MCB 4934 / 6937
Special Topics – Molecular Bioinformatics in UNIX
3 credits

This course is taught completely online through UF e-learning; no scheduled meeting times

Prerequisites

Undergraduate: MCB 3020 or MCB 3023 or BSC 2011 or BCH 4024 or CHM 3218, with a minimum grade of C

Graduate: None

Although familiarity with the UNIX operating system is not required prior to taking this course, if you do not have prior experience using UNIX, it is highly recommended that you complete the "UNIX Basics" module during the first two days of class (see weekly schedule) and before attempting the other course modules. The "UNIX Basics" module will cover what you need for this course.

In addition to course access through e-learning, all students will be provided with user accounts on our course UNIX server. Access to the course UNIX server is required to complete the weekly laboratory exercises (see weekly schedule). Please consult the "How to Access the Course UNIX Server" documentation during the first week of class to verify that you can connect to and complete laboratory exercises on our server.

The skills you will learn in this course are used in a wide variety of fields, including molecular and evolutionary biology, structural and functional biochemistry, structure-based drug design and protein engineering.

Instructor

Bryan Kolaczkowski
1250 Microbiology & Cell Science
352 392 5925
bryank@ufl.edu
Office hours by appointment; email contact preferred

Course Description

Introduction to major command-line bioinformatics tools used to study protein sequences, structures and molecular functions. We will focus on which tools are available to ask specific scientific questions, how these tools work and how tools can be combined into a comprehensive analysis of protein family sequence, structure and function.

Learning Objectives

1. Design and conduct sequence-similarity searches to identify and collect protein family members from sequence databases
2. Annotate protein sequences using functional domain models
3. Align protein sequences and process alignments for phylogenetic analysis
4. Infer protein family phylogenies, evaluate support for phylogenetic hypotheses and identify patterns of gene duplication and speciation on phylogenetic trees

5. Reconstruct ancestral protein sequences and evaluate support for ancestral sequence reconstructions
6. Infer 3D structures of protein sequences using structural homology modeling and evaluate model accuracy
7. Predict protein function using structure-based affinity prediction and visualize functional changes on a phylogenetic tree

Weekly Schedule of Topics and Assignments

Module Zero UNIX Basics – Complete this module in the first 2 days if you are not comfortable working in the UNIX operating system

Lectures:
1. Overview of the UNIX operating system
2. The UNIX directory hierarchy
3. Basic UNIX commands
4. Working with text files in UNIX
5. Secure file transfer protocol (sftp)

Readings: none

Lab Exercise: Using the UNIX command line interface (not graded)

Assigned Problem Set: none

<u>Module 1: Protein Sequences</u>

Week 1 The sequence search: Genes, proteins, sequences, homology and similarity

Lectures:
1. Overview of central dogma
2. Why work in protein space?
3. Sequence similarity and (local) alignment
4. Similarity and homology
5. Why finding homologs can be so difficult

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Altschul et al. (1990) Basic Local Alignment Search Tool. *J Mol Biol* 215(3):403-410.

Graduates:

Altschul et al. (1990) Basic Local Alignment Search Tool. *J Mol Biol* 215(3):403-410.

Altschul et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.

Lab Exercise: Sequence similarity search – Due Friday by 5pm

Week 2 What's in a protein sequence? Functional domains and protein domain architecture

Lectures:
1. What is a 'functional domain?'
2. Structure-based domains
3. Models of functional domains
4. Domain function and protein function
5. Do domains always do what they are supposed to?

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Marchler-Bauer et al. (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43(Database Issue):D222-D226.

Graduates:

Marchler-Bauer et al. (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43(Database Issue):D222-D226.

Barrera et al. (2014) Analysis of the protein domain and domain architecture content in fungi and its application in the search of new antifungal targets. *PLoS Comput Biol* 10(7):e1003733.

Lab Exercise: Domain architecture annotation – Due Friday by 5pm

Assigned Problem Set: Sequence similarity search and domain architecture – Due Friday by 5pm

<u>Module 2: Sequence Alignment</u>

Week 3 Pairwise and multiple sequence alignment

Lectures:
1. What is a sequence alignment?
2. Pairwise alignment
3. Profile alignments, guide trees and multiple alignment
4. Problems with indels
5. Iterative refinement and simultaneous tree-alignment

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Katoh and Standley. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772-780.

Graduates:

Katoh and Standley. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772-780.

Needleman and Wunsch. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48(3):443-453.

