DNAtools : Tools for empirical testing of DNA match probabilities

James M. Curran
University of Auckland

Torben Tvedebrink
Aalborg University

Abstract

DNA evidence is the pre-eminent tool in the modern forensic scientists toolbox. It is widely accepted by the public, scientific and legal communities and it has been instrumental in determining both the innocence and guilt of individuals involved in the legal process. Despite this widespread acceptance there is unease regarding the statistical measures used to evaluate DNA evidence amongst some of members of all these communities. In particular, some people regard the random match probabilities associated with DNA evidence as just too small or basically unsupportable. In this article we discuss what it means for a pair of DNA profiles to match or partially match, and we present an R package that allows a rational examination of the statistical properties of a DNA database.

1 Introduction

In 2001, a poster was presented by a forensic scientist from Arizona (Troyer *et al.*, 2001) at a scientific meeting on human identification. This poster reported a nine locus match between two unrelated men, one white and one black (Kaye, 2009). It was not a full match. Both men had been typed at thirteen loci in total, and "partially matched" at three of the remaining four loci. These partially or non-matching loci would have excluded either man as a suspect if the other was the true offender. However, such a match seemed to be at odds with the random match probabilities. On one hand, these two men were in a DNA database which consisted of approximately 65,000 profiles, and on the other hand, the random match probabilities for the nine locus genotype were "1 in 754 million in Caucasians, 1 in 561 billion in African Americans, and 1 in 113 trillion in Southwest Hispanics", Troyer *et al.* (2001).

As we will show later this is, in effect, an example of the "birthday problem" and therefore is regarded as completely predictable from a statistical perspective. However, most us who have taught a class on the birthday problem know that our students are initially sceptical.

1.1 DNA evidence, matches and partial matches

Forensic genetics has its terminology which we briefly explain here. Human DNA consists of 23 pairs of chromosomes and those chromosomes are composed of a sequence of nucleotides which are labelled A, G, C and T after the bases adenine, guanine, cytosine and thymine that are used to form them. Modern DNA typing uses short tandem repeats (STRs). These are regions of DNA which are highly variable, but are patterned in that they consist of repeats of a short sequence of DNA bases. The locations at which this information is collected are called loci, and the (length) variations in the patterns observed at each locus are called alleles.

We have two alleles at each locus, because humans are a diploid species, meaning they have two copies of each chromosome. One allele comes from our mother, and the other from our father. A pair of alleles at a locus is called a genotype, and therefore a DNA profile is actually a multi-locus genotype. Modern forensic laboratories genotype DNA evidence using commercial kits, called multiplexes which consist of 9–17 loci. The multiplex currently used in the United Kingdom (and until recently New Zealand and Denmark) is called AmpF*l*STR® SGM Plus™, or SGM Plus for short, and consists of 10 loci, plus one sex specific locus, Amelogenin. Forensic laboratories in the United States which load profiles into the FBI's Combined DNA Index System (CODIS) collect a core set of thirteen loci, although they are not constrained to use one multiplex.

Locus	vWA	D18	TH01	D2	D8	D3	FGA	D16	D21	D19
Alleles	15,18	14,17	6,9.3	17,23	12,15	15,15	19,23	11,12	28,28	13,14

Table 1: A DNA profile from the AmpF*l*STR® SGM Plus™ multiplex

Table 1 shows a DNA profile from the AmpF*l*STR® SGM Plus™ multiplex. There are two numbers at each locus representing the two alleles that make up the genotype at that locus. The numbers relate to the number of times the pattern or motif that describe the alleles at the locus are repeated. For example, this person's genotype at the locus TH01 is 6,9.3. This means that on one chromosome, the motif for TH01, `TCAT` was repeated 6 times, and on the other chromosome it was repeated 9 times, and then followed by `TCA`. The .3 represents the fact that three of the four bases have been repeated.

A pair of profiles is said to (fully) match if every allele at every locus that occurs in one profile occurs in the other. A pair of profiles are said to partially match if there are allelic matches at a subset of loci. Weir (2004) provided a taxonomy for describing partial matches which depends on the number of fully, partial and non-matching loci between a pair of profiles. For any given pair of (full) profiles from the same multiplex there will be: m_2 loci where both alleles match, m_1 loci where only one of the alleles matches, and m_0 loci where none of the alleles match. For example, the profile that Troyer *et al.* (2001) found was a 9/3/1 partial match - nine fully matching loci, three partially matching loci, and one non-matching locus.

