Jeffrey W. Morris,[1] M.D., Ph.D.; A. I. Sanda,[2] Ph.D.; and Jeffrey Glassberg,[3] Ph.D.

Biostatistical Evaluation of Evidence from Continuous Allele Frequency Distribution Deoxyribonucleic Acid (DNA) Probes in Reference to Disputed Paternity and Identity

REFERENCE: Morris, J. W., Sanda, A. I., and Glassberg, J., "**Biostatistical Evaluation of Evidence from Continuous Allele Frequency Distribution Deoxyribonucleic Acid (DNA) Probes in Reference to Disputed Paternity and Identity,**" *Journal of Forensic Sciences*, JFSCA, Vol. 34, No. 6, Nov. 1989, pp. 1311–1317.

ABSTRACT: We present a development and discussion of the biostatistical evaluation of deoxyribonucleic acid (DNA) probe evidence in forensic science cases of disputed paternity and identity. We restrict ourselves to single-locus codominant systems (highly analogous to more conventional systems) which have the apparently novel complication of an experimentally continuous allele frequency distribution. This complication necessitates reformulations of standard biostatistical summaries of the evidence (the paternity index (PI) and the phenotype frequency, respectively). These reformulations, rather than representing a unique case, have applicability to the evaluation of evidence obtained in standard genetic systems now in widespread use.

KEYWORDS: pathology and biology, paternity, deoxyribonucleic acid (DNA)

Theoretical Considerations

The paternity index (PI) is a classical likelihood ratio that tests the mutually exclusive hypotheses of paternity and nonpaternity [*1*]. It is the quotient of conditional probabilities (likelihoods):

$$\mathrm{PI} = \frac{P\ (\text{genetic observations}|\text{paternity})}{P\ (\text{genetic observations}|\text{nonpaternity})} \tag{1}$$

The relevant genetic observations are the results of phenotyping studies of the trio (mother-child-alleged father); typically, before evaluation of the PI the results of phe-

Received for publication 28 Sept. 1988; revised manuscript received 23 Dec. 1988; accepted for publication 28 Dec. 1988.

[1] Associate clinical professor of pathology, University of California, Irvine, School of Medicine, and Department of Pathology, Memorial Medical Center, Long Beach, CA.

[2] Associate professor of physics, Rockefeller University, New York, NY, and Lifecodes Corp., Valhalla, NY.

[3] Formerly, vice president, Lifecodes Corp., Valhalla, NY, now, president, Nucleics Inc., Chappaqua, NY.

notyping studies have been reduced to apparently discrete phenotypes. This step, which simplifies the numerical evaluation may obscure significant problems in the evaluation of the evidence. Given that discrete phenotypes have been assigned, the PI reduces to:

$$\text{PI} = \frac{P\text{ (phenotypes of trio}|\text{paternity)}}{P\text{ (phenotypes of trio}|\text{nonpaternity)}} \qquad (2)$$

Application of the rules of conditional probability, along with a few reasonable genetic assumptions permits several alternative formulations of the PI, of which the following is the most convenient form:

$$\text{PI} = \frac{P(\text{C}|\text{M,AF paternity})}{P(\text{C}|\text{M,AF nonpaternity})} \qquad (3)$$

where C, M, and AF refer to phenotypes assigned to child, mother, and alleged father, respectively. The numerator of Eq 3 is the frequency with which children of Phenotypes C are found among children whose parents are Phenotypes M and AF, while the denominator of Eq 3 is the frequency with which children of Phenotypes C are found among children whose mothers are of Phenotypes M.

As noted, Eq 3 contains a reduction and simplification of the genetic evidence; discrete phenotypes have been assigned. For reasons that will become apparent, it is convenient to reformulate the PI as:

$$\text{PI} = \frac{P(\text{C}^*|\text{M}^*,\text{AF}^*\text{ paternity})}{P(\text{C}^*|\text{M}^*,\text{AF}^*\text{ nonpaternity})} \qquad (3^*)$$

where the superscript* indicates that it is the actual genetic evidence, rather than assigned phenotypes, which is being evaluated.

