SARS-CoV 推测 N 蛋白功能结构的生物信息学研究*

刘树春[1**] 赵雨杰[2] 张 学[1] 罗 阳[1]
(1 中国医科大学医学基因组学研究室 沈阳 110001 2 中国医科大学基础医学院生物芯片中心 沈阳 110001)

摘要 目的: 利用生物信息学方法理论分析不同地区来源的 SARS 冠状病毒(SARS-CoV) 推断 N 蛋白的基因组与氨基酸序列的差异及分子生物学特征以及基因突变对蛋白结构功能的影响。方法: 针对 GenBank 上发布的来自不同国家地区的 15 条 SARS-CoV 基因组序列, 采用生物信息学软件分析其推测 N 蛋白的 CDS 和氨基酸序列, 分别找出突变位点并预测其等电点及功能结构域。结果: SARS-CoV 推测 N 蛋白基因组序列存在 5 个变异位点导致蛋白序列有 4 个位点发生突变。在该蛋白上发现四个有意义的低成分复杂性区域; 未发现卷曲螺旋、跨膜螺旋和信号肽序列。基因突变造成 4 条序列在功能位点数量上减少, 但未影响抗原决定簇。预测发现两个保守的 Domain 和一个丝氨酸富集区。结论: 不同地区来源的 15 条推测 N 蛋白序列的变异很少。基因突变导致部分序列功能位点数量发生改变, 但未影响抗原决定簇的数量。

关键词 严重急性呼吸综合征 冠状病毒 差异比较 核衣壳蛋白 生物信息学

SARS-CoV (severe acute respiratory syndrome coronavirus) 的 N 蛋白(nucleocapsid protein) 是冠状病毒中一种重要的结构蛋白, 是 SARS-CoV 的四个主要的结构蛋白之一[1]。在冠状病毒颗粒中, N 蛋白处于病毒颗粒的核心部分, 以和 Genomic RNA 结合的形式存在。以前对鼠肝炎病毒(murine hepative virus) 的研究表明, N 蛋白在病毒的包装过程中起着重要的作用[1)]。SARS 基因疫苗的设计是在病毒基因上寻找抗原决定簇, 再把抗原决定簇克隆后转到表达载体上, 再转入表达系统, 使其产生相应的抗原蛋白[2]。因此, 在 Genbank 数据库收录的不同地区来源的 SARS-CoV 全基因组序列中, 从其 N 蛋白的 CDS 及编码氨基酸序列入手分析该蛋白的结构特征, 找出不同地区来源的 SARS-CoV 的 N 蛋白的差异; 以及应用生物信息学理论分析基因突变对这些序列的功能位点、抗原决定簇等功能结构区的影响, 找出不同地区来源病毒株各 N 蛋白序列间在功能结构上的差异, 可以为以后的疫苗开发等提供依据。

1 材料与方法

从 GenBank 数据库(http://www.ncbi.nlm.nih.gov/) 中选取来自中国广州、中国浙江、中国上海、加拿大、越南、德国、意大利、中国香港、中国台湾等不同国家和地区的包括 SARS Coronavirus BJ01、GD01、TOR2、Ref、Urbani、Frankfurt_1、HSR_1、HKU-Su10、CUHK-W1、TW1、FRA、ZJ-HZ01、Shanghai_LY、Singapore NP 及 HPZ-2003 等 15 个 SARS-CoV 病毒株推测 N 蛋白的基因组和氨基酸序列, 利用 CLUSTAL X 1.81 软件进行列队比较, 找出并分析其变异与突变情况。

利用 DNATools 软件 6.0 版、DNASIS v2.5、NCBI Conserved Domain 等生物信息学软件及数据库进行蛋白质成分分析。并对编码的 N 蛋白进行分析, 寻找该蛋白的保守区、变异区、等电点、疏水性及蛋白的分子量等功能与结构特征。

