International Journal of Epidemiology, 2018, 1034–1039
doi: 10.1093/ije/dyy049
Advance Access Publication Date: 5 April 2018
Software Application Profile

Software Application Profile

An open-source, integrated pedigree data management and visualization tool for genetic epidemiology

Thilina Ranaweera, Enes Makalic, John L Hopper and Adrian Bickerstaffe*

Centre for Epidemiology and Biostatistics, University of Melbourne, Carlton, VIC, Australia

*Corresponding author. University of Melbourne, Melbourne School of Population and Global Health, 203 Bouverie Street, Carlton, VIC 3010, Australia. E-mail: adrianb@unimelb.edu.au

Editorial decision 12 March 2018; Accepted 23 March 2018

Abstract

With advances in genetic epidemiology, increasingly large amounts of pedigree-related information are being collected by family studies, including twin studies. To date, biomedical data management systems that cater for family data have usually done so as part of their standard (non-family-centric) data model. Consequently, data managers with computing expertise are needed to extract family datasets and perform family-centric operations. We present a robust approach to handling large family datasets. Our approach is implemented as a new module which extends the capabilities of The Ark, an open-source web-based biomedical data management tool. Using an algorithm designed by the authors, the pedigree module dynamically infers family relationships for any selected subject (not necessarily the proband). A web interface allows researchers to create, update, delete and navigate parental and twin relationships between subjects, and bulk import/export pedigrees. Consanguineous relationships can be captured, and configurable pedigree visualizations generated. A web services interface provides interoperability.

Key words: Database, pedigree, demography, medical informatics, software

Introduction

With advances in genetic epidemiology, increasingly large amounts of pedigree-related information are being collected by family studies,[1–3] including twin and twin family studies.[4] Most family-centric datasets are stored in study-specific database management systems designed for studies of unrelated individuals, and data for specific research studies are later extracted and analysed as a key component of the research process.

Specialized biomedical data management systems have been developed to manage datasets generated by medical studies.[5,6] The pedigree component of these datasets has usually been incorporated as part of standard (non-family-centric) data management models. Consequently, many researchers depend on data managers to extract family datasets and perform family-centric operations. To answer a specific research question, data managers post-process study datasets to extract pedigree data and

Key Messages

- Large amounts of pedigree-related information are being collected by family studies, including twin studies.
- The structure of pedigree data demands a tailored data management solution with algorithmic support.
- Management and visualization of pedigree data in an extensible manner is non-trivial.
- The pedigree module for The Ark provides a user-friendly, web-based solution to pedigree data modelling and visualization.
- Interoperability with the pedigree module is achieved via a web services interface.

re-assemble them into a semantically correct family representation.

Family studies produce data with pedigree structures that are potentially complex and non-trivial to visualize automatically. To address this challenge, computer scientists have developed tools[7–10] and advanced techniques for pedigree visualization.[11–13] Despite these developments, pedigree data management has tended to remain separate from visualization software, such as the Madeline project.[14] (see section on Visualization).

We present a robust approach to handling large family datasets. Our approach is implemented as a new pedigree module which extends the capabilities of The Ark, a study-oriented biomedical data management system that does not require specialized computing/database skills to operate.[15] The pedigree module provides a flexible mechanism to model, harmonize and visualize large family datasets. Our approach eliminates the need to explicitly specify extended family relationships between study participants; only parent-child and twin relationships need to be specified. The module provides a user-friendly web interface that allows researchers to specify relationships between study subjects. A new pedigree inference algorithm presents the researcher with a dynamically generated pedigree with respect to the subject of interest.

Implementation

The Ark

The Ark is an open-source web-based biomedical data management system programmed using Java technologies. This paper describes a new module for the system, integrating pedigree data modelling and visualization with the rest of The Ark's research data management capabilities, e.g. study, subject, phenotype and laboratory data management.

Once logged in to The Ark, the researcher selects a study, then selects a subject participating in the study. We refer to the selected study as the 'study in context' and the selected subject as the 'subject in context'. This ensures that all subject-related data management operations are applied on a 'study-subject context' basis. This is important because a subject could be a participant in more than one study, and study-specific data about the subject must only appear with respect to the relevant study.

