Int J Legal Med (2009) 123:59–63
DOI 10.1007/s00414-008-0239-1

TECHNICAL NOTE

Partial matches in heterogeneous offender databases do not call into question the validity of random match probability calculations

Bruce Budowle · F. Samuel Baechtel · Ranajit Chakraborty

Received: 5 November 2007 / Accepted: 25 March 2008 / Published online: 6 May 2008

Abstract Offender DNA databases have been highly successful tools for generating investigative leads. Due to their success, the database sizes have increased such that some have suggested using the DNA profiles in offender databases for empirical pairwise studies to provide inferences regarding the validity of the current practices for generating random match probability estimates. These critics use observations under the assumption of independence to suggest that the current forensic DNA statistical calculations are invalid. However, some of these databases, such as CODIS, are not appropriate for such studies because they contain duplicate profiles and profiles of close relatives and are highly heterogeneous (i.e., comprised of individuals from many different population groups with unknown proportions). Observed departures from expectations will occur using these databases, but would have no relevance for questioning the reliability of statistical practices because the very heterogeneous data sets would be expected to violate the basic assumptions of independence. In addition, 9-, 10-, 11-, and 12-locus (out of 13 loci) matching profiles have been observed, are expected, and do not call into question the reliability of statistical practices. The phenomenon of matching profiles is similar to the concept of the birthday scenario. Regardless, simple computations under the assumption of independence for guideline purposes only show that partial matches observed in offender databases are not inconsistent with expectations. Indeed, computed random match probabilities that explain the observed matching profiles from pairwise comparisons are smaller than those observed based on routine casework calculations. Data analyses from offender databases based on assumptions of independence do not provide any basis for questioning the legitimacy of computations of random match probability values of any specific target profile based on the modified product rule that are currently followed in the DNA forensic community. Defined population data, which are sufficiently abundant, have already demonstrated the validity of the basic assumptions of DNA forensic statistical assumptions.

Keywords Matching profiles · Statistics · Random match probabilities · Pairwise comparisons · Population heterogeneity · STR loci · Offender databases

B. Budowle (✉) · F. S. Baechtel
FBI Laboratory,
2501 Investigation Parkway,
Quantico, VA 22135, USA
e-mail: bruce.budowle@ic.fbi.gov

R. Chakraborty
Center for Genome Information,
Department of Environmental Health, College of Medicine,
University of Cincinnati,
3223 Eden Avenue,
Cincinnati, OH 45276, USA

Introduction

Because of the success of offender DNA databases [1, 2], their size has increased substantially. For example, CODIS for the US Combined DNA Index System and NDNAD for the UK National DNA Database each contain more than four million profiles. There are suggestions that offender database(s) could be used for empirical studies to provide inferences regarding the validity of the current practices for generating random match probability estimates (as described in [3]). In some US legal proceedings [4], some have suggested that the demonstration of matching pairs of 9 (10, 11, 12, or 13) loci out 13 loci profiles in these large data sets violate the Hardy–Weinberg expectations (HWE). Because the HWE are violated, these critics argue that the currently used statistical approaches are invalid and that DNA evidence should not be admitted in court proceedings. Such suggestions are misleading because current statistical practices do not strictly follow the assumptions of HWE [3], and offender databases, such as CODIS, are heterogeneous and would be expected to depart from HWE.

This technical note is *not* about developing a conservative statistical approach by using θ adjustment formulas for ameliorating the effects of population heterogeneity (substructure) and when using average allele frequencies for statistical calculations. The NRC II Report [3] assumed that departures from HWE would occur and already offers θ adjustment formulas, which are routinely used by the forensic community, for such phenomena. Indeed, Weir [5] demonstrated that θ adjustment using the pragmatic values recommended by the NRC II Report [3] are more than adequate for overcoming effects found in heterogeneous databases similar (but not exactly) to the construct of the CODIS database. In contrast, this paper identifies the flaws in arguments raised by those in recent legal proceedings [4] who do not use θ adjustment formulas but instead strictly assume HWE and misapply average allele frequencies to assess departures from expectations in heterogeneous database.