Evaluation of Continuous Allele Frequency Distribution DNA Probe Evidence

Consider Fig. 1*a*, which is the genetic evidence obtained in the DNA probe system pAC255 in an actual paternity case. Three points are apparent. First, the alleged father is not excluded from paternity, as he possesses an allele indistinguishable from the paternal allele (at 8.40 kb) of the child. Second, referring to Fig. 1*b*, it is clear that a discrete phenotype cannot be assigned to any of the trio. Third, there is no way to determine whether or not the alleged father actually possesses the paternal allele; the operational meaning of nonexclusion is reduced to indistinguishablity of alleles. As we point out below, this final point, while obvious in this genetic system, is common to all genetic systems now used in disputed paternity studies.

As discrete phenotypes cannot be assigned Eq 3* rather than Eq 3 is appropriate. The numerator of Eq 3* for the data of Fig. 1 is straightforward; one quarter of the offspring of parents of the observed phenotypic results will have the phenotypic results observed for the child.

The analysis of the denominator of Eq 3* is more complex. The data to be evaluated consist of the measured molecular weights (MWs) of the alleles of the trio and the indistinguishability of the paternal allele and one of the alleged father's alleles. The denominator of Eq 3* has the form:

$$\text{denominator } 3^* = (1/2)g(S_i,\delta,\sigma)$$

where S_i are the measured allele MWs, δ is the discrimination power of the analytical system (a measurement of the inability to distinguish alleles differing by a few base pairs),

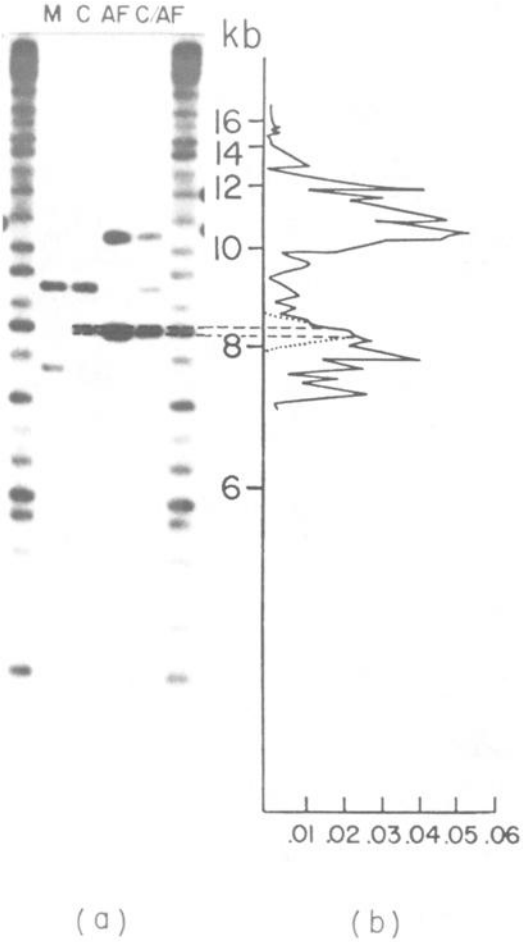


FIG. 1—*Interpretation of a paternity test.* (a): *Results of a Southern Blot and subsequent hybridization with probe pAC255 (see Ref 2 for procedures). The two flanking lanes contain a set of marker DNAs of known lengths. Lane 1 contains Pst digested DNA from the mother, Lane 2 DNA from the child, Lane 3 DNA from the alleged father, and Lane 4 a mixture of DNA from the alleged father and the child.* (b): *The frequency distribution of alleles visualized with pAC255. Size is in kilobase pairs. The width between the dashed lines represents* δ. *The mathematical computations described in the text are schematically demonstrated. The dotted line is graphic representation of the exponential factor of Eq 5 applied to this case. The column labelled C/AF corresponds to the lane that contains DNA from both the child and the alleged father. It helps to give a visual comparison of the DNA from the child and the alleged father.*

