Topics in Computational Biology and Genomics

{MCB, PMB, BioE}{c146, c246}

University of California, Berkeley Spring 2005

"Instruction and discussion of topics in genomics and computational biology. Working from evolutionary concepts, the course will cover principles and application of molecular sequence comparison, genome comparison & functional annotation, and phylogenetic analysis."

4 Units

Instructors.

Steven E. Brenner

Associate Professor, Plant & Microbial Biology
Affiliated Associate Professor, Molecular & Cell Biology, Bioengineering
Faculty Scientist, Lawrence Berkeley National Laboratory

Michael B. Eisen

Assistant Professor, Molecular & Cell Biology
Faculty Scientist, Lawrence Berkeley National Laboratory

Both may be reached by email to c246profs@compbio.berkeley.edu

Teaching assistant.

Liana Lareau, lfl@compbio.berkeley.edu

Class meetings.

Tuesday and Thursday, 11:00-12:30 in 241 Cory Hall
Weekly discussion section: Friday 10:00-12:00, 9 Evans.

Attendance is required.

Prerequisites.

Bioengineering 142, Computer Science 61B, or equivalent ability to write programs in Java, Perl, C, or C++; Molecular and Cell Biology 100, 102, or equivalent; or consent of instructor

Core specialization (Bioengineering).

B (Bioinformatics and Genomics) and D (Computational Bioengineering). It also fulfills biological content.

Core requirement (Molecular and Cell Biology).

This course can fulfill a core course requirement for the graduate program in G&D in the Department of Molecular and Cell Biology by petition.

Textbook.

Durbin R., Eddy S., Krogh A., Mitchison G. Biological Sequence Analysis. Cambridge: Cambridge UP, 1998.

Literature articles found on the course website: http://compbio.berkeley.edu/class/c246

Assigned readings must be completed before the class for which they are assigned.

Optional Additional References.

These books provide additional introductory references to the core topics that will be discussed in the course. Copies will be placed on reserve in the Biosciences Library.

Eidhammer I., Jonassen I., Taylor R. *Protein Bioinformatics.* Wiley, 2004.

Hall B.G. *Phylogenetic Trees Made Easy.* Sinauer Associates, 2001.

Koonin E.V., Galperin M.Y. *Sequence – Evolution – Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, 2002.

Lesk A.M. *Introduction to Bioinformatics.* Oxford: Oxford UP, 2002.

Sebutal J.C., Meidanis J. *Introduction to Computational Molecular Biology*. Brooks/Cole Pub Co, 1997.

Grading.

25% homework
20% midterm exam
20% project
25% final exam (+ resurrection from midterm exam)
10% class participation

Membership.

The six different "versions" of the class. The versions listed in different departments are *identical*. You may sign up for any version.

The undergraduate c146 & graduate c246 versions have the same lectures. However, for the graduate version, students will be required to do additional questions on homework problem-sets and to prepare a paper presentation for the class section.

Auditors are welcome if space allows. Auditors are expected to participate fully in the class

Homework.

Homework will typically be assigned in class on Tuesdays, and it will be due by email to lfl@compbio.berkeley.edu by 6pm the following Monday. Homework should be submitted electronically to the GSI and should be in plaintext, Word or PDF. *Identical* paper copies must be turned in at the beginning of class on Tuesday. Where verbal responses are required, they must be in cogent standard written English.

Homework received between 6pm Monday and 11am Tuesday will be penalized by 10 percent. After that, an additional 10 percent will be deducted for each day late, and no credit will be given for problems that have been discussed in class.

Oral discussion of the class and homework is encouraged. However, all homework questions must be answered in writing alone and must be fully understood. You must also list all the people with whom you discussed the question.

Homework received between 6pm Monday and 11am Tuesday will be penalized by 10 percent. After that, an additional 10 percent will be deducted for each day late, and no credit will be given for problems that have been discussed in class.

The lowest scoring homework will not be included in your grade calculation.

Computer access.

Programs may be written on any computer in Perl, C, C++, or Java.

Class notes.

For lectures given with PowerPoint, the instructor's presentation will be placed on the course website following class.

