Amino Acids (2008) 35: 365–373
DOI 10.1007/s00726-007-0602-4
Printed in The Netherlands

Amino Acids

Prediction of mutations engineered by randomness in H5N1 hemagglutinins of influenza A virus

G. Wu and **S. Yan**

Computational Mutation Project, DreamSciTech Consulting, Shenzhen, Guangdong Province, China

Received August 18, 2007
Accepted September 30, 2007
Published online November 2, 2007; © Springer-Verlag 2007

Summary. This is the continuation of our studies on the prediction of mutation engineered by randomness in proteins from influenza A virus. In our previous studies, we have demonstrated that randomness plays a role in engineering mutations because the measures of randomness in protein are different before and after mutations. Thus we built a cause-mutation relationship to count the mutation engineered by randomness, and conducted several concept-initiated studies to predict the mutations in proteins from influenza A virus, which demonstrated the possibility of prediction of mutations along this line of thought. On the other hand, these concept-initiated studies indicate the directions forwards the enhancement of predictability, of which we need to use the neural network instead of logistic regression that was used in those concept-initiated studies to enhance the predictability. In this proof-of-concept study, we attempt to apply the neural network to modeling the cause-mutation relationship to predict the possible mutation positions, and then we use the amino acid mutating probability to predict the would-be-mutated amino acids at predicted positions. The results confirm the possibility of use of internal cause-mutation relationship with neural network model to predict the mutation positions and use of amino acid mutating probability to predict the would-be-mutated amino acids.

Keywords: Amino acid – Hemagglutinin – Influenza – Mutation – Neural network

Introduction

The unpredictable mutations in the proteins from influenza A virus not only threaten the world with possible pandemics/epidemics, but also raise the issue of how to accurately, precisely and reliably predict the mutations.

Generally speaking, the simplest and best way for prediction of mutations is to find the cause for the mutation. Then, we could build either a qualitative or a quantitative cause-mutation relationship, by which we could predict the mutations. Nevertheless, the current research on prediction of mutations is going along this line of thought.

However, many causes that historically led mutations might never leave any clue due to the great changes in environments. Therefore we would probably have a detailed record of mutations, but a poor record of mutation causes. Moreover, the current version of proteins from influenza A virus might no longer be subject to the causes, which led the mutations in the past, because of the evolution of influenza A virus. The third difficulty is that we cannot determine the historical micro-environment, under which the mutations occurred.

However, randomness should play a role in engineering mutations not only because 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 more importantly because our previous studies show that the randomness is different before and after mutation (Wu and Yan, 2001b, d, e, 2002a, b, 2003a–h, 2004a–c, f, 2005a, c, e) when using our methods to quantify the randomness within a protein. Actually, randomness simply means that an amino acid with a bigger mutation probability would more easily mutate than an amino acid with a smaller mutation probability.

Hence, we can establish a cause-mutation relationship because we have quantified randomness for a partial cause and we have the occurrence or non-occurrence of mutations by comparing parent and daughter proteins along a branch of evolution tree determined by phylogenetics. In addition, we can classify the occurrence or non-occurrence of mutations as unity and zero. This is very suggestive because such a cause-mutation relationship can be switched to the problem of classification, which can be solved using either the logistic regression in statistics

(Draper and Smith, 1981; Hosmer and Lemeshow, 2000) or the neural network model (Demuth and Beale, 2001).

Still, we need to solve the problem of prediction of would-be-mutated amino acids, say, which type of amino acid will an amino acid mutate to? This is because our cause-mutation relationship at this moment deals with binary events, that is, the occurrence or non-occurrence of mutations. Here we face a more complicated problem, because there are at least 20 types of amino acids needed to take into account, which would be too difficult to use the classification method and other methods.

All these imply that we need at least two steps for accurate, precise and reliable prediction of mutations, (i) the prediction of mutation positions and (ii) the prediction of would-be-mutated amino acids at predicted positions.

