The Role of Statistics in DNA Evidence Interpretation

M. A. C. S. S. Fernando[a*]

a Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, P.O. Box 02, Belihuloya 70140, Sri Lanka.

*Correspondence: *sampathf73@gmail.com*

Forensic DNA analysis has received much public attention over the last thirty years because of its incredible usefulness in criminal investigations. It has also had considerable scientific scrutiny, mainly in response to changes in science and legal challenges. The field of statistics is paramount in DNA evidence interpretation because of the intrinsic probabilistic nature of the problem. Statistical DNA interpretation uses knowledge from the fields of: statistics, population genetics, and molecular biology. This has been in existence for longer than any of the current technologies used for typing the evidence, because the same basic ideas generally apply, regardless of how the evidence is typed.

In a typical criminal case, biological materials such as blood, semen, saliva, or other body tissues, may be recovered and between 50 to 100 picograms (10^{-12}g) of DNA is extracted from these materials for polymerase chain reaction (PCR) amplification. PCR amplification allows length variants in the DNA, called short tandem repeats (STRs), to be detected by measuring relative fluorescence when the sample is exposed to laser light. The resulting signal is collected by a photomultiplier and displayed graphically as an electropherogram (epg). The epg consists of a trace signal displayed on a molecular weight axis, which is mostly flat with peaks in various locations. The presence of a peak corresponds to the alleles present in the DNA sample. Crudely, alleles are variants or polymorphisms of a gene, which can be used to describe differences between individuals. The heights of the peaks are approximately proportional to the amount of template DNA present. This quantitative information (as opposed to the discrete allele information) can greatly enhance the interpretation process.

The likelihood ratio (*LR*) approach is (now) the favoured method for presenting forensic evidence in the court in many jurisdictions. It links the evidence related to two hypotheses: prosecution and defence. The prosecution hypothesis claims that the accused is the donor of recovered DNA from the crime scene while the defence claims that an unknown person who is not blood-related to the accused is the donor. Then the ratio between the probabilities of the two hypotheses is defined as *LR*. There are four competing models: classical, binary, semi-continuous, and continuous for the interpretation of DNA evidence, which essentially differ in the definitions of the weights assigned to sets of possible genotypes. Implementing continuous probabilistic models which ensure relatively greater objectivity and consistency between analysts requires statistical models for PCR phenomena such as stutter. This research reviews the existing models and develops new, advanced, Bayesian models for predicting stutter with increased accuracy. The models include nonhierarchical, hierarchical, and infinite mixture models.

Keywords: DNA, Electropherogram, Forensic, Likelihood ratio (LR), PCR, Stutter