

BIOS 784 Introduction to Computational Biology

Course Description and Objectives:

With the recent dramatic increase in many types of biological data due to the human genome project and other high-throughput projects, the scope of research in bioinformatics has expanded to a diverse range of topics including protein, DNA and RNA sequence analysis, microarray analysis, structural and functional predictions, gene finding and phylogeny reconstruction. This one semester course is intended to provide coverage of bioinformatics developments in the past two decades with an emphasis on topics of recent interest. By the end of this class, students are expected to have an in-depth knowledge of computational methods important in bioinformatics, and a thorough grasp of the underlying principles which would be adequate to evaluate and develop novel techniques in scenarios that may arise in the future.

Course Meetings: Mondays and Wednesdays from 12.30-1.45 pm. Separate lab sessions to be announced.

Course requirements: Reading of designated articles. Homeworks (to be assigned every 2-3 weeks), an in-class presentation, and a final project due on the last day of class.

Prerequisites: Biostatistics 661 (previously 161) and 663 (previously 163), or permission of the instructor. Familiarity with a programming language (such as C, PERL or Fortran), and statistical software (such as R or SPlus) is highly desirable.

Main References/Textbooks

- Durbin, R. and Eddy, S. and Krogh, S. and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.
- Waterman, M. (1995). *Introduction to computational biology : maps, sequences and genomes*. Chapman and Hall.

In addition, most topics would substantially draw from articles in the current literature.

Course Contents: The following topics will be covered, in approximately chronological order.

Week 1: Overview of molecular biology and database resources.

Genomics and molecular biology basics; the human genome project. Review of biological databases, data formats, methods for retrieving relevant data.

Week 2: Statistical methodology and theory of sequence alignment.

- Dynamic programming methods and pairwise alignment algorithms.
- Global and local alignment, substitution matrices, gap models, BLOSUM, PAM matrices.

Week 3: Database search and alignment statistics.

- BLAST theory; gapped BLAST, Psi-BLAST.
- Matching statistics and the Poisson approximation, statistical significance of alignment scores, extreme value distributions, p-values and e-values.

Week 4: Essential stochastic processes and probability theory.

- Markov chains and the Poisson process. Substitution models.
- Hidden Markov models for biological sequences and their estimation. Forward-backward algorithm, recursions, Viterbi and Baum-Welch algorithms. Sequence segmentation and database scanning.

Weeks 5/6: Statistical estimation and computation through missing data formulation and iterative algorithms.

- Mixture models, profiles, Dirichlet mixtures. EM algorithm.
- Multiple sequence alignment and profile HMMs. Progressive alignment.
- Gene finding with HMM, modeling protein families, PFAM.

Week 7: Bayesian probability models and Monte Carlo methods in bioinformatics.

- Gibbs sampling approaches for multiple sequence alignment.
- Metropolis-Hastings, parallel tempering, simulated annealing and evolutionary Monte Carlo. Model-based clustering and classification methods.

Weeks 8/9: Statistical models and methods for gene regulatory motif discovery

- Combinatoric, Likelihood-based and Bayesian approaches.
- Regulatory module detection. Gene expression analysis and motif discovery. Functional motif clustering. Phylogenetic approaches.

Week 10: Construction of physical and genomic maps

Genomic mapping by random clones. Fragment assembly, whole genome shotgun assembly and genome annotation. Sequencing accuracy.

Week 11: RNA and Protein structure modeling and analysis

- RNA structure, stochastic context-free grammars.
- Ab-initio protein structure prediction, molecular optimization, Monte Carlo approaches to the protein folding problem, structural alignment.
- Comparative modeling: homology modeling, protein threading.

Week 12: Evolution and phylogeny.

Discrete and continuous time models for nucleotide substitution. Reconstruction of phylogenetic trees: parsimony, maximum likelihood and Bayesian approaches: modeling, estimation and hypothesis testing. Comparative genomics.

Weeks 13/14: Special topics.

SNP detection and haplotype reconstruction: Parsimony, EM, and Gibbs sampling approaches. Bayesian networks and probabilistic graphical models for inferring gene networks.

Week 15: Student presentations.

Other Recommended Readings

- Koski, T. (2001). *Hidden Markov models for bioinformatics*. Kluwer Academic.
- Liu, J. S. (2001). *Monte Carlo Strategies for Scientific Computing*. Springer-Verlag.
- Baxevanis, A. D. and Ouellette, B. F. F. (1998). *Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins*. Wiley-Interscience.
- Mount, D. (2001) *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.
- Lodish et. al (2000). *Molecular Cell Biology*. W. H. Freeman & Co. (available online at <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=mcb.TOC>)
- (Non-technical) Ridley, M. (2000) *Genome: the autobiography of a species in 23 chapters*.