Available online at www.sciencedirect.com

SCIENCE DIRECT®

Peptides 25 (2004) 901–908

PEPTIDES

www.elsevier.com/locate/peptides

Potential targets for anti-SARS drugs in the structural proteins from SARS related coronavirus

Guang Wu*, Shaomin Yan

DreamSciTech Consulting Co. Ltd., 301 Building 12, Nanyou A-Zone, Jiannan Road, Shenzhen, Guangdong Province, CN-518054, China

Received 1 December 2003; received in revised form 1 March 2004; accepted 1 March 2004

Available online 10 April 2004

Abstract

This is a further study on the severe acute respiratory syndrome (SARS) using the probabilistic models. The purpose was to define the potential targets for anti-SARS drugs in the structural proteins from human SARS related coronavirus (SARS-CoV) while knowing little about the functional sites and possible mutations in these proteins. From a probabilistic viewpoint, we can theoretically select the amino acid pairs as potential candidates for anti-SARS drugs. These candidates have a greater chance of colliding with anti-SARS drugs, are more likely to link with the protein functions and are less vulnerable to mutations.

Keywords: Amino acid pairs; Coronavirus; Drug targets; SARS

1. Introduction

To deal with a possible recurrence of the severe acute respiratory syndrome (SARS), the determination of targets in SARS related coronavirus (SARS-CoV) is important and pressing for the development of anti-SARS drugs [1,3,7,20]. Without sufficient knowledge of the SARS-CoV at present, it is quite difficult to define the potential targets in SARS-CoV for anti-SARS drugs. To solve this problem, several approaches, such as the determination of binding sites in SARS-CoV [6], the alignment and multiple comparison among protein data bank with Blastp and other computer software [5,11], the prediction of mutation sites in spike protein from SARS-CoV [35], have been taken to analyze the potential targets in SARS-CoV.

Nevertheless, it is still necessary to use different approaches to discover the potential targets in SARS-CoV for drug design. The structural proteins from SARS-CoV can be primarily considered the potential targets [1], because their functions are comparatively clear. Currently, we are attracted by five structural proteins from SARS-CoV, i.e. the replicase, spike, envelope, membrane and nucleocapsid proteins. These proteins are similar to other coronaviruses in function. 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 protein is responsible for binding to receptors on host cells and for membrane fusion. The envelope and membrane glycoproteins are components of the viral envelope that plays a central role in virus morphogenesis and assembly via their interactions with other viral proteins. The nucleocapsid protein is the major structural component of virons that associates with genomic RNA to form a helical nucleocapsid [4,7,10,14,15].

Without detailed knowledge of functional sites in the structural proteins in SARS-CoV, a probabilistically simple approach for drug efficacy is to target the abundant amino acids in these proteins. We could expect that the anti-SARS drugs have a greater chance to interact with SARS-CoV if the collision between anti-SARS drugs and structural proteins in SARS-CoV is a random event [28]. This being the case, there are two problems: (i) a single amino acid from abundant groups does not directly represent the functional groups in proteins, because a good signature pattern of a protein must be as short as possible, but the conserved sequence is not longer than four or five residues [13]; and (ii) even if these amino acids would be located at the functional sites, there is still a chance of mutation at these amino acids leading to the inefficacy of anti-SARS drugs [17].

Therefore, further effort is made to find potential functional sites with small mutation possibility in the structural proteins from SARS-CoV. Over the last few years we have

* Corresponding author. Tel.: +86-755-2202-9353; fax: +86-755-2520-8256.
E-mail address: hongguanglishibahao@yahoo.com (G. Wu).

0196-9781/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.peptides.2004.03.002

developed three models using random approaches to analyze the functional amino acid pairs in proteins and to evaluate the mutation effects on amino acid pairs [30]. Generally, in terms of actual and predicted frequencies, our models classify the amino acid pairs in a protein into two categories: the randomly predictable and the unpredictable. First, our models suggest that the amino acid pairs with large differences between actual and predicted frequencies should be deliberately developed and probably be located at the functional sites, as the random construction of an amino acid pair is the least time- and energy-consuming. For the functional purpose, nature should have the intention to spend more time and energy to construct amino acid pairs with large differences between actual and predicted frequencies [24–26,29]. Second, our models reveal that the mutations are likely to occur at the amino acid pairs, whose actual frequency is larger than their predicted frequency [31–34]. Finally, our models demonstrate that the amino acid pairs with high Markov transition probability are less sensitive to mutations [21–23,27].

Our models suggest that the ideal targets are based not only on the abundance of amino acids, but also on the potentially functional sites as well as on a small chance of mutations. In this study we used our models to analyze five structural proteins from SARS-CoV to determine the potential targets for anti-SARS drugs.

