Forensic Science International: Genetics Supplement Series xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Research article

Bayesian networks for victim identification on the basis of DNA profiles

C.J. Bruijning-van Dongen [a], K. Slooten [a], W. Burgers [b,*], W. Wiegerinck [b]

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

ARTICLE INFO

Article history:
Received 7 August 2009
Accepted 11 August 2009

Keywords:
Bayesian networks
DNA profile
Forensic identification
Computer programs

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.

1. Introduction

Bayesian networks are very well suited to model statistical relations of genetic material of relatives in a pedigree [1]. They can be applied in kinship analysis such that whole pedigrees of relatives of the missing persons are used in the screening phase. As a result, correct matches can be found at the costs of much less false hits than with methods which do not take complete pedigree information into account. An additional advantage of a Bayesian network approach is that the analysis tool becomes more transparent and flexible, allowing to incorporate other relevant factors such new models for mutation, size bias corrections, measurement error probabilities, missing data, statistics of more genetic markers, etc.

For these reasons we have developed software for Bayesian network kinship analysis based on DNA profiles. The software is called Napoleon/Bonaparte. It consists of separate parts: Napoleon is designed to transport DNA profiles and pedigree information from NFI database to Bonaparte for screening and matching purpose. Bonaparte is the core for computation, database handling and user interface. Bonaparte communicates with Napoleon (and can communicate with other data transport systems) via standard data interfaces. In this paper we will describe its Bayesian network approach and we will discuss some of its additional features to facilitate handling events with many victims.

2. Bonapartes computational core

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

$$LR = \frac{P(E|H_p)}{P(E|H_d)} \tag{1}$$

where P(.|.) 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_i, 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

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_{pa(i)}). \tag{2}$$

Quantities like *LR* as in (1) can be efficiently computed by standard Bayesian network algorithms such as the junction tree algorithm [2].

* Corresponding author. Tel.: +31 24 3614243; fax: +31 24 3541435.
E-mail addresses: c.bruijning@nfi.minjus.nl (C.J. Bruijning-van Dongen), k.slooten@nfi.minjus.nl (K. Slooten), w.burgers@science.ru.nl (W. Burgers), w.wiegerinck@science.ru.nl (W. Wiegerinck).

1875-1768/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigss.2009.08.024

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 determined by the user. Models have been proposed in which founder alleles are assumed correlated [3]. Their disadvantage is the severe increase in computation time in cases with large pedigrees.

2.2.2. Transmission model

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

$$P(x_i|x_{f(i)}, x_{m(i)}) = P(x_i^f|x_{f(i)})P(x_i^m|x_{m(i)}), \qquad (3)$$

where

$$P(x_i^t|x_{t(i)}) = \frac{1}{2}\sum_{s=f,m} P(x_i^t|x_{f(i)}^s) \quad \text{with } t \in \{f, m\}. \qquad (4)$$

The probabilities $P(x_i^t|x_{t(i)}^s)$ 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,

$$P(b|a) = \begin{cases} 1-\mu & \text{if } a = b \\ \mu/(N-1) & \text{if } a \neq b, \end{cases} \qquad (5)$$

with *N* the number of allele-states. The mutation rate μ can be set by the user.

2.2.3. Observation model

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

$$P(\bar{x}_i|y_i) = \begin{cases} 1 & \text{if } \bar{x}_i = \bar{y}_i \\ 0 & \text{otherwise.} \end{cases} \qquad (6)$$

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 $E = \{\bar{x}_i\}$ represents all observations. H_p and H_d are described in Section 2. The final likelihoods are obtained by multiplication of likelihoods per locus.

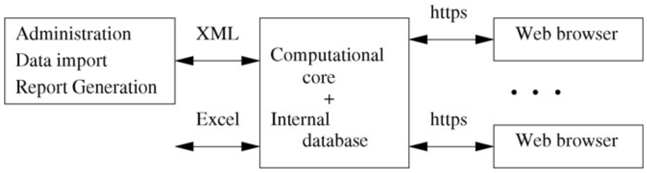

Fig. 1. Bonaparte's basic architecture.

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 *UI*s with *PE*s (Personal Effects)) and indirect (*UI*s with *MP*s, using Bayesian networks).
- A list of *LR*s 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 *LR*'s 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

[1] A. Dawid, et al., Probabilistic expert systems for forensic inference from genetic markers, Scand. J. Stat. (2002) 577–595.

[2] S. Lauritzen, D. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, J. R. Stat. Soc. B Met. (1988) 157–224.

[3] D. Balding, R. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (2–3) (1994) 125–140.

[4] J. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Academic Press, 2005.

[5] N. Friedman, et al., Likelihood computations using value abstraction, in: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, 2000, pp. 192–200.

[6] T. Egeland, et al., Beyond traditional paternity and identification cases: selecting the most probable pedigree, Forensic Sci. Int. 110 (1) (2000) 47–59.

Please cite this article in press as: C.J. Bruijning-van Dongen, et al., Bayesian networks for victim identification on the basis of DNA profiles, Forensic Sci. Int. Gene. Suppl. (2009), doi:10.1016/j.fsigss.2009.08.024