Int J Leg Med (1993) 106:103–106

International Journal of
Legal Medicine

Short communications

On allele frequency computation from DNA typing data

Ranajit Chakraborty[1], Li Jin[1], Yixi Zhong[1], M. R. Srinivasan[1, 2], and Bruce Budowle[3]

[1] Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, Texas 77225, USA
[2] Present address: Department of Statistics, Madras University, Madras, India
[3] Forensic Science Research and Training Center, FBI Academy, Quantico, Virginia, USA

Received February 17, 1993 / Received in revised form May 10, 1993

Summary. Forensic applications of DNA typing data require the estimation of the frequencies of all observed alleles, which is currently done by a fixed set of groupings (binning) of alleles in a database. Recently it's validity has been questioned on the ground that when a DNA fragment size is close to a bin boundary, the frequencies of all adjacent bins should be added. On the contrary, the current forensic database indicates that when the match window of a DNA fragment overlaps 2 bins, it is enough to consider the bin with the larger frequency, and this *never underestimates* the frequency within the match interval with the current choice of fixed-bin widths. On average, the current fixed-bin procedure yields an allele frequency at least 2-fold higher than that of a floating-bin.

Key words: Allele frequency computation – Fixed bin – Floating bin

Zusammenfassung. Forensische Anwendungen von DNA-Befunden erfordern die Bestimmung der Häufigkeiten aller beobachteten Allele. Dies erfolgt gegenwärtig durch einen fixierten Satz von Gruppierungen (binning) von Allelen in einer Datenbank. Kürzlich wurde die Geeignetheit dieses Ansatzes in Frage gestellt mit der Begründung, daß wenn eine Fragmentgröße nahe bei einer bin-Grenze ist, die Frequenzen aller angrenzenden bins addiert werden sollten. Im Gegensatz hierzu weist die gegenwärtige forensische Datenbank darauf hin, daß bei Überlappung von 2 bins durch einen sog. match window es ausreicht, das bin mit der höheren Frequenz zu berücksichtigen und daß dies nie zu einer Unterschätzung der Frequenz innerhalb des match-Intervalls führt, unter Berücksichtigung der gegenwärtigen Auswahl der Weiten der „fixed-bins". Im Durchschnitt führt das gegenwärtige „fixed-bin"-Verfahren zu einer Allelfreqenz, welche mindestens 2fach höher ist als jene eines „gleitenden bin".

Correspondence to: R. Chakraborty

Schlüsselwörter: Allelfrequenzberechnung – fixierte Allelgruppierungen – gleitende Allelgruppierungen

Introduction

The recent report on DNA Technology in Forensic Science [1] acknowledges that the current technique of restriction fragment length polymorphism (RFLP)-based protocols of DNA typing, when appropriately conducted, provides a scientifically reliable method of DNA profiling of forensic samples. However, without any data to support this contention, the report questions only one aspect of the fixed-bin method [2] of determining allele frequencies from DNA typing databases. According to the current practice of the use of the fixed bin method [3] for a given allele in a forensic case, when the window specified by the laboratory's forensic quantitative matching rule overlaps two adjacent bins, the larger bin frequency is taken as the relevant allele frequency for DNA-profile frequency computations [2]. However, without any justification, the NRC report [1] recommends that "All bin frequencies must be added; it is not enough to take the largest bin frequencies". This suggestion is inappropriate and unnecessary. Using the current forensic databases, we show that the fixed bin approach, without any modification, provides more conservative (larger) estimates of allele frequencies than what would be obtained from the quantitative matching rule, even when the alleles observed in forensic cases are close to the fixed bin boundaries.

