J Mol Model (2005) 11: 8–16
DOI 10.1007/s00894-004-0210-0

ORIGINAL PAPER

Guang Wu · Shaomin Yan

Reasoning of spike glycoproteins being more vulnerable to mutations among 158 coronavirus proteins from different species

Received: 18 March 2004 / Accepted: 30 August 2004 / Published online: 9 December 2004

Abstract In this study, we used the probabilistic models developed by us over the last several years to analyze 158 proteins from coronaviruses in order to determine which protein is more vulnerable to mutations. The results provide three lines of evidence suggesting that the spike glycoprotein is different from the other coronavirus proteins: (1) the spike glycoprotein is more sensitive to mutations, this is the current state of the spike glycoprotein, (2) the spike glycoprotein has undergone more mutations in the past, this is the history of spike glycoprotein, and (3) the spike glycoprotein has a bigger potential towards future mutations, this is the future of spike glycoprotein. Furthermore, this study gives a clue on the species susceptibility regarding different proteins.

Keywords Coronavirus · Protein · Probability · SARS

Introduction

With the occurrence of new cases of severe acute respiratory syndrome (SARS), the prognosis of a possible return of SARS in the near future is coming true. Also hypothesis that the new SARS cases could be somewhat different from the previous SARS cases in possible mutated forms appears to be true. Accumulating evidence shows that there are mutations in the SARS-related coronavirus (SARS-CoV), [1, 2] which may lead to difficulties in diagnosis, treatment, and prevention.

The SARS-CoV is an enveloped RNA virus. Naturally, we would expect that the different components in human SARS-CoV would have different sensitivities to mutation, therefore it would minimize the difficulties in identification of SARS-CoV and facilitate diagnosis, treatment and prevention of SARS if we could identify which component of human SARS-CoV is most subject to mutations. Doubtlessly we should not limit ourselves to sole SARS-CoV, not only because many species carry coronaviruses [3, 4], but also, more importantly, because the coronavirus from civets is likely to be the source of SARS [5].

Among various components in coronavirus, we are more interested in the proteins, because over the last several years we have developed three models to analyze the protein primary structure (for a review, see [6]), including the proteins from SARS-CoV [7, 8]. In general, our first model can classify a protein into the randomly predictable and unpredictable portions, and our findings demonstrate that the unpredictable portion is more sensitive to mutations than the predictable one. Thus, we can find which protein is more vulnerable to mutations by comparing the unpredictable portion with the predictable one among proteins.

So far the envelope protein, hemagglutinin-esterase precursor, membrane glycoprotein, nonstructural protein, nucleocapsid protein, spike glycoprotein, replicase polyprotein and hypothetical proteins have been identified in coronavirus [9–12]. These proteins have the following functions: the hemagglutinin-esterase is the major receptor determinant, binding to sialic acid-containing receptors on the host cell and penetrating of virus genome into host cell cytoplasm by fusion of virus and host cell membranes. Both the envelope and membrane glycoproteins are components of the viral envelope that play a central role in virus morphogenesis and assembly via its interactions with other viral proteins. The nonstructural proteins mediate nuclear export of viral RNPs and bind RNA, thereby inhibiting host

Electronic Supplementary Material is available for this article if you access the article at http://dx.doi.org/10.1007/s00894-004-0210-0.

G. Wu (✉) · S. Yan
Computational Mutation Project,
DreamSciTech Consulting Co. Ltd.,
301, Building 12, Nanyou A-zone, Jiannan Road,
Shenzhen, Guangdong Province, 518054, China
E-mail: hongguanglishibahao@yahoo.com
Tel.: +86-755-22029353

mRNA translation, and regulating viral pre-mRNA splicing and translation. The nucleocapsid protein is the major structural component of virons that associates with genomic RNA to form a helical nucleocapsid. The replicase polyprotein is a multifunctional protein containing the activities necessary for the transcription of negative stranded RNA, leader RNA, subgenomic mRNAs and progeny virion RNA as well as proteinases responsible for the cleavage of the polyprotein into functional products. The spike glycoprotein is responsible for both binding to receptors on host cells and for membrane fusion [13–21].

Currently, the sequences of 158 coronavirus proteins from different species have been documented. Each protein must have its own specific sensitivity to mutations otherwise the proteins would have the same ratio of mutations per amino acid sequences. However such an expectation has yet been found, it is therefore important to define which protein is more sensitive to mutations than the others. The aim of the present study is to discover which protein is more sensitive to mutations among 158 coronavirus proteins using the model developed by us over the last several years.

Materials and methods

The amino acid sequences of 158 coronavirus proteins were obtained from the Swiss-Prot databank [22]. These proteins are grouped as envelope proteins, hemagglutinin-esterase precursors, membrane glycoproteins, nonstructural proteins, nucleocapsid proteins, spike glycoproteins and others including replicase polyprotein and hypothetical proteins (for details, see Supplementary Material).

The detailed calculations of randomly predictable and unpredictable portions in proteins have already been published previously (for a review, see [6]). The calculations governed by the simple permutation principle [23] are described for the example of the spike glycoprotein from human SARS-CoV, which consists of 1,255 amino acids. As we know that an amino-acid pair in a protein is composed of any 20 kinds of amino acids, so theoretically there are 400 possible types of amino-acid pairs. In terms of amino-acid pairs, distinguishing proteins is different either in the numbers of possible types of amino-acid pairs or in the frequency of each type, or both.

Randomly predictable present type of amino-acid pair with predictable frequency

There are 39 arginines (R) and 96 serines (S) in spike glycoprotein from human SARS-CoV, the random frequency of the amino-acid pair ''RS'' is 3 (39/1,255×96/1,254×1,254 = 2.983). Actually we find three ''RS''s in the spike glycoprotein, so the type of ''RS'' is present and its frequency is 3. In such a case, both the presence of type ''RS'' and its frequency are randomly predictable, and the difference between actual and predicted values is 0.