Lab Exercise: Multiple sequence alignment – Due Friday by 5pm

Week 4 Alignment processing, domain-wise alignment and structure-based sequence alignment

Lectures:
1. Ambiguity in alignments – removing it
2. Ambiguity in alignments – integrating it
3. Ambiguity in alignments – avoiding it
4. Aligning 3D structures
5. Structure-based sequence alignment

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Castresana. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540-552.

Graduates:

Castresana. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540-552.

Madhusudhan et al. (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng Des Sel* 22(9):569-574.

Lab Exercise: Domain-wise alignment and GBlocks processing – Due Friday by 5pm

Assigned Problem Set: Sequence alignment – Due Friday by 5pm

<u>Module 3: Phylogenetic Inference</u>

Week 5 Evolutionary models and the likelihood of a tree

Lectures:
1. L(t|d)=P(d|t) part 1 – single site, single branch
2. L(t|d)=P(d|t) part 2 – multiple sites on a tree
3. Markov transition models
4. Continuous time Markov models and branch lengths
5. Formalizing the likelihood of a tree

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Felsenstein. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17(6):368-376.

Graduates:

Felsenstein. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17(6):368-376.

Huelsenbeck and Rannala. (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276(5310):227-232.

Lab Exercise: Phylogenetic tree inference – Due Friday by 5pm

Week 6 Finding the 'best' tree and evaluating its 'significance': Tree search and statistical support measures

Lectures:
1. Tree search – finding the best tree
2. Tree search – finding plausible trees
3. Clade support – bootstraps and aLRTs
4. Bayesian approaches to clade support
5. Statistical topology tests

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Huelsenbeck et al. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294(5550):2310-2314.

Graduates:

Huelsenbeck et al. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294(5550):2310-2314.

Anisimova et al. (2011) Survey of branch support methods demonstrates accuracy, power and robustness of fast likelihood-based approximation schemes. *Syst Biol* 60(5):685-699.

Lab Exercise: Clade support and topology testing – Due Friday by 5pm

<u>Module 4: Ancestral Sequence Reconstruction</u>

Week 7 Testing evolutionary hypotheses in the lab: Ancient genes, where to find them, and what you can do with them once you have them

Lectures:
1. What does 'ancestral' mean on a tree?
2. The Functional Synthesis paradigm
3. How basic ancestral sequence reconstruction works

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Dean and Thronton. (2007) Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet* 8(9):675-688.

Graduates:

Dean and Thornton. (2007) Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet* 8(9):675-688.

Randall et al. (2016) An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat Commun* 7:12847.

Lab Exercise: Simple ancestral sequence reconstruction – Due Friday by 5pm

Week 8 Probability distributions across amino-acid residues, alignment positions and nodes on the phylogenetic tree: How to infer ancestral sequences

Lectures:
1. Understanding the ancestral reconstruction algorithm
2. Assessing confidence in ancestral sequences
3. The problem of gaps
4. Alignment and tree ambiguity
5. Potential problems and limitations of ancestral reconstruction

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Yang et al. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141(4):1641-1650.

Graduates:

Yang et al. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141(4):1641-1650.

Arenas et al. (2017) ProtASR: an evolutionary framework for ancestral protein reconstruction with selection on folding stability. *Syst Biol* (epub ahead of print)

Lab Exercise: Ancestral sequence reconstruction with gaps and confidence assessment – Due Friday by 5pm

Assigned Problem Set: Ancestral sequence reconstruction – Due Friday by 5pm

<u>Module 5: Structural Modeling</u>

Week 9 Does protein sequence determine structure? Does structure determine function? The sequence-structure-function triangle

Lectures:
1. Sequence, structure and molecular function
2. Protein sequences and folding
3. Effects of mutations on structure
4. Pockets, active sites and molecular interactions
5. What is a structure's function?

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Roy et al. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4):725-738.

Graduates:

Roy et al. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4):725-738.

Khoury et al. (2014) Protein folding and *de novo* protein design for biotechnological applications. *Trends Biotechnol* 32(2):99-109.

Lab Exercise: Swiss-Model – Due Friday by 5pm

Week 10 Structural homology modeling and evaluation

Lectures:
1. Protein folding and free energy
2. De-novo structure prediction
3. Why can de novo structure prediction be unreliable?
4. Structural homology modeling
5. How do you know your model is 'good?'

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Sali and Blundell. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779-815.

Graduates:

Sali and Blundell. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779-815.

