Comparison of Different Methods for the Calculation of Indices of Paternity

R. Fimmers*, P.M. Schneider**, M.P. Baur*

* Institute for Medical Statistics, University of Bonn, Sigmund-Freud-Str. 25, Germany

** Institut für Rechtsmedizin der Universität Mainz, am Pulverturm 3, Germany

INTRODUCTION

The qualitative decision about paternity in trio cases on the basis of DNA multilocus profiles is no problem. If there are more than 1 or 2 exclusion patterns (band present in child, which is neither present in mother or alleged father), the putative father has to be excluded. The problems arise, if, in the case of a non exclusion, one wants to quantify the evidence for paternity. Different statistics have been proposed for this purpose. This paper discusses three of these statistics and demonstrates their application to real data.

NUMBER OF INFORMATIVE BANDS

The simplest statistic is the number of informative bands, i.e. bands, which are present in the child and the alleged father, but not in the mother. Let b_i be the frequency of the i-th informative band for a given case, then

$$\prod_{i=1}^{k} b_i \tag{1}$$

is the chance of finding this set of bands in a random man (disregarding all other band positions) and

$$A = 1 - \prod_{i=1}^{k} b_i \tag{2}$$

is the exclusion chance for such a random man. The application of a conservative equal frequency b for all bands reduces 2 to

$$A = 1 - b^k \tag{3}$$

were k is the number of informative bands. The average posterior probability W assuming equal priors then is given by

$$\overline{W} = \frac{1}{1 + \frac{1-A}{1}} = \frac{1}{2 - A}. \tag{4}$$

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

BAND SHARING

Band sharing is another straightforward measure of similarity between DNA multilocus profiles. It can be determined for any pair of individual band patterns, which normally should be placed in two adjacent lanes on the same blot. Let n_A and n_B be the number of bands in individual A and in individual B, and s_{AB} the number of bands shared between A and B.

From the simple question: "What is the probability, that B has a band, which is present in A", we get a natural definition of the band sharing rate

$$r^*_{AB} = \frac{s_{AB}}{n_A}. \tag{5}$$

This approach has the problem, that it is not symmetric in A and B (i.e. $r^*_{AB} \neq r^*_{BA}$), whenever $n_A \neq n_B$.

The most commonly used way to define band sharing is

$$\tilde{r}_{AB} = \frac{2s_{AB}}{n_A + n_B}, \tag{6}$$

which of course is symmetric, but is biased to an underestimation of the band sharing rate, if the number of bands n_A and n_B are different.

An unbiased and symmetrical estimator of the band sharing rate is obtained by

$$r_{AB} = \frac{1}{2}(r^*_{AB} + r^*_{BA}) = \frac{s_{AB}}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right), \tag{7}$$

the measure of band sharing, which will be used in the subsequent text.

Band sharing depends on the degree of relationship between the two individuals. Based on the kind of relationship and the expected band sharing between unrelated individuals, which stands for the chance of a random match, it is possible to calculate expected band sharing rates for all types of kinship. Honma et. al. [5] have given formulas for the most relevant types of kinship. For the parent child relationship the result was reformulated by Jeffreys et. al. [8] on the basis of band frequencies and transmission probabilities.

The grandparent grandchild and the uncle nephew cases seem to be wrong in the Honma paper. Describing the kinship in terms of the probabilities p_0, p_1 and p_2 of having 0, 1, or 2 alleles identical by descent, the expected band sharing rate can be calculated as (a = allele frequency)

$$p_0(2a - a^2) + p_1\frac{1 + a - a^2}{2 - a} + p_2. \tag{8}$$

For $p_0 = \frac{1}{2}, p_1 = \frac{1}{2}, p_2 = 0$ (grandparent-grandchild as well as uncle-nephew) this evaluates to

$$1 + 5a - 5a^2 + a^3, \tag{9}$$

deviating from Honma et. al.. Formula 8 can be applied for all kinds of kinship. For parent child we get ($p_0 = 0, p_1 = 1, p_2 = 0$)

$$\frac{1 + a - a^2}{2 - a}, \tag{10}$$

equivalent to Honma. Jeffreys et. al. give in this case the formula (b = band frequency)

$$\frac{1}{b}(2b - 1 + \sqrt{1 - b}^3), \tag{11}$$

which is equivalent to formula 10 using the relation between band and allele frequency $b = a(2 - a)$. It has to be stressed, that the results are the same, though the approach by Jeffreys et. al. does not require the assumption of one locus with defined alleles producing the band, but works with band frequencies and transmission probabilities, which may be a more general approach.

