CS 685-001/PPA 784-003/STA 695-001 Fall 2009

- Course Number: CS 685 - 001/STA 695 - 001/PPA 784 - 003
- Title: Phylogenetic Analysis and Molecular Evolution
- Semester: Fall, 2009
- Time and Date: Tuesdays and Thursdays, 3:30pm to 4:45pm
- Room: Whitehall Classroom Bldg-Rm.233-CB with some meetings in RGAN 103
- Instructors: Drs R. Yoshida (STA 695), C. Schardl (PPA 784), and J. Jaromczyk (CS 685)
- Emails: ruriko.yoshida@uky.edu (Dr Yoshida), chris.schardl@uky.edu (Dr Schardl), jurek@cs.uky.edu (Dr Jaromczyk).
- Offices: 805A POT (Dr Yoshida), 201F Plant Sciences Bldg. (Dr Schardl), 775 Anderson Hall (Dr Jaromczyk).
- Office Hours: After class or by appointment (the location of the OH is TBA).
- Website: `www.cophylogeny.net/courses/F09`
- TA: Dr David Haws: dchaws@gmail.com
- Open computer Lab: RGAN 103 – Tuesdays from 5 PM to 7 PM; CB 307 – Tue/Thursdays from 5:00 PM to 5:30 PM and Mondays from 12:00 PM to 1:00 PM.

Course Description

One of the key tasks to study molecular sequence data is phylogenetics, the study of evolutionary relatedness among various groups of organisms from molecular sequence data. Finding out evolutionary relatedness among various groups of organisms and the reconstruction of the ancestral relationships have applications including predicting evolution of fast evolution species, such as Human Immunodeficiency Virus (HIV), finding the origin of life (the tree of life project, http://www.tolweb.org/tree/) and coevolutions among different species. This highly interdisciplinary course is self-contained and is designed to introduce graduate and advanced undergraduate-level students (permission to enroll is required) to the fast growing field of Bioinformatics (with a lot of research opportunities and funding).

The course will cover methods underlying molecular phylogeny studies of protein and nucleic acid sequences to elucidate evolutionary histories and relationships of genes and organisms, as summarized in phylogenetic trees. Students will learn theory of molecular sequence evolution, methods of data acquisition, utilization of sequence databases, methods of phylogenetic analysis, and the interpretation and evaluation of phylogenetic trees.

The course is interdisciplinary and collaborative in nature, with students and instructors from three relevant disciplines: statistics, computer science, and the life sciences. Students will be introduced to the basics of phylogenetics and phylogenomics from the perspectives of all three disciplines, and will gain more in-depth knowledge through instruction and assignments tailored to their own areas of specialization. Projects that team students with complementary skills will illustrate the interdisciplinary nature of this science and the potential that can be realized through such collaborations.

Selected topics

- Dogma of molecular biology (information flow)
- Nature of mutations
- Neutral theory
- Homology
- Evolution models (such as GTR, HKY, JC etc)
- Hidden Markov Model and Alignment problem
- Maximum likelihood phylogeny inference
- Distance based methods such as NJ method, BME method, and UPGMA.

Prereqs

Interest in phylogenetics and phylogenomics.

Grading

Students performance will be graded based on two groups of assignments:

Common assignments for all students – 80% homework assignments, ability of using/applying bioinformatics tools, a written review of a recent paper in bioinformatics, and a final group project: these will count 20%, 20%, 10%, and 30% of your grade, respectively.

Area-specific assignments – 20% Students in computer science, life sciences and statistics will be additionally graded on area-specific assignments to cover and test the depth of knowledge and understanding of area-specific topics in the scope of the course. Such additional assignments may include quizzes/exams and project assignments. This component will count for 20% of the grade. Students should declare their area of specialization at the beginning of the semester.

Correctness, completeness and presentation of turned-in assignments affects the grade. In general, late assignments will not be accepted. The grading scale: 90% and above – A grade, above 80% – B grade, above 70% – C grade; failing grade otherwise.

Students conduct, academic honesty

All students must adhere to the university policies described in Senate regulations. In particular, all the submitted assignments should represent student's individual work with all sources clearly identified. Academic offenses such as plagiarism and cheating (see `http://www.uky.edu/Ombud/Plagiarism.pdf`) are penalized. Consult Senate rules (see SR 6.4) (see `http://www.uky.edu/USC/New/SenateRulesMain.htm`), Student Code (see `http://www.uky.edu/StudentAffairs/Code/`), and discuss it with your instructors if you have any doubts or questions.

Final Group Project

Students will work on a group project in a team consisting of students representing different areas. Proposals of group projects are due by the end of the third week and should be discussed with the instructors.

We list some project ideas below:

- Reconstruct trees from data sets, such as endophytes, downloaded from Gene Bank using different methods and then explain what you think. Student(s) from Biology: determine whether the trees can be used to help identify phylogenetic species; Statistics: Which statistical model is best for the given data sets (model selection)? Why do you think so?
- Do computer simulations using PAML and compare various implementations of ML methods (such as ML method implemented in RaXml, Girli, fastDNAml, DNAml from phylip), NJ, BME, under some evolution model. Compare their running time performance. Discuss accuracy of the results. How does PAML generate the DNA sequences with given tree? Why ML works better than the other methods? Why ML is slower than the other methods? How could you improve the algorithms or their implementations?
- Given two genomes (for example, available through NCBI), find the number of ortholog pairs by developing and implementing suitable computer search methods. Also identify any paralogs of these genes, and present your evidence for orthology or paralogy.

Schedule

Table 1: Tentative schedule

Week	Topic	Instructor(s)
1	Introduction	Drs Schardl, and Jaromczyk, and Yoshida
2	Dogma of Molecular Biology	Dr. Schardl
3	Phylogeny and Genealogy.	Drs Schardl, and Jaromczyk, and Yoshida
4	Nature of Mutation	Dr. Schardl
5	Homology (orthology and paralogy)	Dr. Schardl
6	Alignment	Dr. Yoshida
7	Alignment	Dr. Jaromczyk
8	Cladistics and Parsimony	Dr. Schardl
9	Distance Based Methods	Dr. Yoshida
10	Distance Based Methods	Dr. Jaromczyk
11	Evolution Models	Dr. Yoshida
12	Evolution Models	Dr. Jaromczyk
13	Maximum Likelihood Trees	Dr. Yoshida
14	Maximum Likelihood Trees	Dr. Jaromczyk
15	Coalescence	Dr. Weisrock (guest presentation)

Recommended books and reading materials

- R. Durbin, S. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis*, Cambridge University Press (1998)- Application of graphical models to problems in biological sequence

analysis.

- J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, Mass (2004).
- Selected journal articles; references provided by the course instructors.