The Journal of Open Source Software

DNAtools: Tools for Analysing Forensic Genetic DNA Data

Torben Tvedebrink[1], **Mikkel Meyer Andersen**[1], **and James Michael Curran**[2]

1 Department of Mathematical Sciences, Aalborg University, Denmark **2** Department of Statistics, University of Auckland, New Zealand

DOI: 10.21105/joss.01981

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @standage
- @tomsing1

Submitted: 19 December 2019
Published: 16 January 2020

Summary

DNA evidence is the pre-eminent tool in the modern forensic scientist's toolbox. It is widely accepted by the public as well as in the 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 members of these communities.

The prevailing technology in forensic genetics is that of capillary electrophoresis (Butler, 2010), which measures the lengths of so-called *short tandem repeat* regions on the DNA (STR markers). One quantity of interest is the random match probability. The random match probability is defined as the probability that a randomly chosen individual has *a specific* DNA profile (G_C), given that we know that at least one other (usually person S of interest with genotype G_S) has this profile (Steele & Balding, 2015). We write this as

$$\Pr(G_C \mid G_S \equiv G_C).$$

Estimates of this probability become very small (in the order of 10^{-20}) as the number of STR markers increases. Some people regard the small random match probabilities associated with DNA evidence as just too small or basically unsupportable.

In 2001, Karen Troyer and others (Troyer, Gilboy, & Koeneman, 2001) published a poster reporting the results of a *database matching* exercise. In such an exercise, every profile is compared to every other profile and the number of loci where the two profiles match is recorded. This is a very useful exercise, because, amongst other things it helps laboratories detect potentially erroneous entries in their databases. The Arizona laboratory, where Troyer worked, used the CODIS set of loci, which was a standardized set of 13 STR markers used in many jurisdictions across the US, including federally by the FBI. Troyer et al.'s poster (Troyer et al., 2001) reported that a 9 locus match had been found between two apparently unrelated individuals. This information was seized upon by an enterprising defence lawyer, because at first glance, it seemed to cast doubt on extremely small match probabilities. That is, how could two unrelated individuals, in a database of 65,000 people, have the same (partial) profile, when the probability of this profile was at most 1 in 754 million (7.54×10^{-8})? This issue is nicely summarized by Charles Brenner on his webpage "Arizona DNA Database Matches".

Weir (Weir, 2004, 2007) and others pointed out that this degree of matching is not surprising when one takes into account the total number of comparisons being made (about 4 billion in the Arizona case), and recognize that it is not the probability of a specific profile that is of interest, but rather the probability that *any* two loci would match at 9 loci. Hence, one has to use the correct probabilities and also account for the fact the number of comparisons to be made between all pairs of profiles for a database of size N is $N(N+1)/2$. The `DNAtools`

package implements the methodology of Tvedebrink, Eriksen, Curran, Mogensen, & Morling (2012) for efficient computations of the expectation and variance of the number of matches. To our knowledge, `DNAtools` is the only software that can perform such computations.

The analysis of mixed DNA traces has proven to be one of the most challenging tasks in forensic genetics. DNA mixtures, as they are referred to, are observed biological traces which are comprised of biological material from two or more individuals. Assessing the number of contributors to a DNA mixture is difficult. One indicator is the number of distinct alleles in the stain – the more alleles the more contributors. `DNAtools` implements the expression of Tvedebrink (2013), where the distribution of the number of distinct alleles can be computed, while accounting for subpopulation effects by the θ-correction (Tvedebrink, 2013). Equally, researchers looking at the efficacy of new multiplexes (in this context this is mainly about the number and frequencies of alleles in newly included loci) are interested in understanding the probability that a mixture which truly consists of n individuals appears to consist of $n-1$ individuals. This might happen, for example, when a two person mixture shows no more than two alleles per locus at every locus in a multiplex. `DNAtools` allows the rapid, and exact, computation of such probabilities for any number of individuals. To our knowledge, `DNAtools` is the only software that can perform such computations.

The documentation of `DNAtools` consists of manual pages for the various available functions, articles describing how to perform contiguous analyses (*vignettes*), and unit tests.

References

Butler, J. M. (2010). *Fundamentals of Forensic DNA Typing* (1st ed.). Academic Press. doi:10.1016/C2009-0-01945-X

Steele, C., & Balding, D. (2015). *Weight of evidence for forensic DNA profiles* (2nd ed.). Wiley. doi:10.1002/9780470867693

Troyer, K., Gilboy, T., & Koeneman, B. (2001). A nine STR locus match between two apparently unrelated individuals using AmFISTR Profiler Plus and Cofiler. In Y. Berbers & W. Zwaenepoel (Eds.), *Proceedings of the 12th international symposium on human identification*.

Tvedebrink, T. (2013). On the exact distribution of the numbers of alleles in DNA mixtures. *Forensic Science International: Genetics Supplement Series*, *4*(1), e278–e279. doi:10.1016/j.fsigss.2013.10.142

Tvedebrink, T., Eriksen, P. S., Curran, J. M., Mogensen, H. S., & Morling, N. (2012). Analysis of matches and partial-matches in a Danish STR data set. *Forensic Science International: Genetics*, *6*(3), 387–392. doi:10.1016/j.fsigen.2011.08.001

Weir, B. (2004). Matching and partially-matching DNA profiles. *Journal of Forensic Sciences*, *49*(5). doi:10.1520/jfs2003039

Weir, B. (2007). The rarity of DNA profiles. *The Annals of Applied Statistics*, *1*(2), 358–370. doi:10.1214/07-AOAS128