Forensic Science International: Genetics Supplement Series 8 (2022) 131–132

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Bayesian networks for DNA-based kinship analysis: Functionality and validation of the GENis missing person identification module

Ariel Chernomoretz [a,b,*], Franco Marsico [c], Javier Iserte [b], Mariana Herrera Piñero [d], Maria Soledad Escobar [e], Manuel Balparda [e], Gustavo Sibilla [e]

[a] Phys. Department, School of Sciences, University of Buenos Aires/IFIBA CONICET, Argentina
[b] Leloir Institute Foundation, Argentina
[c] UNPAZ, Argentina
[d] Banco Nacional de Datos Geneticos, Argentina
[e] Fundación Sadosky, Argentina

A B S T R A C T

GENis is a recently published open-source multi-tier information system developed to run forensic DNA databases. It relies on a Bayesian Networks framework and it is particularly well suited to efficiently perform large-size queries against databases of missing individuals. In this contribution we present a validation of the missing person identification capabilities of GENis. To that end we introduce *fbnet*, a free-software package written in the R statistical language that implements the complete GENis functionality to perform kinship analysis based on DNA profiles. With the aid of *fbnet*, we could validate likelihood ratios against estimations draw with *Familias* and *forrel* (two well-recognized R packages for kinship quantification) for complex pedigrees provided by the Argentinian reference databank (Banco Nacional de Datos Geneticos, BNDG). We found that our methodological approach presented an excellent performance in terms of accuracy and computation times.

1. Introduction

GENis is a highly customizable system composed of three different modules related to: (a) person identification and analysis of forensic evidence, (b) missing person identification, and (c) disaster victim identification [1]. The missing person identification (MPI) module was specifically developed to run automatic queries on family and missing person databases. For each family, the pedigree structure and available genotypes are integrated into a Bayesian network (BN). These modeling tools serve to represent joint probability distributions in a compact and efficient way, explicitly taking into account the statistical independence between random variables [2]. In our approach to the kinship analysis problem we used BN to infer genotype probability tables for the queried missing person (MP). The availability of these probability tables allows for subsequent rapid likelihood estimations for large MP databases.

2. Methods

We implemented the complete GENis functionality to perform kinship analysis in *fbnet*, an open-source package written in the R statistical language, freely available from CRAN [3]. To speed-up calculations GENis and *fbnet* implemented a truncated version of the canonical stepwise mutational model where probabilities for rare mutations, i.e. more than a given number (L) of steps away from the original allelic value (i.e. the diagonal term), are neglected.

Throughout this contribution we used R packages *Familias*-v.2.4 [4] and *forrel*-v1.3 [5] in order to get reference values against which our LR estimations could be compared. Molecular markers considered for LR calculations are summarized in Sup Table 1.

3. Results

3.1. Precision of the truncated step-wise mutational model

In order to assess for the accuracy of the truncated stepwise mutational model, we considered L = 1, 2 and 4 and estimated with *fbnet* LR values for 1000 simulated profiles (marker setA in Sup Table 1) of unidentified persons (UPs). We considered ensembles for alternative scenarios, H1: UP is MP, and H0: UP is not MP. In Fig. 1. A we showed the precision of these estimations, at a given tolerance level, compared against the software *Familias* LR values for familyFF (Sup Fig. 1.C). It can be appreciated that while L= 1 is a rather rough approximation, LR estimations very rapidly level off at high precision levels for L= 4. For instance, 97.7 % and 99.6 % agreement at a 10 % tolerance level

* Corresponding author at: Phys. Department, School of Sciences, University of Buenos Aires/IFIBA CONICET, Argentina.
E-mail address: ariel@df.uba.ar (A. Chernomoretz).

https://doi.org/10.1016/j.fsigss.2022.10.008
Received 16 September 2022; Accepted 5 October 2022
Available online 6 October 2022

