SARS-CoV 编码蛋白质的识别

梅 虎 孙立力 周 原 熊 清 李志良 *

(重庆大学化学化工学院, 重庆 400044; 生物医学工程教育部与重庆市重点实验室, 重庆 400044; 重庆大学生物工程学院, 重庆 400044)

摘要 选取非典型肺炎病毒(SARS-CoV, NC 004718)和其他 6 种冠状病毒共计 100 个编码蛋白为样本, 从 172 个变量中经逐步回归选取 23 个变量建立多元线性回归模型, 模型的复相关系数 $R^2 = 0.645$, 交互检验 $R_{cv}^2 = 0.375$. 剔除 4 个异常样本后, 模型的复相关系数 $R^2 = 0.743$, 交互检验 $R_{cv}^2 = 0.543$.

关键词 SARS-CoV 冠状病毒 多元线性回归 多元线性逐步回归 编码蛋白 识别

冠状病毒(coronavirus)属于巢状病毒目(order *Nidovirales*)、冠状病毒科(family *Coronaviridae*)、冠状病毒属(genus *Coronavirus*)的一大类有蛋白质包被的单链正义 RNA 病毒(single-stranded, (+)sense RNA Virus), 具有 3 万个左右的核苷酸, 在已知 RNA 病毒的基因组中是最大的. 冠状病毒可以导致动物的许多疾病, 一些病毒例如传染性支气管炎病毒、猫的传染性腹膜炎病毒和传染性胃肠炎病毒都是典型的动物病原体^[1]. 人类的冠状病毒分属于第 1 类(HCoV-229E)和第 2 类(HCoV-OC43), 与 30% 温和性上呼吸道疾病有关. 已经证实 SARS-CoV (severe acute respiratory syndrome (SARS) associated coronavirus) 是一类新的冠状病毒, 是引发人类严重急性呼吸系统综合症的病原体^[2,3]. 目前对 SARS-CoV 的分子生物学研究主要集中在基因组结构、编码蛋白产物的空间结构预测、病毒的复制与转录过程以及致病机理等方面的研究. 而本文从一种新的角度出发, 选用反映氨基酸组成、物化和空间结构性质的一些参数, 对 SARS-CoV 和已知完整基因组序列的其他 6 种冠状病毒的 100 个编码蛋白样本, 采用统计分析和逐步回归的方法建立识别模型, 并在此基础上试图寻找 SARS-CoV 编码蛋白区别与其他冠状病毒编码蛋白的总体特征, 以期对 SARS-CoV 的蛋白质识别及相关基因组和蛋白质组研究提供一些有益的信息和参考依据.

1 原理与方法

SARS-CoV 编码蛋白区别于其他冠状病毒编码蛋白的特点不仅表现在氨基酸序列的差异上, 同时

也会在氨基酸组成、物化和空间结构性质上表现出一定的差异, 从而最终导致其生物学功能的差异. 鉴于此, 本文从上述性质入手, 以 SARS-CoV 和其他 6 种冠状病毒共计 100 个编码蛋白为样本通过逐步回归建立多元线性识别模型, 并从模型出发来探寻功能差异背后的性质差异. 与此同时, 运用该模型也可对 SARS-CoV 基因识别的结果进行辅助分析和判断.

2 样本来源和变量的选择

SARS-CoV (NC 004718) 的编码蛋白样本共计 30 个; 其他冠状病毒编码蛋白样本共计 70 个, 由已知完整基因组序列的 6 种冠状病毒的 140 个编码蛋白样本随机抽样而来. 这 6 种冠状病毒为 Group 1: 人类冠状病毒 229E (Human coronavirus 229E NC 002645)、猪的传染性腹泻病毒 (Porcine epidemic diarrhea virus NC 003436) 以及传染性胃肠炎病毒 (Transmissible gastroenteritis virus NC 002306); Group 2: 牛冠状病毒 (Bovine coronavirus NC 003045) 和鼠的肝炎病毒 (Murine hepatitis virus NC 001846); Group 3: 鸟类传染性支气管炎病毒 (Avian infectious bronchitis virus NC 001451)¹⁾. 参与建模的 100 个编码蛋白的详细资料参见附表 1²⁾.

运用逐步回归方法对反映编码蛋白物化和空间结构性质的 170 个变量进行筛选. 这 170 个变量主要反映以下的一些信息: 氨基酸频率和分子量, 蛋白质紫外吸收, 蛋白质的电性、疏水性、立体几何性质, 蛋白质的空间二级结构等. 逐步回归共筛选出 23 个变量 (表 1) 参与建模. 23 个变量的原始数据参见附表 2²⁾. 逐步回归和交互检验 (留一法) 采用 SPSS 11.0 统计分

1) <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>

2) 请向作者索取

<https://engine.scichina.com/doi/10.1360/csb2004-49-15-1503>

析软件包和本实验室自编软件完成.