2. Materials and methods

The amino acid sequences of the replicase (access number: P59641), spike (access number: P59594), envelope (access number: P59637), membrane (access number: P59596) and nucleocapsid (access number: P59595) proteins are obtained from the SWISS-PROT data bank [2].

2.1. Determination of potentially functional amino acid pairs

For the determination of the potentially functional amino acid pairs in the structural proteins, the actual and predicted frequencies of amino acid pairs were calculated and the difference between them compared. The detailed calculations and their rationales with examples are described below.

2.1.1. Amino acid pairs in SARS-CoV spike protein

The spike protein from human SARS-CoV consists of 1255 amino acids. We count the first and second amino acids as an amino acid pair, the second and third as another pair, the third and fourth, until the 1254th and 1255th, thus there are 1254 pairs. As there are 20 types of amino acids and any amino acid pair can be composed from any of these 20 types of amino acids, so theoretically there are 400 possible types of amino acid pairs. Again there are 1254 pairs in the spike protein, more than the 400 types of theoretically possible pairs. Clearly some of the 400 types should appear more than once. Meanwhile, it is reasonable to expect that some of the 400 types are absent from the spike protein.

2.1.2. Actual frequency and randomly predicted frequency in SARS-CoV spike protein

The randomly predicted frequency is governed by the simple permutation principle [8]. For instance, there are 39 arginines (R) and 96 serines (S) in the spike protein. The predicted frequency of amino acid pair “RS” would be 3 $((39/1255) \times (96/1254) \times 1254 = 2.983)$. Actually we can find three “RS”s in the spike protein, so the actual frequency of “RS” is 3.

2.1.3. Randomly predictable present amino acid pairs

As described in the last section, the predicted frequency of a randomly present amino acid pair “RS” would be 3 and “RS” really appears three times in the protein, so the presence of “RS” is randomly predictable.

2.1.4. Randomly unpredictable present amino acid pairs

There are 84 alanines (A) in the spike protein, the frequency of a random presence of amino acid pair “AA” would be 6 $((84/1255) \times (83/1254) \times 1254 = 5.555)$, i.e. there would be six “AA”s in the spike protein. But in fact the “AA” appears 10 times in the protein, so the presence of “AA” is randomly unpredictable. This illustrates the case that the actual frequency of “AA” is larger than its predicted frequency. Another case is that the actual frequency is smaller than the predicted frequency, for example, there are 91 valines (V) in the spike protein and the predicted frequency of “AV” is 6 $((84/1255) \times (91/1254) \times 1254 = 6.091)$, while its actual frequency is only 3.

2.1.5. Difference between actual and predicted frequencies

Hence, there are three relationships between the actual and predicted frequencies, i.e. the actual frequency is smaller than, equal to or larger than the predicted frequency. Our previous studies suggest that the amino acid pairs with a big difference between actual and predicted frequencies can be deliberately developed and probably are located at the functional sites.

2.2. Calculation of Markov transition probability

To minimize the chance that the mutations occur at the targets of anti-SARS drugs, we need to find the amino acid pairs with high Markov transition probability. The Markov transition probability calculates the probability from one state to another state [18]. For an amino acid pair, an amino acid has a certain probability to follow a certain preceding amino acid, which constructs a conditional probability (the first-order Markov chain), i.e. the probability of an amino acid occurs in an amino acid pair given a certain first amino acid [P(second amino acid|first amino acid)]. The calculation of this probability is the transition from the state of one amino acid to the state of an amino acid pair. More practically, in the case of the English language, the first-order Markov chain analyzes the probability, for example, that the "e" follows "w" from 26 available letters. Once again we use the spike protein as an example, there are 39 "R"s and 39 "C"s in this protein. Intuitively either "R" or "C" would have the same probability to follow "N", actually the Markov transition probability shows that the "C" has a bigger probability to follow "N" rather than "R", so the amino acid pair "NC" is more stable and less sensitive to mutations than "NR".