LIKELIHOOD APPROACH

The method to quantify paternity, which is here called the likelihood approach, was proposed independently by Evett et. al. [1], Honma et. al. [4] and Hummel et. al. [6]. The main idea is to assume independence between the bands of a DNA multilocus profile. This allows to treat the band positions separately and to get an overall likelihood by multiplication. The assumption of independence is very strong. A genetic model for the band pattern, which leads to this independence, has to postulate, that the bands of a DNA multilocus profile are not allelic and that the loci are not linked. In practice, these requirements can only be fulfilled to a certain degree.

The special aspect of the Evett approach is, that he formulates the problem in terms of band frequencies and transmission probabilities. Therefore he needs no further assumptions except that independence between band position holds, and that bands are transmitted from parent to child with a certain probability (e.g. it does not bother how many loci are involved in the production of one band).

Honma et. al. [4] and Hummel et. al. [6] use a model, which seems formally more restrictive. They interpret the presence or absence of a band at a single position as coming from a diallelic locus with one recessive allele (no band). Consequently the calculation of likelihoods for the single band position is a standard procedure. The difference between Honma and Hummel is the question, which types of band pattern should be evaluated at the different band positions. The results are identical to the Evett approach. An additional assumption, which is not necessary, but often made, is to assume equal band or allele frequencies for all band positions. This may lead to wrong (anti conservative) results, if the overall estimation of the allele frequency is not chosen with caution.

For a usual trio case, with mother, child and alleged father and a fixed band position, the vector $(\cdot, \cdot, \cdot)$ will give the information, whether or not mother (first position), child (second position) and alleged father (third position) have a band at this position of their DNA profile. $+$ will indicate presence, $-$ will indicate absence of a band. E.g. $(-, +, +)$ means, that the mother has no band, but child and alleged father have a band at the position in question. Using this notation we have eight different types of band patterns which can occur in the different positions. Some of these are worth to be discussed in more detail.

Depending on whether the alleged father is excluded or not excluded the pattern $(-, +, -)$ can be designated as "exclusion" or as "mutation" band pattern. The qualitative decision about paternity can be based exclusively on this type of band pattern. The formal models, which are used here to calculate likelihoods, do not allow the occurrence of $(-, +, -)$ under the assumption of paternity.

In a case with one or two of these patterns, and the alleged father not excluded, it is necessary to include this information into the calculation of the likelihoods. Similar to the singlelocus case [3] one can use a kind of global mutation rate (the probability of the occurrence of $(-,+,-)$ patterns in the complete band pattern of a triplet). Together with the allele frequency a one can calculate an approximate paternity index of $\frac{\mu}{a}$ for the $(-,+,-)$ pattern.

The other important pattern with high information for paternity is $(-,+,+)$. Depending on the band frequency the evidence for paternity from this pattern is high in comparison to all other band patterns (except $(-,+,-)$). Honma et.al. [4] propose, to use this type of pattern only. The number of these "informative bands" can be regarded as an independent characteristic value for paternity (see above) and Honma's proposition leads to the above defined exclusion chance. The number of $(-,+,+)$ patterns is also the fundamental part of band sharing.

The most controversial pattern is $(-,-,-)$. This patterns, in contrast to all other patterns, are not defined by the occurrence of a band. Their number is gained by way of estimation of the potential number of band positions. Following Hummel [7], the effective number of bands N_{eff} can be estimated from the average number of bands per individual n and the average number of bands shared between pairs of two unrelated individuals s, as follows

$$N_{eff} = \frac{n^2}{s}. \tag{12}$$

A different way to define the number of $(-,-,-)$ patterns, is to define the number of band positions using a binning approach.

The question is, whether or not these patterns should be regarded as information about paternity and can be included into the calculation of the likelihoods? The strong independence assumptions, which had to be made (especially for the Hummel version) lead to a divergence between the model and the unknown genetic reality. The $(-,-,-)$ patterns can be regarded as an artifact introduced by the model. A meaningful approximation to the genetical background of a multilocus profile is to understand it as an overlay of an unknown number of singlelocus patterns. From this point of view there is no reason to assume information in positions without a band. A discussion about the usage of these patterns will always lead to a discussion, whether or not the assumptions of independence and multiple diallelic systems are appropriate. From a pure formal genetic point of view they are obviously not.

