Applied Genomics: An Introduction to Bioinformatics and Network Modeling

Course Description

This course will introduce fundamental methods of analyzing large data sets from genomics experiments. Through a combination of lectures, hands-on computational training, and in-depth discussions of current scientific papers, students will learn the conceptual foundations of basic analytical methods, the computational skills to implement these methods, and the reasoning skills to read critically the primary literature in genomics. Analysis will focus on data from genome-wide studies of gene expression and molecular and genetic interactions. Methods covered will include clustering, multiple-hypothesis testing, and network inference. The course will also be open to advanced undergraduates by permission of the instructors. A large part of the course is dedicated to students completing an individual project that will be tailored to meet their background and training.

Aims of Course

Students who intend to work on wet bench problems or pursue biomedical careers need to understand and, in some cases, incorporate bioinformatics techniques into their work because of the prevalence of large datasets in all aspects of biological and biomedical research. The aim of this course is prepare both first or second year graduate students and upper-level undergraduate students to gain a hands-on understanding of computational research. The course is not necessarily designed for computer science students but is geared toward students who may pursue experimental or clinical careers. In addition, the course is appropriate for computer science students with some biology background who wish to improve their skills in translating biological problems into computational approaches. The course is based on the premise that biological and computational research is now highly intertwined.

The first goal of the course is to prepare students to read and understand the bioinformatics literature and work with large scale datasets. Five basic topics in bioinformatics will be explored in depth (see “Course Modules” in the attached syllabus). Toward this aim, students will extract data from these publications and learn to perform the tests presented in papers, apply different tests to the same data, and explore the assumptions of the original analyses. Thus, the course structure moves from lecture to paper discussion to data manipulation and, finally, a computational project. The second goal of the course is to familiarize students with a high level programming language (R) that will enable them to perform basic analyses on large datasets.

On a conceptual level, the course will teach students to synthesize data from the literature and devise novel computational experiments to test a new idea or hypothesis. For the final project, students will be required to read the primary biological literature, selection of a biological problem and an available dataset to work with, and design an algorithm to test that problem. Finally, students will present their projects to instructors and fellow students to improve their presentation skills.

Grading Scheme

35% class participation
25% midterm exam
40% final project

Syllabus
Applied Genomics: An Introduction to Bioinformatics and Network Modeling

Instructors: David Gresham and Ken Birnbaum

Course Number: G23. 1130

Credits: 4

Lecture/Computational lab/Discussion: Thu 9:30-12:15 p.m.
Recitation: 3:30-4:30 p.m.

Week 1 - KB

Class 1 (9/8)	**Lecture:** Introduction to genomic scale data. First steps in using microarray and RNA-seq data. The importance of normalization.
Class 2	**Computer Lab** Skill Building: The R environment, Importing and parsing row by column data, matrix manipulation, calling functions in R,
Recitation 1 (9/12)	**Computational Exercises:** Build a matrix, write a simple script to index specific arrays. Practice manipulating matrices and calling functions.
Assignment	Sugino et al (2006).

Week 2 -- KB

Class 3 (9/15)	**Lecture:** Grouping Data: Hierarchal Clustering and K-Means Analysis.
Class 4	**Computer Lab** Importing raw microarray data from Sugino et al (2006), normalizing data using RMA.
Recitation 2 (9/19)	**Computational Exercises** R skill building, control statements. If, for, while.
Assignment	Alizadeh et al. (2003)

Week 3 -- KB

Class 5 (9/22)	**Discussion: "Neurons and Taxon:"** How data clustering has been put to use in diverse fields. Do we approve?
Class 6	**Computer Lab** Skill Building: Implementing clustering routines on Sugino data in R
Recitation 3	**Computational Exercises**: Clustering continued.
Assignment (9/26)	RNA-seq paper (Birnbaum).

Week 4 -- KB

Class 7 (9/29)	**Discussion: What does RNA-seq tell us that microarrays did not?**
Class 8	**Computer Lab**: Working with RNA-seq data in Galaxy and mapping and normalizing issues.
Recitation 4	R skill building continued
Assignment (10/3)	Handout: Introduction to Statistical Analysis; Tusher et al. (2001)

Week 5 -- KB

Midterm Exam 10/6	**In class midterm; notes are allowed in any form (electronic or paper); no communication in any form is allowed (e.g., no phone, no text, no talking)**

No Recitation (Columbus Day Holiday)

Module III: Hypothesis Testing

Week 6 -- KB

Class 9 (10/13)	**Lecture** Basic statistical problems in gene expression analysis; multiple testing, and permutation testing.
Class 10	**Computer Lab** Use of R's statistical functions for gene expression analysis and the design of bootstrapping routines.
Recitation 5	**Computational Exercises:** Further implementation of statistical functions and use of permutation routines.
Assignment	Subramanian et al., Storey and Tibsharani