表 1 多元线性回归模型中 23 个自变量

变量编号	变量描述
V_1	pmol/1 μg
V_2	A[280] 1 mg/mL ^{a)}
V_3	带电氨基酸(RKHYCDE)质量百分比
V_4	酸性氨基酸(DE)质量百分比
V_5	碱性氨基酸(KR)质量百分比
V_6	疏水性氨基酸(AILFWV)质量百分比
V_7	$\ln Z_1 ^{b)[4]}$
V_8	$\ln Z_3 ^{b)}$
V_9	$(\ln Z_1 ^{b)})^2$
V_{10}	侧链杂原子与碳原子的质量比率 ^{b)[5]}
V_{11}	摩尔份数 ^{b)[6]}
V_{12}	β -折叠频率 ^{b)[7]}
V_{13}	识别因子得分 ^{b)[8]}
V_{14}	$\ln A^{c)}$
V_{15}	$\ln F^{c)}$
V_{16}	$\ln G^{c)}$
V_{17}	$\ln H^{c)}$
V_{18}	$\ln L^{c)}$
V_{19}	$\ln N^{c)}$
V_{20}	$\ln P^{c)}$
V_{21}	$\ln Q^{c)}$
V_{22}	$\ln S^{c)}$
V_{23}	$\ln V^{c)}$

a) 该参数为 Vector NTI suite 8.0 生物信息软件包的预测值.
b) 变量的计算方法为文献中每个氨基酸的取值与其频率乘积的累加和.
c) 氨基酸频率(%)

3 结果与分析

运用 23 个自变量和 1 个因变量(SARS-CoV 编码蛋白的因变量定为 1, 非 SARS 的冠状病毒编码蛋白的因变量定为 0)对 100 个冠状病毒编码蛋白样本进行回归分析, 统计结果见表 2 的模型 1 部分. 结果表明: 模型 1 的稳定性较好, 预测能力较强. 其中 SARS-CoV 编码蛋白的识别正确率为 83.3%(回归值 0.5 为正确), 非 SARS-CoV 编码蛋白的识别正确率

为 95.7%(回归值<0.5 为正确); 交互检验 SARS-CoV 编码蛋白的识别正确率为 70.0%, 非 SARS-CoV 编码蛋白的识别正确率为 90.0%. 模型 1 的回归方程和各变量的偏相关系数(括号内为偏相关系数)如下:

$$Y = 54.085 + 0.002V_1 + 1.115V_2 - 0.209V_3 + 0.187V_4 + 0.283V_5$$
$$- 0.192V_6 - 0.162V_7 + 0.408V_8 + 0.045V_9 + 0.183V_{10}$$
$$+ 0.013V_{11} + 0.469V_{12} - 0.014V_{13} + 0.758V_{14} + 0.398V_{15}$$
$$- 0.747V_{16} + 0.635V_{17} + 2.281V_{18} + 0.670V_{19} + 1.202V_{20}$$
$$+ 0.396V_{21} + 0.641V_{22} - 0.730V_{23}$$

(0.29) (0.49) (-0.50) (0.51) (0.47) (-0.52) (-0.36) (0.44) (0.37) (0.58) (0.30) (0.57) (-0.51) (0.42) (0.27) (-0.39) (0.49) (0.55) (0.50) (0.59) (0.35) (0.29) (-0.30)

将模型 1 的残差值与样本号作图, 发现样本号为 19, 53, 74, 79 的残差相对较大, 远远偏离其他样本(图 1). 将各样本的 Cook 距离对中心化杠杆值作图可以看出, 这四个样本虽然对回归估计影响相对较大, 但是对模型的斜率影响相对较小; 而第 28 号样本同时具有较大的 Cook 值和中心杠杆值, 为模型的影响点(图 2). 剔除前 4 个样本并赋予 28 号样本较小的权重(0.5), 我们重新建立多元线性回归方程, 统计结果见表 2 的模型 2 部分. 不难看出模型 2 的回归和预测效果都得到了较大的提高, 其中 SARS-CoV 编码蛋白的识别正确率为 86.2%, 非 SARS-CoV 编码蛋白的

