Amino Acids (2008) 34: 81–90
DOI 10.1007/s00726-007-0579-z
Printed in The Netherlands

Amino Acids

Prediction of mutations engineered by randomness in H5N1 neuraminidases from influenza A virus

G. Wu and **S. Yan**

DreamSciTech Consulting, Guangdong Province, China

Received June 25, 2007
Accepted July 3, 2007
Published online August 28, 2007; © Springer-Verlag 2007

Summary. In this proof-of-concept study, we attempt to determine whether the cause-mutation relationship defined by randomness is protein dependent by predicting mutations in H5N1 neuraminidases from influenza A virus, because we have recently conducted several concept-initiated studies on the prediction of mutations in hemagglutinins from influenza A virus. In our concept-initiated studies, we defined the randomness as a cause for mutation, upon which we built a cause-mutation relationship, which is then switched into the classification problem because the occurrence and non-occurrence of mutations can be classified as unity and zero. Thereafter, we used the logistic regression and neural network to solve this classification problem to predict the mutation positions in hemagglutinins, and then used the amino acid mutating probability to predict the would-be-mutated amino acids. As the previous results were promising, we extend this approach to other proteins, such as H5N1 neuraminidase in this study, and further address various issues raised during the development of this approach. The result of this study confirms that we can use this cause-mutation relationship to predict the mutations in H5N1 neuraminidases.

Keywords: Influenza – Logistic regression – Mutation – Neuraminidase – Prediction

Introduction

In preparation for the possible epidemics and pandemics of influenza, an important issue is the prediction of mutated proteins of influenza A virus, because the unpredictable mutations lead humans to have little immunity against this deadly disease. Among various subtypes of influenza viruses, the H5N1 viruses are highly pathogenic (Lee et al., 2005; Chen et al., 2006), of which the mutations mainly occur in the RNA genes coding for ten virus proteins (Hilleman, 2002).

Neuraminidase is a sialidase (Gottschalk, 1957) that prevents virion aggregation by removing cell and virion surface sialic acid (Paulson, 1985), is the major antigen for neutralizing antibodies and is involved in the binding of virus particles to receptors on host cells (Zambon, 1999). Still, neuraminidase is the target of several anti-influenza drugs (Hochgürtel et al., 2002; Garman and Laver, 2004; Oxford et al., 2004). Of subtypes, H5N1 neuraminidase is important as H5N1 virus is currently threatening humans and the mutations in neuraminidase may lead to the dysfunction of anti-influenza drugs.

The preparedness is currently conduced along various approaches, of which the modeling is playing its role in this battle against influenza A virus. A prominent approach in developing inhibitors is conducted at several levels. At receptor protein level, the modeling helps to determine the ''binding pocket'' of the receptor protein with its ligands (Chou, 2004a–e, 2005; Chou et al., 1997, 1999, 2000, 2003, 2006; Li et al., 2007; Wang et al., 2007a, c). At ''cleavage-site'' level, the modeling is trying to find the target residue for mutagenesis (Poorman et al., 1991; Elhammer et al., 1993; Chou, 1993a, b, 1996; Thompson et al., 1995). Upon two levels above, it is generally possible to find the target residues, the next level study is directed to the mutagenesis and the designing of effective inhibitors (Althaus et al., 1993a–c; Chou et al., 1994; Du et al., 2005, 2007; Gan et al., 2006; Gao et al., 2007; Wei et al., 2007). The fourth level of modeling is the determination of 3D structure of binding interaction in proteins of interests (Wei et al., 2006; Wang et al., 2007b).

Recently we have tried to use the modeling approach to predict mutations in proteins from influenza A virus, which is related to several levels too, say, the prediction of mutation positions, the prediction of would-be-mutated amino acids at predicted positions, the timing of mutations, and the prediction of new functions resulting from

the mutations. The first three types of predictions are relevant to the primary structure of proteins, while the last one is relevant to 3D structure of proteins.

It is no doubt that the best way for the prediction of mutation is to find the cause for mutations, and then we can build a cause-mutation relationship, and predict the occurrence of mutation when its cause appears. This approach is quite straightforward, however it is challenged by three facts. First, many causes, which led mutations in the past, might nerve leave any trace due to the huge changes in environments, so we would have a relatively detailed record of mutations, but a poor record of their causes. Thus, we could not establish the one-to-one relationship between causes and mutations. Consequently, we even could not define the scale of causes for monitoring. Second, the current version of proteins might not be subject to the causes, which led the historic mutations, because of evolution, the multi-drug resistance could be an example for the evolution of bacteria. Third, it is difficult to find the historically macro- and micro-environmental surroundings, under which historic causes triggered the mutations.

As the searching of each historically instant cause appears difficult, we might need to direct our effort forward the searching of constant causes, because the protein constantly evolutes although its evolutionary speed is not constant. Randomness should be one of such constant causes, which engineer mutations through generation, not only because the pure chance is now considered to lie at the very heart of nature (Everitt, 1999) and the occurrence of mutation is generally considered a random event (Fitch et al., 1997), but also because randomness suggests that an event does not occur deliberately, but naturally. This further suggests that the event with a bigger probability would occur more easily than the event with a smaller probability. Although nature should deliberately construct the absolutely necessary structure for a protein with more time and energy, there must be some structures that can be explained by random mechanism, because not only nature follows parsimony, but also nature cannot predict the future by constructing the structure for the future, which are currently useless.

Once we could measure and quantify this randomness, we could compare the quantified randomness before and after mutations to determine whether randomness plays a role. If so, we could build a quantitative cause-mutation relationship to predict the mutations engineered by randomness.

Although it is difficult to measure and quantify the randomness in nature, we could measure and quantify the randomness in a protein, which mirrors the randomness in nature. Since 1999, we have developed three methods to quantify the randomness in protein, and find the quantified randomness sensitive to mutations. This means that the randomness does play an important role in engineering mutations or we can use the random mechanism to explain some mutations.

Furthermore, this also means that we can build a cause-mutation relationship accounting for the mutations engineered by randomness. This is possible, because we can classify the occurrence and non-occurrence of mutation as unity and zero. This way, the cause-mutation relationship is switched into the classification problem, which can be solved either using logistic regression or neural network. However, the occurrence and non-occurrence of mutation is a binary event, which means that we can only use this cause-mutation relationship to predict the mutation positions rather than the would-be-mutated amino acids at predicted positions.

For prediction of would-be-mutated amino acids, we have more difficulties to build a deterministic relationship or classification model. However, there are several common ways (Dayhoff et al., 1978; Feng et al., 1985; Karlin and Ghandour, 1985; Müller et al., 2002) as well as the amino acid mutating probability developed by us (Wu and Yan, 2005g, 2006a, 2007b) to solve this issue.

