

New Mathematical Model of BIS for Influenza a Virus Pattern Prediction

Se-Eun Bae *

*KM Fundamental Research Division, Korea Institute of Oriental Medicine, 1672
Yuseong-daero, Yuseong-gu, Daejeon, Korea
Seeun80@kiom.re.kr*

Abstract

Influenza virus undergoes continuous changes, evolving into different types every year; in fact, there are up to 198 subtypes of the influenza virus. Each subtype infects a different animal species and has different levels of infectivity. The infection pattern and time of outbreak of each subtype can be analyzed to predict its infectivity. Conventional approaches involve the use of simple DNA or protein sequencing methods to evaluate the viral toxicity. However, this approach cannot explain complex biological behavior resulting from various interactions. In this study, a new mathematical model was developed to evaluate the central dogma, and the model was used to predict the pandemic, epidemic, or reassortment nature of the virus.

Keywords: *Entropy, Infectious disease, pandemic, epidemic, reassortment*

1. Introduction

Various types and combinations of the influenza virus emerge worldwide every year. It is often classified as an epidemic rather than a pandemic; however, major changes in the viral protein or a combination of various subtypes by reassortment can lead to strong pandemic-like infectivity [1-2]. Deadly influenza virus outbreaks recorded in history include the 1918 Spanish flu, 1950 flu (higher death toll than the Asian flu and Hong Kong flu [3]), 1957 Asian flu, and the 2009 novel swine-origin influenza A infection. The infectivity and pathogenicity of most of the causative viruses (of these diseases) were higher than those of ordinary strains because of the reassortment of various subtypes [2]. Infectious diseases can re-emerge with a higher infectivity because of the higher population density, which increases the likelihood of causing a pandemic. However, very few studies have attempted to predict pandemics at the early stages, or to evaluate the information gain or loss in the virus genome because of mutations in the virus [4]. The easiest approach to examining such materials (known to date) is the entropy-based approach. Entropy, a concept typically used in physics, can be applied to biology to reveal biological information such as the information on conformational constraints on side chains in protein residues [5], correlation between entropy, structure, and sequence [6], protein-protein interactions (antigen-bound immunoglobulin) [7], sequence conservation measures [8], and genomic sequence analysis [9]. Xiao detected sensitive proteins responsible for mutagenesis by genomic sequence analysis, based on the theory of entropy. In this study, analysis of the nucleic acid sequences of the coding and non-coding regions of the classical swine fever virus (CSFV) genome demonstrated that the first position of each triplet showed maximal entropy, according to the code theory. A majority of the studies conducted so far have attempted to explain the interactions between proteins and the characteristics of gene data based on the sequence information. Entropy was introduced to explain viral infection, and to characterize the important proteins involved in the infection process, or to interpret the prediction information. Resch used entropy to predict the phenotype of human immunodeficiency virus type 1 [10], and Koo measured

the diversity of the West Nile virus (WNV) sequence using an entropy-based analysis [11]. The concept of entropy was also used to analyze the co-evolution of the human influenza A/H3N2 virus, which specifically affects the antigenic variants [12]. Moreover, the entropy of the segment 7 nucleotide sequence of influenza A virus has been analyzed [13]. The aim of this study was to evaluate the complexity of information loss or gain when the influenza virus protein undergoes mutation or reassortment, as entropy values. First, the rate of changes in the amino acid composition was examined to predict the changes in virus patterns [14]. Subsequently, the virus protein and DNA entropy were calculated and compared. Changes in the protein entropy more accurately reflected the changes in viral information than the changes in DNA entropy. However, whereas DNA or protein entropy may indicate the degree of change in each, any information loss or gain resulting from the correlation between the two was not reflected. In other words, as a protein is synthesized from DNA, DNA can induce changes in the proteins by altering the base frequency, but this information is not considered while calculating the entropy. As entropy is calculated using only the DNA or protein sequence, the pandemic pattern caused by mutation and reassortment is not accurately described. Therefore, in this study, a new method was proposed that could account for all information.