图 1 各样本的残差分布图

表 2 多元线性回归模型及其交互检验的统计结果

模型	R	R^2	SD	U	Q	F	R_{cv}	R_{cv}^2	SD_{cv}	U_{cv}	Q_{cv}	F_{cv}
1	0.803	0.645	0.313	13.550	7.450	6.010	0.612	0.375	0.415	7.879	13.121	1.984
2	0.862	0.743	0.268	15.044	5.196	9.064	0.737	0.543	0.358	10.987	9.253	3.717

图 2 Cook 距离-中心化杠杆值的关系

识别正确率为 100.0%; 交互检验 SARS-CoV 编码蛋白的识别正确率为 86.2%, 非 SARS-CoV 编码蛋白的识别正确率为 94.0%. 模型 2 的多元线性回归方程和各变量的偏相关系数(括号内为偏相关系数)如下:

$$\begin{aligned}
 Y = & 51.518 + 0.002V_1 + 1.152V_2 - 0.215V_3 + 0.198V_4 + 0.284V_5 \\
 & (-0.32) \quad (0.57) \quad (-0.58) \quad (0.60) \quad (0.54) \\
 & - 0.205V_6 - 0.228V_7 + 0.413V_8 + 0.058V_9 + 0.181V_{10} \\
 & (-0.61) \quad (-0.50) \quad (0.51) \quad (0.49) \quad (0.65) \\
 & + 0.015V_{11} + 0.466V_{12} - 0.014V_{13} + 0.748V_{14} + 0.431V_{15} \\
 & (0.39) \quad (0.64) \quad (-0.57) \quad (0.47) \quad (0.34) \\
 & - 0.745V_{16} + 0.719V_{17} + 2.418V_{18} + 0.689V_{19} + 1.263V_{20} \\
 & (-0.45) \quad (0.58) \quad (0.64) \quad (0.57) \quad (0.67) \\
 & + 0.445V_{21} + 0.648V_{22} - 0.744V_{23} \\
 & (0.45) \quad (0.39) \quad (-0.34)
 \end{aligned}$$

通过各自变量在模型 2 中的偏相关系数, 我们可以就每个自变量对模型的影响作一些初步的分析. 在 23 个变量中 V_{10} , V_{12} , V_{18} , V_{20} 这四个变量的偏相关系数均大于 0.60, 它们分别为侧链中杂原子对 C 原子的质量比、 β -折叠频率、亮氨酸频率的自然对数值、脯氨酸频率的自然对数值, 由此我们推断以上性质对 SARS-CoV 编码蛋白与其他 6 种冠状病毒编码蛋白存在较大的影响, 高的端基杂原子质量比、 β -折叠频率、亮氨酸和脯氨酸频率有形成 SARS-CoV 编码蛋白的倾向. 其中 V_{20} 即脯氨酸频率的自然对数值的偏相关系数绝对值为 0.67, 在所有的 23 个变量中最大, 这一点引起了我们的注意. 我们知道脯氨酸是非常有特点的一个氨基酸, 它与其他氨基酸形成酰胺键有明显的特点, 较易变成顺式的肽键. 顺式的肽

键由于大基团间的作用, 而处于高能态, 不稳定, 同时也会影响到肽链的空间构象. 蛋白质的变性和复性以及新生肽链折叠时发生的一些肽键的顺反异构体的互变, 也往往是动力学上的限速的一步. 于是我们又想到了甘氨酸, 甘氨酸是所有氨基酸中惟一在 α 碳原子上只有两个氢原子而没有侧链的氨基酸, 所以它和其他残基间的相互作用相对较小, 也不会产生大的空间位阻. 有趣的是, 甘氨酸频率的自然对数值(V_{16})也出现在模型中, 且偏相关系数为 -0.45, 这是否是偶然的巧合还有待进一步的研究. 对模型有负性影响的主要有 3 个变量, 其偏相关系数绝对值均大于 0.50, 分别为带电氨基酸的质量百分比(V_3)、疏水性氨基酸质量百分比(V_6)和识别因子得分(V_{13}). 说明这三个变量对 SARS-CoV 编码蛋白的形成不利. 通过分析 V_3 的偏相关系数, 我们发现高的 D, E, K, R, H, Y, C 含量对 SARS-CoV 编码蛋白的形成不利, 但是结合到 V_4 , V_5 , V_{17} 的偏相关系数我们又发现高的 D, E, K, R, H 含量却对 SARS-CoV 编码蛋白的形成有利. 惟一合理的解释便是高含量的 Y(酪氨酸)和 C(半胱氨酸)可能对 SARS-CoV 编码蛋白的形成极为不利, 以至将 D, E, K, R, H 对模型的影响所“屏蔽”. 其中半胱氨酸又是非常特殊的一个, 它可以通过在肽链内和肽链间形成二硫键的方式影响蛋白质的空间构象. 同时它又具有高度的化学反应活性, 可以进一步影响蛋白质的一些生物学性质. 其他变量偏相关系数在此不作分析.

