New Horizons in Evolutionary Linguistics
Journal of Chinese Linguistics Monograph Series 27 (2017): 67-85

2409-2878/2017/27-04

METRIC LEARNING FOR PHYLOGENETIC INVARIANTS: AN ALGEBRAIC APPROACH TO EVOLUTIONARY TREE CONSTRUCTION

Nicholas Eriksson *Stanford University*

Yao Yuan *Stanford University; Peking University*

ABSTRACT

Construction of phylogenetic trees from observations is a fundamental challenge in both evolutionary biology and evolutionary linguistics. Here we approach the problem from a new perspective by adopting algebraic invariants associated with topological structures of phylogenetic trees. Our key development is based on machine learning to optimize the power of phylogenetic invariants for the construction of phylogenetic tree quartets, the

Acknowledgment The research was conducted while both authors were in Stanford University with an early version of this paper deposited as preprint at http://arxiv.org/abs/q-bio/0703034. This work was supported by NSF grant DMS-0603448 and DARPA grant 1092228. We wish to thank MSRI and the IMA for their hospitality. We also wish to thank E. Allman, M. Drton, D. Ge, J. Löfberg, F. Memoli, L. Pachter, J. Rhodes, and Y. Ye for helpful comments, and especially G. Carlsson for his encouragement and computational resources.

Nicholas Eriksson is currently senior computational biologist at Calico in South San Francisco; [nick.eriksson@gmail.com]. The work of this co-authored paper was mainly done when he was at Department of Statistics, Stanford University. His recent research interests include learning, statistics, genetics, and computational biology, with papers published in *Nature Genetics*, *PLoS genetics*, *BMC Bioinformatics*.

Yao, Yuan 姚远 (corresponding author) has been associate professor of statistics in School of Mathematical Sciences, Peking University; [yuany@math.pku.edu.cn]. He currently is associate professor in the Department of Mathematics at Hong Kong University of Science and Technology; [yuany@ust.hk]. By courtesy he also works with Biomedical Engineering Research Unit, and Department of Computer Science and Engineering at HKUST**.** The work of this co-authored paper was mainly finished when he was at Department of Mathematics, Stanford University, with further revision done at PKU. He is the author of *A Dynamic Theory of Learning: Online Learning and Stochastic Algorithms in Reproducing Kernel Hilbert Spaces* (2008) and numerous articles on statistical ranking via Hodge Theory, online learning, and differential inclusion approach to high dimensional statistics. His recent research focuses on topological and geometric methods for data analysis, statistical machine learning, and applications in computer and life sciences.

building blocks of general evolutionary trees. Phylogenetic invariants are polynomials in the joint probabilities which vanish under a model of evolution on a phylogenetic tree. We give algorithms for selecting a good set of invariants and for learning a metric on this set of invariants which optimally distinguishes the different models. Our learning algorithms involve linear and semidefinite programming on data simulated over a wide range of parameters. We provide extensive tests of the learned metrics on simulated data from phylogenetic trees with four leaves under the Jukes-Cantor and Kimura 3-parameter models of DNA evolution. Our method greatly improves other uses of invariants and is competitive with or better than the popular neighbor-joining method. In particular, we obtain metrics trained on trees with short internal branches which perform much better than neighbor joining on this region of parameter space. These results exhibit potential advantages of applying the new methodology to evolutionary linguistics.

KEYWORDS
Phylogenetic invariants **A**lgebraic statistics **S**emidefinite programming **F**elsenstein zone

1. INTRODUCTION

Evolution is the change in the inherited characteristics of biological populations, or taxonomic groups (taxa), over successive generations (Hall and Hallgrìmsson 2008). Evolutionary processes occur at every level of biological organization, including species, individual organisms and molecules, such as DNA and proteins. Language, as a behavior, has been studied from an evolutionary point of view at least since August Schleicher (1869) who showed the relations among Indo-European languages by drawing an evolutionary tree (Saitou and Nei 1987; Wang 2010). Cavalli-Sforza et al (1988) drew a diagram of the populations of the world, where a tree based on genes is compared against a tree based on languages. Evolutionary linguistics, in spite of the absence of a fossil record, benefits greatly from mapping the linguistic data to evolutionary or phylogenetic trees. In such trees, each leaf node represents an existing language, a branching node sheds light on some ancestral but extinct languages, and lengths of branches correspond to elapsed time.

用于演化树代数不变量的度量结构学习—构造进化树的一种代数方法

Nicholas Eriksson
斯坦福大学

姚远
斯坦福大学；北京大学

提要

从观测数据中构建演化树是生命演化和进化语言学的一个基础问题。本文试图从一个新角度来研究这个问题，即通过演化树的代数不变量来重建演化树的拓扑结构。我们关键的新发展是基于机器学习来优化选择演化树的代数不变量，针对四元演化树发展了一种新的构造方法。演化树代数不变量是指关于联合分布的一种特殊的代数多项式，其在树上的演化模型下恒等于零。本文主要贡献在于发展了一类算法，用于选择一组更好区分不同演化树模型拓扑结构的代数不变量以及相应的度量结构。我们的算法基于给定演化模型下的广泛参数变化而产生的仿真数据，采用线性规划和半正定规划来学习。文中对于 DNA 演化的 Jukes-Cantor 模型和 Kimura 三参数模型进行了广泛的仿真试验测试。试验表明：本文方法整体上同目前广泛使用的Neighbor－Joining 算法相比，具有相似或者更好的性能;特别是对于四元树具有较短内部分支的Felsenstein参数区，本文方法远远超过后者的性能。这些结果展示了将我们的新方法应用于进化语言学研究时可能具有的优势。

关键词

演化树代数不变量　代数统计量　半正定规划　Felsenstein参数区