Use of large offender databases to question strict rule of independence

Pairwise profile analysis can be a meaningful test of the reliability of the basic statistical assumptions used for generating DNA profile frequencies [6]. Therefore, someone might perceive the available profile data in these large offender databases as an opportunity to carry out pairwise profile comparisons studies (and compare the observations with expectations under the assumption of HWE). Some [4] have suggested that observing partial matching profiles, such as nine-locus profiles (out of 13 loci) sharing the same genotype, invalidate the manner that forensic laboratories calculate the rarity of a DNA profile. However, such logic is flawed. Departures from HWE are expected and any results obtained from such studies would not be relevant and would be misleading. The CODIS DNA database is comprised of very diverse populations, and the profiles are not apportioned into population categories such as is used for routine casework statistical calculations. Therefore, any analyses under the assumption of HWE are not particularly informative because departures are expected in heterogeneous data sets and the results do not assess the impact of using the population data sets employed for current statistical calculations.

Another obstacle to using the CODIS database for evaluating (under assumptions of HWE) the statistical legitimacy of using allele frequency estimates under current forensic practices is that the databases contain duplicate profiles and profiles of close relatives. Before any such inferences could be drawn from a pairwise database analysis, it would be imperative to remove matching or partially matching profiles contributed by relatives. The θ adjustment approach does not address directly the presence of relatives. Moreover, the removal of such profiles would be a monumental task that would have to be coordinated by all 50 states.

The impact of the presence of relatives in the database may be illustrated by Fig. 1 where the distribution of matching loci in pairwise comparisons of DNA profiles in four hypothetical databases is plotted. The distributions clearly show that the number of matched loci becomes highly distorted toward the direction of a larger number of matched loci in the presence of relatives in the database, and the deviation depends on the extent of the number of relatives as well. Note that while the distribution of matched loci is affected by the presence of relatives in the database, as long as the loci do not have any viability or fertility consequences, the presence of relatives in a database does not influence allele frequency estimates of the loci (data not shown). Thus, a few relatives in a database will increase the number of matching loci observed by pairwise comparisons.

An advocate of using an offender database for evaluating the validity of the statistical practices would now have to ignore concerns about using a very heterogeneous database [7]. The CODIS database qualifies as one of the most heterogeneous DNA profile databases available. It is comprised of individuals from many different population groups (African American, Asian, Caucasian, Hispanic, Native American, and Oceanian), the proportions are unknown, and it is likely that the proportions of these groups in the database are not the same as they are in the greater US population.

Frequency of pairwise comparison matches within offender databases as a surrogate for legitimate profile frequency estimates

When conducting pairwise DNA profile comparisons using offender database data, it is important to recognize that the number of profile matches that might be found at nine or more loci (out of 13 loci) is predictable. Troyer et al. [8] reported a nine-locus (i.e., partial out of 13 loci) match between an African American and a Caucasian profile in the Arizona offender database. The random match proba-