This class will use the scribe system. *Failure to adhere to the following requirements will impact the student's class participation grade*. One student ("scribe") will be designated to take notes each week, while another ("reader") will review these notes for accuracy and work with the scribe to correct any errors or omissions. The scribe must provide notes to the reader by the following lecture. By the lecture thereafter, the reader must submit the notes by email to the teaching assistant. All notes must be electronic so they may be placed on the website.

Office hours.

Office hours for Steven Brenner will be 1:00 to 2:00pm on Tuesdays in Koshland 461East. Office hours for Michael Eisen will be by appointment. Any changes in office hours will be announced.

Project.

Pairs (or for exceptionally complex projects, triples) of students will undertake a substantial research project, creating new computational biology methodologies or performing a significant genomic analysis. The final project will be presented at a class poster session and written up as a brief (roughly 3 page) report. Electronic versions of both the poster and report must be submitted, along with supplementary information including figures, references, datasets, and custom software.

Website.

The course website is http://compbio.berkeley.edu/class/c246. Consult the page regularly for homework, class notes, and updated information.

Course Schedule

(Timing and details subject to change)

Lecture 1 [Brenner]
Introduction to Sequence Analysis; **Scoring Alignments; Dynamic Programming**
DOTTER, dot plots, local & global alignments

Lecture 2 [Brenner]
Sequence Evolution

This lecture will discuss evolution at the sequence level and the importance of understanding evolution for sequence analysis. Read Walter Fitch's modern discussion of homology and terminology used in discussing sequence evolution. Then read short note by Winter et al. which argued the reconstruction of sequence evolution is impossible, and Fitch's rebuttal. Begin reading the ISMB tutorial on protein evolution by William Pearson (complete by Lecture 3).

Pearson W.(2000). **Protein sequence comparison and protein evolution** This is the ISMB tutorial.
Fitch WM.(2000). **Homology a personal view on some of the problems**. *Trends in Genetics* 16:227-31.
Winter WP, Walsh KA and Neurath H.(1968). **Homology as applied to proteins**. *Science* 162:1433.
Fitch WM.(1970). **Distinguishing homologous from analogous proteins**. *Systematic Zoology* 19:99-113.For now, just read pages 99-102, 112-113.

Lecture 3 [Brenner]
Sequence Evolution

Start reading chapter 2 of the Durbin, Eddy, Krogh & Mitchison (DEKM) book.
Finish reading pages 103-111 of the Fitch article. Focus on understanding principles, but not the details.
Continue reading the Pearson ISMB tutorial.

Lecture 4 [Brenner]
Dynamic Progamming with General Gap Penalties
Affine gaps, dynamic programming, gap parameters

Finish reading DEKM sections 2.1, 2.2, 2.3, 2.4

Lecture 5 [Brenner]
Efficient & effective scoring: Matrices and Gap Parameters

DEKM section 2.8
Henikoff S and Henikoff JG (1992). **Amino acid substitution matrices from protein blocks**. *Proceedings of the National Academy of Sciences of the United States of America* 89:10915-19.
Dayhoff MO, Schwartz RM and Orcutt BC (1978). **A model of evolutionary change in proteins**.
Yu YK, Wootton JC, Altschul SF. 2003. The compositional adjustment of amino acid substitution matrices. *Proc Natl Acad Sci U S A.* 100: 15688-93.

Lecture 6 [Brenner]
Heuristic Alignment Methods [FASTA and BLAST]

DEKM section 2.5
Gallison F (2000). **The Fasta and Blast programs**

Lecture 7 [Brenner]
Statistical Significance of Alignments

DEKM section 2.7
Pagni M and Jongeneel CV (2001). **Making sense of score statistics for sequence alignments**. *Briefings in Bioinformatics* 2:51-67

Lecture 8 [Brenner]
Gene annotation based on homology

Brenner SE (1999). **Errors in genome annotation**. *Trends in Genetics* 15:132-3.
Ashburner M, Ball CA, Blake JA, Botstein D, Butler H *et al.* (2000). **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 25:25-29.
Single family studies (Kinome)
Eisen JA (1998). **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis**. *Genome Research* 8:163-7.
Phylogenomics & annotation [papers TBD]
Orengo TBA
Todd, Valencia, Rost TBA
Engelhart B and Brenner SE, TBA

Lecture 9 [Brenner]
Scoring multiple alignments

DEKM Chapter 6
Gonnet GH, Korostensky C, Benner S (2000). **Evaluation measures of multiple sequence alignments.** *Journal of Computational Biology* 7:261-276.

