SIMULATING FAMILIES FROM STR DATA DERIVED FROM FILES EXPORTED FROM CODIS

Daniel Myers, New York State Police Forensic Investigation Center

A familial simulation macro was written in Microsoft Visual Basic for Applications (VBA) for Excel for use with offender data in CODIS Interpol export format for validating a familial searching procedure. The offender data consisted of 1,100 genotypes (i.e., seeds) used to reverse engineer three simulated, biological relatives of each seed. The simulator first creates a set of parents by separating the seed alleles into two parent genotypes at each locus. The simulator then randomly chooses the second allele for each parent genotype from a table of alleles distributed according to an aggregate of the FBI African-American, Caucasian, and Southwestern Hispanic populations. The macro further randomly chooses one allele from each of these parent genotypes according to the Mendelian rules of inheritance to create a full sibling of the seed. The result is a family of 4 individuals, two simulated biological parents and two children (the seed and a simulated full-sibling of the seed). The ability to produce a large number of families was found to be particularly useful in estimating the sensitivity and specificity in the NYS Offender Index for familial searching of first order relatives. The simulations were processed in Excel worksheets, and then written to an XML file in CODIS Interpol format for upload into familial searching software. The usefulness of the macro warrants further development such that it will work directly with a document object model rather than an Excel file, potentially with VBA, Python or R programming languages, while maintaining compatibility with CODIS CMF export and Interpol format files. In addition to a familial searching validation tool, this simulator could be used as a basis for creating a simulated population of randomly mating individuals and for quality assurance testing of the sensitivity and specificity. Further, the random number generator could be used for creating a simulated database of individuals following the distribution of any entered population.