Forensic Science International: Genetics 20 (2016) 30–44

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

A probabilistic approach for the interpretation of RNA profiles as cell type evidence

Jacob de Zoete [a,*], James Curran [b], Marjan Sjerps [c]

[a] University of Amsterdam, Korteweg de Vries Instituut voor Wiskunde, Postbus 94248, 1090 GE Amsterdam, The Netherlands
[b] The University of Auckland, The Department of Statistics, Private Bag 92019, Auckland 1142, New Zealand
[c] Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB Den Haag, The Netherlands, Postbus 24044, 2490 AA Den Haag, The Netherlands

ARTICLE INFO

Article history:
Received 25 April 2015
Received in revised form 28 July 2015
Accepted 17 September 2015
Available online 21 September 2015

Keywords:
RNA profiles
Probabilistic approach
Naïve Bayes
Multinomial logistic regression
Cell type
Body fluid

ABSTRACT

DNA profiles can be used as evidence to distinguish between possible donors of a crime stain. In some cases, both the prosecution and the defence claim that the cell material was left by the suspect but they dispute which cell type was left behind. For example, in sexual offense cases the prosecution could claim that the sample contains semen cells where the defence argues that the sample contains skin cells. In these cases, traditional methods (e.g. a phosphatase test) can be used to examine the cell type contained in the sample. However, there are some drawbacks when using these methods. For instance, many of these techniques need to be carried out separately for each cell type and each of them requires part of the available sample, which reduces the amount that can be used for DNA analysis.

Another option is messenger RNA (mRNA) evidence. mRNA expression levels vary among cell types and can be used to make (probability) statements about the cell type(s) present in a sample. Existing methods for the interpretation of RNA profiles as evidence for the presence of certain cell types aim at making categorical statements. Such statements limit the possibility to report the associated uncertainty. Some of these existing methods will be discussed. Most notably, a method based on a 'n/2' scoring rule (Lindenbergh et al. [2]) and a method using marker values and cell type scoring thresholds (Roeder et al. [3]).

From a statistical point of view, a probabilistic approach is the most obvious choice. Two approaches (multinomial logistic regression and naïve Bayes') are suggested. All methods are compared, using two different datasets and several criteria regarding their ability to assess the evidential value of RNA profiles.

We conclude that both the naïve Bayes' method and a method based on multinomial logistic regression, that produces a probabilistic statement as measure of the evidential value, are an important improvement of the existing methods. Besides a better performance, they are flexible and can be adapted to other situations. For example, they could potentially assist in the combination of RNA with DNA evidence.

1. Introduction

In forensic cases, it is often insufficient to only consider the question who donated DNA material. To evaluate what actually happened, information on when or what type of material was donated is needed. Other evidence related to the crime stain could be used to assist in answering these questions. For example, the location and/or the amount of cell material that was found might be relevant. Another option is to include evidence regarding the type of cells (body fluid)[1] in the crime stain. This can be done by using traditional methods (microscopy/immunological/chemical/enzymatic) of cell type investigation for forensic purposes [1]. However, there are some drawbacks when using these methods. For example, many of these techniques need to be carried out separately for each cell type and each of these tests requires part of the available sample, which in turn limits the amount that can be used for DNA analysis.

* Corresponding author.
E-mail addresses: c.dezoete@uva.nl">j.?c.dezoete@uva.nl (J. de Zoete), j.curran@auckland.ac.nz (J. Curran), m.sjerps@nfi.minvenj.nl (M. Sjerps).

[1] In this study, we will use the term cell type instead of the other commonly used term body fluid. Neither of them covers all considered classes, but we stick with one of them for clarity reasons.

http://dx.doi.org/10.1016/j.fsigen.2015.09.007
1872-4973/

Another option is using messenger RNA (mRNA) evidence. mRNA expression levels vary among cell types and, therefore, analyses for the presence of particular mRNAs can be used to make (probabilistic) statements about the cell type(s) present in the crime scene stain. One benefit of using RNA evidence is that it is possible to co-extract RNA and DNA from the same sample. This is an advantage when the amount of biological material available for analysis is limited. In addition, there exists the ability to simultaneously analyse multiple markers and tissue types within one run which saves time and preserves sample.

