Chinese Science Bulletin 2004 Vol. 49 No. 19 2037—2040

Identification of encoding proteins related to SARS-CoV

MEI Hu[1,2,3], SUN Lili[2,3], ZHOU Yuan[1,2], XIONG Qing[2,3] & LI Zhiliang[1,2]

1. College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China;
2. Key Laboratory of Biomedical Engineering of Ministry of Education and Chongqing City, Chongqing 400044, China;
3. College of Bioengineering, Chongqing University, Chongqing 400044, China

Correspondence should be addressed to Li Zhiliang (e-mail: zlli2662@163.com)

Abstract **By sampling 100 encoding proteins from SARS-coronavirus (SARS-CoV, NC 004718) and other six coronaviruses and selecting 23 variables through stepwise multiple regression (SMR) from 172 variables, the multiple linear regression (MLR) model was established with good results of the quantitative modelling correlation coefficient R^2 = 0.645 and the cross-validation correlation coefficient R^2_{CV} = 0.375. After removing 4 outliers, the quantitative modelling and cross-validation correlation coefficients were R^2 = 0.743 and R^2_{CV} = 0.543, respectively.**

Keywords: SARS-CoV, coronavirus, multiple linear regression (MLR), stepwise multiple regression (SMR), encoding protein, identification.

DOI: 10.1360/03wb0198

The coronaviruses (order *Nidovirales*, family *Coronaviridae*, genus *Coronavirus*) are a diverse group of large, enveloped, positive-stranded RNA viruses. At ~30000 nucleotides, their genome is the largest found in any of the RNA viruses. The viruses can cause severe disease in many animals, and several viruses, including infectious bronchitis virus, feline infectious peritonitis virus, and transmissible gastroenteritis virus, are significant veterinary pathogens[1]. Human coronaviruses are found in both group 1 (HCoV-229E) and group 2 (HCoV-OC43) and are responsible for 30% of mild upper respiratory tract illnesses. The severe acute respiratory syndrome (SARS) associated with coronavirus (SARS-CoV) has proved to be a new group and the prime criminal for SARS infection[2,3]. At present, researches in molecular biology related to SARS-CoV are mainly focused on genome organization, virus replication, transcriptions, pathogenesis and protein structure prediction. In this paper, by stepwise multiple regression (SMR), 23 variables are selected to establish multiple linear regression (MLR) model from 172 variables which are mainly about amino acids constitutions, physicochemical and 3-D structural properties of 100 encoding proteins from SARS-CoV and other 6 coronaviruses. Form the model established, we can derive some overall characters, which distinguish the SARS-CoV from other coronaviruses, and can provide some valuable information for protein recognition, genome approaches and promote researches.

1 Principle and methodology

The characteristics that distinguish the encoding proteins of SARS-CoV from those of other coronaviruses are related to not only the sequence of amino acids but also amino acids constitution, physicochemical and 3-D structural properties. These characteristics finally induce the functional difference between them. So, here we make a systematic study on 172 variables, which mainly describe amino acids constitution, physicochemical and 3-D structural properties of 100 encoding proteins from SARS-CoV and other 6 coronaviruses. By stepwise multiple regression, the model is established from which we can deduce the most important variables that contribute significantly to the functional difference between encoding proteins from SARS-CoV and other 6 coronaviruses.

2 Selection of samples and variables

Total 100 encoding proteins, i.e. 30 of SARS-CoV (NC 004718) and 70 of other 6 coronaviruses, are used as calibration samples (please refer to supplemental materials for details). The latter 70 samples are randomly selected from 140 encoding proteins of 6 coronaviruses. The 6 coronaviruses comprise Group 1: human coronavirus 229E (NC 002645), porcine epidemic diarrhea virus (NC 003436) and transmissible gastroenteritis virus (NC 002306); Group 2: bovine coronavirus (NC 003045) and murine hepatitis virus (NC 001846) and Group 3: avian infectious bronchitis virus (NC 001451)[1)]. The original 172 variables mainly described the amino acids frequency, molecular weight, violet absorbing, hydrophobicity, bulk, and electronic properties, 3-D structural properties of encoding proteins. By stepwise multiple regression, 23 variables (available as supplemental materials) are selected to establish the calibration model (Table 1). The stepwise multiple regression and cross-validation technique with the leave-one-out procedure are performed with SPSS 10.0 package and an in-house program, respectively.

3 Results and discussion

For two classes of encoding proteins under consideration, we use a categorical variable Y where one class is set to 1 for the encoding proteins of SARS-CoV and the other is set to 0 for the others. Then, the categorical variables together with the original 172 variables are modeled by stepwise multiple regression. The results are shown in model 1 (Table 2). From the results of model 1, we can see that the model has both the modeling robustness and pre-

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

Table 1 The 23 independent variables used in the multiple linear regression modeling processes

