Information Entropy Analysis of the H1N1 Genetic Code ANDY
MARTWICK, Portland State, WILLIAM THOMPSON COLLABORATION[1],
JOEL WELTMAN COLLABORATION[2] — During the current H1N1 pandemic,
viral samples are being obtained from large numbers of infected people world-wide
and are being sequenced on the NCBI Influenza Virus Resource Database. The in-
formation entropy of the sequences was computed from the probability of occurrence
of each nucleotide base at every position of each set of sequences using Shannon's
definition of information entropy,

$$H = \sum_b p_b \, \log_2 \left(\frac{1}{p_b} \right)$$

where H is the observed information entropy at each nucleotide position and p_b is the
probability of the base pair of the nucleotides {A, C, G, U}. Information entropy
of the current H1N1 pandemic is compared to reference human and swine H1N1
entropy. As expected, the current H1N1 entropy is in a low entropy state and has
a very large mutation potential. Using the entropy method in mature genes we can
identify low entropy regions of nucleotides that generally correlate to critical protein
function.

[1]Department of Applied Mathematics, Brown University
[2]Department of Medicine, Brown University

Andy Martwick
Portland State

Date submitted: 20 Nov 2009   Electronic form version 1.4