Mass casualty identification through DNA analysis:

overview, problems and pitfalls

Mark W. Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA

29 August 2007

Book chapter appears in:

Perlin, M. W. (2007). Mass casualty identification through DNA analysis: overview, problems and pitfalls. Forensic Investigation and Management of Mass Disasters. M. I. Okoye and C. H. Wecht. Tucson, AZ, Lawyers & Judges Publishing Co: 23-30.

Corresponding author contact information:
Dr. Mark W. Perlin
Cybergenetics
160 North Craig Street, Suite 210
Pittsburgh, PA 15213 USA
412.683.3004
412.683.3005 FAX
perlin@cybgen.com

Introduction

When people die in a mass disaster, they leave behind biological material. This biological material may be their entire body, or body parts. One important task of mass casualty identification is to identify these victim remains by associating them with missing people. Identification of victim remains from a mass casualty site is critical for bringing closure to family and loved ones.

A complementary forensic task is to identify the people who are missing after a mass casualty, and associate them with victim remains. These missing people typically leave behind biological material in their homes and with their families. One source of biological material for missing people is their personal effects, such as toothbrushes, hairbrushes and clothing. Another source of biological material is the missing people's family references, since relatives have biological features which are similar to those of the missing person.

The key task of mass casualty identification is to match the biological material from the mass casualty site to the missing people. This is done by identifying the biological features of each of the victim remains found at the mass casualty site. Separately, the biological features of the personal effects and of the family references are analyzed to form a biological profile of each missing person. By comparing the biological features of each of the victim remains against those of each of the missing people, a match may be

obtained between a particular victim remain found at the mass casualty site and one of the missing people.

DNA Identification

The DNA genetic code contained within each person is ideally suited for mass casualty identification. Each person's genome contains three billion DNA letters spread out over twenty-three pairs of human chromosomes. This genetic code contains features that are unique to each individual, and is therefore is ideal for forensic identification. Moreover, the chromosome pairs come from an individual's parents with one copy coming from the father and one copy coming from the mother. Therefore, an individual's genetic identity can be partially reconstituted from their family DNA.

Forensic identification is best done using distinguishing features that are relatively unique to an individual. In the human genome, there are hundreds of thousands of such highly polymorphic genetic locations or loci which differ greatly between individuals. One type is known as the short tandem repeat (or STR) which is an excellent marker that is now widely used for human identity. At each STR locus, a short DNA phrase comprising about four letters is tandemly repeated about ten to twenty times. At an STR locus, an individual will have two different copies of DNA, each having their own specific length. These pairs of DNA fragment lengths are the features which form the basis for STR genetic identity, as it is currently used in forensics.

STR data is generated from a biological specimen in a molecular biology laboratory in three steps. First, the DNA is extracted from the biological material. Next, ten to fifteen STR loci are amplified simultaneously in one tube one million-fold using the polymerase chain reaction (or PCR). Finally, the amplified PCR product is detected and separated on an automated DNA sequencer. The DNA signal that is produced contains peaks that correspond to the DNA fragments. The peak location on the x-axis corresponds to the length of the DNA fragment, while the peak height on the y-axis corresponds to the amount of DNA present. When an individual has two different STR lengths at the same locus (i.e., heterozygote), two peaks appear. When an individual has two copies of the same STR lenght at a locus (i.e., homozygote), one peak appears in the data.

The DNA profile of an individual is the listing of these pairs of allele lengths at each STR locus. With the ten to fifteen STR loci currently used in forensic practice, these pairs of allele lengths produce a virtually unique digital signature for an individual. Each STR locus provides (on average) a factor of ten of distinguishability from a person who might be randomly selected from the human population. Therefore, twelve STR loci provide (on average) 10^{12}, that is, a 1 followed by twelve 0's, giving a trillion to one relative uniqueness of DNA identifiability for that individual.

In principal then, the task of identifying human remains is arduous, but straightforward. A DNA laboratory produces an STR profile from each of the victim remains, and, separately, produces a DNA profile for each of the missing people. By comparing all the victim remains DNA profiles against all the missing person profiles, matches can be

obtained; these matches pair particular victim remains with particular missing people. This approach assumes, however, that perfect DNA profiles can be obtained from victim remains and from missing people.

Uncertain DNA Evidence

The damaged victim remains found at a mass casualty site do not produce pristine biological specimens. Rather, DNA that has been burned, degraded, mixed with other samples, or environmentally decomposed does not amplify well. Instead of producing pristine DNA data, damaged victim remains can produce multiple low level peaks that are not amenable to standard DNA interpretation methods.

