BONNEAU LABORATORY

Structure Prediction | Protein Design | Biclustering + cMonkey | Network Inference | Social Networks | Publications | Links + Collaborations | People | Opportunities | Software + Code | Teaching

Bioinformatics and Genomes (Gs23.1127)

Course Links

External

Course Wiki

Internal

Lecture Structure

Grading

Course Project

Participation

Topics

Textbooks

Past Papers

Introduction

The recent explosion of the availability of genome-wide data has also let to a vast increase in bioinformatics research and tool development. Bioinformatics is becoming a cornerstone for modern biology, especially in fields such as genomics. It is thus crucial to for biologist and computer scientists interested in biology to understand the basic ideas and to learn fundamental bioinformatics techniques. Emphasis will be given to not only provide understanding of existing tools but also to provide understanding of underlying statistical and algorithmic principles such that students will be able to solve new problems in creative ways.

This course is one of three courses currently offered by the center for competitive functional genomics as a joint venture between Courant and the biology dept. This course will focus on sequence based methods for aligning proteins and finding small functional sequence motifs (TF binding, microRNA binding sites, etc.), phylogeny (building species trees, phylogenetic footprinting, etc.) and finally (if we have time) methods for modeling metabolic flux that scale to very large networks.

My goals are to get as many of you to understand this stuff, and I'm going to place a bigger emphasis on getting the class to understand this as a whole (and on building a shared foundation) than on covering every last bit of current systems biology. Thus we'll assume very little knowledge going into the class and build a common framework. We start with biological sequence analysis as this gives us a principled and well developed place to start. Feedback from you is very important.

Biologists with little CS background should take this class.

CS grad students with little Biology should take this class.

Trust me.

Structure of Lectures

Each lecture will adhere roughly to:

- 11:45 - 12:15: math intro for day's topic
- 12:15 - 1:05: lecture on day's method
- 1:05 - 1:15 : break
- 1:15 - 1:45 : discussion

Each lecture one or more of the people taking the course for credit will be charged with leading the discussion and posting a written copy of their notes for that days lecture and discussion. Each person will only have to do this once, and depending on attendance, multiple people may share a day.

Grading

Grading this course poses some philosophical problems for me due to the fact that there will be such a wide variation in the incoming skills. I want CS people to learn some specific applications and get more in tune with the main bio-relevant problems and I want bio people to learn some of the algorithms behind tools they use every day. In the end, you'll be graded on not your comprehensive mastery of all of these topics but on your progress towards your own intellectual goals, this is a grad class after-all. SO, when I ask you what you want out of this class on the info sheet give a careful answer.

- Participation: 40%
- Project: 40%
- Discussion Leader Notes: 15%
- Honesty / politely declaring your confusion when something is unclear: 5%

Course Project

Class will divide into teams. Teams will depend on number of students enrolled. Those auditing the class will also be divided into Audit teams and can do fake-o projects if they wish. The project is quite open ended so that we can tune the project to the goals of each team, but should be one of two kinds:

1. Algorithmic : Team will implement / experiment with/develop new algorithm. Examples would include implementation of a motif finding algorithm, or modification of an alignment algorithm.
2. Tool pipeline / Tool Use : Take problem relevant to research of one or more team members and use existing tools to solve the problem OR organize several tools into a pipeline to automate slow step in analysis.

Given that we have CS and Bio people in the same class we'll iteratively discus your team's project until we find something of correct scope and difficulty.

Project Timeline

Lecture 1: Make teams such that biology and CS skills are (ideally) evenly distributed.

Lecture 2-4: Hand in single paragraph describing general topic (ideally in line with a topic for which i have a lecture planned. At this point we will also discuss what might be required in terms of CS/programming introduction+review OR other background info needed for project. People with less CS experience should make sure to check in with me early to discuss an appropriate intro to scripting languages, unix, R, or whatever makes sense given their knowledge,project and team. What you need to review will depend on your team.

Lecture 5: Project map due. This should be a short document describing your project, including what language(s) you wish to use, Data sets you will need, nd expected outcomes.

Lecture 13: Project Due. Each team will demonstrate their project to instructor.

Participation

The last portion of each lecture will be a discussion. In addition each student will be responsible for leading one discussion. Discussion leaders will prepare notes on the discussion they lead that will represent 15% of the points for this class. Participation is not based on wizardly mastery of all topics but on honest feedback.

Topics / Schedule

(notes may change or be out of sync with lecture number)

lecture 1 : Organization of genomes and Fundamental bioinformatics sequence analysis techniques.

lecture 2 : Sequence Alignments I: Smith Waterman algorithm, BLAST, PSI-BLAST, Fasta

lecture 3 : Sequence Alignments II: Hidden Markov Models, HMMER.

lecture 4 : Sequence Alignments III: conserved domain databases (COG + Pfam).

lecture 5: Protein structure prediction. Secondary Structure Prediction. Transmembrane Helix prediction. Fold recognition.

lecture 6: Gene finding. Repeats, RepeatMasker.

lecture 7 : Transcriptional Regulation I: basic ideas

lecture 8 : Transcriptional Regulation II: cross-species comparisons, transcription regulator binding motifs.

lecture 9 : Translational Regulation III: microRNA gene predictions

lecture 10: Translational Regulation IV: microRNA targets (includes RNA secondary structure predictions) lecture 11: Phylogeny I. Building species trees.

lecture 12: Phylogeny II.

lecture 13: Flux balance analysis I. methods for global modeling of metabolism and signaling.

lecture 14: Flux balance II. & Description of Projects

Textbooks

Required: Durbin, Eddy, Krogh, Mitchison "Biological Sequence Analysis", Cambridge University Press, 2002, ISBN 0521629713

Suggested: T. A. Brown "Genomes" 2nd edition, paperback, ISBN:1859960294, Wiley-Liss 2002

Past Papers for Class Review

- Improved Tools for Biological Sequence Comparison
- Gapped BLAST and PSI-BLAST: a new generation of protein database search programs
- Prediction of Translation Initiation Sites on the Genome of Synechocystis sp. Strain PCC6803 by Hidden Markov Model
- A hidden Markov model for predicting transmembrane helices in protein
- Hidden Markov models
- Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments
- Evolutonary Trees from DNA Sequences: A Maxiumum Likelihood Approach
- Approaches to microRNA discovery
- Large Punctuational Contribution of Speciation to Evolutionary Divergence at the Molecular Level
- Pfold: RNA secondary structure prediction using stochastic context-free grammars