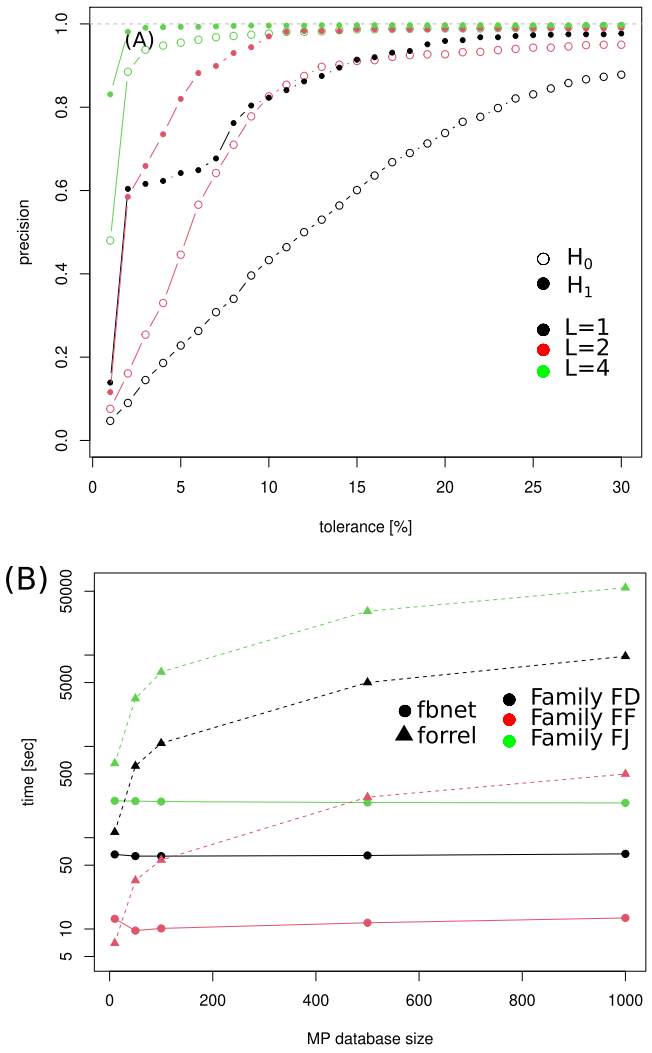

Fig. 1. Panel A. Precision of *fbnet* LR estimation (i.e. fraction of acceptable estimations) for Family FJ (see SupFig. 1) as a function of the considered tolerance percentage. Values for L= 1, 2 and 4 truncated stepwise mutational model are displayed with black, red and green symbols respectively. Solid and empty symbols were used for ensembles of simulated genotypes generated under H1 and H0 hypothesis respectively. Panel B: Running times for *fbnet* and *forrel* LR estimation as a function of the considered ensemble size.

between *fbnet* and *Familias* LR estimations were observed for H0 and H1 simulated UP's respectively.

3.2. Speed-up factor of the bayesian network approach

LR estimations for large databases can be done very efficiently within the GENis/*fbnet* bayesian network approach, as the MP genotype probabilities, conditioned by the available evidence and pedigree relationships, are estimated only once for each family. In this way, *fbnet* presented near constant LR computational times for ensembles of simulated UP profiles. On the contrary, running times for other kinship analysis software, like *Familias* or *forrel* linearly increased with the size of the simulated ensemble. This behavior could be verified in Fig. 1. B, where we showed *fbnet* and *forrel* LR computational times for 1000 UP simulations for families FD, FF and FJ (marker setB in Sup Table 1, pedigrees in Supplementary Fig. 1). Noticeably, large speed-up factors (146x, 37x and 227x for families FD, FF and FJ respectively) were obtained using *fbnet* for simulations involving 1000 UPs (cpu: ryzen 5 2600 ×3.6Ghz, 32 gb ram).

3.3. Systematic comparisons of LR values

We considered 24 familial pedigrees from BNDG and randomly generated the genotype of available family contributors (marker setB in Sup Table 1, pedigrees in Supplementary Fig. 2). For each pedigree we conditionally simulated 1000 MP genotypes and compared LR values against *forrel* estimations considering a stepwise mutational model (we used the truncated L=4 model for *fbnet*). Results were included in Sup. Table 1, where we reported the correlation of log(LR) *fbnet* and *forrel* estimations, the mean and maximum observed percentage differences, and the number of discrepancies at a given tolerance level. An overall high accordance level between reference and *fbnet* estimated values could be observed. In particular, almost perfect correlation values were reported for every analyzed pedigree and mean errors were well below 0.2 % (they remained lesser than 0.1 % for the majority of the analyzed families). In addition, through the 28,000 sampled genotypes only 16 profiles (i.e. less than 0.06 % of cases) presented a discrepancy greater than 5 %.

4. Discussion and conclusions

Overall we found an excellent agreement between GENis/*fbnet* estimations and the corresponding reference LR values for complex pedigrees. In particular we showed that the truncated stepwise mutational model with L= 4 produced highly accurate LR estimations.

Particularly interesting for DNA database applications, our Bayesian network approach to the kinship analysis problem provided an extremely efficient methodology to evaluate LRs for large MP databases. This is so because MP genotype probabilities are estimated only once, at pedigree creation time. Of course, it could be eventually the case that MP genotypes presented rare alleles to be considered for LR estimations. For these uncommon cases a re-calculation step of the probabilities should be executed to accommodate the allelic probability value of the new rare allele. Still, an overall net speed-up gain is expected to occur as these uncommon circumstances would become rarer as the size of the database increases.

Last but not least, in this contribution we introduced the R open-source package *fbnet* [3] in order to share with the community the functionality and main design principles behind GENis MPI module.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigss.2022.10.008.

References

[1] Chernomoretz, et al., GENis, an open-source multi-tier forensic DNA information system, Forensic Sci. Int. (2020), https://doi.org/10.1016/j.fsir.2020.100132.
[2] A. Darwiche, Modeling and Reasoning With Bayesian Networks, Cambridge University Press, 2009.
[3] ⟨https://CRAN.R-project.org/package=fbnet⟩.
[4] Thore Egeland, Daniel Kling, Petter Mostad, Relationship Inference with Familias and R, Elsevier, 2015.
[5] ⟨https://CRAN.R-project.org/package=forrel⟩.