COMP 691R

Bioinformatics Algorithms

Lecture 1a

Course Outline

Classes

Wednesdays 17:45 to 20:15 in H-523

Instructor

Professor Greg Butler

Room: ER-603-53

Tel: 848 2424 ext 3031

gregb@cs.concordia.ca

http://www.cs.concordia.ca/~gregb/

Office Hours and Discussion

In class (preferable).

By appointment (email me with suggested time)

Wednesdays 16:00-17:00 in ER-603-53

Bioinformatics Algorithms

— to cover the major algorithms used in bioinformatics

— emphasize algorithmic principles

Important

— algorithms use available databanks!

— practical performance of algorithms

— use on "farms" or "clusters"

Information Sources

See web site:
`http://www.cs.concordia.ca/~gregb`

Books in Webster Library Reserve.

Selected journal and conference articles. Web sites.

Evaluation

three assignments (60%)

three-hour final examination (40%)

Bioinformatics Algorithms — Real Objectives

Cover the necessary background on bioinformatics so that you can understand the problems of multiple sequence alignment (MSA), the applications of MSA to understanding protein families, and the emerging field of phylogenomics.

In particular, to cover the background and details for

L. Duret and S. Abdeddaim, *Multiple alignments for structural, functional, or phylogenetic analyses of homologous sequences*. In **Bioinformatics: Sequence, Structure and Databanks**, editted by D. Higgins and W. Taylor, Oxford University Press, 2000.

C. Notredame, *Recent progresses in multiple sequence alignment: a survey*, Pharmacogenomics **3**(1) (2002) 131–144.

Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry J-C, Thompson JD, Wicker N and Poch O. *PipeAlign: a new toolkit for protein family analysis*. Nucleic Acids Research 2003 (31) 3829-3832.

Sjölander K. *Phylogenomic inference of protein molecular function: advances and challenges*. Bioinformatics 2004 (20) 170-179.

Thomas PD, Campbell MJ, Kejariwa Al, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. *PANTHER: a library of protein families and subfamilies indexed by function*. Genome Research 2003 (1) 2129-2141.

Lecture Schedule — Tentative

Week 1:
Course Outline. Biology and Genomics Introduction.

Weeks 2 – 6:
Sequence Analysis and Annotation.

- Sequence properties, scanning.
- Alignment. Blast.
- Base calling, trimming, assembly.
- Gene Ontology, InterProScan, PSORT II, KEGG.

Weeks 7–12:
Phylogenomics

- Multiple Sequence Alignment (MSA).
- Phylogenetic trees.
- Outlier removal and region masking.
- Tree splitting, speciation vs duplication.

Assignments

All course work is individual work.

Three assignments, either programming or data analysis.

Programming assignments:

— require C and/or C++

— use existing libraries and software
PipeAlign, EMBOSS

— submit written report for each
— describe algorithm, design, testing, results
— source code listing as appendix to report

Data analysis assignments:

— use available tools and libraries

— given datasets

— submit 5–10 page written report for each

Assignments due about every 3–4 weeks.

— precise assignments and schedule to come

Final Examination

Focus is on material in listed papers and the required background.

Must know basic concepts of genomics.

Must know major algorithms

— purpose, how it works, complexity, limitations

— data structures, heuristics

Should be able to compare major algorithms

Should be able to propose a new algorithm for a data analysis problem using the major existing algorithms and information resources