2. Method

2.1. New Mathematical Model of Biological Information Stream (BIS) for the Prediction of Pandemics and Epidemics

The simplest form of representing a system is by calculating its entropy, which was the goal of this study. Therefore, the conventional entropy equation was transformed to explain biological phenomena. Biological Information Stream (BIS) is an equation that integrates all information that changes continuously. Unlike the conventional entropy calculation, the DNA and protein information is not separate; therefore, all information is represented. In other words, one equation can provide a range of information, from the DNA to the protein level. The BIS developed in this study is represented by the following equation:

$$BIS = \sum_{i=1}^4 \sum_{j=1}^{20} Pb_i \log P(n_j | Pa_j \times Pb_{ci}) \quad (1)$$

The log value of the original Shannon entropy equation simultaneously describes the state of highest probability and the state of equal temperature. In this study, log values were used to account for the probability and quantity of information flow. Shannon also used the log₂ transformation of all information for both factors. On the other hand, information used in this study involves more than two systems (pandemic, epidemic, endemic, and reassortment); therefore, the log values were used instead of the log₂ values.

The equation subjected to log is used to calculate the conditional probability of change in 'j' amino acid, which combines the probability of finding the amino acid 'j' and the frequency of the base 'i' in the codon. In other words, the probability of change in a specific amino acid is the product of the probability of finding that amino acid in the selected protein multiplied by the probability of the base that form the codon which produces the amino acid. The change in amino acid is affected by the frequency of the specific amino acid and the change in codon composition (A, G, C, or T); therefore, a conditional probability method was employed. The values differ when the changes in amino acid are affected by the base in the codon. Therefore, the numerator in the above equation indicates the change in 'i' caused by changes in the amino acid 'j'. The numerator, which is the common set in both cases, is described in detail, as follows:

$$\log P(\text{changed amino acid number} \times \frac{\text{changed number by specific base in changing amino acid}}{\text{total changed base number when changing amino acid}}) \quad (2)$$

Where $P_{b_{i1}}$ represents the probability of finding the base 'i' in the DNA sequence. n_j denotes the number of changes in the amino acid 'j' (compared to the original sequence), P_{a_j} is the probability of occurrence of 'j', and $P_{b_{ci}}$ is the probability of using the base 'i' in the codon.

2.2. Using Shannon Entropy to Analyze the Properties of DNA and Protein

The primary structure of the protein is determined by the amino acid sequence; as the DNA sequence determines the basic amino acid sequence, its biological value is much higher. Rudolf Julius Emanuel Clausius introduced the concept of entropy in the 1950s to propose the first and second laws of thermodynamics (<http://www-history.mcs.st-and.ac.uk/Biographies/Clausius.html>). Claude E. Shannon explained the theory of entropy by integrating statistical concepts into the Ludwig Eduard Boltzmann theory [15-16]. Information theory states that entropy is a measure of uncertainty among random variables. In other words, entropy represents the degree of disorder as a specific quantity. This quantity is calculated from probability values, with the possible values being calculated in each state. Lower entropy indicates no randomness, which allows for the prediction of information from the standard biological properties of various phenomena. System maintenance requires a large amount of energy, and insufficient energy can lead to disorder. The Shannon entropy is calculated by the following equation:

$$H = -K \sum_{i=1}^k p(i) \log p(i) \quad (3)$$

Where H represents the entropy of information theory and $p(i)$ denotes the probability. Therefore, as shown in equation (4), the probability of a system is represented at each position 'i'.

$$p(i) = \frac{a_i}{n} \quad (4)$$

In conclusion, entropy is a set of the probabilities in each state ($p_1, p_2, p_3, \dots, p_k$). The entropy with the probabilities of "k" events is calculated as follows:

$$H = -(p_1 \log p_1 + p_2 \log p_2 + \dots + p_k \log p_k) \quad (5)$$

The entropy of amino acid at each position was calculated in this study; therefore, the value of k was 20. The calculated Shannon entropy provides information on genes, as well as the resultant proteins. For example, the sequence analysis of amino acid was used to measure the conservation of protein sequence [8]. Information can be also be obtained by aligned sequence [17].

