Application of neural network to predict mutations in proteins from influenza A viruses – A review of our approaches with implication for predicting mutations in coronaviruses

Shaomin Yan, Guang Wu

National Engineering Research Center for Non-Food Biorefinery, State Key Laboratory of Non-Food Biomass and Enzyme Technology, Guangxi Biomass Engineering Technology Research Center, Guangxi Key Laboratory of Bio-refinery, Guangxi Academy of Sciences, 98 Daling Road, Nanning, 530007, China

hongguanglishibahao@gxas.cn

Abstract. The recent outbreak of COVID-19 pandemic is attributed to cross-species transmission of new coronavirus from bats to humans through unknown intermediate hosts, and the essence of the transmission is closely related to the mutations in coronaviruses. Furthermore, the effort to develop the vaccines against coronaviruses always faces the challenge of unexpected mutations in coronaviruses. In fact, it is very difficult to predict the mutations in any virus and bacterium, although mutations are a process of evolution. Over years, we have been applied the neural network to predict the mutations in proteins from influenza A viruses in comparison with the predictions using logistic regression. Our results are encouraging, but our approaches still need the improvements, for example, to upgrade to using machine learning and artificial intelligence instead of neural network. In this review, we summarize the rationales of neural network modelling, its strength and weakness, with the hope that we can apply the improved methods to predict the mutations in coronaviruses, thus to explore the origin of SARS-CoV-2, to find its intermediate host, and eventually to predict its mutations.

1. Introduction

Although mutations lead to the evolution in each species, some mutations in bacteria and viruses, even humans, can bring harmful effects on humans. The current COVID-19 pandemic is such a case where the mutations in coronavirus, SARS-CoV-2, made cross-species transmission from bats to humans possible. Also, mutations create big challenges in developments of vaccines and drugs, since anti-viral vaccines and drugs may lose their targets due to mutations.

In general, mutations are unpredictable, because a mutation depends upon many known and unknown factors. From the view of mathematical modelling, a quantitative cause-mutation relationship could be such an equation: the occurrence of mutation = f(x₁, x₂, x₃, ...), where x₁ is cause 1, x₂ is cause 2, x₃ is cause 3 and so on. In such a way, we can simply assign zero to a cause, although we actually do not know how many factors leading to a mutation, and what conditions being necessary for a mutation.

Taking a protein sequence as an example, a mutation means that an amino acid at certain position changes to another one. For the prediction, this simplest case is related to two predictions, i.e. (i) the prediction of mutated position, and (ii) the prediction of mutated amino acid. Furthermore, we can include the prediction of the time of mutation, and the prediction of the effect of a mutation on humans.
Thus, whether or not a mutation can occur at certain position can be formulated as the occurrence of mutation at the position \( f(x_1, x_2, x_3, \ldots) \). At this point, it necessarily quantifies the mutation and its causes for this equation. Indeed, there are some ways to quantify a mutation cause, such as the amount of X-ray leading to mutation, and the amount of a chemical leading to mutation. Needless to say, X-ray and chemicals are the external causes for mutation. In reality, bacteria and viruses can mutate in laboratory without any external causes, therefore there must be certain internal causes leading to mutations.

We have defined three internal causes based on random principle, because the occurrence of mutation is generally considered as a random event [1]. We applied three causes together to predict mutation positions in a protein, and a single cause to predict the mutated amino acid at the mutation position, especially in the proteins from influenza A viruses [2-14]. Our results are encouraging, though not perfect. This is understandable as we defined only three causes for mutation.

The current COVID-19 pandemic requires us to use all available methods and to develop new methods to against the disease. Here, we briefly review our approaches and explore the possibility to improve them with the aim to use them to combat the SARS-CoV-2.

2. Prediction Methods

2.1. Prediction model

We can use a neural network to build a cause-mutation relationship, where the inputs and output should be the causes and the mutation. Here, the output can be either the mutation position or the mutated amino acid or both, and in our case the output is the mutation position in a protein. We used the feedforward backpropagation neural network as prediction model [15], whose network structure is 3-6-1 (Fig. 1), i.e. the first layer contains three neurons corresponding to three inputs, the second layer contains six neurons, and the last layer contains one neuron corresponding to the output (target). The transfer functions for three layers are tan-sigmoid, tan-sigmoid and log-sigmoid, respectively, which result in the output between zero and unity, i.e. the probability of mutation. The training algorithm is the resilient backpropagation, which is the fastest algorithm on pattern recognition [16].