1.2 DNA database comparison exercises

The Troyer match came from a database matching exercise. In such an exercise every profile is compared with every other profile in the database. This type of comparison exercise is absolutely essential and, in addition, can provide some interesting information about the statistical properties of the population under consideration. We say that database comparison is essential in the first instance for the detection of duplicates. Duplicates may arise in a number of different ways. For example, an offender may provide a false name or an offender's name may be entered incorrectly. Alternatively, an offender may have an identical twin who is already in the DNA database. There are six pairs of identical twins in the New Zealand National DNA Database (NZDNADB). Forensic scientists are also interested in 'very close' matches. For example, a pair of profiles might fully match at nine loci out of ten and partially match at the remaining locus. This may happen either because the donors of the samples are very close relatives. It is more likely, however, that the profiles do not match because of allelic dropout, primer binding site mutations, nomenclature changes or somatic mutation.

1.3 The birthday problem

Weir (2007) and others (Brenner, 2007; Curran *et al.*, 2007; Mueller, 2008; Kaye, 2009) note that the presence of matching profiles in a DNA database is effectively an instance of the well-known "birthday problem" (Wikipedia, 2010) where, in a group of at least 23 randomly chosen people, there is a greater than 50% chance that one pair of them will have the same birthday. Early critics, implicitly calculating the expected number of matches as Np, used the wrong value for N and the wrong value for p. Firstly, the number of pairwise matches, not the size of the database, is the relevant quantity. Although the database size is relatively small, the number of pairwise comparisons is very large. The Arizona database contained of $N = 65,493$ profiles (Brenner, 2007). Therefore, there are

$$N_{\text{Comparisons}} = \frac{N(N-1)}{2} = 2,144,633,778$$

or approximately two billion, possible pairwise comparisons. Secondly, the random match probability is not the probability we need. The random match probability for the pair of profiles in question answers the question "What is the probability that someone other than these two men would have this particular nine locus profile." The probability we actually want is "What is the probability that two randomly selected profiles would match at nine loci, partially match at three loci, and not match at one locus.". Weir (2004), working on an unrelated case, showed that this probability can be calculated by

$$P_{m_0,m_1,m_2}(\theta) = \sum_{m_{l0},m_{l1},m_{l2}} \prod_l P_{l2}(\theta)^{m_{l2}} P_{l1}(\theta)^{m_{l1}} P_{l0}(\theta)^{m_{l0}} \tag{1}$$

where m_{l0}, m_{1l} and m_{l2} are indicator variables that are equal to one if the individuals share zero, one or two alleles in common respectively and zero otherwise. The expressions $P_{li}(\theta)$ are the probability of sharing $i = 0, 1, 2$ alleles in common at locus l for a given degree of population substructure θ and are given explicitly in Weir (2004, 2007). The coancestry coefficient, θ or F_{ST} models low levels of relatedness between individuals in the same subpopulation, and is typically between 0 and 0.03.

1.4 Modelling the observed data

Weir's original paper (Weir, 2004) contained an informal analysis where the minimum level of θ required to explain the observed counts was calculated. For example using the FBI Caucasian data (Budowle and Moretti, 1999) a θ value of 0.005 is needed to explain the 679 observed one locus matches (at locus FGA). That is, if $\theta > 0.005$, then the expected count at this locus will exceed the observed count. Curran *et al.* (2007) formalised and extended this analysis in the following way. We model the expected number of pairs of profile which fully match at m_2 loci and partially match at m_1 loci for a given value of θ, $E_{m_2/m_1}(\theta)$, as

$$E_{m_2/m_1}(\theta) = \alpha E^U_{m_2/m_1}(\theta) + \beta E^B_{m_2/m_1}(\theta) + \delta E^C_{m_2/m_1}(\theta) + \gamma E^P_{m_2/m_1}(\theta)$$

where $0 \leq \alpha, \beta, \delta \leq 1$ and $\gamma = 1 - \alpha - \beta - \delta$. The quantities $E^R_{m_2/m_1}(\theta),\ R \in \{U, B, C, P\}$ are the expected number of matching pairs of profiles calculated under four relationship categories: unrelated, full siblings (brothers), cousins, and parent/child. These expressions are derived in Curran *et al.* (2007), with a typographical mistake corrected in Curran and Buckleton (2010).

