BIOL-GA.1009 BIOLOGICAL DATABASES AND DATAMINING

Sessions: One 165 minutes class (Thursday 2:00pm-4:45pm)
Office Hour (Monday 2:00pm-3:00pm)

Location: Computer laboratory classroom

Recommended texts:

Marketa Zvelebil, Jeremy O. Baum. (2008) *Understanding Bioinformatics*. Garland Science*:*
Chapter 3 (Dealing with Databases)
https://store.vitalsource.com/show/978-1-1369-7696-4

Phil Spector (2008) *Data Manipulation with R*. Springer

Paul D. Lewis (2010) *R in Medicine and Biology* , Jones & Bartlett Publishers

Instructor: Dr. Manpreet S. Katari

Course Aims: The goal of the course is to provide students with the skills to integrate the different types of biological data and databases and learn how to mine them. Students will learn to create their own database using MySQL and SQLite containing different types of biological data and then use packages available in the programming language R to mine them. To mine the heterogeneous biological data, students will use machine-learning methods such as Support Vector Machines and Multiple Regressions on experimental data to classify and predict gene function and regulation.

Prerequisites:

Undergraduates:

Principles of Biology I
Principles of Biology II
Molecular and Cellular Biology I
Molecular and Cellular Biology II

Graduates:

None:

ALL:

Prior programming experience required.

Grading:

Weekly assignments:	40%
Midterm:	25%
Final Project:	25%
Attendance/Participation:	10%

Course Description:

Biological databases and data mining is a graduate course that is also offered to exceptional undergraduate students. The course is divided into three sections: 1) Introduction to MySQL and R. 2) Introduction to different data types, and 3) Machine learning methods for data mining. Students will learn to create their own database using MySQL and SQLite containing different types of biological data. Students will also learn to mine the heterogeneous biological data using machine-learning methods such as Clustering, Decision Trees and Multiple Regressions. We will apply these methods on experimental data in order to classify and prediction gene function and regulation.

Assignments and Projects:

Homework will be assigned at the end of class every week and will be due at the beginning of following class. Late assignments and projects will be penalized 10% for each day it is late. The midterm will be a 2-hour in-class examination, which will test the students on their computational skills and their understanding of the different biological datasets. For the final project, the class will be divided into 5 groups. Each group will be assigned an experimental dataset and a machine-learning method. The task will be to create a database that contains all necessary information and write a function to analyze the experimental data using the machine-learning method assigned. During the finals week each group will submit a written report and give a 15-minute presentation on what they found.

Course Syllabus

Part I: Computational tools

Week 1: **Introduction to databases**

Basic SQL: MySQL vs SQLite

Reading: Understanding Bioinformatics Ch. 3

Week 2: **Complex SQL queries**
Using indexes

Reading: PhP Chapter related to Database.

Week 3: **Introduction to R**
Connecting to Databases using R

Reading: Data Manipulation with R: Chapters 1, 2 & 3

Week 4: **Genome Databases (Browsers, Resources, File Formats)**
Functional Annotations : (GO-terms)
Writing Functions in R

Reading: R in Medicine and Biology: Chapters 6 & 12
Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: a building block for a model organism system database.Genome Res. 2002 Oct;12(10):1599-610.

Week 5: **Transcriptome Databases (Resources for retrieval and analysis)**

Pathway and Gene Regulatory databases (Agris, TransFac)

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10

Davuluri, R.V.,Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E. AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics (2003), 4:25.

Week 6: **Protein Interaction Databases (Biogrid, String)**

Building and Querying an Interaction Network Database

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.Nucleic Acids Res. 2011 Jan;39(Database issue):D561-8. Epub 2010 Nov 2.

Week 7: **Creating a robust database for integrating different Biological data types.**

*****Midterm assigned*****

<u>Part II: Data Mining</u>

Week 8: **Information Gain / Differntial Gene Expression / Clustering**

******GROUPS ARE FORMED AND PROJECTS ARE ASSIGNED******

Week 9: **Differential Gene Expression / Correlation / Clustering**

Reading: R in Medicine and Biology: ch 10-11

Week 10: **Decision Trees**

Installing RWeka

Reading: R in Medicine and Biology: pp. 208-211

Kingsford C, Salzberg SL. What are decision trees? Nat Biotechnol. 2008 Sep;26(9):1011-3.

Week 11: **Logistic Regression.**

Reading: R: pp. 195-207

Week 12: **Evaluating predictions (GO-term enrichment and ROC curves)**

Reading: Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

Week 13: **Case study discussion**

Reading: Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) Support Vector Machines and Kernels for Computational Biology. PLoS Comput Biol 2008, 4(10)

Reading: Jain P, Hirst JD. Automatic structure classification of small proteins using random forest. BMC Bioinformatics. 2010 Jul 1;11:364.

Week 14: Final Project Presentations