Available online at www.sciencedirect.com
ScienceDirect
Science and Justice 48 (2008) 168–177
science&justice
www.elsevier.com/locate/scijus

A MCMC method for resolving two person mixtures

James M. Curran *

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

Received 3 April 2007; received in revised form 10 August 2007; accepted 23 September 2007

Abstract

In this paper a Monte Carlo Markov Chain (MCMC) method for resolving DNA mixtures containing at most four peaks per locus into a major and a minor contributor is presented. Unlike previous methods, this method can provide posterior probability assessments of the most probable genotype *and* a likely range for the mixing proportion. The proposed method is applied to two DNA mixtures where the true genotypes of the contributors are known. The method provides posterior probabilities of the genotypes of the contributes which concord strongly with the known facts.

Keywords: DNA; Bayesian; Mixture; MCMC; Pendulum

1. Introduction

In some recent publications [1–4,13][1] there have been efforts to use quantitative DNA information to aid in the resolution of "two person" DNA mixtures into a major and minor component. This quantitative information comes in the form of either peak height or peak area. It is often the case that the major contributor to a mixed stain will be known to the police and of relatively low interest — for example the victim in a rape case without a suspect. However, even if the police do have a suspect, the true genotype of the minor contributor is unknown. Additionally, there are cases where there is no clear major or minor, or neither the (suspected) major and minor contributors are known. Pendulum [2] is a guided expert system with attempts to determine the genotypes of the major and minor contributors to a "two person" mixed stain. By "two person" it is meant that there are at most four peaks which have been designated as alleles (in whole or in part). This means that the peaks are considered to correspond to alleles of contributors, and hence are not stutters or other PCR artefacts. For example consider the following (idealised) one locus profile.

Assume that the furthermore that the peak areas associated with each of the alleles in Fig. 1 are $\phi_a=990$, $\phi_b=1010$, $\phi_c=260$ and $\phi_d=240$.This best fit is often interpreted as the most likely or most probable genotypic combination. Whilst it is reasonable to believe that there a correspondence between the best fitting combinations and the most probable combinations, there is not a linear relationship. That is, the difference in probability between the "best" fit and the "second best" fit, may be very small, but if we look at the say tenth possibility, then it may be an order of magnitude less probable than the "best". The traditional statistical approach to these types problem is to perform a "goodness of fit" test using the χ^2 distribution, where the probability of the data is evaluated under the assumption that the true expected peak height is known. In practice of course,

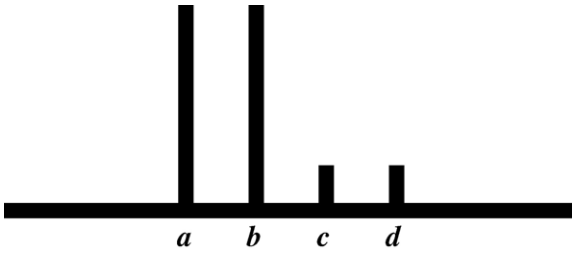

Fig. 1. An idealised two person mixture at one locus.

* Tel.: +64 9 3737 599; fax: +64 9 3737 018.
E-mail address: curran@stat.auckland.ac.nz.

[1] There have been two additional publications which have appeared since the initial submission of this manuscript [5,14].

1355-0306/$ - see front matter © 2007 Forensic Science Society. Published by Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.scijus.2007.09.014

Table 1
Six possible assignments of four alleles to a major and minor contributor

	Major	Minor
G_1	A,B	C,D
G_2	A,C	B,D
G_3	A,D	B,C
G_4	B,C	A,D
G_5	B,D	A,C
G_6	C,D	A,B

we don't know the true peak heights, but we can estimate them by assuming a model.

In this paper use is made of a goodness of fit (GOF) statistic in conjunction with the χ^2 distribution. The χ^2 distribution is used to model the likelihood of the data, as summarised by the GOF statistic, given a mixing proportion and a genotype combination for the major and minor contributors. This likelihood along with some prior assumptions allows a MCMC method to be developed for sampling from the full posterior distribution of the genotypes, mixing proportions and associated hyper-parameters, which in turn will allow probabilistic assessments on the genotype of the major or minor contributor to be given.

The paper of Cowell et al. [5] also offers a Bayesian solution to this problem. The model used in this paper is quite different from that of Cowell et al. Furthermore in this paper, the allele peak areas are not scaled by their repeat number. This was a crude attempt to deal with heterozygous imbalance and is no longer used.

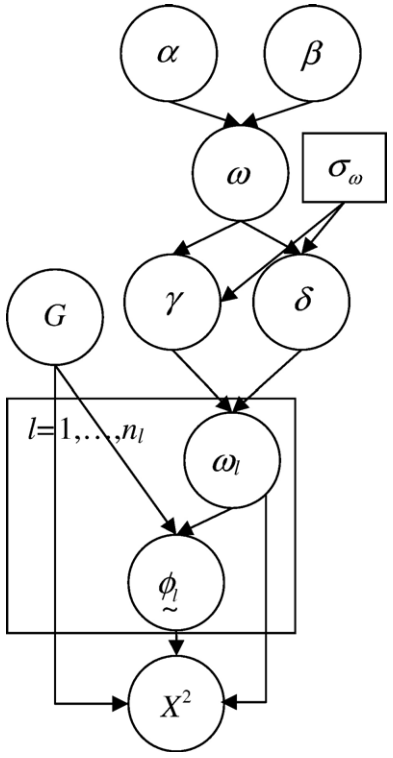

Fig. 2. A graphical model for peak area information.

2. Method

2.1. Modelling peak areas

Assume there are four peaks A, B, C, D with associated areas $\phi_A,\phi_B,\phi_C,\phi_D$ and that the total area is $\phi = \sum_i \phi_i$. There are six different ways to assign the genotype combinations to two people. These are listed in Table 1.

If both contributors contributed an equal amount of DNA to the mixture, then we would expect half the peak area, $\frac{\phi}{2}$, to come from contributor one and the remaining area, $\frac{\phi}{2}$, to come from contributor two. However, if for example contributor one only contributed 25% of the total DNA, then $0.25 \times \phi$ would be attributed to contributor one (and $0.25 \times \frac{\phi}{2}$ of the peak area for each allele from contributor one), and $(1 - 0.25) \times \phi =$

Table 2
13 locus mixture example from Wang et al

			True genotype combination	
Locus	Alleles in the mixture	Allele Peak Area	Victim	Offender
D3S1358	15	1989	15	15
	16	739	16	
	18	1550		18
vWA	15	1318		15
	16	621	16	
	18	793	18	
	19	1200		19
FGA	21	2414	21	21
	22	1461		22
	23	687	23	
D8S1179	12	1431		12
	13	603	13	
	14	560	14	
	16	986		16
D21S11	28	1410		28
	30	1199	30	
	32.2	1506		32.2
D18S51	12	471	12	
	13	386	13	
	17	1181		17
	18	1029		18
D5S818	12	2561	12	12
	13	463	13	
D13S317	11	1607	11	11
	12	834		12
D7S820	8	723		8
	10	1203	10	10
	11	289	11	
D16S539	11	1262		11
	12	515	12	
	13	1253		13
	14	514	14	
THO1	5	944		5
	6	935		6
	8	633	8	
TPOX	8	1257	8	8
	10	984		10
	11	447	11	
CSF1PO	10	482	10	
	11	697		11
	12	617		12