Bulletin of Mathematical Biology (2019) 81:313–315
https://doi.org/10.1007/s11538-018-00551-3

Society for Mathematical Biology

PREFACE

Algebraic Methods in Phylogenetics

Marta Casanellas[1] · **John A. Rhodes**[2]

To those outside the field, and even to some focused on empirical applications, phylogenetics may appear to have little to do with algebra. Probability and statistics are clearly important ingredients, as modeling and inferring evolutionary relationships motivate the field. Combinatorics is also an obvious component, as the graph-theoretic notions of trees, and more recently networks, are used to describe the relationships. But where does the algebra arise?

The models used in phylogenetics are necessarily complex. At the simplest, they depend on a tree structure, as well as Markov matrices describing changes in nucleotide sequences along the edges. These two components result in probability distributions given by rather complicated polynomials on the parameters of the models, whose precise form reflects the structure of the tree. Even following standard statistical paradigms for inference, efficient calculation, such as by the Felsenstein pruning algorithm Felsenstein (1981) used in likelihood calculations, depends on understanding this algebraic structure.

But in the late 1980s, the algebraic structure also suggested alternative inference frameworks to some researchers. These included the *phylogenetic invariants* of Cavender and Felsenstein (1987), and of Lake (1987), and the Hadamard transform framework of Hendy and his colleagues Hendy and Penny (1989), Hendy et al. (1994). While this early explicitly algebraic work resulted in a number of interesting mathematical explorations, perhaps culminated in Evans and Speed's invariants work Evans and Speed (1993), it had little impact on practical inference as simulations studies seldom showed good performance Huelsenbeck (1995).

In the early 2000s, works of Allman and Rhodes (2003) and of Sturmfels and Sullivant (2005) revived interest in invariants. Interest in applying algebraic perspectives to statistical problems, especially in computational biology, was exemplified by the

✉ Marta Casanellas
marta.casanellas@upc.edu

✉ John A. Rhodes
j.rhodes@alaska.edu

[1] Department of Mathematics, Universitat Politècnica de Catalunya, Edifici H Despatx 3.23, Avda. Diagonal, 647, 08028 Barcelona, Spain

[2] Department of Mathematics and Statistics, University of Alaska Fairbanks, P.O. Box 756660, Fairbanks, AK 99775-6660, USA

book of Pachter and Sturmfels (2005), which helped draw new researchers to the field. Of course, algebra in statistics has been present from the beginning, such as in Pearson's work Pearson (1894), but as theoretic and computational tools of algebra have developed, they had remained largely outside of the inference toolbox.

In recent years, algebraic methods have been crucial to advances in the theory of phylogenetic inference [in particular, parameter identifiability of phylogenetic models Allman and Rhodes (2009), Allman et al. (2018)] and in new methods of tree reconstruction Fernández-Sánchez and Casanellas (2016), Chifman and Kubatko (2014) that are competitive with traditional frameworks. The tools that have been used draw from algebraic geometry, commutative algebra, computational algebra and algebraic statistics as well as group representation theory and algebraic combinatorics.

The works in this volume showcase the varied directions in which algebra is playing a role in current phylogenetic research.

Algebraic varieties underly the investigation of mixture models by Gross et al., as well as the study of maximum likelihood inference using recently developed numerical algebraic geometry tools by Kosta and Kubjas. Sumner and Woodhams focus more tightly on the modeling of sequence evolution, and the algebraic origin of nicely structured models.

A number of works move beyond simple evolution on a tree. The multispecies coalescent model, which describes the biological process by which gene trees may differ from species trees, is analyzed by Disanto and Rosenberg with tools of algebraic combinatorics. Long and Kubatko also consider this model, greatly weakening the assumptions necessary to justify the invariant-based SVDquartets method of species tree inference. Durden and Sullivant give an identifiability result for a k-mer based distance under the coalescent.

Moving from trees to networks, Kim et al. investigate the impact of admixture on phylogenetic distances and tree reconstruction. Considering both the coalescent and the hybridization, Baños mixes algebraic and combinatorial approaches to show the identifiability of many network features from gene tree data.

Two works highlight other algebraic tools. Terauds and Sumner apply representation theory to study improving distance estimates based on gene order through maximum likelihood. Yoshida et al. bring tropical geometry and algebra to bear on summarizing collections of trees, through a new form of principal component analysis.

Finally, Huber et al.'s work highlights the role of submodularity, a concept appearing widely in combinatorics and optimization, while Wicke and Fischer address open questions on the Shapely value of trees.

Acknowledgements Marta Casanellas is partially funded by AGAUR Project 2017 SGR-932, MINECO/FEDER Projects MTM2015-69135 and MDM-2014-0445. John A. Rhodes is supported by NIH grant R01 GM117590.

References

Allman ES, Degnan JH, Rhodes JA (2018) Split probabilities and species tree inference under the multispecies coalescent model. Bull Math Biol 80(1):64–103

Allman ES, Rhodes JA (2003) Phylogenetic invariants of the general Markov model of sequence mutation. Math Biosci 186:113–144

Allman ES, Rhodes JA (2009) The identifiability of covarion models in phylogenetics. IEEE ACM Trans Comput Biol Bioinform 6:76–88

Cavender JA, Felsenstein J (1987) Invariants of phylogenies in a simple case with discrete states. J Class 4:57–71

Chifman J, Kubatko L (2014) Quartet inference from snp data under the coalescent model. Bioinformatics 30(23):3317–3324

Evans SN, Speed TP (1993) Invariants of some probability models used in phylogenetic inference. Ann Stat 21(1):355–377

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Fernández-Sánchez J, Casanellas M (2016) Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. Syst Biol 65(2):280–291

Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Zool 38:297–309

Hendy MD, Penny D, Steel M (1994) A discrete Fourier analysis for evolutionary trees. Proc Natl Acad Sci 91:3339–3343

Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation. Syst Biol 44:17–48

Lake JA (1987) A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol 4:167–191

Pearson K (1894) Contributions to the mathematical theory of evolution. Philos Trans R Soc Lond A Math Phys Eng Sci 185:71–110

Pachter L, Sturmfels B (2005) Algebraic statistics for computational biology. Cambridge University Press, New York

Sturmfels B, Sullivant S (2005) Toric ideals of phylogenetic invariants. J Comput Biol 12:204–228