All these indicate that the prediction of mutations includes at least two steps, say, the prediction of mutation position and the prediction of would-be-mutated amino acids at predicted positions. Along this two-step frame, we very recently conducted several concept-initiated studies to test whether we can apply this cause-mutation relationship with logistic regression as well as neural network to predicting the mutation positions, and then apply the amino acid mutating probability to predicting the would-be-mutated amino acids at predicted mutation positions in hemagglutinins from influenza A virus (Wu and Yan, 2006e, f, 2007a, c, d).

As the results of these concept-initiated studies appear promising, we need to conduct many more proof-of-concept studies to determine whether this approach is dependent on different proteins, subtypes, etc., and to refine the approach and to clarify the related issues. Hence, we attempt to apply this approach to predicting the mutations in H5N1 neuraminidase from influenza A virus in this study.

Materials and methods

429 H5N1 neuraminidases from influenza A virus from 1996 to 2006 are obtained from influenza virus resources (Influenza virus resources, 2006).

As our approaches are not familiar to most researchers, we will explain them in great details.

Prediction model

In our two-step frame, the prediction model is only related to the cause-mutation relationship, which is switched to the classification problem. Thus, we use the logistic regression, whose output ranges between zero and unity, $P(y) = \frac{1}{1+e^{b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4+b_5x_5+b_6x_6+b_7x_7}}$, where x_i is the independent, y is the dependent, and b_i is the model parameters. As our previous studies shows that seven independents work better than six independents (Wu and Yan, 2006f, 2007a, d), we will use seven independents in modeling of neuraminidase in this study.

Independent I – amino-acid pair predictability

This quantification is calculated according to permutation, and we have used it to study various proteins (Wu, 1999, 2000a–g; Wu and Yan, 2000a–c, 2001a–c, 2002a–d, 2003a–h, 2004a–e, 2005a–d, f, 2006b, d–f, 2007a, c). Its rationale includes: (i) this is the simplest way to quantify the randomness in a protein, (ii) the counting of amino-acid pairs was inspired from modern encryption technology by counting the frequency of basic unit in an unknown language, and (iii) 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 (PROSITE, 2002), while our previous studies show the amino-acid pair the best for our aim. The practical meanings are that this amino-acid pair predictability is very sensitive to the change in neighboring amino acids, and answers why a type of amino acid is adjacent to a certain type of amino acid but not to the others.

The simplest calculations are as follows: according to the permutation, for example, there are 44 glycines (G) and 34 isoleucines (I) in 2005 AB239126 neuraminidase, the randomly predicted frequency of amino-acid pair "GI" is 3 ($44/449 \times 34/448 \times 448 = 3.3318$), that is, "GI" would appear three times in this neuraminidase, which is the predicted frequency and is the reference for comparison. Actually we do find 3 "GI", so "GI" is predictable and the difference between its actual and predicted frequency is 0. Again, there are 28 threonines (T) in AB239126 neuraminidase, and the randomly predicted frequency of "TI" is 2 ($28/449 \times 34/448 \times 448 = 2.1203$), i.e. there would be two "TI" in the neuraminidase. But the "TI" appears five times in reality, so the difference between its actual and predicted frequency is 3. After such calculations, each amino-acid pair has its difference between actual and predicted frequency. As a point mutation is related to a single amino acid, it connects with two neighboring amino acids except for the terminal one and constructs two amino-acid pairs, so each amino acid can have the sum of difference between actual and predicted frequency in two neighboring amino-acid pairs (SDAPF).

Independent II – percentage of SDAPF

This is a derivate quantification, because our previous studies show that the mutation minimizes the difference between actual and predicted frequency, and the bigger the difference is, the more vulnerable the amino-acid pair to mutation is (Wu and Yan, 2002a–c, f, 2003a–h, 2004a–c). As each amino acid in neuraminidase has its SDAPF, which generally ranges from -5 to 9, we count how many amino acids have SDAPF of 1, SDAPF of 2, and so on, then calculate their percentage with respect to all amino acids, and each amino acid has this percentage.

Independent III – interaction between SDAPF and its percentage

This quantification is based on the common consideration in regression, that is, the first-order interaction between independents is frequently included in regression analysis (Draper and Smith, 1981; Hosmer and Lemeshow, 2000). In our case, SDAPF and its percentage are closely related one another, and our previous studies suggest that the interaction significantly enhances the predictability (Wu and Yan, 2006e, f, 2007a, d). We therefore assign the first-order interaction to each amino acid.

Independent IV – amino-acid distribution probability

This quantification is calculated according to the occupancy of subpopulations and partitions (Feller, 1968), and we have used this quantification to study various proteins (Gao et al., 2006; Wu and Yan, 2000d, 2001d, e, 2002c–f, 2004f, 2005d, e, 2006c–f, 2007a, c, d).

The quantification is developed along such line of thought, for example, there are two methionines (M) among 141 amino acids in human hemoglobin α-chain (Wu and Yan, 2000d). With regard to their random distribution, our intuition may suggest that there would be one "M" in the first half of the chain and another "M" in the second half, which is true in real-life case. In fact, there are only three possible distributions of "M"s in human hemoglobin α-chain, i.e. (i) both "M"s are in the first half, (ii) one "M" is in each half and (iii) both "M"s are in the second half. If we do not distinguish either first half or second half but are simply interested in whether both "M"s are in both halves or in any half, we will have the probability of $1/2$ for each distribution.

If we are interested in the distribution probability of three amino acids in a protein sequence, we naturally imagine to group the protein into three parts, and our intuition may suggest that each part contains an amino acid. If we do not distinguish the first, second and third part, actually there are three types of distributions, i.e. (i) each part contains an amino acid, (ii) two amino acids are in a part and an amino acid in another part, and (iii) three amino acids are in a part. However, the distribution probabilities are different for them, say, 0.2222 for (i), 0.6667 for (ii) and 0.1111 for (iii). Clearly the protein can only adopt one type of distribution for these three amino acids, which is the actual distribution probability, and we may guess that the distribution (ii) is more likely to happen because of its biggest probability, which is the predicted distribution probability and is the reference for comparison.

For four amino acids, we will have five distribution probabilities, i.e. (i) each part contains an amino acid, (ii) a part contains two amino acids and two parts contain an amino acid each, (iii) two parts contain two amino acids each, (iv) a part contains an amino acid and a part contains three amino acids, and (v) a part contains four amino acids. Their distribution probabilities are 0.0938, 0.5625, 0.1406, 0.1875, 0.0156, respectively. Further, we have seven distributions for five amino acids, we have 11 distributions for six amino acids, we have 15 distributions for seven amino acids, and so on.

So we view the positions of each kind of amino acids in a protein as a certain distribution, whose probability can be calculated according to the equation of $r!/(q0! \times q1! \times \cdots \times qn!) \times r!/(r1! \times r2! \times \cdots \times rn!) \times n^{-r}$ (Feller, 1968), where ! is the factorial function, r is the number of a kind of amino acid, q is the number of parts with the same number of amino acids and n is the number of grouped parts in the protein for a kind of amino acid. In fact, this distribution probability can be referred to the statistical mechanics, which classifies the distribution of elementary particles in energy states according to three assumptions of whether or not distinguishing of each particle and energy state, i.e. Maxwell-Boltzmann, Fermi-Dirac and Bose-Einstein assumptions (Feller, 1968). In plain words, this distribution probability is the probability if we would receive seven letters in a week but the letters distribute randomly.