and σ is the standard deviation of the measurement of MW. Both δ and σ are functions of a number of experimental parameters, including the particular conditions of electrophoresis and blotting and the molecular weight of the alleles. The value δ is also a function of the measurement technique, including the quality of the molecular weight markers selected and the method of measurement. The values δ and σ must be determined individually in each laboratory. For the data of Fig. 1, the frequency with which children of the observed alleles are produced by mothers of the observed alleles is equal to ½ × the frequency of alleles in the appropriate population which are indistinguishable from the paternal allele. This immediately suggests that denominator 3* is one half the area under the allele frequency density ± δ from the measured paternal allele:

$$\text{denominator } 3^* = 1/2 \int_{Spa-\delta}^{Spa+\delta} f(S)dS \qquad (4)$$

where $f(S)$ is the allele frequency density [$\int f(S)\, dS = 1$], S_{pa} is the measured paternal allele, and δ is as described above. However, the uncertainty in assignment of S_{pa} means that the assigned midpoint of the range of integration is approximate. One way to account for this uncertainty is to allow the midpoint of the integrated range to vary about S_{pa}, and to weigh such contributions by an appropriate exponential factor. This results in the approximation:

$$\text{denominator } 3^* = (1/2)g(S_{pa},\delta,\sigma)$$

$$= \frac{1}{2\sqrt{2\pi\sigma}} \int_0^{\infty} dS \int_{s-\delta}^{s+\delta} dS' \exp[-(S - S_{pa})^2/2\sigma^2]\, f(S') \qquad (5)$$

For Eqs 4 and 5, δ and σ are fixed at the values assigned for S_{pa}; they are not variables in the integration. As σ → 0, Eq 5 reduces to Eq 4. For discrete alleles, the integral of Eq 4 is replaced by a sum.

$$\text{denominator } 3^* = 1/2 \sum_{i=1}^{n} h(S_i)\Delta(S_i,S_{pa})$$

where $h(S_i)$ is the population frequency of the i^{th} allele and $\Delta(S_i, S_{pa})$ = probability that the i^{th} allele would not be distinguished from the paternal allele. If classification is unambiguous and perfect, $\Delta(S_i, S_{pa}) = 1$ if $i = pa$, and 0 for $i = pa$, yielding:

$$\text{denominator } 3^* = 1/2\, h(S_{pa})$$

the classical result.

The integral of the PI for the various nonexcluding phenotypic observations is given in Fig. 2. Note that the example shown in Fig. 1*a* corresponds to Pattern 8 in Fig. 2.

Extension to Disputed Identity

For disputed identity, the basic biostatistic is the probability that a randomly chosen individual would match a given phenotype. For continuous allele frequency DNA probes this phenotype frequency is given by:

$$2g(S_1,\delta,\sigma)\; g(S_2,\delta,\sigma)$$

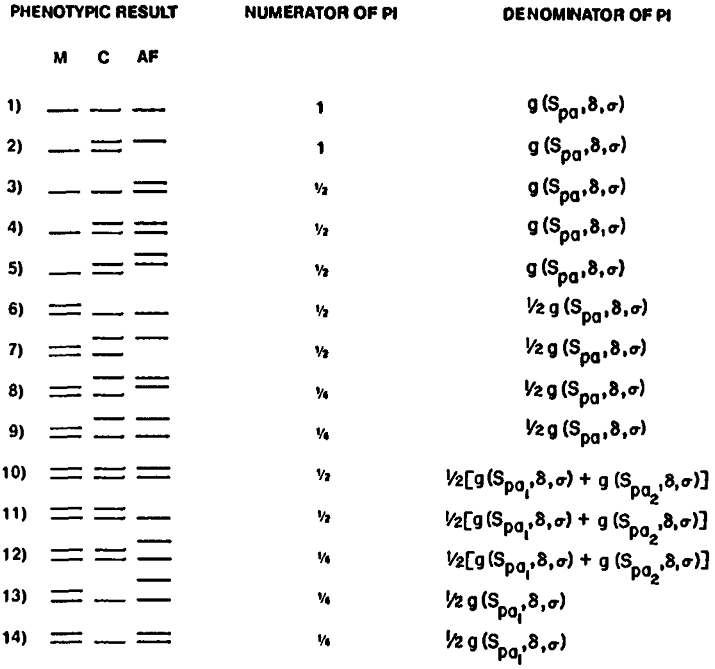
for a match to a heterozygous phenotype S_1S_2, and

$$g^2(S_1,\delta,\sigma)$$

for a match to an apparently homozygous phenotype S_1. The mathematical expression for $g(S, \delta, \sigma)$ is given by Eq 5.

