**Syllabus for STA5934 Statistical Genomics**

**Course Goals**

This course will give an introduction of the types of data in high throughput genomics experiments and statistical problems in analyzing such data. Basic ideas of key methods will be developed with considerable attention to analysis of large scale public accessible data (sequences, structures, gene expressions, SNPs…). Students should gain sufficient background to start exploring their own research questions in the area. Projects are open problems from currently actively studied topics and designed to explore how to extend current methods to novel questions with an objective to experience fruitful cross-disciplinary work.

**Target Students**

This course is aimed at statistics graduate students with interests in genomics and biological graduate students who want to learn statistical methods used in genomics.

**Teaching Approach**

Course will have a fairly fixed syllabus (below) with lectures. Reading and smaller assignments will be given for each segment. Larger term assignments will be collaborative projects in subject area of interest to student teams, leading to a paper and presentation.

**Course Outline**

Below is an outline of topics that will be covered.

Introduction to Biology

Central dogma: DNA/RNA/proteins/traits

Recent massive high-throughput technologies

Statistical issues commonly encountered in genomics and genetics

Biological sequence analysis

 DNA sequence analysis

 Protein sequence analysis

 Hidden Markov Models (HMM)

Gene transcription regulation and regulatory motif finding

 Gibbs sampling and related approaches

 ChIP-chip experiments and data analysis

Comparative genomics

 Gene annotation

 Structural genomics, structure alignment, protein function prediction

 UCSC genome browser

Single-nucleotide polymorphism (SNP) and association studies

High throughput -omic data analysis including Microarray data analysis

Normalization/pre-processing and data smoothing

Multiple testing and false discovery rates

Machine learning

Discriminant gene analysis

Analysis for emerging biotechnological -omic experiments

ChIP-chip, expression tiling, CGH, CSI

Gene selection and grouping

Biological Networks

 Gene regulatory networks

Other biological networks such as metabolism networks, protein-protein interaction networks

Phylogeny & Trees

**Projects**

In each project, students will review current literature, propose their own approach to the problems, work on the project, and present the result of their work.

Project 1. Epitope-Antibody Recognition (EAR) Challenge

<http://wiki.c2b2.columbia.edu/dream/index.php/D5c1>

Project 2. Network Inference Challenge

<http://wiki.c2b2.columbia.edu/dream/index.php/D5c4>

**Tentative Schedule**

|  |  |  |
| --- | --- | --- |
| Week | Tue Lecture | Thu Lecture |
| 1(8/23) | Introduction to Biology I | Sequence Data Analysis - Dynamic Programming |
| 2(8/30) | HMM I | Project I Literature review |
| 3(9/06) | Labor day | HMM II |
| 4(9/13) | HMM III | Regulatory motif finding |
| 5(9/20) | Project I Proposal  | High-throughput experiment |
| 6(9/27) | Microarray Data Analysis I | Microarray Data Analysis II |
| 7(10/04) | Microarray Data Analysis III | Microarray Data Analysis IV |
| 8(10/11) | Project I presentation | Project I presentation |
| 9(10/18) | Bayesian Networks | Bayesian Networks |
| 10(10/25) | Association studies  | Association studies |
| 11(11/01) | Association Studies | Protein Structure Comparison and Alignment  |
| 12(11/08) | Project II Literature review and Proposal | Project II Literature review and Proposal  |
| 13(11/15) | Protein structure prediction | Comparative Genomics, Structural Genomics, Function Prediction |
| 14(11/22) | Biological Networks | Thanksgiving |
| 15 (11/29) | Project II Presentation | Project II Presentation |
| 16(/) |  |  |