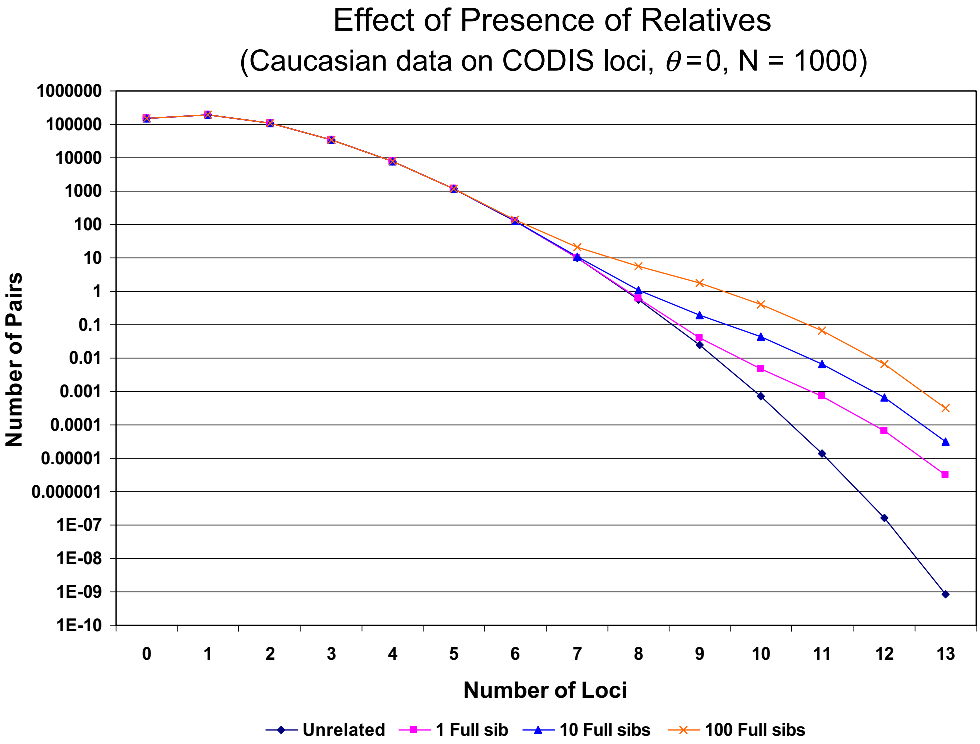


Fig. 1 Effect of the presence of relatives on the number of matching loci. The four hypothetical databases illustrated consist of 13 CODIS STR loci profiles (based on Caucasian allele frequencies, as reported in [10]) on 1,000 individuals in which all 1,000 individuals were unrelated (*filled diamonds*), 998 were unrelated and 1 pair of full siblings was included (*filled squares*), 980 were unrelated and 10 pairs of full siblings were included (*filled triangles*), and 800 were unrelated and 100 pairs of full siblings were included (*x*). Note that the *Y*-axis is in logarithmic scale, visually decreasing the degree of deviation

bility (RMP) frequency under the assumption of HWE for the nine loci was approximately 1/500,000,000. Yet, the database size contained only 8–10,000 profiles. The “coincidental match” did seem surprising to Troyer et al. [8] as the observation appeared to associate two unlikely individuals at a frequency more likely than seemed plausible. However, such a finding is entirely expected and predictable based on probability theory.

The matching pair at nine out of 13 loci that was observed is analogous to the well-known phenomenon the “birthday scenario” [9]. Similarly, matches will occur in offender databases, although at first glance it may seem counterintuitive with profile frequencies estimated to be less than 10^{-12}. The total number of pairwise comparisons for an *n* individual size database is $n(n-1)/2$. So for a database with approximately 3,000,000 profiles, there are more than four trillion pairwise comparisons. Although the birthday scenario and the observation of matching profiles at 9, 10, 11, or 12 (out of 13 loci) in large-sized offender databases would seem obvious to the informed, such statistics from pairwise comparisons of profiles (again under the assumptions of HWE) are cited as a rationale, although erroneous, to question current statistical practices for estimating the rarity of a DNA evidence profile. Although average allele frequencies are used, a threshold may be determined as a guide to determine whether such observations can be expected. The computations below illustrate that such statistics of partial matches in offender databases do not dispute the legitimacy of the reported RMP values in casework analyses that use the modified product rule. We do not advocate the following calculations as accurate, they are merely used to show that the expectations even when assuming HWE and average allele frequencies (as some critics might use) of observing partial matches are well within the plausible range.

RMP and partial match computations

The computations are based on partial matches in the Arizona State offenders' database. The Arizona offender database contained 65,493 offender 13-locus DNA profiles. In a pairwise comparison of these profiles, there were observed 122 pairs of profiles that matched at 9 loci, 20 pairs matched at 10 loci, and 1 pair each matched at 11 and 12 loci (K. Troyer and D. Duplissa, Arizona Department of Public Safety, Phoenix, Arizona, personal communication).

To exemplify what range of values of RMP (under HWE) would support such observations, the steps of the computations are: (1) compute the number of pairwise comparisons for the databases; (2) compute the number of possible combinations of loci (out of 13 loci) with reference to which the autosearch statistics were reported; (3) compute the range of possible number of distinct genotypes for the combination of loci; and (4) finally, compute a value of RMP that would be consistent with the observed number of partial matches.

The results are: (1) With n=65,493 profiles there are 2,144,633,778 pairwise comparisons of profiles in the AZ

Table 1 Number of alleles with frequencies ≥0.01 and genotypes for the STR loci

Loci	Number of segregating alleles (k)[a]	Possible number of genotype, $k(k+1)/2$
CSF1PO	8	36
FGA	21	231
TH01	6	21
TPOX	7	28
vWA	9	45
D3S1358	8	36
D5S818	8	36
D7S820	8	36
D8S1179	10	55
D13S317	7	28
D16S539	7	28
D18S51	15	120
D21S11	17	153

Data from [10].

[a] Number of segregating alleles is based on total observed at each locus.