Along this two-step frame, we have recently applied the logistic regression to predicting the mutation positions (Wu and Yan, 2006e, f, 2007a–c) and then applied the amino acid mutating probability (Wu and Yan, 2005g, 2006a, 2007a–d) to predicting the would-be-mutated amino acids at predicted positions in proteins from influenza A virus.

The results show our logic very convincing. This leads us to consider using a more powerful classification method, neural network, to enhance the predictability regarding the prediction of mutation positions to further confirm our logic on the cause-mutation relationship before large-scale and full detailed studies.

In this proof-of-concept study, we attempt to use the neural network to predict the mutation positions and then apply the amino acid mutating probability to predict the would-be-mutated amino acids at predicted positions in 5HN1 hemagglutinin from influenza A virus, because the hemagglutinin is the major surface antigen of influenza virus, against which neutralizing antibodies are elicited during virus infection and vaccination (Wiley and Skehel, 1987). The hemagglutinins include many subtypes, of which the H5N1 hemagglutinin is the one currently threatening humans.

Materials and methods

The amino acid sequences and corresponding RNA sequences of 339 H5N1 hemagglutinins from influenza A virus from 1996 to 2005 are obtained from the influenza virus resources (Influenza virus resources, 2006). As our approach is not familiar with most researchers yet, we will describe the methods in more detail.

Prediction model

As the cause-mutation relationship couples three types of quantified randomness developed by us with the occurrence and non-occurrence of mutation, we would expect the model to have three inputs and one output. After elaborations, we finally use the feedforward backpropagation neural network as prediction model (MathWorks Inc., 2001), whose network structure is 3-6-1 (Fig. 1), i.e. the first layer contains three neurons corresponding to three inputs (or three elements of input in neural network terminology), the second layer contains six neurons, and the last layer contains one neuron corresponding to the target (output). The transfer functions for three layers are tan-sigmoid, tan-sigmoid and log-sigmoid, respectively. The training algorithm is the resilient backpropagation, which is the fastest algorithm on pattern recognition (Demuth and Beale, 2001).

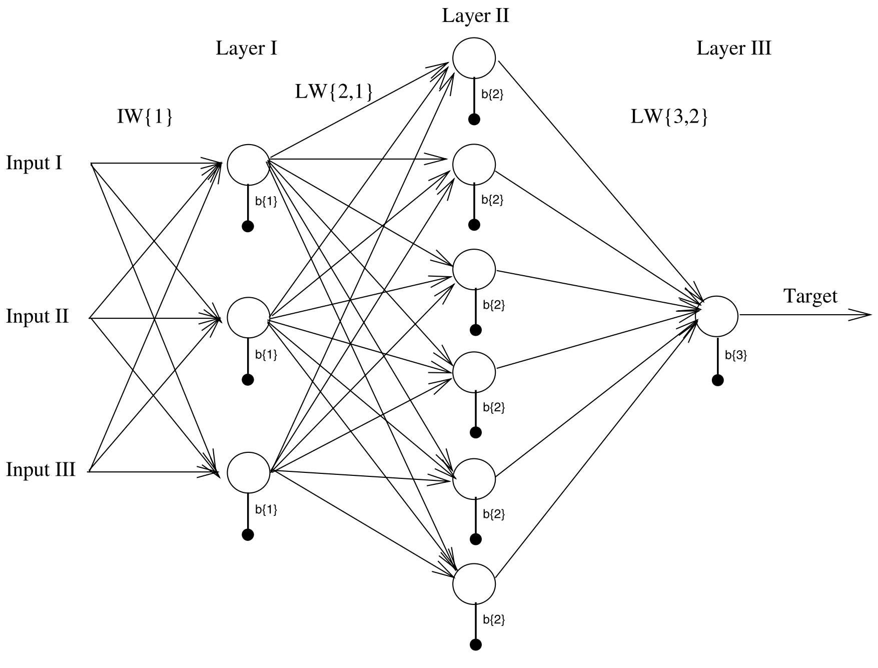

Fig. 1. The 3-6-1 feedforward backpropagation neural network. Each circle presents a neuron. IW{1} is the input weights, LW{2, 1} is the layer weights to the second layer from the first layer, and LW{3, 2} is the layer weights to the third layer from the second layer. b{1}, b{2} and b{3} are the biases related to each neuron at the first, second, and third layers

Input 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). In general, 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. Besides, the reason for using amino acid pair is 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 (Prosite, 2002).

