GNET 744 – Biological Sequence Analysis, Protein Structure and Genome-Wide Data

This course provides an introduction to basic protein structure/function analyses, sequencing informatics and macromolecular structure analysis. In the second half the focus will switch to analysis of genome wide data sets and methods used for the analysis of "big data". Topics covered include: genome databases; sequence retrieval, alignment and analysis; and macromolecular structure retrieval, visualization and analysis, and next-generation sequencing.

This course is designed for a first- or second-year graduate student who is interested in an introductory course in bioinformatics and structural biology computational tools and analysis of genomic data. The goal of this course is to give the student a basic understanding of available tools, but will not fully train them in the application of any one of the tools. At the end of the module, the student should know which computational tools are available, how they might be applied to a variety of research projects, what a typical analysis would consist of, and what results are expected from the analysis. We will also touch on some available resources for students who need further training in any of the computational analyses.

Class description

This class is designed to present the fundamentals of bioinformatic analyses utilizing a combination of lectures and practical "hands-on" computer instruction. In the first three weeks of the class, Dr. Brenda Temple will cover fundamentals of sequence analysis (DNA and protein) and the use of sequence data to infer phylogenetic relationships. Dr. Temple will also cover the basics of macromolecular structural analysis and visualization. Dr. Temple will conclude her section with a primer on how phylogenetic relationships can be utilized for determination of protein structure and for inference of function. In the second part of this module, a guest speaker will focus on analysis of "genome wide" data starting with "next-generation sequence" analysis technologies. This will be followed by microarray data. Genome re-sequencing, RNA-seq, ChIP-seq and their practical applications in biomedical research will be discussed.

Key learning objectives include: 1) Utilization of web-based bioinformatics, genomics and macromolecular structure databases; 2) Identification of homologous proteins/genes using sequence similarity searches; 3) Deriving evolutionary relationships from multiple sequence alignments and phylogenetic analyses; 4) Visualization of macromolecular structures; 5) Combining sequence and structural analyses to derive functional inferences; 6) Alignment, analysis and visualization of genomic data generated through short-read (next-gen) sequencing efforts; 7) Determining the sequence of DNA fragments bound to proteins in a living cell (ChIP data); 8) Identifying DNA sequence variants; and 9) Analysis of expression data (RNA-seq).

The student enrolled in this course should already have a basic understanding of the role of protein and DNA polymers in biochemistry. This should include basic knowledge of amino acids and nucleotides, protein and DNA sequences, macromolecular structure, and how protein and DNA function in a living organism. We will review some material related to amino acids, sequences and structure in class, highlighting relevant issues, but the material will not be taught in any great depth.

Class Syllabus (subject to change)

Calendar

Date (2014)		Instructor	Topic
	M	Brenda Temple	NAR, GenBank, Uniprot
25-Mar	Tu	Brenda Temple	Protein Data Bank (PDB)
26-Mar	W	Brenda Temple	PyMOL Molecular Visualization *In-Class Assignment (5 Points)*
31-Mar	M	Brenda Temple	Pairwise Sequence Similarity Searches *Assignment 1 (25 points) [Due April 10]*
1-Apr	Tu	Brenda Temple	MSA *In-Class Assignment (5 Points)*
2-Apr	W	Brenda Temple	Phylogeny *In-Class Assignment (5 Points)*
7-Apr	M	Brenda Temple	PSSM-HMM-Domains *In-Class Assignment (5 Points)*
8-Apr	Tu	Brenda Temple	Predictors *In-Class Assignment (5 Points)*
9-Apr	W	Brenda Temple	Molecular Evolution *Assignment 2 (50 Points) [Due April 19]*
14-Apr	M	Joel Parker	Next-generation sequencing intro
15-Apr	Tu	Joel Parker	Short Read DNA Alignment & Evaluation
16-Apr	W	Joel Parker	RNA-sequence analysis
21-Apr	M	Joel Parker	RNA-sequence Analysis *Assignment 3 (25 Points) [Due May 3]*
22-Apr	Tu	Joel Parker	DNA variant identification and analysis
23-Apr	W	Joel Parker	ChIP-chip/ChIP-seq Peak Calls

- NAR, GenBank, UniProt
  - Nucleic Acids Research Database of Databases
  - NCBI, Entrez, GenBank
  - UniProt
- Protein Data Bank (PDB)
  - PDB Database
  - X-Ray Crystallography
  - NMR Spectroscopy
- PyMOL Molecular Visualization System
  - Selections
  - Interfaces
  - Alignments

- Pairwise Sequence Similarity Searches
    - Homology
    - Substitution Matrices
    - Extreme Value Distributions & Statistics
- Multiple Sequence Alignments (MSA)
    - Dynamic Programming
    - Multiple Sequence Alignments
    - MSA Programs
- Phylogeny
    - Phylogenetic Tree
    - Clade
    - Tree Inference
- PSSM-HMM-Domains
    - PSSMs & PSI-Blast
    - HMMs & Domain Databases
    - Profiles & Fold Recognition Databases
- Predictors
    - Structural Neighbors Databases
    - Transmembrane Predictions & Order and Disorder Predictions
    - Structural Conservation & Fold Recognition
- Molecular Evolution
    - Gene Duplication & Neofunctionalization
    - Gα evolution
    - ODCase evolution
- Next-generation sequencing
    - Next-Gen sequencing: principle and application
    - Overview of protocols
    - Sequencing strategy
- Short Read DNA Alignment & Evaluation
    - Current alignment strategies
    - Post-alignment data processing
    - Quality evaluation
    - Data visualization using UCSC Genome Browser
- RNA-seq analysis
    - Current alignment strategies
    - Post-alignment data processing
    - Quality evaluation
    - Testing for differential expression
- DNA variant identification and analysis
    - Calling germline and somatic mutations
    - Annotation
    - Association testing and downstream analysis
- ChIP Peak Analysis
    - ChIP peak call using MACS2
    - Peak overlaps using BEDTools
    - Identify genes near peaks (BEDTools & Galaxy)
    - Search for motifs at peaks (MEME)
    - Metagene analysis

Location and time

Time: 10 am – 12 pm, MTueW
Lectures and computer lab: Biogen Idec Classroom 307, Health Sciences Library.
Materials: All learning materials will be posted on "Sakai" https://sakai.unc.edu.