2.3. Statistics

The actual frequency and predicted frequency can be compared as follows. Generally each of the 20 types of amino acids has a chance of 1/20 ($P = 0.05$) to repeat once, and an amino acid pair has a chance of 1/400 ($P = 0.0025$) to repeat once in the protein primary structure. In case of the spike protein from human SARS-CoV, there are 99 threonines (T) and 99 leucines (L), the most abundant amino acids, and 11 tryptophans ("W"s), the least amino acid. If the first amino acid is "T", then the chance of the second amino acid being "T" is 98/1254 ($P = 0.078 > 0.05$). If the first amino acid is "W", then the chance of the second amino acid being "W" is 10/1254 ($P = 0.008 < 0.01$). Accordingly the chance of the first amino acid pair being "TT" is $(99/1255) \times (98/1254)$ ($P = 0.0062 < 0.01$), and the chance of the second amino acid pair being "TT" is $(97/1253) \times (96/1252)$ ($P = 0.0059 < 0.01$). For the least frequent amino acid "W", the

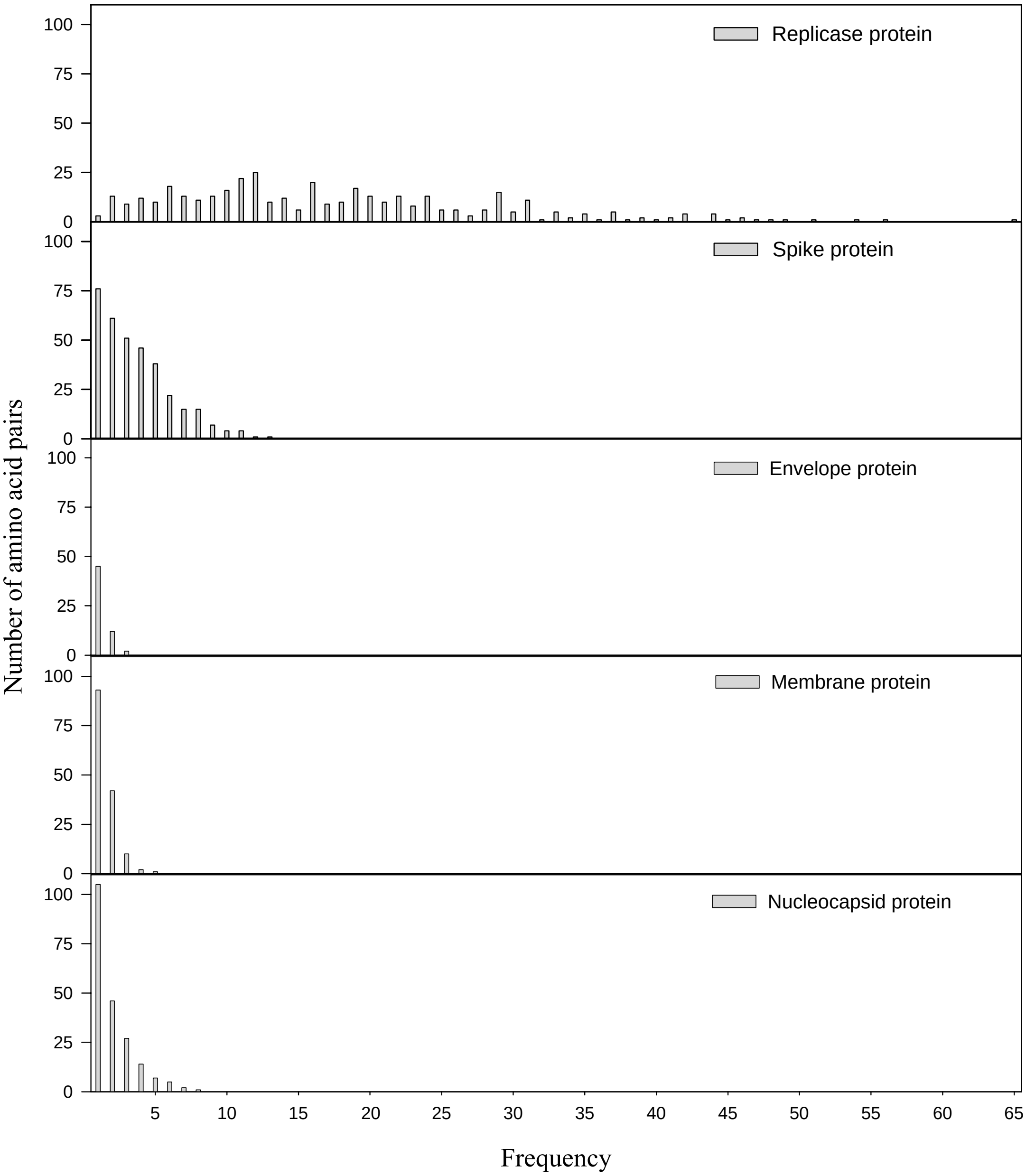

Fig. 1. Repetition of amino acid pairs in the structural proteins from SARS-CoV.

chance of first amino acid pair of “WW” is (11/1255) × (10/1254) ($P = 0.00007 < 0.001$), and the chance of second amino acid pair of “WW” is (9/1253) × (8/1252) ($P = 0.00005 < 0.001$). For that reason, the probability is <0.05 if the difference between any amino acid pairs is greater than or equal to one. This statistical consideration is also applied to the Markov chain transition probability.