The simplest calculations are as follows. According to the permutation, for example, there are 45 serines (S) and 48 leucines (L) in the 2004 chicken H5N1 hemagglutinin (accession number AY653200), the frequency of amino acid pair "SL" is 4 ($45/568 \times 48/567 \times 567 = 3.803$), that is, the "SL" would appear four times in this hemagglutinin, which is also the reference for comparison. Actually we do find 4 "SL", so the amino acid pair "SL" is predictable and the difference between its actual and predicted frequencies is 0. Again, there are 30 alanines (A) and 39 isoleucines (I) in AY653200 hemagglutinin, and the frequency of random presence of "AI" is 2 ($30/568 \times 39/567 \times 567 = 2.060$), i.e. there would be two "AI" in the hemagglutinin. But the "AI" appears seven times in reality, so the difference between its actual and predicted frequencies is 5. After such calculations, each amino acid pair has its difference between actual and predicted frequencies. As a point mutation is relevant to a single amino acid, which connects with two neighboring amino acids except for the terminal one and constructs two amino acid pairs, we use the sum of difference between actual and predicted frequencies in two neighboring amino acid pairs to each amino acid.

Input II – amino acid distribution probability

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

The quantification is developed along such line of thought, for example, there are two methionines (M) among 142 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 Ms in human hemoglobin α-chain, i.e. (i) both Ms are in the first half, (ii) one M is in each half and (iii) both Ms are in the second half. If we do not distinguish either first half or second half but are simply interested both Ms 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, 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 total three types of distributions, i.e. (i) three amino acids are in each part, (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 these three types of distributions, 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.

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 for (i), 0.5625 for (ii), 0.1406 for (iii), 0.1875 for (iv) and 0.0156 for (v). 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 $r!/(q_0! \times q_1! \times \ldots \times q_n!) \times r!/(r_1! \times r_2! \times \ldots \times r_n!) \times n^{-r}$ (Feller, 1968), where r is the number of amino acids, n is the number of parts and is equal to r in our case, r_n is the number of amino acids in the n-th part, qn is the number of parts with the same number of amino acids, and ! is the factorial function. 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 distinguishing 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.

With respect to hemagglutinins in this study, for instance, there are 20 glutamines (Q) in AY653200 hemagglutinin. Their predicted and actual distribution probabilities are 0.0965 and 0.0128, so the ratio of predicted versus actual distribution probabilities is 7.539, whose natural logarithm is 2.0201, which can be assigned to each Q in the sequence.

Input III – 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, 2007e), and we have used it to study various proteins (Wu and Yan, 2005g, 2006a, e, f, 2007a–e). In general, this quantification is mainly subject to the future mutation trend, and answers what probability an amino acid mutates to another type of amino acid.

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, that 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 lead threonine to mutate to isoleucine, asparagine, and serine, the mutation at the third position of ACU can lead threonine to mutate to 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 types 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, we know that there are 30 "A", 24 "R", 47 "N", 26 "D", 15 "C", 40 "E", and so on in AY653200 hemagglutinin. (iii) So we can estimate how many "A" can be mutated using $30 \times 12/36 + 24 \times 0 + 47 \times 0 + 26 \times 2/18 + 15 \times 0 + 40 \times 2/18 +$, and so on. In total, this is the future composition of amino acid "A". (iv) After calculating all 20 kinds of amino acids, "A" contributes 5.9077% of future composition in the hemagglutinin. (v) On the other hand, "A" contributes 5.2817%