Tvedebrink (2010) also derived expressions for avuncular relationships. Curran *et al.* (2007) estimated $\alpha, \beta, \delta, \gamma$ and θ by using a combination of a line search (across θ) and a Monte Carlo steepest descent method to find the values that minimised several different distance metrics applied to the observed and expected value. Curran *et al.* (2007) recommended minimising

$$C_3(\theta) = \sum_{i=0}^{L} \sum_{j=0}^{L-m_2} \frac{|E_{i/j}(\theta) - O_{i/j}|}{O_{i/j}}$$

where $O_{i/j}$ is the observed number of pairs of profiles fully matching at $m_2 = i$ and $m_2 = j$ loci using a multiplex consisting of L loci. This metric was chosen because of the belief that it puts emphasis on "explaining" the higher order matches. Further research into this by Tvedebrink (2010); Tvedebrink *et al.* (2012) has shown that Mahalanobis distance

$$T_2(\theta) = \left(\vec{E}(\theta) - \vec{O}\right)^\top \Sigma(\theta)^- \left(\vec{E}(\theta) - \vec{O}\right)$$

actually provides a "better" fit to the data and does not drive the value of θ to zero. $\Sigma(\theta)^-$ is a pseudo inverse because of constraint

$$\sum_{m_2=0}^{L} \sum_{m_1=0}^{L-m_2} P_{m_2,m_1,m_0} = 1$$

2 The DNAtools package

The aim of the **DNAtools** package is to provide statisticians and forensic scientists with access to the procedures described in the previous sections. Early implementations by Weir (2004) and then Curran *et al.* (2007) required custom written code for each new database and, in the case of Curran *et al.* (2007), generation of at least half a dozen precursor files and a significant amount of memory. Tvedebrink (2010); Tvedebrink *et al.* (2012) reduced the computational effort of Weir (2004) and Curran *et al.* (2007) by deriving recursion formulas for Equation 1, improved the optimisation procedures through the use of the package **Rsolnp** (Ghalanos and Theussl, 2012), and derived the variances of the probabilities which allowed both the computation of Mahalanobis distances and asymptotic confidence intervals. **DNAtools** aims to make all of these procedures easier to use in R (R Development Core Team, 2010).

3 Using the package DNAtools

The expected data format of the databases used as input for the functions in **DNAtools** is a data frame, which is constituted by a column of DNA profile identifiers (the first column) and two columns per typed DNA marker. An example is given below:

```
data(dbExample)
head(dbExample)[,1:9]
id D16S539.1 D16S539.2 D18S51.1 D18S51.2 D19S433.1 D19S433.2 D21S11.1 D21S11.2
 1        11        11       15       21        14        14       28       29
 2        13        12       15       14        16        16       29       28
```

```
 3          9          9         13         17         14         14         28         27
 4         11         12         14         15         15         13         32         29
 5         12         12         17         12       15.2         13       31.2         28
 6          9         13         17         14         13         14       30.2         28
```

Budowle and Moretti (1999) published data from six US subpopulations of different ethnicity (Caucasians, Hispanics, African Americans, Bahamians, Jamaican and Trinidad). We demonstrate here our package **DNAtools** using the Caucasian profiles typed at nine forensic STR markers.

```
(caucasian.summary <- dbCompare(caucasian,hit=5))
```

```
Summary matrix
     partial
match    0    1    2    3    4    5    6    7    8    9
    0   17  145  628 1531 2416 2516 1822  752  170   26
    1   28  178  733 1426 1902 1455  727  211   40
    2   13  121  303  530  492  310  108    9
    3    5   32   64   99   52   23    6
    4    0    6    6    7    2    1
    5    0    1    0    1    1
    6    0    0    0    0
    7    0    0    0
    8    0    0
    9    0

Profiles with at least 5 matching loci
  ID1 ID2 match partial
1  10  29     5       4
2  77 116     5       3
3  64 170     5       1
```

There is a `plot` method for the returned object. Applying this method to `caucasian.summary` yields the "dropping ball"-picture of Figure 1. The right end of the "distribution" is interesting part, due to the larger number of coinciding loci between profile pairs.

