William A. Thompson,
Andy Martwick, and Joel K. Weltman

Examining H1N1 Through Its Information Entropy

As reported by CNN in October 2009 [1], the declaration of the H1N1 (swine flu) pandemic as a national emergency in the United States highlights the magnitude and continuing spread of the first pandemic flu outbreak in 40 years. The H1N1 (swine flu) pandemic that began in March 2009 is the most recent outbreak of type A influenza to threaten the human species [2]. Type A influenza viruses, including the H1N1 of the current pandemic, are the cause of the most frequent and most serious influenza infections in humans [3].

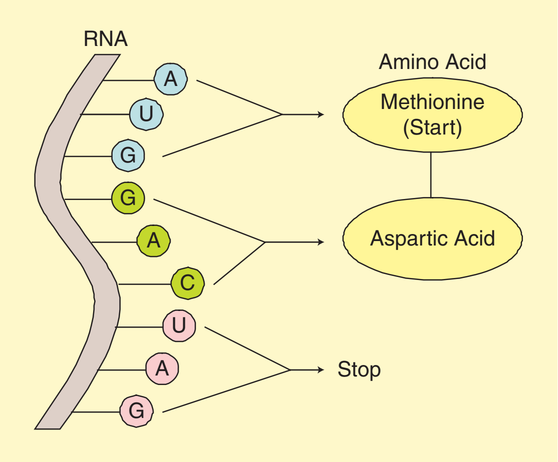
Epidemics of influenza have occurred throughout history. The great influenza pandemic of 1918–1919 killed more than 500,000 people in the United States and 50 million people worldwide [4]. More than 40,000 people are killed each year by influenza in the United States. In this article, we discuss how signal processing techniques can be used to analyze the information content of H1N1 genomic sequences, which could help in understanding this and future flu outbreaks. We first explain how the genetic code is used by the influenza virus to encode the information that is essential to viral function. Next, we demonstrate how the digital signal processing tools of decimation, periodicity, and integration are used to reveal the patterns of entropic uncertainty of that information. We propose that analysis of these patterns changing over time and space can give insight into mechanisms of regional influenza epidemics and global influenza pandemics that may help provide a basis for design of counter measures.

Digital Object Identifier 10.1109/MSP.2010.936017

INFLUENZA AND THE GENETIC CODE

H1N1 is a subtype of the influenza A virus. Influenza viruses are classified into three types: A, B, and C, depending upon their ribonucleic acid (RNA) sequences. The genetic information of influenza is stored and transmitted as RNA and not as deoxyribonucleic acid (DNA) as it is in DNA viruses and in higher organisms such as plants and animals. The RNA molecules of the influenza virus consist of separate single stranded nucleotide chains, referred to as "segments." In H1N1, the sequences of nucleotides in the RNA encode the genes required to produce the viral proteins [5]. Each nucleotide consists of a base, a ribose sugar, and a phosphate. The phosphate connects one ribose to the next. The bases are ring structures containing nitrogen attached to the ribose sugars. In RNA, the bases are adenine (A), cytosine (C), guanine (G), and uracil (U). The sequences of these bases encode the information content of the RNA molecule.

In biological systems, the information stored in the nucleic acid sequences is translated into sequences of amino acids that make up the proteins carrying out the functions of the cells. Proteins are thus the chief action molecules of life. The rules specifying the translation of nucleotide sequences to protein sequences are called the genetic code. The code defines the mapping between sequences of three nucleotides, called codons, and 20 different amino acids plus three stop signals. Translation typically begins with the codon AUG and continues until one of the three stop signals is encountered. The occurrence of any of the three stop signals (UAA, UGA, or UAG) halts the translation of the nucleotide sequences into amino acids. This is illustrated in Figure 1 for three codons translating into a hypothetical protein of two amino acids, e.g., a dipeptide.



[FIG1] Example RNA segment of three codons {AUG, GAC, and UAG} translating into two amino acids.

Using a computer analogy, the genetic code is base four {U, A, G, C} and the genetic symbol code word, a codon, is 3-b long. The three nucleotides are the bits of the codon. The sequence of codons in a gene determines the structure of the protein that is to be assembled from the 20 different amino acids. There are 64 possibilities from the three nucleotide combinations (i.e., 4^3 possibilities) and only 20 amino acids to encode; therefore, the genetic code is redundant. Invoking the computer analogy again, this level of redundancy is what would be expected for the encoding of some type of error correcting code. Most of this redundancy is in the third bit of the codon.