Randomly predictable present type of amino-acid pair with unpredictable frequency

There are 84 alanines (A) in the spike glycoprotein from human SARS-CoV. The frequency of random presence of ''AA'' is 6 (84/1,255×83/1,254×1,254 = 5.555). In fact ''AA'' appears ten times. Thus the presence of type ''AA'' is randomly predictable, but its frequency is randomly unpredictable, and the difference between actual and predicted values is 4.

Randomly unpredictable present type of amino-acid pair

There are 11 tryptophans (W) in the spike glycoprotein from human SARS-CoV, the frequency of random presence of ''WR'' is 0 (11/1,255×39/1,254×1,254 = 0.342), i.e. the type ''WR'' would not appear in the spike glycoprotein. However ''WR'' appears once in reality, so the presence of type ''WR'' is randomly unpredictable. Naturally its frequency is unpredictable too, and the difference between actual and predicted values is 1.

Randomly predictable absent type of amino-acid pair

The frequency of random presence of ''RW'' is 0 (39/1,255×11/1,254×1,254 = 0.342), i.e. the type ''RW'' would not appear in the spike glycoprotein, which is true in the real situation. This is the case that the absence of type ''RW'' with its frequency is randomly predictable, and the difference between actual and predicted values is 0.

Randomly unpredictable absent type of amino-acid pairs

There are 99 threonines (T) in the spike glycoprotein, the frequency of random presence of ''RT'' is 3 (39/1,255×99/1,254×1,254 = 3.076), i.e. there would be three ''RT''s in the spike glycoprotein. However no ''RT'' is found, therefore the absence of ''RT'' from the spike glycoprotein is randomly unpredictable. Naturally its frequency is unpredictable too, and the difference between actual and predicted values is −3.

Statistics

With respect to actual and predicted values in a single protein, the statistical inference is carried out as follows. Generally, each of 20 kinds of amino acids has a chance

of 1/20 ($p = 0.05$) to repeat once, and a type of amino-acid pair has the chance of 1/400 ($p = 0.0025$) to repeat once. In case of the spike glycoprotein from human SARS-CoV, there are 99 Ts, the most abundant amino acid, and 11 Ws, the least abundant amino acid. If the first amino acid is "T", then the chance of the second amino acid to be "T" is 98/1,254 ($p = 0.078 > 0.05$), if the first amino acid is "W", then the chance of the second amino acid to be "W" is 10/1,254 ($p = 0.008 < 0.01$). Thus, the chance of first "TT" is 99/1,255×98/1,254 ($p = 0.0062 < 0.01$), and the chance of second "TT" is 97/1,253×96/1,252 ($p = 0.0059 < 0.01$). If we consider the lowest occurring amino acids "W", the chance of first "WW" is 11/1,255×10/1,254 ($p = 0.00007 < 0.001$), and the chance of second "WW" is 9/1,253×8/1,252 ($p = 0.00005 < 0.001$). Clearly, the probability is less than 0.05 if the difference between actual and predicted values is equal to or larger than 1.

With respect to the comparisons among proteins, the statistical inference is conducted as follows. All the data are examined by the Kolmogorov–Smirnov test to determine their distribution properties. For normal distributions, the data are presented as mean ± SD. For non-normal distributions, the data are presented as median with interquartile range. Outliers are detected according to Healy's method [24]. The one-way ANOVA and the Friedman ANOVA rank tests are used for parametric and non-parametric tests, respectively, followed by comparison tests. SigmaStat for Windows (SPSS Inc, 1992–2003) is used to perform all the statistical tests, and the $p < 0.05$ is considered statistically significant.

Results

After such calculations, the amino-acid pairs in a protein are classified into randomly predictable and unpredictable portions. By comparing the percentages of predictable and unpredictable portions among different proteins, we can find which protein has a larger unpredictable portion than others. Consequently this protein is more sensitive to mutations according to our previous studies [25–32].

Figure 1 shows the predictable and unpredictable portions in coronavirus proteins. This figure can be read as follows. The length of each bar presents 100%, which is located at both unpredictable and predictable sites separated by dotted line. For example, the unfilled bar in spike glycoprotein group presents the absent types, which are composed of 19.70% randomly predictable portion with interquartile range from 16.67 to 26.89% (right panel) and 80.30% randomly unpredictable portion with interquartile range from 73.11 to 83.33% (left panel). The statistical inference in Fig. 1 as well as Fig. 2 is conducted by using the ANOVA test to detect whether or not there is a difference among different proteins in a panel followed by a comparison test. For example, regarding the absent type in Fig. 1, at first we use the Friedman ANOVA rank test whether or not there is a difference among different protein groups. Taking three bars in Fig. 1 into account, the spike glycoproteins have a larger unpredictable portion than others. These results suggest that the spike glycoprotein is more sensitive to mutations than other coronavirus proteins.