No.	Variable description
V_1	pmol/μg
V_2	A[280] of 1 mg/mL[a)]
V_3	weight percentage of charged AA (RKHYCDE)
V_4	weight percentage of acidic AA (DE)
V_5	weight percentage of basic AA (KR)
V_6	weight percentage of hydrophobic AA (AILFWV)
V_7	ln $\|Z_1\|$[4]b)]
V_8	ln $\|Z_3\|$[b)]
V_9	$(\ln\|Z_1\|)^2$ [b)]
V_{10}	atomic weight ratio of hetero elements in end group to C in side chain[5] b)]
V_{11}	molar fraction (%) of 2001 buried residues[6] b)]
V_{12}	conformational parameter for beta-sheet[7] b)]
V_{13}	recognition factors[8] b)]
V_{14}	ln A[c)]
V_{15}	ln F
V_{16}	ln G
V_{17}	ln H
V_{18}	ln L
V_{19}	ln N
V_{20}	ln P
V_{21}	ln Q
V_{22}	ln S
V_{23}	ln V

a) Predicted by vector NTI suite 8.0 package, b) the values of variables are weighted mean by frequency, c) natural logarithm of the frequency of amino acids (V14—V23).

Table 2 The summary results of multiple linear regression and cross validation procedures

Model	R	R^2	SD	U	Q	F	R_{CV}	R^2_{CV}	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

dictive capability. For the encoding proteins of SARS-CoV, the calculated values, which are equal to or larger than 0.5, are thought to be predicted correctly, or else incorrectly. Conversely, for the encoding proteins of other 6 coronaviruses, the calculated values, which are less than 0.5, are thought to be predicted correctly, or else incorrectly. So when model 1 is considered, the percentage of correct prediction for SARS-CoV and other 6 coronaviruses samples are 83.3% and 95.7% respectively and the percentage in cross-validation results are 70.0% and 90.0% respectively. The equation of model 1 and partial correlation coefficient of each variable are listed below:

$$\begin{aligned} Y = {} & 54.085 + \underset{(0.29)}{0.002V_1} + \underset{(0.49)}{1.115V_2} - \underset{(-0.50)}{0.209V_3} + \underset{(0.51)}{0.187V_4} + \underset{(0.47)}{0.283V_5} - \underset{(-0.52)}{0.192V_6} - \underset{(-0.36)}{0.162V_7} + \underset{(0.44)}{0.408V_8} + \underset{(0.37)}{0.045V_9} + \underset{(0.58)}{0.183V_{10}} \\ & + \underset{(0.30)}{0.013V_{11}} + \underset{(0.57)}{0.469V_{12}} - \underset{(-0.51)}{0.014V_{13}} + \underset{(0.42)}{0.758V_{14}} + \underset{(0.27)}{0.398V_{15}} - \underset{(-0.39)}{0.747V_{16}} + \underset{(0.49)}{0.635V_{17}} + \underset{(0.55)}{2.281V_{18}} + \underset{(0.50)}{0.670V_{19}} + \underset{(0.59)}{1.202V_{20}} \\ & + \underset{(0.35)}{0.396V_{21}} + \underset{(0.29)}{0.641V_{22}} - \underset{(-0.30)}{0.730V_{23}}. \end{aligned}$$

When plotting residual against observation ID (Fig. 1), we find that residuals of the observations labeled as 19, 53, 74, 79 are relatively very large and the absolute standardized residuals of these 4 samples are 2 times larger than standard deviation. So we identify these 4 observations as outliers. To measure the influence of a point on regression fitness, we plot cook's distance against centered leverage values (Fig. 2), from which we can see that the observation labeled as 28 has a high leverage and high influence. Its high leverage gives it extra weight in the computation of the regression line, and the high influence indicates that it did affect the slope of the regression line. So we examine this influential point by using a weighting variable (0.5) that gives the influential point less weight. After removing 4 outliers and giving observation 28 less weight, we rebuilt the multiple linear regression model,

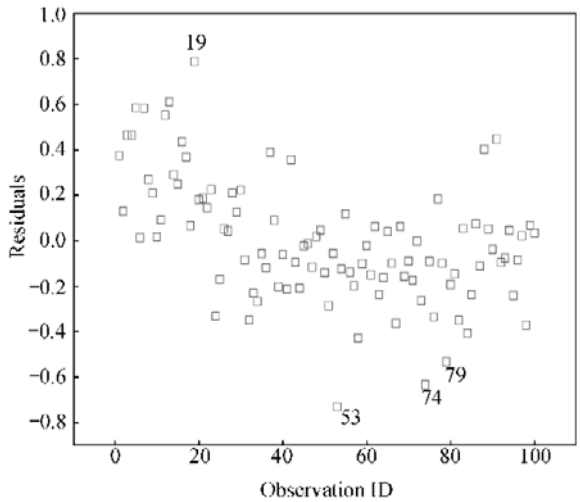

Fig. 1. The distribution of residuals in observations.