在同一时间分别将这些病毒株 N 蛋白序列提交给因特网上蛋白质序列分析软件系统 The

收稿日期: 2003-10-28
* 辽宁省及中国医科大学 SARS 研究专项基金资助项目
** 电子信箱: scliu@mail.cmu.edu.cn
1) 石磊, 张其鹏, 黄伟, 等: SARS 冠状病毒 N 蛋白结构和功能初步分析. http://cmbi.bjmu.edu.cn/cmbidata/sars_secstructure/nucleocapsid%20protein.htm

Predict Protein Server (http://cubic.bioc.columbia.edu/predictprotein/predictprotein.html)、Predicting Antigenic Peptides(http://mif.dfci.harvard.edu/Tools/antigenic.html)、SMART 3.4(http://smart.ox.ac.uk/)、TMHMM Server v. 2.0(http://genome.cbs.dtu.dk/services/TMHMM-2.0/)、ProtParam Tools (http://cn.expasy.org/tools/protparam.html)等软件分析 SARS-CoV 的 N 蛋白的不稳定指数、消光系数等生物学特征,并预测各条 N 蛋白序列的 motif、低复杂度区域(low-complexity regions)、卷曲螺旋(Coiled coil)、跨膜螺旋(Transmembrane Helix, TMH)、信号肽及抗原决定簇(antigenic determinants)等结构功能域,比较不同地区来源的 N 蛋白的变异情况并分析基因突变对功能位点及抗原决定簇的影响。

表 1 15 个 SARS-CoV 推测 N 蛋白及其序列来源

protein name	Protein Gi number	Accession	Genome name	Position in Genome	Sequence source
nucleocapsid protein N	31416302	AAP51234.1	SARS coronvirus GD01	28133-29401	Guangzhou
nucleocapsid protein	29836503	NP_828858.1	SARS coronavirus Ref	28120-29388	Toronto
nucleocapsid protein	30795155	AAP41047.1	SARS coronavirus Tor2	28120-29388	Toronto
putative nucleocapsid protein	30027611	AAP13814.1	SARS coronavirus CUHK-W1	28105-29373	Hong Kong
putative nucleocapsid protein N	30698336	AAP37024.1	SARS coronavirus TW1	28120-29388	Taipei
putative nucleocapsid protein	30421455	AAP30714.1	SARS coronavirus CUHK-Su10	28105-29373	Hong Kong
nucleocapsid protein N	30275676	AAP30037.1	SARS coronavirus BJ01	28101-29369	Beijing
N protein	30027624	AAP13445.1	SARS coronavirus Urbani	28120-29388	Hanoi
nucleocapsid protein N	31581515	AAP33707.1	SARS coronavirus Frankfurt 1	28106-28111	Frankfurt
nucleocapsid protein N	32187355	AAP72984.1	SARS coronavirus HSR_1	28120-29388	Milano
nucleocapsid protein	33578028	AAP50495.1	SARS coronavirus FRA	28120-29388	Siena
nucleocapsid protein	32454352	AAP82974.1	SARS coronavirus Shanghai LY	755-2023	Shanghai
nucleocapsid protein	31505970	AAP44772.1	SARS coronavirus ZJ-HZ01	75-1343	Hangzhou
nucleocapsid protein	31540949	AAP49024.1	SARS coronavirus NP	1-1269	Singapore
nucleocapsid protein	34329619	AAQ63890.1	SARS coronavirus HPZ-2003	1-1269	Hangzhou

2 结 果

2.1 SARS-CoV 推测 N 蛋白的变异

来自不同国家和地区的 15 个推测 N 蛋白具有相同长度的 CDS,皆为 1 269bp,编码的 N 蛋白也具有相同的长度,均为 422 个氨基酸。利用 Clustal X 1.81 软件分别对不同地区来源的 15 个 N 蛋白的 CDS 和氨基酸序列进行列队比对分析,结果发现在 1269bp 的序列上有 5 个变异位点,包括 FRA、Frankfurt_1 和 Singapore_NP 序列在第 149 位碱基的变异(C→T);CUHK-Su10 序列在 577 位碱基的变异(G→T);Shanghai_LY 序列在第 792 位碱基(T→A)、第 974 位碱基(T→C)和第 976 位碱基的变异(A→G)。