This DNA data uncertainty can produce considerable uncertainty in the inferred DNA profile. For example, an ambiguous DNA peak pattern may be consistent with many different allele pair genotype alternatives. The data, however, do contain considerable DNA identification information, and a key forensic objective is to preserve evidence. Given the different DNA interpretation alternatives available, what, then, is the best DNA review method for examining compromised DNA data at a mass casualty site?

Match Information

One approach might be to allow a DNA profile inferred from ambiguous DNA data to match all possible DNA profiles. For example, the inferred profile might be comprised of

two wild card alleles that can match any possible reference profile. This DNA profile inference and matching approach is entirely non-specific, and is not at all useful.

This lack of utility is because inferring all genotype possibilities yields no match information. The "all possibilities" genetic profile, comprised of all possible genotypes, matches all other genotypes. Therefore, a match would occur with a 100% probability. When the relative frequency of an inferred profile is a 100%, the match information is 0 (where match information is the logarithm of one over the relative frequency).

More precisely, match information is defined as the ratio of two probabilities. The numerator is the probability of a specific match of the inferred DNA profile against a particular missing person. The denominator is the probability of a random match between the inferred victim remains profile and a random genotype in the population. This ratio, when used with unambiguous DNA profiles, is often referred to as a random match probability. Taking the logarithm of this ratio measures information. Therefore, match information is defined as the logarithm of the "specific to random" probability match ratio.

Human Review

Conservative review of DNA data is used by human STR analysts to avoid overstating the match results. When applied to uncertain STR data, conservative review can provide more than zero information, but not the full information that is present in the

DNA data. This "conservative" interpretation method entails reporting alleles, rather than genotypes. The genotypes are inferred afterwards by forming all possible pairs of these alleles. Wild card alleles are also allowed. For example, when there is one strong peak and other smaller peaks, the analyst would call the one "obligate" allele corresponding to the one tall peak. Then, any genotype allele pair that contains this tall peak allele, as well as any other allele, would be considered part of this genotype set.

Comparing a genotype that contains one allele and a wild card allele against a known or referenced genotype that also includes that allele will produce a match. Therefore, the numerator of the match ratio information ratio is 1, because the probability of a match is 1. The denominator of the match ratio is determined by the relative frequency of the inferred profile. In conservative review, for this example that would include all possible genotypes that contain this one allele, as well as any other allele. With a common allele, that possibility of all combinations which include this allele may encompass half (or 50%) of the genotypes in the population for that locus. Therefore, the match information (which is the base 10 logarithm of 1 over 0.50) increases from 0 to a small number such as 0.3.

Aggressive human review is a more informative approach to inferring STR profiles. The goal here is to try ruling out unlikely combinations of genotypes from the set of all possible allele pairs. Suppose that the uncertain STR data at the locus is comprised of a small peak 'a' and a large peak 'b'. Then the analyst might designate the genotype

possibilities [a b], as well as [b b], thereby producing just two of the possible genotypes. The inferred profile would be reported out as a list of the possible allele calls.

When this shorter list of inferred profiles from the victim remains sample is compared against a profile from a known missing person, there are fewer match possibilities. Therefore, when there is a match, the numerator (which is the probability of the match) is 1. However, the denominator contains a smaller number which reflects the reduced relative frequency of this more specific DNA feature. For example, a small set of human genotypes may only match 20% of the random population. Therefore, the match information would be the logarithm of 1 over 0.20. This increased specificity in the inferred genotype would therefore increase the match information value to 0.7.

Genetic Profile - Victim Remains

A more scientific review of the STR data would infer DNA profiles in a way that preserves all the match information. This is done by assigning each candidate genotype a probability. Probabilistic genotypes have been used in genetics for over a century. When parent genotypes are known, Mendel's Laws dictate the possible genotypes of the offspring, as well as the probability of each candidate. For example, when a heterozygote father is combined with a homozygote mother having a different allele, there are exactly two genotype possibilities for the child, each having a probability of one half.

Just as there is genetic uncertainty in human inheritance, there can be data uncertainty in the STR peak information. The conservative and aggressive human review methods create lists of genotypes that assign each candidate on the list the same uniform probability. A scientific review of the data also creates a list of genotype candidates, but instead using the same probability for every genotype, it assigns an individualized probability to each genotype based on the data. The result is that more likely genotypes will have a higher probability than less likely genotypes.