Although different proteins have different types of unpredictable absent amino-acid pairs, some types are absent from all members of a group of proteins. For

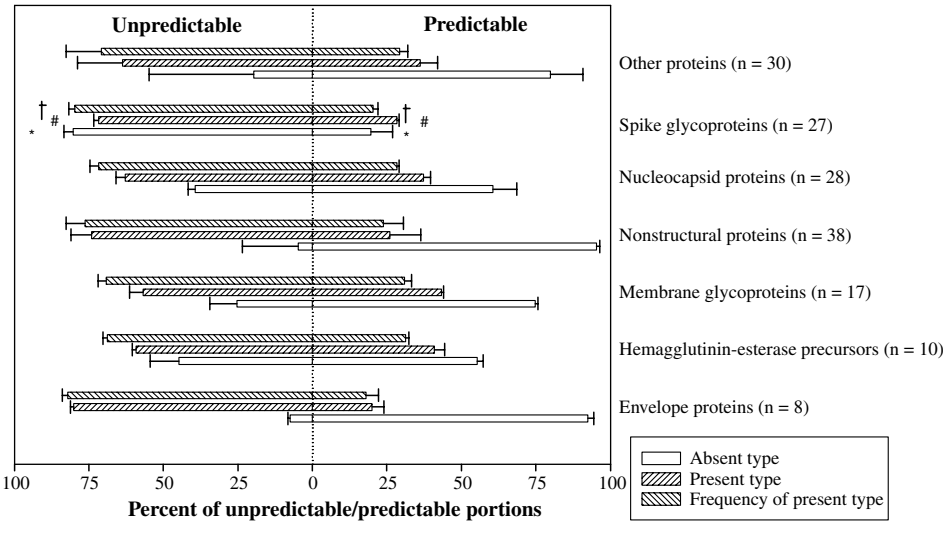

Fig. 1 Predictable and unpredictable portions in coronavirus proteins. The data are presented as median with interquartile range. * the predictable and unpredictable portions in spike glycoprotein group are statistically different from any other protein groups at $p < 0.05$ level, except for hemagglutinin-esterase precursor group. # the predictable and unpredictable portions in spike glycoprotein group are statistically different from hemagglutinin-esterase precursor, membrane protein and nucleocapsid protein groups at $p < 0.05$ level. † the predictable and unpredictable portions in spike glycoprotein group are statistically different from hemagglutinin-esterase precursor, and membrane protein groups at $p < 0.05$ level

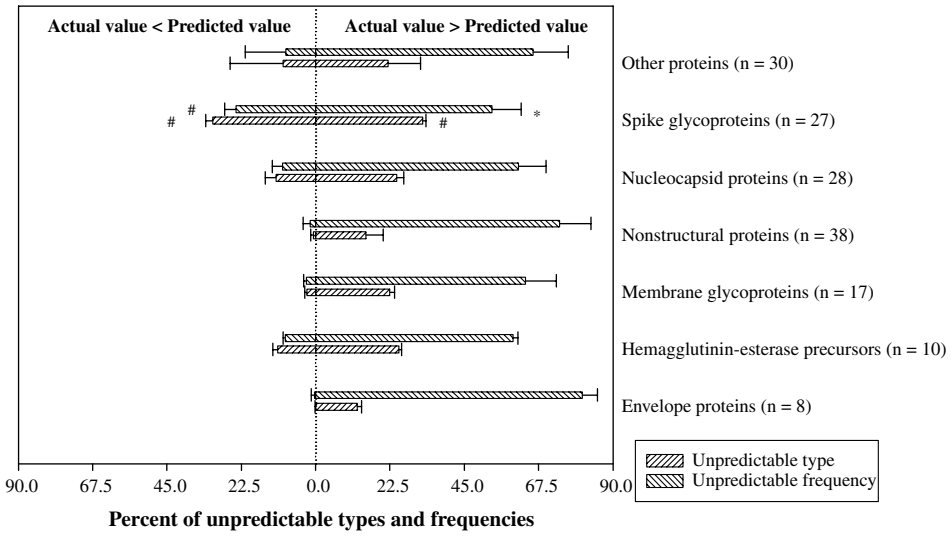

Fig. 2 Percent of unpredictable types and frequencies with respect to whether the actual value is larger or smaller than the predicted value in coronavirus proteins. The data are presented as mean ± SD. * the percents of unpredictable types/frequencies in spike glycoprotein group are statistically different from other protein groups at $p < 0.05$ level. # the percents of unpredictable types in spike glycoprotein group are statistically different from any other protein groups at $p < 0.05$ level, except for hemagglutinin-esterase precursor and nucleocapsid protein groups

example, the amino-acid pair "WI" is absent from all 27 spike glycoproteins no matter which kind of species or strain (Table 1).

Thereafter, we are particularly interested in the unpredictable portions (left panel in Fig. 1), because they are not engineered by randomness. As mentioned under Materials and methods, an unpredictable portion includes the unpredictable types and predictable types with unpredictable frequency, which can be presented as the actual values either larger or smaller than its predicted values. Our previous studies reveal that the unpredictable types whose actual value is larger than its predicted value are highly likely to be targeted by mutations, whereas the unpredictable types whose actual value is smaller than its predicted value are highly likely to be formed after mutations [25–33].

Figure 2 illustrates the percentage of unpredictable types and frequencies with respect to whether the actual value is larger or smaller than its predicted value in coronavirus proteins. Technically Fig. 2 is a subset of Fig. 1 obtained by classifying the data in the left panel of Fig. 1 into two criteria, i.e., the actual value is larger than the predicted value, or vice versa. In view of the unpredictable portion whose actual value is smaller than its predicted value (left panel), the spike glycoproteins have the largest percentages in both unpredictable type and frequency among different coronavirus proteins. Whereas in view of the unpredictable portion whose actual value is larger than its predicted value (right panel), the spike glucoprotein group reveals a larger percentage of unpredictable type accompanied by a smaller percentage of unpredictable frequency. This means that the spike glycoprotein might have undergone more mutations in the past than others.

Subsequently, we are still more interested in the magnitude of difference between the actual and predicted values because our previous studies show that the larger the difference between actual and predicted values, the bigger the potential towards future mutations [25–33].

Figure 3 displays the magnitude of difference between actual and predicted values in coronavirus proteins. It can be seen that the difference between actual and predicted values is larger in the spike glycoprotein group than in others. This implies that the spike glycoproteins have a high potential for future mutations.

