Proc. Assoc. Advmt. Anim. Breed. Genet: 26: 303-306

IDENTIFICATION OF RELATIONSHIPS AND PEDIGREE SIMPLIFICATION USING GRAPH THEORY

M.H. Ferdosi and D. Johnston

Animal Genetics Breeding Unit*, University of New England Armidale, NSW, 2351 Australia

SUMMARY

Reliable pedigree and genomic data are essential for single-step genomic best linear unbiased prediction. Discrepancies between pedigree and genomic relationships must be identified and resolved before analysis. While parent-progeny relationships are typically verified using genomic data, other relationships, such as those between progeny and grandparents or half-sibs, are often unchecked. The numerator relationship matrix and genomic relationship matrix further obscure the precise nature of individuals' relationships. For example, half-sibs and grandparent-progeny pairs appear identical in the numerator relationship matrix. This article identifies these relationships among individuals in the numerator relationship matrix that are inconsistent with the genomic relationship matrix, applying graph theory to identify the precise type of their relationships and simplify pedigrees to pinpoint and correct errors. The results demonstrate that graph theory algorithms can accurately identify pedigree relationship types and extract any individual's ancestors, descendants, or other relationships from the pedigree.

INTRODUCTION

The single-step genomic best linear unbiased prediction (ssGBLUP) integrates genomic and pedigree information to estimate breeding values accurately (Legarra *et al.* 2014). A critical component of quality control in this process is ensuring alignment between the genomic relationship matrix (GRM) and numerator relationship matrix (NRM). Pedigree errors are common in industry data, and some can be identified and corrected through DNA parentage verification and discovery. However, certain relationships, such as progeny-grandparent or cousin relationships, are largely unchecked (Connors *et al.* 2017). Incomplete pedigrees often lead to GRM values exceeding NRM values. However, when NRM values are significantly higher than GRM values (by more than 0.25, with GRM values near zero), pedigree errors are likely the cause. The scatter plot of NRM versus GRM values highlights these over-specified relationships. However, correcting these issues using additional information, such as herd and multi-sire grouped mating, requires identifying the precise relationship type. For instance, NRM values between individuals and their grandparents are identical to those among half-sibs, i.e., 0.25. The graph theory in mathematics considers structures called graphs, which consist of vertices (nodes) connected by edges (links) to analyse relationships, connectivity, and networks. Graph theory is used in this paper to resolve pedigree inconsistencies and precisely identify the nature of relationships in the pedigree that are not supported by relationships defined by the NRM and GRM. In addition, applying this theory simplifies the pedigree by extracting the ancestors or descents of individuals with issues to facilitate locating and fixing the pedigree errors.

* A joint venture of NSW Department of Primary Industries and Regional Development and the University of New England

MATERIALS AND METHODS

Test pedigree. A small pedigree with 15 individuals was created to test the graph theory approach. The R package visPedigree (Luan 2018) was used to plot this pedigree (Figure 1).

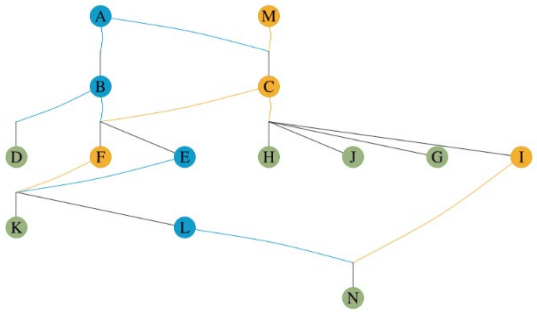

Figure 1. A simple pedigree used to check the graph theory algorithm for the identification of various relationships. Dark blue, golden yellow and olive-green shows, sire, dam and individuals with unknown sex, respectively

Simulated genotype and pedigree. This study utilised QMSim v2 (Sargolzaei and Schenkel 2009) and its first example to generate pedigrees and genotypes, producing a population spanning 10 generations. Genotypes from generations 8 and 10 were used. In contrast, those from generation 9 were omitted to demonstrate how inconsistencies between NRM and GRM can reveal pedigree errors, particularly where parent-progeny relationships cannot be verified genomically. In each generation, 20 males were randomly mated with 400 females, with each female producing two progenies. The genome consisted of 30 chromosomes, each spanning 100 cM and containing 333 markers. The sires of two individuals in generation 8 were altered to simulate pedigree errors.