3. Results

To determine the amino acid pairs which have a greater chance of random collision with the anti-SARS drugs, we count the frequency of each amino acid pair in these proteins. Fig. 1 shows the number of amino acid pairs versus their frequency. For example, the bottom panel indicates that 105 amino acid pairs appear once, 46 pairs twice, 27 pairs three times, 14 pairs four times, seven pairs five times, two pairs six times and one pair eight times in the nucleocapsid protein. Clearly, the frequency of amino acid pairs is higher in the replicase protein than in the other four proteins.

Although frequently appearing amino acid pairs have a greater chance of interaction with anti-SARS drugs, they are possibly not located at functional sites or well exposed to anti-SARS drugs. The differences between actual and predicted frequencies were calculated in order to determine the amino acid pairs with high probability at functional sites (Fig. 2). Comparing the amino acid pairs in Figs. 1 and 2, it

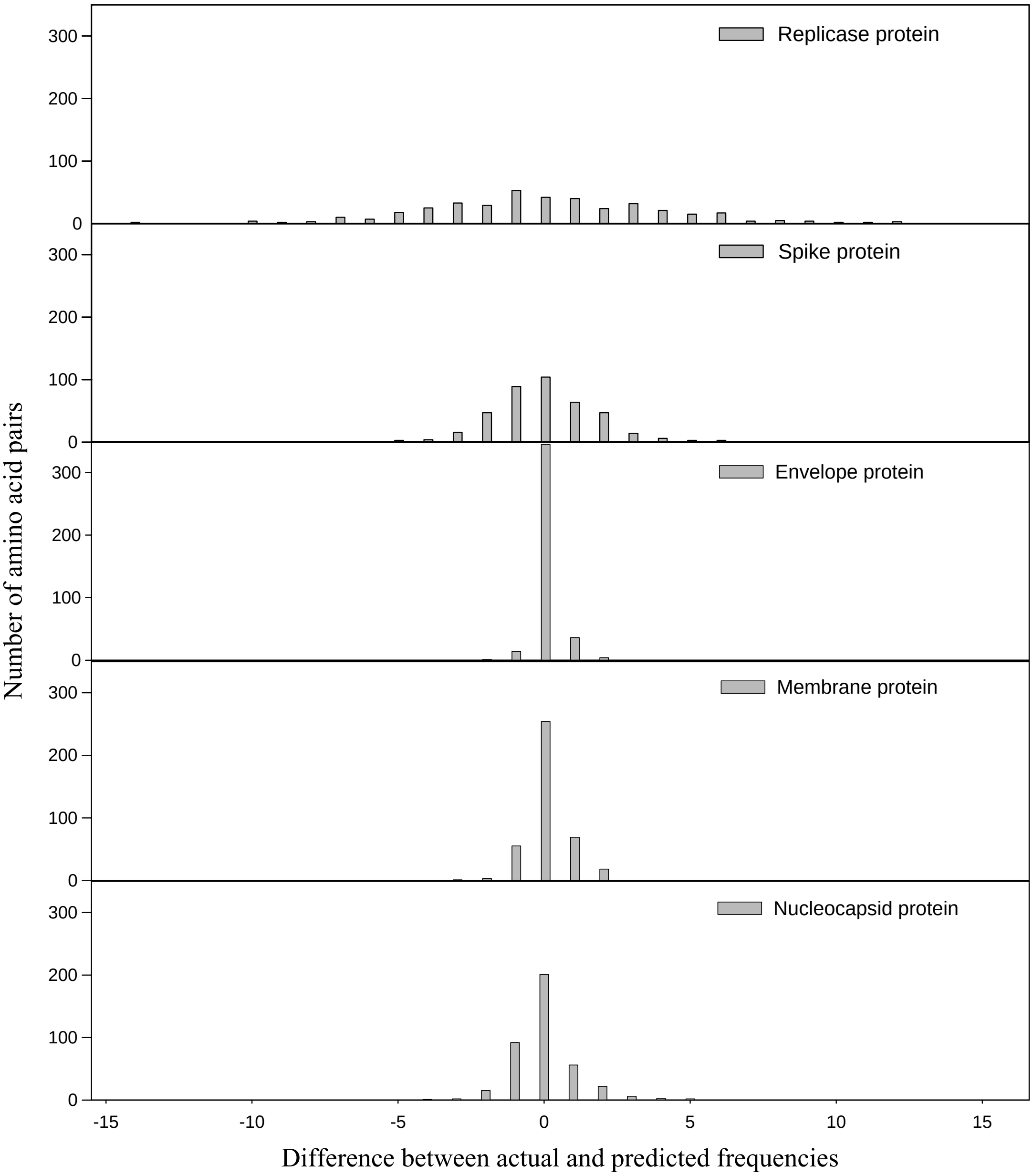

Fig. 2. Difference between actual and predicted frequencies in the structural proteins from SARS-CoV.

can be found that the some of frequently appearing amino acid pairs are not the ones with large differences between actual and predicted frequencies. For example, the amino acid pair "LL" appears 65 times in the replicase protein, while there is only one difference between actual and predicted frequencies. This means that "LL" is unlikely to be located at the functional sites although its frequency is the highest in the protein.