The pedigree module for the Ark

The pedigree module was designed to operate within the existing framework of The Ark's study-subject context-based approach. This approach differs from other common pedigree management systems, e.g. Progeny[16] and Cyrillic,[17] which explicitly delimit families using stored identifiers, and operate by identifying a proband and then defining specific relationship types (e.g. mother, cousin etc.). The Ark pedigree module does not define families using a family identifier; instead, it automatically infers pedigree structures for the subject in context.

Pedigree data input and export

Pedigree structures are formed in the pedigree module by defining parental (mother and father) relationships, and where appropriate, monozygotic/dizygotic twin relationships, for each subject. The web interface enables researchers to define these relationships directly for each subject. The interface provides functionality to search for mothers and fathers based on a unique identifier, first/last name and date of birth. The search algorithm automatically restricts the search for mothers and fathers to female and male subjects, respectively. The interface allows subjects to be specified as members of twin pairs or higher-order multiples only when they share both father and mother relationships. Thus, the pedigree module's user interface enforces the entry of valid data, pre-empting data cleaning to fix rudimentary pedigree errors.

To support data migration from potentially large family studies, The Ark pedigree module is capable of batch importing family datasets using a bulk uploading facility. To import pedigree data, the researcher first converts the data

to The Ark pedigree format (template downloadable), then uploads the data using a web-based wizard that includes a validation step, and concludes with the import running as a background process. To ensure interoperability with different visualization engines, the module includes a data export mechanism capable of exporting a pedigree dataset for the subject in context. The Ark and Madeline pedigree file formats are supported.

The data model

The key challenge in building a pedigree data management system is the modelling and storage of complex family relationships in a general form that is extensible to *nth*-degree relatives. The challenge is simplified if we consider that every family member is genetically linked to other relatives by his/her parent relationships alone. A complete pedigree kinship structure can be constructed using only parental and twin relationships. Therefore, we designed a normalized relational table structure (see Supplementary Figure 1, available as Supplementary data at *IJE* online) to store parent and twin relationships in a format compatible with The Ark's study-subject context-based data model. The key advantage of this data model is that it separates the modelling of pedigree data from the explicit statement of relationships with respect to a particular proband. Additionally, the data model is vastly simplified by considering only parent and twin relationship types.

Pedigree inference

We created the pedigree module's inference algorithm, BloodLine, to operate in conjunction with the system's study-subject context design and the pedigree data model. The objective of BloodLine is to identify dynamically all genetically related relatives of the proband (subject in context) and group them into a single family. BloodLine eliminates the need to manually delimit families and define all subject relationship types.

The BloodLine algorithm operates in two main stages (see Supplementary Figure 2, available as Supplementary data at *IJE* online). The first stage discovers the ancestors of the proband. Beginning with the proband, the algorithm traverses parental relationships and stores these ancestors in memory, after checking them for duplication (which occurs if consanguineous relationships exist). The second stage begins with a list of the ancestors (found in the first stage) and identifies all of their children. The family is defined as the amalgamation of ancestors and their children. Finally, BloodLine labels twin pair relationships in the inferred pedigree using data drawn from the data model.

The computational complexity of BloodLine is linear in terms of the number of family members. Using modest computer hardware, a typical run of the BloodLine algorithm for a pedigree comprising 30 subjects takes less than 1 s.

Pedigree validation

The BloodLine algorithm and Madeline visualization tool rely on clean data to operate correctly and deliver accurate pedigree diagrams. Therefore, it is crucial to validate family relationships before storing them in database tables. We implemented a set of validations at the web interface and batch-import levels. The validations ensure that fathers and mothers are male and female, respectively, and that twins share the same parents. Where date of birth and vital status information is present, validations ensure that a living parent of a subject is older than the subject.