In Table 2, the estimated parameters are reported using the different object functions implemented in the `optim.relatedness`-function of the **DNAtools**-package. Only for T_2 the estimate of θ is different from 0.

The fitted values can be used to compute $E_{m_2/m_1}(\theta)$. This is done using `dbExpect` which takes θ and a list of locus specific probability vectors as input. The function efficiently computes the expectation using a recursion relation (Tvedebrink *et al.*, 2012). Similarly, the superimposed confidence intervals is Figure 1 were computed using the `dbVariance`-function, which computes the covariance matrix of the summary statistic (also by recursion over loci Tvedebrink *et al.*, 2012). The confidence intervals are based on a normal approximation such that the width of the interval around the expectations is computed as $\pm 2\sqrt{\text{diag}\{\Sigma(\theta)\}}$. This approximation is asymptotic, hence the coverage accuracy decreases with the (expected) cell count.

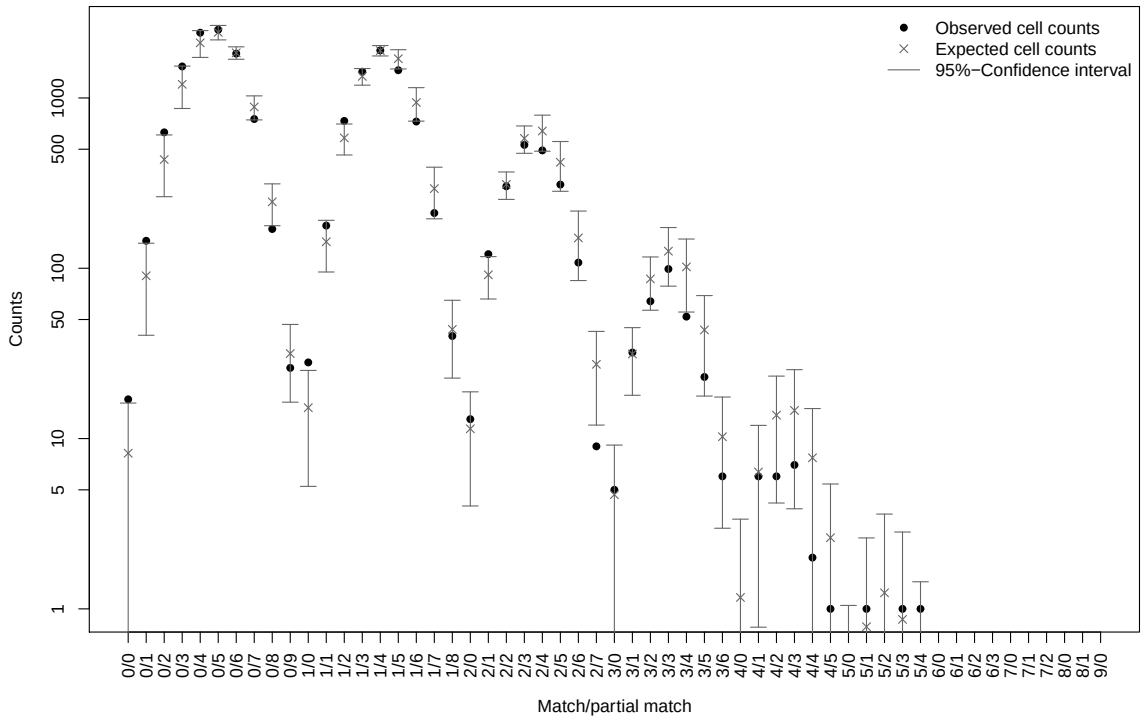

Figure 1: Plot produced by `plot(caucasian.summary,pch=16)`. Superimposed are the expected counts and associated 95%-confidence intervals. The labels of the first axis denote the number of matching and partial-matching loci. The second axis is on a $\log_{10}$-scale.