由于 CDS 碱基的变异导致在 N 蛋白的 422 个氨基酸残基的序列中存在 4 个位点的突变,即 FRA、Frankfurt_1 和 Singapore_NP 三条序列的第 50 位氨基酸由 T 变为 I(苏氨酸→异亮氨酸);CUHK-Su10 序列的第 193 位氨基酸由 G 变为 C(甘氨酸→半胱氨酸);Shanghai_LY 序列在第 325 位氨基酸由 V 变为 A(缬氨酸→丙氨酸)以及在第 326 位氨基酸由 T 变为 A(苏氨酸→丙氨酸)。而 Shanghai_LY 序列在其基因组序列中第 792 位碱基的变异未导致其编码氨基酸的突变。

2.2 SARS-CoV 推测 N 蛋白结构特征的预测

利用 DNATools6 软件及 ExPASy 服务器上的网络生物信息学软件-ProtParam Tools 对 SARS-ref 推测 N 蛋白序列(gi_29836503_ref_NP_828858.1)进行分析,结果表明:SARS-CoV 推测 N 蛋白的 422 个氨基酸残基中包含 6393 个原子,其成分包括:碳原子(carbon)1985 个、氢原子(hydrogen)3150 个、氮原子(nitrogen)618 个、氧原子(oxygen)633 个、硫原子(sulfur)7 个。结构式为:$C_{1985}H_{3150}N_{618}O_{633}S_7$。该蛋白的分子量为 46025.0,理论 pI 值为 10.11,等电点为 10.93。估计半衰期分别为 30h(哺乳动物网状细胞,体外)、20h 以上(酵母菌,体内)和 10h 以上(大肠杆菌、体内);根据 Guruprasad[3] 方法确定该蛋白的不稳定指数(instability index, II)为 52.28,脂肪指数(aliphatic index)为 49.81。在标准条件下[4],当波长为 276nm 时,蛋白的消光系数(extinction coefficients)为 42950,光密度(optical density)为 0.933。溶解度为 94% 不可溶。

2.3 SARS-CoV 的 N 蛋白结构功能域分析

利用 SMART v3.4 (simple modular architecture research tool) 软件对 SARS-ref 推测 N 蛋白序列进行蛋白结构功能域分析。结果表明: 不同地区来源病毒株的 N 蛋白预测获得的低成分复杂度区、卷曲螺旋和跨膜蛋白结果没有差别, 即在所有 15 条 SARS-CoV 的 N 蛋白序列中, 皆发现四个有意义的低成分复杂度区域, 分别位于 176~ 207、220~ 231、233~ 251 和 362~ 379 位残基区间内; 预测未发现卷曲螺旋和跨膜螺旋序列。利用 TMHMM Server v. 2.0 软件和 PredictProtein 服务器进行 SARS-CoV 的 N 蛋白的跨膜螺旋预测, 也证实了 SARS-CoV 的 N 蛋白没有 TMH, 序列全部位于膜外。

利用 NCBI 的 Conserved Domain 数据库搜索获得两个保守的 domain, 分别位于 40~ 175 和 252~ 361 位氨基酸, 预测分值分别为 101 和 66.9。而利用 SignalP 预测结果表明推测 N 蛋白中不存在信号肽序列。