Discussion

Our analysis has explicitly taken into account two properties of genetic evidence obtained in systems with continuous allele frequency distributions. First, alleles are shown



PHENOTYPIC RESULT	NUMERATOR OF PI	DENOMINATOR OF PI
1)	1	$g(S_{pa}, \delta, \sigma)$
2)	1	$g(S_{pa}, \delta, \sigma)$
3)	½	$g(S_{pa}, \delta, \sigma)$
4)	½	$g(S_{pa}, \delta, \sigma)$
5)	½	$g(S_{pa}, \delta, \sigma)$
6)	½	$½ g(S_{pa}, \delta, \sigma)$
7)	½	$½ g(S_{pa}, \delta, \sigma)$
8)	¼	$½ g(S_{pa}, \delta, \sigma)$
9)	¼	$½ g(S_{pa}, \delta, \sigma)$
10)	½	$½[g(S_{pa_1}, \delta, \sigma) + g(S_{pa_2}, \delta, \sigma)]$
11)	½	$½[g(S_{pa_1}, \delta, \sigma) + g(S_{pa_2}, \delta, \sigma)]$
12)	¼	$½[g(S_{pa_1}, \delta, \sigma) + g(S_{pa_2}, \delta, \sigma)]$
13)	¼	$½ g(S_{pa_1}, \delta, \sigma)$
14)	¼	$½ g(S_{pa_1}, \delta, \sigma)$

FIG. 2—*Under "phenotypic result" we show all possible nonexcluding patterns. The numerator and the denominator of the PI corresponding to each pattern are also given.*

only to be indistinguishable rather than identical, and second, discrete alleles cannot be assigned. The first property is common to all genetic systems and examples are well known. In the red cell antigen system ABO the paternal allele may be A and the alleged father AB, but identity of alleles has not been demonstrated as further investigation may demonstrate that the paternal allele is A_1 while the alleged father is A_2B. Even if such an example did not exist, resolution of alleles is dependent on experimental technique, and failure to discriminate does not prove identity. In some standard electrophoretic systems it is well known that assignment of phenotypes by conventional techniques lumps phenotypes distinguishable by improved resolution (for example, isoelectric focusing) [*3*]. In the HLA system broad specificities may be resolved into "splits"; variants of several well-defined specificities exist and segregate in a Mendelian fashion [*4*]. "Splits" and variants may or may not be resolved on any given day by any given set of reagents. In these conventional systems, the standard biostatistical evaluation is based on indistinguishability (nondistinguished alleles are lumped together) rather than identity, although it is commonplace to speak of identifying the paternal allele in the phenotype of the alleged father. For DNA probe systems the criterion of indistinguishability is based on coelectrophoresis experiments. We have provided a basis for biostatistical evaluation of the evidence based on resolving power.

The second property, that of the experimentally continuous allele frequency distribution, appears at first glance to have no counterpart in conventional genetic systems. For red cell antigen and protein and enzyme systems one could strain the analogy by assuming that experimental results on the mother, child, and alleged father are determined at the same time, so that indistinguishability of alleles is determined by direct comparison of results, rather than independent assignment of phenotype. The counterpart of the uncertainty in assigned phenotype as a result of measurement error in DNA systems would be the explicit acknowledgment of the possibility of misclassification of phenotypes in discrete systems, and a discrete analysis analogous to the integrals developed above could be made. However, in expert laboratories, misclassification of phenotypes in such systems occurs with negligible frequency.