Although the amino acid pairs with a big difference between actual and predicted frequencies are more likely to be located at functional sites, they may be subject to the mutations. Our previous studies show that the amino acid pairs with a big difference between actual and predicted frequencies are more sensitive to new mutations, especially for the amino acid pairs whose actual frequency is larger than their predicted frequency [31–34]. In this case the anti-SARS drugs would be ineffective if their targets undergo mutations. But the anti-SARS drugs will still have effects if the targets are the amino acid pairs formed through mutations. So, the amino acid pairs with smaller actual frequency than their predicted frequency is preferred, as they are more likely to be formed through mutations [31–35]. In addition, this does not require knowledge of the possible outcome of a mutation.

To find the amino acid pairs less sensitive to mutations, we calculate the first-order Markov transition probability (Fig. 3). For instance, the amino acid pairs "AN" and "PL" have the same values of both actual frequency (AF = 31)

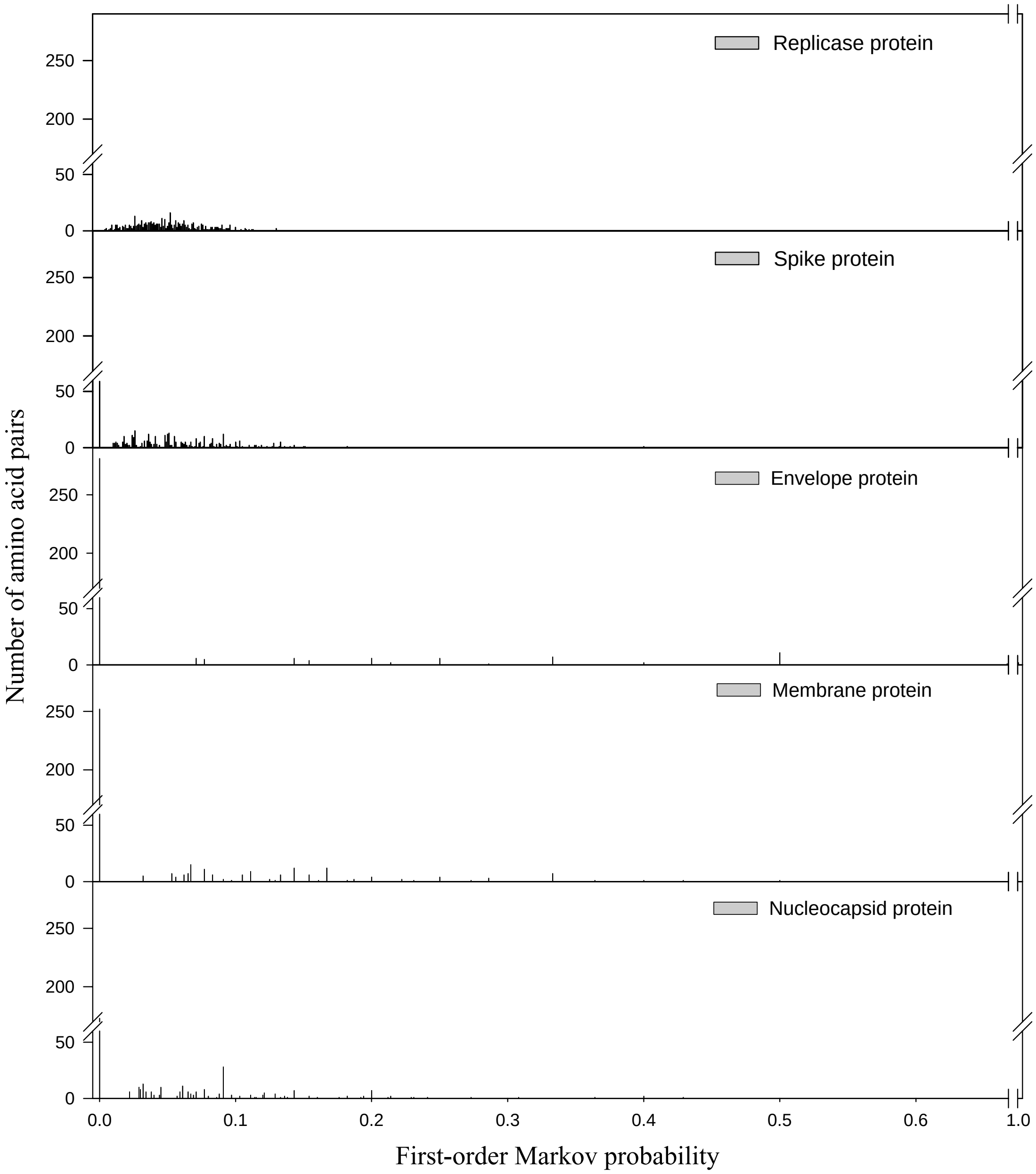

Fig. 3. The first-order Markov transition probability in the structural proteins from SARS-CoV.

Table 1
Some possible targets for anti-SARS drugs in structural proteins from human SARS-CoV

Protein	Pair	Repetition	AF–PF	Markov probability
Replicase	LV	41	−14	0.061
	GL	29	−11	0.069
	LT	40	−7	0.059
	KV	24	−10	0.058
	TA	21	−15	0.042
	DL	34	−4	0.086
	VS	31	−7	0.053
	TV	37	−4	0.075
	VG	24	−10	0.041
Spike	LS	5	−3	0.051
	SL	6	−2	0.063
	TL	6	−2	0.061
	FL	5	−2	0.06
	TV	4	−3	0.041
	LA	4	−3	0.04
Envelope	GQ	3	−1	0.067
	GS	3	−1	0.067
	KG	2	−1	0.069
Membrane	IL	1	−2	0.056
	AL	1	−2	0.053
Nucleocapsid	VL	1	−2	0.077