Table 1. Amino-acid mutating probability

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	18/54R + 2/54C + 2/54Q + 6/54G + 2/54H + 1/54I + 4/54L + 2/54K + 1/54M + 4/54P + 6/54S + 2/54T + 2/54W + 2/54STOP
N	2/18N + 2/18D + 2/18H + 2/18I + 4/18K + 2/18S + 2/18T + 2/18Y
D	2/18A + 2/18N + 2/18D + 4/18E + 2/18G + 2/18H + 2/18Y + 2/18V
C	2/18R + 2/18C + 2/18G + 2/18F + 4/18S + 2/18W + 2/18Y + 2/18STOP
E	2/18A + 4/18D + 2/18E + 2/18Q + 2/18G + 2/18K + 2/18V + 2/18STOP
Q	2/18R + 2/18E + 2/18Q + 4/18H + 2/18L + 2/18K + 2/18P + 2/18STOP
G	4/36A + 6/36R + 2/36D + 2/36C + 2/36E + 12/36G + 2/36S + 1/36W + 4/36V + 1/36STOP
H	2/18R + 2/18N + 2/18D + 4/18Q + 2/18H + 2/18L + 2/18P + 2/18Y
I	1/27R + 2/27N + 6/27I + 4/27L + 1/27K + 3/27M + 2/27F + 2/27S + 3/27T + 3/27V
L	4/54R + 2/54Q + 2/54H + 4/54I + 18/54L + 2/54M + 6/54F + 4/54P + 2/54S + 1/54W + 6/54V + 3/54STOP
K	2/18R + 4/18N + 2/18E + 2/18Q + 1/18I + 2/18K + 1/18M + 2/18T + 2/18STOP
M	1/9R + 3/9I + 2/9L + 1/9K + 1/9T + 1/9V
F	2/18C + 2/18I + 6/18L + 2/18F + 2/18S + 2/18Y + 2/18V
P	4/36A + 4/36R + 2/36Q + 2/36H + 4/36L + 12/36P + 4/36S + 4/36T
S	4/54A + 6/54R + 2/54N + 4/54C + 2/54G + 2/54I + 2/54L + 2/54F + 4/54P + 14/54S + 6/54T + 1/54W + 2/54Y + 3/54STOP
T	4/36A + 2/36R + 2/36N + 3/36I + 2/36K + 1/36M + 4/36P + 6/36S + 12/36T
W	2/9R + 2/9C + 1/9G + 1/9L + 1/9S + 2/9STOP
Y	2/18N + 2/18D + 2/18C + 2/18H + 2/18F + 2/18S + 2/18Y + 4/18STOP
V	4/36A + 2/36D + 2/36E + 4/36G + 3/36I + 6/36L + 1/36M + 2/36F + 12/36V
STOP	2/27R + 1/27C + 2/27E + 2/27Q + 1/27G + 3/27L + 2/27K + 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

(30/568) of current composition in AY653200 hemagglutinin. (vi) Thus, we have the ratio of future versus current compositions, for example, the ratio of "A" is 1.1185 (5.9077%/5.2817%), which can be assigned to each "A" in AY653200 hemagglutinin.

Target – occurrence or non-occurrence of mutation

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

Method for prediction of would-be-mutated amino acids at predicted positions

Currently, we have no explicit idea to build a cause-mutation relationship between an original amino acid and its mutated amino acids. However, we can make the estimation according to the amino acid mutating probability based on the translation probability between RNA codons and translated amino acids (Wu and Yan, 2005g, 2006a, 2007e) in Table 1. For instance, if we predict that the possible mutation position is 196, which houses amino acid "H", from Table 1 we know that "H" has the largest chance of mutating to "Q", and the equal chance of mutating to other seven amino acids. In this manner, we make the prediction.

Statistics

The MatLab software (MathWorks Inc., 2001) is used for the model development and prediction. The outlier (3SD) is detected according to Healy (1979). The calculations of prediction sensitivity, specificity and total correct rate are according to the published method (Systat Software Inc., 2004).

Results

Perhaps, we could stratify the model development into following steps, establishing the model, finding model parameters and determining if the model can explain or capture the data.

