The Robustness of Models for Evaluating Patterns of DNA Multi-Locus Probes

C.H. Brenner

2300 Grand Canal, Venice, California 90291, USA

INTRODUCTION

DNA multi-locus probe patterns clearly have the potential to be powerful evidence in clarifying issues of identity, paternity, or other questions of kinship. Unfortunately the evidence is difficult to quantify. Any computation rests on assumptions built into a mathematical model. Of the assumptions usually built into models of multi-locus profiles, some are improbable and some -- like independence -- are at best nearly impossible to justify.

Computer simulations and other arguments presented here suggest encouragingly that some of the difficult assumptions may not be essential assumptions.

In particular, in considering independence a distinction has to be made between haplotypes and phenotypes. Of course, there can be 100% independence among the band positions (phenotypic information) only if the underlying haplotypes are also independent. However, redundant information in the haplotypes is by no means reflected in equal measure as redundancy in the phenotypes.

ASSUMPTIONS

The following simplifying assumptions define what will be referred to as the "ideal" multi-locus genetic model:

A. Unambiguity:
   1. There is a discrete collection of possible band positions, rather than a continuum. Therefore one can always make an unequivocal decision as to match / no match, and a correct interpretation of possible double bands.
   2. The patterns are clear in that one can clearly differentiate between a band and a stray mark.
   3. Fragments always manifest bands.

B. Genetic simplicity:
   1. Each position represents a +- system (i.e. + dominates -).

Advances in Forensic Haemogenetics 4
Edited by Ch. Rittner and P. M. Schneider

2. There is no mutation.

C. Independence of bands:
   1. There is no observable allelism -- bands at two different positions won't come from the same locus.
   2. There is no linkage between loci -- neither a physical linkage nor statistical correlation.

D. Constant band frequency
   1. There is no overlap between fragment sizes from different loci.
   2. Every fragment size occurs with the same frequency.

On the basis of these assumptions formulae can be derived for the correct likelihood ratio (i.e. significance of the evidence), as Evett et al (1989) and Hummel et al (1990, "Biostatistical ...") have done in the case of using the multi-locus pattern of a trio for evaluation of paternity. Hummel's analysis includes an ingenious maximum likelihood method to estimate the parameters of band frequency, and number of potential band positions (N). In consequence however, the analysis relies more heavily on the idealized assumptions.

Formulae for evaluating identity can easily be derived.

Of course no one believes that the ideal model is a correct description of nature. However, that does not necessarily invalidate the calculations. They may still be approximately right, or they may be conservative. Some of the assumptions can be tested by computer simulations, or by explicit analysis.

Constant Band Frequency

For evaluating identity, this is a conservative assumption as can be shown by calculating a simple example. There is one hidden point, however: there can be a difference between average band frequency and average band-sharing frequency.

Suppose there are only two band positions, G and H, that have different band frequencies, 1/2 and 1/4 respectively. If both bands appear in a crime stain, a random suspect will match only one time in eight, so the correct likelihood ratio is 8.

The likelihood ratio computed by assuming the idealized model will depend on how a value is chosen for the average band frequency, and there are two possibilities.

The easier value to determine in practice is the average rate of band sharing, which is equal to the band frequency when the band frequency is constant. To see what band sharing rate would be observed, imagine a population of 1000 people. Of these, 500 have a G band and 250 have an H band. Hence if we pick a band at random,

- 2/3 of the time we pick a G band, which is shared with 1/2 of the population;
- 1/3 of the time we pick an H band, which is shared with 1/4 of the population.

Therefore the average rate of band sharing would be

$$(2/3)(1/2) + (1/3)(1/4) = 5/12.$$

Using this figure, we would calculate a likelihood ratio of $(12/5)^2 = 6$ for matching two bands between stain and suspect, which is duly conservative compared to 8.

Alternatively, if we take the ideal model very seriously then we might imagine that we arrive at the correct average band frequency of 3/8 by using Hummel's method. In that case we would compute a likelihood ratio of $(8/3)^2 = 7$, still conservative.

For paternity evaluation purposes the situation is rather more complicated. Given the same example of two band positions, G and H, and assuming that the average band frequency, 3/8, is used for calculations according to the method of Evett et al and Hummel et al, the idealized assumptions result in a conservatively small likelihood ratio for some combinations of patterns, but an unfairly high number in other cases. For example, when the "incriminating" -++ pattern (no band in mother, band in child, band in man) occurs at the G position and not at the H position, the idealized number is unfair typically by 40%. Choosing the average band-share rate of 5/12 for band frequency reduces this particular over-estimate, but at the cost of unfairly exaggerating the evidentiary significance when child and man share non-band (+--, and the controversial --- case).

On the other hand, when -++ occurs at both loci the net effect is safely conservative. There is therefore a temptation to consider -++ patterns only. Also, by limiting analysis to that pattern one avoids the pitfall (alluded to by Evett et al, 1990) of needing an accurate value for band frequency; an overestimate of the band frequency is good enough to ensure a conservative likelihood ratio for the -++ pattern.

