BIO/CSE 597F
Bioinformatics I: basic analysis of DNA and protein sequences

Fall semester, 2001
Tues/Thurs 2:30-3:45, 112 Thomas Building.
3 credits
Schedule numbers: BIO 597F - 822867; CSE 597F - 908214

Prerequisites: Graduate standing or consent of instructors. Familiarity with molecular biology or computer methods is not assumed (though a willingness to learn a little about them is required).

Topics: (estimated time for discussion)

- Introduction to molecular biology for the non-biologist. (0.5 week)
- World Wide Web resources for identifying genes in a genomic sequence and for predicting a gene's function. Students will use RepeatMasker, GenScan, Blast, ClustalW, Pfam or Interpro, and PipMaker. Analysis of chloroplast genome sequences will serve as a focal point for illustrating concepts and methods (and probably for many of the term projects). (3 weeks)
- Phylogenetic analysis of molecular sequences with an emphasis on methods of phylogenetic inference and hypothesis testing. Gene and genome history, gene family evolution, inference of ancestral proteins, and phylogenetic analysis as a predictive tool. (2 weeks)
- Survey of interspersed repeats in mammals. (0.5 week)

  Search methods for sequence databases. The Blast family of programs, including psi-blast and phi-blast. Introduction to dynamic programming and determination of substitution scores. (2 weeks)
- Multiple sequence alignment, including ClustalW. (0.5 week)
- Comparison of a genomic sequence to a spliced gene sequence. The Sim4 program and greedy alignment algorithms. (0.5 week)
- Introduction to hidden Markov models. (1 week)
- Methods to search databases of protein motifs or domains, particularly Pfam. (0.5 week)
- *Ab initio* gene prediction, i.e., predicting the location of genes within a given genomic sequence using only intrinsic sequence properties. Methods employed by GenScan (currently the best program of this type) will be studied in some detail. (1.5 weeks)
- Students' in-class reports on their term projects. (3 weeks).

The course will cover neither protein structure prediction nor ''post-sequencing'' bioinformatics, such as analysis of expression data (e.g., from micro-arrays), proteomics, and analysis of regulatory networks. Protein structure prediction is covered in courses offered by the Chemistry Department, and post-sequencing bioinformatics is the topic of Bioinformatics II, which is tentatively scheduled for Spring Semester.

Grading will be based on (1) a term project of the student's choice (60%), (2) an exam over the lecture material (30%), and (3) a few homework assignments covering World Wide Web resources for analyzing genomic sequences (10%).

Required text:

Bioinformatics: Sequence and Genome Analysis
David W. Mount
Cold Spring Harbor Laboratory Press, 2001
ISBN 0-87969-608-7 (paperback)

For more information:
Webb Miller, 326A Pond Lab, 865-4551, webb@cse.psu.edu
Claude dePamphilis, 212 Mueller Lab, 863-6412, cwd3@psu.edu