Availability The R package **DNAtools** is available at CRAN: **DNAtools**

4 Conclusion

In this paper we have described an R package which allows statisticians and forensic scientists to easily examine the properties of a forensic DNA database. In particular, our package makes it simple to carry out a database comparison exercise where every DNA profile in the database is compared to every other database, and compare the resulting numbers of observed pairs of matching and partially matching profiles to expectation under a set of population genetic assumptions. There are potential limitations on the use of this package in that it may not scale well to extraordinarily large databases (>100,000 profiles), but we expect that this will

	θ	Unrelated	First-Cousins	Avuncular	Parent-child	Full-siblings
C_1	0	9.99e-01	2.32e-13	1.15e-08	5.01e-04	1.57e-04
C_2	0	9.99e-01	6.90e-08	3.99e-08	1.16e-08	3.68e-14
C_3	0	9.99e-01	2.16e-08	1.01e-11	1.08e-04	7.03e-04
T_1	0	9.99e-01	3.97e-09	5.64e-12	4.63e-04	1.58e-05
T_2	0.015	9.99e-01	1.67e-10	4.60e-06	7.47e-04	6.15e-06

Table 2: The estimated parameters of the model for the Caucasian subsample.

be remedied by further development.

Acknowledgements

We wish to acknowledge Professor Chris Triggs and the Department of Statistics at the University of Auckland for financial support for accommodation and travel for T. Tvedebrink. We also wish to thank Dr. John Buckleton for initially introducing us to this problem. T. Tvedebrink would like to thank Poul Svante Eriksen, Aalborg University, for his contributions to the recursion formulas for computing expectations and covariance matrices.

References

Brenner C (2007). "Arizona DNA Database Matches." Accessed 4-January-2010, URL http://dna-view.com/ArizonaMatch.htm.

Budowle B, Moretti TR (1999). "Genotype Profiles for Six Population Groups at the 13 CODIS Short Tandem Repeat Core Loci and Other PCR-Based Loci." *Forensic Science Communications*, **1**(2). Accessed 5-January-2010, URL http://www2.fbi.gov/hq/lab/fsc/backissu/july1999/budowle.htm.

Curran JM, Buckleton JS (2010). "Re: Sign mistake in allele sharing probability formulae of Curran, et al." *Forensic Science International: Genetics*, **4**(3), 215–217.

Curran JM, Walsh SJ, Buckleton J (2007). "Empirical testing of estimated DNA frequencies." *Forensic Science International: Genetics*, **1**(3-4), 267–272. ISSN 1878-0326. URL http://www.ncbi.nlm.nih.gov/pubmed/19083772.

Ghalanos A, Theussl S (2012). ***Rsolnp**: General Non-linear Optimization*. R package version 1.14, URL http://CRAN.R-project.org/package=Rsolnp.

Kaye DH (2009). "Trawling DNA Databases for Partial Matches: What Is the FBI Afraid Of?" *Cornell Journal of Law and Public Policy*, **19**(1).

Mueller LD (2008). "Can simple population genetic models reconcile partial match frequencies observed in large forensic databases?" *Journal of Genetics*, **87**(2), 101–108. ISSN 0022-1333. URL http://www.ncbi.nlm.nih.gov/pubmed/18776637.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Troyer K, Gilboy T, Koeneman B (2001). "A nine STR locus match between two apparently unrelated individuals using AmpF*l*STR® Profiler Plus™ and Cofiler™." In *Genetic Identity Conference Proceedings*, 12th International Symposium on Human Identification. Accessed 4-January-2010, URL http://www.promega.com/geneticidproc/ussymp12proc/abstracts/troyer.pdf.

Tvedebrink T (2010). *Statistical Aspects of Forensic Genetics – Models for Qualitative and Quantitative STR Data*. Ph.D. thesis, Department of Mathematical Sciences, Aalborg University.

Tvedebrink T, Eriksen PS, Curran JM, Mogensen HS, Morling N (2012). "Analysis of matches and partial-matches in a Danish DNA reference profile data set." *Forensic Science International: Genetics* **6**(3), 387–392.

Weir BS (2004). "Matching and partially-matching DNA profiles." *Journal of Forensic Sciences*, **49**(5), 1009–1014. ISSN 0022-1198. URL http://www.ncbi.nlm.nih.gov/pubmed/15461102.

Weir BS (2007). "The rarity of DNA Profiles." *The Annals of Applied Statistics*, **1**(2), 358–370. ISSN 1941-7330. URL http://www.ncbi.nlm.nih.gov/pubmed/19030117.

Wikipedia (2010). "Birthday problem." Accessed 5-January-2010, URL http://en.wikipedia.org/wiki/Birthday_problem.