Nevertheless it is interesting to look at the empirical distribution of the resulting likelihood ratios in case of paternity and non paternity [2]. The $(-,-,-)$ patterns have to be judged according to their impact on the resulting likelihood values. Depending on the band frequencies the $(-,-,-)$ pattern suggests evidence for paternity, which increases exponentially with the number of postulated positions N_{eff}. The inclusion of this "information" into the calculation of overall likelihood will bias the result towards assuming paternity. Not to use the $(-,-,-)$ patterns is therefore conservative in the sense of the non excluded father. The effect of the use of the $(-,-,-)$ patterns will be demonstrated in the following application of the method to empirical data.

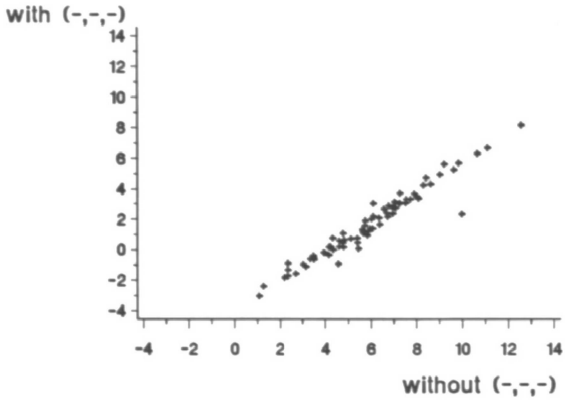

Figure 1: EM-values, with vs. without $(-,-,-)$ patterns

APPLICATION TO EMPIRICAL DATA

The objective of this paper is the comparison of different characteristics for paternity based on DNA multilocus profiles in their application to real data. The cases, which were used came from the MZ 1.3 study (see Schneider [9]). 76 Mother-child-alleged-father triplets were taken from the data of one participating laboratory. Paternity was confirmed by the results from blood-group, HLA and singlelocus DNA systems. The migration distances had been read to an accuracy of 0.5mm. Bands in adjacent lanes were regarded to come from equal sized fragment, if their running distances did not differ more than 0.5mm. An algorithm was used to correct for these differences. The numbers of the eight $+/-$ -patterns were counted for each of the 76 families. Based on this "$+/-$ -statistic" it was easily possible to calculate the paternity characteristics. For the application of the likelihood approach (incl. $(-,-,-)$) an overall allele frequency a and the effective number of bands were estimated as proposed by Hummel [7].

RESULTS

Figure 1 shows the comparison of the Essen-Möller-values (EM-values) with and without the $(-,-,-)$ combinations. The values from a more or less straight line, which shows, that the inclusion of $(-,-,-)$ does not give additional (different) information. The EM-values are shifted about 4 units towards assuming paternity, which means, that paternity indices are inflated by a factor of 10000.

Figures 2,3,4 show all bivariate plots for the number of $(-,+,+)$, band sharing and EM-values (without $(-,-,-)$). As expected from the theoretical considerations, the values are highly correlated. The correlation is smallest for band sharing and the number of $(-,+,+)$ patterns.