通过以上的多元线性回归模型, 我们可以对 SARS-CoV 编码蛋白与其他冠状病毒编码蛋白在物化性质和一些二级结构特征上的总体区别有个初步的认识. 同时希望本文能够对 SARS-CoV 的蛋白质和基因识别以及基因组和蛋白质组研究提供一些有益的信息和参考依据.

此外, 我们运用该模型对 SARS-CoV BJ01 (AY278488) 基因预测的结果^[9]进行了分析. 在 SARS-CoV BJ01 的 35 个 ORFs (Open Reading Frame, 开放阅读框) 中, 有 14 个 ORFs 为已知 ORFs, 另有 21 个为未知 ORFs, 可能为新基因. 对于 21 个未知的 ORFs, 我们将启发式模型 (Heuristic models)、基因识别 (Gene Identification)、开放阅读框识别 (ORF Finder) 3 种基因预测模型的结果和利用模型 2 进行多元线性回归的预测结果列入表 3 以供参考.

<https://engine.scichina.com/doi/10.1360/csb2004-49-15-1503>

表 3 SARS-CoV BJ01 基因组中 21 个未知 ORFs 的基因预测结果和多元线性回归预测结果

ORF	编码位置(bp)	氨基酸长度	开放阅读框识别	基因识别	启发式模型	多元线性回归预测结果 ^{a)}	多元线性回归预测值
ORF2	715~1216	163	+	—	—	+	0.80
ORF3	952~1216	84	+	—	—	+	0.72
ORF4	2974~3276	100	+	+	—	—	-0.36
ORF5	3547~3762	71	+	—	—	—	0.03
ORF6	3883~4035	50	+	—	—	—	0.28
ORF7	7480~7665	61	+	—	—	—	0.01
ORF8	8572~8736	54	+	—	—	+	0.78
ORF9	8809~8994	61	+	+	—	+	1.67
ORF10	10048~10311	87	+	+	—	+	1.43
ORF11	11002~11157	51	+	+	—	+	0.70
ORF12	12406~12570	54	+	—	—	—	0.10
ORF13	12670~12852	60	+	—	—	—	-0.29
ORF14	12676~12852	58	-	+	—	—	-0.28
ORF15	13580~21466	2628	+	+	+	+	1.00
ORF16	14169~14339	56	+	+	—	—	0.15
ORF17	16494~16676	60	+	—	—	+	0.93
ORF18	20502~20690	62	+	—	—	—	-1.13
ORF20	22713~22907	64	+	+	—	—	0.39
ORF21	24138~24296	52	+	—	—	+	1.57
ORF22	24591~24761	56	+	—	—	—	-0.19
ORF23	24798~24998	66	+	—	—	+	1.38

a) 多元线性回归预测值 0.5 判定为 SARS-CoV 编码蛋白质, <0.5 判为非 SARS-CoV 编码蛋白质

致谢 本工作为国家“春晖计划”教育部启动基金(批准号: 99-04+99-37)、霍英东青年教师基金(批准号: 98-7-6)、重庆市应用基础基金(批准号: 01-3-6)及重庆大学研究生院抗“非典”科技攻关 (批准号: 03-03)资助项目.

参 考 文 献

1 Rota P A, Oberste M S, Monroe S S, et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 2003, 300(5624): 1394~1399[DOI]

2 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(20): 1967~1976[DOI]

3 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(20): 1953~1966[DOI]

4 Hellberg S, Sjostrom M, Skagerberg B, et al. Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem*, 1987, 30(7): 1126~1135

5 Grantham R. Amino acid difference formula to help explain protein evolution. *Science*, 1974, 185(4154): 862~864

6 Janin J. Surface and inside volumes in globular proteins. *Nature*, 1979, 277(5696): 491~492

7 Levitt M. Conformational preferences of amino acids in globular proteins. *Biochemistry*, 1978, 17(20): 4277~4285

8 Fraga S, San-Fabian E, Thornton S, et al. Prediction of the secondary structure and functional sites of major histocompatibility complex molecules. *J Mol Recognit*, 1990, 3(2): 65~73

9 贺福初主编. SARS——严重急性呼吸系统综合症. 北京: 科学出版社, 2003. 61~69

(2003-11-07 收稿, 2004-06-17 收修改稿)