In addition, the difference between the actual and predicted values can tell us which species is more subject to mutations if we arrange the number of amino-acid pairs with respect to the difference between the actual and predicted values in each group of proteins from different species.

Figures 4, 5, 6, 7, 8, 9 and 10 show the difference between the actual and predicted values in each group of proteins from different species. The scale of the vertical axes in Figs. 4, 5, 6, 7, 8, 9 and 10 is shown logarithmically in order to emphasize the amino-acid pairs with large differences between the actual and predicted values. Due to the limitation of the graphic software, the filled forms are duplicated in one or two bars. However the data used in these figures can be found in the Supplementary Material. These figures can be understood as follows, the bars at two extremes along the horizontal axis present the amino-acid pairs sensitive to mutations, because our previous studies have shown that the larger the difference between actual and predicted values is, the more sensitive to the mutations is [25–33]. By comparing the scales of horizontal axes from Figs. 4, 5, 6, 7, 8, 9 and 10, we can see that the spike glycoproteins are more sensitive to mutations than other proteins because Fig. 9

Table 1 Unpredictable absent amino-acid pairs that disappear from a group of proteins

Hemagglutinin-esterase precursor	Spike glycoprotein
RA, RD, NQ, DR, CA, CS, QF, IK, LK, FA, FC, FQ, FP, VK	WI

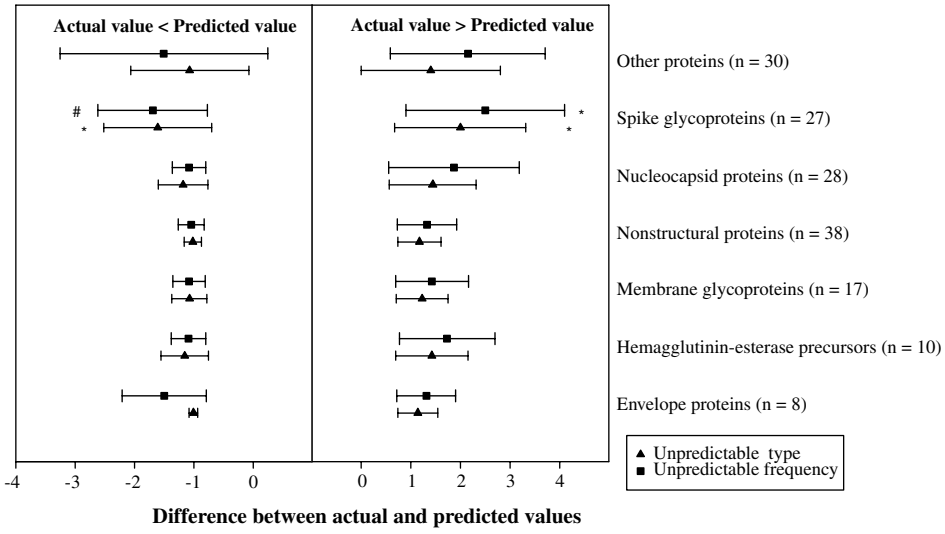

Fig. 3 Magnitude of difference between actual and predicted values in coronavirus proteins. The data are presented as mean ± SD. * indicates the difference between actual and predicted values in spike glycoprotein group is statistically different from any other protein group at $p < 0.05$ level. # indicates the difference between actual and predicted values in spike glycoprotein group is statistically different from other protein groups at $p < 0.05$ level, except for envelope protein group

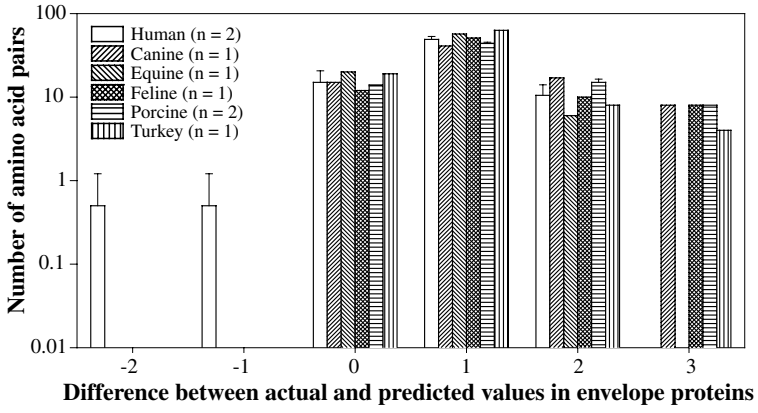

Fig. 4 Number of amino-acid pairs in envelop proteins from different species with respect to the difference between their actual and predicted values. The data are presented as mean ± SD

has a largest scale for the horizontal axis. Still we can see which species is more sensitive to mutations in each figure. For instance, the human spike glycoprotein is more sensitive to mutation in Fig. 9.

Discussion

Without clearly identifying the source of SARS-CoV, its fast-spreading process, and its mutations, the battle with SARS is unlikely to be finished soon, therefore sooner or later we would expect to see new mutated forms of SARS-CoV. In such a case the determination of vulnerable proteins in SARS-CoV is important and pressing.

The coronaviruses exhibit considerable serologic and sequence variation, with the most extreme variability being within S genes [3]. Variant spike glycoproteins [34] are now known to impact pathogenic outcome [15, 35–37].

This study provides three lines of evidence that suggest that the spike glycoprotein is different from the others: (1) the spike glycoprotein is more sensitive to mutations, this is the current state of spike glycoprotein, (2) the spike glycoprotein had experienced more mutations in the past, this is the history of spike glycoprotein, and (3) the spike glycoprotein has a bigger potential towards future mutations, this is the future of spike glycoprotein.