With respect to neural network, the model parameters are the weights and biases, which need to be determined

Table 2. Inputs and target of AY653200 hemagglutinin sequence

Position	Amino acid	Input			Target
		I	II	III	
1	M	6	0.4700	0.7361	0
...	...	...	...	...	...
153	Y	2	0.7622	0.7454	0
154	Q	4	2.0206	0.8917	1
155	G	1	5.1098	0.8947	0
156	K	1	1.4116	0.7061	1
157	S	5	1.5562	1.0111	0
...	...	...	...	...	...
568	I	2	0.6821	0.8526	0

using historical data. This process is similar to use a pharmacokinetic model to fit the drug concentration-time curve.

After a huge amount of calculations, we have three inputs and one target in each amino acid for all parent hemagglutinins. Table 2 shows such a fraction of a hemagglutinin, where each amino acid is associated with three inputs and one target determined by comparing two 2004 chicken hemagglutinins (AY653200 and DQ080022). This format is used for input into computer for training the neural network.

After trying different neural network models with different numbers of layers, neurons, transfer functions, training algorithms, the 3-6-1 feedforward backpropagation neural network appears to be a suitable model without compromising predictability (Fig. 1), the tan-sigmoid, tan-sigmoid and log-sigmoid as suitable transfer functions and the resilient backpropagation as suitable training algorithm. In principle, the cause-mutation relationship exists between three inputs and target, and we hope the neural network can model this implicit relationship.

When using a pharmacokinetic model to fit the drug concentration-time curve, the initial model parameters can be determined through various methods. However, we have to use the random initialization function to initiate the neural network weights and biases because no historical data on the initial weights and biases are available for our neural network. The question raised here is whether the neural network can converge during its training with a limited number of epochs. Figure 2 shows the convergence of mean squared error performance function with 100 different initial weights and biases generated by random initialization function in using DQ334760 hemagglutinin. As seen, the neural network converges during its training within 250 epochs although the initial weights and biases were randomly generated by the initialization function. Hence, we can use the random initialization function to train the neural network to find the suitable weights and biases.

In order to determine whether the neural network model can capture the cause-mutation relationship, we compare the predicted with the actual mutation positions by classifying the predicted mutation positions as the positives, false positives, negatives and false negatives. Then we calculate the prediction sensitivity, specificity and total correct rate (Fig. 3). As seen, the prediction specificity and total correct rate are quite high while the prediction sensitivity is low.

Until this point, we are step by step approaching to the possibility of using neural network to predict the mutation positions. With this possibility in mind, we used the trained weights and biases to predict the mutation positions, and then predict the would-be-mutated amino acids. Figure 4 shows this two-step prediction in DQ497705 hemagglutinin, A/duck/Vietnam/283/2005 (H5N1). The solid line in the lower panel is the predicted mutation probability by the neural network, 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

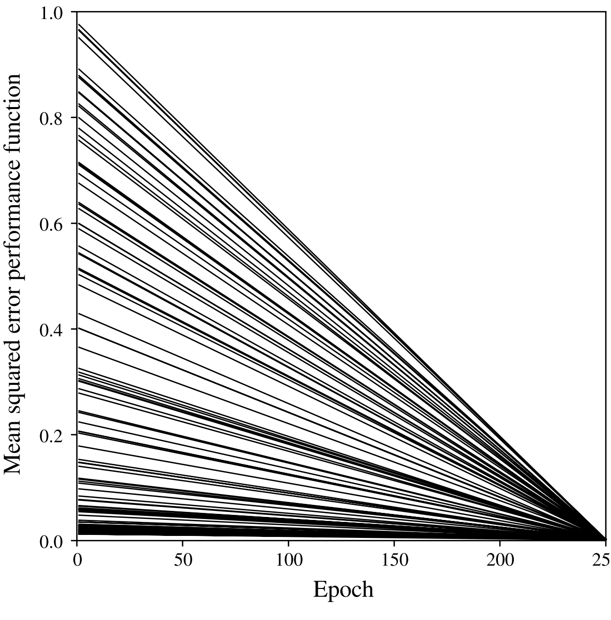

Fig. 2. Convergence of mean squared error performance function with 100 different initial weights and biases generated by random initialization function in using DQ334760 hemagglutinin