The practical meanings are that this quantification is mainly subject to any change in the position of amino acid, and answers why the majority of amino acids cluster in some regions rather than homogenously distribute along the primary structure of a protein.

With respect to neuraminidases in this study, for instance, there are 18 cysteines (C) in AB239126 neuraminidase. Its predicted and actual distribution probabilities are 0.1246 and 0.0138, so the ratio of predicted versus actual distribution probabilities is 9, whose natural logarithm is 2.1972 (LRPADP). In this way, each amino acid has its LRPADP.

Table 1. Amino acid mutating probability based on the translation probability between RNA codons and translated amino acids

Amino acid	Mutated amino acid with its translation probability
A	$12/36A + 2/36D + 2/36E + 4/36G + 4/36P + 4/36S + 4/36T + 4/36V$
R	$2/54C + 6/54G + 2/54H + 1/54I + 2/54K + 4/54L + 1/54M + 4/54P + 2/54Q + 18/54R + 6/54S + 2/54T + 2/54W + 2/54STOP$
N	$2/18D + 2/18H + 2/18I + 4/18K + 2/18N + 2/18S + 2/18T + 2/18Y$
D	$2/18A + 2/18D + 4/18E + 2/18G + 2/18H + 2/18N + 2/18V + 2/18Y$
C	$2/18C + 2/18F + 2/18G + 2/18R + 4/18S + 2/18W + 2/18Y + 2/18STOP$
E	$2/18A + 4/18D + 2/18E + 2/18G + 2/18K + 2/18Q + 2/18V + 2/18STOP$
Q	$2/18E + 4/18H + 2/18K + 2/18L + 2/18P + 2/18Q + 2/18R + 2/18STOP$
G	$4/36A + 2/36C + 2/36D + 2/36E + 12/36G + 6/36R + 2/36S + 4/36V + 1/36W + 1/36STOP$
H	$2/18D + 2/18H + 2/18L + 2/18N + 2/18P + 4/18Q + 2/18R + 2/18Y$
I	$2/27F + 6/27I + 1/27K + 4/27L + 3/27M + 2/27N + 1/27R + 2/27S + 3/27T + 3/27V$
L	$6/54F + 2/54H + 4/54I + 18/54L + 2/54M + 4/54P + 2/54Q + 4/54R + 2/54S + 1/54W + 6/54V + 3/54STOP$
K	$2/18E + 1/18I + 2/18K + 1/18M + 4/18N + 2/18Q + 2/18R + 2/18T + 2/18STOP$
M	$3/9I + 1/9K + 2/9L + 1/9R + 1/9T + 1/9V$
F	$2/18C + 2/18F + 2/18I + 6/18L + 2/18S + 2/18V + 2/18Y$
P	$4/36A + 2/36H + 4/36L + 12/36P + 2/36Q + 4/36R + 4/36S + 4/36T$
S	$4/54A + 4/54C + 2/54F + 2/54G + 2/54I + 2/54L + 2/54N + 4/54P + 6/54R + 14/54S + 6/54T + 1/54W + 2/54Y + 3/54STOP$
T	$4/36A + 3/36I + 2/36K + 1/36M + 2/36N + 4/36P + 2/36R + 6/36S + 12/36T$
W	$2/9C + 1/9G + 1/9L + 2/9R + 1/9S + 2/9STOP$
Y	$2/18C + 2/18D + 2/18F + 2/18H + 2/18N + 2/18S + 2/18Y + 4/18STOP$
V	$4/36A + 2/36D + 2/36E + 2/36F + 4/36G + 3/36I + 6/36L + 1/36M + 12/36V$
STOP	$1/27C + 2/27E + 1/27G + 2/27K + 3/27L + 2/27Q + 2/27R + 3/27S + 2/27W + 4/27Y + 4/27STOP$

A alanine; *R* arginine; *N* asparagine; *D* aspartic acid; *C* cysteine; *E* glutamic acid; *Q* glutamine; *G* glycine; *H* histidine; *I* isoleucine; *L* leucine; *K* lysine; *M* methionine; *F* phenylalanine; *P* proline; *S* serine; *T* threonine; *W* tryptophan; *Y* tyrosine; *V* valine

Independents V and VI – percentage of LRPADP and interaction between LRPADP and its percentage

With the similar consideration of independents II and III, we give the percentage of LRPADP and the first order interaction between LRPADP and its percentage to each amino acid.

Independent VII – future composition of amino acids

This quantification is calculated according to the translation probability between RNA codons and translated amino acids (Wu and Yan, 2005g, 2006a, 2007b), and we have used this quantification to study various proteins (Wu and Yan, 2005g, 2006a, f, 2007a–d).

This quantification is developed along such line of thought, for example, we are interested in the amino acid threonine and its mutated amino acids with their mutating probability. As the RNA codons have the unambiguous relationship with their translated amino acids, we can extend this question to RNA level, this is, a point mutation in RNA codon leads to the mutation at amino acid level.

Threonine is related to RNA codons ACU, ACC, ACA, and ACG, the mutation at the first position of ACU can lead ACU to mutate to CCU, GCU, and UCU, which correspond to threonine to mutate to proline, alanine, and serine at amino acid level. Similarly, the mutation at second position of ACU can result in isoleucine, asparagine, and serine, the mutation at the third position of ACU can result in threonine, threonine and threonine. Taken four RNA codons together, threonine would mutate in such a way, say, 4 alanines + 2 arginines + 2 asparagines + 3 isoleucines + 2 lysines + methionine + 4 prolines + 6 serines + 12 threonines. Thus we have the threonine mutating probability to these amino acids, say, $4/36 + 2/36 + 2/36 + 3/36 + 2/36 + 1/36 + 4/36 + 6/36 + 12/36$. For all 20 kinds of amino acids, we have the amino acid mutating probability in Table 1.

For the calculation of future composition of amino acids, we have the following steps: (i) We would expect that "A" has the 12/36 chance of mutating to "A" (line 2 in Table 1), "R" and "N" have no chance of mutating to "A" (lines 3 and 4 in Table 1), "D" has 2/18 chance (line 5 in Table 1), "C" has no chance (line 6 in Table 1), "E" has 2/18 chance, and so on. (ii) Meanwhile, AB239126 neuraminidase has 18 "A", 16 "R", 29 "N", 21 "D", 18 "C", 20 "E", and so on. (iii) So we can estimate how many "A" can be mutated using $18 \times 12/36 + 16 \times 0 + 29 \times 0 + 21 \times 2/18 + 18 \times 0 + 20 \times 2/18+$, and so on. In total, this is the future composition of amino acid "A". (iv) After calculated all 20 kinds of amino acids, "A" contributes 6.3374% to the future composition of neuraminidase, which is the predicted composition and is the reference for comparison. (v) On the other hand, "A" contributes 4% (18/450) to the current composition of AB239126 neuraminidase. (vi) Thus, we have the ratio of future versus current compositions, for example, the ratio of "A" is 1.5844 (6.3374%/4%), which can be assigned to each "A" in AB239126 neuraminidase.

