PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a postprint version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/80458

Please be advised that this information was generated on 2018-01-07 and may be subject to change.
Bayesian networks for victim identification on the basis of DNA profiles

C.J. Bruijning-van Dongen\textsuperscript{a}, K. Slooten\textsuperscript{a}, W. Burgers\textsuperscript{b,\ast}, W. Wiegerinck\textsuperscript{b}

\textsuperscript{a}Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB Den Haag, The Netherlands
\textsuperscript{b}SNN Adaptive Intelligence, Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands

\textbf{ABSTRACT}

We have developed software to improve screening and matching routine for victim identification based on DNA profiles. The software, called Napoleon/Bonaparte, uses Bayesian networks for the analysis. It is designed for effective handling of the identification process in case of a large disaster with many victims and can be applied in the missing person program. In this paper we will describe the Bayesian network approach and we will discuss some of the additional features to handle events with many victims.

2. Bonapartes computational core

Bonaparte’s computational core is designed to calculate the likelihood ratio (LR):

\[ LR = \frac{P(E|H_p)}{P(E|H_d)} \]

where \( P(\cdot|\cdot) \) denotes the conditional probability. \( H_p \) (cq. \( H_d \)) is the hypothesis that missing person MP is equal (cq. not related) to unidentified individual UI; \( E \) consists of DNA profiles of UI, a pedigree \( Ped \) of which MP is a member and at least one DNA profile in \( Ped \). In addition, based on prior odds \( P(H_p)/P(H_d) \), Bonaparte will return posterior odds \( P(H_p|E)/P(H_d|E) \). Bonaparte uses Bayesian networks to compute these quantities.

2.1. Bayesian networks

A Bayesian network is a probability model \( P \) on a directed acyclic graph with \( n \) nodes. For each node \( i \) in the graph, there is (1) a random variable \( X_i \) that can assume a finite number of states \( X \), and (2) a conditional probability distribution \( P(X_i|X_{pa(i)}) \), where \( X_{pa(i)} \) are the states of the variables corresponding to the nodes \( pa(i) \) that point towards \( i \) in the graph. The joint distribution of the Bayesian network is the product of the conditional probability distributions

\[ P(X_1,\ldots,X_n) = \prod_{i=1}^n P(X_i|X_{pa(i)}). \]

Quantities like \( LR \) as in (1) can be efficiently computed by standard Bayesian network algorithms such as the junction tree algorithm [2].
2.2. Bayesian networks for kinship analysis

We now will describe how Bayesian networks are used to compute likelihoods of DNA profiles (currently restricted to short tandem repeat (STR) markers). The likelihoods for each locus are computed independently. We consider a given locus and a pedigree Ped with individuals $i$. Each non-founder individual $i$ has two parents in Ped: father $f(i)$ and mother $m(i)$. Founders are individuals without parents in Ped. The model variables are the alleles of the individuals and the genotypes. The pair of alleles of individual $i$ is denoted as $x_i = (x_i^f, x_i^m)$, with paternal allele $x_i^f$ and maternal allele $x_i^m$. Alleles are only indirectly observable as genotypes $\bar{x}_i$ in which the parental origin is lost. So, $x_i = (a, b)$ and $x_i = (b, a)$ both lead to the same genotype $\bar{x}_i$.

2.2.1. Prior model

In the current version of Bonaparte, the alleles in founders are assumed to be independent. The probability of observing an allele is determined from the population database and prior parameters assumed to be independent. The probability of observing an allele $x$ in the population database is $P(x)$ where $\bar{x}_i$ with $\bar{x}_i = (\bar{x}_i^f, \bar{x}_i^m)$.

The probabilities $P(x_i^f|x_i^m)$ are given by a mutation model $P(b|a)$, which encodes the probability that allele $a$ is transmitted as allele $b$. The current version of Bonaparte uses the uniform mutation model,

\[
P(b|a) = \begin{cases} 
1 - \mu, & \text{if } a = b \\
\mu/(N - 1), & \text{if } a \neq b,
\end{cases}
\]

with $N$ the number of allele-states. The mutation rate $\mu$ can be set by the user.

2.2.2. Transmission model

For a non-founder individual $i$, the allele probability given the alleles of its parents is

\[
P(x_i|x_f^i;x_m^i) = P(x_i^f|x_f^i)P(x_i^m|x_m^i).
\]

where

\[
P(x_i^j|x_i) = \frac{1}{2} \sum_{t=f,m} P(x_i^j|x_i^t) \quad \text{with } t \in \{f, m\}.
\]

The probabilities $P(x_i^j|x_i^t)$ are given by a mutation model $P(b|a)$, which encodes the probability that allele $a$ is transmitted as allele $b$. The current version of Bonaparte uses the uniform mutation model,

\[
P(b|a) = \begin{cases} 
1 - \mu, & \text{if } a = b \\
\mu/(N - 1), & \text{if } a \neq b,
\end{cases}
\]

with $N$ the number of allele-states. The mutation rate $\mu$ can be set by the user.

2.2.3. Observation model

The observation model follows straightforwardly from the definition of the genotype $\bar{x}_i$,

\[
P(\bar{x}_i|y_i) = \begin{cases} 
1, & \text{if } \bar{x}_i = y_i \\
0, & \text{otherwise.}
\end{cases}
\]

In Bonaparte, the observation model actually includes the possibility of allele loss [4]. Due to space limitations the precise model specification will be described elsewhere.

2.2.4. Value abstraction

In a naive implementation, the allele-state space will be prohibitively large and make computation infeasible in larger pedigrees. We apply value abstraction [5]. This is an exact procedure that reduces the allele-state space by abstracting from unobserved allele values and treating them as a single state $z$.

2.2.5. Likelihood computation

With the prior, transmission model, and the observation model, we have a complete description of a Bayesian network. Within this model we can compute $P(E|H_p)$ and $P(E|H_d)$, where $E = \{x_i\}$ represents all observations. $H_p$ and $H_d$ are described in Section 2. The final likelihoods are obtained by multiplication of likelihoods per locus.

3. Features for large scale analysis

Bonaparte has the following additional features to facilitate large scale matching.

- A scalable, multi-user client-server based architecture, see Fig. 1. Computational core and the internal database runs on a server. Via an XML and secure https interfaces, the server connects to other systems.
- Users connect via a web browser, so that there is no need for additional software on the clients.
- Imports/exports data through XML files facilitates data exchange in new environments. In addition, data can be imported from Excel.
- Pedigrees can be imported using XML files, or created and edited using the pedigree editor.
- Matching can be scheduled for large sets of cases, but also performed manually for individual cases.
- Project structure facilitates parallel handling of different (small and/or large scale) cases by multiple users.
- Matching can be direct (e.g. to match UEs with PEUs (Personal Effects)) and indirect (UEs with MPUs, using Bayesian networks).
- A list of LRs is presented to the user. Filter options facilitate browsing through the results.
- All match results are stored in internal database. Rewind to any point in back in time is possible.

A live demo version will be made available on www.dnadvi.nl.

4. Validation

Currently, Napoleon/Bonaparte is under validation. We have defined a set of test-cases with reference LRs’ computed by closed form formulas for the simpler cases, or by brute force summation for the more complex ones. In addition, we have compared the results with those of Familias [6]. Details will be published elsewhere.

Role of funding

This research was partly funded by The Intelligent Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant BSIK03024. The ICIS project and the Ministry of Economic Affairs had no involvement in this research and resulting article.

Conflict of interest

None.

References


