

QSB/ES 244 - Phylogenetics

August 24, 2017

Instructor: Emily Jane McTavish Office: SE1 284, Email: ejmctavish@ucmerced.edu
Course website: <https://mctavishlab.github.io/GradPhylo> and [catcourses](#) site

Introduction and Course Goals:

This is a graduate course on phylogenetic methods covering theory, statistics and practice. We will study the evolutionary models used to construct phylogenies (evolutionary trees), how these phylogenetic estimates may be used to understand evolutionary processes. We will focus on Maximum Likelihood and Bayesian approaches to inferring phylogenetic trees. We will develop our computational skills to allow us to use High Performance Computing (HPC) resources to perform phylogenetic analyses. The course provides hands-on experience with several important phylogenetic software packages (PAUP*, GARLI, RAxML, RevBayes, BEAST2). By the end of the course students should be able to understand much of the primary literature on modern phylogenetic methods and know how to apply these methods to their own problems.

Textbook

Paul Lewis, Phylogenetics (Sinauer) Forthcoming

Learning outcomes

By the end of this course you will:

- Understand the theory and statistics underpinning phylogenetic analysis.
- Apply this theory to computational phylogenetic analyses and learn to use remote computing resources (MERCED cluster)
- Assess and discuss primary research papers with a phylogenetic focus

Assessment

25% Attendance and participation

15% Paper discussion or research presentation

30% Lab assignments and homeworks

30% Final paper

Attendance and participation

Come to class and discussion prepared, with assigned readings completed. Attend at least 10 out of 15 Friday discussion sections. Participate in paper and research discussions

Paper Discussion or Research Presentation

Friday meetings can be a great opportunity to get feedback on your ongoing work and practice presentation skills.

You must either sign up to present your ongoing work (can be very informal!), or jointly sign up to present a paper (2 people).

When choosing a paper, select one that is:

- RECENT - ask permission for papers more than 5 years old (may be totally fine, but need a bit of a reason).
- DATA or METHODS - Not review papers. They don't provide much to discuss. If the value of a paper is taxon specific, then it won't be broadly interesting. (or the taxon has to be REALLY COOL:)

If you are having trouble selecting a paper, browse through recent issues of Systematic Biology, Molecular Ecology, or Evolution.

Applications papers in other disciplines incorporating phylogeny are welcome!

Paper presentations

- Send out the paper to the Phylogenetics discussion listserve at latest **before class on Tuesday**.
- Presentations should be 10-20 minutes.
- Focus on the methods and figures.
- Set up questions for discussion.

Research presentations

- Send out the title to the Phylogenetics discussion listserve at latest **before class on Tuesday**.
- Presentations should be 20-30 minutes.
- Specify questions or topics you would like advice/comments on.

Lab assignments and Homeworks

Labs and or homeworks will be assigned most weeks, and will be turned in via the catcourses class site. Pacing of the semester will be determined as we go, so a full assignment schedule will not be available in advance. The homework assignments should be worked on your own.

Paper

A final paper will be due at the the time of the scheduled exam, Dec 11, 2017 at 6:30 pm. The term paper will make up 40% of your grade. The project can consist of a new phylogenetic analysis (of your own data or published data) or a paper reviewing a research topic in phylogenetic analysis. Please talk to me about your planned project before spending too much time so that we can agree that the scope is appropriate.

QSB Program Learning Outcomes

- “Background definitions and motivation of fundamental statistical and modeling concepts in the context of quantitative and statistical challenges that exist for certain topics or approaches.”
- “Assessments of student understanding of fundamental statistical and modeling concepts and the ability to conceive, plan, execute and/or interpret the applications of these approaches to research questions.”
- “In-class discussion of primary research papers with major conclusions that depend on quantitative, statistical, or model-dependent approaches.”

ES Program Learning outcomes

- Learning outcomes 1, 2 and 3, fulfill the ES Core Knowledge program learning outcome “Graduates will be knowledgeable, skillful and self-directed in the observation and analysis of environmental systems in terms of their capacity to independently identify important research questions, develop experimental plans, analyze data, and formulate conclusions in the context of a doctoral dissertation”

- Learning outcomes 1 and 2 develop applied statistical and computation skills which fulfill the the ES Career Placement and Advancement program learning outcome “Graduates will find suitable career placement and achieve advancement in government agencies, non-government organizations, private industry, and/or academic teaching and research institutions”
- Learning outcome 3 fulfills the ES “Communication Skills” program learning outcome. “Graduates will be conversant in at least two areas of environmental systems, and be adept at oral, written and visual communication of research results to peers and non-technical decision makers”

Course Policies

1. Classroom interaction. I encourage personal views and critical inquiry based on the material and topics at hand. Equally, I expect that the viewpoints of others will be respected. Consider this course to be valuable practice to engage with your peers through professional communication and scholarly discourse.
2. Special accommodations. Students who need special accommodations are required to submit the form to me in person, preferably outside of class (e.g. office hours) within the first two weeks of the quarter. If you will be requesting academic accommodations, you must first contact the Disability Services (<http://disabilityservices.ucmerced.edu/>) to make arrangements.
3. Academic integrity. The University has established codes concerning proper academic conduct and the consequences resulting from improper behavior. Please be aware of these policies. The official UC Conduct Standards can be found at: <http://studentlife.ucmerced.edu/content/uc-conduct-standards>
4. Life as a UC-Merced Student. Your course facilitators are aware of the many pressures we all face. There are many campus services specifically suited to help you throughout your university career, please take advantage of your resources, including: University academic advising (<http://advising.ucmerced.edu/>), Health Services (<http://health.ucmerced.edu/>), and University Counseling and Psychological Services (<http://counseling.ucmerced.edu/>).