The interpretation of RNA profiles, in terms of evidential value, can be difficult. For example, the level of expression (peak height) for different markers can differ substantially due to numerous variables, such as the physical condition of the donor, spurious transcription that occurs whenever RNA polymerase binds to DNA, or the cell type of the tested sample. Furthermore, peaks for distinct markers for the same cell type may differ in heights (or may drop out) due to the different expression levels for the specific mRNAs and to the regulation of mRNA by biological, physiological or environmental factors. Moreover, markers that are used as cell type specific markers (markers that only amplify given that the sample contained a specific cell type) infrequently amplify on non-target cell types. So, making a (probabilistic) statement regarding the cell type of the examined crime stain based on marker expression levels seems very problematic. However, one could use present/absent data of the markers (ignoring the corresponding peak heights).

Several methods to interpret the results in this format obtained from mRNA research have been suggested. Amongst these are a $x = n/2$ scoring system that was first suggested in Ref. [2], and a method that combines marker values with a threshold score to distinguish between cell types described in Ref. [3]. Both of these methods aim to make a categorical identification statement regarding the presence/absence of different cell types in a sample. It is common in DNA casework to express the uncertainty regarding the evidential value of a DNA match in the form of a likelihood ratio or a (random) match probability. Both quantities are probabilistic and their value depends on the amount of available information. For instance, a DNA profile utilizing 5 loci (usually) carries less information/evidential strength than a DNA profile on 20 loci. A probabilistic statement used to report the findings will distinguish between these situations, and is preferable in this respect to a categorical identification.

In this paper, we will examine and discuss some of the existing methods to interpret RNA profiles. Furthermore, we suggest two new methods that, unlike the existing methods, result in a probabilistic statement rather than an categorical identification statement. The paper is structured as follows. In Section 2 the existing methods for the interpretation of RNA profiles are discussed, with special attention given to the methods proposed in Refs. [2,3]. In Section 3 the software packages and datasets used in this study are mentioned. In Section 4, a naïve Bayes method based on Bayesian networks and a method based on multinomial logistic regression (MLR) are introduced. In Section 5 we compare the methods on several criteria. The conclusion and discussion can be found in Section 6.

2. Literature overview of RNA interpretation methods

2.1. Multiplex mRNA profiling for the identification of body fluids

Juusola and Ballantyne developed a multiplex reverse transcription-polymerase chain reaction (RT-PCR) method for the identification of the cell types that are commonly encountered in forensic casework analysis, namely blood, saliva, semen, and vaginal secretions [4]. The authors describe two cell type specific genes for each cell type. These are β-spectrin (SPTB) and porphobilinogen deaminase (PBGD) for blood, statherin (STATH) and histatin 3 (HTN3) for saliva, protamine 1 (PRM1) and protamine 2 (PRM2) for semen, and humanbeta-defensin 1 (HBD-1) and mucin 4 (MUC4) for vaginal secretions. The method that is used to designate a cell type as being present in a sample consists of checking whether the selected markers amplify on an unknown sample. If the markers corresponding to one of the cell types amplify, then this cell type is identified as being (one of) the cell type(s) of the sample.

The interpretation of RNA profiles with this method is a very simple one. An identification of cell types can be made given the amplified markers. However, there is a problem. Although the tests within this study did not show any false positives, other studies did. Both Refs. [2,3] (see Tables A15 and A16 in Appendix A) show that some of these markers also amplified on non-target bodily fluids (for example, MUC4 on saliva). A question not addressed by Juusola and Ballantyne (presumably because the study was proof of concept, not an exhaustive approach) is how to interpret the results when only one of the two cell type specific markers amplifies.

2.2. n/2 scoring method

Lindenbergh et al. [2] address the problem where one is unable to make a presence/absence statement when some of the cell type specific markers fail to amplify, or when non-target cell type markers sporadically amplify. This paper describes a procedure that "...accommodates unbiased analysis and interpretation of RNA profiles." This procedure reduces the marker information obtained from a RNA profile to a table which summarizes which cell types were *observed*/*observed and fits*/*sporadically observed and fits*/*not observed*/*sporadically observed*, *not reliable*/*non-specific due to high input* (see Table 1).

The authors suggest performing multiple PCRs and selecting, from the obtained RNA profiles, a set of 'informative' profiles for

Table 1
Example of a table to insert observations from obtained RNA profiles, as suggested in Ref. [2].

Cell type	Observed	Observed and fits	Sporadically observed, and fits	Not observed	Sporadically observed, not reliable	Non-specific due to high input
Case #1						
Blood					x	
Menstrual secretion	x					
Vaginal mucosa		x				
Mucosa		x				
Saliva				x		
Semen				x		
Skin					x	