CALIFORNIA STATE SCIENCE FAIR
2014 PROJECT SUMMARY

Name(s) **Min Jean Cho**	**Project Number** 34491

Project Title

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

Abstract

Objectives/Goals

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.

Methods/Materials

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) X 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.

Results

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. All test sequences showed the highest P(Si|Q) values for the species to which they belong, which indicated that the developed method correctly identified the test sequences (accuracy = 100%).

Conclusions/Discussion

The hypothesized algorithm was proven to work in the experiments carried out with DNA sequences of bacterial species. Dividing the unknown DNA sequence Q into small-size words (wj) was especially important to determine P(Q|Si) and P(Q). An unknown sequence should be classified into the species with the highest P(Si|Q) value (rank-based identification), which indicated the most probable species among the species included in the database.

Summary Statement

Bayes¡# theorem was applied to DNA sequence analysis in order to determine the conditional probability that a DNA sequence from an unknown species belongs to a particular species.

Help Received

My father assisted me with researching, and my mother read and edited my writing. I would also like to thank Ms. Julia Newman for her valuable advice and suggestions.