and predicted frequency (PF = 26) in the replicase protein, but a difference can be found in their Markov transition probability, which is 0.061 for “AN” and 0.113 for “PL”. Therefore, the amino acid pair “PL” rather than “AN” is preferable as the potential target for drugs, because the “PL” is more stable than the “AN”.

Taking the above three factors into account, we can theoretically define the possible targets for the development of anti-SARS drugs. Table 1 lists the amino acid pairs which can serve as the potential targets for anti-SARS drugs in case about 10% of amino acid pairs are chosen in each protein.

4. Discussion

In this study we outline the selection of the potential targets for anti-SARS drugs in the structural proteins from SARS-CoV at the stage of no detailed knowledge on these proteins, on their functional sites, and on their mutation patterns.

In such a situation, we can assume that the interaction between anti-SARS drugs and SARS-CoV is a random collision [28]. The abundant amino acids in the structural proteins from SARS-CoV would have a greater chance to collide with anti-SARS drugs [9,12] if each amino acid had an equal chance to be exposed to the drugs. In this manner we would have the first group of candidates as potential targets for anti-SARS drugs.

Although a single amino acid can be the target of anti-SARS drugs [19], the targeted amino acid (except for the one at terminal) has a connection with two neighboring

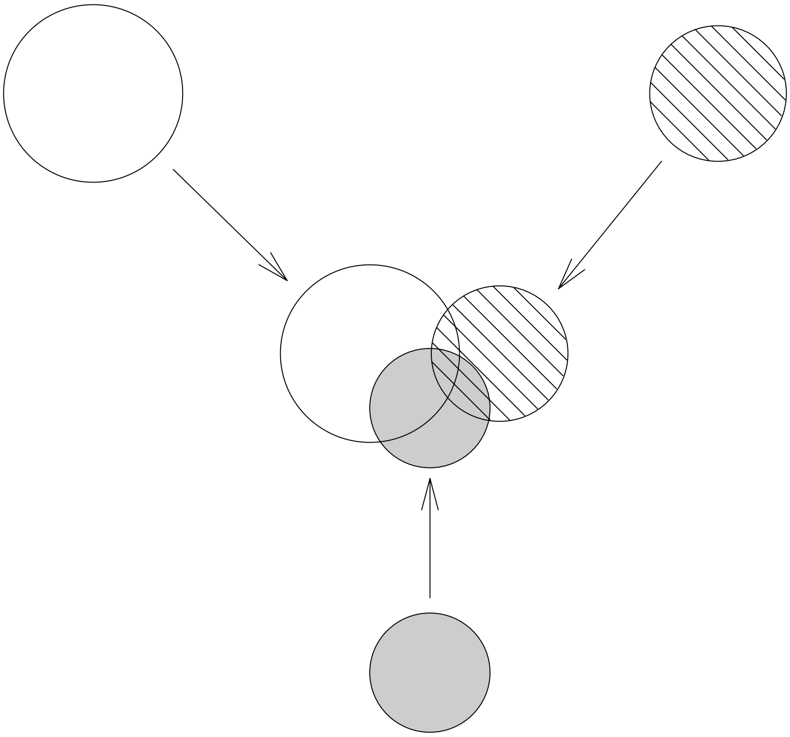

Fig. 4. The ideally targets (intersection among three circles) for anti-SARS drugs in relation to the amino acid pairs (open circle) grouped according to their frequencies in proteins, the amino acid pairs (lined circle) grouped according to the difference between actual and predicted frequencies and the amino acid pairs (gray circle) grouped according to their first-order Markov probability.