2.3. Data Collection for Protein Analysis (1918 H1N1 Influenza A virus ~ 2009 H1N1 Influenza A virus and H2N2, H3N1, H5N1)

In 2008, reassortment events in the H1N1 influenza A virus gene were phylogenetically analyzed using the evolutionary history of 8 constituent genes (*PB2, PB1, PA, HA, NP, NA, M1, NS*) from 1918 to 2006 [18]. Additionally, this study analyzed data from Nelson and that of the 2009 novel swine-origin influenza A virus epidemic. For the proteins, only the CDS of the complete genome sequence of the H1N1 influenza A

virus from 1918 to 2009 was used, as the protein properties have been previously compared and analyzed. All protein sequences and DNA sequences were downloaded from the Influenza Virus Resource, available on GenBank (<http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html>).

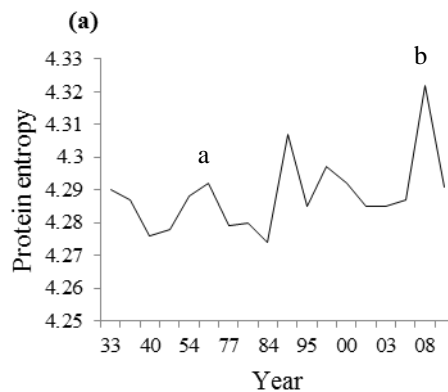
3. Result

3.1. Pandemic and Epidemiology of Influenza Virus

The most prominent characteristic of the influenza virus is its reassortment ability. An examination of the incidence of H1N1 over the years showed certain specific trends: the 1918–1919 Spanish flu outbreak resulted in 50 million deaths worldwide [3], and continuous cases of the flu were reported from 1920 to 1950. Although no flu pandemic occurred during the 1950–1951 period, the death tolls were higher than those recorded for the Asian flu and Hong Kong flu; therefore, this was classified as a flu epidemic. The incidence of H1N1 increased temporally during 1957 with the reassortment of H1N1 to H2N2. This led to the H2N2 Asian flu pandemic, which subsequently co-circulated with the H3N2 type flu. Human infections began in 2007, resulting in a pandemic by 2009. Analysis of the 8 proteins of the 2009 H1N1 strain showed that it contained NA and M protein information of the Eurasian avian-like swine A (H1N1) virus, whereas the other 6 proteins contained reassortment of 3 different subtypes. Reassortment of *PB2* and *PA* from the avian virus, *PB1* from human A (H3N2) virus, and *HA*, *NP*, *NS* from classical swine A (H1N1) virus resulted in a triple-reassortment virus. The antigenic shift that occurred during the Hong Kong flu (1968–1969) led to the evolution of the H3N2 strain. Triple-reassortment of the H3N2 strain with the H1N1 and H1N2 occurred during 1998. The strain that emerged during 1968 evolved from H2N2. Finally, H5N1 infection of humans started in 1997. Probing of the H5N1 sequence identified a reassortment with a portion of the *HA* sequence from H1N1. This virus then re-emerged in Hong Kong during 2003.

3.2. Biological Information Stream (BIS) of H1N1 Hemagglutinin

Influenza can occur annually as a pandemic, epidemic, or endemic via mutations in the causative virus or antigenic shifts. Figure 1(c) depicts the information stream of influenza virus H1N1 hemagglutinin developed using the biological information system (BIS) developed in this study.



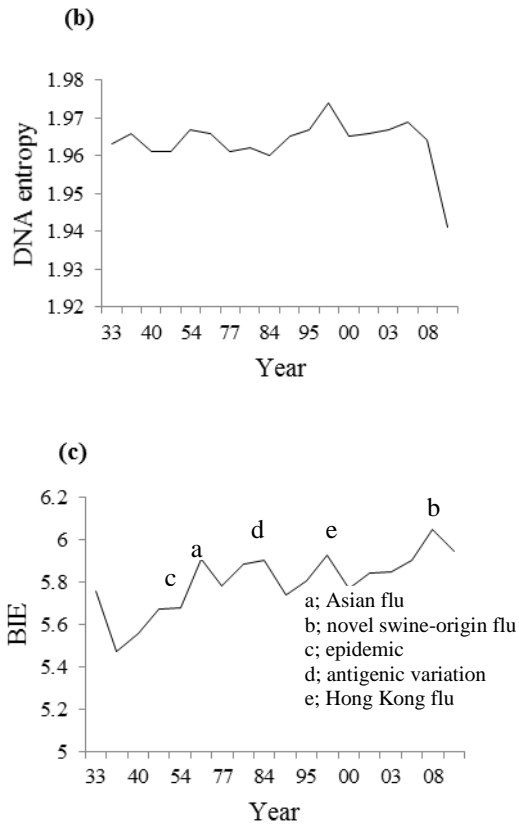


Figure 1. Entropy of the H1N1 Influenza Virus

Information on the changes in amino acids in the H1N1 hemagglutinin protein and their causative bases during DNA-protein synthesis during each year was obtained using the MEGA 4 software. The amino acid undergoing the maximum number of changes each year is presented in Figure 2. These changes can help identify which amino acids had an influence on viral pandemic, epidemic, and endemic outbreaks.

