Department of Computer Science and Software Engineering
COMP 6811 Bioinformatics Algorithms (Reading Course)
Fall 2019 Section AA Instructor: Gregory Butler

Curriculum Description

COMP 6811 Bioinformatics Algorithms (4 credits)

The principal objectives of the course are to cover the major algorithms used in bioinformatics; sequence alignment, multiple sequence alignment, phylogeny; classifying patterns in sequences; secondary structure prediction; 3D structure prediction; analysis of gene expression data. This includes dynamic programming, machine learning, simulated annealing, and clustering algorithms. Algorithmic principles will be emphasized. A project is required.

Outline of Topics

The course will focus on algorithms for *protein sequence analysis*. It will not cover genome assembly, genome mapping, or gene recognition.

- Background in Biology and Genomics
- Sequence Alignment: Pairwise and Multiple
- Representation of Protein Amino Acid Composition
- Profile Hidden Markov Models
- Specificity Determining Sites
- Curation, Annotation, and Ontologies
- Machine Learning: Secondary Structure, Signals, Subcellular Location
- Protein Families, Phylogenomics, and Orthologous Groups
- Profile-Based Alignments
- Algorithms Based on k-mers

Texts — in Library

D. Higgins and W. Taylor (editors). **Bioinformatics: Sequence, Structure and Databanks**, Oxford University Press, 2000.

A. D. Baxevanis and B. F. F. Ouelette. **Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins**, Wiley, 1998.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids**. Cambridge University Press, 1998.

Pierre Baldi and Søren Brunak. **Bioinformatics: The Machine Learning Approach**. MIT Press, Cambridge, second edition, 2001.

and other reading material.

Evaluation

Project 1	25%
Project 2	25%
Assignment 1	10%
Assignment 2	10%
Assignment 3	10%
Assignment 4	20%
Total	100%

Assignments

The assignments will be a mix of programming algorithms, programming scripts, writing reports, and comparing algorithms.

Assignment 1: Write a Python program that implements dynamic programming to perform global pairwise sequence alignment of protein sequences in `fasta` using a subsitution matrix and gap penalties (as parameters).

Assignment 2: Write a technical report explaining the `UPGMA` and `mbed` algorithms for constructing phylogenetic trees that are used as guide trees in some multiple sequence alignments. Compare and contrast their approaches. [Up to 8 pages in IEEE two-column format.]

Assignment 3: Write Python scripts to compare `Clustal Omega` and `T-Coffee` performance on the Balibase benchmark, and produce appropriate tables and figures for a technical report. What other methods are available to compare the algorithms? Write a technical report comparing the two algorithms. [Up to 8 pages in IEEE two-column format.]

Assignment 4: Run `eggNOG-mapper` on the protein sequences from the plants: corn (*Zea mays*), bread wheat (*Triticum aestivum var. Chinese Spring*), and rice (*Oryza sativa*); fungi: the common mushroom (*Agaricus bisporus*), noble rot (*Botrytis cinerea*), *Penicillium chrysogenum*, and wheat blight (*Gibberella zeae*); and animals: chicken (*Gallus gallus*), pig (*Sus scrofa*), cow (*Bos taurus*). Compare the annotations using the Biological Process (BP) aspect of an appropriate GO Slims. Write a technical report. [Up to 10 pages in IEEE two-column format.]

Projects

Project 1: A take-home project to develop Python scripts to cluster the sequences in TCDB (Transporter Classification Database) based on similarity measures derived from `blast` pairwise alignment of all pairs of sequences in the TCDB, and using at least two clustering algorithms. Do clusters conform to the TC nomenclature? Are clusters consistent with the GO annotation of the Swissprot entries in the cluster?. Do the different clustering algorithms agree, or disagree, in their clusters?

Project 2: A take home project to develop Python scripts for a HMM-based classifier for each of (a) ion channels, (b) receptors, and (c) GPI-anchored peripheral membrane proteins.

Other Reading

Faizah Aplop (2016), **Computational Approaches for Improving the Reconstruction of Metabolic Pathways**. Department of Computer Science and Software Engineering, Concordia University.

Christine Houry Kehyayan (2009–2013), **Using Synteny in Phylogenomics Algorithms to Cluster Proteins**. Department of Computer Science and Software Engineering, Concordia University.

Qing Ye (2019), **Classifying Transport Proteins Using Profile Hidden Markov Models and Specificity Determining Sites**, Department of Computer Science and Software Engineering, Concordia University.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

Stephen F. Altschul and Warren Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 1996.

Marco Pagni and C. Victor Jongeneel. Making sense of score statistics for sequence alignments. *Briefings in Bioinformatics*, 2(1):51–67, 2001.

Cedric Notredame, Recent progresses in multiple sequence alignment: a survey. *Pharmacogenomics* 3(1) (2002) 131144.

Cedric Notredame, Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, 2007 Aug 31;3(8):e123

Julie D. Thompson, Frédéric Plewniak, and Olivier Poch. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research* 27.13 (1999): 2682-2690.

Timo Lassmann and Erik LL Sonnhammer. Quality assessment of multiple alignment programs. *FEBS letters* 529.1 (2002): 126-130.

Sean R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531, 1994.

Elin Teppa, Angela D. Wilkins, Morten Nielsen, and Cristina Marino Buslje. Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. *BMC Bioinformatics* 13, no. 1 (2012): 235.

Abhijit Chakraborty and Saikat Chakrabarti. A survey on prediction of specificity-determining sites in proteins. *Briefings in Bioinformatics* 16, no. 1 (2014): 71-88.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Natural Genetics*, 25(1):25–29, 2000.

Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41(D1):D456–D463, 2013.

Rachael P Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J Martin, and Claire O'Donovan. The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063, 2015.

Marco Punta, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, Alex Bateman, and Robert D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, 2012.

Sean Powell, Kristoffer Forslund, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Jaime Huerta-Cepas, Toni Gabaldón, Thomas Rattei, Chris Creevey, Michael Kuhn, et al. eggNOG v4. 0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, page gkt1253, 2013.

Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* 34, no. 8 (2017): 2115-2122.

Adrian M Altenhoff and Christophe Dessimoz. Inferring orthology and paralogy. *Methods in Molecular Biology*, 855:259–279, 2012.

Natasha Glover, Christophe Dessimoz, Ingo Ebersberger, Sofia K Forslund, Toni Gabaldón, Jaime Huerta-Cepas, Maria-Jesus Martin, Matthieu Muffato, Mateus Patricio, Cécile Pereira, Alan Sousa da Silva, Yan Wang, Quest for Orthologs Consortium, Erik Sonnhammer, Paul D Thomas. Advances and Applications in the Quest for Orthologs. *Molecular Biology and Evolution*, 2019, msz150s.

Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, CJ Adams-Collier, and Kenta Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, 35(suppl 2):W585–W587, 2007.