![Fig.1 The 3-6-1 feedforward backpropagation neural network. Each cycle 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, respectively.](image-url)
2.2. Input I - Amino-acid pair predictability
This input is calculated according to permutation, i.e. how can an amino-acid pair be composed according to amino acid composition in a protein. For example, the neuraminidase from strain A/swine/Spain/51915/2003(H1N1) with accession number CY010574 has 469 amino acids, among which there are 57 serines (S) and 40 asparagines (N). The frequency of amino-acid pair “SN” is 5 (57/469×40/468×468=4.861), that is, the “SN” would appear five times, which is the predicted frequency. Actually we find 5 “SN” in this neuraminidase, so the amino-acid pair “SN” is predictable, and the difference between its actual and predicted frequencies is 0. Again, there are 19 cysteines (C) in this neuraminidase, and the frequency of random presence of amino-acid pair “SC” is 2 (57/469×19/468×468=2.309), i.e. there would be two “SC” in the neuraminidase. But the “SC” appears six times in reality, so the difference between its actual and predicted frequencies is 4. Consequently, each amino-acid pair can be compared between its actual and predicted frequencies. The online calculation can be available at http://www.gxas.cn/calculation/pp.htm. The biological 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 [17].

2.3. Input II - Amino-acid distribution probability
This parameter is calculated according to the occupancy of subpopulations and partitions [18], and 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 [18]. The distribution probability is r!/(q0!×q1!×...×qn!)×r!/(r1!×r2!×...×rn!)×nr, where ! is the factorial function. For our purpose, we define r is the number of a type of amino acid, q is the number of partitions with the same number of amino acids, and n is the number of grouped partitions in the protein for a type of amino acid. For instance, there are 17 phenylalanines (F) in CY010574 neuraminidase, so we can divide this 469 amino-acid neuraminidase into 17 partitions, give the number of phenylalanines in each partition into the equation, and to calculate the actual distribution probability. The online calculation can be available at http://www.gxas.cn/calculation/dp.htm. Meanwhile, we can calculate the predicted distribution probability for a type of amino acid, which is the biggest one in a particular partitioning of a protein. For this case, the predicted and actual distribution probabilities are 0.1280 and 0.0366.

2.4. Input III - Ratio of future versus current amino acid compositions
This input comes out from the observation that 64 RNA codons are not proportional to 20 types of amino acids. For example, only one type of RNA codon AUG can translate to methionine, but different types of RNA codons can translate any of other types of amino acids. For example, threonine is related to 4 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, respectively, at amino acid level. Similarly, the mutation at second position of ACU can lead threonine to mutate to isoleucine, asparagine and serine, respectively. The mutation at the third position of ACU can lead threonine to mutate to threonine, threonine and threonine, respectively. Taking 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 compiled a table of the amino acid mutating probability [2-9, 12-14].

Again taking CY010574 neuraminidase as an example, according to the table of amino acid mutating probability: 1) alanine has the 12/36 chance of mutating to alanine; 2) cysteine, arginine and asparagine cannot mutate to alanine; 3) aspartic acid and glutamic acid has 2/18 chance of mutating to alanine; 4) there are 16 alanines, 18 arginines, 40 asparagines, 20 aspartic acids, 19 cysteines and 20 glutamic acids in CY010574 neuraminidase. So we can estimate how many alanines can be mutated,
using \(16 \times 12/36 + 18 \times 0 + 40 \times 0 + 20 \times 2/18 + 19 \times 0 + 20 \times 2/18\). In total, this is the future composition of alanines. After calculated all 20 types of amino acids, alanines contribute 5.9228\% of future composition in this neuraminidase. On the other hand, alanines contribute 3.4115\% (16/469) of current composition. The online calculation can be available at http://www.gxas.cn/calculation/fc.htm. Thus, the ratio of future versus current compositions of alanines is 1.7361 (5.9228\% / 3.4115\%) in CY010574 neuraminidase.

2.5. Training and prediction
In order to use neural network for prediction, it needs to train the neural network. We follow the phylogenetic tree to compare sequential pair of proteins to find the mutations, then these pairs are transferred as numeric sequence using the methods we defined above, i.e. each position in protein is represented using three numbers, and occurrence and non-occurrence of mutation are presented as unity and zero. After training, a protein can be predicted to see at which position a mutation will occur.

