DETERMINING THE STATISTICAL SIGNIFICANCE OF OBSERVED FREQUENCIES OF SHORT DNA MOTIFS IN A GENOME

Philip E. Pfeiffer
Department of Computer Science
E-mail: pfeiffpe@notes.udayton.edu

Peter W. Hovey
Department of Mathematics
E-mail: Peter.Hovey@notes.udayton.edu

Sudhindra R. Gadagkar
Department of Biology
E-mail: gadagkar@notes.udayton.edu

University of Dayton
Dayton, OH 45469

Abstract: Until recently over 90 percent of the DNA in the human genome was considered junk DNA, with no known function. However, this non-coding DNA is now known to harbor elements that perform important functions in gene regulation. In particular, there is currently much interest in the search for short DNA motifs collectively known as *cis*-regulatory elements. Most studies attempt to identify these elements by means of cross-species comparisons. We have approached the problem of finding cis-regulatory elements by searching for conserved DNA motifs within genomes. This requires searching for DNA motifs that are repeated in the genomes either more or less frequently than expected by random chance. However, the usual chi-squared test cannot be used to test for the statistical significance of any observed frequency since overlapping regions of the genome are checked for DNA motif matches. We present here a statistical measure that has been developed to quantify the expectation and variance of the frequency of a given DNA motif in a given target sequence that may contain overlapping regions.