Fixed bin versus floating bin

The purpose of binning (grouping) of DNA fragment sizes is to take into account measurement errors when sizing DNA fragments obtained by RFLP-based protocols. Comparisons of differences in size measurements

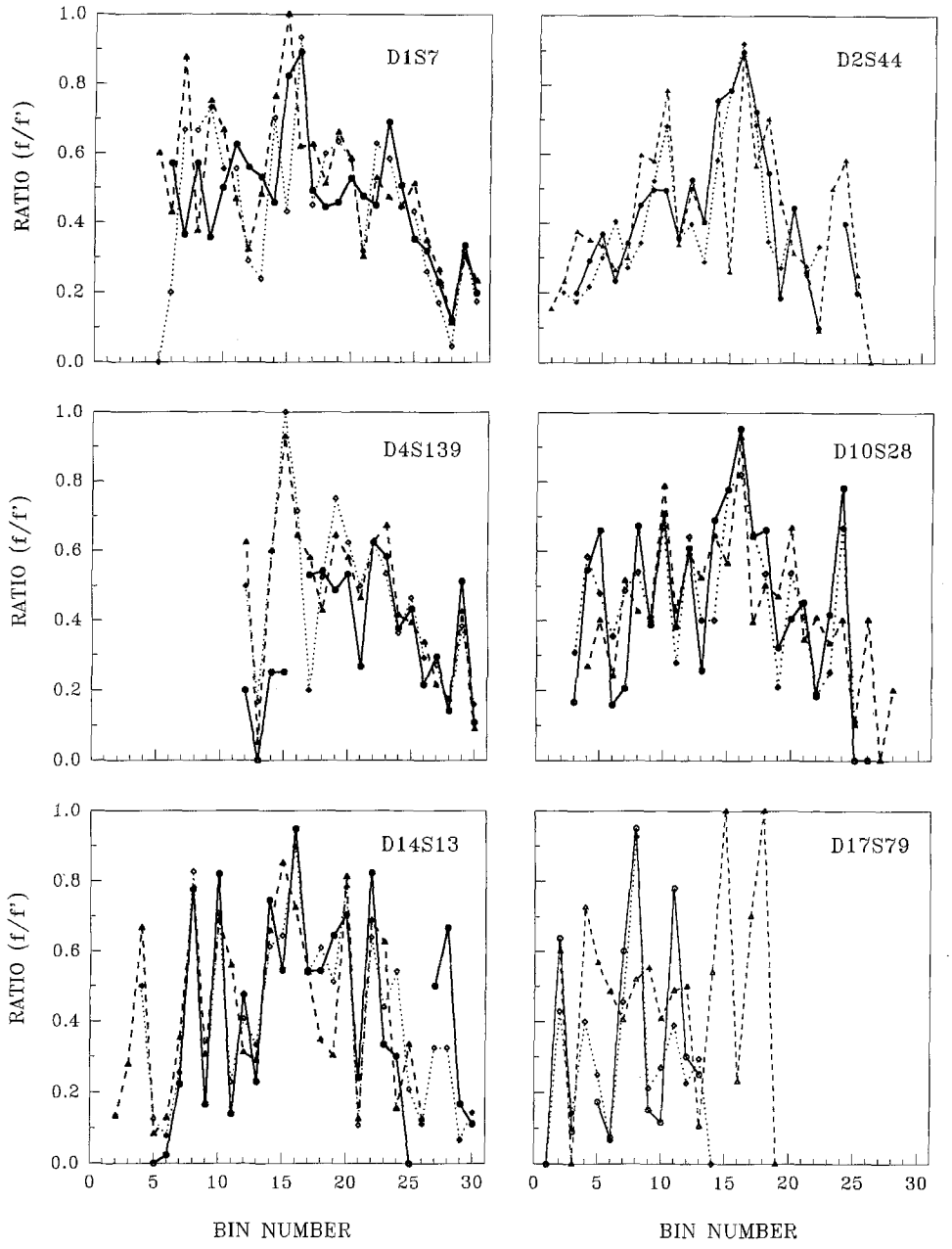

Fig. 1. Ratio of ±2.5% floating bin (f) and fixed bin (f') frequencies of DNA fragment sizes in the current forensic database for any DNA fragment size whose ±2.5% matching window crosses the boundaries of any of the 31 bins. Each panel represents the binned data for the locus indicated on the panel. For each panel, the lines joining solid circles (●) represent the data for Caucasians, those joining the solid triangles (▲) are for the Black data, and the ones joining the solid diamonds (◆) represent the composite Hispanic population, as defined in the data presented in ref. [3]