2.4 不同地区来源 SARS-CoV 的 N 蛋白中 Motif 分析

通过 PredictProtein 服务器的 PROSITE 的 motif (基序) 搜索, 在 15 条不同国家地区来源的 SARS-CoV 病毒株的 N 蛋白氨基酸序列中, 除了 Singapore _ NP、FRA、Frankfurt _ 1 和 CUHK-Su10 四条序列以外的 11 条序列预测得到的 motif 数目和序列片段皆相同, 包括 N-糖基化位点 (ASN _ GLYCOSYLATION) 2 个、黏多糖附着位点 (glycosaminoglycan) 2 个、cAMP-与 cGMP 依赖性蛋白激酶磷酸化位点 (CAMP _ PHOSPHO _ SITE) 2 个、蛋白激酶 C 磷酸化位点 (PKC _ PHOSPHO _ SITE) 9 个、酪蛋白激酶 II 磷酸化位点 (CK2 _ PHOSPHO _ SITE) 4 个、N-肉豆蔻酰化位点 (MYRISTYL) 9 个及酰胺化位点 (AMIDATION) 1 个, 合计 29 个。

而 Singapore _ NP、FRA、Frankfurt _ 1 三条序列比上述 11 个序列缺少一个位于第 48-51 位氨基酸的 N-糖基化位点。CUHK-Su10 较以上 11 条序列缺少一个位于 193~ 198 位氨基酸的 N-肉豆蔻酰化位点。

2.5 不同地区来源 SARS-CoV 的 N 蛋白的抗原决定簇的比较

利用 Predicting Antigenic Peptides 软件[5],2) 分别预测不同来源的 15 个 SARS-CoV 的 N 蛋白的抗原决定簇。15 个病毒株的 N 蛋白序列皆获得相同的 16 个抗原决定簇。N 蛋白的两个位点突变未发生在抗原决定簇片段上。

3 讨　论

3.1 SARS-CoV 的 N 蛋白的突变分析

通过比对可见, 虽然该 15 条 N 蛋白序列分别来自不同的国家或地区, 但其编码序列及编码蛋白质的长度皆完全一致, 而且这些序列的相似程度达到 99% 以上, 可以证实该蛋白推测的准确性并且全部来自于同一种病毒。和以前的研究 (本文首页脚注: 石磊, 等) 有所不同的是, 我们发现在 15 个不同地区来源病毒株推测 N 蛋白基因组序列间存在 5 个变异位点, 分别发生在 FRA、Frankfurt _ 1、Singapore _ NP、CUHK-Su10 和 Shanghai _ LY 等 5 条序列上。并且由于 CDS 的变异而导致其编码氨基酸序列发生了 4 处突变, 另有一处为沉默突变。虽然这种序列间的差异可能是测序的误差, 但也不能排除基因突变的结果。值得注意的是, 所发现的突变主要发生在 Shanghai _ LY 序列上, 包括 CDS 序列上 3 个位点 (792、974 和 976 位碱基) 和氨基酸序列上的 2 个位点 (326 和 792 位氨基酸)。可见, 虽然在 15 条 N 蛋白序列中有 10 条序列完全一致, 但随着病毒株的来源地区不同, 其 N 蛋白也存在少量突变。

3.2 SARS-CoV 的 N 蛋白结构与功能分析

SARS-CoV 的 N 蛋白的不稳定指数为 52.28, 确定该蛋白属于不稳定类蛋白。而可作为球蛋白耐热性上升的正指数的脂肪指数为 49.81。该蛋白的半衰期最长在体外可达到 30 个小时。

在 15 条不同地区来源病毒株的 N 蛋白序列中皆发现 4 个有意义的低成分复杂度区域, 而均未预测获得跨膜螺旋序列和卷曲螺旋序列, 这和其功能是相对应的 (本文首页脚注: 石磊, 等)。该蛋白也不存在信号肽序列。而利用 Prosite 的 motif 分析发现在第 176~ 207 位氨基酸之间存在一个丝氨酸富集区, 可能是磷酸化的重要区域。