When an organism reproduces, the genes are replicated. In higher organisms, there are extensive error correcting and repair mechanisms during replication. In RNA viruses there is effectively no error checking of the code when the gene copies itself [6]. Genetic mutations during replication are quite likely at all positions in the gene. Some changes in the gene do not change the amino acid sequence because of the redundancy in the genetic code. For example, both AAU and AAC encode the amino acid Isoleucine. Such changes in the codon nucleotides are called synonymous mutations.

Nonsynonymous mutations in the gene alter the amino acid sequence. Changes in the amino acid sequence can potentially change critical parameters of the protein such as shape, dipole moment, surface charge, and folding abilities. Such changes may affect a protein's ability to function or to interact with other molecules in the biological network. Some changes to the amino acid sequence are not critical to the function of the protein.

The principle of natural selection dictates that mutations reducing the fitness and function of the organism are unlikely to survive and will not be passed on to future generations. Other changes may improve fitness, while some do not affect fitness and merely produce greater genetic variation with no apparent survival benefit. Fitness is usually measured by the organism's ability to reproduce. The lack of error checking is what accounts for the rapid mutations found in RNA viruses. It is these mutations in the noncritical regions of proteins that cause our immune system to no longer recognize a particular virus. The critical function of the protein does not change so the virus survives, but the outer proteins of the virus looks different, and our antibodies no longer recognize the strain of virus. This is why we need a new flu shot every year to protect against these mutations [7].

In the genetic code, 57 of the 61 (93.4%) codons specifying amino acids allow synonymous mutations at position three. We are studying these synonymous mutations both as a measure of genetic diversity and of critical function. One would expect some genetic regions to be highly conserved (critical function) and other regions to be more highly mutated (genetic diversity). Due to the redundancy in the codons encoding the amino acids, there are significant differences in the number of mutations between the first two nucleotides and the third.

H1N1 SPECIAL PROPERTIES

Influenza A viruses consist of eight separate segments of RNA potentially encoding 11 distinct proteins, labeled: PB2, PB1, PB1-F2, PA, HA, NP, NA, M1, M2, NS1, and NS2. At the time the studies presented here were conducted (October 2009) the pandemic H1N1 virus was expressing only ten of these proteins [2]. This is fortunate because the unexpressed protein (PB1-F2) may have been a significant virulence factor in the H1N1 virus that caused the 1918–1919 pandemic. PB1-F2 sequences are switched off by a stop signal early in the genomic sequence encoding PB1-F2. However, by January 2010, the FLU project database contained one reported PB1-F2 sequence in a virus that had been isolated in July 2009. It will be vital to see whether reports of PB1-F2 increase over the coming months.

During the current H1N1 pandemic, viral samples are being obtained from large numbers of infected people worldwide. The sequences of the nucleotides in these H1N1 influenza A viruses are being determined and made available on the The National Center for Biotechnology Information (NCBI) Influenza Virus Resource Database (http://www.ncbi.nlm.nih.gov/genomes/FLU/). We downloaded the nucleotide sequences for the genes of the 2009 pandemic H1N1 virus (1,563 sequences), archival human H1N1 (7,917 sequences) and archival swine H1N1 (1,191 sequences). The archival H1N1 virus sequences spanned the time from 1931 to 2008. We analyzed the information content in these viral sequences using decimation to separate the bits and integration to elucidate the genetic organization of the pandemic influenza viruses, as previously reported [8].

INFORMATION ENTROPY OF H1N1

For each of the three sets of H1N1 viruses (pandemic, human archival, and swine archival), the nucleotide sequences for each protein gene were aligned and the number of occurrences of each nucleotide in each aligned column was counted. The probability of a particular nucleotide occurring was calculated as the number of occurrences of the nucleotide in the column divided by the total number of sequences in the set. Following the procedure first given by Shannon, the information entropy was then computed from the probability of occurrence of each nucleotide base at every position of each set of sequences. The result is an information entropy vector for each gene, for each of the three sets. The greater the entropy, the greater the diversity of the nucleotides. The diversity of nucleotides, and of the proteins encoded by those nucleotides, helps explain the ability of the influenza virus to evade our immune responses to infection and to vaccines. The meaning of the diversity of synonymous nucleotide substitutions remains unclear. Diversity at the level of the nucleotide position, as measured by entropy, may reflect levels of organization and biological interactions still to be explored.

