Molecular Phylogenetics ECEV 36400
Spring 2016
W 100-400
Joe Thornton, Professor, joet1@uchicago.edu
Office hours M 200-300 CDT or by appointment
Office: W504 GCIS
Course website: on Chalk (UChicago)
TA: Roberto Márquez rmarquezp@uchicago.edu

Description

All comparative analysis in biology takes place in the context of an understanding of phylogenetic relationships, whether explicit or implicit. The advent of sequencing technology has provided a wealth of data for recovering phylogenetic relationships and tracing the genetic changes that have occurred over evolutionary time. Methods for inferring phylogeny and reconstructing the evolutionary process from these data have also proliferated.

In this course you will learn the fundamental concepts and current techniques for inferring evolutionary relationships from gene sequence data and testing hypotheses about molecular evolution using those phylogenies as a scaffold. We will cover the theoretical basis of phylogenetic methods in evolutionary and statistical theory, including the justifications and applications for maximum parsimony, evolutionary distance, maximum likelihood, and Bayesian analysis. We will discuss cases in which these methods have been applied to understand the evolution of taxa, genes, and diseases.

One thing that makes phylogenetics a stimulating field is the intellectual debate about the validity and performance of various methods. By the end of the course, you should be able to critically evaluate the use of phylogenetic techniques in your own work and in papers you read. You should also have a working knowledge of the most important techniques and computer programs for phylogenetic analysis.

Prerequisites

This course is open to graduate students. No prior experience with phylogenetics is required. Experience with evolution, statistics and probability is expected. The math will not be too heavy: algebra, a bit of calculus, and a little basic linear algebra (matrix operations). Facility with the command line is necessary.

Class format

This is a graduate course. The way to gain advanced understanding in a scientific field is to wrestle directly with the field's seminal concepts and their implementation in case studies from the literature.

The course will consist of interactive lectures, discussions of primary readings from the literature, and problem sets. How productive and interesting the class will be is determined in significant part by how much everybody participates.

You are primarily responsible for your own learning: you have to work through the readings and problems until you really understand them, ask questions in class about anything you couldn't figure out, and engage in discussions about the meaning, importance, and validity of each paper.

There are many issues that will only be covered in class, so your regular attendance is essential and required. This is not the kind of class in which you can learn or succeed by staying home and reading the book.

Readings

The bulk of the course readings will be from the primary literature, including both methodological discussions and empirical analyses that apply the methods, illustrating their potential advantages and disadvantages. Some are easier than others, but you should, once you have been introduced to the terms in class, be able to absorb all of them. We will work through the papers in class, so you will have a chance to bring up any issues that are not clear to you, as well as points that you think are worth discussing or challenging.

All papers are available on the course Blackboard/Chalk site as pdfs for download.

Readings must be completed by the time of the class meeting in which they will be discussed. Class participation is essential to an intellectually lively course; it will also contribute significantly to your grade.

Two textbooks that are likely to be helpful for understanding methodology and algorithms are *Inferring Phylogenies*, by J. Felsenstein and *Tree Thinking* by D. Baum and S. Smith.

Homework

Several brief problem sets will illustrate the concepts and techniques we have learned in class. Some are logical exercises; for others, you will use computer software to phylogenetically analyze data I will supply. The assignments will include detailed instructions about how to solve these problems. Answer keys will be placed on the website.

The computer programs are generally relatively easy to use, and the manuals will be available on the course website. I encourage you to work with and help each other on the problems, and you can also ask for my assistance. I will be available during office hours for consultation and help, and you can make an appointment to see me at another time if necessary. For the weeks in which problem sets involve computer work, I will hold my office hours in the biology computer lab. Problem sets are due one week after the lecture in which the associated material is taught.

Computer programs

The major programs to be used in this course are RaxML, PAUP*, MrBayes, Phyml, Mesquite, PAML, ClustalX, MUSCLE, Figtree, and Modeltest. All but PAUP are available for free.

There are numerous additional useful programs that will be mentioned, but there won't be time to teach their use in the course. Links will be available on the website to Garli, Hyphy, and others.

Presentations

Each student will be required to host the discussion of one paper during the quarter. The host is responsible for leading the discussion and presentation of a paper in an interactive **chalk talk.** We believe that this format will be more rewarding and engaging for all students than the typical electronic presentation.