3.3 基因突变对 SARS-CoV 的 N 蛋白的功能结构域及抗原决定簇的影响

在不同国家地区来源的 SARS-CoV 病毒株的 N 蛋白氨基酸序列中, 绝大多数序列间的功能位点数量是一致的。但在 29 个功能位点中, 有 2 个发生

2) Prediction of Antigenic Peptides WWW Server, http://mif.dfci.harvard.edu/Tools/antigenic.pl

了突变,造成 Singapore _ NP、FRA 和 Frankfurt _ 1 三条序列缺少一个 N-糖基化位点以及 CUHK-Su10 序列缺少一个 N-肉豆蔻酰化位点。这是前三条序列中第 50 位氨基酸突变和 CUHK-Su10 序列第 193 位氨基酸的突变所造成的。可见,由于个别序列发生了少量突变也导致其功能结构域的数量发生了少量变化。

但是,基因突变未发生在抗原决定簇片段上。因此也未导致 N 蛋白的抗原决定簇在数量和构成上发生变化。

4 结论

研究表明,SARS-CoV 的 N 蛋白属于不稳定类蛋白,在其序列的第 176~ 207 位氨基酸之间存在一个丝氨酸富集区,可能是磷酸化的主要区域。在其序列上存在两个保守 Domain。不同地区来源的 SARS-CoV 病毒株的推测 N 蛋白预测获得的低成分复杂度区、卷曲螺旋和跨膜蛋白、信号肽等结构没有差别,即在 N 蛋白序列中存在 4 个有意义的低成分复杂度区域,没有卷曲螺旋、跨膜序列和信号肽。但不同地区来源的序列间也存在少量突变,包括在其基因组序列上的 5 个位点变异而导致氨基酸序列上的 4 个位点的突变。基因突变也导致该蛋白在功能结构域上的改变。Singapore _ NP、FRA 和 Frankfurt _ 1 三条序列较其他序列缺少一个 N-糖基化位点以及 CUHK-Su10 序列缺少一个 N-肉豆蔻酰化位点。但突变对抗原决定簇的数量和构成未造成影响。

参考文献

[1] Ruan YJ, Wei CL, Ee AL, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet, 2003, 361(9371): 1779~ 1785

[2] 赵雨杰, 何群, 马佳明, 等. SARS 冠状病毒基因组及其所编码蛋白质生物信息学分析. 中国医科大学学报, 2003, 32(3): 193~ 195

[3] Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence. Protein Engineering, 1990, 4(2): 155~ 161

[4] Gill SC, von Hippel PH. Calculation of protein extinction coefficients from amino acid sequence data. Anal. Biochem, 1989, 182(2): 319~ 326

[5] Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett, 1990, 276(1-2): 172~ 174

Bioinformatic Analysis of Function and Structure of Putative Nucleocapsid Protein Sequences of SARS-CoV

Liu Shuchun[1] Zhao Yujie[2] Zhang Xue[1] Luo Yang[1]

(1 The Department of Medical Genome Research China Medical University Shenyang 110001)

(2 The Center of Biochip College of Basic Medical Sciences, China Medical University Shenyang 110001)

Abstract Objective: The objective of the paper was to analyze the characteristics of putative nucleocapsid protein sequence of Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and provide information for the further study of their structure and function. Methods: A series of bioinformatic tools and software such as Clustal X, DNATools, DNASIS, SMART, TMHMM Server, ProtParam Tools and so on were used to compare and analyze the SARS-CoV nucleocapsid protein CDS as well as coding protein sequences from GenBank. Results: The gene sequence of 15 different-sourced SARS-CoV putative nucleocapsid proteins has 5 variances resulting in 4 mutations occurred in its amino acid sequence. Four significant low compositional complexities were predicted and no coiled coil, TMH and signal peptide were found. Conclusion: The gene mutation resulted in the functional sites changed in some sequences and did not influence the antigenic determinants.

Key words Severe acute respiratory syndrome SARS Coronavirus Nucleocapsid protein Bioinformatics