Shen and Sali. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* 15(11):2507-2524.

Lab Exercise: Simple structural homology modeling – Due Friday by 5pm

<u>Module 6: Protein Function Prediction</u>

Week 11 What is a protein's function?

Lectures:
1. Informal ideas about protein function
2. Can we formalize general protein function?
3. Biological function, genomic context, and environment
4. Molecular function

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Botstein et al. (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25(1):25-29.

Graduates:

Ashburner et al. (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25(1):25-29.

Radivojac et al. (2013) A large-scale evaluation of computational protein function prediction. *Nat Methods* 10(3):221-227.

Lab Exercise: Gene ontology analysis – Due Friday by 5pm

Week 12 Structure-based function prediction

Lectures:
1. Protein sequence, structure and function
2. Does protein dynamics impact molecular function?
3. Small-effect and large-effect mutation models
4. Ligand affinity as generalizable molecular function
5. Structure-based drug design and beyond

Readings – Literature Summaries due Friday by 5pm

Undergraduates:

Dias and Kolaczkowski. (2015) Different combinations of atomic interactions predict protein-small molecule and protein-DNA/RNA affinities with similar accuracy. *Proteins* 83(11):2100-2114.

Graduates:

Dias and Kolaczkowski. (2015) Different combinations of atomic interactions predict protein-small molecule and protein-DNA/RNA affinities with similar accuracy. *Proteins* 83(11):2100-2114.

Yugandhar and Gromiha. (2014) Protein-protein binding affinity prediction from amino acid sequence. *Bioinformatics* 30(24):3583-3589.

Lab Exercise: Affinity prediction by machine learning – Due Friday by 5pm

Assigned Problem Set: Protein function and function prediction – Due Friday by 5pm

Critical Dates

Friday of each week – Weekly literature summaries and laboratory exercises due by 5pm

Friday, week 02 – Module 1 assigned problem set due by 5pm
Friday, week 04 – Module 2 assigned problem set due by 5pm
Friday, week 06 – Module 3 assigned problem set due by 5pm
Friday, week 08 – Module 4 assigned problem set due by 5pm
Friday, week 10 – Module 5 assigned problem set due by 5pm
Friday, week 12 – Module 6 assigned problem set due by 5pm
– Final project due by 5pm

Textbooks

Required readings will be posted on the course website (see weekly schedule, above, for complete reading list).

No additional textbook is required. The following are recommended textbooks that might be helpful or interesting:

Bioinformatics: Sequence and Genome Analysis by David Mount. Cold Spring Harbor Press. ISBN 0879697121

Introduction to Bioinformatics by Teresa K Atwood and David Parry-Smith. Pearson Education. ISBN 0582327881

Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins by Andreas D Baxevanis and BF Francis Ouellette (Eds). John Wiley & Sons. ISBN 0471478784

Evaluation and Grading

1. Literature Summaries Due Friday of each week, by 5pm
   For each assigned reading, students will write a short, structured summary. Points will be awarded according to the following rubrics:

   Undergraduates (10pts each; 120pts total):
   - 5pts – Identify the main conclusion(s) of the paper
   - 5pts – Identify the evidence supporting the main conclusion(s)

   Graduates (5pts each; 120pts total):
   - 1pt – Identify the main conclusion(s) of the paper
   - 2pts – Identify the evidence supporting the main conclusion(s)
   - 2pts – Summarize the methodologies used to generate supporting evidence

2. Laboratory Exercises (10pts each; 120pts total) Due Friday of each week, by 5pm
   Students will complete each laboratory exercise on the course UNIX server, where it will be graded automatically after the due date/time.

3. Assigned Problem Sets (10pts each; 60pts total) Due the last Friday of each module, by 5pm
   At the conclusion of each module, students will answer a short series of conceptual problems derived from material presented in the video lectures. Problem sets will be 'open book' and 'open notes;' you may use any provided or additional information available to you, but you may not ask any unauthorized third party for help on assigned problem sets, consistent with the UF Academic Honesty policy (see below).

   Note that problem sets for Graduate Students Only may contain questions derived from assigned readings; undergraduate problem sets will be derived solely from lecture material.

4. Final Project (140pts total) Due Wednesday of finals week, by 5pm

Students will complete a final laboratory project integrating approaches and data from each of the weekly laboratory exercises. Specific requirements for Undergraduate and Graduate final projects are available on the Final Project page.