Some family datasets contain consanguineous relationships. Therefore, we incorporated a configuration option to allow or disallow consanguineous relationships on a per-study basis (default: disallow). To minimize input errors, the module displays a confirmation dialogue to the researcher if the setting of a mother/father relationship will lead to consanguinity. This warning is replaced with an error dialogue for studies in which consanguineous relationships are disallowed. The pedigree bulk uploader incorporates equivalent error reporting.

When consanguineous relationships are allowed, The Ark pedigree module restricts the researcher from later toggling the setting to avoid a situation in which consanguineous relationships have been established, but the study is reconfigured to disallow them. The module displays a confirmation dialogue to the researcher when consanguineous relationships are first enabled, alerting him/her of this behaviour. To detect consanguineous relationships, we represent the pedigree as an undirected graph[18] and perform a depth-first search to discover cycles (see Supplementary Figure 3, available as Supplementary data at *IJE* online).

Visualization

A major challenge for implementing the pedigree module is that of pedigree visualization. Rather than implement a new visualization solution, we used Madeline, an open-source pedigree visualization tool. Because Madeline is implemented in C++ and The Ark is based on Java programming technologies, we repackaged Madeline as a native library that is integrated with the pedigree module using customized Java Native Interface classes. This complex programming is hidden from the researcher. To generate a visualization of a pedigree inferred by BloodLine, the researcher simply clicks a button on the web interface. The visualization produced by Madeline is received by the pedigree module and rendered directly by the browser. Visualization is configurable on a per-study basis, with