With respect to the first line of evidence, the argument is that the randomly unpredictable portion is larger in spike glycoproteins than in others (Fig. 1). If we compare the unpredictable portion in spike glycoproteins with the proteins we have studied in the past (columns I and II in Table 2, similar to the left panel in Fig. 1), we find that the unpredictable portion of the present types is statistically larger in spike glycoproteins than in others, and statistically similar in the unpredictable portion of the present frequencies. This suggests that the spike glycoprotein is not only more sensitive to mutations than other coronavirus proteins, but also more sensitive than the proteins in Table 2.

With respect to the second line of evidence, we find that the spike glycoprotein has a larger percentage of unpredictable types and frequencies whose actual values are smaller than the predicted values in Fig. 2. Actually, 172 mutations have currently been documented in coronavirus proteins, of which 153 occur in spike glyco-

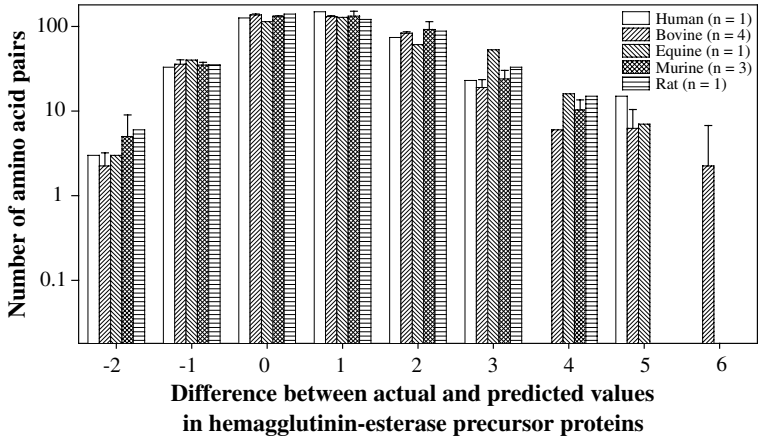

Fig. 5 Number of amino-acid pairs in hemagglutinin-esterase precursor proteins from different species with respect to the difference between their actual and predicted values. The data are presented as mean ± SD

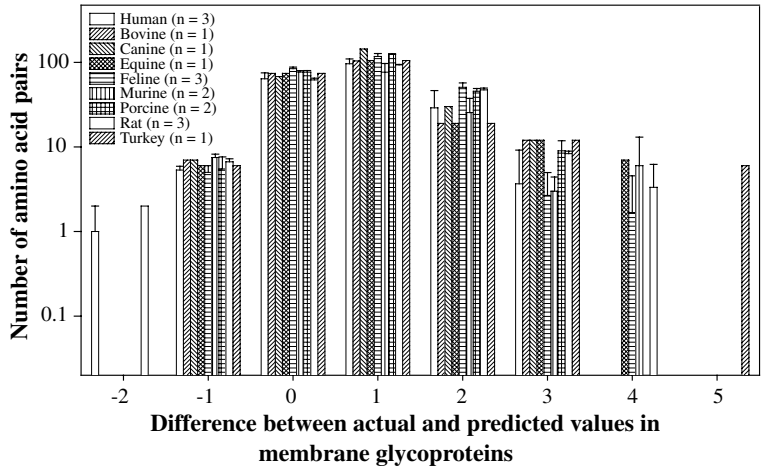

Fig. 6 Number of amino-acid pairs in membrane glycoproteins from different species with respect to the difference between their actual and predicted values. The data are presented as mean ± SD

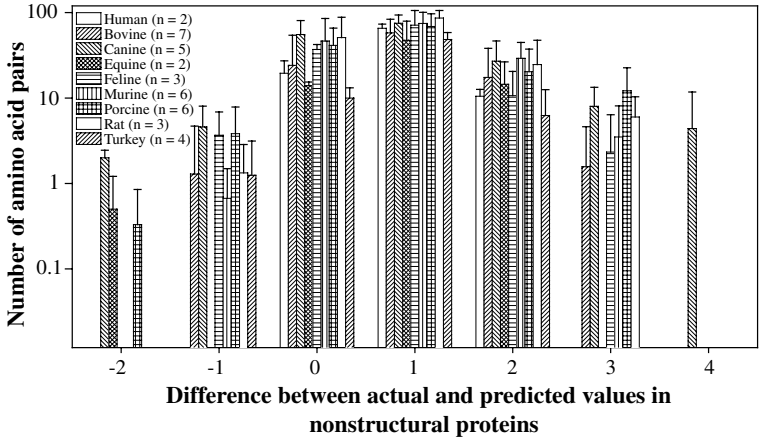

Fig. 7 Number of amino-acid pairs in nonstructural proteins from different species with respect to the difference between their actual and predicted values. The data are presented as mean ± SD

proteins. This supports our argument that the spike glycoprotein has undergone more mutations in the past. Moreover, if we look at the nine proteins which have been documented with more mutations (column IX in Table 2), we find that the percentage of unpredictable type in spike glycoproteins is statistically similar to the proteins in Table 2 (columns III and IV in Table 2, similar to right panel in Fig. 2), but the difference regarding the percentage of unpredictable frequencies is statistical significant. This suggests that the intensity of mutations in spike glycoproteins is weaker than the first nine proteins listed in Table 2.