The human lymphocyte antigen (HLA) system, however, is quite analogous. Consider for example, the HLA B locus, and the cross-reacting alleles B5, B35, and BW53. B5 often can be "split" into B51 and BW52 and further subdivisions are possible [4]. Typically, as many as ten reagents with specificities for one or more of these alleles are used, and reactions for each are scored by an essentially continuous scale [5]. It is the property of the reagents that none of them have 100% specificity and sensitivity for any allele [5]. Suppose, for example, a paternal allele is identified within this cross-reacting group, and that the results of typing the alleged father yield the identical result for each of ten relevant reagents as obtained for the child; there would be no question regarding indistinguishability. However, the particular pattern observed in the child and alleged father may be seen commonly with B35 and less commonly with BW53. Thus, while there is no difficulty with indistinguishability of alleles, the difficulty with assignment of phenotype and calculation of PI is quite analogous. On the other hand, the paternal allele may have a pattern of reaction quite characteristic of say, B35, and typing of the alleged father may also be quite characteristic of B35, but the results may suggest that the alleles are distinguishable, even though phenotypes may be assigned unambiguously. The HLA situation is analogous because the actual data is essentially continuous in nature, although the assignment of (most probable) phenotypes obscures this difficulty. We suggest that analogous biostatistical evaluation of HLA data may be both feasible and useful. In this regard, it is comforting that our integral equations reduce to the classical formulations in the limiting case of discrete alleles.

Finally, for a discussion of the effect of null alleles, if any, in these systems, see Ref *6*.

Note Added in Proof

Charles Brenner has noted that Eq 5, while providing some correction for measurement uncertainty, does not appropriately account for the situation in which, for S within a few standard deviations of S_{pa}, $f(S)$ differs significantly from $f(S_{pa})$. He proposes that denominator 3* be computed as

$$\frac{1}{2N}\int_0^\infty dS \int_{s-\delta}^{s+\delta} dS' \exp\,[-(S - Spa)^2/2\sigma^2]f/S')f(S) \tag{6}$$

where

$$N = \int_0^\infty dS \exp\,[-(S - S_{pa})^2/2\sigma^2]f/(S)$$

In practice, this modification results in significantly (and appropriately) lower *PI*s when S_{pu} lies in a relative minimum between nearby relative maxima. Thus, Eq 6 takes into account that S_{pa} may be an atypical measurement of a frequent allele, as well as that it may be a typical measurement of an infrequent allele, and is therefore the preferred form.

References

[1] Nijenhuis, L. E., "A Critical Evaluation of Various Methods of Approaching Probability of Paternity," *Inclusion Probabilities in Parentage Testing*, R. Walker, Ed., American Association of Blood Banks, Arlington, VA, 1983, pp. 103–114

[2] Baird, M., Balazs, I., Giusti, A., Miyazaki, L., Nicholas, L., et al., "Allele Frequency Distribution of Two Highly Polymorphic DNA Sequences in Three Ethnic Groups and Its Application to the Determination of Paternity," *American Journal of Human Genetics,* Vol. 39, 1986, pp. 489–501.

[3] Dykes, D., "The Use of Frequency Tables in Parentage Testing," in *Probability of Inclusion in Parentage Testing,* H. Silver, Ed., American Association of Blood Banks, Arlington, VA, 1982, pp. 15–44.

[4] Naik, S. and Mittal, K., "Segregation of BW51, BW52, BW53 and BW35 in Families: Splitting of BW51 into "Long" and "Short" Variants," in *Histocompatibility Testing,* P. I. Terasaki, Ed., UCLA Tissue Typing Laboratory, Los Angeles, 1980, p. 759.

[5] Mickey, M. R. and Terasaki, P. R., " The Serological Data of the 8th Workshop and Summary Analysis," in *Histocompatibility Testing,* P. I. Terasaki, Ed., UCLA Tissue Typing Laboratory, Los Angeles, 1980, pp. 21–136.

[6] Balazs, I., Baird, M., Clyne, M., and Meade, E., "Human Population Genetic Studies of Five Hypervariable DNA Loci," *American Journal of Human Genetics,* Vol. 44, 1989, pp. 182–190.

Address requests for reprints or additional information to
Jeffrey Morris, M.D., Ph.D.
Department of Pathology
Memorial Medical Center
Long Beach, CA 90806