Edgar B, TBA

Lecture 10 [Brenner]
Progressive multiple alignments with heuristics

Thompson JD, Higgins DG, and Gibson TJ (1994). **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Research* 22:4673-4680.
Notredame C, Higgins DG and Heringa J (2000). **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *Journal of Molecular Biology* 302:205-217.

Lecture 11 [Brenner]
Recent developments in multiple alignment

Katoh K, Misawa K, Kuma K and Miyata T (2002). **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Research* 30:3059-3066.
Edgar R. MUSCLE (paper TBA)
Partial Order Alignment (paper TBA)
ProbCons (TBA)

Lecture 12 [Brenner]
Profiles, Iterated searching, and PSI-BLAST

Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, et al. (2001). **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements**. *Nucleic Acids Res* 29:2994-3005.
Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs** *Nucleic Acids Research* 25:3389-3402.

Lecture 13 [Eisen]
Comparative genome analysis
Readings TBA: Pipmaker, VISTA, phylogenetic shadowing, searching whole genomes, BLAT, SPIDEY

Lecture 14 [Eisen]
Whole genome alignment methods

Readings TBA: Mummer Glass Avid Chained-BLASTZ LAGAN Inferring genome rearrangements

---

Lecture 15 [Brenner]
Midterm review

Lecture 16

MIDTERM EXAM

Lecture 17 [Eisen]
Finding Motifs I: MEME, EM and Gibbs Sampling

Stormo GD (2000). DNA binding sites: representation and discovery *Bioinformatics* 16:16-23.
Bailey TL and Elkan C (1994). **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.
Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC (1993). **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment**. *Science* 262:208-14.

Lecture 18 [Eisen]
What are Hidden Markov Models and Why Use Them?

Pfam & HMMs DEKM Chapter 3

Lecture 19 [Eisen]
Hidden Markov Models & Biological Applications; also Dirichlet priors

DEKM Chapter 5

Lecture 20 [Liana]
Project discussion

Lecture 21 [Eisen]
More HMMs

Lecture 22
Final projects discussion

Lecture 23 [Eisen]
Why Phylogeny Matters; Distance methods
UPGMA, NJ, NJ variants: bioNJ, Weighbor, FastME
Doolittle WF (1999). **Phylogenetic classification and the universal tree**. *Science* 284:2124-2129.
DEKM Chapter 7
Page RDM and Holmes EC (1998). **Molecular evolution : a phylogenetic approach**, (Oxford; Malden, MA, Blackwell Science), pp. 172-227.

Lecture 24 [Eisen]
Phylogeny: Parsimony, Likelihood, Bootstrapping

PAUP*, Long branch attraction

Felsenstein J (2002). **Bootstrap and randomization tests.** Chapter 20 of **Inferring Phylogenies**, (Cambridge, MA, Sinnauer).
Huelsenbeck JP and Rannala B (1997). **Phylogenetic methods come of age: testing hypotheses in an evolutionary context**. *Science* 276:227-32.
DEKM Chapter 8
Page RDM and Holmes EC (1998). **Molecular evolution : a phylogenetic approach**, (Oxford ; Malden, MA, Blackwell Science), pp. 193-200.
Felsenstein J (2002). **Likelihood methods.**. Chapter 16 of **Inferring Phylogenies**, (Cambridge, MA, Sinnauer).

Lecture 25 [Eisen]
Phylogeny: Bayesian approaches and the unique problems they solve.
Mr. Bayes

TBA Huelsenbeck JP, Ronquist F, Nielsen R and Bollback JP (2001). **Bayesian inference of phylogeny and its impact on evolutionary biology**. *Science* 294:2310-4.
[Nat Rev Genet review?]

Lecture 26 [Eisen]
Whole genome phylogeny

Lecture 27 [Eisen]
Finding general & specific RNA structures

SCFGs snoRNAs

Lecture 28 [Eisen]
Micro RNAs
Sankoff algorithm (paper TBA) Heuristic implementations (papers TBA)

Lecture 29 [Brenner & Eisen]
All questions answered / Final exam review

Lecture 30
Poster presentations

Lecture 31
Poster presentations

FINAL EXAM
Time and location TBA