Applying Bayes' Theorem to DNA Sequence for Identification of Pathogenic Bacteria

Cho, Min Jean

To develop an easy, simple method for identifying microorganisms based on their DNA sequences, Bayes' theorem was applied to DNA sequence analysis. It was hypothesized that the conditional probability of a DNA sequence from an unknown bacterial species being a member of a particular species could be the posterior probability, which could be estimated from prior probability and likelihood function using Bayes' theorem. To test the hypothesis, 16S rRNA gene sequences of foodborne pathogens (eight bacterial species) were downloaded from NIH's GenBank (45 sequences from each bacterial species, 360 sequences in total) to construct a database. Bayes' theorem was used to estimate the posterior probability of a bacterial specie "Si" given an unknown sequence "Q", P(Si|Q) = P(Q|Si)×P(Si)/P(Q). To determine the likelihood, P(Q|Si), the DNA sequence "Q" was divided into words (k-size DNA sequence fragments), and P(Q|Si) was measured from the average probability of observing the word j from species Si, P(wj|Si). The prior probability, P(Si), and P(Q) were calculated from the database sequences. The size of word (k) affected values of P(Q|Si) and P(Q). The optimum size of word (k) was determined to be 39 nucleotides. The overall performance of the developed method was evaluated by simulation tests using selected DNA sequences, and all tested sequences were correctly identified (accuracy, 100%). The developed method was considered to be especially useful to determine the taxonomy of bacteria. It may also be applied to human DNA sequences for forensic analysis.