Instructors

Brenda Temple, Ph.D.
Research Associate Professor
Director, R.L. Juliano Structural Bioinformatics Core
Office, 4062 Genetic Medicine Building
Office Hours, Available by appointment
919-843-9399, brenda_temple@med.unc.edu
http://www.med.unc.edu/csb/SBI/index.html

Joel Parker, Ph.D.
Research Assistant Professor
Department of Genetics
Lineberger Comprehensive Cancer Center
Office Hours, Available by appointment
919-966-9614
parkerjs@email.unc.edu

Class grade (subject to change)

80% Assignments

20% In-Class Assignments

There will be NO FINAL EXAM for this class. Late work will be accepted without penalty only if the student gets approval from the instructor before the due date of the assignment. Late work submitted within a week of the due date but without prior approval will be docked 10%. No late work will be accepted after one week past the original due date of the assignment. All work must be submitted by May 3. Students may discuss the assignments with each other if they desire, but each student must write-up their own results independently. The assignments are considered open book.

Many class sessions will consist of a combination of both lecture and "hands-on" training. Since the in-class "hands-on" computational training is so critical to acquiring the desired skills, it is important that students attend the classes and participate in the computer work. If the student has a legitimate reason for not being able to attend the classroom session, the instructor will work with the student to provide the needed training.

Homework assignments

All assignments involve computational analyses that must be completed and the results and interpretation turned in on the sakai site. Homework assignments must be submitted in portable document format (PDF). Homework assignments will be due by 6 pm on Wednesday or Friday of the following the week they were originally assigned in. You can generally expect to have the grades for the assignments available a week after the due date.

Class Examinations

There will be **NO FINAL EXAM** for this class.

Auditing and class size

Class is limited to twenty (20) students. Auditing is generally discouraged but will be considered in certain circumstances.

Recommended readings

General readings

Mount, DW. (2004). Bioinformatics: sequence and genome analysis. 2nd ed. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press

Baxevanis, AD and Francis Ouelette, BF (2004) Bioinformatics: *A Practical Guide to the Analysis of Genes and Proteins*, 3rd Edition, Wiley.

Sequence similarity

Pearson, WR. (2000). "Protein sequence comparison and protein evolution." http://www.people.virginia.edu/~wrp/papers/ismb2000.pdf

Domain architecture & databases

Ponting, CP and Russell, RR (2002). "The natural history of protein domains." *Annu Rev Biophys Biomol Struct* **31**: 45-71.

Fold recognition and structural proteomics

Baker, D and Sali, A (2001). "Protein structure prediction and structural genomics." *Science* **294**(5540): 93-6.

Thornton, JM, Todd, AE, Milburn, D, Borkakoti, N and Orengo, CA (2000). "From structure to function: approaches and limitations." Nat Struct Biol 7 Suppl: 991-4

Homology Search

Pearson W, online tutorial at http://www.people.virginia.edu/~wrp/papers/ismb2000.pdf

Pertsemlidis A, Fondon JW (2002) Having a BLAST with bioinformatics (and avoiding BLASTphemy) Genome Biology 2:reviews2002.1-2002.10

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14, 755-763.

Alignment

Nicholas HB, Ropelewski AJ, Deerfield DW (2002) Strategies for multiple sequence alignment. Biotechniques 32, 572-578.

Phylogeny

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nature Rev Genetics 4, 275-284.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) pgs 407-514 in Hillis DM, Mortiz C, Mable BK (eds) Molecular Systematics. Sinauer.

Next Generation Sequencing (NGS)

Reviews of applications of NGS: http://www.nature.com/nrg/series/nextgeneration/index.html

Meyerson M, Gabriel S, Getz G (2010). Advances in understanding cancer genomes through second-generation sequencing. Nature Review Genetics 11: 685-696.

Pepke S, Wold B, Mortazavi A (2009). Computation for ChIP-seq and RNA-seq studies. Nature Methods 6:S22-S32.

Linux Server for class

A dedicated linux server will be provisioned for this class with assistance from ITS-Research Computing. You will be provided information about the server at the beginning of the class. Logins will be based on your "onyens". Since instructors are going to have you work with "real" data sets these files are going to be very large. Most of the applications that you will be using will be compiled and pre-installed on this sever. This server will remain available through the duration of the course and will allow you to work outside the class hours for assignments/projects.

Additional Software (Most of this will be installed on classroom PC's)

NOTE: This list of programs is only provided as a guide in case you want to install them on your personal PC/Mac.

Programs are pre-installed on the computers in the library classroom. Versions below are for mainly for PC's but a number of programs have Mac versions available.

Students are expected to the use the computers in the library (and the class server) for the "hands-on" labs.

Cluster: http://rana.lbl.gov/EisenSoftware.htm

Java Treeview : http://sourceforge.net/projects/jtreeview/files/

PyMol: http://pymol.org/

ClustalX: Get ClustalX1.83.XP.zip from ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/

Cn3D - http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dinstall.shtml