The practical meanings are that this quantification is mainly subject to the future mutation trend, and answers with what probability an amino acid mutates to another type of amino acid.

Dependent – occurrence or non-occurrence of mutation

The phylogenetics analyses the evolutionary process of neuraminidases in question. Along same branch of the evolutionary tree, we can compare the parent and daughter neuraminidases, the difference between them indicates the occurrence of mutation, which is marked as unity, whereas no difference between them indicates the non-occurrence of mutation, which is marked as zero.

Would-be-mutated amino acid

To predict the would-be-mutated amino acids at predicted positions, we can also use Table 1 to make the estimation, for example, we would like to know which type of amino acid "T" would mutate to, according to Table 1, we find that "T" has the highest mutating probability (12/36), however this is only the case that "T" mutates to "T", then the next to the highest probability is the one that would be likely to be mutated, that is, "S" has 6/36 probability of occurrence. This way, we can approximately predict the would-be-mutated amino acids at the predicted positions.

Statistics

The SigmaStat (SPSS 1992–2003) and Systat (Systat Software 2004) are used to conduct all the logistic regressions. The outlier (3SD) is calculated according to Healy (1979). The prediction sensitivity, specificity and total correct rate are calculated according to the method mentioned in Systat software (Systat Software 2004). The Chi-square test is performed for comparison.

Results and discussion

After all the calculations, each parent neuraminidase has seven independents and one dependent for each amino acid of its sequence, for example, Table 2 shows a fraction of a neuraminidase after the calculation, where each amino acid is associated with seven independents and one dependent, which is determined by comparing 1996 AAD51926 and 1997 AAK38299 neuraminidases. Thus, we can input this format of data into the logistic regression to obtain the model parameters.

In modeling, we use the so-called population estimates to make predictions, and we have technically two ways to obtain the population estimates, either by calculating mean $\pm$ SD of all obtained model parameters or by pooling the data into a representative. We use the second method in this study because the logistic regression does not appear powerful enough to capture the mutation in each sequence, which is particularly related to the ratio of number of mutations to the length of sequence, although neuraminidase is generally longer than hemagglutinin and the logistic regression functions better.

In our previous studies (Wu and Yan, 2006e, f), we used the linear regression to evaluate the prediction performance, which is a very traditional method for evaluation of prediction performance. However, we soon realized the limitation of linear regression in context with the prediction of mutation. This is because we generally have the paired datasets for linear regression, for example, we might have the measured and predicted blood drug concentrations at certain time points, and then we can use the linear regression to regress them and get the correlation coefficient. However, this is not suitable for the prediction of mutation, for example, we might have five actual mutation positions, but four predicted mutation positions. In such a case, it would be difficult to use the linear regression because of unpaired datasets. Still, we also cannot use the linear regression for evaluation of would-be-mutated amino acids, because an amino acid can mutate to several different types of amino acids, which cannot be considered as paired cases.

To overcome this difficulty, we used the percentage of captured positions for evaluation (Wu and Yan, 2007a, d), and more recently we use the prediction sensitivity, specificity and total correct rate (Wu and Yan, 2007c) according to the method mentioned in Systat software (Systat Software 2004) because we can classify the predicted mutation positions as the positives, false positives, negatives and false negatives when comparing the predicted with the actual mutation positions. Thus, the percentage of captured positions in our previous studies (Wu and Yan, 2007a, d) is in fact equal to the total correct rate.

As can be seen in Fig. 1, the prediction pattern of H5N1 neuraminidase is similar to the prediction patterns of other hemagglutinins although we can find the statistical difference. However, the statistical difference is mainly found between the prediction in hemagglutinins with distinguishing arginine, leucine and serine and others. This is very suggestive because it implies our research direction in near future, say, to conduct the prediction at RNA codon level as a single mutation in RNA codon level may not lead to the mutation at amino acid level such as "A" has $12/36$ chance of mutating "A" in Table 1.

Table 2. Independents and dependent of AAD51926 neuraminidase

Position	Amino acid	Independents							Dependent
		I	II	III	IV	V	VI	VI	
1	M	2	22.3404	44.6809	0.4055	1.4894	0.6039	1.1468	0
...	...								
16	V	1	21.4894	21.4894	1.4962	6.8085	10.1868	0.9826	0
17	V	0	12.3404	0.0000	1.4962	6.8085	10.1868	0.9826	1
18	G	0	12.3404	0.0000	2.1752	9.5745	20.8265	0.7309	0
19	I	2	22.3404	44.6809	2.4812	8.0851	20.0611	0.7020	0
20	I	5	3.6170	18.0851	2.4812	8.0851	20.0611	0.7020	1
21	S	3	16.5957	49.7872	1.0498	11.4894	12.0618	0.8347	0
...	...								
469	K	2	22.3404	44.6809	2.9957	4.0426	12.1104	0.9620	0

