42%。

核壳体蛋白 N 在冠状病毒的转录与复制过程中起着重要的作用[13]。N 蛋白与已知冠状病毒 N 蛋白保守序列 Corona_nucleoca 相似，序列比对分为两段(图 4 C)。SARS-CoV 的 N 蛋白与其他冠状病毒的相同部分一般为 33%左右，其中与鼠肝炎病毒 N 蛋白的相似程度最大，从第 15 个氨基酸到第 388 个氨基酸的这段长度为 374 的区域内，共有 145 个氨基酸是相同的，相同比率达到 36%。

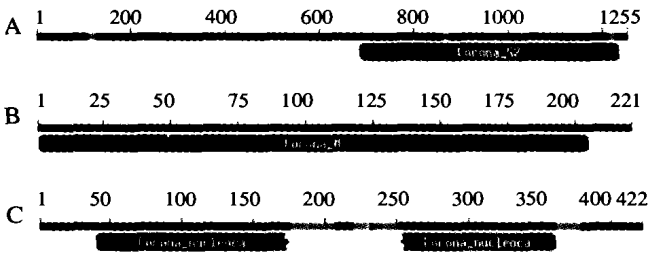

图 4 SARS-CoV 结构蛋白保守区域

Fig.4 Conserved regions of structural proteins of SARS-CoV

A, protein S; B, protein M; C, protein N.

E 蛋白是病毒包膜形成过程中的重要蛋白[14,15]。利用程序 BLASTp 搜索，没有发现显著相似的蛋白质。进一步利用程序 FASTa[16]搜索，发现 SARS-CoV 的 E 蛋白与一些已知冠状病毒的小膜蛋白相似，比对的部分覆盖全序列，相同部分为 28%-30%。

综上所述，SARS-CoV 的 M 蛋白最保守，S 蛋白的 S1 结构域变化最大。另外，相对于 ORF1b，ORF1a 变化也比较大。因此，SARS-CoV 对人类的危害性可能与 S1 和 ORF1a 的变化密切相关。

在 SARS-CoV 基因组有 5 个长度超过 150nt(50 个氨基酸）的 ORF，它们可能为 5 个非结构蛋白编码。对于这些 ORF，无论是用 BLASTn 搜索核酸数据库，还是用 BLASTp 搜索蛋白质数据库，都没有发现显著的相似序列。因此，这些非结构蛋白是 SARS-CoV 所特有的，可能是 SARS 发病机理与传染特性区别于其他冠状病毒的分子基础之一。

4 基因组组成分析

我们前期的研究结果表明，核酸单词的出现频率可以作为核酸序列的特征，可以将该方法用于 SARS 基因组的序列特征分析。因此，根据 Blastx 的搜索结果，针对所找到的与 SARS 相关的冠状病毒基因组进行分析，分析这些基因组在核酸单词出现频率方面的差别，由此判断 SARS-CoV 与已知冠状病毒的进化关系。

所谓核酸单词是由 k 个连续核苷酸组成的片段。在一个长度为 N 的核酸序列中，单词 w_i（i=0,...,4^k-1）的出现频率按下式进行计算：

$$p_{w_i} = \frac{N_{w_i}}{N-k+1}$$

其中，Nw_i 是单词 w_i 在整个序列中出现的次数。可以针对不同的 k 值进行统计分析，发现序列的组成特征。若 k 等于 3，则所有单词的出现频率形成一个 64 维向量：

$$(p_0, p_1, \ldots, p_{63})$$

以该向量反映序列的组成特点。

分析的对象分别是 SARS-CoV、鼠科肝炎病毒、人类冠状病毒、鸟类传染性支气管炎病毒、牛冠状病毒、猪流行痢疾病毒及可遗传的肠胃炎病毒，共 7 个完整基因组。计算三联核苷酸在各个基因组中的出现频率，以一个 64 维向量表示一个基因组的组成特点。计算结果表明，各基因组存在着明显差异。图 5(A)和(B)分别显示 SARS-COV 与牛冠状病毒基因组的单词出现频率，可以看出，两者差别比较大。

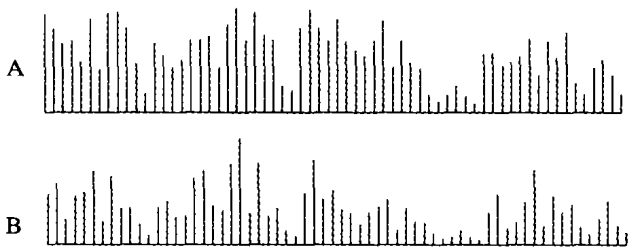

图 5 冠状病毒基因组三联核苷酸出现频率

Fig.5 Word frequency of two coronavirus genomes

Each vertical line represents the frequency of an nucleic acid word. A, *SARS-CoV*; B, *Bovine coronavirus*.