Our digital signal processing analysis of the information entropy of influenza A is based upon the periodicity of the distribution of that entropy, as revealed by fast Fourier transform (FFT) of vectors of entropy values. Figure 2 shows an entropy vector and a FFT of the vector of entropy values for the 2,277 nucleotide positions of the human archival H1N1 PB2 gene. The data in Figure 2 are based upon 776 sequences of the PB2 gene, spanning the years 1933–2008. The periodicity, i.e., one/frequency, of the transformed entropy is equal to three. This periodicity provides the factor used for the decimation of the sequences of entropy values, as next described.

In the decimation process, the data sampling rate was reduced by a factor of three by sampling every third nucleotide position throughout the entire length of the entropy vector. Decrease in sampling rate is referred to as "decimation" in digital signal processing. The decimation in this study was performed by array multiplication to maintain the correct alignment with respect to nucleotide position number. To perform the decimation, entropy vectors were element-by-element multiplied by one of the following vectors:

$$\text{array001} = [0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, \ldots 0]$$

$$\text{array010} = [0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, \ldots 0]$$

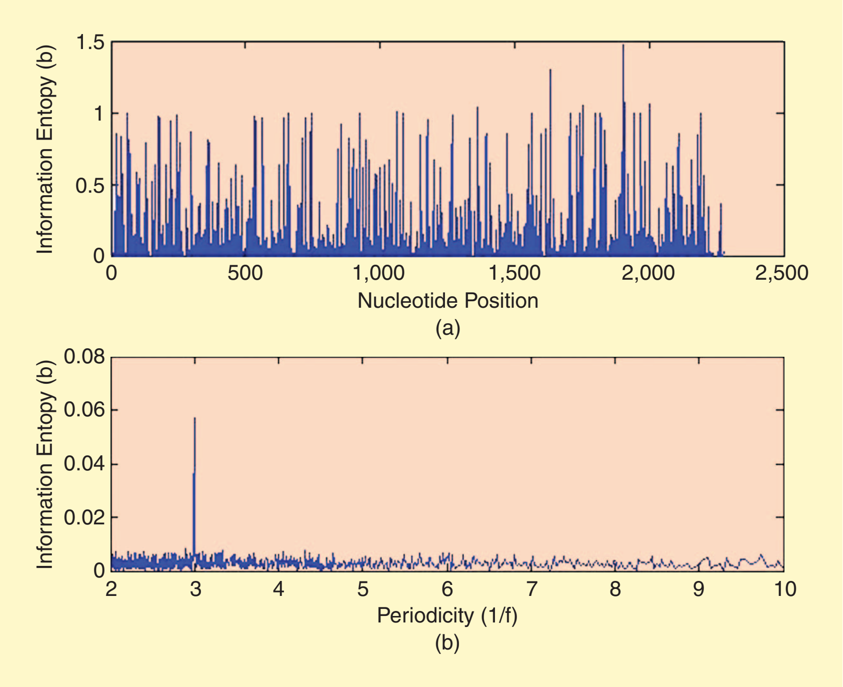
$$\text{array100} = [1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, \ldots 1].$$

Decimation produced by multiplication by array001 is designated 0, 0,1-decimation, decimation by array010 is designated 0, 1, 0-decimation, and decimation by multiplication by array100 is designated 1, 0, 0-decimation. The entropy vector for each gene was decimated to separate the nucleotides into three vectors for each gene. The vectors consisted of entropy values for the nucleotides of position one (100 frame), two (010 frame), and three (001 frame) of each codon. The cumulative entropy was then computed along each decimated vector. Python code for information entropy and for the decimation procedure is available at http://www.brown.edu/Research/Allergy/info_entropy_py.html.

RESULTS

Cumulative entropy in the third position (001 frame) is shown in Figure 3 for the ten expressed genes of the pandemic human H1N1influenza virus data set and for the corresponding genes of the archival human and swine H1N1 data sets.

In all cases, the total cumulative entropy of the pandemic influenza genes was less than that of the corresponding archival human and swine H1N1 genes. This indicates that the pandemic H1N1 genes have not mutated to the degree seen overall in the archival genes. Thus, each of the genes of the pandemic virus is currently in a low entropic state, suggesting a possible adaptation of the virus to biological constraints. Increases in entropy of the pandemic virus can be studied with respect to time and space, i.e., geographic location. Emerging entropic patterns and correlations can be detected and investigated. It is by such means that we will learn the rules determining the biology of the influenza A virus and its genetic changes and will devise more methods to prevent and treat infection.



[FIG2] Information entropy and FFT of the PB2 gene of archival human H1N1 influenza A viruses. (a) The information entropy at each nucleotide position. (b) FFT of the information entropy vector. Periodicity = one/(normalized frequency).