Independence of Bands

The most extreme example of interdependence would be two band positions that are always merely copies of one another. Clearly this would be a bad thing if ignored, resulting in a likelihood ratio that is too high by the square.

However, there are some surprises.

Imagine band positions, A and Z, that are 100% linked, but complementary. That is, considering the pair as a haplotype the only possibilities are A-Z+ and A+Z-. Suppose that these two

haplotypes occur with equal frequency. As a measure of the effect of the linkage we can calculate the power of exclusion of this pair of systems compared with the power of exclusion of a similar but non-linked pair.

Position A excludes exactly when the mother has genotype A-A-, the child A-A+, and the tested man has A-A-, a 1/32 chance. Also, position Z excludes when the A locus genotypes are A+A+, A+A-, A+A+, which is an additional 1/32 chance, for a total exclusion probability of exactly 1/16. On the other hand, if A and Z were independent the combined exclusion probability would actually be a shade less -- $1/16-(1/32)^2$.

So lack of independence is not necessarily damning.

Nor is the example artificial. In fact, it is exactly a single locus with two equally frequent alleles, A and Z. Incidentally, if the bands are known to represent such a system, a further 1/8 of cases are exclusions: A-A- in the child with Z-Z- in the man, or the opposite (Jeff Morris, personal communication).

In order to investigate independence further, simulations were carried out with the aid of a computer. The idea is to create simulated data that violates the independence assumption, make calculations as if independence held, then see how far wrong the calculations are. Each simulation was a Monte Carlo construction of the following sort:

1. A value is chosen for N, the number of band positions.
2. A restricted set of permissible N-position haplotypes is chosen.
3. Mother-Child-Father true trios are generated using the haplotypes. False trios are generated by displacing the fathers to different cases.

This is a flexible scheme, as will appear. The analysis includes some of the following statistics:

4. For each false trio an "ideal" L value (=X/Y) is calculated by counting the incidence of each of the eight patterns ---, --+, ... and assuming the idealized model (i.e. using the Hummel/Evett formulae).
5. The average ideal L for false trios is computed.[1] As noted in Nijenhuis (1983), the average value of L for false trios is 1. If the average computed for the simulation is greater than 1, then L from the idealized model must on the average be unfair by that ratio.

---

[1] Equivalently, this number is the ratio of actual to predicted numbers of Random Men Not Excluded. The prediction is made by trusting the idealized L values and averaging 1/L among non-exclusion cases, per Nijenhuis (1983).

6. In some cases a "redundancy factor," f, was computed as follows: Let the ideal L values for the simulated false trios be denoted L_i. The factor f (usually f>1) is determined such that the average of the numbers $L_i^{1/f}$ is 1. This is a measure of the effective number of times the information at each band position is duplicated. Restated, a system with N bands and redundancy f is about equivalent in evidentiary power to an ideal system with N/f bands.
7. The true trio ideal L values are computed, then adjusted by taking the f-th root.

Simulation (1) has N=5 band positions, haplotypes +----, -+---, --+--, ---+-, ----+, thus modelling a locus with five alleles and allele frequency 0.2. Average ideal L for 2000 false trios was 1.15.

Simulation (2) models a locus with ten alleles, allele frequency 0.1. The average ideal L over 1000 false trios was 1.29.

Simulation (3) has N=10 and 55 haplotypes, namely those with only one or two +'s. This models two overlapping 10-allele loci. The average ideal L over 1000 false trios was 1.09.

For simulations (1)-(3), band position by position the ideal model L values are correct. However in multiplying them together statistical independence is incorrectly assumed, and that is why the average L values are a bit too large. That they are only a little too large suggests that allelism and overlap are not serious problems, and that it will be possible to determine compensating factors depending on N and on the band frequencies.

Simulation (4) has N=6 and only the two haplotypes ---+++ and +++---. Thus, band positions 1 and 4 are the "AZ" system described above, while the remaining band positions are redundant copies of these two informative ones. The simulation comprised 1667 false trios, with an average ideal L value of 1.9. The best fit for the redundancy factor f was 2.7, though the correct value is obviously 3. This simulation is included to indicate the limitation of the method for determining f.

Experiment (5) is a series of simulations designed to assess the impact of random linkage as N grows, and to overcome the difficulty of determining f accurately.

The series consists of simulations with N=3, 4, ... up to 31. For each simulation only 8 of the possible 2^N haplotypes are permitted. For N=3 these haplotypes are the complete set of possible combinations, so this simulation models complete independence. For each larger value of N each of the 3-haplotypes is augmented to length N by a random selection of + and -'s, as in Fig. 1. Since the entire haplotype is thus determined by the first three symbols, a naive hypothesis might be that the

augmented haplotype is no more informative than when N=3. That is, one might predict that N/f=3. Such a view overlooks the situation illustrated in Fig. 2, where the phenotype at the bottom band position reveals information (an exclusion) not apparent from the first three band positions. In fact N/f gradually grows to about 8.