With respect to the third line of evidence, we find that the difference between actual and predicted values in spike glycoproteins is larger than in others (Fig. 3). Comparison with the first nine proteins in Table 2 (columns V, VI, VII and VIII in Table 2, similar to Fig. 3) shows that the difference between actual and

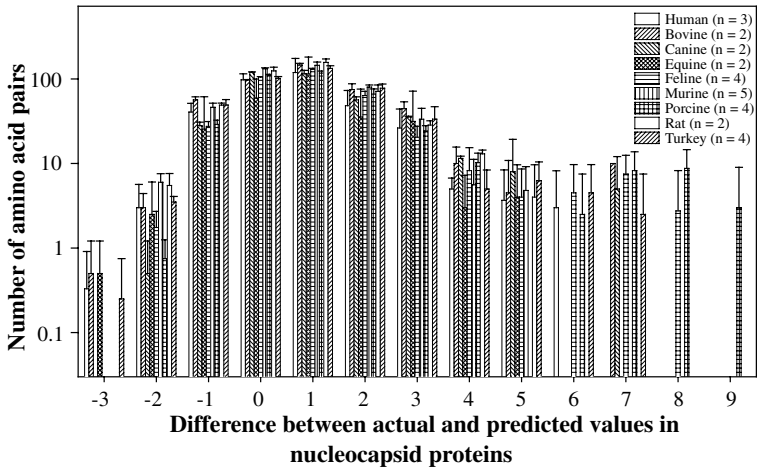

Fig. 8 Number of amino-acid pairs in nucleocapsid proteins from different species with respect to the difference between their actual and predicted values. The data are presented as mean ± SD

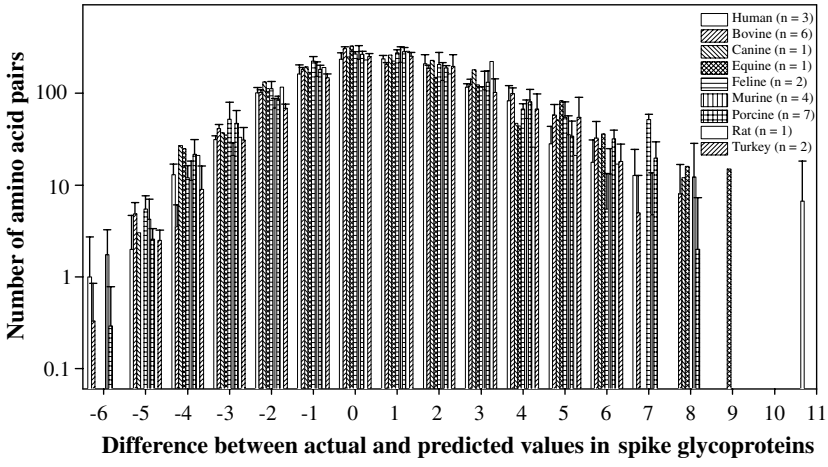

Fig. 9 Number of amino-acid pairs in spike glycoproteins from different species with respect to the difference between their actual and predicted values. The data are presented as mean ± SD

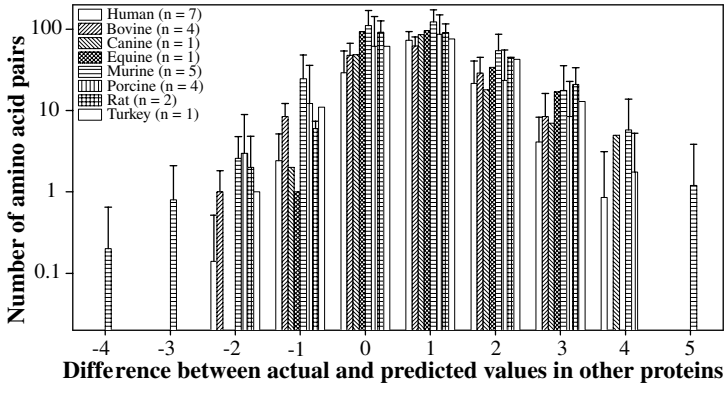

Fig. 10 Number of amino-acid pairs in other proteins from different species with respect to the difference between their actual and predicted values. The data are presented as mean ± SD

predicted values is statistically larger in spike glycoproteins regarding unpredictable types and is statistically smaller regarding unpredictable frequency. This suggests that the spike glycoprotein still has more potential for mutations than the first nine proteins in Table 2.

For the species susceptibility, the vulnerability of species depends on the number of amino-acid pairs with the largest difference between actual and predicted values. Figures 4, 5, 6, 7, 8, 9 and 10 may, at least partly, highlight the species susceptibility. For example, why have so many mutations been found in the human spike glycoproteins?

Although it is obvious that an individual protein is different from the other proteins of a genome, our results quantitatively and systematically determine the difference between the spike and other proteins by comparing their predictable and unpredictable portions of amino-acid pairs. One may argue that it is also known that spike proteins interact with the host, the environment and the immune system and so their structure is par-

Table 2 Characteristics of the proteins that we have studied in the past

Protein	I	II	III	IV	V	VI	VII	VIII	IX	Reference
BTK	62.46	71.88	36.25	12.77	−1.26	−1.30	1.46	1.72	112	[32]
CA54	73.75	93.47	36.50	20.31	−3.86	−10.96	4.68	41.52	151	[28]
FA9	62.35	72.83	32.00	9.35	−1.13	−1.09	1.37	1.60	99	[30]
GLCM	59.77	71.59	37.25	14.39	−1.12	−1.12	1.51	1.93	109	[31]
HBA	61.62	68.57	10.75	4.29	−1.02	−1.00	1.19	1.49	133	[27]
LDLR	69.61	80.21	40.50	18.74	−1.43	−1.32	1.86	2.43	127	[26]
Human p53	57.14	68.37	30.75	5.87	−1.15	−1.13	1.45	1.84	190	[29]
PH4H	59.83	71.84	28.50	7.10	−1.16	−1.06	1.39	1.62	187	[25]
VHL	72.46	78.30	18.00	9.91	−1.07	−1.05	1.24	1.634	109	[33]
RUN1	64.22	75.22							6	[41]
ADHA	55.98	64.61								[42]
CTGF	58.46	70.40								[43]
GSHR	57.70	68.71							1	[44]
AO FB	62.40	73.94								[38]
LIS1	56.76	71.32							5	[45]
TNFA	59.24	69.40								[46]
TYRO	45.45	58.14							64	[47]
ATTY	53.36	67.55							1	[48]
Bovin p53	62.44	71.95								[49]
Mouse p53	60.85	74.29							3	[50]
Sheep p53	60.19	70.34								[51]
AMPC	54.63	66.32							9	[52]
DOPO	61.13	73.75							8	[53]