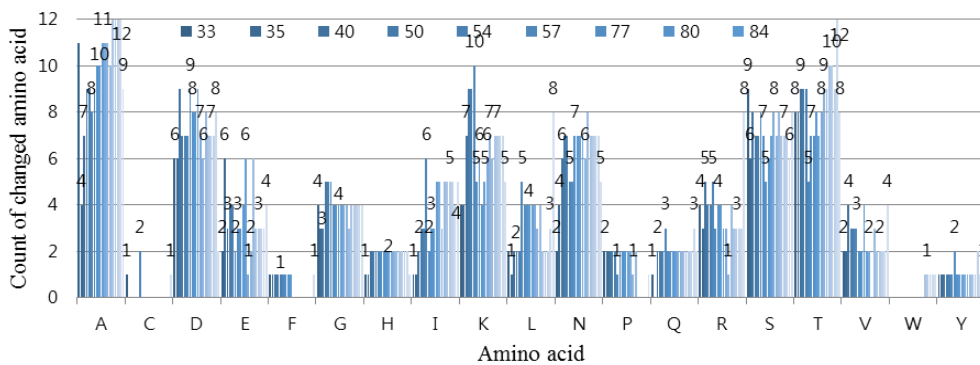


Figure 2. Number of Changed Amino Acids Based on the 1918 H1N1 Virus

Amino acid and base probabilities were calculated by the JAVA program developed to calculate Shannon entropy. BIS was also calculated using this program. The amino acid sequence for each year was compared to the 1918 viral amino acid sequence. Results with higher impact than the Shannon entropy-based protein and DNA (Figures 1a and 1b) were obtained. BIS indicated that the number of deaths caused by H1N1 infection increased in

1950 and temporarily in 1957. Influenza-related information increased steadily with slight changes every year; these were characterized by abrupt changes caused by the increase in death rate, which gradually decreased thereafter. The flu of 1950 resulted in a greater number of deaths than that reported for the Hong Kong flu; therefore, the entropy value increased steadily before the outbreak and peaked at 1950. Another peak was noted in 1957, during the Asian flu H2N2 pandemic resulting from a reassortment with H1N1. Therefore, the change was higher for information related to *HI*. Another increase was observed in 1957 during the outbreak of H5N1 influenza; although this may seem unrelated to the outbreak of H1N1, the development of this virus was attributed to a portion of the *HA* sequence in H1N1 being reassorted into H5N1. In other words, there was a change in *HA* in the H1N1 virus, which was reflected in the graph. Although no significant peak was observed in 1980, another epidemic could be identified during this period. We also observed a stage change in information pertaining to the novel swine-origin influenza A virus, reaching a peak during the 2008 pandemic, indicated that changes in the viral genome were recorded steadily until a pandemic or epidemic event (with a high number of deaths), which occurred when the change was at its peak. Although reassortment with other types does not lead to H1N1 infection, an increase in entropy was also observed during the outbreak of infections caused by a type other than the H1N1 protein. In this case, the number of deaths is not related to H1N1; therefore, the entropy increase is steady (neither an increase nor a decrease was observed after the peak). Small changes in the amino acid can lead to an increase in the total entropy prior to a pandemic; moreover, this leads to an increased chance of human infection during the pandemic. The viral protein is assumed to change to an appropriate form during this time. A widespread pandemic infection is accompanied by various changes in the protein, leading to an increase in BIS. This may contradict the protein or DNA entropy distribution information obtained via conventional Shannon entropy. Table 2 lists the values of biological information entropy over the years. These were not obtained by a conventional method, wherein the DNA entropy is calculated from the DNA or protein sequence alone, without taking into consideration the changes in DNA information or protein entropy. Such a method does not integrate the information regarding changes, and is limited to separating each condition. On the other hand, BIS integrates all information from DNA to protein. In other words, it represents the information entropy that represents biological phenomena. Therefore, this method accurately records the changes in protein information resulting from changes in DNA compared to the Shannon entropy. The information is not simply calculated by amino acid or base probability, but using biological information entropy, which represents all information of the central dogma in biology. In addition, the extent of differences of a newly discovered DNA sequence from the original sequence can be properly interpreted by implying the changes in amino acid and DNA. BIS can predict the extent of change in virus, which in turn can predict the likelihood of a viral pandemic or epidemic. Furthermore, increased values indicate continual mutation or reassortment; therefore, specific amino acids with large changes or specific bases that cause these changes can be identified. Table 1 presents the amino acids that underwent yearly changes, with larger changes indicating the occurrence of a high death toll or pandemic. The bases that resulted in these changes were almost identical. This information can help determine the time point of the viral mutation. In other words, by looking at the amount of change in the information at the point of mutation, the probability of viral mutation or viral reassortment and whether a pandemic would occur or not can be predicted.