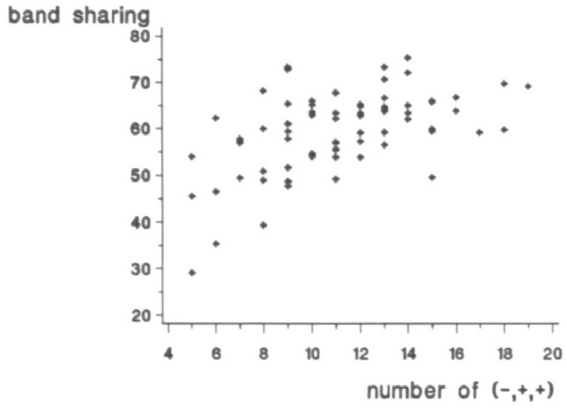

Figure 2: Band sharing vs. number of $(-,+,+)$ patterns

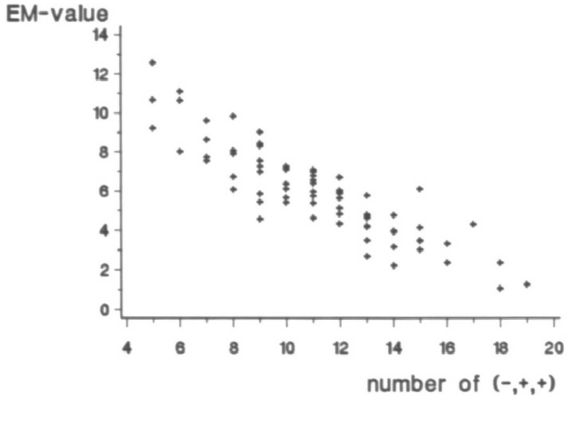

Figure 3: EM-values vs. number of $(-,+,+)$ patterns

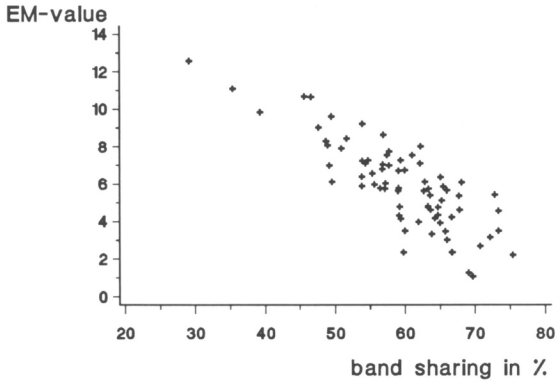

Figure 4: Band sharing vs. EM-values

DISCUSSION

The application of the three different paternity statistics for DNA multilocus profiles shows two important results. The inclusion of $(-,-,-)$ patterns into the calculation of likelihood values results in a systematical bias for likelihood ratios and EM-values. The interpretation for the $(-,-,-)$ patterns is questionable and may be only understood in connection with the assumptions of the "multiple diallelism model". The justification to chose this approach is its practicability and simplicity. Because of the strong anticonservative bias, which is introduced through the $(-,-,-)$ patterns, we strongly argue against this approach. The extension of this method to complex deficiency cases, which would be a standard procedure as in any other well defined Mendelian system, is not possible due to the strong impact of the false assumptions.

The comparison of the number of $(-,+,+)$ patterns, the band sharing rate and the EM-values showed that all three characteristics are highly correlated. Their informational contents is approximately the same. The larger difference between the band sharing and the number of $(-,+,+)$ patterns may be due to the fact, that band sharing also takes $(+,+,+)$ patterns into account and that the number of shared bands is put into relation to the total number of bands.

Nevertheless there is quite a large difference in the interpretation of the three statistics. Band sharing has expected values which depend on the degree of relation between the compared persons. High band sharing (above 50%) is information for a first degree relationship. Low band sharing is characteristic for unrelated individuals. The number of $(-,+,+)$ patterns can be transformed into an exclusion probability. The W-value (equivalent to the likelihood ratio and the EM-value) has the strongest interpretation as a Bayesian aposteriori probability for paternity. If one uses the W-value interpretation, one should be aware, that the resolution power of this statistic cannot be greater than for the band sharing rate, with all implications for the interpretation of the result.

References

[1] Evett IW, Werrett DJ, Buckleton JS (1989) Paternity calculation from DNA multilocusprofiles. J Forens Sci Soc 29: 249-254

[2] Fimmers R, Epplen JT, Schneider PM, Baur MP (1989) Likelihood calculation in paternitytesting on the basis of DNA fingerprints. In: Polesky HF, Mayr WR (eds) Advances in forensic haemogenetics 3. Springer, Berlin Heidelberg New York, pp 14-16

[3] Fimmers R, Henke L, Henke J, Baur MP (1991) How to Deal with Mutations in DNA-Testing. This volume

[4] Honma M, Ishiyama I (1989) Probability of paternity in paternity testing using the DNA fingerprint procedure. Hum Hered 39:165-169

[5] Honma M, Ishiyama I (1990) Application of DNA fingerprinting to parentage and extended family relationship testing. Hum Hered 40:356-362

[6] Hummel K, Fukshanski N (1990) Biostatistical approaches using minisatelite DNA patterns in paternity cases (mother-child-putative father trios). In: Polesky HF, Mayr WR (eds) Advances in forensic haemogenetics 3. Springer, Berlin Heidelberg New York, pp 17-19

[7] Hummel K (1991) Biostatistische Auswertung von DNA-Bandenmustern in Fällen strittiger Identität und Blutsverwandtschaft. Klin Lab 37:252-258

[8] Jeffreys AJ, Turner M, Debenham P (1991) The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework. Am J Hum Genet 39: 11-24

[9] Schneider PM, Fimmers R, Bertrams J, Birkner P, Braunbeck K, Bulnheim U, Feuerbach M, Henke L, Iten E, Osterhaus E, Prinz M, Simeoni E, Rittner C (1991) Biostatistical basis of individualization and segregation analysis using the multilocus DNA probe MZ 1.3: Results of a collaborative study. (submitted for publication)