J Forensic Sci, Sept. 2002, Vol. 47, No. 5
Paper ID JFS20002092_475
Published 21 Aug. 2002
Available online at: www.astm.org

Commentary on Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. J Forensic Sci 2001;46(6):1372–78.

Sir:

The article, "A Mathematical Approach to Resolving Mixed DNA Samples," by Dr. M. Perlin and Dr. B. Szabady, in the November, 2001 issue, presented a "heuristic search algorithm" to resolve a mixture DNA using a best-fit and search approach. An incorrect equation appeared on page 1374, which can lead to serious confusion for those readers who wish to follow the rationale and algorithm of the authors' Linear Mixture Analysis (LMA) approach.

The incorrect equation referred to is the first equation in the left column of page 1374, shown below:

$$w = \frac{G}{d} \qquad \text{Eq 1}$$

where ". . . column vector *d* describes the mixture profile's peak quantitation data, matrix *G* represents the genotypes, and *w* is the weight column vector that reflects the relative proportions of template DNA or PCR product." This follows from the equation on page 1372 in which the authors first state the formulation of the LMA least square problem. This equation is duplicated below:

$$d = G \cdot w \qquad \text{Eq 2}$$

From examining Eq 2, *w* is to be obtained by somehow "dividing *d* by *G* (as in $w = d/G$), but certainly not by dividing *G* by *d*" (as in $w = G/d$ as stated by the authors). Of course, since *G* and *d* are a matrix and a vector respectively, the division by *G,* has to be carried by multiplying *d,* by some kind of inverse of *G* through matrix operation. Even if the correct division was expressed in Eq 1 (as in $w = d/G$), the division operation as expressed by "—" by the authors applies only when both *G* and *d* are scalars. When *G* and *d* are not scalars, the expression for division cannot simply be represented by "—." The correct solution and matrix inverse expression in the solution to Eq 2 is given by

$$w = G^{+} \cdot d \qquad \text{Eq 3}$$

where G^{+} stands for the "pseudoinverse" of *G* (Ref 1), which is the left inverse of *G* if *G* is overdetermined, with more rows than columns and full rank, such that $G^{+} \cdot G = I$, where *I* is an identity matrix of size equal to the number of columns of *G.* Below where Eq 1 appeared on page 1374, the authors state that the solution to Eq 2 above can be computed by $(G^T G)^{-1} G^T$ which is correct when the inverse exists, but would fail if $\mathrm{G}^{\mathrm{T}}\mathrm{G}$ is not invertible. Regardless, posing the solution as in Eq 1 is not correct, because the expression is inverse to what it should be.

The authors further stated that in the MATLAB programming language, one uses the built-in left-division command of "\" (the backslash) to compute the least square solution *w,* as in $w = G/d.$ (Note that this is not the same as $w = G/d.$) The backslash division operation gives the same least square solution as that using the pseudoinverse only when *G* is full rank. When *G* is not full rank (as may happen in a 2-allele locus case in which the two genotypes are identical), the solution using the "\" operator does not give the correct least square solution, and thus not equal to that given by the pseudoinverse approach (2). In this case when *G* is not full rank, $(\mathrm{G}^{\mathrm{T}}\mathrm{G})$ is not invertible, and computation of the least square solution by way of $(G^T G)^{-1} G^T$ cannot proceed, but the pseudoinverse of G still exists and is the only way to compute the least square solution. Regardless of the rank of *G,* $w = \frac{G}{d}$ is not the correct expression for the computation of *w* as stated in the journal article.

A second comment is in order with respect to the inconsistent use of variable notation in the article, which leads to confusion for a novice reader wishing to understand the approach. That is, the variable *w* is used in two different ways: one denotes a vector, while the other denotes a scalar. The "*w*" that appears in the first equation on page 1372: $d = G \cdot w,$ and in other related equations refers to the weight vector, with as many elements as the number of contributors to the mixture DNA. However, the "*w*" that appears in the first equation in the right column on page 1374: $g(w) = (d - w \cdot gA)/(1 - w)$ denotes the weight fraction of one of the DNA in a two-people mixture, and is a scalar, and thus, is very different from the entity the first "*w*" represents.

A third comment refers to the authors' results in using their LMA in resolving a three person mixture, as presented on the bottom of page 1377. They stated "We then performed mixture deconvolution on the three person mixture data *d.* We used genotypes *gA* and *gB* as known references, but . . .". This suggests that one has to know the genotypes of two of the three people (person A and person B) in order to use LMA to compute and search for the optimum genotype for the third person as well as for the mixture weights. It would seem that having to know the genotype of two of the three persons involved in a crime, such as a rape case, limits the practicality of the use of LMA. Can it realistically be expected that one of the two perpetrators of the crime will always be known with certainty (the third person being the victim)? The authors' remarks under the "Other Analysis" section on page 1377 stating to the effect that "LMA model is also useful for resolving mixtures when there are no reference profiles available" reflects the more realistic scenario, but details and testing with data are not given in the article.

Reference

1. Strang G. Linear algebra and its applications. 3rd ed. New York: Harcourt Brace Jovanovich, Inc., 1998.

Tsewei Wang, Ph.D.
Associate Professor
Department of Chemical Engineering and Laboratory of Information Technology
Knoxville, TN, 37996-2200
E-mail: twang@utk.edu