Automatic pedigree reconstruction for genetic studies in isolated populations

ID - 138

Milani Giuseppe[1], Larizza Cristiana[1], Buetti Iwan[2], Nuzzo Angelo[1], Sala Cinzia[2], Toniolo Daniela[2], Bellazzi Riccardo[1]

[1]Department of Computer Science and Systems, University of Pavia, Pavia, Italy

[2]DIBIT - San Raffaele Scientific Institute, Milan, Italy

Motivation

A specific interest of the post-genomic era is the correlation of genotypical and phenotypical information with the aim of investigating the genetic components of complex traits [1]. In the present work we will describe an algorithm for the pedigree reconstruction of the isolated population of the Val Borbera studied within an Italian genetic study started in 2005. The project involves the DIBIT-San Raffaele of Milan and the Laboratory for BioMedical Informatics of the University of Pavia.

Here we present the computational solution adopted to solve the problems encountered in deriving the population pedigree. The algorithm analyzes the municipal and parishes archives and tries to automatically reconstruct the pedigree resorting to record linkage methods, in order to reduce the manual work of the archivists. Until now the pedigree of the population born after 1838 has been completely reconstructed and now we are starting the reconstruction of the pedigree from 1600 until 1838.

Methods

The demographic information coming from municipal and church archives stored in an MS-Access database represent the input data to the algorithm. Each certificate contains information about several individuals that we distinguish into three categories: Registered Persons, Relatives, and Spouses. Registered Persons are identified trough their birth certificate or reported in the electronic municipal demographic archives. Relatives are the persons reported in the certificates as parents or grandfathers of registered persons or spouses. Spouses are the individuals registered in the marriage certificates. The overall data processing covers the following sequence of steps: 1) Data import from the heterogeneous databases into a unified format. The unified structure is filled in by extracting and pre-processing the source data, in order to reduce its imprecision and complete the partial information. Such step represents a sort of standardization process useful to make next steps independent from the input database structure. After the data import into the unified data structure each individual record has the same format. 2) Data cleaning and merge. The goal is to eliminate from the working database possible duplicated individuals, by merging into a single record all the information available. Problems during this step are mainly due to errors or imprecision in the attributes used for record linkage (name, surname, birth dates/age).

3) Pedigree reconstruction. The goal is identifying the parents of each Registered Person.

This phase requires multiple record linkage operations, since the same person can play different roles in different certificates (spouse, father, grandfather, etc).

For each individual of the population the algorithm covers the following steps: 1. Extraction from the data base of the list of registered persons potentially matching with the individuals father/mother. This phase is based on a blocking strategy using the attributes: parents sex, name, surname and birth date/age. If the resulting list is empty, the blocking phase can be reiterated by relaxing some constraints to find possible parents. 2. For every possible parent included in the list, a cross check, based on his/her marriage certificate, is performed in order to confirm the true match.

3. If the marriage data confirm the parental relationship, the corresponding relation parent-child is created. Otherwise the algorithm creates a fictitious individual whose demographical data are derived from the data declared in the sons birth certificate.

The fictitious individual created could be included as potential parent during the next steps of the algorithm. The algorithm provides as output a series of debug reports useful to detect inconsistencies in the data.

Results

We reconstructed the pedigree of 19.139 individuals, which includes a very big family of 10.634 persons. The total number of fictitious individuals created is 12.111, of which only 2.454 are part of the biggest family. It, therefore, covers the 42.7% of the whole population analyzed. The quality of the reconstruction algorithm was measured considering the number of true relations, false relations (number of multiple relations occurring when individuals are related to more than one father and/or mother) and false positives (equal to the number of fictitious individuals).

Considering the number of expected relations as twice the number of the population individuals, we calculated the Recall and Precision of the algorithm [2] respectively equal to about 67.7% and 97.1%. At a first evaluation, it seems that a problem not very frequent is the separation of a unique family into many family groups, while the aggregation of individuals belonging to different families into a single one does not occur. After the analysis of the debug reports, the number of records requiring manual review is about 3398.

Email: milani.giuseppe@gmail.com