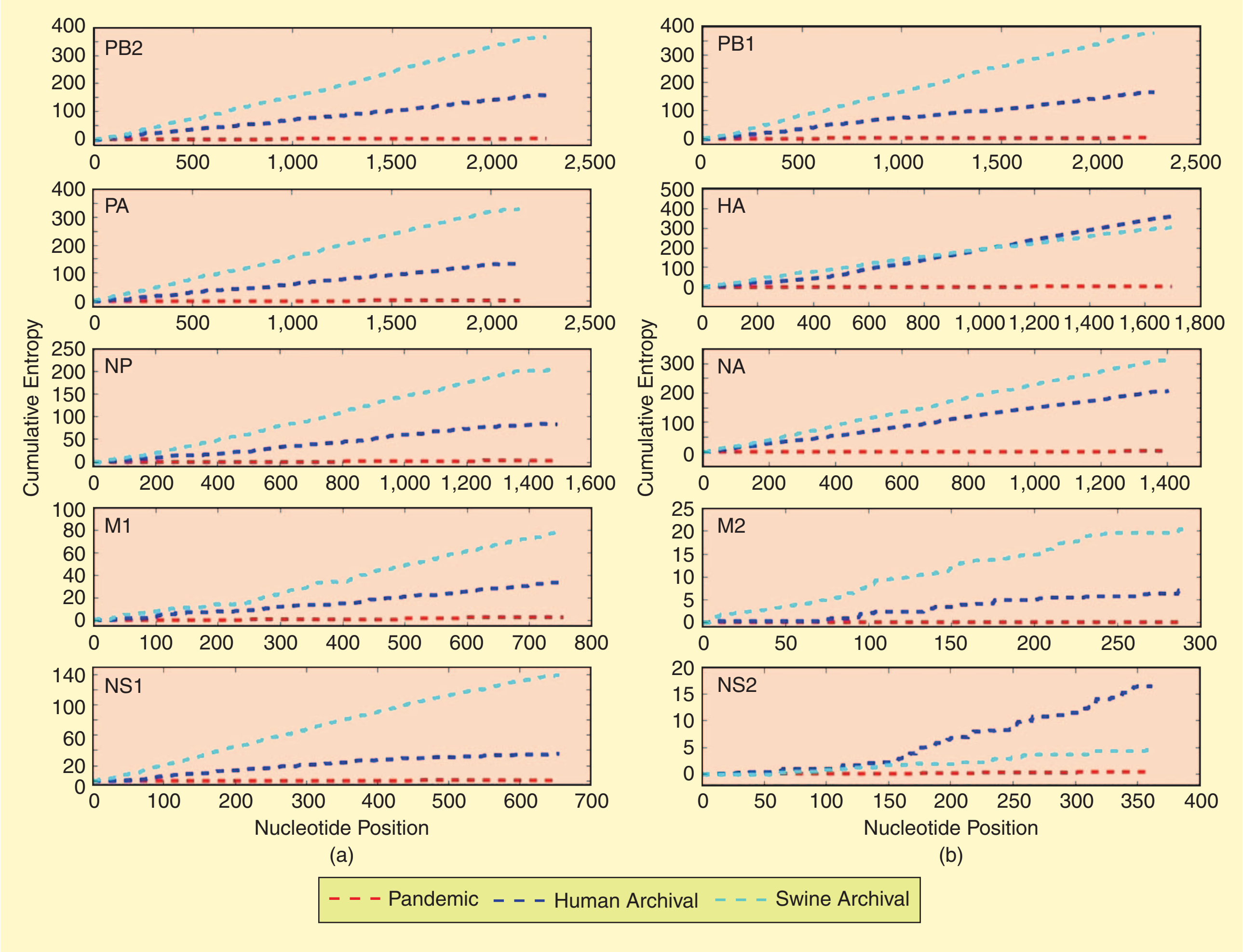
Analysis of the distribution of information entropy of the nucleotide sequences in the 001 frame by the decimation-integration (summation) method will be a practicable tool for studying the time course of the pandemic influenza virus' responses to those constraints. It should be noted that virtually all of the entropic changes in the 001 frame are undetectable at the protein level.

