FROM PEPTIDES TO RANDOM MATCH PROBABILITIES

August E. Woerner[1,2], Myles W. Gardner[3], F. Curtis Hewitt[3], Michael A. Freitas[4], Liwen Zhang[3], Maryam Baniasad[3], Kathleen Q. Schulte[3], Alan R. Smith[3], Nicolette C. Albright[3], Danielle S. LeSassier[3], Katharina Weber[3], Tara E. Manley[3], Leah W. Allen[3], Megan E. Powals[3], Benjamin C. Ludolph[3], Bruce Budowle[1,2]
[1]Center for Human Identification, University of North Texas Health Science Center
[2]Department of Microbiology, Immunology and Genetics, University of North Texas Health Science Center
[3]Signature Science, LLC
[4]The Ohio State University

The first genetic markers used in forensic science were based not on DNA sequences but instead on differences at the protein level, such as those inferred from allozymes. Over the last 40 years, population and forensic genetics has changed its medium from protein to DNA, for reasons both biological (e.g., the ability to amplify by PCR highly polymorphic loci such as VNTRs, STRs and SNPs) and methodological (e.g., capillary electrophoresis, sequencing). However, some evidentiary samples may have little DNA (e.g., hair and trace samples) and may instead have ample genetic signal in the form of proteins. Recent advances in peptide sequencing have enabled the identification of single amino acid polymorphisms (SAPs, which derive from SNPs); these SAPs can be used to identify individuals from challenged samples such as bone, hair, and ancient remains. Despite these advances in chemistry and signal detection, relatively little research has been invested in the statistical interpretation of peptide sequencing. There are several challenges in interpreting peptide expression—such as, contamination may cause the appearance of peptides that are not inherent, low template issues may cause alleles to dropout, additional variations in the peptide sequence may cause the sequence to be undetectable, protein expression may be allele-specific and linkage disequilibrium (LD) coupled with population structure/inbreeding may impact interpretation of the rarity of a given proteomic signal. We propose a method for estimating a random match probability (RMP) that addresses these issues. LD within chromosomes is accounted for by considering the expected protein products from exome resequencing databases (e.g., the 1000 genomes project). The rules for a correspondence between exomes and proteomes are relaxed to permit drop-out that can be random (e.g., arising from low template issues) or deterministic (e.g., allele-specific expression). Further, a Monte Carlo procedure is devised to simulate the effects of drop-in and a custom θ correction is implemented to address population structure. RMPs were estimated from peptide markers mined from 60,706 whole exomes from the ExAC project, typed in skin samples from 25 individuals and characterized by LC-MS/MS, yielding a median RMP of ~10^{-7}.