BTK human Bruton's tyrosine kinase, *CA*54 human collagen α5(IV) chain precursor, *FA*9 human coagulation factor IX precursor, *GLCM* human β-glucocerebrosidase, *HBA* haemoglobin α chain, *LDLR* human low-density lipoprotein receptor, *PH*4*H* human phenylalanine hydroxylase protein, *VHL* Von Hippel-Lindau disease tumor suppressor, *RUN*1 human acute myeloid leukemia 1 protein, *ADHA* human alcohol dehydrogenase α-chain, *CTGF* human connective tissue growth factor, *GSHR* human glutathione reductase, *AOFB* human monoamine oxidase B, *LIS*1 human platelet-activating factor acetylhydrolase α-subunit, *TNFA* human tumor necrosis factor, *TYRO* human tyrosinase, *ATTY* human tyrosine aminotransferase, *AMPC_ CITFR Citrobacter Freundii* β-lactamase, *DOPO* human dopamine β-hydroxylase, *I* percent of unpredictable portion of present types, *II* percent of unpredictable portion of present frequencies, *III* percent of unpredictable present types whose actual values are smaller than predicted values, *IV* percent of unpredictable present frequencies whose actual values are smaller than predicted values, *V* difference between actual and predicted values in unpredictably present types whose actual values are smaller than predicted values, *VI* difference between actual and predicted values in unpredictably present frequencies whose actual values are smaller than predicted values, *VII* difference between actual and predicted values in unpredictably present types whose actual values are larger than predicted values, *VIII* difference between actual and predicted values in unpredictably present frequencies whose actual values are larger than predicted values, *IX* number of mutations

ticularly vulnerable to mutations both in the past and in the future and also regarding its specific phenotypic effects in the numerous interactions it is involved in. However we would like to argue that the host, the environment and the immune system are the external factors imposed on the spike proteins, while the internal factor in the spike proteins, which is particularly interesting to us, is the structure that can be partially explained by our random approach. In another study on the spike protein, we specifically discussed the spike proteins from three human coronaviruses classified with our approach and gave predictions of possible and potential mutation forms regarding the spike protein structure [7].

At this stage of study, it is still difficult to define the reason and to give a biological explanation to the results that the absent types in the spike protein behave differently from and opposed to other proteins, although we have discussed the biological explanation in the present types in rat monoamine oxidase B in the past [38]. However it is certain that the randomly unpredictable absent types should be deliberately eliminated from a protein rather than being self-organized and self-empowered. This is so because such an absence cannot be explained by randomness which suggests the least time- and energy-consuming.

In this study, we do not consider the situation that individual variation within the other protein groups could not in specific cases lead to similar values as observed for specific spike proteins. This is so because the individual variation within the other protein groups would lead to a mutated form of a protein, while this study deals with proteins without mutations. However, a mutated form of protein may lead its predictable and unpredictable portions to shift to similar values as observed for specific spike proteins. In the current form of this study, we cannot make any solid prediction from the present analysis for the behavior of individual proteins, but only observe an overall trend.

The medical implication is that the mutation sensibility in spike glycoprotein leads to the difficulties in producing vaccines that provide us with long-lasting protection against SARS. This finding can be correlated with hemagglutinin and neuraminidase from influenza A virus. Both hemagglutinin and neuraminidase are surface proteins, and subject to the pressure of the antibody and the selective pressure for the appearance of host cell variant with altered receptor binding specificity. Meanwhile the spike glycoprotein is responsible for both binding to receptors on host cells and for membrane fusion. In this viewpoint, the spike glyprotein is quite similar to hemagglutinin and neuraminidase.

The multiple sequence alignments are a phenomenological technique by comparing the similarity among proteins. The phenomenological analogy can be classified into at least three types. For the simplest example, we compare the letters that construct a word to guess the meaning of the word. Another type of phenomenological analogy is equivalent in physical laws, for example, Fick's law and Kirchhoff's law are equivalent to the law of conservation. The third type of phenomenological analogy is mathematically similar, for example, the transfer of energy, mass, heat and momentum can be described by using similar differential equations. [39, 40] In fact, what the multiple sequence alignments are doing is language similarity. On the other hand, our approach is a mechanism-driven technique by calculating the randomly predictable and unpredictable portions in a protein. Our approach is not a phenomenological tool, and is studying the internal power engineering the mutations. Multiple sequence alignments cannot predict

the future, while our approach can predict the likelihood of future mutations. Technically, multiple sequence alignments need a large database for searching, while our approach needs a few data but a large amount of calculations. In general, multiple sequence alignments are the first step for the understanding of proteins, DNA, etc., and science must advance to seek other new techniques for the understanding of proteins, DNA, etc. However, our approach at this moment is only related to the primary structure, therefore it cannot give information on loop regions, as multiple sequence alignments also cannot. With respect to the evolutionary pressure, our approach is using the randomly unpredictable portion to account, as we argue that the randomly unpredictable portion should be deliberately developed through the evolutionary process. This is so because randomness suggests the least time- and energy-consuming to construct proteins.

In conclusion, our results suggest that the spike glycoproteins are more vulnerable to mutations among coronavirus proteins, however the chance of occurring of mutations would be less in spike glycoproteins than in highly-frequently-mutated proteins, e.g. the human p53 protein.

Acknowledgements The authors wish to thank the anonymous referees for their insightful comments, which sharpen up the points presented in this study.

