AN APPROACH TO SEARCHING THE US Y-STR DATABASE

Cristina P. Somolinos, MLIS, D-ABC, NJ State Police Office of Forensic Sciences

The US Y-STR database may be used to aid in estimating a haplotype frequency by way of the counting method and application of a confidence interval.

The US Y-STR database contains profiles which reflect the legacy of ever-expanding locus subsets in multiplex kits over the years. This leads to some difficulty when attempting to search all haplotypes, especially with partial profiles. Because of a quirk in the design of the search mechanism, the search on a locus will only compare to haplotypes that reported alleles in that locus.

I will discuss the philosophy behind and practical approach to the search protocol used in our laboratory. By searching the SWGDAM core loci of a haplotype, we are able to access all matching haplotypes on the database. Thereafter we may view all matching haplotypes and exclude some of these matches based on the alleles in the loci not originally searched. The resultant number of matches (and the size of the population searched) will be used to calculate the confidence interval, ensuring that the full population set of the database was appropriately searched to generate the best estimate of the frequency of a given haplotype.

We encountered other technical issues with the database's summary of results – including significant digits, rounding and truncation, as well as server timeouts with partial haplotype searches. In response, we developed our own Excel- and Visual Basic-based worksheet to calculate the Clopper-Pearson/exact interval.

Examples will be shown of the different results that may be obtained for the same haplotype in the database, depending on the size of the population subset being searched (or ignored). Haplotypes from proficiency tests and the Romanov haplotype will be used to demonstrate.
Our approach is in keeping with the 2014 SWGDAM Interpretation Guidelines for Y-Chromosome STR Typing by Forensic DNA Laboratories; however we will also examine the theta correction, as it is dependent on the number of loci we are searching.