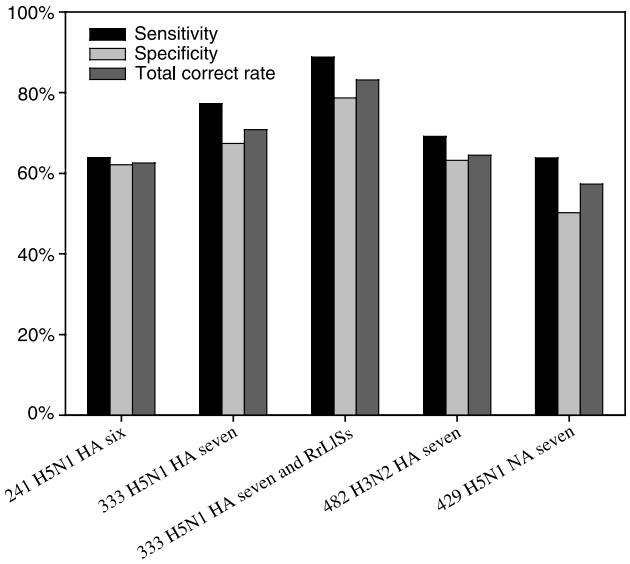

Fig. 1. Prediction performance in studied proteins. The sensitivity is equal to the predicted positives/the actual mutations (%), the specificity is equal to the predicted negatives/the actual non-mutations (%), and the total correct rate is equal to (predicted positives + predicted negatives)/length of hemagglutinin (%). 241 H5N1 HA six is the predictions using 241 H5N1 hemagglutinins with six independents; 333 H5N1 HA seven is the predictions using 333 H5N1 hemagglutinins with seven independents; 333 H5N1 HA seven with RrLlSs is the predictions using 333 H5N1 hemagglutinins with seven independents with distinguishing arginine, leucine and serine; 482 H3N2 HA seven is the predictions using 482 H3N2 hemagglutinins with seven independents; 429 H5N1 NA seven is the predictions using 429 H5N1 neuraminidases with seven independents. The Chi-square test indicates the statistically significant difference in sensitivity between 241 H5N1 HA six and 333 H5N1 HA seven with RrLlSs, between 333 H5N1 HA seven and 333 H5N1 HA seven with RrLlSs, between 333 H5N1 HA seven with RrLlSs and 482 H3N2 HA seven, between 333 H5N1 HA seven with RrLlSs and 429 H5N1 NA seven; the statistically significant difference in specificity between 241 H5N1 HA six and 333 H5N1 HA seven with RrLlSs, between 333 H5N1 HA seven and 429 H5N1 NA seven, between 333 H5N1 HA seven with RrLlSs and 482 H3N2 HA seven, between 333 H5N1 HA seven with RrLlSs and 429 H5N1 NA seven; the statistically significant difference in total correct rate between 241 H5N1 HA six and 333 H5N1 HA seven with RrLlSs, between 333 H5N1 HA seven with RrLlSs and 482 H3N2 HA seven, between 333 H5N1 HA seven with RrLlSs and 429 H5N1 NA seven

Still, Fig. 1 suggests that the cause-mutation relationship defined is independent of subtypes of protein as well as proteins, at least for hemagglutinins and neuraminidases. In this case, it means that the randomness does play a mutating role not only in hemagglutinins but also in neuraminidases although we need to conduct more proof-of-concept studies to further determine this issue.

This way, we can obtain the population estimates from regressing historical data. Thereafter, we can input seven independents of recent neuraminidases, whose mutations are yet to know, into the logistic regression with population estimates and get the output, which is ranged from 0 to 1 in each position. For example, the human H5N1 neuraminidase (AB239126) is a relatively new sequence, which can serve for our prediction. For this neuraminidase, we have seven independents (Table 3), and then we put them into $P(y) = \frac{1}{1+e^{0.42-0.214x_1-0.061x_2-0.156x_3-0.021x_4+0.11x_5+0.016x_6+0.035x_7}}$ which is based on the population of 90 neuraminidases from 2000 to 2004.

Figure 2 displays the prediction of mutation in AB239126 H5N1 neuraminidase according to our two-step frame. The solid line in the lower panel is the predicted mutation probability with respect to each position, and the dash-dotted line is the cut-off mutation probability of 0.5, that is, the amino acid whose mutation probability is larger than 0.5 risks mutation. The pie picture in the upper panel shows how to predict the would-be-mutated amino acid from serine at position 319 according to the amino acid mutating probability in Table 1.

With the population estimates as model parameters for prediction, an important issue is the sampling strategy (Wu et al., 1995, 1996; Wu, 1997), that is, from which population we get the population estimates, not only because there are many subtypes in neuraminidases (Air et al.,

Table 3. Independents of AB239126 neuraminidase

Position	Amino acid	Independents						
		I	II	III	IV	V	VI	VI
1	M	1	22.2222	22.2222	2.1972	5.7778	12.6951	0.9433
2	N	1	22.2222	22.2222	0.3365	6.4444	2.1684	0.6475
3	P	4	8.8889	35.5556	3.5766	4.8889	17.4854	0.9764
4	N	5	4.6667	23.3333	0.3365	6.4444	2.1684	0.6475
5	Q	2	21.3333	42.6667	2.8134	2.4444	6.8772	0.9832
6	K	0	12.8889	0.0000	0.8755	4.6667	4.0855	0.7919
7	I	−1	5.1111	−5.1111	1.0527	7.5556	7.9535	0.7410
8	I	1	22.2222	22.2222	1.0527	7.5556	7.9535	0.7410
...	...							
449	K	2	21.3333	42.6667	0.8755	4.6667	4.0855	0.7919

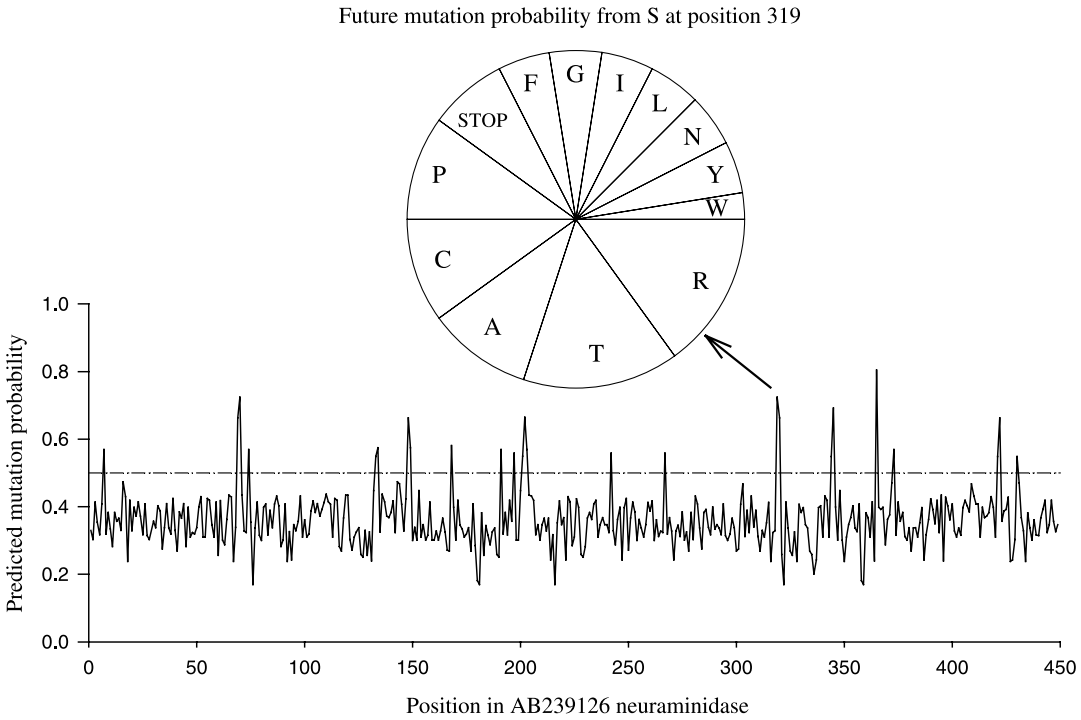

Fig. 2. Prediction of mutations in AB239126 neuraminidase based on logistic regression (lower panel) and amino acid mutating probability (upper panel)

1985; Schreier et al., 1988; Harley et al., 1989; Liu et al., 2003; Campitelli et al., 2004; Suzuki et al., 2004; Bragstad et al., 2005) but also the migration of wild birds is different one from another (Donis et al., 1989; Rohm et al., 1995; Hoffmann et al., 2000; Guan et al., 2004; Krauss et al., 2004; Wu and Yan, 2005e). This implies that the population estimates obtained from Asian wild bird may not be suited for the prediction of mutation in wild bird in North America, which nevertheless needs more studies. Suggestive is that we may have many different population estimates, based on which we make the predictions, which of course needs more studies.