of DNA fragments from victim's blood and vaginal epithelial cells indicate that the standard deviation of size differences between samples from the same source is approximately 1.5% of the size of the DNA fragment being measured (computed from data given in Table 3 of ref. [2]). From this, and other visual characteristics of DNA fragment sizes from the same source, the quantitative match criterion of ±2.5% was adopted [2]. In comparison, the fixed bin widths practiced in the FBI protocol for forensic applications are between ±3.0% to ±9.4% of the mid-points for each of the 31 bins [2]. Therefore, in any sample analysis when the ±2.5% window around an observed fragment size resides completely within any single fixed bin, the fixed bin allele frequencies are always conservative.

When the matching window for any fragment size overlaps adjacent bins, current protocol (of selecting the bin with the larger frequency) may be questioned only if the ±2.5% window contains more alleles or chromosomal counts than that predicted by the fixed bin procedure. To show that this never occurs in the current forensic database, we considered all 3 databases (Caucasian, Black, and composite Hispanic) of the FBI laboratory for 6 loci (D1S7, D2S44, D4S139, D10S28, and D17S79). For each locus-population combination, we consider a fragment size x encompassing the entire range of a band size distribution and open a window of $x \pm 0.025x$ (i.e., a ±2.5% floating bin). Let f denote the number of chromosomal counts in the interval, and n, the total number of chromosomes sampled. If this interval is completely within a bin (say, B_i), the frequency of fragments in bin B_i (say, f_i) is already at least as large as f. When the interval $x \pm 0.025x$ overlaps 2 bins (say, B_i and B_{i+1}), we define f' to be the maximum of f_i and f_{i+1}. If $f/f' \leq 1$

Table 1. Summary statistics of the ratio of ±2.5% floating bin and fixed bin allele frequencies (*f/f'*) in the current FBI forensic database of DNA typing. Only bins containing 5 or more fragments are used in this table, since for others the ratio is necessarily smaller than 1.0. (*n* = number of alleles in the sample, Min. = minimum, Max. = maximum, and Av. = average for the specific locus-population combinations across all bins containing 5 or more observations)

Locus	Population											
	Caucasian				Black				Hispanic			
	n	Min.	Max.	Av.	*n*	Min.	Max.	Av.	*n*	Min.	Max.	Av.
D1S7	1,190	0.121	0.892	0.474	718	0.113	1.000	0.509	1,042	0.000	0.935	0.447
D2S44	1,584	0.100	0.897	0.430	950	0.000	0.900	0.418	1,030	0.174	0.921	0.418
D4S139	1,188	0.000	0.625	0.352	896	0.049	0.929	0.467	1,044	0.144	1.000	0.464
D10S28	858	0.000	0.952	0.460	576	0.000	0.929	0.437	880	0.111	0.821	0.465
D14S13	1,502	0.000	0.947	0.427	1,048	0.083	0.851	0.434	988	0.067	0.897	0.413
D17S79	1,552	0.000	0.951	0.344	1,100	0.000	1.000[a]	0.490	1,042	0.000	0.926	0.290
Mean				0.415				0.459				0.416