When an inferred victim remain genetic profile is compared against a known missing person reference, the more probable genotypes in the genetic profile will have greater weight in the match. This increased weighting will cause the relative frequency of the true genotype to predominate in the match statistic. The result is that the match information will reflect the relative frequency of the true matching inferred profile, instead of a composite that is less specific. For example, if the relative frequency of the true genotype is 10%, that will lead to a match information of equal to the logarithm of 1 over 10% (i.e., $\log_{10}(1/(1/10)) = \log_{10}(10) = 1$), which equals 1 match information unit. In this way, we see that using probabilities of genotypes to represent a genetic profile will tend to preserve match information, and therefore produce a greater match association statistic between victim remain biological specimens and missing people.

Genetic Profiles - Missing People

The genetic profiles of missing people can be formed using family DNA reference material. Geneticists have long been able to reconstruct the genetics profiles of individuals from their family members. For every mother-father-child relationship triple, there is potentially useful genetic information. The resulting genetic profile is a probability distribution over genotypes. These genetic profiles of missing people can be compared against victim remain genetic profiles to produce matches. When personal effects are not available for an individual, such kinship derived genetic profiles may be the only way to establish a match with victim remains.

As with victim remains, personal effects from missing people can also produce compromised biological material, such as DNA mixtures. These materials can therefore produce uncertain data which similarly requires a scientific review. By assigning probabilities to the genotypes at a locus, such as 90% for one possibility and 10% for another possibility, a more specific genetic profile is produced. Comparing a more specific missing person genetic profile against a victim remains genetic profile can produce more match information. Whereas approximate DNA profile inference using conservative or aggressive human methods tends to reduce match information with damaged personal effects, scientifically inferred profiles can preserve this match information.

DNA Mixtures

DNA mixtures occur when a biological specimen contains material from more than one individual. Such DNA mixtures can produce uncertain data that contains more than two allele peaks at each locus. In particular, human identification of the individual component contributors may not be possible from this data. DNA mixture analysis has been a longstanding limitation of human review. However, in mass casualty DNA analysis, DNA mixtures are found both among the victim remains, as well as in the personal reference material. For example, a missing person's toothbrush may have been used by that person and by their spouse.

Fortunately, mathematical and computer solutions have been developed to resolve mixed DNA samples. One such approach is linear mixture analysis [1], which can represent known and unknown genetic profiles using coupled linear equations. Computer solution of these mathematical equations can produce the genetic profiles of the contributors. At each contributor locus, the genetic profile is represented by a probability distribution of genotypes. These inferred mixture genetic profiles can then be used to match victim remains against missing people.

Computer based mixture resolution is about a thousand times more informative than conventional human review methods [2]. This has been demonstrated in scientific studies that have compared dual human review versus computer interpretation on the same data [2]. In addition to being more informative than human review, computer

mixture interpretation with linear mixture analysis is more reproducible than human review [3]. Whereas human analysts can often generate different genetic profiles from the same mixture data, this variation is much less with computer implementation with appropriate mathematical models [3].

Inferred contributor profiles from DNA data can be matched just like any other genetic profile. When a victim remains biological specimen contains two contributors, two contributor profiles may be produced instead of one. Similarly, when the personal effects from a missing person contain more than one contributor, distinct genetic profiles for that individual can be produced. The matching comparison then proceeds by comparing one or more inferred profiles from the victim remains against the one or more genetic profiles that have been inferred from the missing person. A match between the genetic profiles in this case indicates that some component of a biological specimen matches some component of another biological specimen.

Scientific Calculator

It would be helpful to have a scientific calculator that could assist the human DNA analyst in examining evidence for a mass casualty. A typical scientific calculator permits a DNA analyst to add, subtract, multiply, divide and perform other straightforward operations. However, three functionalities are missing from this quantitative device. The first is a button to interpret DNA by conducting a scientific review of the original quantitative STR data. A second useful button would be one

which matches DNA by comparing two different profiles. Third, it would be helpful to have a button for a database engine that could record the results of all these profile interpretations and the matches between them. We shall discuss the development of such a scientific calculator device in the next chapter.

Conclusion

Uncertain DNA data does not pose a problem for computer based massed casualty analysis. The proper scientific approach is to use DNA profiles with probability. These probability based genetic profiles contain more human identification information than profiles produced with conventional human review. These general DNA profiles permit informative DNA analysis of damaged victim remains, damaged personal effects, integrated family references, DNA mixture samples, and multiple sources of DNA data. Using more informative DNA profiles produces a more informative DNA match between victim remains and missing people that preserves the DNA evidence.

References

[1] Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences* 2001;46(6):1372-1377.

[2] Perlin MW. Real-time DNA investigation. In: *Promega's Sixteenth International Symposium on Human Identification*; Dallas, TX; 2005.

[3] Perlin MW. Scientific validation of mixture interpretation methods. In: *Promega's Seventeenth International Symposium on Human Identification*; Nashville, TN; 2006.