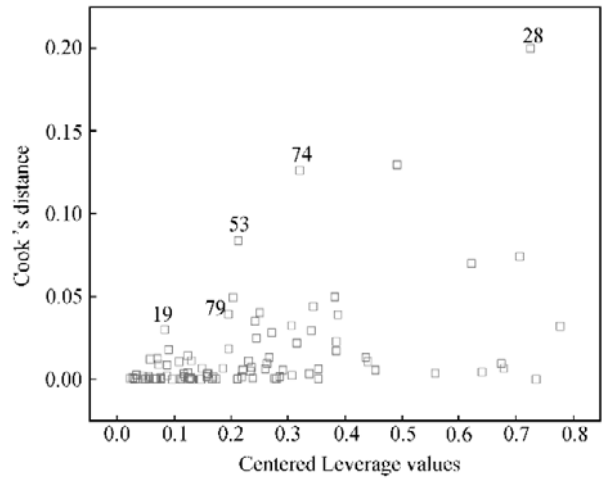

Fig. 2. Cook's distance vs. Centered Leverage values.

the results of which are shown in model 2 (Table 2). As we anticipate, the robustness and predictive capability of model 2 are superior to that of model 1. The percentage of correct prediction for SARS-CoV and other 6 coronaviruses samples are 86.2% and 100.0% respectively and the percentage in cross-validation results are 86.2% and 94.0% respectively. The equation of model 2 and partial correlation coefficient of each variable are listed below:

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

From the partial correlation coefficient, we can obtain some useful information, i.e. the direction and strength of relationships between X and Y variables. In model 2, there are 4 independent variables, whose partial correlation coefficients are larger than 0.6. These 4 independent variables are as follows: atomic weight ratio of hetero elements in end group to C in side chain (V_{10}), conformational parameter for beta-sheet (V_{12}) and natural logarithm of the frequency for both leucine (V_{18}) and proline (V_{20}). The high partial correlation coefficient indicates that these 4 variables give more influence on the encoding protein of SARS-CoV and other 6 coronaviruses. The positive correlation indicates that higher values of these 4 variables tend to form encoding proteins of SARS-CoV. However, there is an interesting thing to note, the absolute value of partial correlation coefficient of V_{20} (the natural logarithm of the frequency of proline) is the largest among 23 variables. As we know, proline is very particular amino acids. When formed amide linkage or peptide bonding with other amino acids, it can easily form cisoid conformation which can cause increased interactions among groups and influence the conformation of backbone at the same time. In the process of protein denaturalization and renaturalization as well as folding of peptide chain, it is a restrictive step of dynamics for converting from transoid conformation to cisoid conformation. Then, how about glycin? Glycin is only amino acids with 2 hydrogen atoms linking to its alpha-carbon atom. So for glycin, there are fewer interactions among groups as well as less steric hindrance. Interestingly, the natural logarithm of the frequency of glycin is also included in model 2, of which the partial correlation coefficient is –0.45. Further investigation is required for reasonable explanation about that. There are 3 variables in model 2 of which the partial correlation coefficients are less than –0.50. These 3 variables are as follows: weight percentage of charged amino acids (V_3), weight percentage of hydrophobic amino acids (V_6), and recognition factors (V_{13}). The negative correlation indicates that higher absolute values of these 3 variables tend to give more negative impact on the tendency of forming encoding proteins of SARS-CoV. The partial correlation coefficients of V_3 indicate that low weight percentage of charged amino acids, i.e. R, K, H, Y, C, D and E tends to form encoding protein of SARS-CoV. However, the partial correlation coefficients of V_4, V_5 and V_{17} indicate that high weight percent

Table 3 Results of gene identification and multiple linear regression prediction in 21 unknown ORFs of SARS-CoV BJ01 genome

ORF	Genome location/bp	Number of amino acids	ORF finder	Gene identification	Heuristic models	Multiple linear regression[a)]	Predicted value of multiple linear regression
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) The calculated values, which are equal to or larger than 0.5, are thought to be predicted correctly, or else incorrectly.

age of acidic amino acids (D, E), basic amino acids (K, R) and H tends to form encoding protein of SARS-CoV. The only reasonable solution is that high weight percentage of Y and C tend to give extremely negative impact on the tendency of forming encoding proteins of SARS-CoV which mask the influence of D, E, K, R and H. As for cysteine (C), it is another particular amino acids, which can influence the conformation of protein by form disulfide bond. Furthermore, cysteine has relatively high reaction activity, which can affect the biological properties of proteins.

From model 2, we can get overall characteristics about difference between encoding proteins of SARS-CoV and those of other 6 coronaviruses. However, for more information about it, further investigation is required. Besides, we compare the results of gene recognition related to SARS-CoV BJ01[9]. In the genome of SARS-CoV BJ01, there are 35 Open Reading Frames (ORFs), in which 14 ORFs are identified and other 21 ORFs are not confirmed, which maybe are new genes. Table 3 lists the results of 3 gene recognition approaches, i.e. Heuristic models, Gene Identification, ORF Finder together with established model 2.

Acknowledgements This work was supported by the National Chuihui Project Foundation (Grant No. 99-04+99-37), the Fok-Yingtung Educational Foundation (Grant No. 98-7-6), the Chongqing Applied Fundamental Science Fund (Grant No. 01-3-6), and Chongqing University Innovation Foundation of Science and Technology (Grant No. 03-03).

References

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.
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.
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.
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. He, F. C., SARS——the Severe Acute Respiratory Syndrome (in Chinese), Beijing: Science Press, 2003, 61—69.

(Received November 7, 2003; accepted June 17, 2004)