References

Air GM, Ritchie LR, Laver WG, Colman PM (1985) Gene and protein sequence of an influenza neuraminidase with hemagglutinin activity. Virology 145: 117–122

Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1993a) Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem 268: 6119–6124

Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1993b) Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32: 6548–6554

Althaus IW, Gonzales AJ, Chou JJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1993c) The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem 268: 14875–14880

Bragstad K, Jorgensen PH, Handberg KJ, Mellergaard S, Corbet S, Fomsgaard A (2005) New avian influenza A virus subtype combination H5N7 identified in Danish mallard ducks. Virus Res 109: 181–190

Campitelli L, Mogavero E, De Marco MA, Delogu M, Puzelli S, Frezza F, Facchini M, Chiapponi C, Foni E, Cordioli P, Webby R, Barigazzi G, Webster RG, Donatelli I (2004) Interspecies transmission of an H7N3 influenza virus from wild birds to intensively reared domestic poultry in Italy. Virology 323: 24–36

Chen H, Smith GJ, Li KS, Wang J, Fan XH, Rayner JM, Vijaykrishna D, Zhang JX, Zhang LJ, Guo CT, Cheung CL, Xu KM, Duan L, Huang K, Qin K, Leung YH, Wu WL, Lu HR, Chen Y, Xia NS, Naipospos TS, Yuen KY, Hassan SS, Bahri S, Nguyen TD, Webster RG, Peiris JS, Guan Y (2006) Establishment of multiple sublineages of H5N1 influenza virus in Asia: implications for pandemic control. Proc Natl Acad Sci USA 103: 2845–2850

Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 268: 16938–16948

Chou JJ (1993a) A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. Biopolymers 33: 1405–1414

Chou JJ (1993b) Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. J Protein Chem 12: 291–302

Chou KC (1996) Prediction of HIV protease cleavage sites in proteins. Anal Biochem 233: 1–14

Chou KC (2004a) Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. Biochem Biophys Res Commun 319: 433–438

Chou KC (2004b) Insights from modelling the tertiary structure of BACE2. J Proteome Res 3: 1069–1072

Chou KC (2004c) Molecular therapeutic target for type-2 diabetes. J Proteome Res 3: 1284–1288

Chou KC (2004d) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11: 2105–2134

Chou KC (2004e) Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. Biochem Biophys Res Commun 316: 636–642

Chou KC (2005) Modeling the tertiary structure of human cathepsin-E. Biochem Biophys Res Commun 331: 56–60

Chou KC, Jones D, Heinrikson RL (1997) Prediction of the tertiary structure and substrate binding site of caspase-8. FEBS Lett 419: 49–54

Chou KC, Kezdy FJ, Reusser F (1994) Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal Biochem 221: 217–230

Chou KC, Tomasselli AG, Heinrikson RL (2000) Prediction of the tertiary structure of a caspase-9/inhibitor complex. FEBS Lett 470: 249–256

Chou KC, Watenpaugh KD, Heinrikson RL (1999) A model of the complex between cyclin-dependent kinase 5(Cdk5) and the activation domain of neuronal Cdk5 activator. Biochem Biophys Res Commun 259: 420–428

Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ (2006) Progress in computational approach to drug development against SARS. Curr Med Chem 13: 3263–3270

Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. Biochem Biophys Res Comm 308: 148-151 (Erratum: ibid, 2003, 310, 675)

Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in protein. Atlas Protein Seq Struct 5: 345–352

Donis RO, Bean WJ, Kawaoka Y, Webster RG (1989) Distinct lineages of influenza virus H4 hemagglutinin genes in different regions of the world. Virology 169: 408–417

Draper NR, Smith H (1981) Applied regression analysis, 2nd ed. Wiley, New York

Du QS, Mezey PG, Chou KC (2005) Heuristic molecular lipophilicity potential (HMLP): a 2D-QSAR study to LADH of molecular family pyrazole and derivatives. J Comput Chem 26: 461–470

Du QS, Sun H, Chou KC (2007) Inhibitor design for SARS coronavirus main protease based on ''distorted key theory''. Med Chem 3: 1–6

Everitt BS (1999) Chance rules: an informal guide to probability, risk, and statistics. Springer, New York

Feller W (1968) An introduction to probability theory and its applications, 3rd ed, Vol. I. Wiley, New York

Feng DF, Johnson MS, Doolittle RF (1985) Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 21: 112–125

Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 94: 7712–7718

Garman E, Laver G (2004) Controlling influenza by inhibiting the virus's neuraminidase. Curr Drug Targ 5: 119–136

Gao N, Yan S, Wu G (2006) Pattern of positions sensitive to mutations in human haemoglobin α-chain. Protein Pept Lett 13: 101–107

Gao WN, Wei DQ, LiY, Gao H, Xu WR, Li AX, Chou KC (2007) Agaritine and its derivatives are potential inhibitors against HIV proteases. Med Chem 3: 221–226

Gottschalk A (1957) The specific enzyme of influenza virus and Vibrio cholerae. Biochim Biophys Acta 23: 645–646

Guan Y, Poon LL, Cheung CY, Ellis TM, Lim W, Lipatov AS, Chan KH, Sturm-Ramirez KM, Cheung CL, Leung YH, Yuen KY, Webster RG, Peiris JS (2004) H5N1 influenza: a protean pandemic threat. Proc Natl Acad Sci USA 101: 8156–8161

Harley VR, Ward CW, Hudson PJ (1989) Molecular cloning and analysis of the N5 neuraminidase subtype from an avian influenza virus. Virology 169: 239–243

Healy MJR (1979) Outliers in clinical chemistry quality-control schemes. Clin Chem 25: 675–677

Hilleman MR (2002) Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control. Vaccine 20: 3068–3087

Hoffmann E, Stech J, Leneva I, Krauss S, Scholtissek C, Chin PS, Peiris M, Shortridge KF, Webster RG (2000) Characterization of the influenza A virus gene pool in avian species in southern China: was H6N1 a derivative or a precursor of H5N1? Virology 74: 6309–6315

Hochgürtel M, Kroth H, Piecha D, Hofmann MW, Nicolau C, Krause S, Schaaf O, Sonnenmoser G, Eliseev AV (2002) Target-induced formation of neuraminidase inhibitors from in vitro virtual combinatorial libraries. Proc Natl Acad Sci USA 99: 3382–3387

Hosmer DW Jr, Lemeshow S (2000) Applied logistic regression, 2nd ed. Wiley, New York

Influenza virus resources (2006) http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/multiple.cgi

Karlin S, Ghandour G (1985) Multiple-alphabet amino acid sequence comparisons of the immunoglobulin kappa-chain constant domain. Proc Natl Acad Sci USA 82: 8597–8601