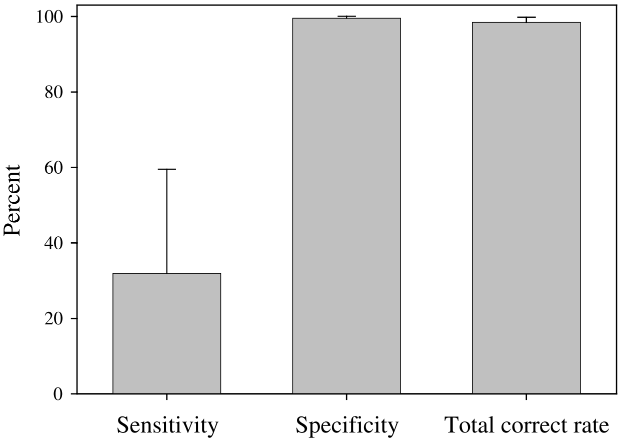

Fig. 3. Prediction sensitivity, specificity and total correct rate for the self-validation. The data are presented as mean ± SD ($n = 110$). The sensitivity is equal to the predicted positives/the actual mutations (%), the specificity is equal to the predicted negatives/the actual non-mutations (%), and the total correct rate is equal to (predicted positives + predicted negatives)/length of hemagglutinin (%)

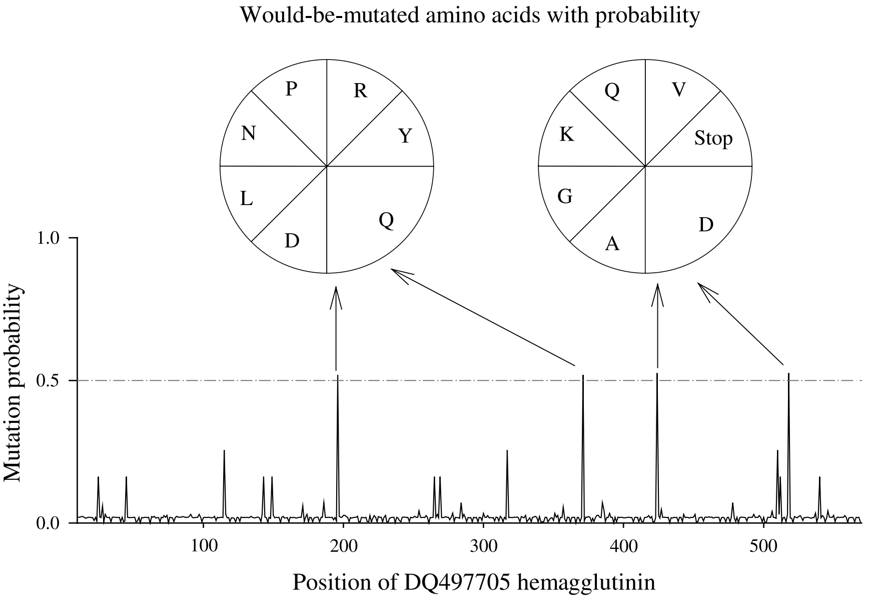

Fig. 4. Prediction of mutations in DQ497705 hemagglutinin. For neural network, IW{1, 1} = [0.2155 1.0388 2.7988; −0.0719 0.6052 0.9833; −0.1510 0.6066 −2.4175], LW{2, 1} = [−1.1000 1.6000 1.2000; 3.2000 1058.0000 1.3000; −0.9000 1.0000 0.3000; −1.0000 −1.7000 5.1000; −1.8000 −0.5000 948.0000; 3.8000 31.3000 1.4000], LW{3, 2}′ = [−1.2252 −2.4157 1.8549 2.3875 −1.6867 1.7738], b{1} = [−5.9882 −2.8807 0.1485], b{2} = [2.8083 −1.4939 −22.3105 −0.8560 −1.3713 1.8262], b{3} = [−0.9614]

0.5 risks mutation. For this hemagglutinin, there are four positions whose mutation probability is larger than 0.5. At these four positions, the would-be-mutated amino acids are predicted using the amino acid mutating probability in Table 1, which is the upper panel of Fig. 4.