The host must be prepared to communicate the important aspects of the paper -- including its purpose, strategy, experimental/analytical methods, results, and conclusions -- using only notes and drawings to be made on the classroom whiteboard during the discussion. It will be impossible to recreate every figure from a paper in its original form. Rather, the host's goal should be to identify the essential aspects of each analysis and to display them in a simplified graphic format that can form the basis for a group discussion and assessment. If you are unsure as to why a particular paper is selected for discussion, and you wish to confirm that you are picking up on the proper thematic content, please reach out to the TA prior to your discussion section.

The presenter alone is not solely responsible for the discussion. Everyone should participate in the conversation to reach mutual understanding of the paper and to assess its contribution to the field. The professors may also spontaneously ask other students in the class to explain a result or analysis, so please be prepared to do so.

Exam
There will be one take-home exam -- a late midterm -- in multiple short essay format. This is an open book exam but, unlike the problems, it is absolutely not collaborative. You are to work on this exam entirely on your own, without any discussion with other students. There is no final exam.

Paper
Each student is responsible for a final paper. This is a brief proposal with preliminary data, written in the style of an NSF doctoral dissertation improvement grant, for a phylogenetics-based research project of your choice. Your paper should be no more than 10 double-spaced pages not included figures and references. The paper will be in the form of a grant proposal with preliminary data and analyses. It should have the following structure (with very rough page guidelines provided):

1) Specific aims (1 page). A summary of the importance and overall goal of the project and the specific aims that will be pursued and the hypotheses that will be tested. (There are typically 2-4 such aims in a project; each can have subparts if necessary.)
2) Background and justification (1-2 pages). The purpose of this section is to justify your project via an explanation of its importance to the relevant scientific fields and a critical evaluation of the published literature on your topic. The goal is to explain clearly why the question is important and the approach or model system being studied appropriate, what is known about it, how we know it, what questions are currently unresolved, and why the approach you propose will contribute to our knowledge.
3) Preliminary data/analysis (3-4 pages, plus figures). This section should include detailed description of the methods you have used to make preliminary progress towards your aims and testing your hypotheses, and the justification for using this approach.
4) Proposed research (~3-5 pages per aim, depending on the intricacy and number of aims). You should clearly state the questions you seek to answer, describe and justify your general strategy, and describe and justify the specific kinds of data you would gather and what techniques you will use to analyze it. Most important is how you will interpret the data. Be sure to explicitly explain what you will conclude given the different possible outcomes of your analysis. What you propose should be practical, given the kinds of tools that you have learned about in the course. If you have not done so in the preceding sections, be sure to justify the choices you make about the tools and types of analyses you will use.
5) Figures and References—not included in page limit.

At the end of week 5, you must turn in at least one substantial idea for your proposal, summarized in a short one paragraph excerpt. If you would prefer to turn in multiple ideas and receive feedback from the instructor and TA, you are free to do so. As always, we are all open for consultation at any point in the process of generating themes for your project.

A one-page prospectus of specific aims is due by Tuesday of week 7. It should describe the problem you will be addressing, provide a very brief justification of its importance, and list the aims (in 2-3 sentences each).

The final paper is due by 5pm on Thursday of finals week.

The paper is an exercise in scientific thinking and in mastering a subtopic relevant to the course. Key aspects are 1) how clearly and persuasively you argue for the importance of your topic and methodology in its scholarly context, 2) the clarity and testability of your hypotheses, 3) the suitability of the analyses you propose for testing these hypotheses and the rigor of your discussion of how you will interpret the data, and 4) the design and rigor of your preliminary analysis and interpretation of the data it produces.

The clarity of your writing and the organization of your ideas is crucial. Scientists use the written word—in grant proposals and manuscripts—to convince each other of the value of their work. As graduate students, honing your skills in this area should be a top priority.

Paper workshops

One of the best ways to improve one's scientific thinking and writing is to read others' works-in-progress and comment on them; the feedback helps the author, too, of course. In week 8, we will assign members of the class into workshop groups of 3-4 students each. You will meet in week 8 with your group to discuss your Specific Aims page. You will also meet by Thursday of week 10 in these groups to review entire drafts of each others' proposals. Everyone will read the papers of the other members of the group in advance and bring comments about what you liked and constructive suggestions for how it might be improved, structured through a provided form.

For each workshop, in addition to sharing your comments with your group, you must email a brief written review to the course TA by 10 am of the day the workshop is scheduled.

Academic integrity

I have a zero tolerance policy for breaches of academic integrity. I assume that you do too.

Grades

This is a pass-fail course. There will be no letter grades. You are graduate students who are presumably in school because you want to become scientists and contribute to your chosen field; you are taking this course presumably because you believe that learning this material will facilitate your development as a scientist. That should be sufficient motivation to do your best work. Instead of grades, we will provide extensive feedback on your work to assist you in these efforts. Passing requires doing work at a level of quality and commitment appropriate to graduate students at the University of Chicago.