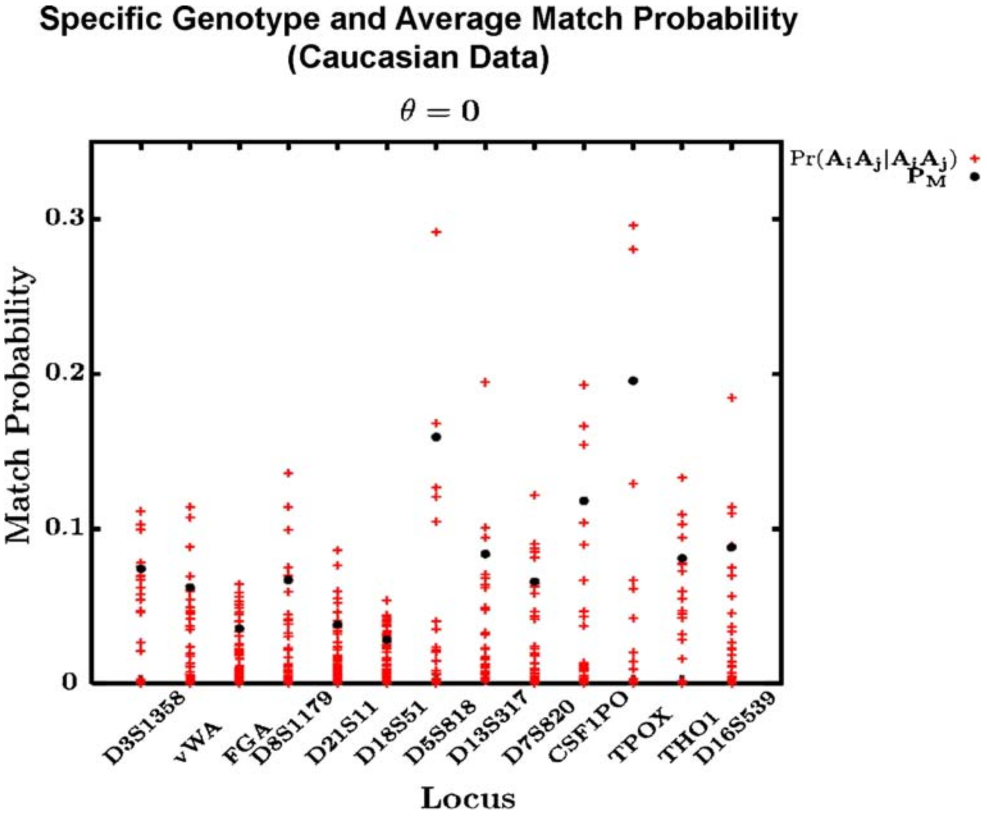
database; (2) There are 715 combinations of 9 loci out of the 13 CODIS STR loci used for DNA profiling in the database; (3) Using the statistics of the number of segregating alleles observed in the DNA forensic databases (e.g., [10]) and noting that with k segregating alleles at a locus, one can observe $k(k+1)/2$ possible distinct genotypes at that locus, the range of possible multilocus genotypes can be computed from Table 1; and (4) The RMP (p) values for 122 matching 9 locus pairs are between 1.523×10^{-12} and 6.769×10^{-14} (the same logic can be used for computing 10 or more locus matches [data not shown]).

These RMP values are typically smaller (i.e., rarer) than those reported for target profiles in casework using the modified product rule (as per recommendations in [3]). The above computations clearly indicate that the number of partial matches in pairwise comparisons of DNA profiles in the Arizona offender database, even when they are looked at without the inherent complicacies of the databases, does not, in general, provide any basis for questioning the legitimacy of computations of RMP values of any specific target profile based on the modified product rule that are currently followed in the DNA forensic community.

Given the three levels (minimum allele frequency, θ adjustment, and tenfold rule) of conservatism built into the current calculations used by the forensic community [3, 11], the predicted number of partial matches based on forensic calculations would be larger that that observed. Even with the large size of the offender database, many of the rare DNA profiles (out of all possible DNA profiles) would not be observed in the database. This makes observing pairs of partial matches appear smaller than expected (based on an average allele frequency).

Using θ adjusted calculations with reasonable θ values more than compensate for the degree of substructuring such as

Fig. 2 Specific genotype and average match probability using Caucasian population data [8]



that encountered in heterogeneous offender databases (note that reference databases for population inferences are not so heterogeneous) [5]. It is important to reemphasize that the criticisms that have arisen in US legal proceedings are not based on analyses such as those carried out by Weir [5]. The θ adjustment approach is ignored and instead tests demonstrating a violation of independence are sought.