2.6. Prediction of would-be-mutated amino acids at predicted positions
To predict the mutated amino acids at predicted positions, we can also use the table of amino acid mutating probability as mentioned above [2-9, 12-14].

3. Worked Example

3.1. Data
258 H1 neuraminidases from influenza A viruses were obtained from the influenza virus resources [19].

3.2. Phylogenetics
Phylogenetic analysis can stratify the data into phylogenetic relationships, which deemed to have a direct mutation relationship, i.e. father and daughter relationship.

3.3. Computation of inputs and output
All father H1 neuraminidases were computed for the three inputs of prediction model, i.e. amino-acid pair predictability, amino-acid distribution probability and the ratio of future versus current amino acid compositions, and then three numeric numbers were assigned to each position of each H1 neuraminidase. The output is determined by comparing father H1 neuraminidase with its daughter H1 neuraminidase in order to find the mutated position, which was marked as unity, whereas unmaturated positions were marked as zero.

3.4. Training of neural network with inputs and outputs
The 3-6-1 feedforward backpropagation neural network (Fig. 1) was trained with these inputs and outputs from each father H1 neuraminidase in order to get the parameters of neural network model, i.e. weights and biases. Generally, neural network can converge during its training within 250 epochs although the initial weights and biases were randomly given by the initialization function. Hence, we can use the random initialization function to train the neural network to find the suitable weights and biases. The MatLab software [15] was used.

3.5. Prediction of mutation position and mutated amino acid
After getting mean±SD of all weights and biases of neural network model from training, we used mean of weights and biases to predict mutation in H1 neuraminidases, whose mutations have yet to occur. Fig. 2 shows an example of prediction, where x-axis is numbered as amino acid position in CY014009 H1 neuraminidase, y-axis is the mutation probability, and the pie in the top is the probability of mutated amino acids. The prediction was based on 3-6-1 neural network with IW\{1,1\}=[-4.6696 -2.0449 0.1747; 3.5841 1.0104 -0.8998; 1.1665 0.2746 -0.9222],
LW{2,1}=[0.3156 2.6191 -0.5951; 0.1253 1.8284 -0.1963; -3.5293 -2.1733 21.3117; -4.1504 -3.8525 21.9636; -3.5088 1.0315 -1.2338; -2.2361 3.3929 5.0140], LW{3,2}=[21.9953 3.7891 1.0323 1.0824 -0.9664 -1.5816], b{1}=[1.6297 1.5949 0.4574], b{2}=[-2.5399 -1.7397 -14.1173 -13.4918 -0.6663 -1.0234], and b{3}=[-0.6037]. As seen in Fig. 2, the mutation probability in three positions, 82, 146 and 285, is larger than 0.5, so the amino acid, threonine (T), at these three positions are likely to mutate. The size of each piece of pie on the top indicates the probability that the threonine is likely to mutate to other amino acid. For example, the threonine has the chance of 1/4 to mutate to serine (S).

3.6. Summary of our previous predictions
In the past, we conducted studies along this research line [2-14], and used sensitivity (predicted positives/actual mutations (%)), specificity (predicted negatives/actual non-mutations (%)) and total correct rate ((predicted positives + predicted negatives)/protein length (%)) to evaluate the predictions. As can be seen in Fig. 3, the predictions are very encouraging in terms of specificity and total correct rate, but the sensitivity needs to improve. Anyway, the predictions made by neural network provide higher sensitivity than that made by logistic regression.
4. Prospective
Computational approach was used to predict conserved epitopes in influenza viruses [20], however, few studies are conducted to predict future mutations. The current COVID-19 pandemic provides an opportunity to apply our approaches to predict the mutations in SARS-CoV-2, especially, to reversely track its origin and determine how the occurrence of mutations led to cross-species transmission, i.e. how coronavirus jumped from bats to humans through intermediate hosts.

(1) Anyway such idea still needs more elaborations, because SARS-CoV-2 does not have many historic data, thus it is hard to build a detailed phylogenetic tree to determine father-daughter relationship, i.e. direct inherit relationship.

(2) Our approaches define just three forces that drive amino acids to mutate, but there are surely many other forces that can drive the mutations, which need search or make assumptions.

(3) Neural network comparing with artificial intelligence (AI) appears weaker, so it is absolutely necessary to develop AI methods for the predictions of mutations in various species, for the time being, for coronaviruses.

