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

Fall semester, 2000
Tues/Thurs 2:30-3:45, 302 Pond Lab.
3 credits
Schedule numbers: CSE 597F - 771102; BIO 597F - 770313

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, emphasizing World Wide Web resources for identifying genes in a genomic sequence and for predicting a gene's function. (2 weeks)
- Search methods for sequence databases. The Blast family of programs, including psi-blast and phi-blast — how to use them effectively and the fundamentals of how they work. 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)
- Comparison of homologous genomic sequences to identify functional regions, with an emphasis on PipMaker and human-mouse comparisons. Survey of interspersed repeats in mammals. (1 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)
- 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. (3 weeks)
- Students' in-class reports on their term projects. (3 weeks).

The course will not cover in detail either the prediction of protein structure from amino acid sequence or ‘‘post-sequencing’’ bioinformatics, such as analysis of expression data (e.g., from micro-arrays), proteomics, and regulatory networks. (Protein structure prediction is covered in courses offered by the Chemistry Department, and analysis of micro-array data will be discussed in Francesca Chiaromonte's Statistics course during Spring semester.)

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

Required texts:

Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids
R. Durbin, S. Eddy, A. Krogh and G. Mitchison
Cambridge University Press 1998
ISBN 0 521 62971 3 (paperback)

Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins
A. D. Baxevanis and B. F. F. Ouellette
Wiley-Interscience 1998
ISBN 0 471 19196 5 (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