Genomic and pedigree relationship matrices. The GRM was built using VanRaden first method (VanRaden 2008), and the NRM was built using the pedigree package (Coster 2022) in R for animals with genotypes.

Identification of the relationships among individuals. The package igraph in R (Csardi and Nepusz 2006), was used for graph theory analysis. In this study, we use graph theory to identify Parent-Offspring (PO), Full-sibs (FS), Half-sibs (HS), Progeny-Grandparent (PG), and Cousins (CO) relationships and their extended relationships. The pedigree data typically has three columns for individuals, sires, and dams. The first step in using graph theory to analyse pedigrees was to convert the pedigree into a direct graph, i.e., each parent-offspring relationship was represented as a directed graph where the 'edge' indicates a flow from one node to the other. The directed graph identified groups of full-sibs, half-sibs, grandparents in common, cousins, and more. The closely related individuals can be identified with the neighbourhood function, and since the pedigree is converted to a directed pedigree, the neighbour incoming function can identify the parents, and the outgoing function can identify the progenies of specific individuals. The intersect function was used to find common parents, which can be used to distinguish full sibs from half-sibs. These procedures can be extended with recursive functions to identify the great-grandparents, other ancestral relationships, and even further ones. So, the pedigree of an individual and its ancestral individuals can easily be extracted and visualised.

RESULTS AND DISCUSSION

The test pedigree was used to assess the methodology. The relationships for some individuals are shown in Table 1. In this table, although A—D and H—G relationships were 0.25 and cannot be

differentiated in NRM, the precise type of their relationships has been identified with graph theory, i.e., A—D and H—G were Progeny-Grandparent and Half-sibs, respectively.

Table 1. Description of the relationships in the test pedigree (Figure 1) based on the numerator relationship matrix (NRM)

ID1	ID2	NRM Value	Relationship Type
A	B	0.5	Parent-Offspring
A	D	0.25	Progeny-Grandparent
H	G	0.25	Half-sibs
B	J	0.125	Uncle/Aunt
G	D	0.0625	Cousins
E	F	0.625	Full sibs
A	N	0.375	Progeny-GreatGrandParent

Figure 2 displays all individuals whose NRM values exceeded their GRM values by more than 0.25, highlighting potential pedigree issues. As indicated by the blue and olive colours, most of these individuals were half-sibs or cousins. By applying graph theory to extract ancestral or descendant pedigrees, individuals with incorrect pedigrees can be identified for further investigation. The individual with the most frequent inconsistency between NRM and GRM was the '6109' (the replaced sire with '6104'). Additionally, individuals '8361' and '8362' exhibited the highest differences between NRM and GRM values. Figure 3 presents a simplified pedigree where grandsire' 6104' was replaced with '6109', revealing that '8361' and '6109' stood out. This approach, when combined with additional information such as flock profiles or multiple-sire mating pedigrees, enhances the efficiency and accuracy of pedigree correction. While other methods, such as recursive algorithms, can also identify these relationships, graph theory algorithms are already well-developed and optimised, making them highly feasible for implementation in scripting languages like R (R Core Team 2021).