Table 1. Biological Information Stream (BIS) of H1N1 Hemagglutinin

	(BIS)1918	Protein	DNA
1933	5.7556	4.29	1.963
1935	5.4733	4.287	1.966
1940	5.5581	4.276	1.961
1950	5.6752	4.278	1.961
1954	5.6760	4.288	1.967
1957	5.9107	4.292	1.966
1977	5.7708	4.279	1.961
1980	5.8825	4.28	1.962
1984	5.9022	4.274	1.96
1991	5.7376	4.307	1.965
1995	5.8092	4.285	1.967
1997	5.9277	4.297	1.974
2000	5.7696	4.292	1.965
2001	5.8438	4.285	1.966
2003	5.8511	4.285	1.967
2006	5.9060	4.287	1.969
2008	6.0522	4.322	1.964
2009	5.9476	4.291	1.941

3.3. Calculation of Shannon Entropy for H1N1 Protein and DNA Sequences

The Shannon entropy plots of H1N1 protein and DNA (Figure 1a and 1b) were very different. An increase in protein entropy corresponded to a decrease in DNA entropy, or vice versa. Alternately, no special pattern was noted in DNA entropy during a pandemic or epidemic; therefore, proteins underwent a greater number of changes than DNA. In fact, we observed multiple peaks in protein entropy values over the years; specifically, abrupt increases were recorded in 1957 (Asian influenza-H2N2 virus), 1991, 1997 (H5N1-Hong Kong), and 2008, corresponding to pandemics or epidemics, except in 1991. The pattern cannot explain the abrupt increase in 1991, similar to the pattern exhibited by H1N1. Moreover, it was difficult to represent the years wherein a large number of deaths occurred as a result of infection with different types of the virus, rather than the occurrence of a pandemic or epidemic due to mutations or reassortment with another subtype. In addition, the year wherein a large number of deaths occurred due to H1N1, but where the spread was not classified as a pandemic, cannot be explained.

3.4. A Comparison of the BIS and Shannon Entropy

Figures 1a and 1c appear to be similar, excluding the DNA-based information, with slight differences. The years 1950–1954, 1977 ~ 1980, 1991, and 2001 specifically indicated an abrupt spread of disease. Whereas the Shannon entropy (Figure 1a) increased during 1950–1954, the BIS curve showed a peak during 1950, followed by a steady-state curve. This can be attributed to an epidemic, and not a pandemic spread of disease. The 1977 ~ 1980 Russian flu epidemic, or the abrupt epidemic that occurred in Russia and China in 1977 after a peak in 1957 (which was followed by a decrease in the number of deaths) was represented well by both the Shannon and NIS curves. No noticeable event was found prior to 1991, occurring as a seasonal flu until 1997 during the outbreak of H5N1. Finally, the steady increase in H1N1 that reached a peak during the pandemic stage was described differently by BIS and Shannon entropy. Moreover, the 1968 Hong Kong influenza pandemic was not represented in the graph. As this was caused by a

reassortment of the H3N2 subtype with the NA of H2N1, it may not have been represented in the HA graph.