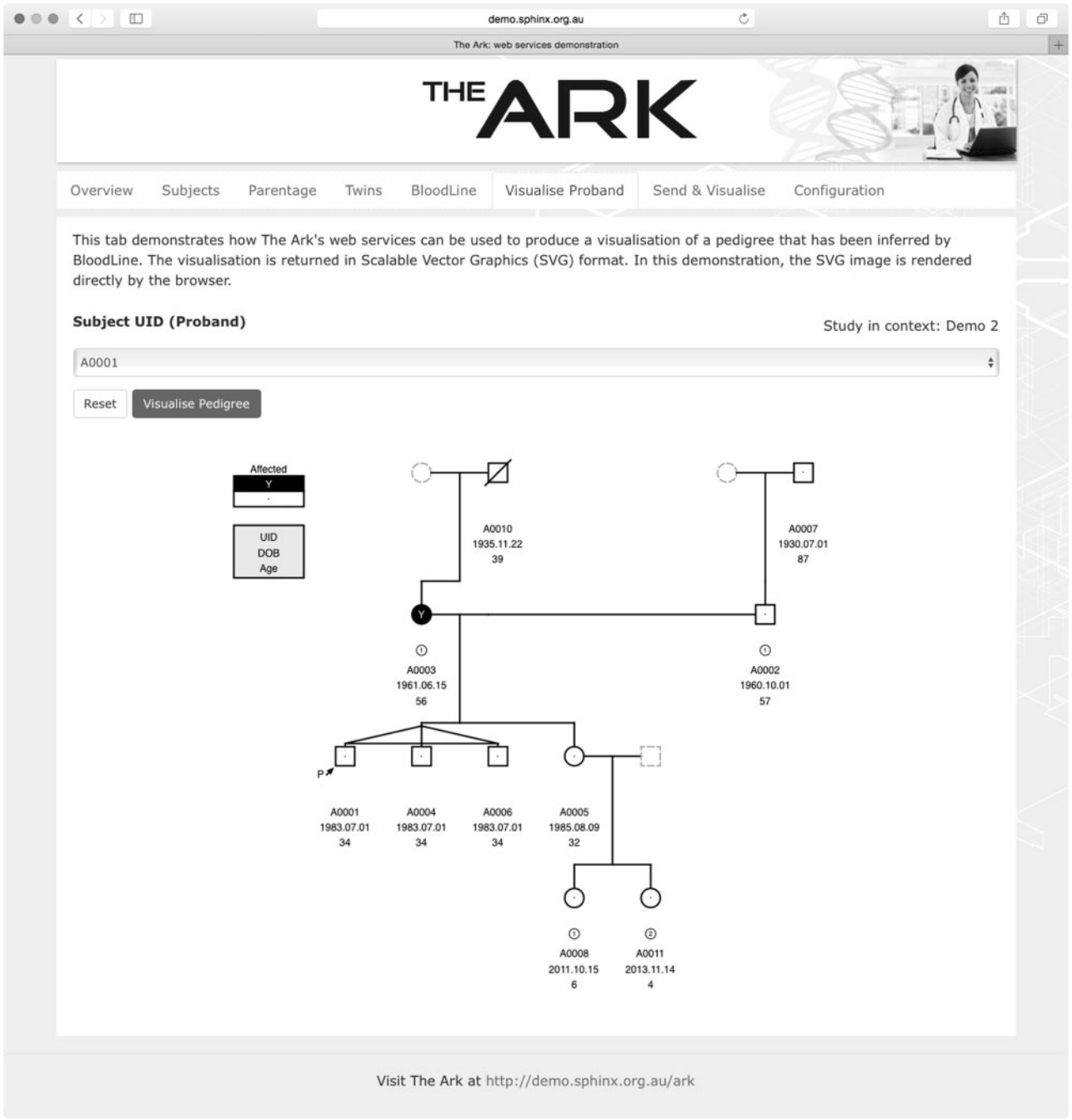

Figure 1. A sample family dataset visualized by a web-based client interoperating with the pedigree module via web services.

options to toggle date of birth and current age annotations, in addition to assigning a researcher-defined custom data field as the source of affected status information (coded as yes/no). Researchers can download the visualization in Portable Network Graphic, Portable Document Format or Scalable Vector Graphic file formats.

Interoperability

A secure web services interface allows external computer systems to interoperate with the pedigree module and its related functionality within The Ark. The interface allows external systems to: add/edit study subjects; create/delete parental and twin relationships between study subjects; retrieve an inferred pedigree membership (with relationship labels) or visualization for a given subject (proband) (see Figure 1); and configure the pedigree module's settings. The interface also provides a means for external systems to send pedigree data to the module and retrieve a corresponding visualization without saving or drawing upon pedigree data in The Ark. This is useful for researchers

who would like to use the pedigree module's configurable visualization capabilities but do not want to use The Ark as their data management system (perhaps having an existing system in place).

Useage

We migrated participant and family data from the Australian Mammographic Density Twins and Sisters study[19] to The Ark. This dataset was formerly managed using a combination of Microsoft SQL Server and Access databases. Using The Ark's bulk uploading capabilities, we imported subject and pedigree data for 1564 families and 5120 subjects, including 544 monozygotic and 339 dizygotic twin pairs. The process brought into The Ark: study-specific information such as unique identifier strings, recruitment date, date last known alive etc.; basic demographic and contact data including date of birth, sex, vital status (alive/deceased), first/last names, residential address etc.; and pedigree data to link subjects within families and establish twin relationships, including zygosity.

A postgraduate without specialized computing expertise took about 3 days to export the data via Microsoft Access and produce files in The Ark's import formats. Using modest server hardware (4 1-GHz processor cores and 8 GB of memory), the pedigree module completed the data import in approximately 5 min.

Conclusion

This paper has presented a novel open-source, web-based approach to pedigree data management and visualization. Our pedigree module for The Ark enables researchers to establish complex family structures using simple parent and twin pair assignments on a per-subject basis. The module is highly scalable, capable of inferring *nth*-degree relatives and modelling very large families efficiently. The size of pedigrees is limited only by database storage.

Future work will focus on implementing an interactive within-browser approach to constructing pedigree data structures and visualizations. To achieve this, we will leverage pedigreejs, a web-based graphical pedigree editor available as an open-source Javascript library.[19] The integration of pedigreejs will allow researchers to build pedigrees within an interactive pedigree diagram setting, including: selecting study subjects; setting mother/father/twin relationships; and directly entering annotations, e.g. affected status values. The researcher's manipulation of a pedigree diagram will be persisted (saved) using the pedigree module's existing data model. The BloodLine algorithm will operate to infer pedigree membership, as before, and feed these data to pedigreejs as the basis for the interactive diagram. The strengths of Madeline, e.g. generalized modelling of consanguineous relationships, mean that it will remain a visualization option available in parallel with pedigreejs.