Week 7 – DG

Class 11 (10/20)	False discovery rate, Gene Ontology, Gene set enrichment analysis, Fisher's exact test, hypergeometric distribution
Class 12	**Computer Lab:** q-values, GO graphs, statistical tests for enrichment analysis
Recitation 6	**Computational Exercises:** Gene set analysis in R
Assignment	Johnson et al Bussemaker et al

Week 8 -- DG

Class 13 (10/27)	**Lecture:** ChIP-seq analysis, motif analysis
Class 14	**Computer Lab:** ChIP-seq and motif analysis in R
Assignment	**Review of Midterm:** Prepare a reading list for final project research
Reciation 7	**Motif analysis in R**

Module IV: Networks (static properties)

Week 9 -- DG

Class 15 (11/3)	**Lecture** Representation of biological networks as directed or undirected graphs. Global measures of network structure (connectivity, betweeness, etc.)
Class 16	**Computer Lab** Skill-building: Graph representation, file I/O
Recitation 8	**Computational Exercises** Use of Cytoscape to represent and navigate networks
Assignment	Research reading list: papers 1- 5

Week 10 – DG & KB

Class 17 – DG & KB (11/10)	**Student Presentation of Project Proposals – Class/KB/DG feedback and critique**
Class 18 -- DG	**Paper Discussion** Properties of Biological Networks: Jeong et al. (2001) and Shen-Orr et al. (2002) **Computational Lab** Skill Building: Finding and analyzing network motifs
Recitation 9	**Computational Exercises** Assessing connectivity in networks
Assignment	Research reading list: papers 6- 10 **1-page final project proposal**

FINAL PROJECT PROPOSAL DUE

Module V: Networks (dynamic properties)

Week 11 – DG

Class 19 (11/17)	**Lecture** Implications of local circuitry for network dynamics. Modeling network dynamics with differential equations, Boolean logic or difference equations.
Class 20	**Computer Lab** Dynamics in Boolean network models
Recitation 10	**Computational Exercises** Programming Boolean circuits. Albert and Othmer (2003).
Assignment	Work on final project: programming

THANKSGIVING BREAK

Week 11

Class 21 (12/1)	**Discussion** TBD
Class 22	**Computer Lab** TBD
Recitation 11	**Computational Exercises** TBD
Assignment	Work on final project: programming

Week 12 – KB & DG

Class 23 (12/8)	Troubleshooting: Work on computational projects with instructors
Class 24	Troubleshooting: Work on computational projects with instructors
Recitation 12	Troubleshooting: Work on computational projects with instructors
Assignment	Finish final project, validation

Week 13 – DG & KB

Class 25 (12/15)	Presentations 1-4
Class 26	Presentations 5-8
Recitation 13	Finish or correct final projects
Assignment	Annotate programs, write up paper

Week 14 – DG & KB

Class 27	Presentations 9-12
Class 28	Presentations 13-16
Recitation 14	Finish and correct final projects
Assignment	Annotate programs, write up paper

Assigned Readings:

Week 1 & 2

Sugino K, Hempel CM, Miller MN, Hattox AM, Shapiro P, Wu C, Huang ZJ, Nelson SB. (2006) Molecular taxonomy of major neuronal classes in the adult mouse forebrain. Nat Neurosci. (1):99-107.

Week 3 & 4

Alizadeh et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 403(6769):503-11.

Handout: Introduction to Statistical Analysis

Week 5 & 6

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.

Week 7 and 8

Storey, JD & Tibshirani R. (2003). Statistical significance for genomewide studies. *PNAS* 100:9440-5

Subramanian et al., (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102:15545-50

Johnson DS, Martazavi A, Myers RM and Wold B. (2007). *Science* 316:1497-1502

Bussemaker HJ, Li H, Siggia ED. (2001) Regulatory element detection using correlation with expression. *Nature Genetics.* 27: 167-71

Week 9 & 10

BarabásiA-L and Oltvai ZN, 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5:101–113. (background)

Jeong H, Mason SP, Barabási A-L and Oltvai ZN, 2001. Lethality and centrality in protein networks. *Nature* 411:41–42.

Shen-Orr SS, Milo R, Mangan S and Alon U, 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics* 31:64–68. (this module focus on static measures; transition to dynamics in next module)

Texts and Software:

Software: R

- www.r-project.org
  - Main website for R project
- http://lib.stat.cmu.edu/R/CRAN/
  - One of the mirrors where you can download R
- http://rstudio.org/
  - A recommended programming environment for R

R Programming Manual

http://manuals.bioinformatics.ucr.edu/home/programming-in-r

No special computer hardware is required for this course but convenient access to a PC or McIntosh computer is highly recommended.