Forensic Science International: Genetics 37 (2018) 29–36

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Modelling the dependence structure of Y-STR haplotypes using graphical models

Mikkel Meyer Andersen[a,*], James Curran[b], Jacob de Zoete[c], Duncan Taylor[d], John Buckleton[e]

[a] Department of Mathematical Sciences, Aalborg University, Skjernvej 4A, DK-9220 Aalborg, Denmark
[b] Department of Statistics, University of Auckland, PB 92019 Auckland, New Zealand
[c] School of Electronic Engineering and Computer Science, Queen Mary University of London, 10 Godward Square, London E1 4FZ, England, United Kingdom
[d] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide SA 5001, Australia
[e] ESR, hPB 92021 Auckland, New Zealand

ARTICLE INFO

Keywords:
Y-chromosome
Forensic genetics
Lineage markers
Weight of evidence
Chow-Liu algorithm
Discrete Laplace method

ABSTRACT

Many methods have been suggested for evaluating the evidential value of a matching Y-chromosomal DNA profile obtained from a biological stain associated with a crime scene and the Y-chromosomal DNA profile of a suspect. Most of these methods are based on estimating the population frequency of the Y-profile. The common independence assumption between loci for autosomal DNA profiles cannot be used for Y-chromosomal DNA profiles. In this paper we reconsider the problem of population frequency estimation by application of Bayesian networks and the Chow-Liu algorithm to model dependencies between loci. We found that the method based on the Chow-Liu algorithm performs almost as well as the discrete Laplace method. We have also made comparisons to the independence model and we have demonstrated once again that the independence method for Y-profiles cannot be supported.

1. Introduction

The common independence assumption between loci for autosomal DNA profiles cannot be used for Y-chromosomal DNA profiles [1]. Many methods have been suggested for evaluating the evidential value of a matching Y-chromosomal DNA profile obtained from a biological stain associated with a crime scene and the Y-chromosomal DNA profile of a suspect. These include the kappa (κ) method [2], the coalescent method [3], discrete Laplace method [4], the generalized Good method [5], and distribution of number of matching males [6].

The results from Andersen and Balding (2017) ("How convincing is a matching Y-chromosome profile?" on the distribution of number of matching males) [6] suggest that for new modern kits, population frequencies may not be the most informative statistic. This is because the population frequency is believed to vary greatly depending on the reference population with matches most likely to occur to an individual only in the tens of meioses away from the suspect.

Nevertheless we reconsider the problem of population frequency estimation in this paper as it contributes to both the understanding of previous methods (such as the discrete Laplace method) and also to the application of Bayesian networks and the Chow-Liu algorithm in forensic science. It may still have some practical value for some practitioners, or with profiles from older kits.

2. Method

2.1. Structure of the joint probability mass function

Constructing a statistical model for Y-STR haplotypes is equivalent to describing the joint probability mass function (jpmf)

$$\Pr(X_1, X_2, \ldots, X_L),$$

where X_i is a random variable representing the allele(s)[1] at the ith locus. Frequencies can then be obtained by plugging in alleles $x_1, x_2, \ldots, x_L$ into this function

$$f(x_1, x_2, \ldots, x_L) = \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_L = x_L),$$

and obtaining the numerical value.

There are many ways to model the jpmf. In this paper, we focus on the structural form of the jpmf. One such way is to assume

* Corresponding author.
E-mail addresses: mikl@math.aau.dk (M.M. Andersen), j.curran@auckland.ac.nz (J. Curran), jcdezoete@gmail.com (J. de Zoete), Duncan.Taylor@sa.gov.au (D. Taylor), John.Buckleton@esr.cri.nz (J. Buckleton).