The source code for the pedigree module, and The Ark as its foundation, can be obtained freely from the website [https://github.com/The-Ark-Informatics/ark/]. The source code can be modified and redistributed under the terms of the GNU GPL v3 licence. Documentation and a publicly accessible demonstration of the software, including an example client to demonstrate the web services interface (see section on Interoperability), can be found at [https://sphinx.org.au/the-ark/]. A pre-configured virtual appliance is also provided.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported by the Twins Research Australia, a national resource in part supported by a Centre of Research Excellence Grant [grant number 1079102 to J.L.H.] from the National Health and Medical Research Council.

Conflict of interest: None declared.

References

1. Lindee S. *Moments of Truth in Genetic Medicine*. Baltimore, MD: Johns Hopkins University Press, 2005.
2. John EM, Hopper JL, Beck JC *et al*. The breast cancer family registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res* 2004;**6**:375–89.
3. Newcomb PA, Baron J, Cotterchio M *et al*. Colon cancer family registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev* 2007;**16**:2331–43.
4. Martin N, Boomsma D, Machin G. A twin-pronged attack on complex traits. *Nat Genet* 1997;**17**:387–92.
5. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;**42**:377–81.
6. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. *Int J Epidemiol* 2017;**46**:1372–78.
7. Brun-Samarcq L, Gallina S, Philippi A, Demenais F, Vaysseix G, Barillot E. CoPE: a collaborative pedigree drawing environment. *Bioinformatics* 1999;**15**:345–46.
8. Loh AM, Wiltshire S, Emery J, Carter KW, Palmer LJ. Celestial3D: a novel method for 3D visualization of familial data. *Bioinformatics* 2008;**24**:1210–11.

9. Paterson T, Graham M, Kennedy J, Law A. VIPER: a visualisation tool for exploring inheritance inconsistencies in genotyped pedigrees. *BMC Bioinformatics* 2012;**13**:S5–16.
10. Carver T, Cunningham AP, de Villiers CB *et al*. pedigreejs: a web-based graphical pedigree editor. *Bioinformatics* 2018;**34**:1069–71.
11. Tuttle C, Nonato LG, Silva CT. PedVis: a structured, space-efficient technique for pedigree visualization. *IEEE Trans Visual Comput Graph* 2010;**16**:1063–72.
12. Santos JM, Santos BS, Dias P, Silva S, Ferreira C. Extending the H-tree layout pedigree: an evaluation. *Proceedings of the 17th International Conference on Information Visualisation (IV); 2013, London, 16–18 July 2013*. London; Institute of Electrical and Electronics Engineers, 2013.
13. Sallaberry A, Fu YC, Ho HC, Ma KL. ContactTrees: a technique for studying personal network data. ArXiv e-prints 2014;abs/1411.0052.
14. Trager EH, Khanna R, Marrs A *et al*. Madeline 2.0 PDE: a new program for local and web-based pedigree drawing. *Bioinformatics* 2007;**23**:1854–56.
15. Bickerstaffe A, Ranaweera T, Endersby T *et al*. The Ark—a customizable web-based data management tool for health and medical research. *Bioinformatics* 2017;**33**:624–26.
16. Genetic Pedigree Software. *Progeny*. 2016. http://www.progenygenetics.com/ (16 September 2016, date last accessed).
17. AP Benson. *About Cyrillic 3*, 2016. —http://www.apbenson.com/about-cyrillic/ (16 September 2016, date last accessed).
18. Sedgewick R, Wayne K. *Algorithms*, 4th edn. Boston, MA: Addison-Wesley, 2011.
19. Nguyen TL, Schmidt DF, Makalic E *et al*. Explaining variance in the cumulus mammographic measures that predict breast cancer risk: a twins and sisters study. *Cancer Epidemiol Biomarkers Prev* 2013;**22**:2395–403.