CONCLUSIONS

Information entropy vectors of the 2009 pandemic H1N1 influenza genes and of the corresponding genes of archival H1N1 human and swine influenza A reference data sets were separated into bits using decimation. The cumulative sum curves generated from the decimated data show that the information entropy in the third codon position of each gene of the pandemic H1N1 virus is less than that of the corresponding genes in the reference human and swine H1N1 data sets. Thus, the pandemic H1N1 influenza virus currently occupies a low entropy evolutionary niche. Decimation and integration, applied to viral information entropy, comprise a convenient and practicable method for following the evolution of the pandemic H1N1 influenza virus out of that niche.

Given the high error rate of the influenza RNA replication process [5], the low entropic state of all of its genes suggests interactions of the pandemic H1N1 influenza virus with biological constraints, i.e., resistance or biological push-back. The combined use of the digital signal processing method of decimation and the information entropy bioinformatic parameter provides a simple tool for analyzing those interactions in the viral sequences accumulating in unprecedented large numbers from throughout the world during this pandemic. We are currently analyzing correlation patterns in all three frames of the codon information entropy to help identify the underlying biological processes. By identifying and understanding these processes, new tools may become available for the treatment, control and tracking of the H1N1 virus.

It is important to analyze the entropy and diversity of influenza virus at the



[FIG3] Decimated cumulative information entropy arrays of pandemic (2009), human archival (1933–2008), and swine archival (1931–2008) H1N1 influenza virus genes in the 001 frame.

amino acid level and at the RNA segmental level, as has been reported [9], [10]. The digital signal processing tool of decimation provides a practicable and convenient method for expanding those studies to the information entropy at specific locations and codon positions within these sequences with fast computers, almost in real time. The next challenge will be to obtain such decimated vectors from this influenza pandemic and to expeditiously interpret them in a useful manner.

AUTHORS

William A. Thompson (william_thompson_1@brown.edu) is an assistant professor (research) in the Division of Applied Mathematics and Center for Computational Molecular Biology, Brown University.

Andy Martwick (martwick@pdx.edu) is an assistant professor in the Department of Physics, Portland State University.

Joel K. Weltman (joel_weltman@brown.edu) is a clinical professor Emeritus of Medicine in the Department of Medicine, Alpert Medical School of Brown University.

ACKNOWLEDGMENT

The authors thank the Brown University Center for Computation and Visualization for providing ancillary resources.

REFERENCES

[1] CNN. (2009, Oct. 26). Obama Declares H1N1 Emergency [Online]. Available: http://www.cnn.com/2009/HEALTH/10/24/h1n1.obama/

[2] G. Neumann, T. Noda, and Y. Kawaoka, "Emergence and pandemic potential of swine-origin H1N1 influenza virus," *Nature*, vol. 459, no. 7249, pp. 931–939, 2009.

[3] J. J. Treanor, "Influenza viruses, including avian influenza and swine influenza," in *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 7th ed., G. L. Mandell, J. E. Bennett, and R. Dolin, Eds. Philadelphia,PA: Churchill Livingstone/Elsevier, 2010, ch. 165

[4] World Health Organization Health Topics: Influenza (2010) [Online]. Available: http://www.who.int/topics/influenza/en/

[5] J. D. Chappell and T. S. Dermody, "Introduction to Viruses and Viral Diseases," in *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 7th ed., G. L. Mandell, J. E. Bennett, and R. Dolin, Eds. Philadelphia,PA: Churchill Livingstone/Elsevier, 2010, ch. 132.

[6] D. A. Steinhauer and J. J. Holland, "Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA," *J. Virol.*, vol. 57, no. 1, pp. 219–228, Jan. 1986.

[7] R. Dolin. (2008). *Influenza. Harrison's Principles of Internal Medicine*, 17th ed., A. S. Fauci, E. Braunwald, D. L. Kasper, S. L. Hauser, D. L. Longo, J. L. Jameson, and J. Loscalzo, Eds. New York: McGraw-Hill, ch. 180 [Online]. Available: http://www.accessmedicine.com/content.aspx?aid=2895663

[8] W. A. Thompson, A. Martwick, and J. K. Weltman, "Decimative multiplication of entropy arrays, with application to influenza," *Entropy*, vol. 11, no. 3, pp. 351–359, 2009. *Erratum: Entropy*, vol. 11, no. 3, p. 384, 2009.

[9] G. W. Chen and S. R. Shih, "Genomic signatures of influenza A pandemic (H1N1) 2009 virus," *Emerg. Infect. Dis.*, vol. 15, no. 12, pp. 1897–1903, Dec. 2009.

[10] A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes, "The genomic and epidemiological dynamics of human influenza A virus," *Nature*, vol. 453, no. 7195, pp. 615–619, May 2008.

[SP]