Abstract

With the identification of a novel coronavirus associated with the severe acute respiratory syndrome (SARS), computational analysis of its RNA genome sequence is expected to give useful clues to help elucidate the origin, evolution, and pathogenicity of the virus. In this paper we study the collective counts of palindromes in the SARS genome along with all the completely sequenced coronaviruses. Based on a Markov-chain model for the genome sequence, the mean and standard deviation for the number of palindromes at or above a given length are derived. These theoretical results are complemented by extensive simulations to provide empirical estimates. Using a z score obtained from these mathematical and empirical means and standard deviations, we have observed that palindromes of length four are significantly under-represented in all the coronaviruses in our data set. In contrast, length-six palindromes are significantly under-represented only in the SARS coronavirus. Two other features are unique to the SARS sequence. First, there is a length-22 palindrome TCTTTAACAAGCTTGGTAAAGA spanning positions 25962–25983. Second, there are two repeating length-12 palindromes TTATAATTATAA spanning positions 22712–22723 and 22796–22807. Some further investigations into possible biological implications of these palindrome features are proposed.