Points will be awarded according to the following rubrics:

Undergraduates (140pts total):

- 20pt – Overall project objective clearly stated
- 40pts – Objective of each analysis clearly stated
- 40pts – Overall methodology clearly explained
- 40pts – Results of each analysis clearly explained in detail

Graduates (140pts total):

- 20pts – Overall project objective clearly stated
- 20pts – Objective of each analysis clearly stated
- 20pts – Overall methodology clearly explained
- 20pts – Methodology of each analysis clearly explained in detail
- 40pts – Results of each analysis clearly explained in detail
- 20pts – Overall results summarized clearly in conclusion

Grades will be assigned based on a percentage of total available points:

90.0 – 100.0%	A
80.0 – 89.99%	B
70.0 – 79.99%	C
60.0 – 69.99%	D
<60.0%	E

You can always calculate your standing grade by summing the points you have accumulated so far in the course and dividing by the total number of points possible over the duration of the course; this will be the grade you will receive if you stop attending.

Grades and Grade Points
For information on current UF policies for assigning grade points, see https://catalog.ufl.edu/ugrad/current/regulations/info/grades.aspx

Attendance and Make-Up Work
Requirements for class attendance and make-up exams, assignments and other work are consistent with university policies that can be found at: https://catalog.ufl.edu/ugrad/current/regulations/info/attendance.aspx.

Online Course Evaluation Process
Student assessment of instruction is an important part of efforts to improve teaching and learning. At the end of the semester, students are expected to provide feedback on the quality of instruction in this course using a standard set of university and college criteria. These evaluations are conducted online at https://evaluations.ufl.edu. Evaluations are typically open for students to complete during the last two or three weeks of the semester;

students will be notified of the specific times when they are open. Summary results of these assessments are available to students at https://evaluations.ufl.edu/results.

Academic Honesty
As a student at the University of Florida, you have committed yourself to uphold the Honor Code, which includes the following pledge: "*We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honesty and integrity.*" You are expected to exhibit behavior consistent with this commitment to the UF academic community, and on all work submitted for credit at the University of Florida, the following pledge is either required or implied: *"On my honor, I have neither given nor received unauthorized aid in doing this assignment*."

It is assumed that you will complete all work independently in each course unless the instructor provides explicit permission for you to collaborate on course tasks (e.g. assignments, papers, quizzes, exams). Furthermore, as part of your obligation to uphold the Honor Code, you should report any condition that facilitates academic misconduct to appropriate personnel. It is your individual responsibility to know and comply with all university policies and procedures regarding academic integrity and the Student Honor Code. Violations of the Honor Code at the University of Florida will not be tolerated. Violations will be reported to the Dean of Students Office for consideration of disciplinary action. For more information regarding the Student Honor Code, please see: http://www.dso.ufl.edu/sccr/process/student-conduct-honor-code.

Software Use:
All faculty, staff and students of the university are required and expected to obey the laws and legal agreements governing software use. Failure to do so can lead to monetary damages and/or criminal penalties for the individual violator. Because such violations are also against university policies and rules, disciplinary action will be taken as appropriate.

Services for Students with Disabilities
The Disability Resource Center coordinates the needed accommodations of students with disabilities. This includes registering disabilities, recommending academic accommodations within the classroom, accessing special adaptive computer equipment, providing interpretation services and mediating faculty-student disability related issues. Students requesting classroom accommodation must first register with the Dean of Students Office. The Dean of Students Office will provide documentation to the student who must then provide this documentation to the Instructor when requesting accommodation

0001 Reid Hall, 352-392-8565, **www.dso.ufl.edu/drc/**

Campus Helping Resources
Students experiencing crises or personal problems that interfere with their general well-being are encouraged to utilize the university's counseling resources. The Counseling & Wellness Center provides confidential counseling services at no cost for currently enrolled students. Resources are available on campus for students having personal problems or lacking clear career or academic goals, which interfere with their academic performance.

- *University Counseling & Wellness Center, 3190 Radio Road, 352-392-1575,* **www.counseling.ufl.edu/cwc/**
  - Counseling Services
  - Groups and Workshops
  - Outreach and Consultation
  - Self-Help Library
  - Wellness Coaching

- U Matter We Care, **www.umatter.ufl.edu/**

- *Career Resource Center,* First Floor JWRU, 392-1601, **www.crc.ufl.edu/**

Each online distance learning program has a process for, and will make every attempt to resolve, student complaints within its academic and administrative departments at the program level. See http://distance.ufl.edu/student-complaints for more details.