Discussion

The preparedness for possible pandemics of influenza is currently conducted 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 et al., 1997, 1999, 2000, 2003, 2006; Chou, 2004a–e, 2005a, b, 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; Chou, 1993a–c, 1996; Elhammer et al., 1993; 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., 2005a, b, 2007a; 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., 2006a, b; Wang et al., 2007b). In this approach, an important concept is the ‘‘binding pocket’’, which is the cornerstone for modeling. According to Chou et al. (1999), the binding pocket was defined by those residues that have at least one heavy atom (i.e. an atom other than hydrogen) with a distance ≤5 Å from a heavy atom of the ligand. Such a definition has been widely and successfully used for investigating various protein-ligand interactions (see, e.g. Chou et al., 2000; Chou, 2004a–d, 2005a, b; Sirois et al., 2004; Du et al., 2005a, b, 2007b; Wei et al., 2005, 2006a, b, 2007; Zhang et al., 2006; Gao et al., 2007; Li et al., 2007; Wang et al., 2007a, c).

However, it is highly likely that the random power plays a continuous role because randomness suggests the maximal probability of occurrence, by which a protein would be constructed with the least time- and energy-consuming, which could meet the speed of rapidly changing environments, although nature can deliberately spend more time and energy to construct an absolutely necessary structure. Hence, our quantifications at least describe the random power engineering mutations.

With respect to the prediction of mutation positions, we have the following issues that need to be addressed in future.

(1) How can we measure whether the model captures a cause-mutation relationship? Generally we use the correlation coefficient in linear regression between measured and predicted data to make the judgment, which is suited when measured and predicted data are paired. However, this is not the case for actual and predicted mutation positions, because, for example, the actual mutation position is 499 in AF102674 hemagglutinin, while the predicted mutation position is

500. To the best of our knowledge, we cannot pair them, which lead to the asymmetry between actual and predicted positions and the difficulty to use the correlation coefficient of linear regression to evaluate the prediction performance.

(2) Although the low sensitivity in Fig. 3 suggests several possibilities such as the mutations related to few mutations, external causes, sampling strategy, etc., the essential problem is that we have no method to measure the performance that the actual position is 499 whereas the predicted position is 500. This distance might be tolerable for proteins as long as hemagglutinin, but might not be so for proteins as short as human hemoglobin α-chain.

(3) Another issue related to the measurement of performance is that the number of predicted is not equal to the number of actual mutation positions. For example, there are three mutations in AF102674 hemagglutinin, while the model captures two mutation positions. At this stage, we have yet to develop the method to measure them.

Traditionally, we divide the dataset into training, test and validation in neural network modeling, however we consider such division too early because we have yet to have the method to measure the performance regarding actual and predicted positions.

However, our approach is promising because it is based on the kinetics, which drives mutations, while the current methods, which search the similar patterns, sequences, signature, etc., in various databases, are more or less based on phenomenon law. The phenomenon observation is very important, by which we can build a dynamic model as the Kelper's laws describe the dynamics of planetary motion. On the other hand, the kinetic deduction is also very important, by which we can build a kinetic model as the Newton's laws describe the kinetics of planetary motion. Moreover, the dynamic model based on phenomenon observation is more suitable to deal with the repeated events, but is less powerful when dealing with the evolutionary process, which in general cannot be reversed. By contrast, the kinetic model can deal with both repeated events and evolutionary process if we can properly define the driving force behind them. Hence, our approach not only has the advantage of quantifying proteins but also has the advantage of kinetic modeling.