Krauss S, Walker D, Pryor SP, Niles L, Chenghong L, Hinshaw VS, Webster RG (2004) Influenza A viruses of migrating wild aquatic birds in North America. Vector Borne Zoonotic Dis 4: 177–189

Lee CW, Suarez DL, Tumpey TM, Sung HW, Kwon YK, Lee YJ, Choi JG, Joh SJ, Kim MC, Lee EK, Park JM, Lu X, Katz JM, Spackman E, Swayne DE, Kim JH (2005) Characterization of highly pathogenic H5N1 avian influenza A viruses isolated from South Korea. J Virol 79: 3692–3702

Li L, Wei DQ, Wang JF, Chou KC (2007) Computational studies of the binding mechanism of calmodulin with chrysin. Biochem Biophys Res Commun 358: 1102–1107

Liu M, He S, Walker D, Zhou N, Perez DR, Mo B, Li F, Huang X, Webster RG, Webby RJ (2003) The influenza virus gene pool in a poultry market in South central China. Virology 305: 267–275

Müller T, Spang R, Vingron M (2002) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol 19: 8–13

Oxford J, Balasingam S, Lambkin R (2004) A new millennium conundrum: how to use a powerful class of influenza anti-neuraminidase drugs (NAIs) in the community. J Antimicrob Chemother 53: 133–136

Paulson JC (1985) Interactions of animal viruses with cell surface receptors. In: Connor M (ed) The receptors. Academic Press, Orlando, pp 131–219

Rohm C, Horimoto T, Kawaoka Y, Suss J, Webster RG (1995) Do hemagglutinin genes of highly pathogenic avian influenza viruses constitute unique phylogenetic lineages? Virology 209: 664–670

Schreier E, Roeske H, Driesel G, Kunkel U, Petzold DR, Berlinghoff R, Michel S (1988) Complete nucleotide sequence of the neuraminidase gene of the human influenza virus A/Chile/1/83 (H1N1). Arch Virol 99: 271–276

Suzuki T, Takahashi T, Saito T, Guo CT, Hidari KI, Miyamoto D, Suzuki Y (2004) Evolutional analysis of human influenza A virus N2 neuraminidase genes based on the transition of the low-pH stability of sialidase activity. FEBS Lett 557: 228–232

Thompson TB, Chou KC, Zheng C (1995) Neural network prediction of the HIV-1 protease cleavage sites. J Theor Biol 177: 369–379

Wang JF, Wei DQ, Li L, Zheng SY, Li YX, Chou KC (2007a) 3D structure modeling of cytochrome P450 2C19 and its implication for personalized drug design. Biochem Biophys Res Commun 355: 513–519 (Corrigendum: ibid, 2007, 357, 330)

Wang JF, Wei DQ, Lin Y, Wang YH, Du HL, Chou KC (2007c) Insights from modeling the 3D structure of NAD(P)H-dependent D-xylose reductase of Pichia stipitis and its binding interactions with NAD and NADP. Biochem Biophys Res Commun 359: 323–329

Wang SQ, Du QS, Chou KC (2007b) Study of drug resistance of chicken influenza A virus (H5N1) from homology-modeled 3D structures of neuraminidases. Biochem Biophys Res Commun 354: 634–640

Wei DQ, Du QS, Sun H, Chou KC (2006) Insights from modeling the 3D structure of H5N1 influenza virus neuraminidase and its binding interactions with ligands. Biochem Biophys Res Commun 344: 1048–1055

Wei H, Zhang R, Wang C, Zheng H, Chou KC, Wei DQ (2007) Molecular insights of SAH enzyme catalysis and their implication for inhibitor design. J Theor Biol 244: 692–702

Wu G (1997) An explanation for failure to predict cyclosporine area under the curve using a limited sampling strategy: a beginner's second note. Pharmacol Res 35: 547–552

Wu G (1999) The first and second order Markov chain analysis on amino acids sequence of human haemoglobin α-chain and its three variants with low O_2 affinity. Comp Haematol Int 9: 148–151

Wu G (2000a) Frequency and Markov chain analysis of amino-acid sequence of human glutathione reductase. Biochem Biophys Res Commun 268: 823–826

Wu G (2000b) Frequency and Markov chain analysis of amino-acid sequence of human tumor necrosis factor. Cancer Lett 153: 145–150

Wu G (2000c) Frequency and Markov chain analysis of amino-acid sequences of mouse p53. Human Exper Toxicol 19: 535–539

Wu G (2000d) Frequency and Markov chain analysis of the amino acid sequence of human alcohol dehydrogenase α-chain. Alcohol Alcohol 35: 302–306

Wu G (2000e) Frequency and Markov chain analysis of the amino-acid sequence of sheep p53 protein. J Biochem Mol Biol Biophys 4: 179–185

Wu G (2000f) The first, second and third order Markov chain analysis on amino acids sequence of human tyrosine aminotransferase and its variant causing tyrosinemia type II. Pediatr Relat Top 39: 37–47

Wu G (2000g) The first, second, third and fourth order Markov chain analysis on amino acids sequence of human dopamine β-hydroxylase. Mol Psychiatry 5: 448–451

Wu G, Baraldo M, Pea F, Cossettini P, Furlanut M (1995) Effects of different sampling strategies on predictions of blood cyclosporine concentrations in haematological patients with multidrug resistance by Bayesian and non-linear least squares methods. Pharmacol Res 32: 355–362

Wu G, Cossettini P, Furlanut M (1996) Prediction of blood cyclosporine concentrations in haematological patients with multidrug resistance by one-, two- and three-compartment models using Bayesian and non-linear least squares methods. Pharmacol Res 34: 47–57

Wu G, Yan S (2000a) Frequency and Markov chain analysis of amino-acids sequence of human platelet-activating factor acetylhydrolase α-subunit and its variant causing the lissencephaly syndrome. Pediatr Relat Top 39: 513–526

Wu G, Yan S (2000b) Prediction of two- and three-amino acid sequence of human acute myeloid leukemia 1 protein from its amino acid composition. Comp Haematol Int 10: 85–89

Wu G, Yan S (2000c) Prediction of two- and three-amino-acid sequences of Citrobacter Freundii β-lactamase from its amino acid composition. J Mol Microbiol Biotechnol 2: 277–281

Wu G, Yan S (2000d) Prediction of distributions of amino acids and amino acid pairs in human haemoglobin α-chain and its seven variants causing α-thalassemia from their occurrences according to the random mechanism. Comp Haematol Int 10: 80–84

Wu G, Yan S (2001a) Frequency and Markov chain analysis of amino-acid sequences of human connective tissue growth factor. J Mol Model 5: 120–124

Wu G, Yan S (2001b) 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 18: 23–27

Wu G, Yan S (2001c) Prediction of presence and absence of two- and three-amino-acid sequence of human tyrosinase from their amino acid composition and related changes in human tyrosinase variant causing oculocutaneous albinism. Pediatr Relat Top 40: 153–166