既然不同基因组的单词出现频率不一样，那么能否以单词出现频率代表基因组的特征，甚至借用单词出现频率分析基因组之间的进化关系呢？对已

知的 7 个冠状病毒基因组，分别计算它们长度为 3 的单词出现频率，并对所行成的 7 个 64 维向量进行聚类分析。图 6 是利用层次式聚类分析方法所得到的结果，这个结果反映 SARS-CoV 单独成为一类，远离其他的冠状病毒。我们还利用自组织特征映射神经网络（SOM）进行聚类分析，结果仍然表明 SARS-CoV 与其他冠状病毒截然分开，独成一类。

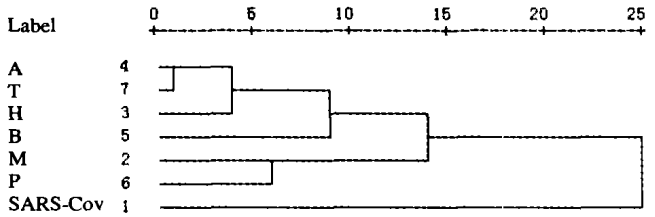

图 6 根据单词出现频率对冠状病毒基因组进行层次式聚类分析的结果

Fig.6 Clustering of coronaviruses according to their genomic word frequency

A-*Avian infectious bronchitis virus*; T-*Transmissible gastroenteritis Virus*; H-*Human coronavirus 229E*; B-*Bovine coronavirus*; M-*Murine hepatitis virus*; P-*Porcine epidemic diarrhea virus*.

5 结论

根据本文的分析，可以得到一些有意义的结论。首先，SARS-CoV 基因组与已知的冠状病毒基因组不仅在结构上一致，而且在多聚蛋白及其他结构蛋白质方面都存在显著的相似性，即对于 SARS-CoV 的主要蛋白质，都可以在其他冠状病毒中找到同源蛋白质。因此，SARS-CoV 在基因组的组织结构上及功能蛋白质方面与现有冠状病毒有比较大的相似性，SARS-CoV 基因组与冠状病毒基因组相关，SARS-CoV 来自于现有冠状病毒。然而，SARS-CoV 基因组还存在一些特异性序列，例如，基因组的 4137-4870 这个区段可能是在 SARS-CoV 基因组中插入的一个大片段。对于基因组 3'端编码非结构蛋白的 ORF，也没有找到显著相似的核酸或蛋白质。对于各蛋白质进行详细分析，发现 SARS-CoV 的 ORF1a 和 S 蛋白（特别是 S1）变化较大。因此，ORF1a 和 S 蛋白（特别是 S1）的变化以及 SARS-CoV 特异性的非结构蛋白可能是 SARS 发病机理与传染特性区别于其他冠状病毒的分子基础。分析 SARS-CoV 及相关冠状病毒基因组的单词出现频率，结果表明，SARS-CoV 远离已知的其他冠状病毒，单独成为一类。该结论与生物分类结果及基于蛋白质序列的系统发生分析结果一致。

参考文献

[1] Ksiazek T G, Erdman D, Goldsmith C S, *et al.* A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome[J]. N Engl J Med, 2003, 348(20): 1953-1966.

[2] Drosten C, Gunther S, Preiser W, *et al.* Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome[J]. N Engl J Med, 2003, 348(20): 1967-1976.

[3] Tsang K W, Ho P L, Ooi G C, *et al.* A Cluster of Cases of Severe Acute Respiratory Syndrome in Hong Kong[J]. N Engl J Med, 2003, 348(20): 1977-1985.

[4] Ziebuhr J, Snijder E J, Gorbalenya A E. Virus-encoded Proteinases and Proteolytic Processing in the Nidovirales[J]. J Gen Virol, 2000, 81(Pt 4): 853-879.

[5] De Vries A A F, Horzinek M C, Rottier P J M, *et al.* The genome organization of the Nidovirales: similarities and diâerences between arteri-, toro- and coronaviruses[J]. Seminars in Virology, 1997, 8:33-47.

[6] Rota P A, Oberste M S, Monroe S S, *et al.* Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome[J]. Science, 2003, 300: 1394-1399.

[7] Qin E, Zhu Q, Yu M, *et al.* A complete sequence and comparative analysis of strain (BJ01) of the SARS-associated virus[J]. Chinese Science Bulletin, 2003, 48(10): 941-948.

[8] Marra M A, Jones S J, Astell C R, *et al.* The Genome Sequence of the SARS-Associated Coronavirus[J]. Science, 2003, 300: 1399-1404.

[9] Altschul S F, Madden T L, Schaffer A A, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Res, 1997, 25: 3389-3402.

[10] van Dinten L C, Rensen S, Gorbalenya A E, *et al.* Proteolytic processing of the open reading frame 1b-encoded part of arterivirus replicase is mediated by nsp4 serine protease and is essential for virus replication[J]. J Virol, 1999, 73(3): 2027-2037.

[11] Gallagher T M, Buchmeier M J. Coronavirus spike proteins in viral entry and pathogenesis[J]. Virology, 2001, 279(2): 371-374.

[12] Krueger D K, Kelly S M, Lewicki D N, *et al.* Variations in disparate regions of the murine coronavirus spike protein impact the initiation of membrane fusion[J]. J Virol, 2001, 75(6): 2792-2802

[13] Baric R S, Nelson G W, Fleming J O, *et al.* Interactions between coronavirus nucleocapsid protein and viral RNAs: implications for viral transcription[J]. J Virol, 1988, 62(11): 4280-4287.

[14] Maeda J, Repass J F, Maeda A, *et al.* Membrane topology of coronavirus E protein[J]. Virology, 2001, 281(2): 163-169.

[15] Kuo L, Masters P S. The small envelope protein E is not essential for murine coronavirus replication[J]. J Virol, 2003, 77(8): 4597-4608.

[16] Pearson W R, Lipman D J. Improved tools for biological sequence comparison[J]. Proc Natl Acad Sci USA, 1988, 85(8): 2444-2448.