Another point to consider is that computations of RMP in the context of specific forensic casework refer to a specific target profile. In contrast, pairwise comparisons of profiles in a database yield statistics of matches and partial matches with regard to any of the possible profiles, a concept that can be related to what may be termed as average match probability. For each of the loci used in current platforms of DNA forensic work, genotype-specific match probability can differ drastically from such average match probability. When summary statistics are used from allele/genotype sharing from pairwise comparisons of n profiles in the database, the summary statistic derived from $n(n-1)/2$ comparisons may approximate the estimate of true averages, if n is sufficiently large. However, if a small number of these pairs of subjects are related (for each of which one needs to invoke kinship adjustment over and beyond θ adjustment), observations on allele–genotype sharing for these pairs would be sensitive to their specific profiles differing substantially from the average which would produce discordances that are irreconcilable without knowing their exact DNA profiles. This is illustrated by Fig. 2 where using the Caucasian allele frequencies (extracted from [10]), the genotype-specific match probabilities were computed for all possible genotypes at each of the 13 loci (represented by the plus sign, and the average locus-specific match probability represented by dots). These show that a target-specific match probability can be drastically different from the average match probability. Although the computations of this figure were done using the HWE of genotype frequencies, the results are qualitatively the same even after adjustments for population substructure effect (i.e., with θ adjustment).

Conclusions

The CODIS database is an excellent database for investigative leads; it is an extremely poor database to analyze for inferences regarding the assumptions of independence. The profiles in CODIS do not lend themselves to good quality population studies; they are not properly annotated and duplicates and relatives reside in the database. If one takes a very simplistic view about the complexity of the heterogeneous offender database and then observes partial matches inconsistent with the assumption of independence, he/she provides no basis to invalidate current forensic practices. Such departures are expected and do not reflect the population statistics databases used by the forensic community for routine statistical calculations.

Extremely important are that concerns exist and will arise about the privacy and confidentiality of data retrieved from matches found during a pairwise comparison of offender DNA profiles. The names of individuals with matching and partial matching profiles would have to be disclosed to scientists and police when there is no criminal investigation underway. The names would be obtained because of a "research experiment." To further annotate such data may not be possible. However, not having annotated data for population studies does not compromise CODIS for its primary purpose that is developing investigative leads.

Acknowledgment This is publication number 07-03 of the Laboratory Division of the Federal Bureau of Investigation. The names of commercial manufacturers are provided for identification only and inclusion does not imply endorsement by the Federal Bureau of Investigation.

References

1. Budowle B, Moretti TR, Niezgoda SJ, Brown BL (1998) CODIS and PCR-based short tandem repeat loci: law enforcement tools In: Second European Symposium on Human Identification 1998, Promega Corporation, Madison, Wisconsin, pp 73–88
2. Martin PD (2004) National DNA databases: practice and practicability. A forum for discussion. Prog Forensic Genet 10:1–8
3. National Research Council II Report (1996) The evaluation of forensic evidence. National Academy Press, Washington, DC
4. The People of the State of Illinois v Juan Luna, In The Circuit Court Of Cook County, Illinois, Criminal Division, No. 02 CR 15430, 2006
5. Weir BS (2004) Matching and partially-matching DNA profiles. J Forensic Sci 49:1009–1014
6. Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (1999) The utility of STR loci beyond human identification: Implications for the development of new DNA typing systems. Electrophoresis 20:1682–1696
7. Shields WM (1992) Problems and solutions associated with matching and generating inclusion probabilities. In: Proceedings of The Third International Symposium on Human Identification, Promega Corporation, Madison, Wisconsin, pp 1–50
8. Troyer K, Kilboy T, Koeneman B (2001) A nine STR locus match between two apparently unrelated individuals using Ampflstr® Profiler Plus™ and Cofiler™. In: Proceedings of the Twelfth International Symposium on Human Identification, Promega Corporation. Available at http://www.promega.com/geneticidproc/ussymp12proc/abstracts.htm
9. Feller W (1968) An introduction to probability theory and its applications, vol. 1. 3rd edn. Wiley, New York, p 33
10. Budowle B, Shea B, Niezgoda S, Chakraborty R (2001) CODIS STR loci data from 41 sample populations. J Forensic Sci 46:453–489
11. Chakraborty R, Lee HS, Budowle B (2004) Response to Krane et al. J Forensic Sci 49:1390–1393