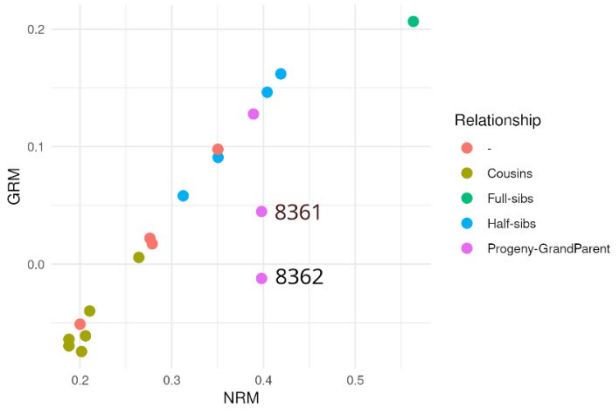

Figure 2. Scatter plot with relationships from the GRM and NRM of simulated data where the pedigree relationships were higher than genomic relationships. The dash symbol represents undetermined relationships. Individuals 8361 and 8362 shows deviation in NRM in Comparison to the GRM

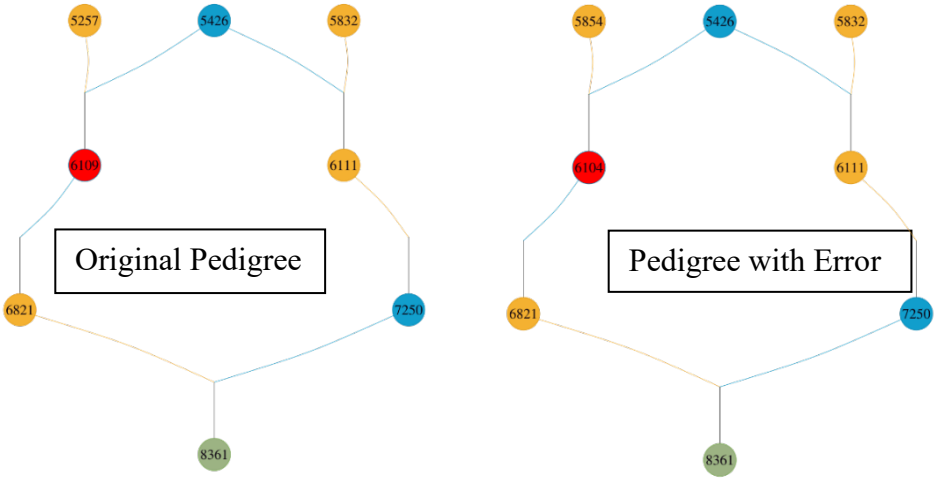

Figure 3 The pedigree shows grandsire '6104' incorrectly assigned to its progeny. Dark blue, golden yellow and olive-green shows, sire, dam and individuals with unknown sex in the pedigree

CONCLUSION

In this study, we applied graph theory to analyse pedigree relationships, specifically identifying cases where the NRM values exceeded the GRM values. Using the same framework, we demonstrated how to simplify the pedigree by isolating a sire's progeny with incorrect pedigree assignment. This approach enables the rapid identification and resolution of pedigree discrepancies using additional data from breed societies and breeders. Correcting pedigree errors and improving the alignment between NRM and GRM values increases the reliability of genomic prediction accuracy. Moreover, the availability of advanced graph theory algorithms and libraries in popular scripting languages further streamlines pedigree analysis, making it more accessible and efficient.

ACKNOWLEDGEMENTS

This study was supported by Meat and Livestock Australia project L.GEN.2204.

REFERENCES

Connors N., Cook J., Girard C., Tier B., Gore K., Johnston D. and Ferdosi M. (2017) *Proc Assoc Advmt Anim Breed Genet.* **22**: 317.

Coster A. (2022) 'pedigree' (*The Comprehensive R Archive Network*).

Csardi G. and Nepusz T. (2006) *InterJournal* **1695**: 1.

Legarra A., Christensen O.F., Aguilar I. and Misztal I. (2014) *Livest Sci* **166**: 54.

Luan S. (2018) 'visPedigree: A package for tidying and drawing animal pedigree'.

R Core Team (2024) 'A Language and Environment for Statistical Computing.' (*R Foundation for Statistical Computing*: Vienna, Austria).

Sargolzaei M. and Schenkel F.S. (2009) *Bioinformatics* **25**: 680.

VanRaden P.M. (2008) *J. Dairy Sci.* **91**: 4414.