Experiment (6) is like (5), but intended to be more realistic by approximating an allele frequency of 0.2 rather than 0.5. In this case N/f grows to about 6, when N=20.

-	-	-	-	+	+	+	+
-	-	+	+	-	-	+	+
-	+	-	+	-	+	-	+
+	+	-	+	+	+	-	-
-	+	-	-	+	-	-	+
h	i	j	k	l	m	n	o

haplotype

Figure 1
The haplotypes that were allowed for experiment (5), N=5. Each column is a haplotype; each row is a band position

	Mother		Child		Man	
genotype	h	j	j	l	m	n
phenotype	-		+		+	
	+		+		+	
	-		-		+	
	+		+		+	
>>>	-		+		-	

Figure 2
A typical false trio in experiment (5), N=5. The haplotypes are from Fig. 1

Simulation (7) introduces a new idea of using the redundancy factor to reevaluate true trios. From 200 false trios using 64 random haplotypes with N=50 and allele frequency 0.2, f=1.96. Likelihood ratios were then calculated for 200 true trios assuming the ideal model. The geometric mean was 5900, and the largest was 230000000. When each was replaced by its 1/f power, the resulting collection had a geometric mean of 84 -- the sort of likelihood ratio that apparently follows even from assuming only a very small degree of independence.

This simulation seems to have the following interpretation: The haplotype information is highly redundant, equivalent of only 6 independent positions. Nonetheless, the phenotypic information is equivalent to that from 25 (50/1.96) independent positions.

Simulation (8) is designed to mimic the case data analyzed in Hummel et al, and to suggest some of the vast possibilities for analysis of such simulations. N=80, and 64 random haplotypes were selected with predominately -'s rather than +'s. From 125 Monte Carlo trios, the allele frequency was 0.054 and band frequency 0.104. Of the 125 false trios, 122 were excluded (not 99.9998% as predicted from the ideal L-values, but still pretty good), with 3.4±1.3 exclusions (-+- patterns) per trio. The true trios had a geometric mean ideal L value of 22000, reduced to 22 by taking the f=3.2 roots. The true trios had 3.8±1.4 -++ patterns per trio (compared to 1.3±0.8 for the false trios). The -++ contributions to the true trio ideal likelihood ratios account for about 1000 out of 22000.

Applying the maximum likelihood method of Hummel et al estimates N=69.9 and band frequency as 0.12, close to the Hummel case

values. When these parameters are used to calculate ideal L values, the low N is safe because it results in underestimating the number of moderately paternity-indicative "---" patterns. Low maximum-likelihood estimates for N seem to be the rule. However, the overestimate for the band frequency inflates the evidentiary significance of most patterns. For example, the -++ pattern is thus counted with a likelihood ratio of 8.37, while it was actually observed only 7.54 times more frequently among the true trios than among the false trios.

Ambiguity, and Mutation

These problems contribute a significant share to the difficulty of analyzing multi-locus patterns and to their acceptability in forensics. The obvious approach is that they should included in the model. With regard to mutation this is often done already. As for questions of ambiguity, which means about the same as objectivity, it might be difficult but it can be done.

CONCLUSION

Linkage and independence are to some extent a bugbear. With 80 variable band positions there are 2^{80} possible haplotypes, all of which should exist assuming independence. But even a handful are enough to exclude nearly all non-fathers, and knowledge of their existence is enough to imply somewhat useful likelihood ratios for fathers.

Haplotypes and phenotypes are different things. Arguments and concerns about linkage, interdependence, and allelic association pertain mostly to haplotypes, and only indirectly to phenotypes. But measurements and observations are of phenotypes. That's why it is possible, and indeed turns out to be the case, that multi-locus measurements are not too tainted by haplotype interdependence.

These results herein are only an indication. It would be desirable to take some hard data that quantifies at least partial independence among some of the band positions, and build it into the simulations. It might turn out that a modest amount of independence is enough to justify very strong inferences from multi-locus probe data.

REFERENCES

Evett IW, Werrett DJ, Buckleton JS (1989) Paternity Calculations from DNA Multilocus Profiles. J Forens Sci Soc 29: 249-254

Fimmers R, Epplen JT, Schneider PM, Baur MP (1990) Likelihood Calculations in Paternity Testing on the Basis of DNA-Fingerprints. Advances in Forensic Haemogenetics 3: 14-16

Hummel K, Fukshansky N, Bär W, Zang K (1990) Biostatistical Approaches Using Minisatellite DNA Patterns in Paternity Cases (Mother-Child-Putative-father Trios). Advances in Forensic Haemogenetics 3: 17-19

Hummel K, Fukshansky N, Bär W (1990) Kinship Plausibilities from DNA Fingerprints. Advances in Forensic Haemogenetics 3: 20-22

Nijenhuis LE, A Critical Evaluation of Various Methods of Approaching Probability of Paternity (1983) Inclusion Probabilities in Parentage Testing, American Association of Blood Banks 103-112