No electronic devices

In an effort to create a more focused and interactive learning environment, use of laptops and tablets will not be permitted during the lectures. We understand that such devices can be useful for note-taking, but they inevitably detract from the discussion-based environment we seek to create. Thus, we ask you to put away your devices during class, to take your notes on paper, and to bring printed copies of the papers assigned to each session.

READING LIST

Books

Felsenstein J. *Inferring Phylogenies*. Sinauer, 2003.

Baum D and Smith S. *Tree Thinking: An Introduction to Phylogenetic Biology.* Roberts and Company 2013.

Also extremely useful for parsimony and cladistics: Kitching IJ, Forey PL, Humphries CH, and Williams DM, eds. *Cladistics, Second Edition: The Theory and Practice of Parsimony Analysis.* Oxford U. Press, 1998.

Papers

1. Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lakes JA. Evidence of a clade of nematodes, arthropods, and other moulting animals. Nature 387:489-492 (1997).
2. Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Systematic Biology 60:685-99 (2011).
3. Baum, DA, Smith, SD, Donovan, SS. Evolution. The tree-thinking challenge. Science 310: 979-80 (2005).
4. Blair, JE, Ikeo, K, Gojobori, T, Hedges, SB. The evolutionary position of nematodes. BMC Evolutionary Biology 2, 7 (2002).
5. Bourlat, SJ et al. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444, 85-8 (2006).
6. Brinkmann, H, van der Giezen, M, Zhou, Y, Poncelin de Raucourt, G, Philippe, H. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Systematic Biology 54, 743-57 (2005).
7. Cunningham CW, Zhu H, Hillis DM. Best-fit maximum likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 52:978-987 (1998).
8. de Queiroz, A., M. J. Donoghue, and J. Kim. Separate versus combined analysis of phylogenetic evidence. Annual Review of Ecology and Systematics 26: 657-681 (1995).
9. Delsuc F, Phillips MJ, Penny D. Comment on “Hexapod origins: monophyletic or paraphyletic.” Science 301:1482 (2003).
10. Delsuc, F, Brinkmann, H, Chourrout, D, Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439: 965-8 (2006).
11. Delsuc, F, Brinkmann, H, Philippe, H. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6: 361-75 (2005).
12. Dunn CW, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-748 (2008).
13. Edgar, RC, Batzoglou, S. Multiple sequence alignment. Current Opinion in Structural Biology 16: 368-73 (2006).
14. Farris J.S. The logical basis of phylogenetic analysis. In: Advances in Cladistics (Edited by Platnick N.I. & Funk V.A.), pp. 1-36 (1983). Columbia U. Press, New York.
15. Farris, JS, M. Källersjö, AG Kluge and C. Bult. Testing significance of incongruence. Cladistics

10:315-320 (1994).

16. Felsenstein J. Confidence limits on phylogenies. Evolution 39:783-781 (1985).

17. Foster PG. The Idiot's Guide to the Zen of Likelihood in a Nutshell in Seven Days for Dummies, Unleashed. Web publication, (2001).

18. Gatesy, J, Hayashi, C, Cronin, MA, Arctander, P. Evidence from milk casein genes that cetaceans are close relatives of hippopotamid artiodactyls. Molecular Biology and Evolution 13: 954-63 (1996).

19. Hanson-Smith V, Kolaczkowski B, Thornton JW. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Molecular Biology and Evolution. 27:1988-99 (2010).

20. Hillis DM, Huelsenbeck JP, Cunningham CW. Application and accuracy of molecular phylogenetics. Science 264: 671-677 (1994).

21. Hillis, D. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Systematic Biology 47: 3-8 (1998).

22. Hillis, D. M. and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic. Biology. 42:182-192.

23. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:310-2314 (2001).

24. Huelsenbeck JP. Systematic bias in phylogenetic analysis: is the strepsiptera problem solved? Systematic Biology 47: 519-537 (1998).

25. Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. Potential applications and pitfalls of Bayesian inference of phylogeny. Systematic Biology 51:673-688 (2002).

26. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? Trends in Genetics 22:225-31 (2006).

27. Kolaczkowski B, Thornton JW. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980-984 (2004).

28. Kolaczkowski, B., Thornton, JW. Long-Branch Attraction Bias and Inconsistency in Bayesian Phylogenetics. PLoS ONE 4: e7891 (2009). With correction.