4. Discussion

The values obtained by BIS corresponded to the pattern of influenza virus infection and the series of significant events leading to changes in the virus. An examination of the influenza subtype with the highest infectivity in humans during a pandemic, epidemic, or reassortment revealed distinctive patterns in H3N2 during the Hong Kong flu, when the pandemic occurred in humans for the first time. H1N1 re-emerged during 1976–1977 and co-circulated with H3N2 in 1977. Novel swine-origin influenza A underwent reassortment with H3N2 *PBI*, leading to protein changes that caused an increase in entropy during 2009 during the occurrence of novel swine-origin influenza A virus. H5N1, the first influenza virus to infect humans (1997) re-emerged in Hong Kong in 2003, accompanied by a corresponding change in entropy. However, the protein entropy was not consistent with the DNA entropy. The differences between DNA and protein entropy were investigated by examining the base compositions for the DNA and the amino acid compositions for the proteins. The results showed that the A, G, C, and T content became constant with time (despite the small quantity). Therefore, this resulted in a decrease in entropy. However, the changes in A, G, C, and T induce a number of changes in the codon, inducing a greater number of changes in the protein level (as each codon uses each base at a different frequency). Therefore, changes in A, G, C, and T content of each virus were compared to the frequency of each base in each codon. This work demonstrates the redundancy possessed by each codon.

Table 2. DNA Bases Frequencies for Amino Acids

Nucleic acid Amino acid	A	G	C	U	Total
Ala (A)	1	5	5	1	12
Cys (C)	0	2	1	3	6
Asp (D)	2	2	1	1	6
Glu (E)	3	3	0	0	6
Phe (F)	0	0	1	5	6
Gly (G)	1	9	1	1	12
His (H)	2	0	3	1	6
Ile (I)	4	0	1	4	9
Lys (K)	5	1	0	0	6
Leu (L)	2	2	5	9	18
Asn (N)	4	0	1	1	6
Pro (P)	1	1	9	1	12
Gln (Q)	3	1	2	0	6
Arg (R)	4	10	3	1	18
Ser (S)	3	3	6	6	18
Thr (T)	5	1	5	1	12
Val (V)	1	5	1	5	12
Trp (W)	0	2	0	1	3
Tyr (Y)	2	0	1	3	6

As shown in Table 2, P can comprise of a total of 4 codons. Moreover, C was the most frequently used (9/12) base; therefore, a decrease in the ratio of C compared to other bases can result in the conversion of P to another amino acid, in turn reducing the P value. In this study, transition and transversion was believed to influence the change in protein quantity or type, as well as the change in bases (*e.g.* codon usage pattern, amino acid

composition, and DNA base content); therefore, all information was represented in one equation. In conclusion, we attempted to represent the amount of information changed by the DNA sequence using one equation. Consequently, we found that the viral protein changed continuously, while the quantity of protein decreased temporally prior to a pandemic, followed by an abrupt increase during the pandemic. A temporary decrease in information change indicates that the strain and sequence composition was almost comparable, as seen during the worldwide Spanish flu pandemic. This led to an exponential increase in the death rate, followed by changes in the viral protein, in turn leading to another pandemic. In other words, pandemics occur when a strain shows a similar degree of sequence change as the previous pandemic or epidemic, as shown by BIS. Substantially, the protein can change exponentially to adapt to a new host. Using these results, the sequence of a new strain can be compared to conventional sequences to predict the progression of an infection to a pandemic level in the future.

Acknowledgments

This research was supported by Development of biofield analysis technology for Qi-Blood status assessment funded by Korea Institute of Oriental Medicine (No.K16012).