[a] Two bins attain the maximum ratio (f/f') of 1.0 for this locus-population combination

for all choices of x (considering all interger values of x from 0 to the largest possible fragment size), the current fixed bin procedure of allele frequency determination would be conservative. Figure 1 shows the plot of the maximum ratio (*f/f'*) for all values of x whose ±2.5% match window cross each of the fixed bin boundaries. As seen from these computations, the ratio never exceeds 1.0. Only in 4 situations (bin boundary of 2,693 bp in the Blacks for D1S7 and D17S79, and in the Hispanics for D4S139, and that of 3,300 bp in the Blacks for D17S79) of the total of 540 ($= 30 \times 3 \times 6$) bin boundaries, does the ratio *f/f'* become 1.0. In other words, in the total sample of 19,186 DNA fragments of the current forensic database, in no situation does the current fixed bin allele frequency computation yield a frequency estimate smaller than that predicted by a ±2.5% floating bin, irrespective of the fragment size found in any forensic analysis. Even if a single case could be found in the current database where a floating bin frequency was lower than a fixed bin, it should be remembered that DNA profile frequencies are based on multiple alleles and multiple loci. Therefore, the conservatism of estimating frequencies (from fixed bins) of other alleles would more than compensate the effect of overestimation of any single allele frequency by the fixed bin procedure, even if any single departure from our observations had occurred.

Table 1 shows the summary statistics of the ratio (*f/f'*) of ±2.5% floating bin frequencies and fixed bin frequencies for all bins (excluding the ones which contain fewer than 5 observations). Firstly, these computations corroborate that no situation was encountered where the current fixed bin procedure of allele frequency computation yields allele frequencies that are smaller than the floating bin frequencies. Secondly, they show that, on average (last row of Table 1), the current practice gives allele frequencies that are more than 2-fold of the ±2.5% floating bin frequencies for each database. In other words, a DNA profile frequency showing k distinct fragments (across all probes examined) would provide a profile frequency that is inflated by a factor of more than 2^k, in the current fixed bin approach of allele frequency computations [5]. This permits an estimation of the average degree of conservatism of the frequency of any multiple locus DNA profile frequency using the fixed bin approach.

Discussion and conclusions

It is true that our results showing that the current fixed bin procedure [2] with bin widths ranging from ±3% to ±9.4% provide conservative estimates of allele frequencies in comparison to floating bins of widths ±2.5% cannot be generalized for any different match window. However, Budowle and Monson [4], in a separate analysis of the same data, showed that the fixed bin windows used above, are on average conservative, even when they are compared with sliding floating bins of ±5% width.

Two important conclusions emerge from these computations. The two statements of the NRC report [1] "the frequency of an allele in a laboratory's databank should be calculated by counting the number of alleles that would be regarded as a match with the laboratory's forensic matching rule" (p. 3–14), and "given an allele in a forensic case, one must then compute its frequency by adding the frequencies of all bins that contain any alleles that fall within the window specified by the laboratory's forensic matching rule" (p. 3–14) are contradictory to each other. Our computations show that this later suggestion, leading to "All bin frequencies must be added; it is not enough to take the largest of the bin frequencies", cannot be justified from the current forensic databases. Also, the fixed bin procedure of taking the largest frequency is inherently conservative, giving allele frequencies that are, on average, at least 2-fold of that predicted by the first suggestion of a floating bin computation.

Acknowledgements. This work was supported by US Public Service Grant 92-IJ-CX-K024 from the National Institute of Justice. The opinions expressed are those of the authors, and they do not con-

stitute any endorsement from the granting agencies. Comments and suggestions from 2 anonymous reviewers are greatly appreciated.

References

1. National Research Council (1992) DNA technology in forensic science. National Academy Press, Washington DC
2. Budowle B, Guisti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, Deadman HA, Monson KL (1991) Fixed-bin analysis for statistical evaluation of continuous distribution of allelic data from VNTR loci, for use in forensic comparisons. Am J Hum Genet 48:841–855
3. U.S. Department of Justice (1991) DNA analysis: a collection of articles 1988–1991, vol 15–18. Crime Laboratory Digest
4. Budowle B, Monson KL (1992) Perspective on the fixed bin method and the floor approach/ceiling principle. In: Proceedings from the Third International Symposium on Human Identification. Promega Corporation, Wisconsin, pp 391–406