Wu G, Yan S (2001d) Analysis of distributions of amino acids, amino acid pairs and triplets in human insulin precursor and four variants from their occurrences according to the random mechanism. J Biochem Mol Biol Biophys 5: 293–300

Wu G, Yan S (2001e) Analysis of distributions of amino acids and amino acid pairs in human tumor necrosis factor precursor and its eight variants according to random mechanism. J Mol Model 7: 318–323

Wu G, Yan S (2002a) Determination of amino acid pairs sensitive to variants in human low-density lipoprotein receptor precursor by means of a random approach. J Biochem Mol Biol Biophys 6: 401–406

Wu G, Yan S (2002b) Estimation of amino acid pairs sensitive to variants in human phenylalanine hydroxylase protein by means of a random approach. Peptides 23: 2085–2090

Wu G, Yan S (2002c) 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 3: 31–37

Wu G, Yan S (2002d) Analysis of distributions of amino acids in the primary structure of apoptosis regulator Bcl-2 family according to the random mechanism. J Biochem Mol Biol Biophys 6: 407–414

Wu G, Yan S (2002e) Analysis of distributions of amino acids in the primary structure of tumor suppressor p53 family according to the random mechanism. J Mol Model 8: 191–198

Wu G, Yan S (2002f) Randomness in the primary structure of protein: methods and implications. Mol Biol Today 3: 55–69

Wu G, Yan S (2003a) Analysis of amino acid pairs sensitive to variants in human collagen α5(IV) chain precursor by means of a random approach. Peptides 24: 347–352

Wu G, Yan S (2003b) Determination of amino acid pairs in human haemoglobulin α-chain sensitive to variants by means of a random approach. Comp Clin Pathol 12: 21–25

Wu G, Yan S (2003c) Determination of amino acid pairs in human p53 protein sensitive to mutations/variants by means of a random approach. J Mol Model 9: 337–341

Wu G, Yan S (2003d) Determination of amino acid pairs in Von Hippel-Lindau disease tumour suppressor (G7 protein) sensitive to variants by means of a random approach. J Appl Res 3: 512–520

Wu G, Yan S (2003e) Determination of amino acid pairs sensitive to variants in human β-glucocerebrosidase by means of a random approach. Protein Eng 16: 195–199

Wu G, Yan S (2003f) Determination of amino acid pairs sensitive to variants in human Bruton's tyrosine kinase by means of a random approach. Mol Simul 29: 249–254

Wu G, Yan S (2003g) Determination of amino acid pairs sensitive to variants in human coagulation factor IX precursor by means of a random approach. J Biomed Sci 10: 451–454

Wu G, Yan S (2003h) Prediction of amino acid pairs sensitive to mutations in the spike protein from SARS related coronavirus. Peptides 24: 1837–1845

Wu G, Yan S (2004a) Amino acid pairs sensitive to variants in human collagen α1(I) chain precursor. EXCLI J 3: 10–19

Wu G, Yan S (2004b) Determination of amino acid pairs sensitive to variants in human copper-transporting ATPase 2. Biochem Biophys Res Commun 319: 27–31

Wu G, Yan S (2004c) Fate of 130 hemagglutinins from different influenza A viruses. Biochem Biophys Res Commun 317: 917–924

Wu G, Yan S (2004d) Potential targets for anti-SARS drugs in the structural proteins from SARS related coronavirus. Peptides 25: 901–908

Wu G, Yan S (2004e) Susceptible amino acid pairs in variants of human collagen α1(III) chain precursor. EXCLI J 3: 20–28

Wu G, Yan S (2004f) Determination of sensitive positions to mutations in human p53 protein. Biochem Biophys Res Commun 321: 313–319

Wu G, Yan S (2005a) Amino acid pairs susceptible to variants in human protein C precursor. Protein Pept Lett 10: 491–494

Wu G, Yan S (2005b) Mutation features of 215 polymerase proteins from different influenza A viruses. Med Sci Monit 11: BR367–BR372

Wu G, Yan S (2005c) Reasoning of spike glycoproteins being more vulnerable to mutations among 158 coronavirus proteins from different species. J Mol Model 11: 8–16

Wu G, Yan S (2005d) Timing of mutation in hemagglutinins from influenza A virus by means of unpredictable portion of amino-acid pair and fast Fourier transform. Biochem Biophys Res Commun 333: 70–78

Wu G, Yan S (2005e) Searching of main cause leading to severe influenza A virus mutations and consequently to influenza pandemics/epidemics. Am J Infect Dis 1: 116–123

Wu G, Yan S (2005f) Prediction of mutation trend in hemagglutinins and neuraminidases from influenza A viruses by means of cross-impact analysis. Biochem Biophys Res Commun 326: 475–482

Wu G, Yan S (2005g) Determination of mutation trend in proteins by means of translation probability between RNA codes and mutated amino acids. Biochem Biophys Res Commun 337: 692–700

Wu G, Yan S (2006a) Fate of influenza A virus proteins. Protein Pept Lett 13: 377–384

Wu G, Yan S (2006b) Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint. Acta Pharmacol Sin 27: 513–526

Wu G, Yan S (2006c) Timing of mutation in hemagglutinins from influenza A virus by means of amino-acid distribution rank and fast Fourier transform. Protein Pept Lett 13: 143–148

Wu G, Yan S (2006d) Determination of mutation trend in hemagglutinins by means of translation probability between RNA codons and mutated amino acids. Protein Pept Lett 13: 601–609

Wu G, Yan S (2006e) Prediction of possible mutations in H5N1 hemagglutinins of influenza A virus by means of logistic regression. Comp Clin Pathol 15: 255–261

Wu G, Yan S (2006f) Prediction of mutations in H5N1 hemagglutinins from influenza A virus. Protein Pept Lett 13: 971–976

Wu G, Yan S (2007a) Improvement of model for prediction of hemagglutinin mutations in H5N1 influenza viruses with distinguishing of arginine, leucine and serine. Protein Pept Lett 14: 191–196

Wu G, Yan S (2007b) Translation probability between RNA codons and translated amino acids, and its applications to protein mutations. In: Ostrovskiy MH (ed) Leading-edge messenger RNA research communications. Nova Science Publishers, New York

Wu G, Yan S (2007c) Improvement of prediction of mutation positions in H5N1 hemagglutinins of influenza A virus using neural network with distinguishing of arginine, leucine and serine. Protein Pept Lett 14: 465–470

Wu G, Yan S (2007d) Prediction of mutations initiated by internal power in H3N2 hemagglutinins of influenza A virus from North America. Int J Pept Res Ther (http://dx.doi.org/10.1007/s10989-007-9104-1)

Zambon MC (1999) Epidemiology and pathogenesis of influenza. J Antimicrob Chemother 44 (Suppl B): 3–9

Authors' address: Guang Wu, Computational Mutation Project, DreamSciTech Consulting 301, Building 12, Nanyou A-zone, Jiannan Road, Shenzhen, Guangdong Province CN-518054, China,
Fax: +86 755 2664 8177, E-mail: hongguanglishibahao@yahoo.com