MUTUAL INDEPENDENCE OF LOCI IN DATABASES OF MULTI-LOCUS GENOTYPES: APPLICATION FOR HUMAN IDENTIFICATION AND RELATIONSHIP TESTING

Bing Song[1], August Woerner[2], Frank R.Wendt[1,2], Shande Chen[3], Bruce Budowle[1,2,4] ,Ranajit Chakraborty[1]
[1]Institute for Molecular Medicine, University of North Texas Health Science Center
[2]Center for Human Identification, University of North Texas Health Science Center
[3]School of Public Health, University of North Texas Health Science Center
[4]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University

Multi-locus genotype data, such as that organized in the Combined DNA Index System (CODIS), are widely used in human identification and relationship testing. Commonly reported statistics in these fields (i.e. random match probability and paternity index) rely on certain underlying assumptions such as the Product Rule and Assumption of Mutual Independence. Hardy-Weinberg Equilibrium and Linkage Disequilibrium of the genotype data are usually tested before multi-locus DNA profiles are used. However the mutual independence of multi-locus data is rarely tested due to the complicated definition and relatively tedious calculation(s). Mutual Independence is defined in statistics as any finite subset of a set of random variables that is independent with any other; that is, each pair, triplet, quartet, etc. of variables must satisfy the multiplication rule:

$$P\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} P(A_k)$$

To avoid cumbersome calculations, two summary statistics are reported: (1) number of heterozygous loci (K); and (2) number of shared alleles (X), whose additive formulas are based on the mutually independent assumptions.

Herein, the “MutualIndepend” package in R was built for building the distribution of observed and expected number of heterozygous loci and number of shared alleles. When applied to loci from the Short Tandem Repeat DNA Internet DataBase of National Institute of Standards and Technology (NIST STRBase), it is possible to determine whether or not to reject the null hypothesis of mutual independence for multi-locus DNA profiles.

By testing the mutual independence of multi-locus databases, it is possible to optimize these repositories to more appropriate utilize the data, as well as provide a method to evaluate markers lately defined. Moreover, by this method, different kinds of loci, like short tandem repeats (STRs), insertion/deletion polymorphisms (InDels), and single nucleotide polymorphisms (SNPs) can be combined into one statistic. If those different kinds of loci are tested balanced, databases can become maximally informative.