[1] At most loci a haplotype exhibits only one allele per locus. Occasionally a locus may exhibit 0, 2, or 3 alleles such as for example the duplications at DYS385a/b.

https://doi.org/10.1016/j.fsigen.2018.07.014
Received 7 May 2018; Received in revised form 29 June 2018; Accepted 16 July 2018
Available online 22 July 2018
1872-4973/ © 2018 Elsevier B.V. All rights reserved.

independence between loci to obtain

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_L = x_L) \overset{\text{indep.}}{\approx} \prod_{i=1}^{L} P(X_i = x_i).$$

For Y-STR profiles the assumptions used to obtain this approximation do not hold [1] and this approach cannot be supported. Without assuming independence the (general) multiplication law of probability allows us to express the joint probability function as a product of conditional probabilities. That is, for any events A, B and C the joint probability $\Pr(A, B, C)$ can be written

$$\begin{aligned}\Pr(A, B, C) &= \Pr(A)\Pr(B, C|A)\\ &= \Pr(A)\Pr(B|A)\Pr(C|A, B)\end{aligned}$$

This expression can be extended for any number of events. Applying this to the problem at hand we have

$$\begin{aligned}\Pr(X_1, X_2, \ldots, X_L) &= \Pr(X_1)\Pr(X_2, \ldots, X_L|X_1)\\ &= \Pr(X_1)\Pr(X_2|X_1)\Pr(X_3, \ldots, X_L|X_1, X_2)\\ &= \Pr(X_1)\Pr(X_2|X_1)\Pr(X_3|X_1, X_2)\Pr(X_4, \ldots, X_L|X_1, X_2, X_3).\end{aligned}$$

Factors, such as $\Pr(X_4,\ldots,X_L|\ X_1, X_2, X_3)$, in this expansion which condition on many variables are undesirable because they are difficult to estimate from any reasonable number of empirical data. The difficulty in estimation arises because, as the number of dimensions in the contingency table increases, there are insufficient data to adequately populate the cells of the table. This in turn means that there are many zero entries and that the estimates are highly variable. There is a general trade-off in modeling: more flexible models (in this case high dimensional contingency tables or factors with many conditional variables) describe the observed data better, but are generally poorer when it comes to prediction. This well-known phenomenon is sometimes called over-fitting.

The independence assumption implies that the conditioning does not matter. That is, if X_1 and X_2 are independent, then

$$\Pr(X_2|X_1) = \Pr(X_2)$$

(and so on). If every pair of loci is assumed to be independent then the *product-rule* estimator is obtained. This is over-simplification for Y-STR. The model is now very simple in that only one dimensional tables are used, but the model is now too simple in that it fails to include potentially large dependencies.

We find ourselves stuck between two extremes; assuming full independence results in a model that is too simplistic, and assuming full dependence results in a situation where we do not have sufficient data to fit the model. We opt here for a compromise strategy – namely to find a model somewhere in between these extremes; the model should not be too simple (independence assumption), nor be too complicated (full jpmf where there is no conditional independence between any pair of random variables).

Take the following example with $L = 4$ loci. Let us assume that

$$\Pr(X_3|X_1, X_2) = \Pr(X_3|X_2).$$

That is, we assume that X_3 is independent of X_1 given X_2. Furthermore if we assume that

$$\Pr(X_4|X_1, X_2, X_3) = \Pr(X_4|X_2),$$

i.e. that X_4 is independent of X_1 and X_3 given X_2, then

$$\Pr(X_1, X_2, X_3, X_4) = \Pr(X_1)\Pr(X_2|X_1)\Pr(X_3|X_1, X_2)\Pr(X_4|X_1, X_2, X_3) \quad (1)$$

$$\overset{\text{by assump.}}{=} \Pr(X_1)\Pr(X_2|X_1)\Pr(X_3|X_2)\Pr(X_4|X_2). \quad (2)$$

This is one such example that is somewhere between the full jpmf and the independence assumption. Ideas similar to this have previously been considered briefly without much success [7,8].