References

- [1] T. T. Wang and P. Palese, "Unraveling the Mystery of Swine Influenza Virus", *Cell.*, vol. 137, (2009), pp. 983-985.
- [2] G. J. D. Smith, D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, S. K. Ma, C. L. Cheung, J. Raghvani, S. Bhatt, J. S. Malik Peiris, Y. Guan and A. Rambaut, "Origins and Evolutionary Genomics of the 2009 Swine-origin H1N1 Influenza A Epidemic", *Nature Letters.*, vol. 459, (2009), pp. 1122-1126.
- [3] G. Neumann, T. Noda and Y. Kawaoka, "Emergence and Pandemic Potential of Swine-origin H1N1 Influenza Virus", *Nature Reviews.*, vol. 459, (2009), pp. 931-939.
- [4] C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani and N. M. Ferguson, "Pandemic Potential of a strain of influenza A(H1N1): Early findings", *Science.*, vol. 324, (2009), pp. 1557-1561.
- [5] A. J. Bojarski, M. Nowak and B. Testa, "Conformational Constraints on Side Chains in Protein Residues Increase Their Information Content", *Cellular and Molecular Life Science.*, vol. 60, (2003), pp. 2526-2531.
- [6] H. Liao, W. Yeh, D. Chiang, R. L. Jemigan and B. Lusting, "Protein Sequence Entropy is Closely Related to Packing Density and Hydrophobicity", *Protein Engineering, Design & Selection.*, vol. 18, no. 2, (2005), pp.59-64.
- [7] J. J. Stewart, C. Y. Lee, S. Ibrahim, P. Watts, M. Shlomchik, M. Weigert and S. Litwin, "A Shannon Entropy Analysis of Immunoglobulin and T cell Receptor", *Molecular Immunology.*, vol. 34, no. 15, (1997), pp. 1067-1082.
- [8] Y. C. Dou, X. Q. Zheng and J. Wang, "Several Appropriate Background Distributions for Entropy-based Protein Sequence Conservation Measures", *Journal of Theoretical Biology.*, vol. 262, (2010), pp. 317-322.
- [9] M. Xiao, Z. Z. Zhu, J. P. Liu and C. Y. Zhang, "A New Method Based on Entropy Theory for Genomic Sequence Analysis", *Acta Biotheoretica.*, vol. 50, (2002), pp. 155-165.
- [10] W. Resch, N. Hoffman, and R. Swanstrom, "Improved Success of Phenotype Prediction of the Human Immunodeficiency Virus Type 1 from Envelope Variable Loop 3 Sequence using Neural Networks", *Virology.*, vol. 288, (2001), pp. 51-62.
- [11] Q. Y. Koo, A. M. Khan, K. O. Jung, S. Ramdas, O. Miotto, T. W. Tan, V. Brusi, J. Salmon and J. T. August, "Conservation and Variability of West Nile Virus proteins", *PLOS One.*, vol. 4, no.4, (2009).
- [12] J. W. Huang, C. C. King and J. M. Yang, "Co-evolution Positions and Rules for Antigenic Variants of Human Influenza A/H3N2 viruses", *BMC Bioinformatics.*, vol. 10, no. 1, (2009).
- [13] W. A. Thompson, S. H. Fan and J. K. Weltman, "Information Entropy of Influenza A Segment 7", *Entropy.*, vol. 10, (2008), pp. 736-744.
- [14] S. E. Bae, "Pattern Analysis of Pandemic/Epidemic/Reassortment of Influenza Virus", *Advanced Science and Technology Letters.*, vol. 128, (2016), pp. 145-148.
- [15] C.E. Shannon, "A Mathematical Theory of Communication", *The Bell System Technical Journal.*, vol. 27, (1948), pp. 379-423.

- [16] C.E. Shannon, "Prediction and Entropy of Printed English", The Bell System Technical Journal., vol. 30, (1951), pp. 50-64.
- [17] P. S. Shenkin, B. Erman, and L. D, "Mastrandrea. Information-Theoretical Entropy as a Measure of Sequence Variability", Proteins: Structure, Function, and Genetics., vol. 11, (1991), pp. 297-313.
- [18] M. I. Nelson, C. Viboud, L. Simonsin, R. T. Bennett, S. B. Griesemer, K. St. George, J. Taylor, D. J. Spiro, N. A. Sengamalay, E. Ghedin, J. K. Taubenberger and E. C. Holmes, "Multiple Reassortment Events in the Evolutionary History of H1N1 Influenza A Virus sin 1918", PLOS Pathogens., vol. 4, no. 2, (2008).

Authors



Se-Eun Bae, she is a Senior researcher from 2015~present: Korea Institute of oriental Medicine; 2013~2014: Korea Advanced Institute of Science and Technology, Research Professor; 2012: National Science and Technology Council, Deputy Director; 2011: Seoul National University, Doctor of Science.