Acknowledgments
This study was supported by National Natural Science Foundation of China (31560315), and Key Project of Guangxi Scientific Research and Technology Development Plan (AB17190534).

References
[1] Fitch Walter M, Bush Robin M, Bender Catherine A, Cox Nancy J. Long term trends in the evolution of H(3) HA1 human influenza type A[J]. Proceedings of the National Academy of Sciences of the United States of America, 1997,94:7712-7718
[2] Wu Guang, Yan Shaomin. Prediction of possible mutations in H5N1 hemagglutinins of influenza A virus by means of logistic regression[J]. Comprehensive Clinical Pathology, 2006,15:255-261.
[3] Wu Guang, Yan Shaomin. Prediction of mutations in H5N1 hemagglutinins from influenza A
virus[J]. Protein and Peptide Letters, 2006,13:971-976.

[4] Wu Guang, Yan Shaomin. Improvement of model for prediction of hemagglutinin mutations in H5N1 influenza viruses with distinguishing of arginine, leucine and serine[J]. Protein and Peptide Letters, 2007,14:191-196.

[5] Wu Guang, Yan Shaomin. Improvement of prediction of mutation positions in H5N1 hemagglutinins of influenza A virus using neural network with distinguishing of arginine, leucine and serine[J]. Protein and Peptide Letters, 2007,14:465-470.

[6] Wu Guang, Yan Shaomin. Prediction of mutations in H1 neuraminidases from North America influenza A virus engineered by internal randomness. Molecular Diversity, 2007,11,131-140.

[7] Wu Guang, Yan Shaomin. Prediction of mutations engineered by randomness in H5N1 hemagglutinins of influenza A virus[J]. Amino Acids, 2008,35:365-373.

[8] Wu Guang, Yan Shaomin. Prediction of mutations engineered by randomness in H5N1 neuraminidases from influenza A virus[J]. Amino Acids, 2008,34:81-90.

[9] Wu Guang, Yan Shaomin. Prediction of mutations initiated by internal power in H3N2 hemagglutinins of influenza A virus from North America[J]. International Journal of Peptide Research and Therapeutics, 2008,14:41-51.

[10] Wu Guang, Yan Shaomin. Prediction of mutation in H3N2 hemagglutinin of influenza A virus from North America based on different datasets[J]. Protein and Peptide Letters, 2008,15:144-152.

[11] Yan Shaomin, Wu Guang. Prediction of mutation position, mutated amino acid and timing in hemagglutinins from North America H1 influenza A virus[J]. Journal of Biomedical Science and Engineering, 2009,2:117-122.

[12] Yan Shaomin, Wu Guang. Prediction of mutation positions in H5N1 neuraminidases by means of neural network[J]. Annals Of Biomedical Engineering, 2010,38:984-992.

[13] Yan Shaomin, Wu Guang. Prediction of mutations in hemagglutinins of EuroAsia H3N2 influenza viruses using neural network[C]. Proceedings of the 3rd International Conference on Intelligent Computing and Intelligent Systems (ICIS), Guangzhou, China, 2011,3:138-142.

[14] Yan Shaomin, Wu Guang. Prediction of Mutations in H7 Hemagglutinins from Influenza A Virus[J]. Journal of Biomedical Science and Engineering, 2020,13:175-186.

[15] Mathworks Inc. MatLab - The Language of Technical Computing (version 6.1.0.450, release 12.1), 1984-2001[CP]. 2001.

[16] Demuth Howard, Beale Mark. Neural Network Toolbox for Use with MatLab, User’s guide[G]. The MATH WORKS Inc., 2010.

[17] PROSITE. A Dictionary of Protein Sites and Patterns User Manual[G]. 2020, https://prosite.expasy.org.

[18] Feller William. An Introduction to Probability Theory and Its Applications[M]. 3rd ed. Vol, I., Wiley, New York. 1968.

[19] The National Center for Biotechnology Information. Influenza Virus Resources[EB/OL]. 2020, https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database.

[20] Wang Xiangyu, Sun Qi, Ye Zhonghua, Hua Ying, Shao Na, Du Yanli, Zhang Qiwei, Wan Chensong. Computational approach for predicting the conserved B-cell epitopes of hemagglutinin H7 subtype influenza virus[J]. Experimental and Therapeutic Medicine, 2016,12:2439–2446.