The dependence structure of any joint probability distribution may be represented as a graph. Each random variable in the distribution appears as a node in the graph. Dependencies between pairs of random variables are represented as edges between the nodes representing those variables. If the edges in the graph are undirected, then the graph describes a Markov Random Field, and there is considerable literature devoted to this [9]. However, if the edges are directed, and there are no cycles, then we find ourselves in the field of Bayesian networks [10].

Bayesian networks (BNs) (also known as directed acyclic graphs (DAGs), or graphical models) can be used for both illustrating and reasoning with conditional dependencies, and have a strong presence in modern forensic literature [11]. Again, each random variable is drawn as a node in the graph. There is a directed edge from X_i to X_j if X_j is dependent on X_i (i.e. $P(X_j|X_i) \neq P(X_j)$), and X_i is said to be a parent of X_j. The term *acyclic* refers to the fact that the graph has no *cycles*. That is, there must be no *path* – sequence of edges – that allows the dependence to return to X_i. The directedness of the graph, and the acyclic restriction means that the jpmf can be represented by the equation

$$P(X_1, X_2, \ldots, X_L) = \prod_{i=1}^{L} P(X_i|\text{pa}(X_i)),$$

where $\text{pa}(X_i)$ denotes the parents of X_i. This representation says that given its parents, the random variable X_i is conditionally independent of all other variables in the distribution. This is a powerful statement which often allows complex problems to become computationally feasible. Note, that it is not allowed to have directed edges in both directions (both from X_i to X_j and from X_j to X_i) as this introduce a cycle of length two.

Using this notation, the independence model consists of the graph with nodes $X_1, X_2, \ldots, X_L$ and no edges. The full jpmf (with no conditional independence between any pair of nodes) for four variables can be represented by the DAG shown in Fig. 1. The jpmf from (2) can then be represented by the DAG shown in Fig. 2.

Lastly, note that some factorisations and thereby DAGs are equivalent in the sense that they look different but actually represent the same distribution (the same probability mass function). For example, one factorisation of $\Pr(X_1, X_2, X_3, X_4)$ based on Fig. 1 is $\Pr(X_1)\Pr(X_2|X_1)\Pr(X_3|X_1, X_2)\Pr(X_4|X_1, X_2, X_3)$. An equivalent factorisation is $\Pr(X_4)\Pr(X_3|X_4)\Pr(X_2|X_3, X_4)\Pr(X_1|X_2, X_3, X_4)$ (DAG not shown).

2.2. The Chow-Liu method

The Chow-Liu (CL) algorithm [12] provides a structured way to approximate the jpmf by only using first-order dependencies (single columns of two-dimensional tables, so for example $\Pr(X_2|X_1)$, $\Pr(X_3|X_2)$). The Chow-Liu algorithm builds a *tree* using the mutual information for pairs of random variables. The mutual information is a formal measure from information theory which quantifies the degree of

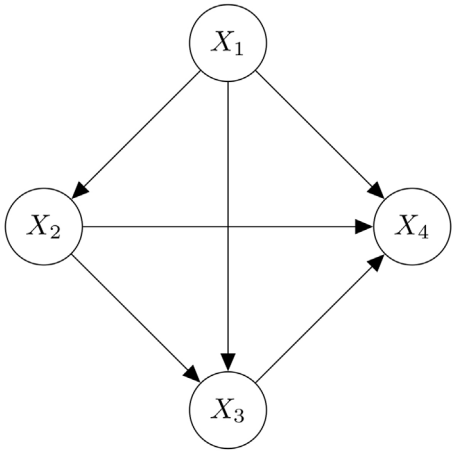

Fig. 1. Representing the full jpmf $\Pr(X_1, X_2, X_3, X_4) = \Pr(X_1)\Pr(X_2|X_1)\Pr(X_3|X_1, X_2)\Pr(X_4|X_1, X_2, X_3)$ (with no conditional independence) as a DAG.

Download English Version:

https://daneshyari.com/en/article/6553175

Download Persian Version:

https://daneshyari.com/article/6553175

Daneshyari.com