amino acids, whose connection constructs two amino acid pairs. We, therefore, regard the amino acid pairs as the basic unit for analysis following the concept that a good signature pattern of a protein must be as short as possible, but the conserved sequence is not longer than four or five residues [13]. Then the finding of abundant amino acids is changed into the finding of abundant amino acid pairs. Taking the frequency of amino acid pairs in proteins as a measure, we scale down the search coverage of candidates as potential targets for anti-SARS drugs. This group of candidates constructs the aggregate of the open circle in Fig. 4. Ideally this circle should only contain the amino acid pairs exposed to anti-SARS drugs.

Although the amino acid pairs selected by abundance have a greater chance to interact with anti-SARS drugs, the abundance does not directly represent the functional activity of amino acid pairs. We need to further narrow down the search coverage by calculating the difference between actual and predicted frequencies, which form the aggregate of lined circle in Fig. 4. Ideally, the lined circle should only include the amino acid pairs at functional sites in proteins. The amino acid pairs with a big difference between actual and predicted frequencies from the abundant amino acid pairs constitute the intersection between the open and lined circles in Fig. 4.

Finally, we cannot neglect the possible mutations in SARS-CoV, which may lead to the inefficacy of anti-SARS drugs as shown in the drug development [16]. The first-order Markov transition probability determines the stability of amino acid pairs that are grouped in the gray circle in Fig. 4. As a result, we have the candidates which are more likely at functional sites and less vulnerable to mutations. These candidates are enclosed in the intersection between the lined and gray circles. We also have the candidates which is the intersection between the open and gray circles. They not only appear more frequently, but also are less vulnerable to mutations. After balancing the three factors, our search is converged to the candidates enclosed in the interaction among three circles, which have a great chance to collide with anti-SARS drugs, and are more likely to link with the functional sites in the structural proteins and less vulnerable to mutations. Our probabilistic model provides, at least partly, a conceptual framework of how to select potential targets for design of anti-SARS drugs.

Acknowledgments

The authors are grateful to the anonymous referees for their deeply insightful comments, which enhance the points presented in this study. The special thanks go to Dr. Adrian Mulchinock at the Portal Co. Ltd., London, UK, the professional interpreter Mr. Arnold Lin at the People's Insurance Company of China and the first referee for correcting the English in the previous versions of the manuscript.

References

[1] Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science 2003;300:1763–7.

[2] Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. Nucl Acids Res 2000;28:45–8.

[3] Bermejo Martin JF, Jimenez JL, Munoz-Fernandez A. Pentoxifylline and severe acute respiratory syndrome (SARS): a drug to be considered. Med Sci Monit 2003;9:SR29–34.

[4] Bosch BJ, van der Zee R, de Haan CA, Rottier PJ. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol 2003;77:8801–11.

[5] Chen LL, Ou HY, Zhang R, Zhang CT. ZCURVE_CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. Biochem Biophys Res Commun 2003;307:382–8.

[6] Chou KC, Wei DQ, Zhong WZ. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. Biochem Biophys Res Commun 2003;308:148–51.

[7] Davidson A, Siddell S. Potential for antiviral treatment of severe acute respiratory syndrome. Curr Opin Infect Dis 2003;16:565–71.

[8] Feller W. An introduction to probability theory and its applications, 3rd ed., vol. I. New York: Wiley; 1968.

[9] Kim MS, Hashemi SB, Spencer PS, Sabri MI. Amino acid and protein targets of 1,2-diacetylbenzene, a potent aromatic gamma-diketone that induces proximal neurofilamentous axonopathy. Toxicol Appl Pharmacol 2002;183:55–65.

[10] Leparc-Goffart I, Hingley ST, Chua MM, Phillips J, Lavi E, Weiss SR. Targeted recombination within the spike gene of murine coronavirus mouse hepatitis virus-A59: Q159 is a determinant of hepatotropism. J Virol 1998;72:9628–36.