Topics

It is very likely that we will run out of time, and not be able to cover all of these topics; so please, speak up and give me some feedback about what topics are most important to you!

Introduction

The terminology of phylogenetics, rooted vs. unrooted trees, ultrametric vs.

unconstrained, paralogy vs. orthology, lineage sorting, “basal” lineages, crown vs. stem groups

Optimality criteria, search strategies

Exhaustive enumeration, branch-and-bound search, algorithmic methods (star decomposition, stepwise addition, NJ), heuristic search strategies (NNI, SPR, TBR), evolutionary algorithms

Consensus trees, the parsimony criterion

Strict, semi-strict, and majority-rule consensus trees; maximum agreement subtrees; Camin-Sokal, Wagner, Fitch, Dollo, and transversion parsimony; step matrices and generalized parsimony

Bootstrapping, distance methods

Bootstrapping; Distance methods: split decomposition, quartet puzzling, neighbor-joining, least squares criterion, minimum evolution criterion

Substitution models

Transition probability, instantaneous rates, Poisson processes, JC69 model, K2P model, F81 model, F84 model, HKY85 model, GTR model

Maximum likelihood criterion

Likelihood: the probability of data given a model, maximum likelihood estimates (MLEs) of model parameters, likelihood of a tree, likelihood ratio test

Rate heterogeneity

Proportion of invariable sites, discrete gamma, site-specific rates

Codon, amino acid, secondary structure models

Empirical amino acid rate matrices, transition probabilities by exponentiating the rate matrix, RNA stem/loop structure, compensatory substitutions, stem models, nonsynonymous vs. synonymous rates, codon models.

Model selection

Likelihood ratio test (LRT), Akaike Information criterion (AIC), Bayesian Information Criterion (BIC)

Expected number of substitutions

An example derivation for the F81 model

Simulation

How to simulate nucleotide sequence data, and why it’s done

Long branch attraction

Statistical consistency, long branch attraction

Topology tests

ILD, KH, SH, AU and SOWH tests

Bayesian statistics

Conditional/joint probabilities

Bayes rule, prior vs. posterior distributions, probability mass vs. probability density

Markov chain Monte Carlo

Metropolis algorithm, MCMC, mixing, heated chains, Hastings ratio

Priors used in Bayesian phylogenetics

Commonly-used prior distributions: Beta, Gamma, Lognormal, Dirichlet

Prior miscellany

Hierarchical models and hyperpriors, Empirical Bayes, Dirichlet process priors,

MCMC without data

Confidence vs. credible intervals

Frequentist confidence intervals differ from Bayesian credible intervals

Bayesian model selection

Marginal likelihoods and Bayes factors

Discrete morphological characters

DNA sequences vs. morphological characters, Symmetric vs. asymmetric 2-state models, Mk model

Correlated discrete character evolution

Pagel's likelihood ratio test

Correlated continuous character evolution

Felsenstein's independent contrasts

Phylogenetic Generalized Least Squares (PGLS)

Linear regression with correlation structure of residuals determined by the phylogeny

Stochastic character mapping

Introduction to the use of stochastic character mapping for estimating ancestral states and character correlation

Mixture models

Mixture of rate matrices, rjMCMC, heterotachy models, covarion models, Dirichlet process models.

Divergence time estimation

Thorne/Kishino autocorrelated log-normal model; BEAST uncorrelated log-normal model; Yule tree priors; Fossilized Birth-Death Prior

Basics of Coalescent theory

Coalescent theory needed for understanding the multispecies coalescent model

Gene trees within species trees

*BEAST, ASTRAL2, SVDQuartets.

Additional Resources:

Bininda-Emonds, Olaf R.P. (Ed.) 2004. Phylogenetic supertrees - Combining information to reveal the Tree of Life Series: Computational Biology , Vol. 4. Springer. ISBN: 978-1-4020-2329-3 - QH367.5 .P475 2004

Felsenstein, J. (2004) Inferring Phylogenies. Sinauer, Sunderland. Product Code: 0-87893-177-5 - QH83 .F45 2004

Hall, B.J. (2004) Phylogenetic trees made easy: a how-to manual 2nd ed. Sinauer, Sunderland. Product Code: 0-87893-312-3 - QH367.5 .H27 2004

Hennig, W. et al. (1999) Phylogenetic Systematics. University of Illinois Press. ISBN 0-252-06814-9. - QL351 .H413 1979

Hillis DM, Moritz C, Mable BK, Graur D. Molecular systematics. Sinauer Associates Sunderland, MA; 1996]Hillis DM, Moritz C, Mable BK, Graur D. Molecular systematics. Sinauer Associates Sunderland, MA; 1996

Page, R.D.M., editor (2002) Tangled Trees: Phylogeny, Cospeciation, and Coevolution. University of Chicago Press. ISBN: 978-0-226-64467-7 (ISBN-10: 0-

226-64467-7) - QH367.5 .T36 2003

Scotland, R. & R.T. Pennington (2000) Homology and Systematics: Coding Characters for Phylogenetic Analysis. Systematics Association Special Volumes Volume: 58. ISBN: 9780748409204 - QH367.5 .H65 2000

Semple, C. Steele, M.A. Phylogenetics Oxford ; New York : Oxford University Press, 2003.

Steel M. Phylogeny: Discrete and random processes in evolution . SIAM; 2016