29. Kolaczkowski, B., Thornton, JW. Effects of branch length uncertainty on Bayesian posterior probabilities for phylogenetic hypotheses. Molecular Biology and Evolution 24, 2108-2118 (2007).

30. Kolaczkowski, B., Thornton, JW. A mixed model of heterotachy improves phylogenetic accuracy. Molecular Biology and Evolution 25:1054-1066 (2008).

31. Lakner C, Holder MT, Goldman N, Naylor GJ. What's in a likelihood? Simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. Systematic Biology 60:161-74 (2011).

32. Lartillot N, Brinkmann H, Philippe H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evolutionary BiologySuppl 1: S4 (2007).

33. Le SQ, Dang CC, Gascuel O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Molecular Biology and Evolution 29: 2921-36 (2012).

34. Le SQ, Gascuel O. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Systematic Biology 59:277-87 (2010).

35. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Molecular Biology and Evolution 25:1307-20 (2008).

36. Medina M, Collins AG, Silberman JD, Sogin ML. Evaluating hypothesis of basal animal phylogeny using complete sequences of large and small subunit rRNA. Proceedings of the National Academy of Sciences of the USA 98: 9707-9712 (2001).

37. Mindell, DP, Thacker, CE. Rates of molecular evolution: phylogenetic issues and applications. Annual. Reviews in Ecology and Systematics 27: 279-303 (1996).

38. Miyamaoto MM, Koop BF, Slightom JL, Goodman M, Tennant MR. Molecular systematics of higher primates: genalogical relations and classifications. Proceedings of the National Acaddemy of Sciences of the USA 85:7627-7631 (1988).

39. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348-51 (2001).

40. Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F. Hexapod origins: monophyletic or paraphyloetic. Science 299:1987-1999 (2003).

41. Nikaido M, Rooney AP, Okada N. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. Proceedings of the National Academy of Sciences of the USA 96:10261-6 (1999).

42. Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature. 470:255-8 (2011).

43. Philippe, H, Lartillot, N, Brinkmann, H. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Molecular Biology and Evolution 22: 1246-53 (2005).

44. Phillippe et al., Phylogenomics revives traditional views on deep animal relationships. Current Biology 19: 706-712 (2009).

45. Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic Biology 56:389-99 (2007).

46. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804 (2003).

47. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327-31 (2013).

48. Sanderson MJ, Kim J. Parametric phylogenetics? Syst Biol 49(4):817-828, 2001

49. Shoemaker JS, Painter IS, Weir BS. Bayesian statistics in genetics: a guide for the uninitiated. Trends in Genetics 15:354-358 (1999).

50. Siddall, ME, Kluge, A. Letter to the editor: the relative merits of likelihood and parsimony. Cladistics 15: 439-440 (1999).

51. Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SC, Rouse GW, Giribet G, Dunn CW. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature.480:364-7 (2011). Erratum in: Nature. 2013 Jan 31;493(7434):708.

52. Thompson, JD, Higgins, DG, Gibson, TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673-80 (1994).

53. Thornton JW, DeSalle R. Gene family evolution and homology: genomics meets phylogenetics. Annual Reviews in Genomics and Human Genetics 1:41-73 (2000).

54. Wenzel JW, Siddall ME. Noise. Cladistics 15, 51–64 (1999).

55. Wheeler WC, Gatesy J, DeSalle R. Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Molecular Phylogenetics and Evolution 4:1-9 (1995).

56. Whiting M. Long-branch distraction and the strepsiptera. Systematic Biology 47:134-138 (1998).

57. Whiting, MF, Bradler, S, Maxwell, T. Loss and recovery of wings in stick insects. Nature 421: 264-267 (2003).

58. Whiting, MF, Carpenter, JC, Wheeler, QD, Wheeler, WC. The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Systematic Biology 46: 1-68 (1997).

59. Yang Z, Kumar S, Nei M. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics. 141:1641-50 (1995).

60. Yokoyama S, Tada T, Zhang H, Britt L. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proceedings of the National Acaddemy of Sciences of the USA. 105:13480-5 (2008).

61. Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution 19:908-17 (2002).

62. Zanis MJ, Soltis DE, Soltis PS, Mathews S, Donoghue MJ. The root of the angiosperms revisited. Proceedings of the National Academy of Sciences of the USA 99: 6848–6853 (2002).

63. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular Biology and Evolution 22: 2472-2479 (2005).

64. Zhuang H, Chien MS, Matsunami H. Dynamic functional evolution of an odorant receptor for sex-steroid-derived odors in primates. Proceedings of the National Acaddemy of Sciences of the USA 106:21247-51 (2009).