[11] Liu S, Pei J, Chen H, Zhu X, Liu Z, Ma W, et al. Modeling of the SARS coronavirus main proteinase and conformational flexibility of the active site. Beijing Da Xue Xue Bao 2003;35(Suppl):62–5.

[12] Moffatt J, Kennedy DO, Kojima A, Hasuma T, Yano Y, Otani S, et al. Involvement of protein tyrosine phosphorylation and reduction of cellular sulfhydryl groups in cell death induced by 1′-acetoxychavicol acetate in Ehrlich ascites tumor cells. Chem Biol Interact 2002;139:215–30.

[13] PROSITE: a dictionary of protein sites and patterns user manual, http://www.expasy.ch/prosite/.

[14] Sanchez CM, Izeta A, Sanchez-Morgado JM, Alonso S, Sola I, Balasch M, et al. Targeted recombination demonstrates that the spike gene of transmissible gastroenteritis coronavirus is a determinant of its enteric tropism and virulence. J Virol 1999;73:7607–18.

[15] Shen X, Xue JH, Yu CY, Luo HB, Qin L, Yu XJ, et al. Small envelope protein E of SARS: cloning, expression, purification, CD determination, and bioinformatics analysis. Acta Pharmacol Sin 2003;24:505–11.

[16] Szabo K, Bakos E, Welker E, Muller M, Goodfellow HR, Higgins CF, et al. Phosphorylation site mutations in the human multidrug transporter modulate its drug-stimulated ATPase activity. J Biol Chem 1997;272:23165–71.

[17] Tsui SK, Chim SS, Lo YM. Coronavirus genomic-sequence variations and the epidemiology of the severe acute respiratory syndrome. N Engl J Med 2003;349:187–8.

[18] van der Lubbe JCA. Information theory. Cambridge: Cambridge University Press; 1997.

[19] Vaughan MD, Sampson PB, Honek JF. Methionine in and out of proteins: targets for drug design. Curr Med Chem 2002;9:385–409.

[20] Williams RK, Yeager CL, Holmes KV. Potential for receptor-based antiviral drugs against SARS. Lancet 2003;362:77.

[21] Wu G. Frequency and Markov chain analysis of amino-acid sequence of human tumor necrosis factor. Cancer Lett 2000;153:145–50.

[22] Wu G. Frequency and Markov chain analysis of amino-acid sequence of human glutathione reductase. Biochem Biophys Res Commun 2000;268:823–6.

[23] Wu G. Frequency and Markov chain analysis of amino-acid sequences of mouse p53. Human Exp Toxicol 2000;19:535–9.

[24] Wu G, Yan SM. Prediction of two- and three-amino acid sequence of human acute myeloid leukemia 1 protein from its amino acid composition. Comp Haematol Int 2000;10:85–9.

[25] Wu G, Yan SM. Prediction of two- and three-amino-acid sequences of *Citrobacter Freundii* β-lactamase from its amino acid composition. J Mol Microbiol Biotechnol 2000;2:277–81.

[26] Wu G, Yan SM. Prediction of presence and absence of two- and three-amino-acid sequence of human monoamine oxidase B from its amino acid composition according to the random mechanism. Biomol Eng 2001;18:23–7.

[27] Wu G, Yan SM. Frequency and Markov chain analysis of amino-acid sequences of human connective tissue growth factor. J Mol Model 2001;5:120–4.

[28] Wu G, Yan SM. Mathematical model of time needed for the immune system to detect and kill cancer cells in blood. Comput Clin Pathol 2002;11:178–83.

[29] Wu G, Yan SM. Random analysis of presence and absence of two- and three-amino-acid sequences and distributions of amino acids, two- and three-amino-acid sequences in bovine p53 protein. Mol Biol Today 2002;3:31–7.

[30] Wu G, Yan SM. Randomness in the primary structure of protein: methods and implications. Mol Biol Today 2002;3:55–6.

[31] Wu G, Yan SM. Estimation of amino acid pairs sensitive to variants in human phenylalanine hydroxylase protein by means of a random approach. Peptides 2002;23:2085–90.

[32] Wu G, Yan S. Analysis of amino acid pairs sensitive to variants in human collagen a5(IV) chain precursor by means of a random approach. Peptides 2003;24:347–52.

[33] Wu G, Yan S. Determination of amino acid pairs sensitive to variants in human β-glucocerebrosidase by means of a random approach. Protein Eng 2003;16:195–9.

[34] Wu G, Yan S. Determination of amino acid pairs in human p53 protein sensitive to mutations/variants by means of a random approach. J Mol Model 2003;9:337–41.

[35] Wu G, Yan S. Prediction of amino acid pairs sensitive to mutations in the spike protein from SARS related coronavirus. Peptides 2003;24:1837–45.