References

1. Hu LD, Zheng GY, Jiang HS, Xia Y, Zhang Y, Kong XY (2003) Acta Pharmacol Sin 24:741–745
2. Ruan YJ, Wei CL, Ee AL, Vega VB, Thoreau H, Su ST, Chia JM, Ng P, Chiu KP, Lim L, Zhang T, Peng CK, Lin EO, Lee NM, Yee SL, Ng LF, Chee RE, Stanton LW, Long PM, Liu ET (2003) Lancet 361:1779–1785
3. Siddell SG (1995) The coronaviridae: an introduction. Plenum, New York, pp 1–10
4. Stavrinides J, Guttman DS (2004) J Virol 78:76–82
5. Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, Luo SW, Li PH, Zhang LJ, Guan YJ, Butt KM, Wong KL, Chan KW, Lim W, Shortridge KF, Yuen KY, Peiris JS, Poon LL (2003) Science 302:276–278
6. Wu G, Yan SM (2002) Mol Biol Today 3:55–69
7. Wu G, Yan S (2003) Peptides 24:1837–1845
8. Wu G, Yan S (2003) Peptides 25:901–908
9. Ismail MM, Cho KO, Hasoksuz M, Saif LJ, Saif YM (2001) Avian Dis 45:978–984
10. Motokawa K, Hohdatsu T, Hashimoto H, Koyama H (1996) Microbiol Immunol 40:425–433
11. Nguyen VP, Hogue BG (1998) Adv Exp Med Biol 440:361–365
12. Vennema H, Godeke GJ, Rossen JW, Voorhout WF, Horzinek MC, Opstelten DJ, Rottier PJ (1996) EMBO J 15:2020–2028
13. Bosch BJ, van der Zee R, de Haan CA, Rottier PJ (2003) J Virol 77:8801–8811
14. Davidson A, Siddell S (2003) Curr Opin Infect Dis 16:565–571
15. Hingley ST, Leparc-Goffart I, Weiss SR (1998) J Virol 72:1606–1609
16. Narayanan K, Maeda A, Maeda J, Makino S (2000) J Virol 74:8127–8134
17. Nguyen VP, Hogue BG (1997) J Virol 71:9278–9284
18. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Gunther S, Osterhaus AD, Drosten C, Pallansch MA, Anderson LJ, Bellini WJ (2003) Science 30:1394–1399
19. Sanchez CM, Izeta A, Sanchez-Morgado JM, Alonso S, Sola I, Balasch M, Plana-Duran J, Enjuanes L (1999) J Virol 73:7607–7618
20. Shen X, Xue JH, Yu CY, Luo HB, Qin L, Yu XJ, Chen J, Chen LL, Xiong B, Yue LD, Cai JH, Shen JH, Luo XM, Chen KX, Shi TL, Li YX, Hu GX, Jiang HL (2003) Acta Pharmacol Sin 24:505–511
21. Taguchi F, Shimazaki YK (2000) J Gen Virol 81:2867–2871
22. Bairoch A, Apweiler R (2000) Nucleic Acids Res 28:45–48
23. Feller W (1968) An introduction to probability theory and its applications, 3rd edn, vol I. Wiley, New York, pp 38–40
24. Healy MJR (1979) Clin Chem 25:675–677
25. Wu G, Yan SM (2002) Peptides 23:2085–2090
26. Wu G, Yan S (2002) J Biochem Mol Biol Biophys 6:401–406
27. Wu G, Yan SM (2003) Comp Clin Pathol 12:21–25
28. Wu G, Yan S (2003) Peptides 24:347–352
29. Wu G, Yan S (2003) J Mol Model 9:337–341
30. Wu G, Yan S (2003) J Biomed Sci 10:451–454
31. Wu G, Yan S (2003) Protein Eng 16:195–199
32. Wu G, Yan S (2003) Mol Simul 29:249–254
33. Wu G, Yan S (2003) J Appl Res 3:512–520
34. Krueger DK, Kelly SM, Lewicki DN, Ruffolo R, Gallagher TM (2001) J Virol 75:2792–2802
35. Leparc-Goffart I, Hingley ST, Chua MM, Phillips J, Lavi E, Weiss SR (1998) J Virol 72:9628–9636
36. Kuo L, Godeke GJ, Raamsman MJB, Masters PS, Rottier PJM (2000) J Virol 74:1393–1406
37. Das Sarma J, Fu L, Tsai JC, Weiss SR, Lavi E (2000) J Virol 74:9206–9213
38. Wu G, Yan SM (2001) Biomol Eng 18:23–27
39. Holman JP (1990) Heat transfer, 7th edn. McGraw-Hill, New York, p 607
40. Wu G (2000) Med Hypotheses 54:748–749
41. Wu G, Yan SM (2000) Comp Haematol Int 10:85–89
42. Wu G (2000) Alcohol Alcohol 35:302–306
43. Wu G, Yan SM (2001) J Mol Model 5:120–124
44. Wu G (2000) Biochem Biophys Res Commun 268:823–826
45. Wu G, Yan SM (2000) Pädiatr Grenzgeb 39:513–526
46. Wu G (2000) Cancer Lett 153:145–150
47. Wu G, Yan SM (2001) Pädiatr Grenzgeb 40:153–166
48. Wu G (2000) Pädiatr Grenzgeb 39:37–47
49. Wu G, Yan SM (2002) Mol Biol Today 3:31–37
50. Wu G (2000) Hum Exp Toxicol 19:535–539
51. Wu G (2000) J Biochem Mol Biol Biophys 4:179–185
52. Wu G, Yan SM (2000) J Mol Microbiol Biotechnol 2:277–281
53. Wu G (2000) Mol Psychiatry 5:448–451