Also the predicted number of mutation positions is reasonable in Fig. 4, because four mutations are similar to the prediction we made using the fast Fourier transform to timing the mutation (Wu and Yan, 2005f).

References

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

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

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 (2005a) Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. J Proteome Res 4: 1681–1686

Chou KC (2005b) 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)

Demuth H, Beale M (2001) Neural Network Toolbox for Use with MatLab. Version 4

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

Du QS, Wang S, Wei DQ, Sirois S, Chou KC (2005a) Molecular modelling and chemical modification for finding peptide inhibitor against SARS CoV Mpro. Anal Biochem 337: 262–270

Du QS, Wang SQ, Jiang ZQ, Gao WN, Li YD, Wei DQ, Chou KC (2005b) Application of bioinformatics in search for cleavable peptides of SARS-CoV Mpro and chemical modification of octapeptides. Med Chem 1: 209–213

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

Du QS, Wang SQ, Chou KC (2007b) Analogue inhibitors by modifying oseltamivir based on the crystal neuraminidase structure for treating drug-resistant H5N1 virus. Biochem Biophys Res Comm; doi: 10.1016/j.bbrc.2007.08.025

Elhammer AP, Poorman RA, Brown E, Maggiora LL, Hoogerheide JG, Kezdy FJ (1993) The specificity of UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase as inferred from a database of in vivo substrates and from the in vitro glycosylation of proteins and peptides. J Biol Chem 268: 10029–10038

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 edn, Vol. I. Wiley, New York, pp 34–40

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

Gan YR, Huang H, Huang YD, Rao CM, Zhao Y, Liu JS, Wu L, Wei DQ (2006) Synthesis and activity assess of an octapeptide inhibitor designed for SARS coronavirus main proteinase. Peptides 27: 622–625

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

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

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

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

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

MathWorks Inc (2001) MatLab – the language of technical computing. version 6.1.0.450, release 12.1

Poorman RA, Tomasselli AG, Heinrikson RL, Kezdy FJ (1991) A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. J Biol Chem 266: 14554–14561

Prosite (2002) A dictionary of protein sites and patterns user manual. http://www.expasy.ch/prosite/

Sirois S, Wei DQ, Du QS, Chou KC (2004) Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. J Chem Inf Comput Sci 44: 1111–1122

Systat Software Inc (2004) Systat for windows, version 11.00.01

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 Comm 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 Comm 354: 634–640

Wei DQ, Sirois S, Du QS, Arias HR, Chou KC (2005) Theoretical studies of Alzheimer's disease drug candidate [(2,4-dimethoxy) benzylidene]-anabaseine dihydrochloride (GTS-21) and its derivatives. Biochem Biophys Res Commun 338: 1059–1064

Wei DQ, Du QS, Sun H, Chou KC (2006a) 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 DQ, Zhang R, Du QS, Gao WN, Li Y, Gao H, Wang SQ, Zhang X, Li AX, Sirois S, Chou KC (2006b) Anti-SARS drug screening by molecular docking. Amino Acids 31: 73–80

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

Wiley DC, Skehel JJ (1987) The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu Rev Biochem 56: 365–394

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. Hum Exp 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, 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) Susceptible amino acid pairs in variants of human collagen α1(III) chain precursor. EXCLI J 3: 20–28

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

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

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) 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 (2005e) 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 (2005f) 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 (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) 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 (2006b) Fate of influenza A virus proteins. Protein Pept Lett 13: 377–384

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) Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint. Acta Pharmacol Sin 27: 513–526

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) 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 (2007c) 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

Wu G, Yan S (2007d) Prediction of mutations engineered by randomness in H5N1 neuraminidases from influenza A virus. Amino Acids, http://dx.doi.org/10.1007/s00726-007-0579-z

Wu G, Yan S (2007e) 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, Chapter 3, pp 47–65

Zhang R, Wei DQ, Du QS, Chou KC (2006) Molecular modeling studies of peptide drug candidates against SARS. Med Chem 2: 309–314

Authors' address: G. 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